ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
MSc Data Science & Marketing Analytics

Date : 25/07/2023

Main Drivers that effect Housing Price: Cellwise vs Casewise Outliers

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Author: Panagiotis Kontogiannis 616994
SUPERVISOR: DR. AURORE ARCHIMBAUD
SECOND ASSESSOR: DR. RADEK KAPRIENKO

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

Θα ήθελα να ευχαριστήσω τον μικρό μου αδερφούλη Θάνο που με βοήθησε στην πτυχιακή μου νοητικά ωστε να μην μάθω στον υπολογιστή μου να πετάει. Σε αγαπάω μπουκλάκια μου ομορφά. Τέλος αξίζω και εγώ ένα συγχαρητήρια που κατάφερα να ολοκληρώσω αυτό το Master και να βρω και δουλειά κατευθείαν. Μπράβο μου!

Abstract

In recent years, the real estate sector has witnessed substantial expansion, establishing itself as an integral component of many economies globally. Possession of residential property, besides satisfying a basic human need, has been progressively acknowledged as a prudent financial venture and a vehicle for securing long-term financial stability. This underlines the crucial necessity of understanding the dynamics of housing prices. To address this, a myriad of methodologies have been deployed to predict house prices and comprehend the variables influencing them, which primarily comprise the property's amenities and the quality of the surrounding neighborhood. However, inherent characteristics of real estate datasets, such as heteroscedasticity and non-normality, impede the accurate modeling of these datasets through non-robust techniques. Certain scholars have attempted to address these issues by employing robust methods. Nonetheless, these methods have their limitations, particularly concerning the types of outliers found in real estate datasets, specifically casewise and cellwise outliers. Casewise outliers represent individual observations diverging from the overall distribution, while cellwise outliers correspond to specific outliers identified within subsets defined by variable combinations. Existing robust methods predominantly address casewise outliers, neglecting cellwise outliers. Addressing this shortfall, the present study applies a robust method developed specifically for managing cellwise outliers. The dataset was acquired from Kaggle.com, and it has been previously utilized by Babb (2019). The kc_house_data dataset contains significant information pertaining to residential properties that were sold within King County, Washington state, during the period spanning May 2014 to May 2015.It contains 12 different variables for 21,601 different houses. The variables were about the characteristics of the houses and some of them were about the quality of the neighborhood the houses were located in. This study incorporated three distinct methodologies on this dataset: the Ordinary Least Squares (OLS) sensitive to outliers, the M Robust regression resistant to casewise outliers (utilizing Huber and Tukey’s loss functions), and a methodology Robust to cellwise outliers (Cellwise M Robust regression). The price-influencing variables demonstrated variations in their effects and significance between OLS and the two robust methodologies. In terms of accuracy, the M robust regression employing the Huber loss function exhibited superior performance, followed by the CRM model. Therefore, the results imply that cellwise outliers do not significantly impact the price, and overall, casewise robust methodologies appear more favorable in the real estate market context

1. Introduction

Homeownership, nowadays, is a focal point of modern living which is reflecting not only an individual's financial stability but also their social status (Kangane et al., 2021). A home is not merely a physical structure, but an investment, a financial asset, and a cornerstone of personal wealth. Consequently, the importance of this asset makes the understanding of its price valuation crucial. The price of a house is affected by various factors, including its location, size, condition, and the state of the market at a given time. Having a comprehensive understanding of how these factors influence the price of a house not only facilitates informed decision-making in buying or selling property, but also assists in predicting future trends, managing financial risk, and maximizing investment returns. A multitude of studies have been embarked upon with the goal of predicting housing prices, as well as discerning the factors that exert influence on these prices. These endeavors have employed a diverse array of methodologies; however, a prevalent limitation shared among the majority of these techniques is their susceptibility to outlier data. Anomalous data points can significantly distort the analytical outcome and bias the interpretation, thereby undermining the efficacy and accuracy of these prediction models. Consequently, the quest for a robust method that remains impervious to outliers continues to be a pressing need within the academic and professional realms of real estate data analysis.

The real estate sector presents a highly intricate and oftentimes capricious market, underpinned by a myriad of factors capable of shaping property valuations. These include, but are not limited to, the broader economic climate, mechanics of supply and demand, geographic location, and intrinsic property features. Conventional regression techniques, such as Ordinary Least Squares (OLS), may not invariably provide adequate modeling for real estate data, due to their pronounced sensitivity to outliers and influential observations. Furthermore, these traditional methods may fail to accommodate the characteristic heteroscedasticity and non-normality often prevalent in real estate data (Yu & Yao, 2017). The influence of outliers on non-robust fit can be of such magnitude as to erroneously classify certain routine observations as outliers, a phenomenon termed as "swamping" (Davies & Gather, 1993). Therefore, navigating the intricacies of this complex market requires advanced analytical tools capable of accommodating these unique characteristics.

Robust high-breakdown methods are more appropriate as they can mitigate the influence of outliers. These robust alternatives demonstrate the ability to accurately estimate the parameters of the assumed model, even when a minority portion (i.e., less than 50%) of the data exhibits considerable deviation from said model. Robust approaches, such the robust linear regression M-estimator with the Huber loss function (Huber, 1973), offer several advantages for modeling real estate data. Primarily, as mentioned by De Menezes et al. (2021b), they can competently manage outliers and influential observations with a higher degree of efficiency as compared to traditional methods. This leads to a more precise estimation of regression parameters and enhances the prediction accuracy of property valuations. Given the real estate market's susceptibility to extreme outliers and influential observations that can substantially affect property values, this attribute of robust methods is of paramount importance. Secondly, robust methods take into account the inherent heteroscedasticity and non-normality frequently observed in real estate data. This is crucial, as the assumption of homoscedasticity and normality is often violated in real estate data sets, leading to potentially biased and inefficient estimations of regression parameters when using conventional methods. Thirdly, robust methods can yield more durable and reliable estimates of model performance measures. This aspect is integral to assessing the efficacy of real estate models and to make informed decisions related to the purchase, sale, and investment in properties, underscoring the necessity of robust data analysis in real estate market dynamics.

Despite the efforts of numerous researchers to incorporate robust methods into real estate analysis for more accurate modeling, the focus has primarily been on casewise outliers, with minimal consideration given to cellwise outliers. The main difference between casewise outliers and cellwise outliers lies in the level at which they are identified and characterized. Casewise outliers are individual observations that deviate from the overall distribution, while cellwise outliers are specific outliers identified within subsets defined by combinations of variables. A more extensive review of those two types of outliers will be reviewed in the next section. Following an exhaustive review of methodologies in the real estate field, which predominantly focus on casewise outliers, it became clear that there was a significant gap in the literature when it comes to addressing cellwise outliers. Recognizing this gap, we aimed to remedy this evident oversight by focusing our analysis on the detection and impact of cellwise outliers. The ordinary least squares (OLS) regression method is widely recognized for fitting a regression line based on all provided observations, without explicitly considering the presence of outliers. Consequently, the resulting regression line may deviate from the underlying trend, as the model aims to minimize the sum of squared differences between the data points and the regression line. In contrast, robust methods take into account the presence of casewise outliers and employ strategies to downweight their influence, aiming to mitigate their impact on the regression line. By doing so, these methods disregard extreme data points to a certain degree, allowing for a better fit of the regression line to the patterns exhibited by the normal observations. It is important to note that there exists a trade-off in the prediction accuracy between inlying cases (non-outliers) and outlying cases. Robust regression, specifically designed to address cellwise outliers, aims to strike a balance between these two types of cases. It places emphasis on downweighting extreme cells that contribute the most to the outlying cases, rather than treating all cases equally. This approach allows for a more targeted and nuanced handling of outliers, offering improved model performance compared to both OLS regression and robust methods that do not account for cellwise outliers. In this research, we endeavor to critically assess the influence of cellwise outliers within the real estate market. We aim to investigate whether a model designed to specifically counteract these outliers yields more accurate results compared to existing robust models. Our study goes beyond just the recognition and handling of these cellwise outliers. We also aspire to illuminate the variances in how different housing variables contribute to the house's pricing structure. This comprehensive analysis will not only allow us to measure the impact of these variables but will also provide a clearer understanding of their individual contributions in shaping the price of a house.

Taking those into consideration, we will apply four distinct analytical techniques to a comprehensive real estate dataset. Initially, we shall employ the widely recognized linear regression, a non-robust approach highly sensitive to the influence of both casewise and cellwise outliers. Subsequently, we will apply the M-estimator method with the Huber loss function, designed to manage the presence of casewise outliers. Finally, the Cellwise Robust M Regression method developed by Filzmoser et al. (2020) will be utilized, thereby incorporating the effect of cellwise outliers into our analysis. The primary aim of this research is to examine the distinct impact of casewise and cellwise outliers on the resultant models. Further, we endeavor to determine which parameters exert the most significant influence on property prices according to these models and whether these determinants diverge from those identified by previously developed non-robust models. This intricate study offers valuable insight into the myriad factors influencing property prices and the distinct role of outliers in shaping these analytical models, thereby enriching our understanding of the complexities inherent to the real estate market. Thus, the main research question is :

"What differences are observed in the estimation of relationships (regression coefficients) between independent and dependent variables when using robust regression methods compared to ordinary linear regression in the real estate market?"

Followed by the sub-questions:

a. **Are robust methods performing better than the OLS in terms of accuracy**?
b. **Do the Robust Methods have differences in their interpretation to OLS**?

Considering computational efficiency, is it worthwhile to employ a complex model such as the Cellwise Robust Regression (CRM) in this field??
Our case study validate our initial hypothesis, that the Cellwise Robust Regression is indeed the middle ground of this tradeoff. There are differences in the coefficients between our models as expected and at the importance of every variables as all the signs of some variables. In the discussion section we will highlight in which circumstances each methodology should be used and the proposed framework that should be followed in case there is need to apply the CRM in a dataset with high variability. The limitations of this study and we will provided as well as the proposed extension of these findings.
The organization of this study is as follows. Section 2 delves into the findings of previous studies, presenting key determinants influencing property prices as identified in earlier research, and elucidating the theory underpinning the concept of outliers. Section 3 elucidates the fundamental theoretical concepts undergirding the methodologies employed in this study, thereby offering a comprehensive understanding of the analytical tools at hand. The data utilized for this research is thoroughly detailed in Section 4. Section 5 unveils the results derived from the developed models, focusing on their interpretability and precision. A comparative analysis between the models is also presented to highlight their differential strengths and shortcomings. Finally, Section 6 provides a conclusive synthesis of the research findings, while Section 7 proposes directions for future investigation, acknowledges limitations of the current study based on the gleaned insights, and suggests potential strategies to address these limitations in subsequent research.

2. Literature Review

As we mentioned before, initially the studies that have been conducted in this field were utilizing non-robust methods. Once more people were starting to do more research on the field, the existence of outliers was acknowledged and the need to counter them emerged. Thus, robust methods started to take the place of being more sensitive to outliers' methods. Firstly, we are going to understand what the outliers are, the types of them and how they are affecting our models. Then we will go through most of the models that have been developed, both robust and non-robust ones and lastly, we are going to see the main drivers that affect the price of the house from previous studies.

2.1. Understanding Outliers

The definition of outliers has been debated by researchers since the early 80s. Several definitions have emerged, but the core meaning remains the same (Papageorgiou et al., 2015). For example, Enderlein (1987) defined outliers as observations that deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism, while Barnet and Lewis (1994) defined outliers as observations inconsistent with the remainder of the dataset. The generation of outliers in a dataset can stem from a variety of sources . Measurement or data entry errors are one such source, where mistakes made during data collection or entry introduce values that are incorrect or inconsistent with the rest of the dataset (Dixon, 1953) . For example, a typographical error or a misplaced decimal point could lead to an outlier in a numerical variable. Another source of outliers is natural variation within the data. In some cases, outliers can be a result of random fluctuations or extreme values that occur due to inherent variability in the phenomenon being measured (Osborne and Overbay, 2004). These outliers may not necessarily indicate errors or problems with the data but rather reflect the inherent diversity or extreme nature of the data. Outliers can also be introduced into a dataset through data contamination or data quality issues (Cousineau & Chartier, 2010). For instance, during data merging or integration processes, outliers may arise if the data from different sources are incompatible or contain inconsistencies. Similarly, data transformation or preprocessing steps that are not properly handled can also generate outliers. Rare events or unusual cases can give rise to outliers as well. These outliers occur when certain data points deviate significantly from the normal behavior observed in the majority of the dataset. Examples include extreme values in financial transactions, abnormal medical test results, or anomalous behaviors in complex systems. Sampling or experimental design issues can also contribute to the generation of outliers. If the sampling process is not representative or suffers from biases, it may inadvertently include extreme observations that are not truly representative of the population. Similarly, in experimental studies, outliers can arise due to experimental errors, uncontrolled factors, or the presence of influential observations.

Certain outliers may deviate substantially across all variables, while others may demonstrate atypical behavior in only a few variables. Hence two types of outliers have been acknowledged, the cellwise outliers and the casewise outliers. Casewise outliers, also known as global outliers or individual outliers, refer to observations that are extreme or deviate significantly from the overall pattern of the data (Wiggins, 2000). These outliers can occur in any variable within the dataset. They are characterized by having unusually high or low values compared to the majority of the data points. In contrast, cellwise outliers, also referred to as conditional or contextual outliers, are observations that are extreme within specific subgroups or categories defined by combinations of predictor variables. Unlike casewise outliers, which are outliers across all variables, cellwise outliers manifest as outliers in specific subsets of the data based on the grouping variables (Filzmoser et al., 2020). Cellwise outliers can provide valuable insights into the relationships between variables within particular contexts or conditions. They indicate that certain subgroups exhibit different characteristics or behaviors compared to the overall population. These outliers can be of interest in exploratory data analysis and may suggest the presence of interactions or contextual effects in the data.

Consequences of Outliers

It is really important to understand the consequences of both types of outliers. Cellwise outliers can have significant consequences on statistical analysis such as t-tests and ANOVAs, as they may not provide reliable results. Osborne and Overbay (2004) noted that the presence of cellwise outliers can lead to incorrect conclusions. In addition, the presence of cellwise outliers can also lead to model misidentification in process modeling and identification, where the model parameters are estimated incorrectly. This can cause incorrect process control decisions, leading to poor predictions and undesirable outcomes (Pearson, 2002).

In contrast, casewise outliers present a different set of issues. They can have significant consequences on regression analysis as they may not provide reliable results. Osborne and Overbay (2004) argue that the presence of casewise outliers can lead to incorrect predictions and conclusions. In process modeling and identification, the presence of casewise outliers can also lead to model misidentification. The estimated model parameters can be biased, and this can lead to poor predictions and incorrect process control decisions (Pearson, 2002).

Elaborating on the consequences of cellwise and casewise outliers, it is essential to emphasize that these outliers can significantly impact the quality of the results of the analysis. These outliers can also result in the violation of the assumptions of the statistical models used. Additionally, cellwise and casewise outliers can reduce the accuracy of predictive models by influencing the training of machine learning algorithms.

2.2. Previously Developed Models

As we mentioned before, in this field, the development of models has evolved in response to emerging challenges, including the presence of outliers in the datasets. Initially, non-robust models were commonly employed, unaware of the potential influence of outliers on the analysis. Those models were mainly developed for two reasons, for predicting the price of a house and for discovering which variables influence price the most. However, given the fact that the models were not robust to outliers, the coefficients were biased, and the accuracy of the predictions was low. Thus, as the significance of outliers became apparent, the need arose for models that were less sensitive to these extreme observations. This section aims to provide an overview of the models used in this field, highlighting the transition from non-robust to robust approaches, followed by an exploration of the main variables identified as significant drivers of house prices based on existing research.

2.2.1. Non-Robust Approaches

Early models in housing price analysis predominantly utilized non-robust methodologies, assuming that the data followed a well-behaved distribution without considering the impact of outliers. The most common of those models are Linear Regression (LR), Multiple Linear Regression (MLR), Penalized Regressions (Lasso and Ridge), Hedonic Regression (HR), Repeat Sales Method (RSM) and the SPAR method.

Linear Regression (LR) and Multiple Linear Regression (MLR): Both these models predict a dependent variable based on one or more independent variables by fitting a linear equation to observed data. The parameters of the model are estimated using the least squares method. A significant limitation of these models is their sensitivity to outliers. The least squares estimation procedure employed in LR and MLR strives to minimize the sum of squared residuals, causing outliers to exert a disproportionate influence on the estimated model parameters, thereby leading to biased results (Su et al., 2012).

Penalized Regressions (Lasso and Ridge): These regression techniques address multicollinearity problems in LR and MLR by adding a penalty term to the loss function, which shrinks the coefficients of less important features towards zero (Fu, 1998). However, similar to LR and MLR, Lasso and Ridge regressions are also vulnerable to outliers. The inclusion of outliers can distort the penalty term, leading to the suboptimal selection of features and compromised predictive performance.

Hedonic Regression (HR): This method is commonly used in real estate economics to relate the price of properties to their characteristics. While useful, HR can be heavily impacted by outliers due to its reliance on ordinary least squares estimation, similar to LR and MLR, leading to biased coefficient estimates and unreliable predictions (De Haan & Diewert, 2013).

Repeat Sales Method (RSM): RSM is typically used to construct real estate price indices. It is based on the idea that the ratio of sales prices of a property sold more than once should represent the overall price trend (Wang & Zorn, 1997). Although it mitigates unobserved property quality issues, extreme sales prices, both casewise and cellwise, can still distort the estimates.

SPAR Method: This method combines the strengths of HR and RSM, using characteristics of a property and repeat sales information to estimate price indices (De Vries et al., 2009). Like RSM and HR, it is not inherently robust to outliers, and extreme observations can significantly affect the estimation of price indices.

2.2.2. Robust Approaches

Recognizing the need for more robust modeling techniques that could mitigate the influence of outliers, researchers began adopting robust regression models. Several have been developed, among them Support Vector Regression (SVR), Decision Tree (DT), Random Forest Regression (RFR), Gradient Boosting (GB), XGBoost and Quantile Regression (QR).

Support Vector Regression (SVR): (SVR) is a powerful method used for predicting continuous outcomes that incorporates the concept of 'epsilon-insensitive loss'. Unlike traditional linear regression, SVR aims to minimize the prediction errors while being less sensitive to outliers and noisy data points. The epsilon-insensitive loss allows the model to tolerate errors that fall within a specified threshold (epsilon) and considers them as negligible, not penalizing them during the optimization process. This feature enables SVR to focus on accurately fitting the majority of data points, particularly those close to the regression line, while reducing the influence of smaller errors and outliers. The choice of epsilon is a hyperparameter that can be adjusted to control the model's sensitivity to errors. A smaller epsilon value makes the SVR more sensitive to errors, while a higher value makes it more tolerant. This adaptability empowers SVR to strike a balance between flexibility and accuracy in predictions, making it a robust tool for handling real-world datasets with heteroscedasticity and non-normality, common in real estate markets, and other domains with complex data patterns.

Decision Tree (DT) and **Random Forest Regression (RFR)**: Both these methods are non-parametric, and their mechanism of partitioning the data space makes them robust against outliers. Decision Trees split the predictor space (set of possible values X can take) into distinct regions based on conditions on the predictors. Random Forests, being an ensemble of Decision Trees, inherit this robustness (Ali et al., 2012). Extreme data points, unless they substantially affect the decision boundaries, do not impact these models' performance (Pranav Kangane et al., 2021).

Gradient Boosting (GB) and XGBoost: As ensemble methods that typically use decision trees as base learners, GB and XGBoost can handle outliers effectively due to the robustness of decision trees (Bentéjac et al., 2020). Casewise outliers have less influence on these models, especially when using tree-based learners (Zhao et al., 2019). The robustness may vary depending on the loss function used.

Quantile Regression (QR): QR predicts a quantile (or percentile) of the response variable rather than the mean, making it inherently robust against both cellwise and casewise outliers (Koenker & Hallock, 2001). By focusing on quantiles rather than means, QR limits the impact of outliers, as the quantiles of the distribution are less influenced by extreme values.

Even though those methods can handle outliers to varying extents, they may not be fully effective in tackling cellwise outliers. For instance, the efficacy of Support Vector Regression (SVR) in addressing cellwise outliers is largely contingent upon the 'epsilon' hyperparameter selection, which defines the width of the insensitive zone in the loss function. An ill-selected 'epsilon' could lead to underfitting or overfitting and may still be susceptible to high-dimensional cellwise outliers. Non-parametric methods like Decision Tree (DT) and Random Forest Regression (RFR) exhibit general robustness to outliers. However, extreme cellwise outliers can alter the partitioning of the data space, potentially resulting in overfitting. The same limitation applies to ensemble methods like Gradient Boosting (GB) and XGBoost, where extreme cellwise outliers can induce skewed splits and overfitting. Lastly, though Quantile Regression (QR) demonstrates robustness to outliers, it does not intrinsically address cellwise outliers effectively, especially if these outliers are influential and present in multiple dimensions. Additionally, the interpretability of QR models can be more complex compared to mean regression models, as QR provides estimates for different quantiles of the response variable distribution.

Throughout the utilization of robust methods, two main disadvantages become evident. Firstly, these methods primarily account for casewise outliers, neglecting the presence of cellwise outliers. Cellwise outliers, which occur within specific subgroups or categories, can significantly affect the relationships between variables and the price. Therefore, the exclusion of cellwise outliers limits the comprehensive understanding of these relationships. Secondly, certain robust methods fall into the category of "black box" models, which can pose challenges in terms of interpretability. While these models may exhibit robustness against outliers, their complex algorithms and intricate internal mechanisms make it difficult to interpret the relationships between variables and the housing price. This lack of interpretability hinders the transparent understanding of the underlying drivers of housing prices.

One technique that addresses the aforementioned concerns is Cellwise Robust M-regression (CRM). By accounting for both casewise and cellwise outliers, CRM offers a more comprehensive and robust approach to modeling the relationships between variables and housing prices. Moreover, CRM aims to balance robustness against outliers with the interpretability of the model, allowing for meaningful insights to be extracted even in the presence of outliers. Cellwise Robust M-regression introduces a distinctive approach by employing a weighting mechanism at the level of individual cells within the dataset. This methodology goes beyond considering outliers solely at the case level and instead assigns weights to individual cells, allowing for the targeted reduction of influence from outlying cells. By adaptively down-weighting both casewise and cellwise outliers, CRM provides a significant advantage when modeling data that may contain both types of outliers. More about the methodology of CRM is explained in the following Section (3.3). Consequently, CRM can offer a more nuanced and effective approach to handling cellwise outliers in high-dimensional data, thus leading to potentially more reliable model estimates and predictions in the presence of such outliers.

2.3. Main Drivers that effect the Price

The real estate market's dynamics are shaped by two principal factors: the intrinsic attributes of the property and the contextual features of its surrounding neighborhood. The characteristics of the house, encompassing its size, number of bedrooms and bathrooms, presence of amenities, and overall quality, exert a profound influence on the property's valuation. Moreover, the neighborhood's geographical location, safety, proximity to essential facilities, and responsiveness to market trends are pivotal determinants of housing prices. This section draws support from various scholarly works that have observed similar findings regarding the impact of these drivers on the real estate market. While the cited papers provide valuable insights, it is noteworthy that numerous researchers have also documented the significance of these factors in shaping housing prices.

2.3.1. Housing Characteristics

In addition to the evolution of modeling techniques, extensive research efforts have focused on identifying the main variables that impact housing prices. Various factors have been recognized as significant drivers of house prices, encompassing both property-specific and external influences. The variables that exhibit a positive relationship, whereby an increase in their value by one unit corresponds to an increase in the housing price, include:

Area in Square Meters: As observed by Neelam Shinde and Kiran Gawande (2018),the size of a house which is usually measured in square feet or meters, is one of the most straightforward factors affecting its price. Larger houses tend to be more expensive because they offer more living space, often come with more rooms, and can accommodate larger or more amenities. They also typically require more materials and labor to build, contributing to their higher cost. In the real estate market, the price per square foot or meter can be a useful measurement to compare the value of different.

Number of Bedrooms and Bathrooms: These are crucial characteristics of a house Sirmans et al. (2005), and they have been for decades as Witte et al. (1979) also obtained the same insights, that more bedrooms can accommodate larger families or can be used for various purposes such as guest rooms or home offices. More bathrooms reduce the inconveniences of sharing, especially in larger households, and can be a luxury feature in themselves. Therefore, houses with more bedrooms and bathrooms usually command higher prices.

Garage Area and Number of Cars Accommodated: Pranav Kangane et al. (2021) note that garages serve dual purposes as storage and parking space. Houses with larger garages offer more storage space and can accommodate more vehicles. This is particularly valuable in areas where on-street parking is limited or unavailable, or where adverse weather makes covered parking desirable. Therefore, houses with larger garages tend to be priced higher.

Overall Quality Rating: The overall quality and condition of a house significantly impact its price. This includes the quality of construction, the condition and quality of the interior finish, the state of repair, and the quality of the appliances and systems like heating and cooling. Babb (2019) found that higher quality houses generally command higher prices, reflecting the materials, workmanship, and maintenance of the property.

Location: According to Bourassa et al. (2006a), location is one of the most important factors in real estate. Houses in desirable locations tend to have higher prices. Desirability can depend on a variety of factors, including proximity to city centers, schools, parks, and shopping centers, as well as the quality of the local school district, the level of local services, and even the prestige associated with a particular neighborhood or area. Houses with better views or in quieter, less congested areas can also command higher prices.

Amenities: Amenities are special features or facilities that add to the comfort, convenience, or luxury of a house. This can include a wide range of features, from swimming pools and large, well-equipped kitchens to home theaters, fireplaces, and advanced home automation systems. Houses with more or better amenities typically command higher prices because they offer more comfort, convenience, or enjoyment to the homeowner (Zietz et al.,2008b).

On the other hand, there are variables that have a negative effect on housing prices. These variables include:

Age and Deterioration: While historical or vintage houses may sometimes be valued for their unique architectural features, generally older houses tend to sell for less, unless they've been substantially renovated. This is due to several reasons. According to several papers (Neelam Shinde and Kiran Gawande, 2018; Pranav Kangane et al., 2021; Bourassa et al., 2006a; Bourassa et al., 2009c) older houses might not meet current building codes, have outdated layouts and designs, or lack modern amenities and conveniences. In addition, they are more likely to need repairs and maintenance, including potentially costly updates to systems like plumbing, wiring, or heating.

2.3.2. Neighborhood Quality

Zabel's (1996) study highlighted the pivotal role of location and neighborhood quality in influencing housing prices. His research identified several variables directly linked to neighborhood characteristics that exert a negative impact on house prices. These findings were consistent with those of Sirmans et al. (2005), reaffirming the significance of these factors in the real estate market. Some of the variables negatively affecting house prices and closely associated with neighborhood attributes include:

Location in a High Crime Area: Crime rates can significantly affect property values. Houses in neighborhoods with high crime rates are often less desirable to buyers, leading to lower prices. Potential homeowners may not feel safe living in such areas, and the crime rate could also affect their property insurance rates. This goes beyond just personal safety - areas with high crime rates often also suffer from issues like poor maintenance and neglect, which can further depress property prices.

Poor School Districts: For families with children, the quality of the local school district is often a key factor in choosing a house. Houses located in poor school districts can therefore have lower prices, as they may be less attractive to these buyers. Even for buyers without children, a poor school district can be seen as a negative feature, as it could make the house harder to sell in the future.

Noise Pollution: Houses located near sources of noise, such as busy highways, airports, train tracks, or industrial areas, can be less desirable to buyers and therefore command lower prices. Noise can disrupt peace and quiet at home, disturb sleep, and generally reduce the quality of life. Noise pollution can also be a sign of other issues, such as heavy traffic, pollution, and a lack of green spaces.

2.4. Relevance - Contribution

The valuation of property holds immense relevance across diverse sectors, including real estate, stock market, taxation, and the broader economy. Four compelling reasons justify the emphasis on housing market analysis.

Firstly, the wealth effect associated with housing exceeds that of financial assets, substantiated by Case et al. (2005) and Benjamin et al. (2004). Secondly, housing, as corroborated by Englund et al. (2002) and Flavin and Yamashita (2002), constitutes the primary asset in household portfolios. Thirdly, housing market downturns wield a significantly stronger impact on the economy than stock market declines. According to Helbling and Terrones (2003), the output effects linked with housing price declines during 1970-2002 were twice as significant as those associated with equity price downturns. Furthermore, the repercussions of housing price drops supersede those of stock market crashes, with the economic slowdown posting a housing market collapse being twice as protracted. Fourthly, housing purchases are primarily driven by consumption motives, coupled with the fact that high transaction costs, the heterogeneous and illiquid nature of housing limit arbitrage. Therefore, any inefficient pricing can endure for uncertain and extended periods, making the correction towards 'true' value a gradual process.

Accurate prediction of house prices serves a plethora of stakeholders. For prospective homeowners, an accurate prediction facilitates informed decision-making regarding the property's purchase. Real estate agents and brokers can utilize these predictions to assist clients in setting reasonable prices, attracting potential buyers, and finalizing sales. Real estate developers can derive insights about construction decisions and pricing strategies by understanding price influencing factors. Other beneficiaries of this prediction include lenders, appraisers, tax assessors, government agencies, and housing market investors.

Moreover, the housing dataset in this study represents multivariate data exhibiting high variability, a common feature in the contemporary world. Consequently, this study is not merely enhancing the predictive accuracy for a vital asset, but also paving the way for fields represented by similar datasets.

In summation, this study proposes a novel approach to discern the relationship between house prices and various influencing variables. It augments existing literature on algorithms specializing in detecting cellwise outliers in real estate, an underexplored area. It is a pioneering effort to apply the CRM's package implementation of the CRM algorithm to real estate data.

The study aims to identify the limitations of current methodologies and presents a comprehensive comparison of methods sensitive to outliers, casewise outliers, and methods accounting for both cellwise and casewise outliers. Lastly, this work contributes a new technique to the current methodologies for predicting a variable in a multivariate dataset with high variability and provides guidance on its application in appropriate scenarios.

3. Methodology

For this research, we are going to use the traditional Multiple Linear Regression, the M Robust Regression with the Hubert loss function, the M Robust Regression with the Bisquared loss function and the Cellwise M Regression.

3.1. Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to model the relationship between a dependent variable and two or more independent variables. It builds upon simple linear regression, which focuses on the connection between a single independent variable and the dependent variable. In MLR, the goal is to estimate the regression coefficients that quantify the numerical relationship between the independent and dependent variables. These coefficients, when other variables are kept constant, indicate the average change in the dependent variable for a one-unit change in each independent variable. The MLR model is commonly represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{\mathrm{p}} X_{\mathrm{p}} + \varepsilon$$

Where:

- Y represents the dependent variable
- $X_1, X_2, \ldots, X_{\mathrm{p}}$ represent the independent variables
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_{\mathrm{p}}$ are the regression coefficients
- ε represents the error term, which captures the unexplained variation in the dependent variable not accounted for by the independent variables

3.1.1. Assumption of MLR

To ensure the validity and reliability of the MLR model, several key assumptions need to be met. These assumptions provide the foundation for statistical inference and accurate interpretation of the regression results. In this subsection, we will discuss and describe each assumption in detail.

Linearity: The relationship between each independent variable and the dependent variable should be linear, according to the linearity assumption in MLR. Additionally, there should be linearity in the overall relationship between the dependent variable and all of the independent variables. The pattern between the residuals and the fitted values can be represented with a straight line in the following Figure (1), demonstrating that the linear assumption is met at the initial plot. The assumption is broken, as can be seen on the right plot where the line representing this relationship is curled. This could lead to skewed standard errors, p-values, R squared, and coefficients.

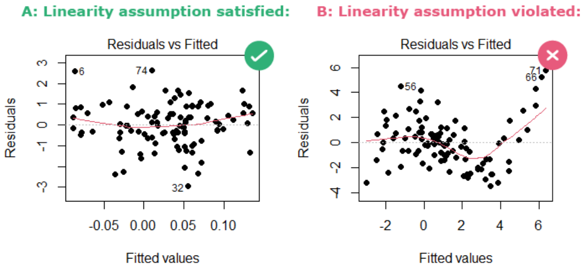


Figure 1: Linearity Assumption of the Linear Regression obtained by Choueiry (2022b)

Independence: MLR makes the assumption that the observations utilized in the regression analysis are independent of one another. In other words, the value of one observation is independent of the value of another. Two linear regression lines are depicted in Figure 2, with the points distributed randomly on the left. The points on the right are obviously impacting one another, which contradicts this presumption. Estimates of the regression coefficients may be skewed and ineffective if the mistakes are not random.

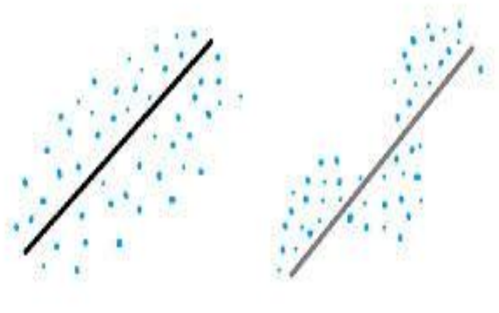


Figure 2: Independence Assumption of Linear Regression

Homoscedasticity: The term "homoscedasticity" describes the presumption that the error (or residual) variability is constant at all levels of the independent variables. In other words, regardless of the values of the independent variables, the spread of the residuals should be constant. The residuals are distributed across the range of predictors in Figure 3's left plot, with no clear trend to be seen. This suggests that the errors are dispersed normally. On the right plot, we can discern a pattern (a straight line) that suggests a non-normal distribution, though. For successful hypothesis testing and trustworthy standard errors, it is crucial to make this assumption.

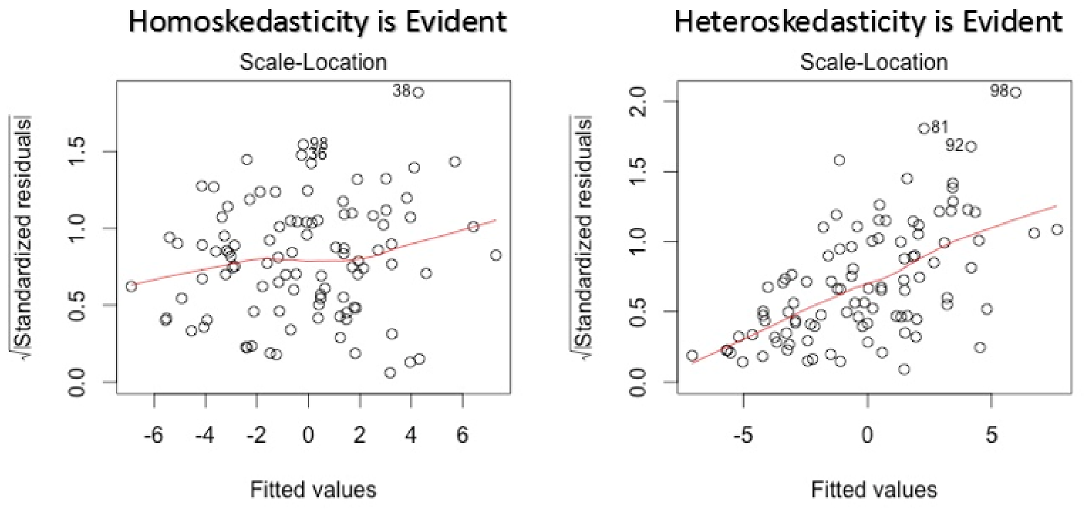


Figure 3: Homoscedasticity Assumption of Linear Regression obtained by Avcontentteam (2023b)

Normality: According to the normality assumption, the residuals in MLR have a normal distribution. For performing hypothesis tests, creating confidence intervals, and drawing statistical inferences, normality is essential. Because the first two plots on the left follow the normality line, they show that the data are normal. The initial plot includes a few minor variations, but given the volume, they are tolerable because there are no significant ones. The data in the figure on the right, on the other hand, are not normal since they vary greatly from the normality line. The accuracy of the regression findings may be impacted by deviations from normality, which may point to potential problems with the model's underlying assumptions.

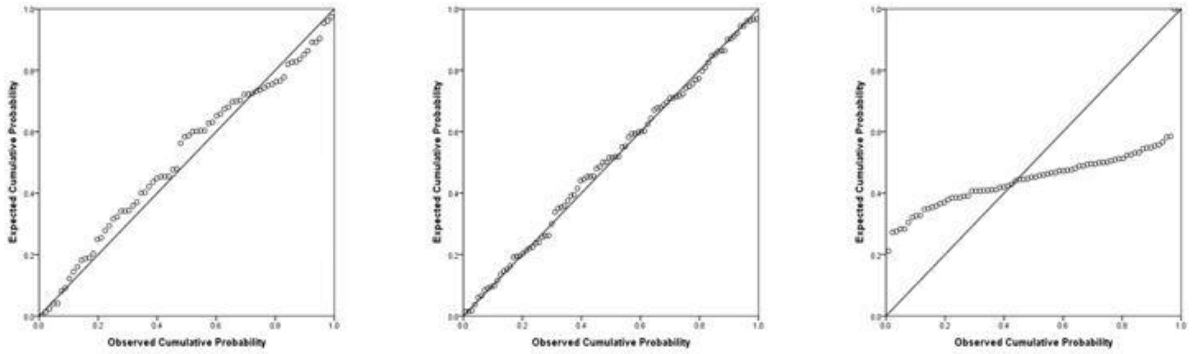


Figure 4: Normality Assumption of the Linear Regression obtained by Moran (2021)

Multicollinearity: The multicollinearity assumption in linear regression refers to the presence of high correlation among the independent variables in the model. This condition can lead to unstable coefficient estimates, making it difficult to determine the unique contribution of each variable to the dependent variable. Interpretation of coefficients becomes challenging as the effects of individual variables are confounded by multicollinearity. Additionally, multicollinearity inflates standard errors, reducing the precision and statistical significance of the coefficient estimates. It also diminishes the predictive accuracy of the model, particularly for new observations.

3.1.2. OLS fit to a sample Dataset

We constructed a representative sample dataset incorporating distinctly observable outliers. Subsequently, we produced a scatter plot of this dataset, supplemented by a simple regression line, as depicted in Figure 5. In the scatter plot, the data points are marked by dots. While the black dots represent the 'normal' observations, the red ones signify the outliers. We fitted an Ordinary Least Squares (OLS) regression line to this dataset. An important characteristic of OLS regression is that it assigns equal weights to each observation, regardless of whether it is an outlier or a regular observation. As a result, the OLS regression model becomes sensitive to these extreme data points. This sensitivity is a double-edged sword: it can provide insights into unique occurrences within the dataset, yet it also has the potential to skew the regression line and thereby distort predictions. In our sample dataset, the three extreme points (indicated in red) significantly influence the OLS regression line, highlighting the impact outliers can have on this form of regression analysis.

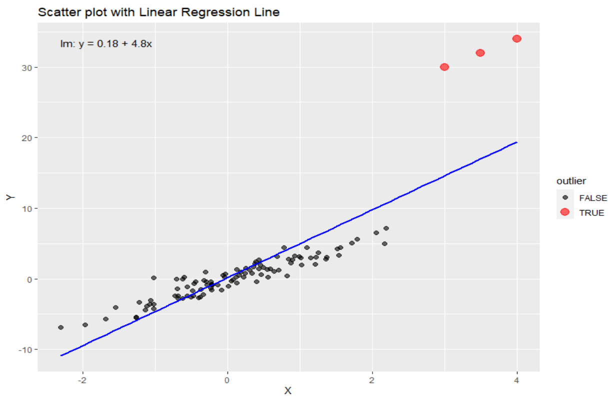


Figure 5: Fit of the OLS Regression Line as an example.
The skewness of the line towards the red dots (outliers) shows the sensitivity of the fit.

3.2. M Robust Regression

M-estimator robust regression (De Menezes et al., 2021b) is a statistical approach employed for parameter estimation in regression models when there are outliers or violations of assumptions. This technique aims to be more resilient to extreme data points and departures from normality in comparison to conventional regression methods like ordinary least squares (OLS). Its primary objective is to provide more robust and reliable parameter estimates in the presence of challenging data characteristics. The general form of the M-estimator can be expressed as follows:

$$S = \Sigma_{i=1}^{n} H(\varepsilon_i)$$

Where:

- H is the loss function
- n represents the number of observations in the dataset

The mean squared error function can be considered an M-estimator in which $H(\varepsilon) = \varepsilon^2$. In robust regression using M-estimators, the goal is to determine the regression coefficients that minimize a particular loss function applied to the residuals. This loss function quantifies the difference between the observed values and the predicted values obtained from the regression model. By selecting an appropriate loss function, the M-estimator approach can mitigate the impact of outliers and handle deviations from underlying assumptions, leading to more resilient parameter estimates. So, in the case of general form of the error function, the derivative with respect to k_{th} parameter takes the following form after applying the chain rule:

$$\frac{\partial S}{\partial \theta_k} = \Sigma_{\mathrm{i}=1}^{n} w_i \varepsilon_i x_{ki} = \Sigma_{\mathrm{i}=1}^{n} w_i \left(\theta_k x_i - y_i\right) x_{ki}$$

Where :

- S represents the objective function or the loss function being minimized in the M-estimator approach
- θ_k is the k_{th} regression coefficient being differentiated with respect to
- w_i is the weight associated with the i_{th} observation
- $\theta_k x_i$ represents the predicted value for the i_{th}observation based on the current regression coefficients
- y_i is the observed value for the i_{th}observation
- x_i represents the input or independent variable of the i_{th} observation

This expression is set to zero and solved simultaneously for all parameters. In the general case, the parameter values are obtained through an iterative reweighting procedure. This involves defining a "weight" variable as:

$$w_i = \frac{1}{\varepsilon} \frac{\partial H}{\partial \varepsilon_i}$$

So, in this case:

$$\frac{\partial S}{\partial \theta_k} = \Sigma_{\mathrm{i}=1}^{n} w_i \varepsilon_i x_{ki} = \Sigma_{\mathrm{i}=1}^{n} w_i \left(\theta_k x_i - y_i\right) x_{ki} = 0$$

In the robust regression procedure, the initial weights are set to a uniform value, typically of 1. This allows us to form a system of linear equations that can be solved for the parameter estimates, denoted as θ. With the newly obtained parameter values, the error terms are computed. These error terms are then used with a chosen residual function, denoted as $H(\varepsilon)$, to calculate a new set of weights, denoted as w. This iterative process continues until convergence is achieved. In robust regression, different loss functions can be employed as the residual function. Two commonly used loss functions are the Huber loss function and the Bisquared loss function.

3.2.1. Huber Loss Function

The Huber loss function (Huber, 1973) is a combination of the squared loss and absolute loss. It provides a compromise between the efficiency , in terms of more precise parameter estimates with smaller standard errors, of the squared loss and the robustness of the absolute loss. The Huber loss function is defined as:

$$\begin{cases} \frac{\varepsilon^2}{2} for|\varepsilon| \leq k, \\ k|\varepsilon| - \frac{k^2}{2} for|\varepsilon| > k \end{cases}$$

For errors which are not bigger than some threshold this function behaves more like the mean squared error, which ensures that the function is continuous at the origin. In case of bigger errors, the function behaves more like the mean absolute error, and the penalization of the outliers becomes proportional to their distance to the mean. As for the k value, Huber proposed 1.345 of standard deviation of a sample, which results in approximately 95% of efficiency which MSE provides. The efficiency of the Huber loss function essentially lies in its capability to strike a balance between the high precision of MSE, which is effective for smaller errors, and the robustness of Mean Absolute Error (MAE), which better accommodates larger errors. When we reference the 95% efficiency of the Huber loss function, we're speaking to its ability to deliver results that stand up to 95% of the efficiency level achieved through the least squares method. This comparative benchmark, the least squares method, is conventionally viewed as 100% efficient under specific circumstances, such as when errors display a normal distribution and exhibit homoscedasticity (constant variance across different values of the independent variables).

3.2.2. Bisquare M - Estimation

The Bisquared loss function developed by Tukey (1960) is another commonly used loss function in M-estimation. It is designed to be more robust to outliers than the Huber loss. The Bisquared loss function is defined as:

$$\begin{cases} \frac{k^2}{6}\left(1-\left(1-\left(\frac{\varepsilon}{k}\right)^2\right)^3\right) for|\varepsilon| \leq k, \\ \frac{k^2}{6} for|\varepsilon| > k \end{cases}$$

This type of function is even more robust than the Huber M-estimator. For the residuals with the values greater than some threshold (the proposed value for k is 4.685 standard deviations) the penalization remains constant.

3.2.3. Comparing Huber and Bisquare

Figure 6 illustrates the behavior of Huber's function and Tukey's estimator in dealing with residuals. Huber's function resembles least squares until the residuals reach a certain magnitude, after which their influence diminishes gradually. In contrast, Tukey's estimator immediately reduces the influence of residuals, and beyond a certain point, they have no impact at all (Fox & Monette, 2003). The weight functions of these estimators, as shown in Table 1, differ in shape and the rate at which outliers are downweighted. Huber's weight function exhibits a smoother transition from quadratic to linear loss, resulting in a gradual reduction in weight for outliers. On the other hand, Tukey's Bisquare weight function shows a more abrupt decline in weight beyond the threshold, leading to a more significant downweighting of outliers. Huber's estimator aims to strike a balance between efficiency and robustness, while Tukey's estimator is more aggressive in downweighting outliers.

Method	Objective function	Weight function
Least Squares	$H_{LS}(\varepsilon) = \varepsilon^2$	$W_{LS}(\varepsilon) = 1$
Huber	$H_H(\varepsilon) = \begin{cases} \frac{\varepsilon^2}{2} for \lvert\varepsilon\rvert \leq k, \\ k\lvert\varepsilon\rvert - \frac{k^2}{2} for \lvert\varepsilon\rvert > k \end{cases}$	$W_H(\varepsilon) = \begin{cases} 1, for\ \lvert\varepsilon\rvert \leq k \\ \frac{k}{\lvert\varepsilon\rvert}, for\ \lvert\varepsilon\rvert > k \end{cases}$
Tukey bisquare	$H_T(\varepsilon) = \begin{cases} \frac{k^2}{6}\left(1 - \left(1 - \left(\frac{\varepsilon}{k}\right)^2\right)^3\right) for \lvert\varepsilon\rvert \leq k, \\ \frac{k^2}{6} for \lvert\varepsilon\rvert > k \end{cases}$	$W_T(\varepsilon) = \begin{cases} \left[1 - \left(\frac{\varepsilon}{k}\right)^2\right]^2, for\ \lvert\varepsilon\rvert \leq k \\ 0, for\ \lvert\varepsilon\rvert > k \end{cases}$

Table 1: Objective function and weight function for the Least Square, Huber, and Tukey bisquare estimators

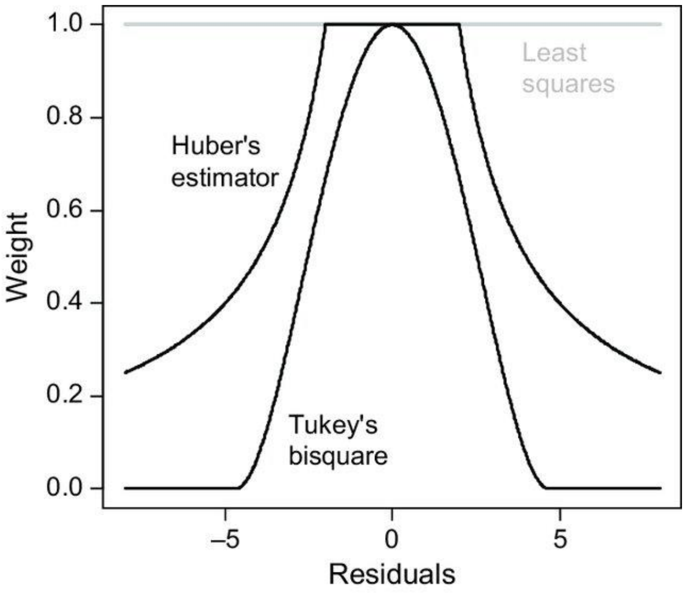


Figure 6: Comparison of Huber's to Tukey's Bisquare and Least squares Loss Functions

3.2.4. Casewise Robust Regression fit to the sample Data

Using the same sample dataset and the initial OLS regression line as referenced, we proceeded to fit an M-regression line to the data. Notably, M regression is recognized for its robustness against casewise outliers. In contrast to OLS, this model strategically downweights these outliers, resulting in less sensitivity to their influence.

As visualized in Figure 7, the M-Regression line lies notably lower than the Ols regression line. This shift is attributable to the counteractive mechanism of M-Regression again casewise outliers, downweighting their impact to minimize distortion. The resulting M-Regression line more closely reflects the pattern of the "normal" observations in the dataset. By mitigating the influence of the outliers, this model achieves a better fit, yielding a more reliable representation of the majority data points.

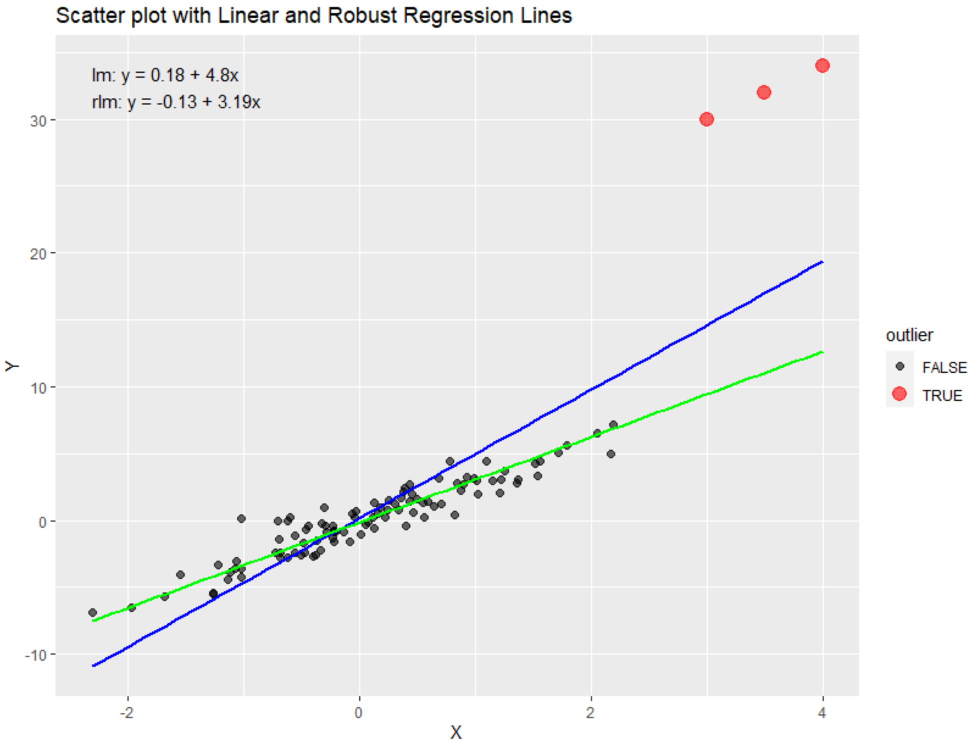


Figure 7: Fit of the Casewise Robust Regression line

3.3. Cellwise Robust M Regression

Cellwise Robust M (Filzmoser et al., 2020) is a robust regression method that effectively handles the presence of outliers or influential data points within a dataset. The fundamental objective of this methodology is to furnish dependable estimates of the regression coefficients by diminishing the impact of extreme observations. Notably, CRM distinguishes itself from other robust regression methods by accounting for both casewise outliers and cellwise outliers. By incorporating cellwise outliers into its framework, CRM enhances the reliability and accuracy of the estimated regression coefficients. This comprehensive approach ensures that the influence of extreme observations is suitably downweighted during the estimation process. Consequently, CRM is better equipped to capture the underlying relationships between variables and provide robust results.

In this section, we present an algorithm that extends cellwise robustness properties to various robust regression methods. As different robust regression methods have distinct characteristics, several details in the algorithm have been adjusted to accommodate specific regression techniques. The algorithm we describe is primarily a cellwise extension of robust MM regression, which is known for its favorable robustness-efficiency tradeoff demonstrated in theory, simulations, and practical applications (Maronna et al., 2006). The MM regression estimators involve two steps: first, a highly robust initial estimate is computed, leveraging its high breakdown point throughout the procedure. This initial estimate then serves as a plug-in estimator for an M-estimator, where it acts as a starting point for an iterative reweighting algorithm to achieve improved efficiency. We delve into the specifics of how this concept is utilized to develop an efficient and highly robust cellwise regression method, providing a detailed explanation of each essential step in the subsequent discussion. The main purposes of the CRM algorithm are to robustly estimate regression coefficients, improve model reliability and accuracy in the presence of outliers, and handle both casewise and cellwise outliers effectively. By incorporating these steps, CRM ensures that extreme observations are appropriately downweighted, enabling the algorithm to capture the underlying relationships between variables and provide robust results. The algorithm's performance is evaluated based on the Robust Mahalanobis distance, and convergence is achieved when the algorithm stabilizes and provides reliable coefficient estimates for all cases.

The initial step of the algorithm involves robustly scaling and centering the data. The authors recommend using estimators with a 50% asymptotic breakdown point, such as the L1 median and Qn scale estimator (Rousseeuw & Croux, 1993), to ensure robustness. Here the asymptotic breakdown point implies that those estimators handle up to 50% of data contamination or outliers without providing arbitrary or completely incorrect estimates. Beyond this point, the reliability of the estimator can significantly deteriorate, and it may yield faulty or misleading conclusions. Subsequently, a robust Linear Regression method like MM robust regression or the LTS estimator is applied. After obtaining the initial estimates, the algorithm proceeds by scaling the residuals using the Median Absolute Deviation (MAD) and determining the weights for each case using the Hampel weights function (Adamczyk, 2017). Cases with absolute standardized residuals exceeding the 95% quantile of the standard normal distribution are identified as outliers. The algorithm then determines whether these cases are casewise outliers or if they contain subsets of cells that exhibit outlying behavior. To address this, the SPADIMO method (Debruyne et al., 2019) is employed. For cases identified as containing cellwise outliers, the outlying variables are treated as missing cells and their values are imputed using the two nearest neighbors method based on the clean cells within the case. After these cells are pinpointed, the algorithm calculates a new value for each of these cells. This new value is determined by finding the two nearest cases (that have not been flagged as outliers), averaging the values of the two neighbors. This imputation procedure is called the Nearest Neighbors and it results in modified cases with reduced residuals, leading to increased case weights. Consequently, the valuable information contained in the non-outlying variables contributes more substantially to the model. The algorithm proceeds by replacing the new values in the outlying cells and utilizing the IRLS procedure to obtain updated coefficients and weights. The iterations are evaluated based on the Robust Mahalanobis distance and the algorithm continues until convergence is achieved. During each iteration, the residuals are recalculated, and casewise outliers are identified based on the magnitude of the residuals. For cases flagged as outliers, the variables contributing to their outlyingness are identified using the SPADIMO method. For cases that are not entirely classified as outliers, the outlying cells are imputed as previously described. The iterations persist until convergence is attained for all cases.

The CRM Algorithm step by step:

1) Robust scaling and centering the data
2) Application of a robust regression estimator
3) Calculate the case weights based on the residuals of the robust regression using Hampel Weight
4) Apply SPADIMO algorithm to detect which cells are contributing most to outlyingness of the flagged cases and impute the flagged cells.
5) Repeat step 2-4 until convergence.
6) Scale the coefficients back to the original scale

The SPADIMO Algorithm:

1) Calculate the Robust Malahanobis distance for every outlying case
2) Identify which cells are outlying and impute them with the Nearest Neighbors Method
3) Repeat steps 1-2 until there are no outlying cases anymore

CRM fit to the sample Data

Once again utilizing the same mock dataset, we applied the Cellwise Robust Method (CRM), a model designed specifically to counteract cellwise outliers. In Figure 8, we can observe a shift in the regression line, positioning it slightly above the casewise robust (M-regression) line. Interestingly, the CRM model's regression line is now situated between those produced by the Casewise and OLS models. The upward shift of the CRM model's regression line, as compared to the M-regression line, points to the model's unique approach in handling outliers. While the M-regression model specifically targets and downweights casewise outliers, the CRM model operates on a broader scope, accounting for both cellwise and casewise outliers. This behavior lends the CRM model a notable degree of flexibility, enabling it to deliver a balanced fit that aligns more closely with the data pattern as a whole, rather than being heavily skewed towards normal observations or unduly influenced by outliers. The juxtaposition of these three regression models - OLS, M-regression, and CRM - on the same dataset provides a comprehensive illustration of how different outlier-handling mechanisms can significantly alter the line of best fit.

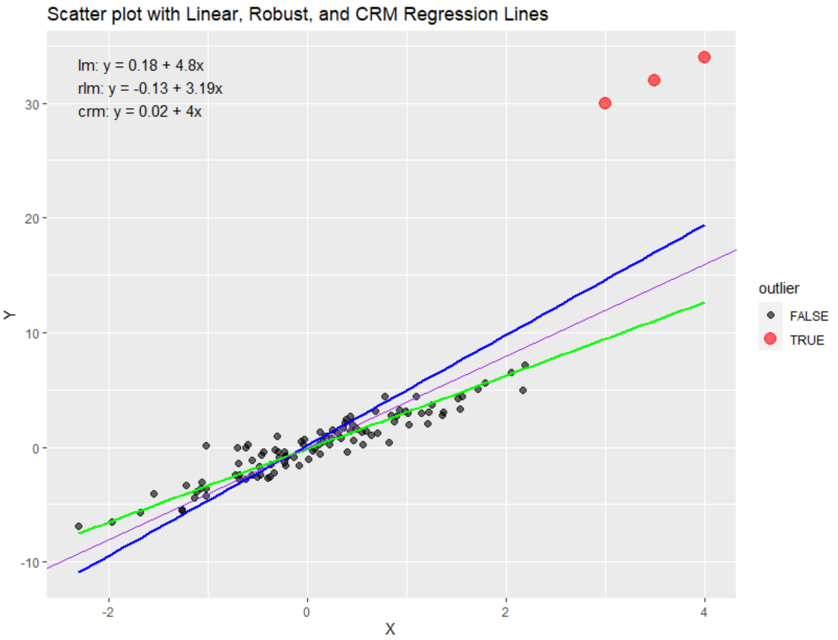


Figure 8: Fit of all 3 models

3.4. Evaluation

To assess the performance of the aforementioned methods, three metrics will be considered: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median Absolute Error (MAE).

Mean Squared Error (MSE)

Mean Squared Error (MSE) is a measure used to evaluate the average squared difference between predicted and actual values in a regression or prediction model. It provides an indication of the overall magnitude of errors or residuals in the model predictions. A lower MSE indicates better model performance, as it signifies a smaller average squared difference between predicted and actual values. The MSE is calculated using the following formula:

$$MSE = \frac{1}{n} x\, \Sigma(y_i - \hat{y}_i)^2$$

Where:

- n is the number of observations.
- y_i represents the actual observed values.
- $\hat{y}_i$ represents the predicted values for the corresponding actual values.

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a metric employed to evaluate the disparity between predicted values and actual values in a regression scenario. It gauges the average difference between the predicted and actual values, denoted in the same units as the target variable. The RMSE is computed as the square root of the mean of the squared differences between predicted and actual values. A smaller RMSE signifies superior model performance. The RMSE is derived by taking the square root of the MSE.

Median Absolute Error (MAE)

Mean Absolute Error (MAE) is a metric utilized to evaluate the average magnitude of errors between predicted values and actual values in a regression scenario. It quantifies the absolute difference between predicted and actual values, using the same units as the target variable. The MAE is calculated as the mean of the absolute differences between predicted and actual values. A lower MAE indicates better model performance. The formula for calculating MAE is as follows:

$$\mathrm{MAE} = \frac{1}{n} \times \sum |\hat{y}_i - y_i|$$

Where:

- n is the number of observations or data points.
- $\hat{y}_i$ represents the predicted values for the corresponding actual values.
- y_i represents the actual observed values.

In the Methodology section, we have elucidated the statistical methodologies that will be adopted in this study. Now, transitioning to the Data section, we will delve into the intrinsic characteristics of the real estate dataset that will serve as the foundation for our investigation. Additionally, we will perform a Descriptive Analysis to unveil inherent relationships and correlations within the dataset, enabling us to gain valuable insights into the interplay between different variables and their impact on housing prices.

4. Data

The dataset was retrieved from the Kaggle.com, an online platform for predictive modeling and analytics competitions. It has also been used by Babb (2019), whose work have been mentioned in the literature review and Sahu et al. (2019b). The kc_house_data dataset provides valuable information on residential properties sold in King County, Washington state, between May 2014 and May 2015. Table 2 presents the description of every variable our dataset contained. These variables can be used to study the relationship between the characteristics of a property and its sale price. Furthermore, sqft_living15 and sqft_lot15 can be considered as measures of the neighborhood's characteristics and can be used to investigate the impact of the neighborhood's features on the sale price of a house.

Variables	Description
id	A unique identifier for each property
date	The date the property was sold
price	The sale price of the property
bedrooms	The number of bedrooms in the property
bathrooms	The number of bathrooms in the property
sqft_living	The total living area in square feet
sqft_lot	The total lot area in square feet
floors	The number of floors in the property
waterfront	A binary variable indicating whether the property has a waterfront view.
view	An index from 1 to 5 of the quality of the view from the property
condition	An index from 1 to 5 of the overall condition of the property
grade	An index from 1 to 13 of the overall grade of the property, based on the King County grading system
sqft_above	The square footage of the house apart from the basement
sqft_basement	The square footage of the basement
yr_built	The year the property was built
yr_renovated	The year the property was last renovated, or 0 if it has not been renovated
zipcode	The ZIP code of the property
lat	The latitude of the property
long	The longitude of the property
sqft_living15	The average interior square footage of living space of the 15 nearest houses to the subject property
sqft_lot15	The average size of land lots of the 15 nearest houses to the subject property
age	The age of the property

Table 2: Description of the Variables

4.1. Data Preparation

Initially, missing values were eliminated from the dataset. Subsequently, we computed the correlation matrix to identify potential collinearity among variables. Interestingly, our analysis revealed a significant correlation among three variables: sqft_living, sqft_basement, and sqft_above. Further investigation unveiled that the sqft_living variable was derived as the sum of sqft_basement and sqft_above. Consequently, we decided to retain only the sqft_living variable, as it encapsulated the information contained within the other two variables.

In addition, several other variables were deemed irrelevant for our study and were excluded. These variables included the house identification (id), as well as the zipcode, longitude, and latitude. The waterfront variable was also removed from the dataset, as it exhibited a mere 200 observations out of the original 21,600. Additionally, we created the age of every house by deducting the year that the house was built and the year it was sold. After the data preparation stage, the dataset comprises 21,601 observations and 12 variables.

4.2. Descriptive Analysis

Table 3 provides an overview of the key characteristics of the variables in the dataset. This table reveals important insights into the remaining variables' distributions and ranges. For instance, the "price" variable exhibits a wide range, with values ranging from 7,500 to 7.7 million dollars. The majority of houses have 3 bedrooms and 2.25 bathrooms on average, with "sqft_living" ranging from 290 to 13,540 square feet. The "sqft_lot" variable shows a similar pattern, ranging from 520 to 1,651,359 square feet. The number of floors in most houses tends to be around 1 to 1.5.

The analysis of the "condition" and "grade" variables reveals that the majority of houses in the dataset exhibit a favorable overall condition and grade, with a concentration in the middle range of the respective scales. These findings suggest that the houses generally possess satisfactory qualities and meet certain standards. Further examination of the summary statistics Table 3 provides valuable insights into the distribution and skewness of the variables. Notably, the "price" variable demonstrates a positively skewed distribution, indicating a clustering of house prices below the mean value. Similarly, variables such as "bedrooms," "bathrooms," "floors," "sqft_living," and "sqft_lot" also exhibit positive skewness, suggesting the presence of longer tails on the right side of their distributions. A comparison of the mean and median values within the summary statistics Table 3 further highlights the extent of skewness present in the dataset. For instance, the median value of "sqft_lot" is 7620, while the mean value is more than twice that figure. Conversely, variables such as "floors" display relatively mild skewness, as evidenced by a median of 1.5 compared to a mean of 1.49. The observed skewness in these variables indicates the potential presence of outliers or extreme values that deviate from the majority of data points.

Variables	Minimum	Quartile_1	Median	Mean	Quartile_3	Maximum
price	75000	321500	450000	540129.5	645000	7700000
bedrooms	0	3	3	3.370909	4	33
bathrooms	0	1.75	2.25	2.11459	2.5	8
sqft_living	290	1430	1910	2079.835	2550	13540
sqft_lot	520	5042	7620	15113.19	10696	1651359
floors	1	1	1.5	1.493866	2	3.5
view	1	1	1	1.234341	1	5
condition	1	3	3	3.409657	4	5
grade	1	7	7	7.656405	8	13
sqft_living15	399	1490	1840	1986.675	2360	6210
sqft_lot15	651	5100	7620	12772.46	10086	871200
age	0	18	40	43.34244	63	115

Table 3: Summary Statistics

Upon examining the distribution of the price (Figure 9a), it is evident that it exhibits that it is skewed to the right. This is indicated by a longer tail on the right side of the distribution, with a few properties having significantly higher prices compared to the majority. The majority of house prices tend to be concentrated towards the lower end of the scale, while a smaller proportion of houses command much higher prices. The distribution of the price variable follows a positively skewed pattern, implying that the mean price is higher than the median which can also been seen from Table 3. This suggests that the presence of a few expensive properties significantly influences the average price.

Furthermore, an interesting observation can be made regarding the distribution of the "View" variable (Figure 9b). Table 3 corroborates this finding, as both the first and third quartiles (Q1 and Q3) are identical, indicating that the variable's value changes exclusively within the fourth quartile. Upon examining the distribution, it is notable that the majority of properties in the dataset have a view rating of 1. Specifically, among the 21,600 houses, nearly 19,500 of them do not have a notable view. The distribution of the view variable is highly skewed towards the lower view ratings, resulting in an imbalanced distribution. Such a high concentration of observations in a single category raises concerns about the variable's impact on the robust scaling of the CRM model. Consequently, this variable was excluded from the subsequent models.

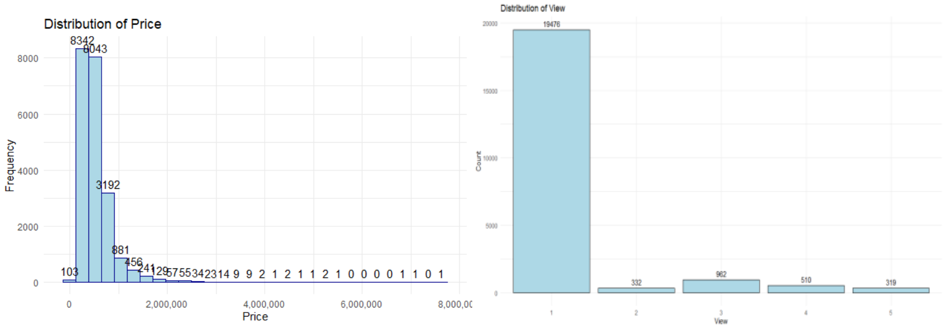


Figure 9: a)Distribution of the Price of the Houses, b)Distribution of the View of the houses scaling from 1 to 5 where 1 is the worst

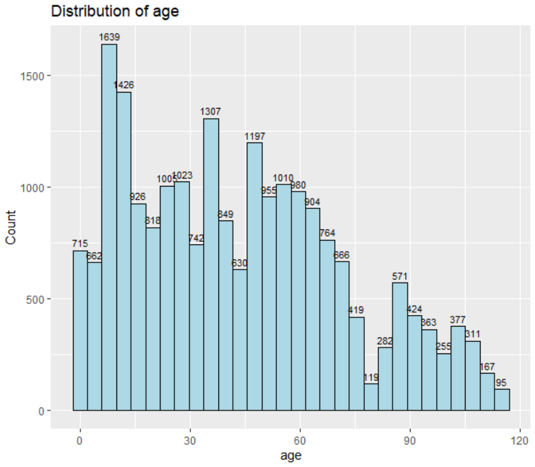


Figure 10 : Distribution of the Age of the Houses

Additional noteworthy insights are revealed through the examination of relationships between variables. Figure 10 depicts the relationship between a house's age and its price. Surprisingly, the graph indicates that the age of a house has minimal effect on its price, contrary to conventional expectations that older houses tend to have lower prices compared to newer ones. This unexpected finding may be attributed to the dataset containing a significant number of new houses and relatively few older ones, as evidenced by the distribution of the age variable in Figure 10.

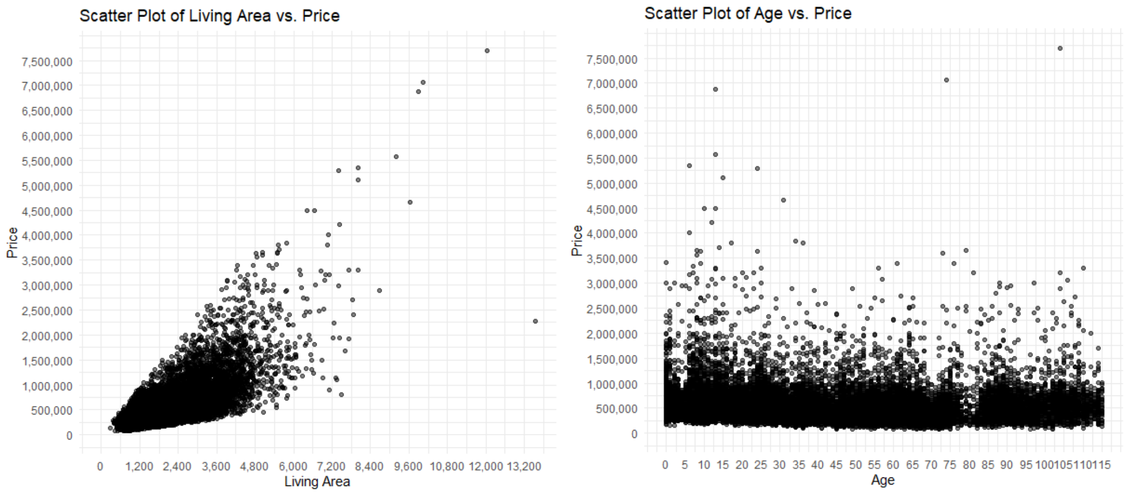


Figure 11 a) Scatter Plot of Living Area in relation to the Price of the Houses, b) Scatter Plot of the Age of the Houses in relation to the Price of the Houses

From Figure 11a) we observe a positive relationship between the squared meters of the house and its price. As the size of the living area increases, there is generally an upward trend in the prices of the houses. This suggests that larger houses with more spacious living areas tend to have higher price tags. Figure 12a) explores the relationship between the number of floors in a house and its price. Upon analyzing the scatter plot, we can observe a somewhat varied relationship between the number of floors and the price of houses. While there is not a clear and consistent linear trend, there are some noticeable patterns. In general, houses with a lower number of floors, such as single-story homes, tend to exhibit a wider range of prices. There is no distinct indication that having fewer floors leads to lower or higher prices. However, as the number of floors increases, we can observe that some houses with two or three floors tend to have higher prices.

Figure 12b) demonstrates a positive correlation between the square meters of the living area of the 15 nearest houses and the price of the subject house. From analyzing the scatter plot, we can observe a positive relationship between the average interior square footage of living space and the price of houses. As the sqft_living15 increases, there is a general upward trend in the prices of the houses. This suggests that houses surrounded by larger living spaces in the neighborhood tend to have higher prices.

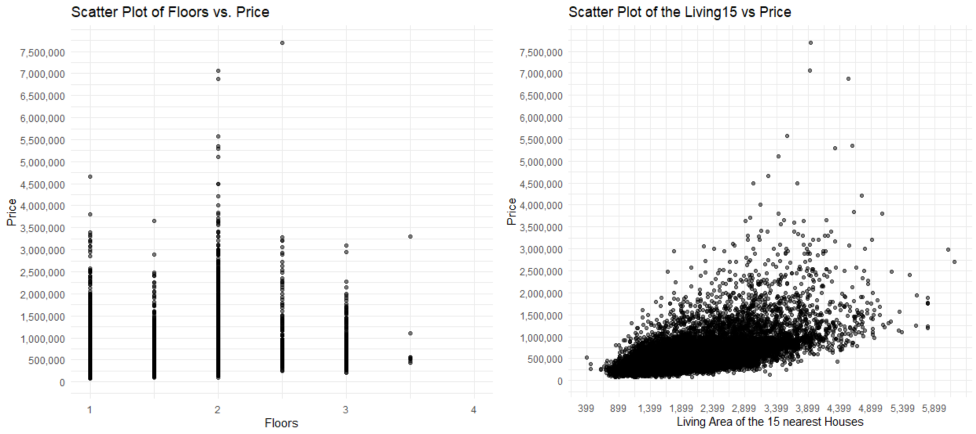


Figure 12 a) Scatter Plot between the Floors and the Price, b) : Scatter Plot between the Living Area of the 15 nearest Houses and the Price

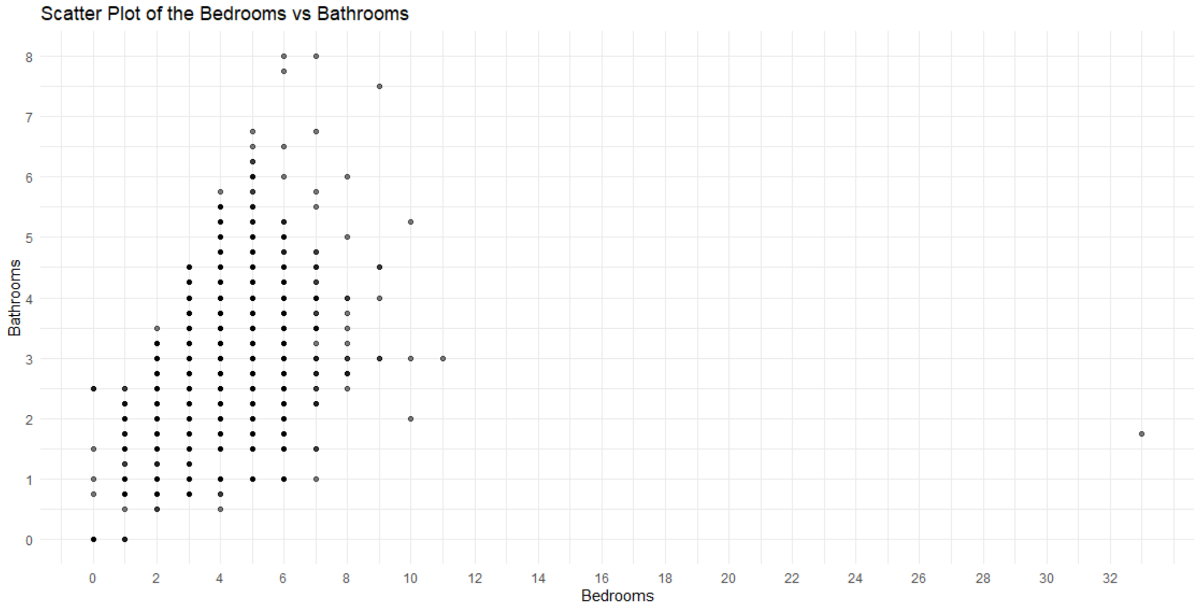


Figure 13: Scatter Plot of Bedrooms and Bathrooms

From the Figure 13, we can observe a positive relationship between the number of bedrooms and the number of bathrooms. As the number of bedrooms increases, there tends to be a corresponding increase in the number of bathrooms. This positive relationship suggests that larger properties with more bedrooms tend to have a greater number of bathrooms. It aligns with the common expectation that larger homes typically feature more bedrooms and consequently require more bathrooms to accommodate the needs of the occupants.

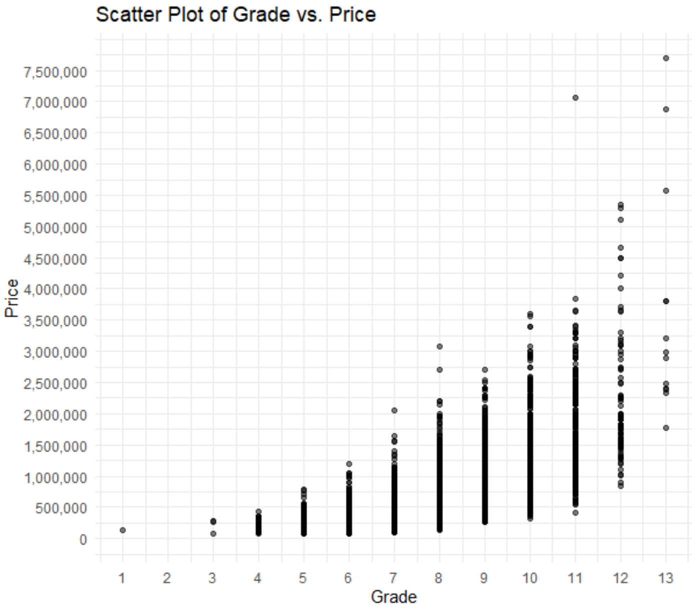


Figure 14 Scatter Plot between Grade and Price

Upon examining the relationship between price and grade from Figure 14, it becomes apparent that there is a positive correlation between these variables. Generally, properties with higher grades tend to command higher prices, while properties with lower grades are typically priced lower. This positive relationship suggests that buyers are willing to pay a premium for properties with higher grades, which reflect superior quality, design, and construction. Properties with higher grades often offer enhanced features, finishes, and overall desirability, leading to their higher market value.

5. Results

A total of four models were developed in this study: a linear regression model, a Robust M regression model using the Huber Loss Function, another Robust M Regression model using the Tukey's Bisquare Loss Function, and the CRM model. All models were implemented in the R Studio environment using specific packages. The "stats" package was utilized for the linear regression model R Core Team (2023), the "MASS" package for both M estimator models (Venables & Ripley, 2002), and the "crmReg" package for the CRM model (Filzmoser et al., 2020).

For the CRM model, the default method for initial scaling was the Qn (quantile normalized) scaling method. The Qn scaling method involves sorting the data in ascending order for each variable, calculating the median of each variable, obtaining the absolute deviations from the median for each observation, and computing the median absolute deviation (MAD). The MAD represents the median value of these absolute deviations. The Qn statistic is derived by dividing the MAD by a constant scaling factor, ensuring robustness against outliers. This scaling factor is based on the asymptotic consistency of the MAD estimator. Subsequently, each observation is divided by the Qn statistic of its respective variable to achieve data scaling.

However, due to the heavy skewness observed in our data, certain variables had MAD values of zero, rendering the division undefined. To address this issue, alternative robust scaling methods were explored. Filzmoser et al. (2020) proposed various alternatives to the Qn method. Considering the skewness of our data, the only feasible robust scaling method was the Interquartile Range (IQR) scaling. In this approach, each observation is divided by the IQR, which represents the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

5.1. Interpretability

Table 4 presents the coefficients of the variables in the three models that were developed as well as the variables importance. It should be mentioned here that the price variable was adjusted for scale and ease of interpretation by dividing each value by 10,000. This transformation does not affect the relative relationships in the data but provides more manageable numbers, making our results easier to interpret and communicate. Consequently, the resulting coefficients associated with the predictors in the regression analysis reflect the change in the house price for each unit change in the predictor, with the house price being in units of 10,000. The initial observation from this analysis is that the signs of the coefficients remain with the same sign across the models, except for the squared feet of the lot area. In linear regression, it demonstrates a negative relationship with the price, which contrasts with the sign observed in the M regression and CRM models. In general, there are noticeable differences in coefficient values between the linear regression model and the CRM model. However, the disparities between the M regression model and the CRM model are relatively minor in some variables. An interesting observation is that the coefficients of the Tukey's model are between the Huber model and the CRM model.

Furthermore, it is noteworthy that the number of bedrooms consistently shows a negative coefficient. This implies that an additional bedroom in a house leads to a noticeable decrease in the house price, varying depending on the respective model. Except for the number of bedrooms and the squared feet of the total lot area of the 15 nearest houses, all other variables have a positive impact on the dependent variable across all models. Interestingly, if the age of the house increases by one unit, the price of the house also increases by several units, with the magnitude varying across the models. Contrary to expectations, where an older house would typically cost less than a newer one, we believe this is due to the dataset containing a substantial number of new houses and a smaller proportion of older ones. Additionally, another factor influencing the relationship between house price and age is whether the house has undergone renovations. Although a variable containing relevant information was excluded due to limited inputs, it plays a role in shaping this relationship.

Another interesting insight is the positive coefficients for bathrooms which indicates that an increase in the number is bathrooms is associated with a higher house price in contrast with the number of bedrooms. This could be attributed to the desirability and convenience of having more bathrooms, which is often considered a luxury or a desirable feature. A great example of that can be seen in Airbnb's where it can be observed that many bedrooms have their own bathrooms so they can offer hospitality to many people. As expected, the squared feet of living area of the houses are associated with a higher house price in all three models. This is intuitive , as larger living areas generally command higher prices due to increased space and potential functionality. The number of floors, the condition of the house and the overall grade of the property have positive coefficients across all three models which suggest that they have a positive impact on house prices. Intuitively more floors, better condition and better overall grade tend to increase the price of house. Finally, the two variables that are associated with the characteristics of the neighborhood the house is located in tend to have a positive impact on the price. This suggest that the characteristics of the neighborhood indeed are affecting the price of the house. From these observations, it can be concluded that various factors such as the number of bedrooms, bathrooms, square footage of living space, lot area, number of floors, condition, grade, neighboring property characteristics, and age influence house prices. The specific magnitudes of the coefficients may differ across models which validates that there are outliers in our dataset hence hypothetically the more robust models should outperform the linear regression.

Variable	Linear Regression	Importance	M Regression Huber	M Regression Tukey's	CRM
Intercept	-1055880	***	-911544	-825359	-770022
bedrooms	-49582.2	***	-30456.3	-22966.3	-23668.8
bathrooms	54132.2	***	41960.26	37969.3	40522.35
sqft_living	176.80	***	106.37	80.32	75.84
sqft_lot	-0.0063		0.11	0.13	0.13
floors	23510.65	***	39992.15	44033.45	43909.25
condition	19966.56	***	21090.02	20474.76	17158.07
grade	123591.6	***	111201.3	102151.2	98214.98
sqft_living15	31.94	***	46.80	54.25	52.32
sqft_lot15	-0.52	***	-0.36	-0.28	-0.24
age	4034.44	***	3412.64	3140.20	3077.20

Table 4: Coefficients of the variables of every model.

5.2. Performance

Table 5 presents the performance metrics of the four models that were developed. The MSE values for the four models range from 512.66 to 608.47. These values indicate that on average the squared difference between the predicted and actual house prices (in units of 10,000) is ranging between those values. The Linear Regression (LM) model has the lowest MSE, suggesting that it provides relatively good predictions compared to the other models. The M Regression with Huber (MH) model has the next lowest MSE, followed by the M Regression with the Tukey's loss function and lastly CRM Model. The RMSE values for the models range from 22,64 to 24,67 measured in 10.000 USD. These values represent the average prediction error in terms of house prices and indicate the dispersion of errors around the true prices.

The Linear Regression (LM) model also has the lowest RMSE which is also measured in 10.000 USD, indicating relatively better accuracy compared to the other models. The M Regression with Huber (MH) model has the next lowest RMSE, followed by the other M Regression model and the CRM. This means that for example, if we use the OLS model to predict the price on average, the predictions will be off by about 22.64 * 10,000 = 226,400 USD from the actual house prices. Given the fact that the price is ranging from 75.000 to 7.700.000 this is a relatively good prediction.

The MAE values range from 13.68 to 14.47,also measured in 10.000 USD, representing the average absolute prediction error in terms of house prices. This metric suggests that the robust methods are performing better than the OLS which contradicts the other two metrics. The two Casewise models are ranking first, followed by the Cellwise model and lastly comes the OLS. This means that if we use the M regression model with the Huber loss function on average the predictions will be off by about 14.47 * 10,000 = 144,700 USD from the actual house prices. MAE provides a measure of the model's typical prediction accuracy. The M Regression with Huber (MH) model has the lowest MAE, indicating relatively better performance in terms of minimizing the average absolute errors. The CRM Model has a slightly higher MAE, while the Linear Regression (LM) model has the highest MAE.

It is pretty clear that there is a conflict in our results. The MSE and RMSE metric suggest that the OLS model performs better than the Robust model while the MAE says otherwise. This discrepancy arises due to the different sensitivity levels of these metrics to outliers. The first two metrics square the residuals (i.e., the differences between the predicted and actual values) before averaging them. This squaring operation makes these metrics very sensitive to outliers. A few large residuals can lead to a much higher MSE or RMSE. This is why your Ordinary Least Squares (OLS) model, which minimizes the sum of squared residuals, tends to perform well in terms of these metrics. MAE takes the average of the absolute differences between the predicted and actual values. It doesn't square the residuals, making it less sensitive to outliers than the MSE or RMSE. This is why the robust models, which aim to fit the majority of the data well without being unduly influenced by outliers, are performing better in terms of MAE.

Sahu et al. (2019b) conducted their study using the same dataset, employing multiple models including Multiple Linear Regression (MLR), Lasso, and Gradient Boosting, and compared their performance using the Accuracy metric. Among these models, the Gradient Boosting model demonstrated superior performance compared to the other two. However, considering that only Lasso can be classified as a robust method, and the accuracy metric is sensitive to outliers, it is crucial to incorporate a metric such as Mean Absolute Error (MAE) to provide a more comprehensive assessment. This suggestion aligns with the findings of Babb (2019), who also investigated various robust and non-robust methods on the same dataset. Babb utilized metrics such as Root Mean Squared Error (RMSE), MAE, Mean Absolute Percentage Error (MAPE), and Adjusted R^2 for comparison. The outcomes of their study corroborate our observations, as per RMSE, MAPE, and Adjusted R^2 which are not robust, the sensitive-to-outliers methods outperformed the robust methods. Conversely, according to MAE, the robust methods demonstrated better performance, with Support Vector Regression (SVR) using a Gaussian Kernel achieving the highest ranking.

In conclusion, the M robust regression model employing the Huber loss function exhibits superior performance and robustness when compared to all other models considered in this study. Notably, the differences observed among the various robust regression models are minimal. The primary emphasis lies in the comparison between robust and non-robust models, where the robust models distinctly outperform the non-robust model in terms of performance. This finding underscores the efficacy of robust regression techniques in handling outliers and influential data points, thereby contributing to more reliable and stable model outcomes.

Model	MSE	RMSE	MAE
Linear Regression (lm)	**512.66**	**22.64**	14.47
M Regression with Huber Loss (MH)	553.88	23.53	**13.68**
M Regression with Tukey's Bisquare	597.03	24.43	13.74
CRM Model	608.47	24.67	13.80

Table 5: Performance Metrics of every model

Diagnostics of the models

Figure 15a) and Figure 15b) illustrate the residual plots for both the Linear Regression and M robust regression models. In these plots, a red line is fitted to the residuals to facilitate the identification of any discernible patterns. The absence of significant patterns in the residuals implies that the assumption of linearity is upheld. Notably, the red line in the robust regression plot appears to exhibit a smoother trend, suggesting an enhanced fit to the data. Moreover, as we traverse along the x-axis, there is an increasing density of residuals. This observation manifests as a tunnel-shaped pattern, indicative of the presence of heteroscedasticity. Furthermore, the figures reveal the occurrence of outliers, characterized by values that deviate considerably from the predictions generated by the models. While there is an improvement in the shape of the residuals, suggesting that the robust regression model handles them more effectively, the discrepancy in performance between the two models is not substantial.

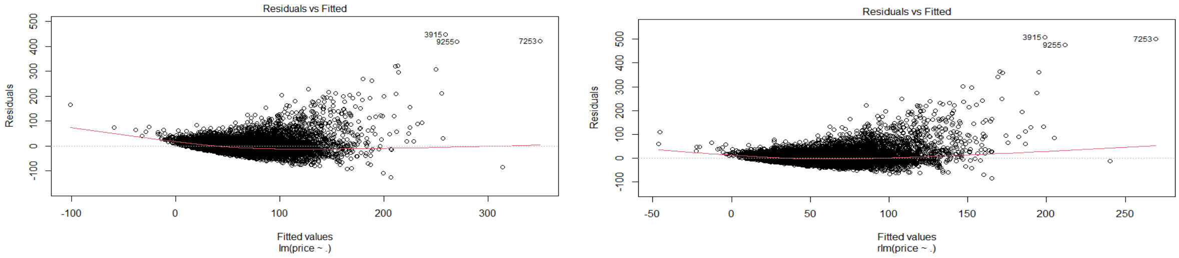


Figure 15 a) Residuals vs Fitted values of the Linear Regression Model, b) Residuals vs Fitted values of the Robust M with the Huber Loss function model

Figure 16a) and Figure 16b) depict the Q-Q plots of the Linear Regression model and the M Robust Regression model respectively. These plots serve as a means to assess whether the data adheres to a normal distribution assumption. Notably, both plots demonstrate deviations from normality, as evident by the data points straying from the diagonal line. Following the initial segment, the data points display significant deviations, signifying departures from normality. Furthermore, an examination of the left tail of both models reveals a heavy distribution, with more data points positioned below the diagonal line. This observation further supports the conclusion that the data violates the assumption of normal distribution. However, it is worth noting that the M Robust Regression model exhibits a relatively closer alignment with the diagonal line in the left tail, indicating an improvement in handling the non-normality of this extreme region. Conversely, the right tail of both models is considered light, as all data points fall above the line. At this juncture, it is challenging to ascertain whether the robust model outperforms the Linear Regression model in handling non-normality in that particular region. Additionally, it is noteworthy that the maximum value of residuals is higher in the robust model. Although this may appear counterintuitive, it suggests that the robust model has encountered more extreme observations or potential outliers that were not adequately addressed by the Linear Regression model.

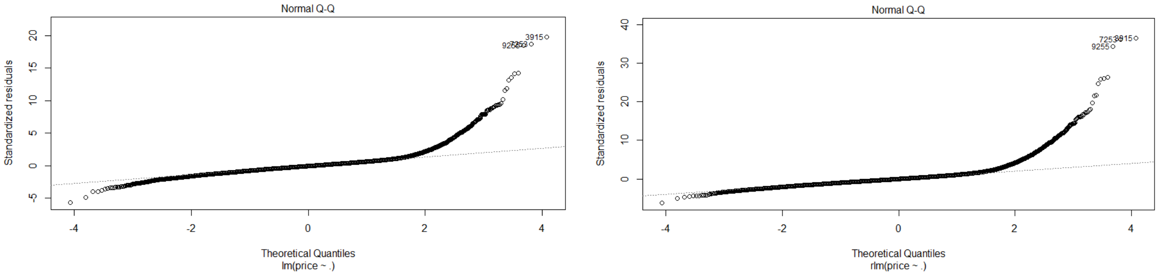


Figure 16 a) QQ plot of Linear Regression, b) QQ plot of Robust M Regression

The scale-location plot serves as a simplified analysis of the homoscedasticity assumption and is closely related to the residuals vs. fitted plot. However, it represents the square root of the absolute value of standardized residuals rather than plotting the residuals themselves. In Figure 17a), we observe that the red line in the Scale-Location plot for the Linear Regression model is not horizontal. The presence of a curved trend line in the Linear Regression model's Scale-Location plot indicates a non-constant spread of residuals across the range of fitted values. This violation of the assumption of homoscedasticity suggests that the spread of residuals systematically varies as the fitted values change. Conversely, the smoothening of the trend line in the Robust Regression model's Scale-Location plot suggests a more stable and consistent spread of residuals across the range of fitted values. This indicates that the Robust Regression model effectively handles heteroscedasticity, resulting in a more even and less variable spread of residuals. From these observations, we can conclude that the Robust Regression model handles outliers better, as anticipated. Although the residuals still exhibit a tunnel-shaped pattern, the smoother pattern in the Robust Regression model's Scale-Location plot suggests improved handling of outliers. This indicates that the Robust Regression model robustly accommodates extreme observation, resulting in a more stable spread of residuals.

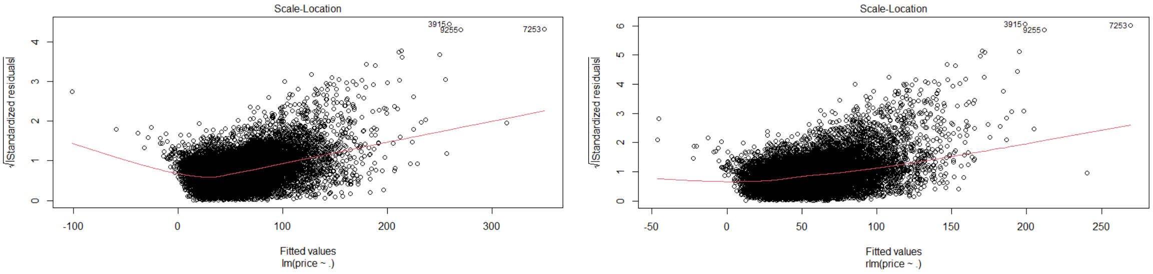


Figure 17 a) Scale – Location plot of Linear Regression, b) Scale – Location plot of M Robust Regression

The residuals vs. leverage plot serves as a diagnostic tool to identify influential observations within a regression model. Each data point from the dataset is represented as a single point on the plot, with the x-axis indicating the leverage of each point and the y-axis representing the standardized residual. In Figure 18*a)*, we observe certain data points that lie in close proximity to the border of Cook's distance, but they do not extend beyond the dashed line. This indicates that there are no influential points present in either of the models under consideration. However, notable differences emerge when comparing the two models. Specifically, the robust regression model exhibits a more effective handling of outliers, as evidenced by the absence of data points in close proximity to the dashed line. Additionally, we observe that in the linear regression model, some leverage points fall within the range of 0 to 0.2 leverage values. However, as observed in Figure 18b), in the robust regression model, these leverage points appear to possess a value of 0. The presence of these values, which are considered less influential due to the robust estimation methods employed, further highlights the superior handling of outliers by the robust regression model. The presence of a smoother red line in the Robust Regression model reflects the model's robustness to outliers and influential points. This indicates that the robust regression model is more resilient to the effects of outliers and can produce more stable and reliable estimates.

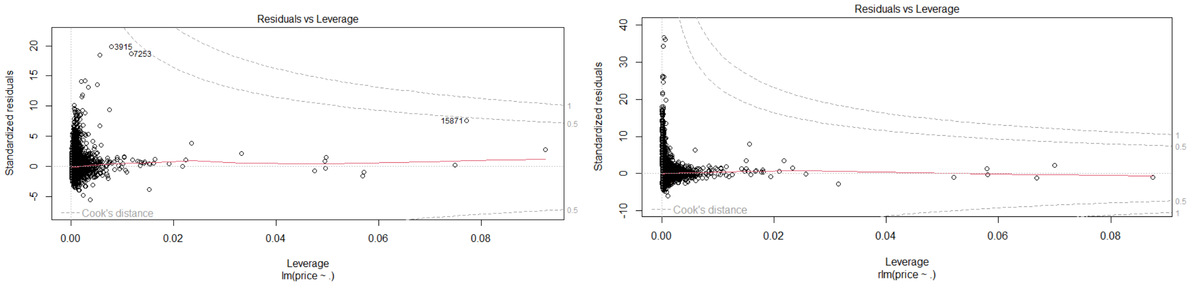


Figure 18 a) Residuals vs Leverage plot of Linear Regression, b) Residuals vs Leverage plot of M Robust Regression

Overall, the diagnostic plots suggest that the Robust Regression model adheres better to the assumptions around the Linear Regression model. The Robust Regression model shows a smoother fit, closer alignment with the diagonal line in the Q-Q plot, and a more consistent spread of residuals in the Scale-Location plot. These observations indicate the robustness of the Robust Regression model in addressing the limitations of traditional Linear Regression and its ability to provide more reliable and robust estimates.

The presented visual representation, denoted as Figure 19, comprises a heatmap that effectively portrays the initial 20 outlying houses as rows. Each row corresponds to a specific house, while the columns represent diverse features of the houses, some of which have been identified as contaminated or exhibiting outlier characteristics. These exceptional features are visually depicted through colored boxes embedded within the heatmap. The color scheme employed in the heatmap serves the purpose of distinguishing distinct types of deviations manifested by the anomalous cells. Specifically, the colors blue and red are utilized to indicate upward and downward deviations, respectively, concerning the original data values. To elaborate, blue cells signify values intended to be replaced with larger counterparts, while red cells will be substituted with smaller values through the application of the CRM model. The intensity or saturation of colors featured in the heatmap provides a visual cue, aiding in the understanding of the magnitude of differences between the imputed values and the original data values. Figure 20 subsequently displays the imputed values generated by the CRM model. For instance, upon scrutinizing the 50th case presented in Figure 19, it becomes evident that both the "sqft_lot" and "sqft_lot15" values have been flagged with red, indicative of their higher-than-expected magnitudes. Additionally, it is pertinent to note that the opacity or intensity of the red color in the flagged cells exhibits variation, signifying that the imputed values will undergo distinctive adjustments. In particular, cells that appear opaquer will experience a comparatively smaller decrease as opposed to those displaying lower opacity. It is evident from Figure 20 that the "sqft_lot" value, which exhibits lesser intensity than the "sqft_lot15," underwent a relatively smaller decrease.

Significantly, the CRM analysis proved successful in identifying a total of 3010 casewise outliers within the dataset, which consists of 21600 houses. Casewise outliers are identified based on the presence of at least one contaminated or outlying cell, as determined by the CRM method. The robust identification of these outliers through the CRM model contributes to the robustness and accuracy of the overall analysis, enabling the detection and treatment of exceptional cases that deviate significantly from the expected data distribution.

Remarkably, a striking observation emerges from the imputation process, as all the newly imputed values appear to cluster around the median of their respective variables. This finding carries significant implications for the performance and efficacy of the algorithm in addressing cellwise outliers. To illustrate this point, let us examine the variable "bedrooms" as an example, where the median value is determined to be 3. Upon close examination of the imputed values, it becomes evident that the algorithm has successfully taken corrective measures to bring the outlier values back in line with the central tendency represented by the median. For instance, at the 52nd observation, the original value of 5 for the "bedrooms" variable was identified as an outlier and consequently reduced to 4 to align with the central tendency. Similarly, in the 54th observation, the initial value of 2 was imputed as 3.5, while in the 70th observation, the original value of 5 was adjusted to 3.2. These examples illustrate the consistent trend of the algorithm in adjusting the imputed values closer to the median.

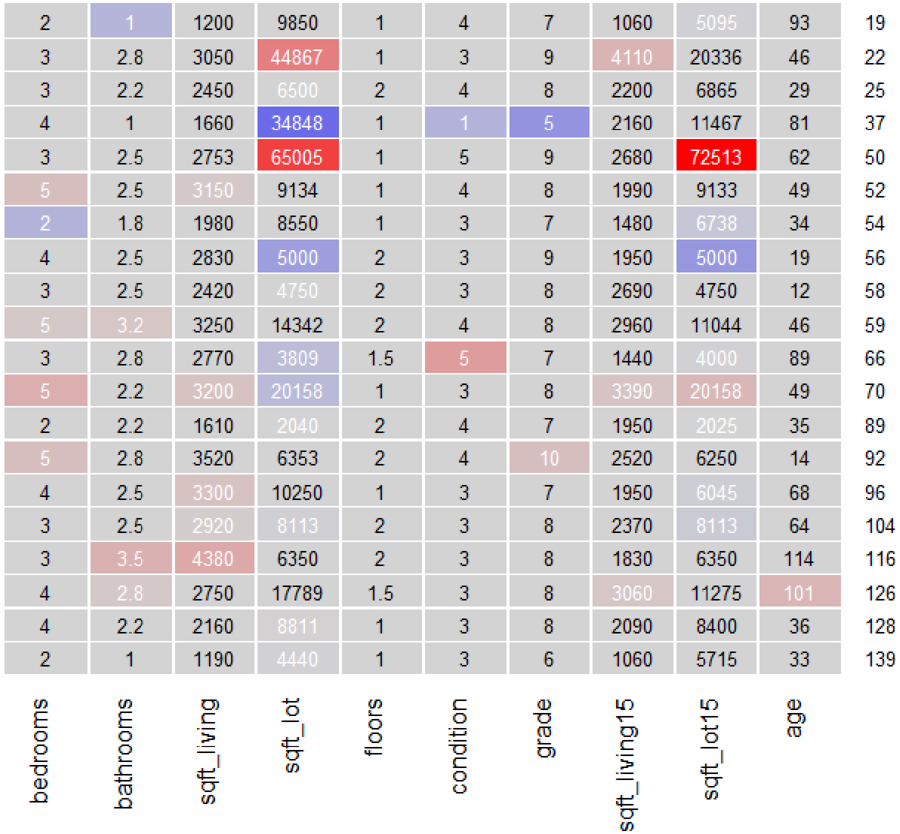
This pattern of imputation, where the algorithm seeks to align the imputed values with the central tendency of each variable, serves as compelling evidence that the model indeed exhibits effective behavior in countering cellwise outliers. By anchoring the imputed values around the median, the algorithm ensures that the overall distribution of the data is preserved while mitigating the impact of outlier values on the analysis. This observation further bolsters the confidence in the reliability and accuracy of the proposed approach in handling cellwise outliers within the dataset. The adherence of the imputed values to the central tendency underscores the algorithm's ability to maintain the integrity of the data while effectively addressing the presence of anomalous observations. Consequently, this outcome enhances the credibility and robustness of the analysis conducted using the CRM method, making it a valuable tool for identifying and handling outliers in diverse datasets.

In this context, the M-robust regression model with the Huber loss function is utilized to assess the weights assigned to the cases that were previously flagged by the CRM model. Table 6 displays these weights for the same 20 cases that are visualized in Figures Figure 19 and Figure 20. The presence of weights equal to 1 in the M-robust regression results indicates that certain cases were not considered outliers according to this robust model. Consequently, these cases were not downweighted, as their residuals were not significantly deviant from the model's predictions. It is essential to recognize that while the CRM model identifies certain cases as outliers based on the contaminated cells, the M-robust regression model evaluates outliers using a different criterion, focusing on the Huber loss function to determine weights.

Particularly, the 50th case, which bears the lowest weight of 0.61 among all the cases, stands out as a highly downweighted observation. Comparing this case's weight with the heatmap visualizations, it becomes evident that two variables, "sqft_lot" and "sqft_lot15," were flagged by the M-robust regression and underwent substantial decreases in their values during the imputation process. Conversely, for cases that were not flagged as outliers by the M-robust regression, limited and non-dramatic changes are observed. Consequently, these cases receive weights equal to 1, signifying that they were not considered outliers by the robust model. The stability of their weights can be attributed to the model's perception of these cases as conforming reasonably well to the overall data distribution. It is important to highlight that the M-robust regression and the CRM model employ distinct approaches to identify and address outliers. While the CRM model relies on the detection of contaminated cells, the M-robust regression focuses on minimizing the impact of outliers through the Huber loss function. The differences in their identification criteria can lead to discrepancies in the flagged cases. Nonetheless, the combination of both methods offers a comprehensive approach to handling outliers and enhances the robustness and accuracy of the overall analysis.

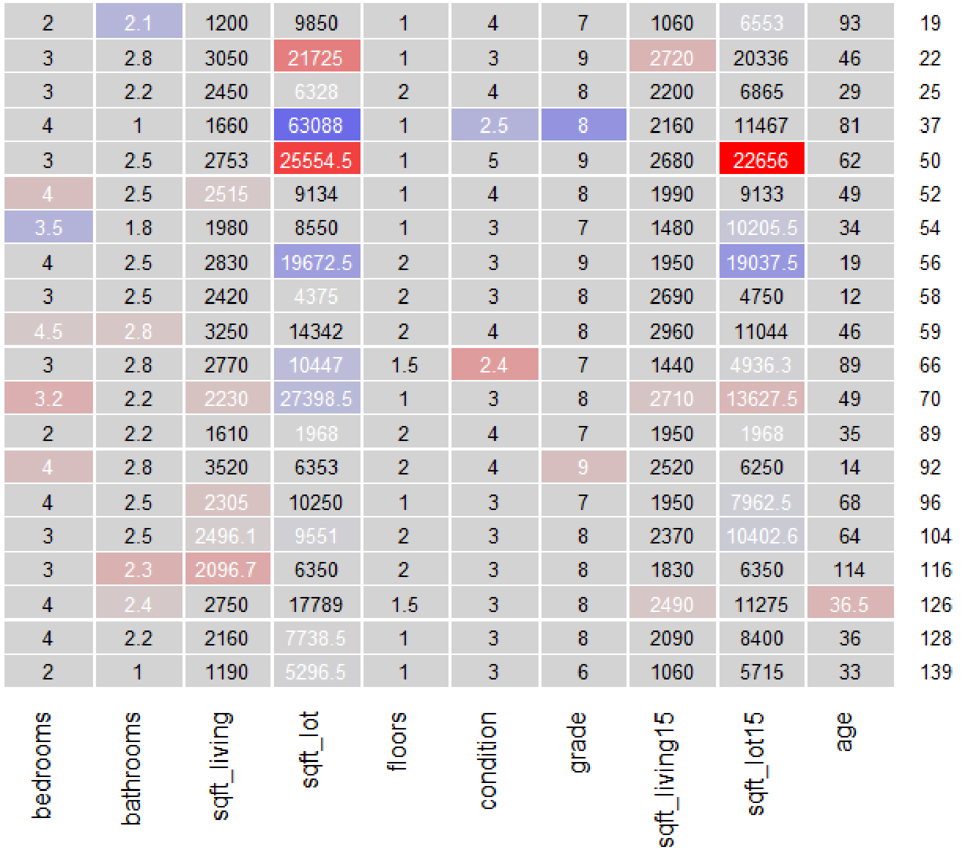
Column name	Weights	Column name	Weights
19	1	**66**	1
22	1	**70**	1
25	1	**89**	1
37	1	**92**	1
50	0.61	**96**	1
52	0.90	**104**	0.85
54	0.87	**116**	1
56	0.73	**126**	0.77
58	1	**128**	1
59	1	**139**	0.78

Table 6: Weights of the M Robust Regression model



bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	grade	sqft_living15	sqft_lot15	age	
2	1	1200	9850	1	4	7	1060	5095	93	19
3	2.8	3050	44867	1	3	9	4110	20336	46	22
3	2.2	2450	6500	2	4	8	2200	6865	29	25
4	1	1660	34848	1	1	5	2160	11467	81	37
3	2.5	2753	65005	1	5	9	2680	72513	62	50
5	2.5	3150	9134	1	4	8	1990	9133	49	52
2	1.8	1980	8550	1	3	7	1480	6738	34	54
4	2.5	2830	5000	2	3	9	1950	5000	19	56
3	2.5	2420	4750	2	3	8	2690	4750	12	58
5	3.2	3250	14342	2	4	8	2960	11044	46	59
3	2.8	2770	3809	1.5	5	7	1440	4000	89	66
5	2.2	3200	20158	1	3	8	3390	20158	49	70
2	2.2	1610	2040	2	4	7	1950	2025	35	89
5	2.8	3520	6353	2	4	10	2520	6250	14	92
4	2.5	3300	10250	1	3	7	1950	6045	68	96
3	2.5	2920	8113	2	3	8	2370	8113	64	104
3	3.5	4380	6350	2	3	8	1830	6350	114	116
4	2.8	2750	17789	1.5	3	8	3060	11275	101	126
4	2.2	2160	8811	1	3	8	2090	8400	36	128
2	1	1190	4440	1	3	6	1060	5715	33	139

Figure 19 Heatmap of the first 20 cases that were not yet imputed by the CRM model



bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	grade	sqft_living15	sqft_lot15	age	
2	2.1	1200	9850	1	4	7	1060	6553	93	19
3	2.8	3050	21725	1	3	9	2720	20336	46	22
3	2.2	2450	6328	2	4	8	2200	6865	29	25
4	1	1660	63088	1	2.5	8	2160	11467	81	37
3	2.5	2753	25554.5	1	5	9	2680	22656	62	50
4	2.5	2515	9134	1	4	8	1990	9133	49	52
3.5	1.8	1980	8550	1	3	7	1480	10205.5	34	54
4	2.5	2830	19672.5	2	3	9	1950	19037.5	19	56
3	2.5	2420	4375	2	3	8	2690	4750	12	58
4.5	2.8	3250	14342	2	4	8	2960	11044	46	59
3	2.8	2770	10447	1.5	2.4	7	1440	4936.3	89	66
3.2	2.2	2230	27398.5	1	3	8	2710	13627.5	49	70
2	2.2	1610	1968	2	4	7	1950	1968	35	89
4	2.8	3520	6353	2	4	9	2520	6250	14	92
4	2.5	2305	10250	1	3	7	1950	7962.5	68	96
3	2.5	2496.1	9551	2	3	8	2370	10402.6	64	104
3	2.3	2096.7	6350	2	3	8	1830	6350	114	116
4	2.4	2750	17789	1.5	3	8	2490	11275	36.5	126
4	2.2	2160	7738.5	1	3	8	2090	8400	36	128
2	1	1190	5296.5	1	3	6	1060	5715	33	139

Figure 20 Heatmap of the first 20 cases that were imputed by the CRM mode

Conclusions

The analysis looked at multiple models to predict house prices based on a range of factors. The models consistently showed that most variables (like the number of bathrooms, living area square footage, number of floors, house condition, and neighborhood characteristics) positively impacted house prices, while the number of bedrooms negatively impacted prices. One point of discrepancy was seen in how the square footage of the lot area was treated across models. In the linear regression model, it had a negative relationship with the price, but in the M regression and CRM models, the relationship was positive. Another unexpected finding was that the housing age was positively associated with price, which is against common expectations. It was hypothesized that this may be due to the dataset containing a large number of new houses and fewer old ones, and the possible impact of house renovations. The coefficients for each variable varied across the models, with the M regressions and CRM models showing more similar results to each other than to the linear model, presumably due to the latter's sensitivity to outliers. In conclusion, the house price is influenced by a variety of factors, including the number of bedrooms and bathrooms, square footage of living and lot area, number of floors, house condition, property grade, and characteristics of the neighborhood. However, the specific influence of these factors may differ depending on the statistical model used, highlighting the importance of model selection, and understanding potential outliers in the data.

Performance wise based on the MSE and RMSE metrics, the LM model showed the highest accuracy, however, when using the MAE metric, which is less sensitive to outliers, the robust models (MH and CRM) performed better. This discrepancy arises because MSE and RMSE square the residuals, making them more sensitive to outliers. A few large residuals can significantly increase MSE and RMSE values. On the other hand, MAE does not square the residuals, making it less sensitive to outliers, hence robust models perform better in terms of MAE. It can be observed that, as the M Regression outperforms the CRM in terms of MAE, the impact of cellwise outliers is not considerable.

In conclusion, an understanding of the specific attributes of the house to be predicted, the nature of the dataset, and the characteristics of different models can significantly enhance the precision of the property valuation process. The choice between the M-regression and CRM models should be guided by these considerations to ensure the most accurate house price prediction. Finally, from this analysis can be highlighted that our models are not affected by cellwise outliers thus a model that is robust to casewise outliers is sufficient.

6. Limitations and Future Work

This study, while insightful and contributive to the field of real estate price prediction, faced several limitations that could inspire future research directions. The primary limitation lies in the data utilized for the analysis. Some variables within the dataset lacked substantial representation, leading to their exclusion. This exclusion potentially resulted in the loss of vital information about specific house features that might have otherwise significantly influenced price prediction. For instance, variables encapsulating certain architectural characteristics or specific amenities might not have been adequately represented. Moreover, the age variable was skewed towards newer houses with a limited number of older ones. This skewness could have influenced the conclusion that a positive relationship exists between the age of a house and its price. This raises a question about whether this relationship would have changed with a more balanced representation of house ages. The quality of a neighborhood is another crucial aspect influencing house prices, yet it was not considered in this study. Variables reflecting the quality of a neighborhood, such as the proximity to schools, hospitals, parks, and the crime rate, can have a significant impact on real estate prices. Future research could benefit from considering these aspects. Turning to the methodologies used, while OLS, M-regression, and CRM have shown efficacy in this context, they are not universally applicable. The choice of method is sensitive to the nature of the data. Therefore, future studies could aim to develop more versatile models or mechanisms that can intelligently choose between models based on the characteristics of the dataset. The study also offers a framework for those intending to apply the CRM model to real estate datasets. On a broader scale, there is ample room to test these methodologies in other fields dealing with multivariate datasets exhibiting high variability. Such studies could help to generalize the findings of this research and might reveal additional insights regarding the performance of these models across diverse contexts. Furthermore, the proposed framework for applying the CRM model to real estate datasets could be empirically tested and refined. For example, the step involving scaling of residuals using the IQR method in place of MAD in the case of skewed data could be subjected to additional testing. Finally, while this study focused primarily on the detection of cellwise outliers, other types of outliers and anomalies might also be relevant in real estate data, such as blockwise or groupwise outliers. Future research could delve into the development or adaptation of methodologies to handle these different types of outliers more effectively. This would contribute to the broader aim of enhancing the robustness and reliability of house price predictions.

7. References

1) Adamczyk, T. (2017). Application of the Huber and Hampel M-estimation in real estate value modeling. *Geomatics and Environmental Engineering*, *11*(1), 15. https://doi.org/10.7494/geom.2017.11.1.15
2) Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, *24*(3), 441–461. https://doi.org/10.1007/s11749-015-0450-6
3) Ali, J., Khan, R. U., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278. http://ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf
4) Archer, W. R., Gatzlaff, D. H., & Ling, D. C. (1996). Measuring the Importance of Location in House Price Appreciation. *Journal of Urban Economics*, *40*(3), 334–353. https://doi.org/10.1006/juec.1996.0036
5) Audibert, J., & Catoni, O. (2011). Robust linear least squares regression. *Annals of Statistics*, *39*(5). https://doi.org/10.1214/11-aos918
6) Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Apress eBooks* (pp. 67–80). https://doi.org/10.1007/978-1-4302-5990-9_4
7) Babb, O. (2019). A Comparison of Machine Learning Approaches to Housing Value Estimation. *SIAM Undergraduate Research Online*, *12*. https://doi.org/10.1137/18s017296
8) Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3, No. 1). New York: Wiley.
9) Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5
10) Bourassa, S. C., Haurin, D. R., Haurin, J. L., Hoesli, M., & Sun, J. (2009c). House Price Changes and Idiosyncratic Risk: The Impact of Property Characteristics. *Real Estate Economics*, *37*(2), 259–278. https://doi.org/10.1111/j.1540-6229.2009.00242.x

11) Bourassa, S. C., Hoesli, M., & Sun, J. (2006b). A simple alternative house price index method. *Journal of Housing Economics*, *15*(1), 80–97. https://doi.org/10.1016/j.jhe.2006.03.001

12) Cohen, V., & Karpavičiūtė, L. (2017c). The analysis of the determinants of housing prices. *Independent Journal of Management &Amp; Production*, *8*(1), 49–63. https://doi.org/10.14807/ijmp.v8i1.521

13) Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research*, *3*(1), 58–67. https://doi.org/10.21500/20112084.844

14) Dataset: *House sales in King County, USA*. (2016, August 25). Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

15) Davies, L., & Gather, U. (1993). The Identification of Multiple Outliers. *Journal of the American Statistical Association*, *88*(423), 782–792. https://doi.org/10.1080/01621459.1993.10476339

16) De Haan, J., & Diewert, E. (2013). Hedonic Regression Methods. In *OECD eBooks* (pp. 49–64). https://doi.org/10.1787/9789264197183-7-en

17) De Menezes, D. Q. F., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021). A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, *147*, 107254. https://doi.org/10.1016/j.compchemeng.2021.107254

18) De Menezes, D. Q. F., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021b). A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, *147*, 107254. https://doi.org/10.1016/j.compchemeng.2021.107254

19) De Vries, P. S., De Haan, J., Van Der Wal, E., & Mariën, G. (2009). A house price index based on the SPAR method. *Journal of Housing Economics*, *18*(3), 214–223. https://doi.org/10.1016/j.jhe.2009.07.002

20) Debruyne, M., Höppner, S., Serneels, S., & Verdonck, T. (2019). Outlyingness: Which variables contribute most? *Statistics and Computing*. https://doi.org/10.1007/s11222-018-9831-5

21) Debruyne, M., Höppner, S., Serneels, S., & Verdonck, T. (2019). Outlyingness: Which variables contribute most? *Statistics and Computing*, *29*(4), 707–723. https://doi.org/10.1007/s11222-018-9831-5

22) Dietz, R. D., & Haurin, D. R. (2003). The social and private micro-level consequences of homeownership. *Journal of Urban Economics*, *54*(3), 401–450. https://doi.org/10.1016/s0094-1190(03)00080-9

23) Dixon, W. J. (1953). Processing Data for Outliers. *Biometrics*, *9*(1), 74. https://doi.org/10.2307/3001634

24) Enderlein, G. (1987). Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 S., £ 14, 50. *Biometrical Journal*, *29*(2), 198. https://doi.org/10.1002/bimj.4710290215

25) Falk, R. F., & Miller, N. H. (1992a). *A Primer for Soft Modeling*. The University of Akron Press: Akron, OH. https://psycnet.apa.org/record/1992-98610-000

26) Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., & Verdonck, T. (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, *147*, 106944. https://doi.org/10.1016/j.csda.2020.106944

27) Fox, J., & Monette, G. (2003). An R and S-Plus Companion to Applied Regression. *Canadian Journal of Sociology*, *28*(1), 110. https://doi.org/10.2307/3341881

28) Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397–416. https://doi.org/10.1080/10618600.1998.10474784

29) Goh, Y. M., Costello, G., & Schwann, G. (2012). Accuracy and Robustness of House Price Index Methods. *Housing Studies*, *27*(5), 643–666. https://doi.org/10.1080/02673037.2012.697551

30) Hermey, D., & Watson, G. A. (1999). Fitting Data with Errors in All Variables Using the Huber M-estimator. *SIAM Journal on Scientific Computing*, *20*(4), 1276–1298. https://doi.org/10.1137/s106482759731823x
31) Hoffmann, I. S., Filzmoser, P., Serneels, S., & Varmuza, K. (2016). Sparse and robust PLS for binary classification. *Journal of Chemometrics*, *30*(4), 153–162. https://doi.org/10.1002/cem.2775
32) Huber, P. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Statistics*, *1*(5). https://doi.org/10.1214/aos/1176342503
33) Huber, P. (1981). Robust Statistics. In *Wiley series in probability and statistics*. Wiley. https://doi.org/10.1002/0471725250
34) Huber, P. (2011). Robust Statistics. *Springer EBooks*, 1248–1251. https://doi.org/10.1007/978-3-642-04898-2_594
35) Isaac F. Megbolugbe. (1995). Editor's Introduction and Summary. *Journal of Housing Research*, *6*(2). https://www.jstor.org/stable/24832824
36) Jorgensen, M. A. (2012). Iteratively Reweighted Least Squares. *Encyclopedia of Environmetrics*. https://doi.org/10.1002/9780470057339.vai022
37) Kiel, K. A., & Carson, R. E. (1990). An Examination of Systematic Differences in the Appreciation of Individual Housing Units. *Journal of Real Estate Research*, *5*(3), 301–318. https://doi.org/10.1080/10835547.1990.12090630
38) Koenker, R., & Hallock, K. F. (2001). Quantile Regression. *Journal of Economic Perspectives*, *15*(4), 143–156. https://doi.org/10.1257/jep.15.4.143
39) Kohl, M. (2005). *Numerical Contributions to the Asymptotic Theory of Robustness*. https://epub.uni-bayreuth.de/839/
40) Leung, A., Zhang, H., & Zamar, R. H. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Computational Statistics & Data Analysis*, *99*, 1–11. https://doi.org/10.1016/j.csda.2016.01.004
41) Marazzi, A. (1993). *Algorithms, Routines, and S-Functions for Robust Statistics*. CRC Press.
42) Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. http://ci.nii.ac.jp/ncid/BA76421258
43) McMillen, D. P. (2008). Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients? *Journal of Urban Economics*, *64*(3), 573–589. https://doi.org/10.1016/j.jue.2008.06.002
44) Muhlbauer, A., Spichtinger, P., & Lohmann, U. (2009b). Application and Comparison of Robust Linear Regression Methods for Trend Estimation. *Journal of Applied Meteorology and Climatology*. https://doi.org/10.1175/2009jamc1851.1
45) NEELAM SHINDE, & KIRAN GAWANDE. (2018). VALUATION OF HOUSE PRICES USING PREDICTIVE TECHNIQUES. *International Journal of Advances in Electronics and Computer Science*, *Volume-5*(Issue-6). https://www.iraj.in/journal/journal_file/journal_pdf/12-477-153396274234-40.pdf
46) Öllerer, V., Alfons, A., & Croux, C. (2016). The shooting S-estimator for robust regression. *Computational Statistics*, *31*(3), 829–844. https://doi.org/10.1007/s00180-015-0593-7
47) Osborne, J. A., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research and Evaluation*, *9*(1), 1–8. https://doi.org/10.7275/qf69-7k43
48) Pearson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions on Control Systems and Technology*, *10*(1), 55–63. https://doi.org/10.1109/87.974338

49) Pranav Kangane, Aadesh Mallya, Aayush Gawane, Vivek Joshi, & Shivam Gulve. (2021). Analysis of Different Regression Models for Real Estate Price Prediction. *International Journal of Engineering Applied Sciences and Technology*, *Vol. 5*(2455–2143), 247–254.
50) R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
51) Raymaekers, J., & Rousseeuw, P. J. (2023b). Challenges of cellwise outliers. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.02156
52) Richards, F. S. G. (1961). A Method of Maximum-Likelihood Estimation. *Journal of the Royal Statistical Society Series B-Methodological*, *23*(2), 469–475. https://doi.org/10.1111/j.2517-6161.1961.tb00430.x
53) Rieder, H. (1994). Robust Asymptotic Statistics. In *Springer eBooks*. https://doi.org/10.1007/978-1-4684-0624-5
54) Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. https://doi.org/10.1080/01621459.1993.10476408
55) Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust Regression and Outlier Detection*. John Wiley & Sons.
56) Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*, *12*(1), 29–45. https://doi.org/10.1007/s10618-005-0024-4
57) Ruppert, D., Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1987). Robust Statistics: The Approach Based on Influence Functions. *Technometrics*, *29*(2), 240. https://doi.org/10.2307/1269782
58) Ruppert, D., Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1987b). Robust Statistics: The Approach Based on Influence Functions. *Technometrics*, *29*(2), 240. https://doi.org/10.2307/1269782
59) Sahu, M., Singh, A., & Chawda, R. (2019). House Price Prediction using Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717–722. https://doi.org/10.35940/ijitee.i7849.078919
60) Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-sts330
61) Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), 1–44. https://doi.org/10.1080/10835547.2005.12090154
62) Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44. https://doi.org/10.1080/10835547.2005.12090154
63) Spiegel, M., & Goetzmann, W. N. (1997). A Spatial Model of Housing Returns and Neighborhood Substitutability. *Social Science Research Network*. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=9273
64) Su, X., Yan, X., & Tsai, C. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198
65) Thode, H. C., Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1986b). Exploring data tables, trends, and shapes. *Technometrics*, *28*(4), 399. https://doi.org/10.2307/1268989
66) Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B-methodological*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
67) Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 448–485. https://ci.nii.ac.jp/naid/20000755025/

68) Velasco, H., Laniado, H., Toro, M., Leiva, V., & Lio, Y. (2020). Robust Three-Step Regression Based on Comedian and Its Performance in Cell-Wise and Case-Wise Outliers. *Mathematics*, *8*(8), 1259. https://doi.org/10.3390/math8081259

69) Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. In *Statistics and computing*. Springer Nature. https://doi.org/10.1007/978-0-387-21706-2

70) Wang, F., & Zorn, P. (1997). Estimating House Price Growth with Repeat Sales Data: What's the Aim of the Game? *Journal of Housing Economics*, *6*(2), 93–118. https://doi.org/10.1006/jhec.1997.0209

71) Wiggins, B. C. (2000). Detecting and Dealing with Outliers in Univariate and Multivariate Contexts. *Annual Meeting of the Mid-South Educational Research Association*. http://files.eric.ed.gov/fulltext/ED448189.pdf

72) Witte, A. D., Sumka, H. J., & Erekson, H. (1979). An Estimate of a Structural Hedonic Price Model of the Housing Market: An Application of Rosen's Theory of Implicit Markets. *Econometrica*, *47*(5), 1151. https://doi.org/10.2307/1911956

73) Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation*, *46*(8), 6261–6282. https://doi.org/10.1080/03610918.2016.1202271

74) Zabel, J. E. (1996). Controlling for Quality in House Price Indices. *Social Science Research Network*. https://autopapers.ssrn.com/sol3/papers.cfm?abstract_id=181529

75) Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. In *R News* (Vol. 2, Issue 3, pp. 7–10). https://CRAN.R-project.org/doc/Rnews/

76) Zhao, Y., Chetty, G., & Tran, D. (2019). *Deep Learning with XGBoost for Real Estate Appraisal*. https://doi.org/10.1109/ssci44817.2019.9002790

77) Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008b). Determinants of House Prices: A Quantile Regression Approach. *Journal of Real Estate Finance and Economics*, *37*(4), 317–333. https://doi.org/10.1007/s11146-007-9053-7