ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Master of Data Science and Marketing Analytics

Pattern Pending™

Exploring Patterns and Innovation Strategy in Patent Applications

Name: Tarun Kuppili Venkata
Student Number: 438196

Thesis Supervisor: S. L. Malek
Second Assessor: P. J. F. Groenen

Date Final Version: 13 August 2021

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.

Abstract

A large number of new products entering the market today end up failing. There have been as many studies navigating the various factors that have an important influence on new product success, however, surprisingly few about the knowledge management side of invention. This is the gap that this thesis aims to bridge. This thesis explores the viability of relying on patent data to help managers make decisions regarding investment in the research and development. The dataset was from the UK IPO office, from 1978 to 2018. From this, all applications pertaining to electricity were collected and investigated. By examining the WIPO classifications assigned to each of the applications, the study aimed to explore emergent patterns and relationships between classifications. Principal component analysis, penalised regression and ensemble bagging methods played a role in helping identify the main and interaction effects. The results pointed to distinct areas of innovation goals and the variables most aligned to each area. While there were many interesting main effects, there were almost no observable interaction effects. Additionally, a number of the classifications overlapped between different areas, alluding to the flexibility afforded by innovation managers and R&D teams around the world. Thus, based on these results, some structure and goal orientation could be introduced into corporate innovation, to better drive investment and ultimately the company's bottom line.

Table of Contents

Introduction

“Ninety percent of everything is crap,” observed the science fiction author Theodore Sturgeon. This colourful observation is famously known, today, as Sturgeon's Law or Sturgeon's Revelation. At the time, the science fiction genre was routinely derided by critics, for the underwhelming quality in an overwhelming number of its works. Sturgeon's Revelation was in response to this wave of negative criticism; according to him, the majority of examples of any art form were rather poor in quality. Thus, he equated science fiction to art.

Following Sturgeon's logic, new product development could also be considered an artform. This is because an overwhelming proportion of new products in the market end up failing. According to Harvard Business School professor, Clay Christensen, 95% of the 30,000 new products introduced every year fail (Nobel, 2011). Nielsen BASES observed the rate to be between 80 and 85% for fast moving consumer goods (FMCG) fail (Melgarejo et al., 2018). As a consequence, vast amounts of investments in both start-ups and corporate innovations fail to make any meaningful returns every year (Dean, 2017; Rowley, 2017).

Despite these extreme failure rates, firms have no choice but to innovate. Both Zahra & Covin (1994) and Bessant et al. (2005) consider innovation to be intrinsically tied to survival of the firm in the marketplace, due to limited growth prospects. In the last decade, this trend was embodied by the smartphone manufacturer HTC. In 2011, HTC commanded over 10% of the market share thanks to its then innovative focus on build quality and design on the burgeoning smartphone device. The company, however, ultimately fell out of public graces once competitors such as Samsung and Google started matching the devices in terms of design and build quality. As a result, HTC’s products did not manage to differentiate themselves in the market anymore and the company quickly faded, with under 1% of market share today in the smartphone category today (Statista, 2021).

Consequently, the high failure rate implies there is some room for improvement. McKinsey, the management consultancy, has turned its attention towards leadership (Cohen et. al, 2021). The Boston Consulting Group took a more holistic view in considering the innovation problem, and observed four barriers (lack of general direction for innovation; failing to take ground realities into account; inadequate testing of new ideas; not allocating enough resources to new ventures) a company must overcome in order to succeed (Harnoss et al., 2019). This appears to be somewhat of a trend, as more consultancies tend to focus within the organisation to identify factors that might result in new product success (NPS).

Similarly, academic literature in the field also points towards factors that might improve the chances of success for an undertaking. Evanschitzky et. al (2012) takes a meta-analytic approach to illustrate the efficacy of a number of highly cited NPS factors. While the results of this study provides a holistic insight into the dynamics of success for new products, it does not help inform an organisation on how best to assess their technical capabilities and deploy resources for research and development.

Consequently, this thesis aims to model the best path for innovation success based on emergent patterns within the technology sector. To do this, it will focus on granted patents and extrapolate their impact through applications generated by them. Based on these, emergent patterns and their components will be studied to generate insight. The results of this investigation could provide structure to innovation efforts within companies. Rather than simply seeking successful project completions, managers could be equipped to question how to proceed based on on-hand expertise.

Research Question

This paper looks into patent data to try and uncover interesting insights regarding innovation and strategy. The existence of a patent indicates a solution devised by an applicant for a problem. Patent analysis is routinely used to examine specific trends in innovation. It has proven instrumental in specific topics such as mapping evolution of particular technologies (Ardito et al, 2018; Cho et al, 2018; Shubbak, 2019; Yuan & Cai, 2021) as well as in broader topics such as impact of policy changes (Costantini et al, 2017; Abraham & Moitra, 2000). As the problems get more complex, more avenues for potential solutions open up for innovators. Based on this, the following research question is formulated:

RQ: How can patent information shape innovation strategy in the technology sector?

Innovation success is difficult to predict. There have been numerous studies about the factors that influence success (Evanschitzky et al. 2012; Henard & Szymanski, 2001), but they tend to focus on the process, rather than the features and ingredients within the invention. In technology, these tend to be represented by patent codes. Thus, the aim of this paper is twofold: to first identify successful combinations of patent codes into innovations and second, to help an enterprising company map out its best path towards a favourable combination. For this, only patents with patent code H (Electricity) by the International Patent Classification (IPC) are considered. Based on this, the following sub-questions have been devised.

SQ: Are there specific patterns within successful patent applications?

Before exploring the viability of patent analysis to explore innovation strategy, it is vital to examine if there are any prevalent patterns that emerge from successful applications. Theory (Fagerberg et al., 2013) suggests that there are four major types of innovation. These four types of innovations are: product, process, organisation and marketing. Considering that patent applications and innovation are closely related, how do these innovation types emerge within patent applications? Identifying patterns would prove instrumental in setting innovation goals for managers.

SQ: Which technological problems tend to be the most and least profitable?

When a patent is applied, it is assigned a collection of codes from the IPC. Depending on the exact nature of the problem, there are subdivisions within each code to inform on specificity. Due to this, it is common for an application to feature multiple codes, as an invention can focus on diverse problems, or a few exceedingly complex problems with multiple aspects. Thus, these applications could provide a more exacting definition of the problem that they are trying to solve. Examining the co-occurrences of particular IPC codes within each application is interesting in identifying which problems are the most valuable for a company to solve i.e. worth the research and development investment.

Subject Relevance for Management

The research is particularly relevant to managers as it helps them assess their in-house competence and plan the best path forward. Particularly, it is helpful in understanding if their current strategy is optimal: if they should focus more on incremental innovation or combine multiple projects into a breakthrough innovation. The thesis can further provide insight into whether it is profitable to collaborate with experts within or across industries, to develop disruptive products. Considering the aforementioned failure rate, the results of this research could provide structure towards a company’s research and development efforts.

Subject Relevance for Academia

This thesis aims to complement existing overview on NPS factors. By using patent analysis to connect with technological competence, the paper aims to connect knowledge management to NPS factors. Furthermore, from a methodological perspective, this thesis aims to forward machine learning tools in patent analysis for more interesting and predictive insights. Typically, patent analysis studies focus on mapping trends and technologies with the help of patents. This paper turns the focus to the future, using existing correlations through time to guide corporate innovation.

Literature Review

Types of Innovation

Satell (2017) offers that innovation is about problem solving. This is the interpretation that this thesis adheres to, as it translates elegantly into the focus of the research: technology. As technology becomes more granular and complex, the problems tend to follow suit. This opens up numerous avenues and multitudes of possible solutions, each addressing different aspects of the problem at hand. For example, upon the advent of smartphones, while Apple may have its patented touchscreen technology, their solution was not the only one; indeed, other mobile phone manufacturers designed or licensed their own solutions towards this one aspect of phones. Thus, it becomes vitally important to investigate different levels of innovation, in attempts to study this phenomenon.

Broadly speaking, the literature around innovation breaks innovation into two levels: incremental and breakthrough innovation. Clausen & Pohjola (2013) suggest broad definitions for these two levels of innovation:

- Incremental innovation: product innovations that are only new to the firm, but not to the market
- Breakthrough innovation: product innovations that are new to both the firm and the market

The implication here is that incremental innovation seems to be preceded by breakthrough innovation; the first mover with an invention is labelled as the breakthrough innovation. The paper takes this into consideration, and thus examines the number of applications with similar classifications generated immediately after the invention. Of course, a possible shortcoming of this approach is that there might be breakthrough innovations that just do not succeed; and therefore generate any subsequent applications. However, this paper is mainly interested in new product success, thus, such applications might be outside the scope of this paper. Innovation in itself is difficult to measure, its dynamics can be quantified through patent classifications.

Fagerberg et al. (2013) identifies 4 types of innovations: product, process, organisational and marketing. As the latter two cannot be patented, innovations in these areas cannot be explored within patent applications. Tavassoli & Karlsson (2015) examined persistence in each of these types of innovations. Persistence is defined as the influence of past and present innovations on future innovations. The paper concludes persistence is strongest in product innovation, followed by process innovations. This is a particularly relevant finding, as it suggests that future innovations are more strongly influenced by product innovations than process innovations. In the context of this paper, that could imply that product innovations would result in more applications within the set of classifications than with process innovations. This particular point will become clearer in the Data section.

Abiodun (2017) further explored the mediation effect of breakthrough innovations on the relationship between product and process innovation and financial performance, and reported positive significant effects. Furthermore, the results from this paper also suggest that the impact for product innovations is greater than for process innovations. This imbalance could imply that investment could be more readily available for a product innovation project than for process innovation. It would be interesting to note if this could be addressed within the framework of this study. Thus, studying breakthrough innovations in this fashion could carry meaningful implications for the firm's bottom line.

Patent Analysis

Using patent data to map innovation is not a new concept. Nagaoka et al (2010) cover rather extensively the strengths of using patent data as an indicator for innovation, despite its shortcomings. They assert that patent data proves to be a more robust source to analyse the innovation process than research and development expenditure within a company. The authors further maintain that patent data is not friendly to rigorous statistical tools, which is less true today than it was a decade ago. Scherer & Harhoff (2000) indicates that about 10% of the most valuable patents account for more than 80% of the value of all patents, in the German market. This aligns rather interestingly with the aspirations of breakthrough innovations versus incremental innovations.

Numerous studies over the years have used patent analysis to study innovation. Abraham & Moitra (2001) used patent information to investigate foreign investment in India. More recently, Ardito et al (2018) successfully used patent analysis to understand how the Internet of Things has evolved over a period of 2002-2012. They have similarly explored inventions through time from the UK IP office data (the data source for this research) in an attempt to plot the direction of development within the IoT. Costantini et al (2017) used patent analysis to measure the impact of policy changes in energy efficient technology. Both of these papers indicate that patents can be indicators in the path of progress. This paper would like to combine that with statistical tools to better predict innovation actions today. Cho et al (2018), Shubbak (2019) and Yuan & Cai (2021) all focus on using patent analysis and trends to observe nascent technologies. However, this thesis focuses more on the mechanism of patents to guide strategy rather than the evolution of a particular technology.

New Product Success

Evanschitzky et al (2012) builds on the meta-analysis foundation laid by Henard and Szymanski (2001). Both of these studies try to combine and explore the most important success factors when it comes to new product success (NPS). For the context of this research question, the most relevant factors are identified as such:

- Product meets customer needs
- Product advantage
- Product technological sophistication
- Product innovativeness

Their research indicates that all factors except *Product meets customer needs* and *Product innovativeness* are significant predictors of NPS. The former suggests that measuring sales, returns or usage performance might not result in meaningful insights in terms of product success. This was an underlying consideration for the development of the dependent variable in this study. This will be elaborated on in the following section. *Product Innovativeness* is defined as the perceived radicalness, originality or newness of the product. This information is difficult to ascertain with the current dataset.

Additionally, the results state that *Product Advantage* has a stronger impact on goods than services, and Evanschitzky et al. (2012) concluded that *Product Technological Sophistication* was insignificant. *Product Technological Sophistication* is defined as perceived technological sophistication and seems to only consider if a product is "high-tech" or "low-tech". This research believes this definition is constricting, and therefore the actual technological sophistication could be better explored and built upon. In this research, the invention's advantage and technological sophistication can be measured by the number and diversity in the assigned patent codes.

Methods

Operationalisation of Theoretical Concepts

The World Intellectual Property Organisation (WIPO) defines a patent as follows:

> *A patent is an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem.*

A patent thus provides the applicant, as stated, exclusive rights for a limited period of time for the solution. The exclusivity is typically the reward for an organisation's research and development; the organisation would be the only one able to monetise on that particular intellectual property. This paper reasons, therefore, that the patent itself is a proof of a possible solution to a problem; inspecting the patent can provide insight into the problem that it is solving.

As WIPO (2020) indicates, the International Patent Classification (IPC) system is a precise tool to classify patents. Every invention is divided into 8 classes, with an estimated 75,000 subdivisions within each class, each represented by a combination of a Latin alphabet and a series of numbers. This system features a hierarchy, meaning each following character indicates a more specific function within the preceding character's function. This paper is primarily interested in the H category, for electricity. If a patent contains multiple classifications, more information can be gleaned regarding the nature of the problem, or problems that the invention is trying to solve.

The research is broken down into two phases, ultimately seeking to answer each of the questions raised.

Phase 1: Identifying levels of innovation

In this phase, the paper attempts to identify different levels of innovation within the Electricity class (H) of the patents data, by focusing on the classifications assigned to each of the applications. The number and diversity of classifications should signal if an application is intended to be incremental or breakthrough innovation by the applicant company. Descriptive analysis could provide indicators towards the general popularity of the classifications.

Upon inspection, certain applications feature more than one IPC code assigned to them. These would be considered as intended to be breakthrough innovations by the applicant. If the codes fall within the same subdivision, it could be argued that the innovation is a breakthrough within the category. For example, a new flagship camera system by Nikon. If there is a diversity in the subdivisions, it could be argued that the invention is intended to be a breakthrough transcending any one category, such as a new iPhone by Apple, which features many different functionalities across the electricity division. In this phase, the paper can potentially identify the frequency of breakthrough innovations, further differentiating between within-category innovations and radical blue ocean innovations. At this stage, applications with three or fewer codes shall be assigned as incremental innovation, while more would be assigned as breakthrough innovation.

There is a very real possibility of misclassifying breakthrough innovations as incremental innovations, if they only have one single code assigned to them. However, from a purely patent perspective, it is truly difficult to discern this and is a risk. In addition, even if an invention is novel, but it only solves one pre-existing code, it could be argued as an incremental innovation, as it is still finding a new way to improve an existing solution.

Phase 2: Identifying winning combinations

Upon successful identification, it is interesting to understand which patent codes tend to be most successful over time. This would be investigated by observing patent applications, following a granted application, featuring similar patent codes. Per the definition from the literature, the first innovation is a breakthrough innovation and all succeeding innovations tend to be incremental innovations. In the case of either breakthrough innovation, the combination of codes would yield a unique signature for the solution that the invention is trying to solve. Thus, subsequent applications featuring a similar combination of IPC codes could safely be concluded as solving the same problem as the original innovation.

To better understand this phenomenon, perhaps it is educational to look at a different category from the past. In 1996, when JK Rowling created what is now known as the Wizarding World of Harry Potter, it “solved” the problem that the young-adult market faced, by creating literature for this category. In the light of the meteoric success of the franchise, publishers scrambled to uncover their intellectual property to capitalise on this untapped market. Thus, following the launch of Harry Potter and the Philosopher’s Stone, there were numerous other authors and publishers penning their own series; Hunger Games, Divergent, Percy Jackson etc. Harry Potter was the breakthrough innovation, and Hunger Games, Divergent and Percy Jackson are incremental innovations. The success of Harry Potter is signalled by the volume of incremental innovations following it.

Similarly, while PDAs and touch screens were not new, the introduction of the iPhone to the market opened the floodgates for the smartphone market as seen today. In this case, the paper would look at the number of applications for patents with codes coinciding with Apple’s own successful application. The quantity would indicate the perceived profitability of the innovation by the competition. This would be done by employing text analytics on the codes. A co-occurrence matrix could prove instrumental in identifying these patterns.

The dependent variables in this case would result in a large table containing the number of times each patent code appears in an application following the successful patent. This would yield a very cumbersome dataframe. Thus, in order to better understand this data, dimensionality reduction would have to be employed. Since the data will have counts for each patent code, Principal Component Analysis (PCA) is the correct choice.

The PCA is an unsupervised learning technique, developed to reduce the dimensionality of data, with minimal loss in information. It does this by constructing the eponymous principal components: lines constructed through the data, that account for the maximum available variance (James et al., 2021). Each subsequent principal component would be perpendicular to the prior components, to reduce overlap in the variance accounted for. Thus, in this case, hundreds of patent codes and their interactions can be reduced to a handful of components which can act as surrogates for the complete information captured by the dataset. These components also yield interesting insights about the general trends that tend to govern patent assignment and their interactions.

The resulting principal components would try to keep as much of the variance in the tallies intact, while reducing the number of variables to explore. Thus, measuring the impact of the combination of patent classifications in each application on the principal component scores could approximate the effect that the application would have on the tallies. Based on this reasoning, the principal component scores would be considered the dependent variables, while the patent classifications would be the independent variables in this analysis. Thus the formula to be explored is as follows:

$$Principal\ Component\ =\ Patent\ Code\ 1\ *\ Patent\ Code\ 2\ *\ Patent\ Code\ 3\ ...$$

Where,

Principal Component refers to the factor loading for each of the applicable dimensions

Patent Code refers to the number of times a code was assigned to a future application, beyond the application in inspection.

Typically, ordinary least squares (OLS) is widely used due to its simplicity. However, simple models such as OLS suffer from poor predictive performance as the number of variables increases, or if there is correlation between the parameters. This is because OLS is unable to select between variables, leading to potentially overfitted or noisy predictions.

In these cases, penalised regression could be considered. Penalised regression techniques introduce bias in order to reduce variance. Thus, in these cases, a penalty term is introduced to punish variables with large coefficients. Depending on the type of penalty, the impact of unimportant variables is minimised (in the case of a ridge regression) or eliminated (in the case of a lasso regression) (James et al., 2021). In this case, a lasso regression would be essential in eliminating unimportant patent codes. This would help the research focus on the most important patent codes and combinations.

Based on this, a simple OLS model and a random forest model could be built. This is to explore if there is a similar predictive power between a complex model and a simple model. In the latter case, the simple linear model would have to be selected for its computational ease and interpretability. However, if a complex model vastly outperforms the simple model, various tools can be used to estimate the impact of each of the important variables, such as Permutation Feature Importance (PFI) and Accumulated Local Effects (ALE). Interaction effects could be observed by Friedman's H-Statistic.

Random Forest is a type of bagging ensemble technique for supervised learning. As with all ensemble techniques, Random Forest attempts to answer the original question: can a group of weak learners create strong predictions?

Bagging is an amalgamation of bootstrap and aggregating. Thus, as the name suggests, a bagging technique would generate numerous bootstrapped samples from the original dataset and train numerous weak learners on these samples. The predictions from each of these weak learners would then be aggregated into a strong prediction. In the case of Random Forest, the weak learners tend to be decision trees. (James et al., 2021)

While in theory we understand how the Random Forest meta-algorithm generates predictions, it is extremely unclear how each prediction is actually generated. This is due to insufficient information on the bootstrapped samples and the decision trees trained. Due to this, the Random Forest is considered a black box model.

Thus in order to better understand the prediction results, model agnostic black box interpretation tools such as PFI and ALE shall be employed. PFI takes a model and intentionally permutes the features in a model and measures the decline in overall accuracy (Molnar, 2020). It reasons, then, that the variable that causes the largest such drop would have to be the most important feature, i.e. the feature that the algorithm depends on the heaviest for a prediction.

The second tool to be used is the Accumulated Local Effects plot. While the PFI gives an impression on which variables are important, the ALE shows how a prediction might change within a small area of a feature (Molnar, 2020). Thus, ALE helps illustrate how the predictions vary along the important variables. Armed with these two black box interpretation tools, the paper should be able to ascertain which patents and combinations tend to be lucrative for a company.

Friedman's H-Statistic is an important tool for exploring potential interactions between variables. The statistic is designed to measure how much of the variation of a prediction depends on the interaction between two variables (Molnar, 2020). It measures the difference between the observed partial dependence functions and the sum of the variables in question. This difference represents the interaction between the variables being investigated. The amount of variance explained by the interaction is used as the strength of the interaction. If it is 0, it indicates no interaction between the variables, and a value of 1 indicates that all of the variance explained by the variables in question is a result of their interaction effect.

Thus, by the end of this phase, the impact of innovations should be codified and a model suggesting successful patent codes and lucrative patent combinations should be revealed. After understanding this, it is important to examine any commonalities in the innovation path taken by a company to reach this stage.

Data

The data source for this thesis comes from the UK Intellectual Property Office. It is a repository of patent applications ranging from 1978 to 2018. The dataset provides a wealth of information in the evolution of technology, which proves indispensable to the perspective of the research, as covered earlier. The full list of variables can be found here. Following is an overview of the variables key to this research.

Variable	Description	Remarks
Application number	The reference number given to a patent at application.	This is a number assigned to each application (even if it were for a revision of an older application). This number would be handy in combining existing applications together to track progress.
Filing date	The date of an earlier application (often called the priority application) containing a disclosure of the invention. If there are no earlier patent applications then the earliest application date is the same as the filing date.	This would be handy, again, in tracking an invention's progress through the IPO's application process. This would be an interesting variable to explore in the final stage. In addition, this would also help track the subsequent applications made following a landmark invention.
B publication date	The date the patent was granted.	This would be the day from which a patent would be in effect. Thus, the applicant can command monopoly on that specific invention or solution for the problem.

IPC 7	International Patent Classification (IPC) 7 shows the technology areas of the invention the subclass level, the first four digits (IPC definitions), version 7). This version of the IPC covers patents from January 2000 to December 2005. Some applications may be classified under both IPC7 and IPC8.	This is for the 7th edition of the IPC codes. The class, subclass level and the numeric digits help narrow the type of problems being solved. As mentioned earlier, a combination of codes with the same subclass would indicate that the problem is very category specific, whereas different subclasses indicate a broader radical innovation.
IPC 8	International Patent Classification (IPC) 8 shows the technology areas of the invention the subclass level, the first four digits (IPC definitions), see the most recent version). Before this version the IPC would be subject to large changes every few years. This version is updated annually.	The same as the IPC 7 but with the 8th edition. There is a dictionary that helps bridge any discrepancies between the two editions.

Table 1: Variables to be used in this research from the base dataset

Data Preparation

As mentioned earlier, the research only selected the applications with at least one H category in them. This narrowed the number of applications from 557,153 to 100,679. The valuable information was stored with two columns named IPC7 and IPC8. Prior to 2006, the WIPO released a new version to the intellectual property classifications at around every five years. From 2006, this update came more frequently, and fewer categories needed to be changed. This allowed the WIPO to be more flexible in adapting to the rapidly changing landscape in technology.

Thus IPC7 contained classifications updated until 2006. The IPC8 column is up to date to the most recent version (2017). Understandably, this indicated that there were gaps to be bridged between 2006 and 2017. This is where the Revision Concordance List (RCL) became instrumental. Thus, accounting for this, both columns were collapsed into one IPC column featuring the updated classifications as of 2017. This allowed for seamless interpretation between the two editions.

Once unified into one column, it was clear that the data needed to be prepared in order to isolate each classification assigned. Thus, tokenising the classifications into a document-term-matrix (DTM) proved to be a satisfactory solution. This matrix is typically employed in text analytics projects, where a variable might feature large bodies of text. A DTM takes this column and creates a matrix which essentially creates tallies for the number of times a term appears within a document. The columns in such a DTM represent the entire set of terms within all documents in a given corpus. While this research is not a text analysis, the resulting DTM succinctly transformed into a matrix where the documents were the applications and the terms were the individual IPC classifications. This proved to be an efficient way to create tallies of all classifications within the 100,679 applications that were initially added.

At this stage, a particular choice had to be made. Each IPC had up to 8 bits, with each bit contributing to a more granular level. Thus, the more bits considered, the more specific the function, the more specific the problem solved. This is best illustrated in the following example:

IPC: **H04N 9/31**

Breakdown:
H refers to the category *Electricity*
04 refers to the subcategory *Electrical Communication Technique*
N refers to *Pictorial Communication e.g. Television*
9/00 refers to *Details of Colour Television Systems*
31 refers to *Projection Devices for Colour Picture Display*

On first glance, the maximum granularity would seem to be ideal, however this leads to numerous practical problems. Firstly, with each level the complexity of problem definition increases almost exponentially. This means that the size of the matrix also increases, leading to practical problems, such as limited hardware. As will be explored a bit more in depth, most of the inventions in the dataset were also incremental. This indicates that not only would the resulting DTM have a large number of terms, the overwhelming majority of them would only be referred to once or twice. This yields an extremely sparse matrix, which also happens to be very resource expensive. Otherwise there might be overfitting, and the number of classifications balloons to an even more sparse matrix with over 17,050 classifications for very specific issues.

Instead of using the full codes, only the first four letters were used. This yielded another considerably sparse matrix, with a total of 527 different codes across all of the applications, in the category. Thus, the resulting matrix yields about 3% of the size of the full matrix. Both DTMs are completely sparse, as the table below suggests. In addition, the larger matrix has a total of 207,325 cells that are non-sparse (having a value greater than 0), and while the smaller matrix, understandably, has fewer non-sparse cells, the loss was not proportional

(150,035). Therefore, only the first four letters (**H04N** as opposed to **H04N 9/31**) were used, as the specificity did not seem to efficiently improve upon the information derived from the dataset, for this thesis.

The full descriptions of classifications can be seen here/in the appendix. The distribution of classifications were as follows:

Number of classifications	Number of applications
1	42,748
2	30,960
3	15,325
>3	11,646

Table 2: Overview of distribution of number of classifications

Thus, it becomes apparent that only 11,646 of the 100,679 applications feature more than 3 classifications, with the maximum having a total of 28 classifications. Furthermore, the oldest entry is from 23/05/1978 and the most recent entry is from 11/04/2018. Thus, the source spans approximately 40 years. The highest number of different classifications was 4, and the highest tally for any single classification in an application was 18. With this, the data was prepared for analysis.

Results

Phase 1

This phase focused on differentiating between incremental and breakthrough innovation. As the table in the previous section revealed, there were only 11,646 applications that had more than 3 classifications. Of these, only 7,155 had an assigned B Publication Date; the date when the patent was granted. Thus, only 7.1% of all applications with at least 1 category H classification were approved applications for breakthrough invention. Furthermore, within all filed applications for breakthrough invention, the approval rate was 61.4%. It is interesting to note that, not only does breakthrough innovation generally tend to be the minority (at least from a patent perspective), an application also tends to be approved less than average (68.5% in category H; 65.9% across all categories).

As mentioned earlier, it is difficult to measure the success of incremental innovations. Thus, this research focuses on the 7,155 innovations with greater than three classifications, to try and map out their success and the factors at play.

Phase 2

After differentiating between the incremental innovation and the breakthrough innovation, it became imperative to understand which of these were most successful. This would be marked by the number of applications filed after these were granted. There was a choice to be made on whether to consider the day the patent was published or approved, but only the latter was considered as it would be rather difficult to predict the success of an invention on the basis of patent publication alone.

For each of the approved breakthrough applications, a period of three years was considered. This time period was derived from Gerkin et al. (2015), where it was observed that the time lag from application from product launch was generally between 2 and 3 years for the specific technology categories listed. For each of the classification assigned to an application, tallies were aggregated over this period in the future. Thus, a measure of the invention's success could be gathered. Based on this, it is interesting to consider what might be the differences between the successful and unsuccessful inventions. More specifically, if there were particular classifications that typically tended to be more successful than others, and if there were interesting interactions between the classifications.

Due to the large number of potential dependent variables, it was imperative to narrow down the scope and try to see if these measures of success exhibited any overarching patterns. As planned, PCA was employed to reveal if any meaning could be distilled by reducing the number of dimensions. The following scree plot succinctly captures the results from the PCA.

As figure 1 illustrates, there are two discernable elbows, one occurring between 2 and 3 dimensions and one occurring between 4 and 5 dimensions. In addition to this, at 4 dimensions, 82.5% of the total variance in the variables is accounted for. Thus, while there is no stated objective criteria for selection, based on the rules of thumb outlined in James, 2021, 4 dimensions were selected as the appropriate number of dimensions to reduce to. Based on the classifications and how they correlate with each of the dimensions, the dimensions could be interpreted. The dimensions were interpreted as Communication, Images, Manufacturing and Transmission, respectively. The reasoning and analysis behind these interpretations are elaborated upon in the following subsections.

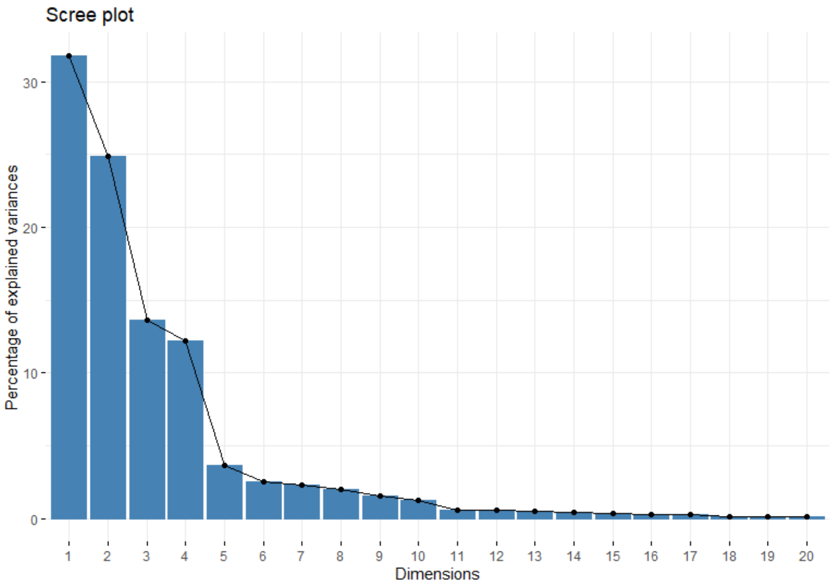

Figure 1: Scree plot of PCA conducted on success measures

Armed with the interpretation and understanding of the variables, it made sense to sift through the dataset to better understand which variables and interactions best contributed to each dimension. However, considering the still large number of terms, it is unfeasible to gauge every potential interaction between each of the terms, as this yields about 4.393×10^{158} terms. Furthermore, as the vast majority of this matrix is sparse, this exercise would prove fruitless for the overwhelming majority of the terms. Thus, a selection had to be made on the terms alone, prior to exploring potential interaction.

Thus, penalised regression was employed to eliminate the extraneous variables. Interesting to note that a large proportion of the variables did not have a variance, meaning they had only a singular value (0). This was conducted for each of the dimensions, and the 10 most important classification terms were examined. These are reported in the following subsections. Subsequently, the interactions between these terms were also explored with a penalised regression. While, again, it was tempting to consider all possible combinations of interactions between the 10 terms, this would still yield 1024 total combinations, with unequal importance on predicting the dimension. However, the sparsity again indicates that a large number of the interactions either do not exist or are simply not important. The most important of these are also reported in the following subsections.

Thus, after exploring 4 dimensions to success, and the terms and interactions contributing to each dimension, the next logical step was to see how these terms and interactions contributed to the dimension. As explored in the previous sections, a random forest model and a linear regression model each were built to facilitate interpretation. However, the random forest model consistently featured lower root mean squared error across all dimensions. These will also be explored in depth in the following subsections.

Principal Component 1 - Communication

The first Principal Component Dimension is assigned the interpretation of "Communication". This dimension accounts for 31.76% of the variation in the data, which is the highest for any singular dimension. No other dimension captures as much of the variation in the rest of the dataset. Taking this into account, it is important to see why this interpretation was assigned, and what it reveals about the nature of the data.

According to the results of the lasso regression, the most important terms and their impact on the dimension are as follows. The model was best at a lambda value of 5.179475.

Variable	**Importance**
H04L - *Transmission Of Digital Information, E.G. Telegraphic Communication*	100.00
H04W - *Wireless Communication Networks*	32.63
G06F - *Electric Digital Data Processing*	24.48
H01L - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	15.61
H04Q - *Selecting*	12.53
H04B - *Transmission*	10.45
H04N - *Pictorial Communication, E.G. Television*	9.89
H03M - *Coding, Decoding Or Code Conversion, In General*	9.13
H04J - *Multiplex Communication*	8.43
H04M - *Telephonic Communication*	7.10

Table 3: Variable Importance from penalised regressions for Dimension 1

As each of the variables were standardised before training the model, the importance is relative to the other variables. Thus, at a cursory glance, it becomes apparent that **H04L** (*Transmission Of Digital Information, e.g. Telegraphic Communication*) is, by far, the most important variable. The next most important variables are relatively similar in impact. Variable importance in the glmnet package is measured by measuring the coefficients of the variables. Seven out of the ten most important variables fall within **H04** classification, which relates to *Electric Communication Technique*. This lends somewhat to the interpretation of this dimension as pertaining to communications technologies. The only non-**H04** codes are **G06F**, **H01L** and **H03M**. All of these appear to relate to either generating or converting electrical signals, which is understandable considering the most important variable.

Judging by the confluence of patent classifications pertaining to transmission of electrical signals and data, it becomes considerably clear why this interpretation has been assigned. Furthermore, as Appendix 2 reveals, each of these variables are negatively correlated with the dimension, with the exception of **H01L**. This seems to indicate that this dimension captures, at some level, the explosion in innovation in communication. Variation along this dimension could possibly relate to the implementation of the Internet-of-Things (IoT); on the negative dimension, devices that are connected and on the positive side, inventions that are stand alone. Once the most important variables were identified, it could be safely concluded that the most interesting interactions would occur amongst these, as recall that the DTM is a sparse matrix.

Thus, the interactions between these 10 variables are tabulated below. Only the top 1 terms are recorded as the importance of the remaining 9 selected variables were close to 0.

Interaction	**Importance**
H04L	100.00
H04W	25.37
G06F	13.71
H01L	12.75
H04L:G06F	8.10
H04L:G06F:H04N	6.96
H04B	6.69
H04L:H04Q	5.70
H04L:H04M	4.74
H04L:H04W	4.07

Table 4: Most important interactions for Dimension 1

The new penalised regression model selected 19 variables, 10 of which were interaction effects. Every selected interaction effect was between **H04L** and other variables, indicating just how important this classification is to predicting the value on the dimension. Furthermore, upon introducing the interaction effects, the main effect of **H04M** (*Telephonic Communication*) is unselected by the model. Instead, its interaction with **H04L** and other variables are selected, with minimal importance (**H04L:H04M** - 4.74; **H04L:G06F:H04M** - 0.73; **H04L:H04Q:H04M** - 0.52). This is somewhat understandable considering the relative importance of the **H04L** classification and the closeness of **H04L** and **H04M**, at least per the definition provided by WIPO. Furthermore, the interaction effects are interesting to investigate, as these will provide insights into the success of the invention.

Based on these main and interaction effects, two models were trained; OLS and Random Forest. This was to see if there was a difference in the predictive power between the two models. 80% of the rows were assigned at random to the training set and 20% to the testing set. On both these sets, the Random Forest outperformed the OLS, on both percentage of variance explained (79.65% to 63.5%) and RMSE (280.08 to 356.17).

While measuring the feature importance of the Random Forest model in figure 2, it was apparent that **H04L** was considered by far the most important variable, followed by **H04W** and **H04N**. Beyond these variables, all other variables seem to produce comparable MSE upon permutations, implying that these variables might not impact the overall predictability of the variance in Dimension 1. Based on this, it was important to measure if there were any interesting interactions between the variables. For this the Friedman's H-statistic was employed. The following figure reveals the interaction strength between the variables.

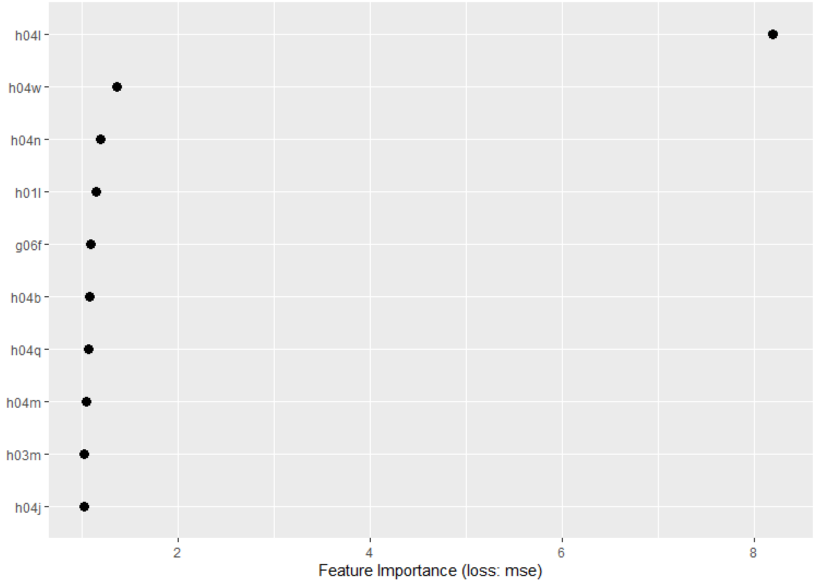

Figure 2: Permutation Feature Importance for Dimension 1

It becomes immediately apparent, from figure 3, **H04L** also has the greatest overall interaction strength (0.05). However a value close to 0 implies no meaningful interactions. This indicates that while **H04L** has the highest potential for interactions between the other variables, it is highly unlikely that any of the most important variables interact with each other, or, apparently, have meaningful impact on the dependent variable, outside of **H04L**.

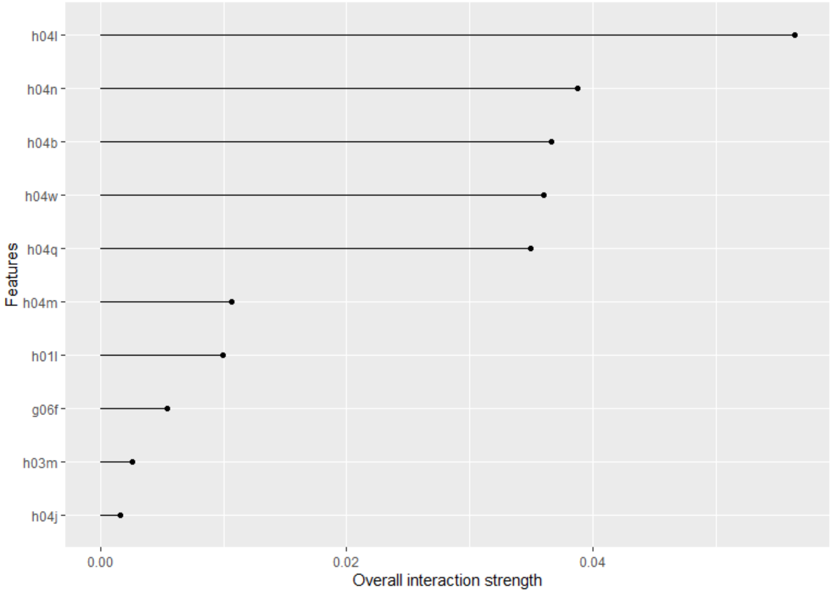

Figure 3: Overall Interaction Strength for Dimension 1

This finding is particularly in contrast with the results from the penalised regression, which indicated the **H04W** and **G06F** also have some level of impact on the dependent variable. However, the random forest seems to disagree with that assessment. Both Random Forest and Penalised Regression do not seem to expect any of the interactions to be meaningfully important to the scores of Dimension 1.

The ALE plots reveal the general direction within which the dependent variable, here the dimension score, varies across different values of each variable. The correlations suggested that every variable had a negative correlation with the principal component, with the exception of **H01L**. However, the ALE plot allows for more granularity in understanding the variation at each possible value of the independent variable.

Looking at the ALE plots in Appendix 2, it becomes immediately apparent how these negative correlations were calculated. Six of the top ten variables feature a steep reduction in the score, between values of 0 and 1. However, for a value of 1 and higher, the dimension score has a positive correlation. This could be attributable to the construction of the dataset. As the dataset is exceedingly sparse, and includes exclusively breakthrough innovations, a value of 0 for any variable means a high score on the other classifiers (as a value under 3 is not possible).

Thus, from a value of 1 and above, **H04L**, **G06F** and **H04N** tend to be positively correlated with the dimension score. **H04W**, **H04Q**, **H04B**, **H04J**, **H04M** and **H03M** tend to be negatively correlated with the dimension score. Furthermore, **H01L** tends to be positively correlated between values of 1 and 3, but negatively correlated beyond a value of 3. The implications of these findings are better discussed in the Conclusions section.

Principal Component 2 - Images

The second Principal Component Dimension is assigned the interpretation of “Images”. This dimension accounts for 24.87% of the variation in the data, which is the second highest. This seems to indicate that the images and the processing thereof is the second most important consideration for a manager, following the communication dimension explored earlier. Following the penalised regression, these are the terms that are most significant to its prediction. The best model was trained with a lambda value of 10.

Variable	**Importance**
H04N - *Pictorial Communication, E.G. Television*	100.00
G03B - *Apparatus Or Arrangements For Taking Photographs Or For Projecting Or Viewing Them; Apparatus Or Arrangements Employing Analogous Techniques Using Waves Other Than Optical Waves; Accessories Therefor*	26.13
G06T - *Image Data Processing Or Generation, In General*	23.26
H04W - *Wireless Communication Networks*	16.21
H01L - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	12.43
G11B - *Information Storage Based On Relative Movement Between Record Carrier And Transducer*	11.81
G03H - *Holographic Processes Or Apparatus*	10.09
H04L - *Transmission Of Digital Information, E.G. Telegraphic Communication*	9.92

G06K - *Recognition Of Data; Presentation Of Data; Record Carriers; Handling Record Carriers*	8.03
G02B - *Optical Elements, Systems, Or Apparatus*	8.02

Table 5: Variable Importance from penalised regressions for Dimension 2

As table 5 reveals, this dimension has a larger focus on image and image processing technology. Accordingly, while the data set was centered around the category H (electricity), 6 of the 10 most important features selected belong to category G (Physics). Due to this increased emphasis on image processing, optics and general storage, the interpretation of Images is afforded to this dimension. The most important classification is **H04N** (*Pictorial Communication, E.G. Television*), which lends the most to the interpretation. As with the previous dimension, the other classifications appear to have considerably lower relative importance compared to this one. Interestingly, the correlation for all variables with the principal component are positive, with the exception of **H04W** (*Wireless Communication Networks*), **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*) and **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*), which are weakly negative. The interactions between these ten variables are recorded in the following table, based on their importance according to penalised regression.

Variable	**Importance**
H04N	100.00
H04N:H04L	25.05
G06T	21.46
H04N:G03B	15.91
H04W	13.57
H04N:G11B	12.78
H04L	12.00
G03B	11.42
H01L	10.92
H04N:G06K	8.55

Table 6: Most important main and interaction effects for dimension 2

In the most important interactions, similar to the previous principal component, most of them feature **H04N**, the most important variable in this regression. However, this effect is not as stark as with **H04L**, in the previous dimension. The importance of H04L also increases when the interactions are introduced (from 9.92 to 12.00), along with the interaction **H04N:H04L** being the second most important variable. On the other hand, the main effects of **G11B** (*Information Storage Based On Relative Movement Between Record Carrier And Transducer*) and **G06K** (*Recognition Of Data; Presentation Of Data; Record Carriers; Handling Record Carriers*) completely disappear from the most important variables as soon as the interactions are introduced. This seems to imply that while both these technologies are not very significant to this dimension on their own, they seem to play a supporting role for the more relevant technologies; **G11B** appears on 2 of the top 20 interactions, while **G06K** appears on 4 such interactions.

As with the previous dimension, OLS and Random Forest models were trained, to try and explain how these interactions impact the main effects. As mentioned earlier, the random forest model outperformed OLS in both variance explained (74.55% to 48.90%) and RMSE (309.52 to 412.23). Figure 4 explores the PFI for all the important variables for dimension 2.

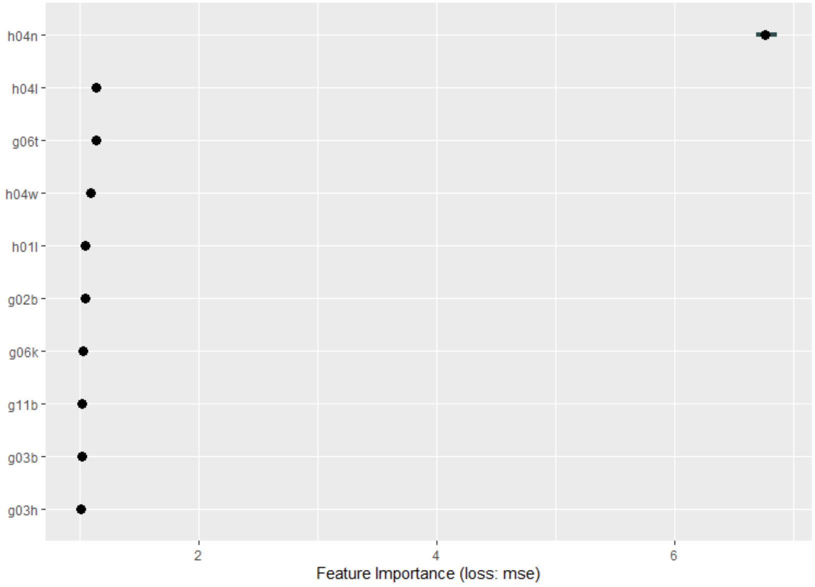

Figure 4: Permutation Feature Importance for Dimension 2

As figure 4 reveals, **H04N** is the most important variable for the prediction of this dimension. However, the other variables seemingly have minimal impact on the loss of MSE. This indicates that the model relies very heavily on **H04N** for the prediction of this variable, however, and less on the other variables. Furthermore, a cursory glance at the interactions reveals another interesting finding. The overall interactions are presented below, through Friedman's H-Statistic.

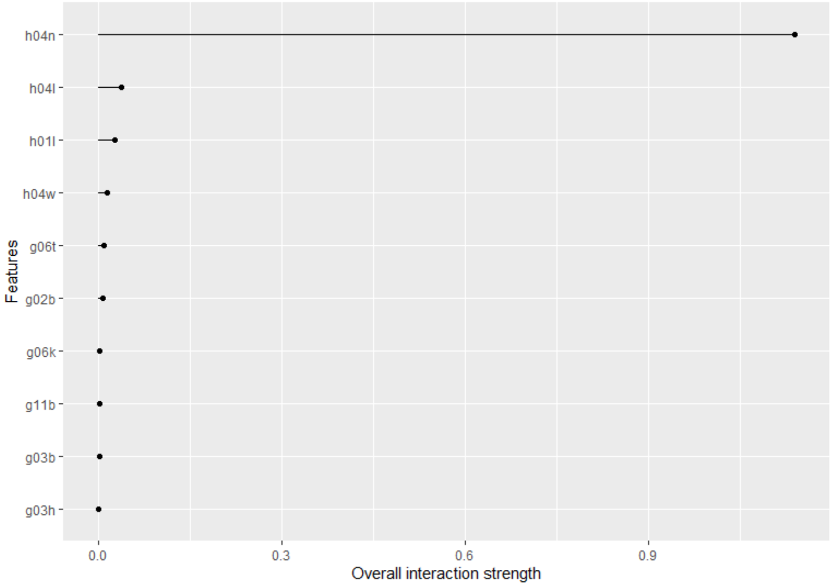

Figure 5: Overall Interactions in Dimension 2

H04N has an extremely high interaction potential, with a score of 1.14. This indicates that there is a very high likelihood that **H04N** interacts with another variable. **H04L** and **H01L** also carry some low interaction potential (0.04 and 0.03). Thus, based on these findings, interactions between **H04N** were investigated in greater depth. Figure 6 reveals the results of possible two-way interactions between **H04N** and other variables.

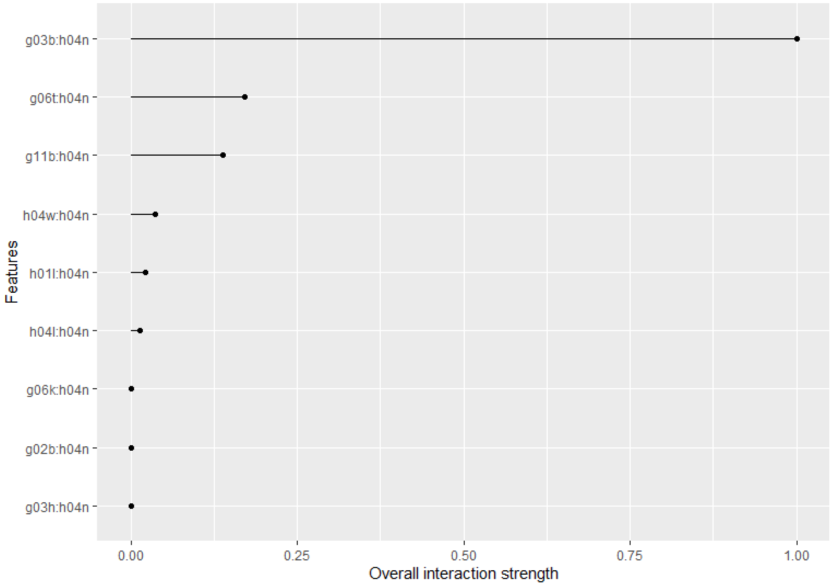

Figure 6: Two way interactions between H04N and other variables

The interactions reveal an extremely high possibility for **G03B:H04N** (1.00). This seems to indicate that, within the technologies, there is a strong interaction between *Holographic Processes Or Apparatus* and *Pictorial Communication, E.G. Television*. Put this way, the interaction seems surprisingly mundane, but it perhaps could further be explained by the numerous forays into 3D televisions in the late 2000s and early 2010s. Interestingly, **G03B** does not seem to have a strong main effect of its own, based on PFI results. This seems to indicate that holographic processes or apparatuses do not seem interesting to develop on their own, but only within a larger, more established technology. Holographic displays are often used within science fiction media, but these results seem to contextualise how such an invention might come to be.

Furthermore, the interactions between **G06T** and **G11B** with **H04N** also indicate some potential (0.17 and 0.14). This could indicate that *Image Data Processing Or Generation, In General* and *Information Storage Based On Relative Movement Between Record Carrier And Transducer* tend to have a somewhat supporting relationship with Pictorial Communication. This could perhaps be better attributable to the rise in AI in image processing, such as Google's Vision AI.

The ALE plots from Appendix 2 for dimension 2 reveal very interesting trends. Similar to dimension 1, the correlations appear to be severely offset by the values assigned to the 0 value to 4 of the 10 variables. **G03B**, **G06T**, **G11B**, **G03H**, **G06K** and **G02B** appear to be positively correlated with the dimension. In the case of **G03B**, the entire effect could be captured by its interaction with **H04N**. **H04W** appears to be negatively correlated with the dimension. **H04N** is weakly positively correlated between the values of 1 and 3, and negatively correlated for values greater than 3. Conversely, **H01L** appears to be weakly negatively correlated between values of 1 and 3, and positively correlated for values greater than 3. **H04L** appears to be strongly positively correlated between values of 1 and 2, and weakly positively correlated for values beyond 2. The results will be fully interpreted in the Conclusions and Discussions section.

Principal Component 3 - Manufacturing

Principal Component 3 is given the interpretation of Manufacturing. As will be apparent later, this dimension focuses heavily on the process and fabrication of materials, particularly metals. This dimension accounts for 13.65% of the variance in the data. Considering the important variables and innovation theory, it is very interesting that industrial innovation is considered a key dimension to capture the dimensions.

The following are the classifications that are most significant to its prediction. These are the interaction effects that are most significant. The best model was trained with a lambda value of 5.179475.

Variable	Importance
H01L - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	100.00
H04W - *Wireless Communication Networks*	27.56
H04L - *Transmission Of Digital Information, E.G. Telegraphic Communication*	26.77
F24J - *Production Or Use Of Heat Not Otherwise Provided For*	11.45
B82Y - *Specific Uses Or Applications Of Nanostructures; Measurement Or Analysis Of Nanostructures; Manufacture Or Treatment Of Nanostructures*	7.20
C07F - *Acyclic, Carbocyclic, Or Heterocyclic Compounds Containing Elements Other Than Carbon, Hydrogen, Halogen, Oxygen, Nitrogen, Sulfur, Selenium Or Tellurium*	7.12
C23C - *Coating Metallic Material; Coating Material With Metallic Material; Surface Treatment Of Metallic Material By Diffusion Into The Surface, By Chemical Conversion Or Substitution; Coating By Vacuum Evaporation, By Sputtering, By Ion Implantation Or By Chemical Vapour Deposition, In General*	6.68
H04N - *Pictorial Communication, E.G. Television*	6.51
G06F - *Electric Digital Data Processing*	6.42
C08G - *Macromolecular Compounds Obtained Otherwise Than By Reactions Only Involving Carbon-to-carbon Unsaturated Bonds*	6.41

Table 7: Variable Importance from penalised regressions for Dimension 3

As table 7 illustrates, the variable most important to dimension 3 is **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*), followed by **H04W** (*Wireless Communication Networks*) and **H04L** (*Transmission Of Digital Information, E.G. Telegraphic*

Communication). Furthermore, there is a considerable drop off from these three to the next most important variable. This could be attributable, again, to the sparsity of the matrix. The implications of this will be discussed later in an appropriate chapter. This dimension features a number of important classifications from category C (Chemistry) and one for category B (*Performing Operations; Transporting*) and F (*Mechanical Engineering; Lighting; Heating; Weapons; Blasting*). Thus, the manufacture and industrial design connotation seems rather appropriate for this dimension. With the exception of **H04N** (*Pictorial Communication, E.G. Television*) and **H04W**, all other variables also have a negative correlation with the dimension score. **H04W** is weakly positively correlated, while **H04N** is almost uncorrelated (Appendix 1).

Based on these 10 variables, the interactions were generated and selected. The 10 most important interactions are as follows:

Variable	**Importance**
H01L	100.00
H04L	25.33
H04W	24.23
H01L:B82Y	13.37
H01L:H04N	11.77
F24J	11.72
H01L:C23C	10.37
H01L:H04L	8.84
C08G	7.54
C07F	6.65

Table 8: Most important main and interaction effects for Dimension 3

Upon introducing interaction effects, the 20 most important selected variables feature 11 interactions. 10 of these 11 are interactions of various variables and the most important variable in this dimension, **H01L**. Interestingly, not only does the main effect of **H04L** become more important, it also features in 6 of the 11 selected interaction effects. Interestingly, the fifth most important variable, **B82Y** (*Specific Uses Or Applications Of Nanostructures; Measurement Or Analysis Of Nanostructures; Manufacture Or Treatment Of Nanostructures*), completely disappears when the interaction effects are introduced. In addition to this, it also appears in only one interaction effect, **H01L:B82Y**, the fourth most important variable overall. This seems to imply that this variable, as observed earlier, might play an important role in supporting the impact of **H01L**, rather than an impact of its own. This might be crucial in differentiating the role of the more ubiquitous classifications that have been noted in every dimension thus far.

In order to better understand the nature of these relationships, two models were again trained. As with each previous dimension, the Random Forest model outperformed OLS in both variance explained (81.16% to 59.99%) and MSE (178.77 to 269.97). The following paragraphs attempt to better interpret the results from the random forest model, beginning with figure 7, which covers the results of PFI in dimension 3.

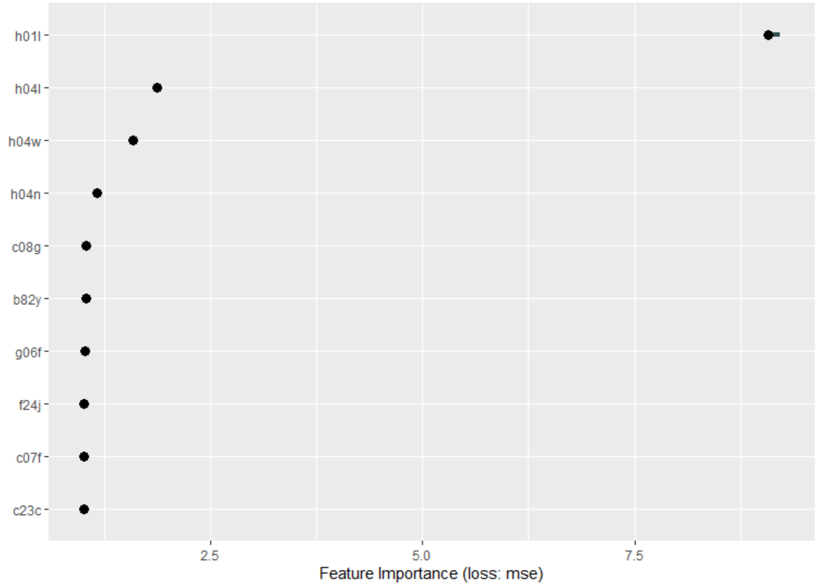

Figure 7: Permutation Feature Importance for Dimension 3

As Figure 7 reveals, the Random Forest model relies extremely heavily on **H01L**. The variables most important after this are **H04L** and **H04W**. This is very consistent with the results from penalised regression. Aside from these three, only **H04N** seems to impact MSE upon permutation, indicating that the Random Forest does not rely on the rest of the selected features. This is similar to the results of the penalised regression, where after **F24J**, all other variables had very low importance in predicting the dependent variable.

Much like with Dimension 1, figure 8 reveals very weak interaction possibilities. **H04L** scores the highest on the H-Statistic, with only 0.06. This indicates that there is very little possibility for interaction between any of the variables. This differs from the results found in penalised regression, which indicated that interactions **H01L:B82Y** and **H01L:H04N** both had some importance on the regression.

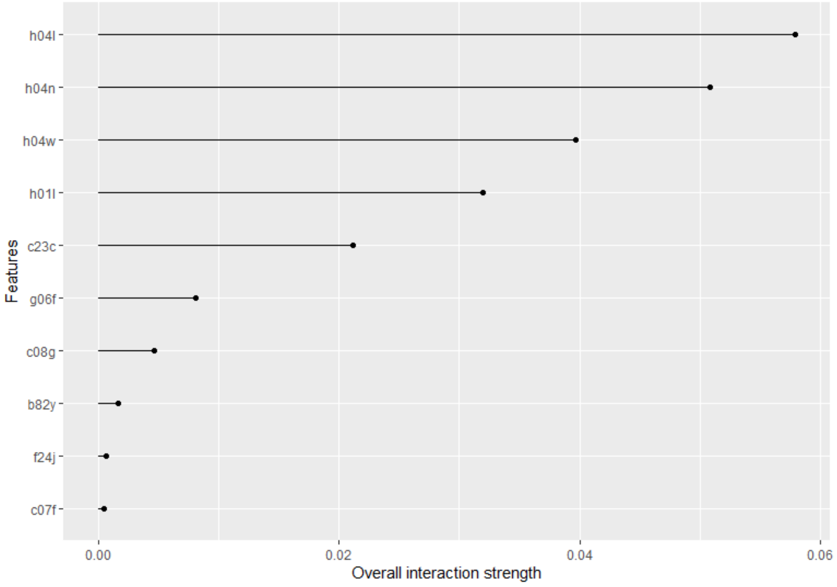

Figure 8: Overall Interactions for Dimension 3

As with the dimensions (Appendix 2), the correlations appear to have been heavily influenced by variables with a value of 0. 5 of the 10 variables experience a sharp shift between 0 and 1, resulting in a negative correlation coefficient for each of the variables. However, as before, discounting for this effect reveals patterns in variation. **H04W**, **B82Y** and **G06F** appear to be positively correlated with the dimension, beyond the value of 1. **F24J**, **C07F**, **C23C** and **C08G** appear to be negatively correlated. Three variables exhibit non-standard behaviour. **H04N** is positively correlated between 1 and 3, however the effect becomes weaker for a value greater than 3. **H04L** presents a weak negative effect on the dimension score between 1 and 2, and a positive effect for a value greater than 2. **H01L** similarly exhibits a decline in the score between values of 1 and 3, before providing a positive effect.

Principal Component 4 - Transmission

The fourth and final dimension considered for this paper is interpreted as Transmission. There is a considerable overlap with dimension 1, Communication, and the exact contributions to the interpretation of this dimension and its differences with the first one shall be explored shortly. It accounts for 12.24% of the variance explained in the data set.

The following are the ten most important classifications as selected by the penalised regression. The best model was trained with a lambda value of 3.727594.

Variable	Importance
H04W - *Wireless Communication Networks*	100.00
H01L - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	31.27
H04L - *Transmission Of Digital Information, E.G. Telegraphic Communication*	23.34
H04Q - *Selecting*	11.86
H04N - *Pictorial Communication, E.G. Television*	11.58
G06F - *Electric Digital Data Processing*	6.34
G01S - *Radio Direction-finding; Radio Navigation; Determining Distance Or Velocity By Use Of Radio Waves; Locating Or Presence-detecting By Use Of The Reflection Or Reradiation Of Radio Waves; Analogous Arrangements Using Other Waves*	3.88
H03M - *Coding, Decoding Or Code Conversion, In General*	3.82
G07G - *Registering The Receipt Of Cash, Valuables, Or Tokens*	3.62
F24J - *Production Or Use Of Heat Not Otherwise Provided For*	2.99

Table 9: Variable Importance from penalised regressions for Dimension 4

Based on table 9, **H04W** (*Wireless Communication Networks*), **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*) and **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*) immediately become apparent as the most important variables. However, where there were enough distinctive important variables to help interpretation, this is, however, somewhat missing in this particular dimension. This could potentially also be an explanation for the low variance explained by this particular dimension, compared to the rest. Following the three most important variables are **H04Q** (*Selecting*) and **H04N** (*Pictorial Communication, E.G. Television*), and the rest of selected variables tend to have a considerably low importance score. The inclusion of **G06F** (*Electric Digital Data Processing*) and **G01S** (*Radio Direction-finding; Radio Navigation; Determining Distance Or Velocity By Use Of Radio Waves; Locating Or Presence-detecting By Use Of The Reflection Or Reradiation Of Radio Waves; Analogous Arrangements Using Other Waves*) implies that this dimension focuses more on signals and transmissions, where the first dimension is more concerned with a larger process, emergent from aggregating signals. The last two important variables **G07G** (*Registering The Receipt Of Cash, Valuables, Or Tokens*) and **F24J** (*Production Or Use Of Heat Not Otherwise Provided Fo*) seem rather confusing for the interpretation, however, as Appendix 1 suggests, along with **G01S** and **H04N**, they are nearly uncorrelated with dimension. Appendix 1 also reveals that, of the remaining 6 important variables, **H04W** is strongly positively correlated, **H01L** is weakly positively correlated, while all others are weakly negatively correlated with the dimension.

Perhaps the interaction effects between these variables would shed some light on the interpretation. The following contains the most important main and interaction effects selected again by penalised regression.

Variable	**Importance**
H04W	100.00
H01L	33.42
H04L	30.30
H04W:H04L	23.25
H04W:G06F	14.79
H04N	9.86
H04Q	8.63
H04L:G06F	8.14
H04W:G01S	5.27
H04W:H04Q:G06F	2.64

Table 10: The most important main and interaction effects for Dimension 4

Based on the interactions, it could be concluded that this particular dimension is defined by **H04W**, **H01L** and **H04L**. The fourth most important variable is the interaction between **H04W** and **H04L**. No other variable has an importance score over 15. Furthermore, the twentieth most important variable (**H04L:G07G**) has an importance of 4.28 × 10^{-14}, which is so low, it is effectively equal to zero, relative to the other terms. Furthermore, 13 of the top 20 selected variables are interactions, and the main effects for **G01S**, **G07G** and **F24J** are dropped. Furthermore, **F24J** does not even appear in an interaction term, while **G07G** appears twice, but both terms with under 0.5 score on importance.

As with every dimension, OLS and Random Forest were constructed, and Random Forest outperformed OLS in both percentage of variance explained (77.16% to 60.45%) and in MSE (174.35 to 245.93). Figure 9 presents the results from PFI from the random forest model on dimension 4.

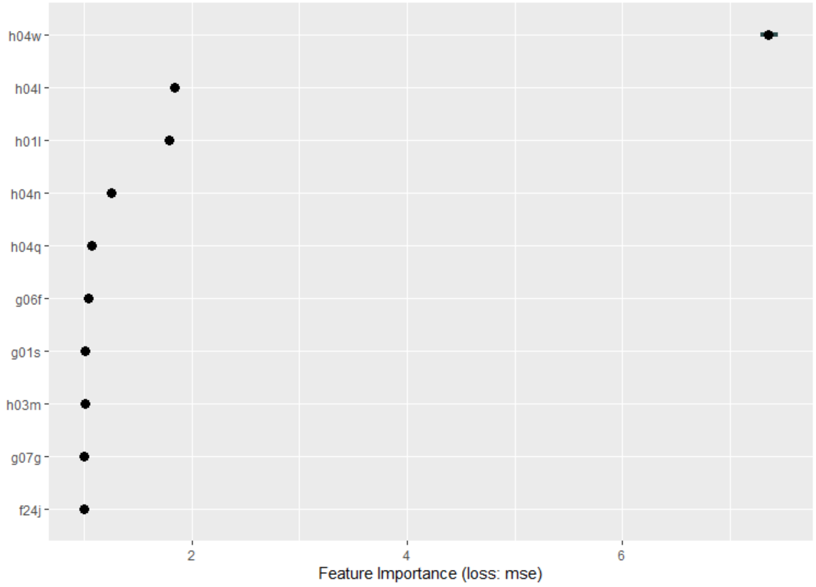

Figure 9: Permutation Feature Importance for Dimension 4

Feature Importance reveals that **H04W** is the most important variable for the model. Following this, **H04L** and **H01L** are the next most important variables, followed by **H04N**. A notable difference between the results from PFI and the penalised regression is the importance of the variable **H04Q** (*Selecting*), which seems to barely have an impact on MSE, upon permutation. This indicates that the random forest does not rely on **H04Q** for predictions in Dimension 4. The following are the results from Friedman's H-Statistic, measuring the interaction potential between the variables.

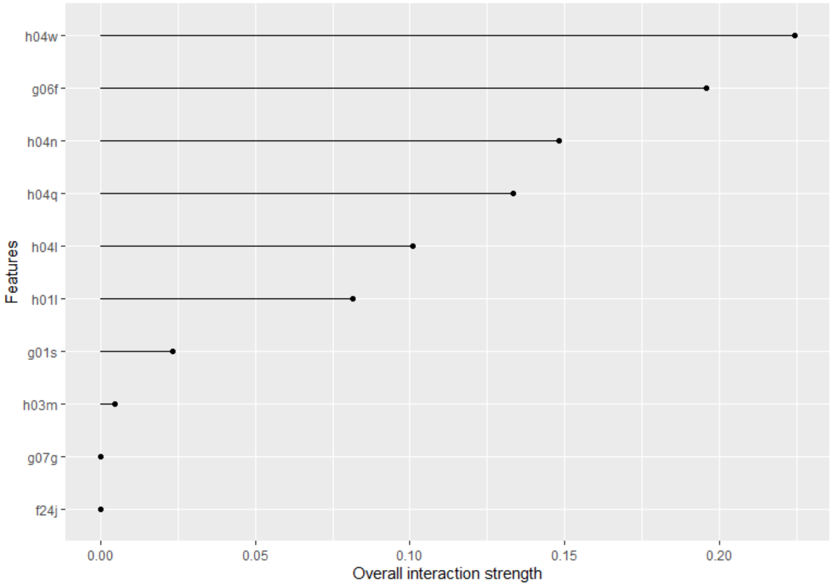

Figure 10: Overall Interactions for Dimension 4

The interactions for Dimension 4 reveal numerous weak to moderate potential interactions. **H04W** holds the strongest interaction potential with a score of 0.22, followed by **G06F** (0.20), **H04N**, **H04Q** and **H04L**. At first glance, it is very interesting that **H04Q** appears here, implying that the variable has important interactions, but not a main effect, quite in contrast to the lasso regression, as mentioned before. Based on this, the main interactions of each of these variables are explored, in figure 11.

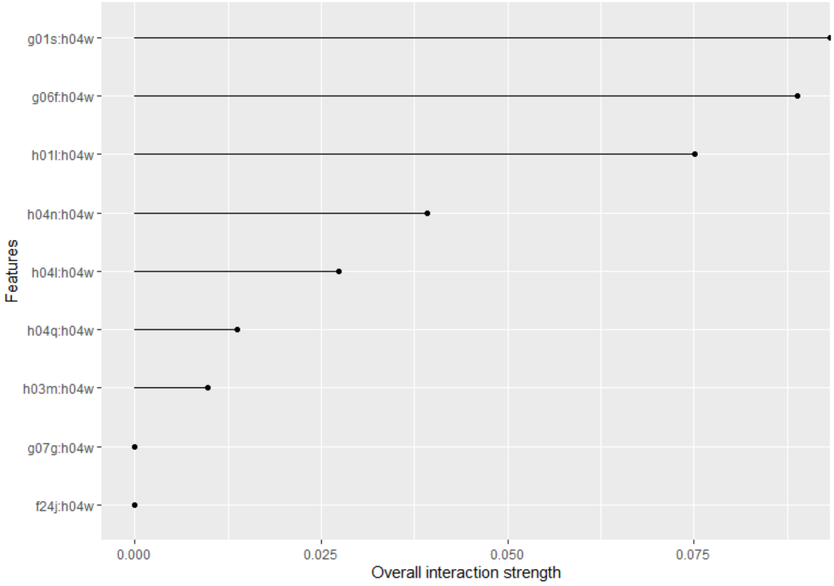

Figure 11: Interactions between **H04W** and other variables in Dimension 4

Interestingly, H04W does not result in any strong possible two way interactions, despite scoring high on the overall interactions. The highest scoring interactions potentials are with **G01S** (*Radio Direction-finding; Radio Navigation; Determining Distance Or Velocity By Use Of Radio Waves; Locating Or Presence-detecting By Use Of The Reflection Or Reradiation Of Radio Waves; Analogous Arrangements Using Other Waves*) and **G06F** (*Electric Digital Data Processing*), with neither rating more than 0.09 on the H-Statistic. This indicates that the chances of interactions between these variables are unlikely. Furthermore, exploring two way interactions for each of the other stated variables reveals similar results; in fact, the only strong interaction with a H-Statistic higher than 0.1 is **F24J:H04Q** (1.37). This is particularly interesting, as the ALE plots (Appendix 2) reveal that variations in **F24J** have no effect on the dependent variable, further reinforcing the relegation of **H04Q** in the PFI.

As observed in each of the dimensions, the correlation coefficients are skewed due to abnormalities with applications with a value of 0 for a specific variable. 5 of the 10 top variables experience distortion due to the 0 value. The classifications **H04W**, **G01S** and **G07G** all appear to positively influence the dimension score. It is important to note that **G07G** features a maximum of 2 instances in an application, with the overwhelming majority not being classified as such. **H04Q**, **G06F** and **H03M** all seem to deduct from the dimension score consistently; they produce a negative slope consistently across all values greater than 1. **H01L** generates a slight increment in the score between 1 and 3, and a negative value for all values above 3. Conversely, **H04L** produces a negative effect on the dimension score between the values of 1 and 2, and a positive score for all values above 2. **H04N** yields a positive score between values of 1 and 3, before generating a negative effect on the dimension score for all values greater than 3. Most interestingly, the tenth most important variable, **F24J**, is approximated to have no effect at all on the score, for any value, being thoroughly uncorrelated with the dimension score. This is consistent with the findings from the penalised regression.

Based on the results that have been compiled, numerous results can be drawn. The results from the penalised regression provide a simpler interpretation towards the relationships between these variables, perhaps it is more worthwhile to consider the consequences by the insight afforded by the Random Forest model and the accompanying black box interpretation tools. In the Conclusions section, this paper will explore compile learnings across all of the dimensions and models to construct answers and actionable insights for managers.

Conclusions and Discussion

Innovation is a difficult endeavour. Despite the volumes of research and best practices, both from theory and empirical research, the success rates for new products in the market remain low; the understanding, elusive. This thesis aimed to deepen understanding in this regard, by investigating the knowledge components of innovations, thereby complementing the existing management research.

This thesis aimed to find out if patent information could shape innovation strategy, within the technology sector. It sought the answer to this within the patent applications filed with the United Kingdom Intellectual Property Office's database. To do this, the paper first uncovered underlying patterns within the applications, based on the IPC codes assigned to each application. These patterns resulted from imbalances in the attractiveness of a problem, from the perspective of an innovating firm. Upon identifying emergent patterns with PCA, the research then attempted to identify which classifications contributed the most to each of these patterns, and the direction in which they contributed. This was achieved by deploying various regression methods, primarily penalised regression and random forest regression.

This section explores how the research has addressed the main and the sub questions put forward in the earlier chapters. Following the results, numerous general conclusions can be drawn, both pertaining to the questions raised, but also some of the more interesting findings that the analysis has revealed. Thus, this section will investigate each of these findings in greater depth. In general, each of the questions have been adequately addressed.

The first sub-question dealt with emergent patterns within the patent data. PCA revealed four dimensions which accounted for most of the variation in the dataset. Furthermore, interpretation of these dimensions coincided with two of the four types of innovation, identified by Fagerberg et al. (2013). This is discussed more in depth later.

The second sub-question dealt with profitable technological problems. The most important variables within each of the dimensions could be considered as the most profitable problems within the category. According to the data and analysis, classifications not mentioned in the results sections are considered to be unprofitable, as they influence the dimension scores in any meaningful way. However, they could provide a menial supportive role. A sufficiently high powered future analysis could focus on this.

Thus, the research concludes that patent information can shape innovation strategy, at least within technology. The following sections will delve deeply into general observations and each specific dimension and the steps that an R&D team could take to maximise chances for new product success.

Dimensions of Innovation

Principal Component Analysis was used in order to help mitigate some of the complexity arising from 527 classifications. However the results of PCA revealed interesting corroborations with theory and trends.

At least in Category H (Electricity) of WIPO's Intellectual Property Classifications, based on patent data from the United Kingdom, 4 separate dimensions of success can be ascertained, which accounted for 82.53% of all the variation in the data. Judging by the correlations, unique interpretations could be afforded for each of the principal components. Dimension 1, which accounted for 31.76% of the total variation, was designated the interpretation of "Communication"; Dimension 2, which accounted for 24.8% of the total variation, was designated the interpretation of "Images"; Dimension 3, which accounted for 13.65% of the variation, was designated the interpretation of "Manufacturing"; and finally, Dimension 4, which accounted for 12.24% of the total variation, was designated the interpretation of "Transmission".

Following theoretical constructs, each of these four dimensions adhered to two of the four types of innovation, first put forward by Fagerberg et al. (2013). Dimensions 1 and 2 seem to coincide with product innovation, while the dimensions 3 and 4 seem to belong to process innovation. This is because for the first two dimensions, the most important variables tend to have a common function. In dimension 1, it relates to communication and signals, while in dimension 2, it pertains to rendering and processing images. In contrast to this, dimensions 3 and 4 feature classifications which appear to have some flexibility in how they are deployed. For example, with **F24J** (*Production Or Use Of Heat Not Otherwise Provided For*), it is not immediately clear how this would connect thematically with the other classifications in dimension 3, giving an idea for the category they represent. While both these dimensions are thematically distinct, the lack of clarity in the intrinsic instrumentality in these themes and its resulting flexibility suggests that both of these dimensions are more process innovation oriented.

Tavassoli & Karlsson (2015) concluded that product innovations tend to have greater persistence than process innovations. This means that product innovation competence tends to carry over to future innovations more consistently than process innovations. Abiodun (2017), similarly, stated that the mediation effect that breakthrough innovations have on the relationship between innovation and financial performance tends to be larger for product innovation, than for process innovations. In the context of patents, an application that tends to feature more product innovations would then tend to be met with a larger number of applications filed, than with process innovations. This is because the competition for any product innovation would be other product innovators, resulting in a higher expected number of applications filed, than compared to process innovators. Thus, it is not surprising that the greatest amount of variation captured, according to how the data set was constructed, was by the two product innovation dimensions.

An R&D team should consider which of these dimensions they would like to prioritise. In general, the recommendation of this paper would be to focus on product innovations, rather than process innovations, however, this just might not be viable based on the company's in-house expertise and investment requirements. Based on how these dimensions are represented in this research, it becomes apparent that they are not mutually exclusive. There are numerous different IPC classifications which consistently appear across each of these dimensions. Thus, the question becomes one of balancing or prioritising dimensions. Once this fundamental question is answered, additional steps can be plotted out.

Ingredients of Innovation

Filtering for category H, there were 100,679 applications, of which the vast majority had only 3 or fewer classifications. Only 7,155 of these applications had more, and were also successful. This means, over a period of 40 years, only 11.6% of the applications could be considered as breakthrough innovations, in category H, and only 7.1% of all applications were approved breakthrough innovations. This has numerous practical implications, as discussed below.

The resulting Document-Term Matrix was understandably sparse. 88.4% of all applications featuring at least one category H classification had 3 or fewer codes, out of a potential 527 distinct classifications. This implies that there are inherent imbalances between each of the classification codes: certain codes feature more prominently than others. From a purely practical perspective, this imbalance somewhat defeats the purpose of a hierarchical classification system as the more commonly employed classification become too general and act as umbrella terms to aid others. In order to rectify this, the IPC has hastened its revisions to an annual rate.

However, as a consequence, each of the principal components had one variable that overwhelmingly contributed to its score. These were **H04L** (*Transmission Of Digital*

Information, E.G. Telegraphic Communication) for the Communication dimension; **H04N** (*Pictorial Communication, E.G. Television*) for the Images dimension; **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*) for the Manufacturing dimension; **H04W** (*Wireless Communication Networks*) for the Transmission dimension. Furthermore, these components also tended to be important within the other variables. The following table reflects the position and relative importance of these variables in each of the dimensions.

Variable	Communication	Images	Manufacturing	Transmission
H04L	100.00 (1)	9.92 (8)	26.77 (3)	23.34 (3)
H04N	9.89 (7)	100.00 (1)	6.51 (8)	11.58 (5)
H01L	15.61 (4)	12.43 (5)	100.00 (1)	23.34 (2)
H04W	32.63 (2)	16.21 (4)	27.56 (2)	100.00 (1)

Table 11: Importance (ranked) of top variables in each PCA dimension

Arguments can be made for why **H04W** and **H01L** appear so highly within each of the dimensions, considering that these inventions are the most important variables in the process innovation dimension. The implied versatility from these dimensions lends an explanation for why these two variables also rate so highly, in terms of importance, in the product innovation dimension. However, this does not explain why **H04N** and **H04L** appear to be very important for the process innovation variables. Potential explanations could possibly stem from the general ubiquitousness of communications and image technologies.

Drivers of Innovation

Following the identification of the dimension and the leading driver within each area, it was important to examine how each dimension was structured. While one classification dominated the prediction, the less contributive elements aided greatly within interpretation of each dimension, providing much needed specificity in each dimension. Correlation coefficients painted a very straightforward picture, however, a more in-depth analysis revealed certain interesting ideas about the influence that each of these variables have on the score, ultimately deepening the interpretation for each dimension score.

Main Effects

With regards to the main effects, the ALE plots yielded a number of commonalities between them. A large number of the plots were distorted between values of 0 and 1 for a classification. In addition to this, a large number of main effects would also change behaviour, typically occurring between values 1 and 2 or 1 and 3, Both of these observations can be explained by the construction of the dataset. Considering the general sparsity of data, an overwhelmingly large number of applications would feature values of 0 in the top variables, and are accordingly assigned a base value on the dimensions. Furthermore, this research defined breakthrough innovation as applications with a total of 4 or greater classifications. Thus, a value on one variable under 4 guarantees that the application has more than one assigned classification code, impacting the main effect of that variable. It is important to note, however, that this does not necessarily imply an interaction effect for these variables; only that the main effect of each variable is more clearly observed beyond the value of 3. The exact directions and implications of this are discussed later.

Evanschitzky et al (2012) suggests that *Product Technological Sophistication* and *Product Advantage* have a strong influence on new product success. In terms of *Product Technological Sophistication*, it was found that inventions with multiple classifications within one category typically yielded steady impact with each additional classification. This is because the ALE plots consistently result in a straight line, for classification values greater than 3. The main effects are more difficult to decipher in the case of multiple such classifications, due to a lack of clear interactions. In terms of *Product Advantage,* it can be gathered that the advantage is typically covered by the most important variables in each dimension, as these variables typically generate the greatest shifts in dimension scores. Classifications not covered in the results appear to not have a predictable impact on success, as defined by the dimensions and the subsequent random forest model. In later subsections, the exact effects that each important classification has on each of the dimensions are covered in greater depth.

Interaction Effects

For the majority of the variables investigated, there were very few interactions. Penalised regression suggests numerous interaction effects with low importance, however Random Forest yielded very few possible interactions, per Friedman's H-Statistic. The Communication and Manufacturing dimensions did not produce a single strong interaction, and the Transmission dimension produced a strong interaction potential with an uncorrelated variable. The one noteworthy interaction was generated by the Images dimension, **G03B:H04N** (**G03B**: *Apparatus Or Arrangements For Taking Photographs Or For Projecting Or Viewing Them; Apparatus Or Arrangements Employing Analogous Techniques Using Waves Other Than Optical Waves; Accessories Therefor*; **H04N**: *Pictorial Communication, E.G. Television*)(H-Statistic of 1.0), indicating that the entire effect could be explained by the interaction effect. Combining this with the reduced importance of the **G03B** variable suggested that the classification best performed as a supporting feature to a core invention.

This indicates that, according to the Random Forest model, most of the observed effects are main effects of the classifications. This lends to the notion that, in the vast majority of the cases, breakthrough innovations tend to specialise within a particular classification, with very little observable integration between them.

Communication Breakdown

Compiling the results of each of the major dimensions, it becomes clear that the relationship between the dimension score and the classification is surprisingly nuanced. In the results section, the paper reported the general tendencies of each of the important variables. Now, the paper will attempt to interpret what the score is trying to capture.

The following table contains terms, based on their slope in the ALE. Terms are arranged by their relative importance to the model, from the results of PFI.

Positively Correlated	**Negatively Correlated**
H04L (1) - *Transmission Of Digital Information, E.G. Telegraphic Communication*	**H04W (2)** - *Wireless Communication Networks*
H04N (3) - *Pictorial Communication, E.G. Television*	**H01L (4)** - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*
G06F (5) - *Electric Digital Data Processing*	**H04B (6)** - *Transmission*
	H04Q (7) - *Selecting*
	H04M (8) - *Telephonic Communication*
	H03M (9) - *Coding, Decoding Or Code Conversion, In General*
	H04J (10) - *Multiplex Communication*

Table 12: Classifications (ranked) positively and negatively correlated with Dimension 1 based on ALE

It appears that, despite a thematic connection between all of the variables, the effects of some of the most important variables are considerably antagonistic to each other. It is difficult to truly interpret what the extremes of the dimensions are truly capturing here. One potential interpretation could be that the variables that impact the dimension score positively appear to capture some form of network effect between devices, while the variables that impact the dimension negatively appear to capture innovations within communicating devices. This seems to indicate that both ends of this dimension could yield unique results for innovation success. Finally, while there might not be overt interaction effects, the aligning direction seems to indicate that it might still be profitable to pursue classifications in either direction.

Thus, for inventions focusing on technology focusing on networks and infrastructure, it would help to focus solely on the **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*), **H04N** (*Pictorial Communication, E.G. Television*) and **G06F** (*Electric Digital Data Processing*) classifications. In order to address more granular device design, it might help to focus more on **H04W** (*Wireless Communication Networks*), **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*), **H04B** (*Transmission*), **H04Q** (*Selecting*), **H04M** (*Telephonic Communication*), **H03M** (*Coding, Decoding Or Code Conversion, In General*) and **H04J** (*Multiplex Communication*) classifications. It is important to bear in mind that the importance of each variable reduces further down the list. Thus, if an invention focuses on a less important classification, it is advisable to either turn it into an incremental innovation, or use a multitude of innovations. An example for an invention for a more positive dimension 1 score would, therefore, be 5G mobile network, whereas for a more negative dimension 1 score would perhaps focus on specialised circuitry.

Image Processing

Similarly, the following table features classifications based on the direction of their effect on the dimension score, according to ranking, for dimension 2, Images.

The most important variable, **H04N** (*Pictorial Communication, E.G. Television*), appears to negatively affect the score as a technology specialises within this classification. Similar conclusions can be drawn for **H04W** (*Wireless Communication Networks*). Interpretation for this extremity of the dimension score is challenging. However, classifications that appear to positively affect the score are all of the category G variables. This suggests that the positive extreme of this dimension focuses more on the optics and optical supporting components within the invention. To tie it up elegantly, one could establish a dichotomy that the positive end of this dimension pertains to capturing image and image signals, while the negative end of this dimension refers to receiving and reproducing these same image signals.

Positively Correlated	**Negatively Correlated**
H04L (2) - *Transmission Of Digital Information, E.G. Telegraphic Communication*	**H04N (1)** - *Pictorial Communication, E.G. Television*
H01L (3) - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	**H04W (4)** - *Wireless Communication Networks*
G06T (5) - *Image Data Processing Or Generation, In General*	
G02B (6) - *Optical Elements, Systems, Or Apparatus*	
G06K (7) - *Recognition Of Data; Presentation Of Data; Record Carriers; Handling Record Carriers*	
G11B (8) - *Information Storage Based On Relative Movement Between Record Carrier And Transducer*	
G03B (9) - *Apparatus Or Arrangements For Taking Photographs Or For Projecting Or Viewing Them; Apparatus Or Arrangements Employing Analogous Techniques Using Waves Other Than Optical Waves; Accessories Therefor*	
G03H (10) - *Holographic Processes Or Apparatus*	

Table 13: Classifications (ranked) positively and negatively correlated with Dimension 2 based on ALE

Thus, for innovations focusing on developing new ways of capturing optical information, it is advisable to focus on **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*), **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*), **G06T** (*Image Data Processing Or Generation, In General*), **G02B** (*Optical Elements, Systems, Or Apparatus*), **G06K** (*Recognition Of Data; Presentation Of Data; Record Carriers; Handling Record Carriers*), **G11B** (*Information Storage Based On Relative Movement Between Record Carrier And Transducer*), **G03B** (*Apparatus Or Arrangements For Taking Photographs Or For Projecting Or Viewing Them; Apparatus Or Arrangements Employing Analogous Techniques Using Waves Other Than Optical Waves; Accessories Therefor*) and

G03H (*Holographic Processes Or Apparatus*). As explained in the previous section, the entire effect of **G03B** could be captured as a result of its interaction with **H04N**. An example of an invention in this area could be a new camera system. For inventions aiming for a negative score on this dimension, it is advisable to focus on **H04N** and **H04W**. An example of a successful innovation in this area would be a new television or projector system.

Manufacturing Insight

The following table focuses on the variables and their impact on the dimension 3 score, based on ALE effects. This proved instrumental in correctly interpreting the direction of variance in this dimension.

Positively Correlated	**Negatively Correlated**
H01L (1) - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*	**C08G (5)** - *Macromolecular Compounds Obtained Otherwise Than By Reactions Only Involving Carbon-to-carbon Unsaturated Bonds*
H04L (2) - *Transmission Of Digital Information, E.G. Telegraphic Communication*	**F24J (8)** - *Production Or Use Of Heat Not Otherwise Provided For*
H04W (3) - *Wireless Communication Networks*	**C07F (9)** - *Acyclic, Carbocyclic, Or Heterocyclic Compounds Containing Elements Other Than Carbon, Hydrogen, Halogen, Oxygen, Nitrogen, Sulfur, Selenium Or Tellurium*
H04N (4) - *Pictorial Communication, E.G. Television*	**C23C (10)** - *Coating Metallic Material; Coating Material With Metallic Material; Surface Treatment Of Metallic Material By Diffusion Into The Surface, By Chemical Conversion Or Substitution; Coating By Vacuum Evaporation, By Sputtering, By Ion Implantation Or By Chemical Vapour Deposition, In General*
B82Y (6) - *Specific Uses Or Applications Of Nanostructures; Measurement Or Analysis Of Nanostructures; Manufacture Or Treatment Of Nanostructures*	
G06F (7) - *Electric Digital Data Processing*	

Table 14: Classifications (ranked) positively and negatively correlated with Dimension 3 based on ALE

When investigating the values of the most important variables in dimensions 3, a dichotomy becomes apparent. The dimension rightfully captures an element of fabrication and manufacturing. However, the ALE suggests that the positive end of the dimension features more on industrial design, while the negative end seems to capture more on the material production process.

Thus, inventions aiming to improve process design and layouts could aim for classifications within **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*), **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*), **H04W** (*Wireless Communication Networks*), **H04N** (*Pictorial Communication, E.G. Television*), **B82Y** (*Specific Uses Or Applications Of Nanostructures; Measurement Or Analysis Of Nanostructures; Manufacture Or Treatment Of Nanostructures*) and **G06F** (*Electric Digital Data Processing*). While an example for innovations that combine these classifications is not readily available, these would be important when a smartphone manufacturer would aim to generate slimmer devices, or fit larger sensors or components without altering the physical dimensions of that device.

Inventions that aim to generate new materials for production efficiencies could focus on the classifications **C08G** (*Macromolecular Compounds Obtained Otherwise Than By Reactions Only Involving Carbon-to-carbon Unsaturated Bonds*), **F24J** (*Production Or Use Of Heat Not Otherwise Provided For*), **C07F** (*Acyclic, Carbocyclic, Or Heterocyclic Compounds Containing Elements Other Than Carbon, Hydrogen, Halogen, Oxygen, Nitrogen, Sulfur, Selenium Or Tellurium*), **C23C** (*Coating Metallic Material; Coating Material With Metallic Material; Surface Treatment Of Metallic Material By Diffusion Into The Surface, By Chemical Conversion Or Substitution; Coating By Vacuum Evaporation, By Sputtering, By Ion Implantation Or By Chemical Vapour Deposition, In General*). An example for an invention within this area could be industrial heatsinks, designed to better manage production temperatures.

Transmitting Knowledge

In table 15, the localised effects of each variable on the score for dimension 4 are examined. This allows for better interpretation for the variance along this dimension. Terms are arranged by their relative importance to the model, from the results of PFI. This helps provide a clearer understanding of which aspect of the variation is being captured by each end of the dimension.

Positively Correlated	**Negatively Correlated**
H04W (1) - *Wireless Communication Networks*	**H01L (3)** - *Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*
H04L (2) - *Transmission Of Digital Information, E.G. Telegraphic Communication*	**H04N (4)** - *Pictorial Communication, E.G. Television*
G01S (7) - *Radio Direction-finding; Radio Navigation; Determining Distance Or Velocity By Use Of Radio Waves; Locating Or Presence-detecting By Use Of The Reflection Or Reradiation Of Radio Waves; Analogous Arrangements Using Other Waves*	**H04Q (5)** - *Selecting*
G07G (9) - *Registering The Receipt Of Cash, Valuables, Or Tokens*	**G06F (6)** - *Electric Digital Data Processing*
	H03M (8) - *Coding, Decoding Or Code Conversion, In General*

Table 15: Classifications (ranked) positively and negatively correlated with Dimension 4 based on ALE

Based on the variables that seem to positively and negatively impact the dimension, perhaps this dimension could be reinterpreted as signals. The variables negatively impacting the dimension score seem to focus on receiving and interpreting signals, where the positive variables appear to focus on the transmissions or network effects (with the exception of **G07G** (*Registering The Receipt Of Cash, Valuables, Or Tokens*)). However, again, unlike in dimension 1, both sides appear to focus on device design.

Thus, for inventions focusing on signal transmission, it is advisable to couple the following classifications: **H04W** (*Wireless Communication Networks*), **H04L** (*Transmission Of Digital Information, E.G. Telegraphic Communication*) and **G01S** (*Radio Direction-finding; Radio Navigation; Determining Distance Or Velocity By Use Of Radio Waves; Locating Or Presence-detecting By Use Of The Reflection Or Reradiation Of Radio Waves; Analogous Arrangements Using Other Waves*). For inventions focusing on signal receivers, it is advisable to focus on **H01L** (*Semiconductor Devices; Electric Solid State Devices Not Otherwise Provided For*), **H04N** (*Pictorial Communication, E.G. Television*), **H04Q** (*Selecting*), **G06F** (*Electric Digital Data Processing*) and **H03M** (*Coding, Decoding Or Code Conversion, In General*).

Furthermore, it is important to note that the total set of important variables across all 4 dimensions comprises only 23 (rather than potentially 40) classifications. This indicates a large amount of overlap between the dimensions. As stated earlier, this gives an innovation manager flexibility; the manager does not have to choose a dimension to forgo others, but opts to prioritise and combine classifications stated above to best suit the goals for their research and development team.

Limitations and Future Research

This research had numerous limitations, and expanding upon each of them would be an avenue for future research.

Based on the IPC classifications, this research had a singular focus on the category H (electricity). This was because the analogy of problems and solutions that is favoured by innovation literature, adheres best within this category. However, this is more a metaphysical consideration, and the methods and results from this research indicate that the same can be done for every category. More interestingly, research could approach from a more consumer focused category or categories, in order to further tie theory from marketing and technology trends with patent analysis.

Furthermore, the research proved feasible with the first four bits of each classification. Thus, a more specialised player could expand further and investigate the last four bits, i.e. the **9/31**in the code **H04N 9/31** used as an example in the earlier sections of this paper. This could be particularly interesting in trying to understand trends within a very particular type of problem being solved. This could prove instrumental in developing a strategy for direct competitors in a category.

In addition, patent analysis via IPC classifications could also benefit from being treated as a Text Analytics project. Latent Dirichlet Allocation, Non-negative Matrix Factorisation and Word Embeddings could be employed in interesting ways to try and construct a language and infer meaning and relationships between each classification in greater depth. N-Gram analysis could also provide interesting ways to study interactions and synergies.

The dataset was also centered around the UK Intellectual Property Office. While this is a major market in global business and the results could be extrapolated somewhat to other geographies, there is no substitute to actual research. In addition, it is extremely interesting to explore the patent trends in diverse geographies, which would introduce more national or regional geopolitical intricacies to the evolution of technology and innovations in that region.

Evolution of patents could also be explored from a cultural perspective, as a number of possibilities become apparent from works such as those of Theodore Sturgeon. A structure of one generic dominant classification, supported by less important, but specialist inventions is very similar to those encountered in branding and marketing theory. As time moves on, the lines between these seemingly discrete fields become blurred; the potential for research increases.

This research also explored a period of 40 years. This has an inherent shortcoming being that the research assumes that the results before and after each invention can be applicable uniformly through this time. However, as any student of history would concur, this is far from the truth. Successful technology in the 1980s could perhaps have a much different path compared to the 2010s. Thus, exploring the many shifting nuances of each narrower time period could reveal interesting patterns to investigate.

Finally, this research was solely interested in successful inventions. By definition, this paper looked at patents that were successfully granted and their performance in the marketplace. Even within category H, that amounted to only 7.1% of the applications. As stated in the results sections, the success rate of breakthrough innovations lags behind the success rate of all innovations, both within category H and within all categories. This research did not explore at all why this was the case; if this was statistically significant; and how the overall rate might be improved. There are numerous studies that focus on new product success, but rather few that explore a successful patent application.

In a similar vein, it would be very interesting to explore how in-house competence versus category competence affects new product success. Evanschitzky et al (2012) suggests that *Technological Synergy* and *Technological Proficiency* are both significantly influential in predicting new product success. Based on how they are defined in Henard & Szymanski (2001), both these variables relate to expertise within a company. Looking back at applications filed by a company and published patent applications by competitors could give insight into the circumstances that influenced the innovation.

References

Abiodun, T. S. (2017). An Examination of the Relationships between Different Types of Innovation and Firm Performance and the Mediating Effect of Radical and Incremental Innovations on These Relationships. International Journal Of Innovation And Economic Development, 3(1), 38-58. doi:10.18775/ijied.1849-7551-7020.2015.35.2003

Abraham, B. P., & Moitra, S. D. (2001). Innovation assessment through patent analysis. Technovation, 21(4), 245-252. doi:10.1016/s0166-4972(00)00040-7

Ardito, L., Dadda, D., & Petruzzelli, A. M. (2018). Mapping innovation dynamics in the Internet of Things domain: Evidence from patent analysis. *Technological Forecasting and Social Change, 136*, 317-330. doi:10.1016/j.techfore.2017.04.022

Cho, H. P., Lim, H., Lee, D., Cho, H., & Kang, K. (2018). Patent analysis for forecasting promising technology in high-rise building construction. *Technological Forecasting and Social Change, 128*, 144-153. doi:10.1016/j.techfore.2017.11.012

Clausen, T. H., & Pohjola, M. (2013). Persistence of product innovation: Comparing breakthrough and incremental product innovation. *Technology Analysis & Strategic Management, 25*(4), 369-385. doi:10.1080/09537325.2013.774344

Cohen, D., Quinn, B., & Roth, E. (2021, March 12). The Innovation Commitment. Retrieved April 29, 2021, from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-innovation-commitment

Costantini, V., Crespi, F., & Palma, A. (2017). Characterizing the policy mix and its impact on eco-innovation: A patent analysis of energy-efficient technologies. *Research Policy, 46*(4), 799-819. doi:10.1016/j.respol.2017.02.004

Dean, T. (2017, June 01). The meeting that showed me the truth about VCs. Retrieved May 5, 2021, from https://techcrunch.com/2017/06/01/the-meeting-that-showed-me-the-truth-about-vcs/

Evanschitzky, H., Eisend, M., Calantone, R. J., & Jiang, Y. (2012). Success Factors of Product Innovation: An Updated Meta-Analysis. *Journal of Product Innovation Management, 29*, 21-37. doi:10.1111/j.1540-5885.2012.00964.x

Fagerberg, J., Mowery, D. C., & Nelson, R. R. (2013). The Oxford handbook of innovation. Oxford: Oxford University Press.

Gerken, J., Moehrle, M. G., & Walter, L. (2015, May 15). One Year Ahead! Investigating the Time Lag between Patent Publication and Market Launch: Insights from a Longitudinal Study in the Automotive Industry. Retrieved August 4, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2606440

Harnoss, J. D., & Baeza, R. (2019, September 24). Overcoming the Four Big Barriers to Innovation Success. Retrieved from https://www.bcg.com/publications/2019/overcoming-four-big-barriers-to-innovation-success

Henard, D. H., & Szymanski, D. M. (2001). Why Some New Products are More Successful than Others. *Journal of Marketing Research, 38*(3), 362-375. doi:10.1509/jmkr.38.3.362.18861

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R. Boston: Springer.

Melgarejo, R., & Malek, K. (2018). *SETTING THE RECORD STRAIGHT ON INNOVATION FAILURE* (Rep.). Retrieved April 29, 2021, from Nielsen BASES website: https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/setting-the-record-straight-common-causes-of-innovation-failure-1.pdf

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Victoria, British Columbia: Leanpub.

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent Statistics as an Innovation Indicator. *Handbook of the Economics of Innovation, Volume 2 Handbook of the Economics of Innovation,* 1083-1127. doi:10.1016/s0169-7218(10)02009-5

Nobel, C. (2011, February 14). Clay Christensen's Milkshake Marketing. Retrieved April 29, 2021, from https://hbswk.hbs.edu/item/clay-christensens-milkshake-marketing

Patents. (n.d.). Retrieved June 14, 2021, from https://www.wipo.int/patents/en/

Satell, G. (2017, June 21). The 4 Types of Innovation and the Problems They Solve. Retrieved June 15, 2021, from https://hbr.org/2017/06/the-4-types-of-innovation-and-the-problems-they-solve

Scherer, F., & Harhoff, D. (2000). Technology policy for a world of skew-distributed outcomes. *Research Policy, 29*(4-5), 559-566. doi:10.1016/s0048-7333(99)00089-x

Shubbak, M. H. (2019). Advances in solar photovoltaics: Technology review and patent trends. *Renewable and Sustainable Energy Reviews, 115*, 109383. doi:10.1016/j.rser.2019.109383

Rowley, J. (2017, May 17). Here's how likely your startup is to get acquired at any stage. Retrieved May 5, 2021, from https://techcrunch.com/2017/05/17/heres-how-likely-your-startup-is-to-get-acquired-at-any-stage/

Statista. (2021, August 11). Smartphone market share 2021. Retrieved August 12, 2021, from https://www.statista.com/statistics/271496/global-market-share-held-by-smartphone-vendors-since-4th-quarter-2009/

Tavassoli, S., & Karlsson, C. (2015). Persistence of various types of innovation analyzed and explained. Research Policy, 44(10), 1887-1901. doi:10.1016/j.respol.2015.06.001

WIPO. (2020). *International Patent Classification (IPC)* [Brochure]. Author. Retrieved June 14, 2021, from https://www.wipo.int/edocs/pubdocs/en/wipo_brochure_ipc.pdf

Yuan, X., & Cai, Y. (2021). Forecasting the development trend of low emission vehicle technologies: Based on patent data. *Technological Forecasting and Social Change, 166*, 120651. doi:10.1016/j.techfore.2021.120651

Appendix

Appendix 1: Correlations of Variables and Principal Components

Principal Component 1: Communication

Variable	Correlation
H04L	-0.70560607
H04W	-0.27200506
G06F	-0.27831404
H01L	0.22214351
H04Q	-0.16742230
H04B	-0.17546282
H04N	-0.02265021
H03M	-0.14923063
H04J	-0.13092374
H04M	-0.11005594

Principal Component 2: Images

Variable	Correlation
H04n	0.60250373
G03b	0.19850656
G06t	0.25720000
H04w	-0.14969167
H01l	-0.13298595
G11b	0.10866729
G03h	0.09137407
H04l	-0.11267608
G06k	0.08951954
G02b	0.11668337

Principal Component 3: Manufacturing

Variable	Correlation
H01L	-0.680927368
H04W	0.227016636
H04L	-0.096690066
F24J	-0.116578043
B82Y	-0.016309174
C07F	-0.091056388
C23C	-0.088082428
H04N	0.002148289
G06F	-0.041012897
C08G	-0.118176919

Principal Component 4: Transmission

Variable	Correlation
H04W	0.66015552
H01L	0.19813241
H04L	-0.14804038
H04Q	-0.14420739
H04N	0.04527944
G06F	-0.12062222
G01S	0.03807514
H03M	-0.08901807
G07G	0.03899765
F24J	0.03422148

Appendix 2: ALE Plots

Principal Component 1: Communication

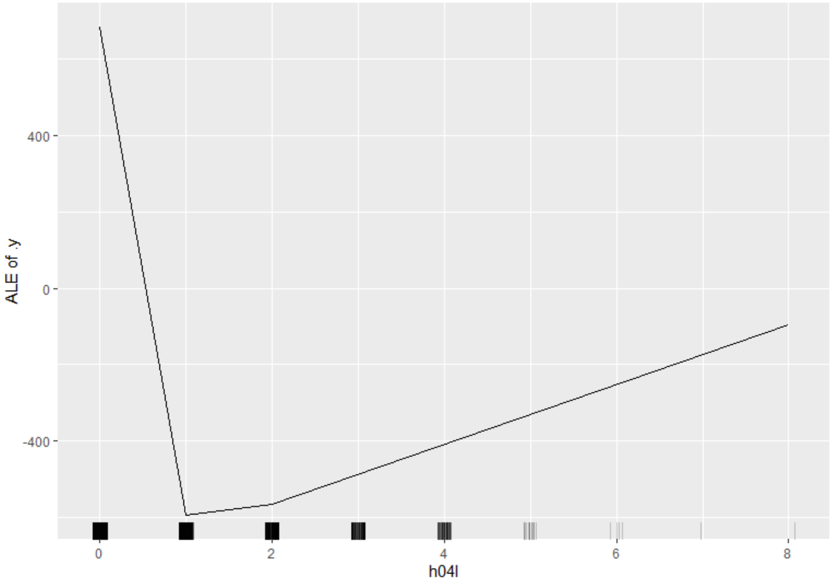

Figure 12: ALE for H04L on Dimension 1

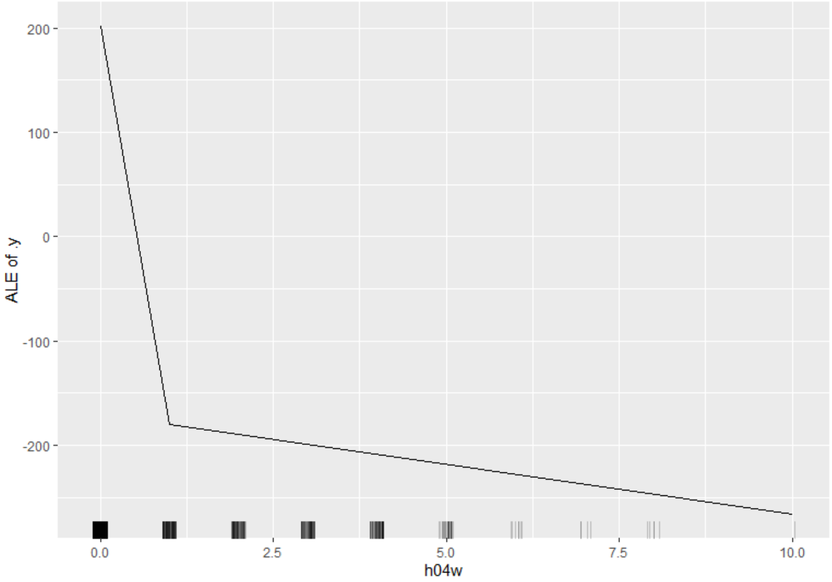

Figure 13: ALE for H04W on Dimension 1

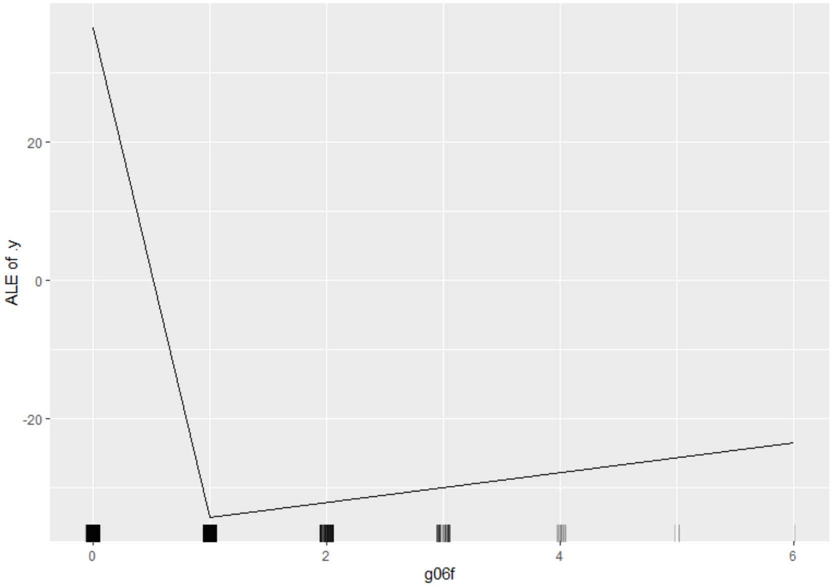

Figure 14: ALE for G06F on Dimension 1

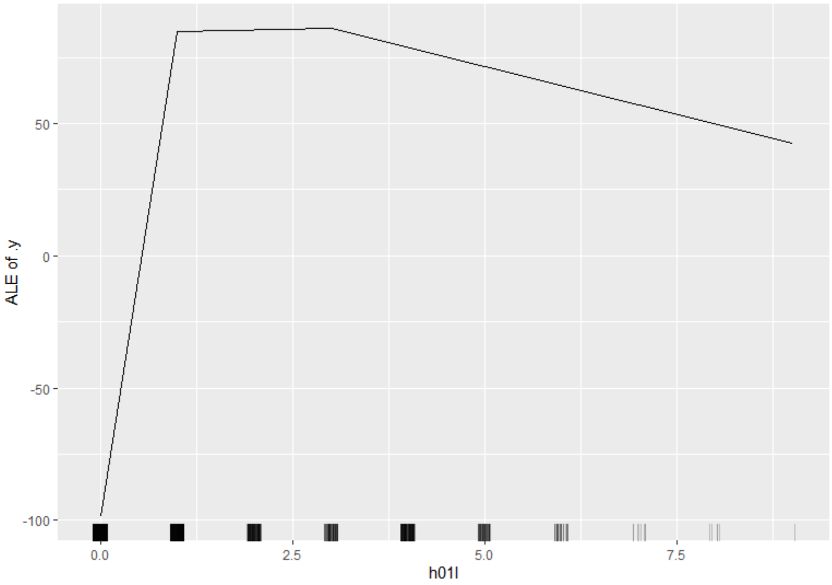

Figure 15: ALE for H01L on Dimension 1

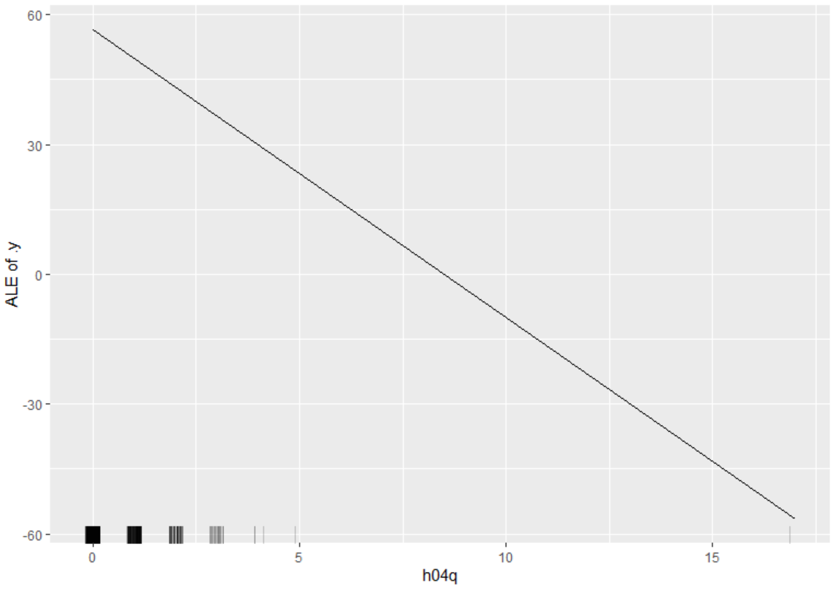

Figure 16: ALE for H04Q on Dimension 1

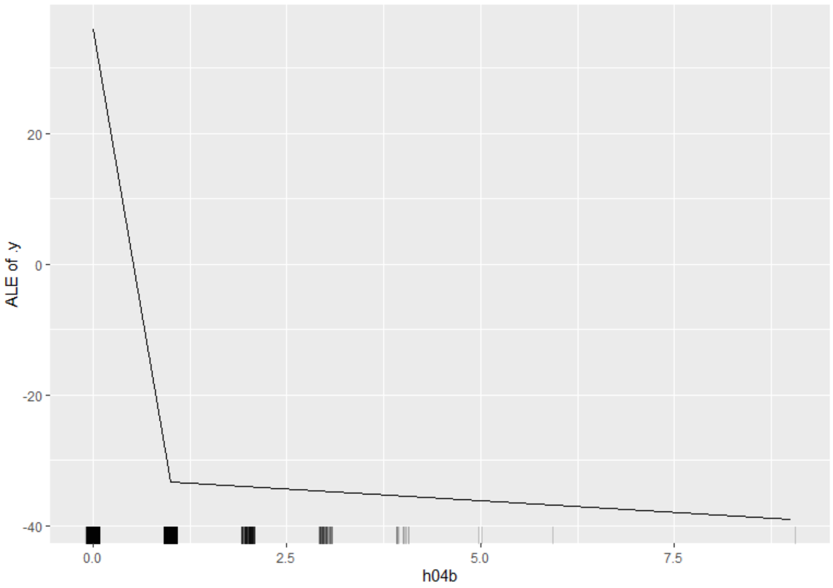

Figure 17: ALE for H04B on Dimension 1

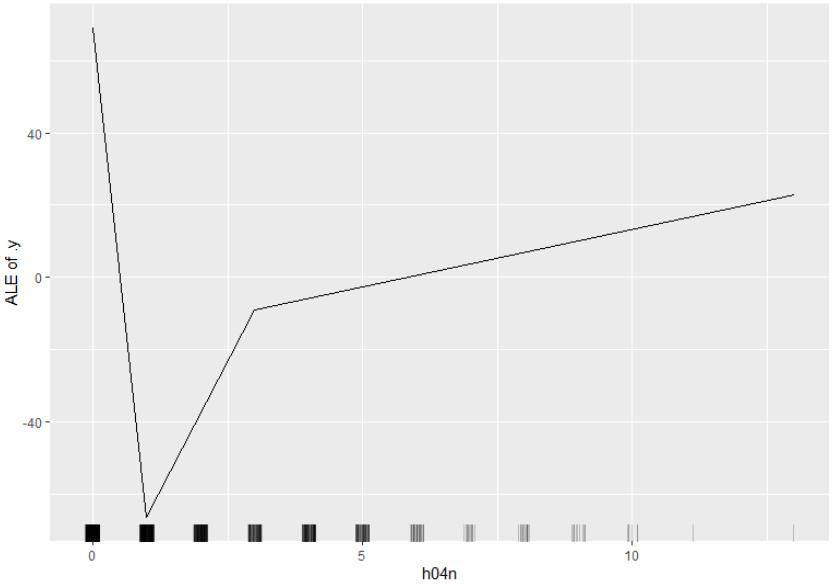

Figure 18: ALE for H04N on Dimension 1

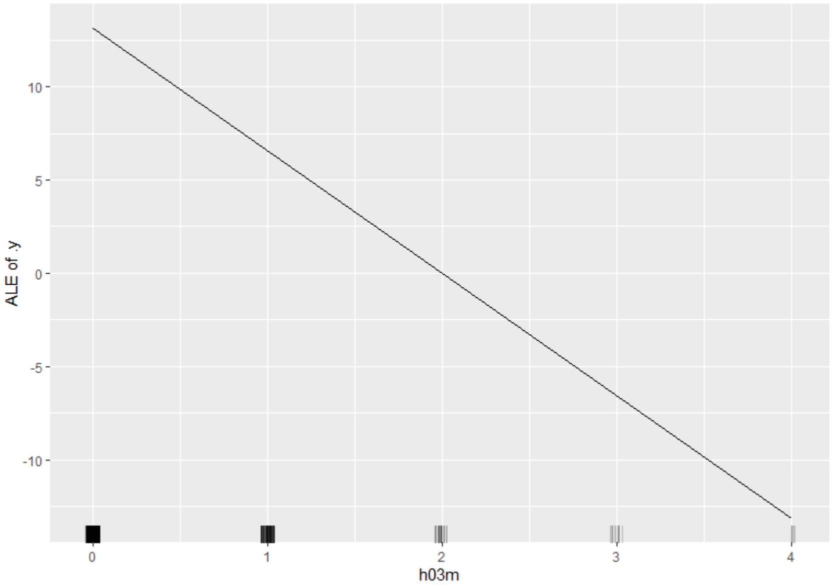

Figure 19: ALE for H03M on Dimension 1

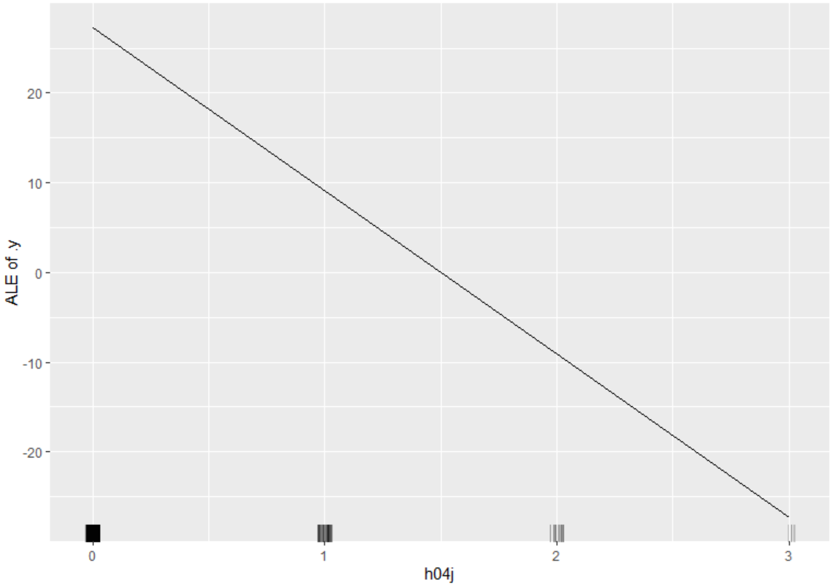

Figure 20: ALE for H04J on Dimension 1

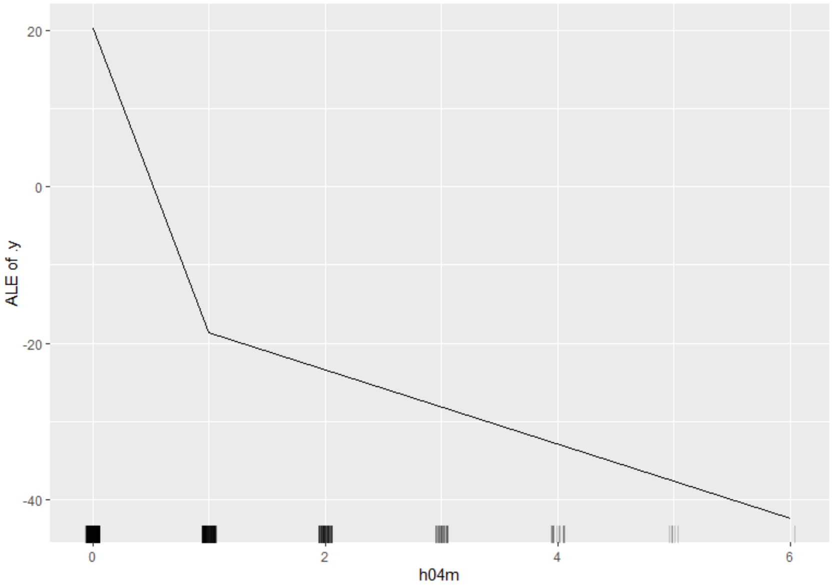

Figure 21: ALE for H04M on Dimension 1

Principal Component 2: Images

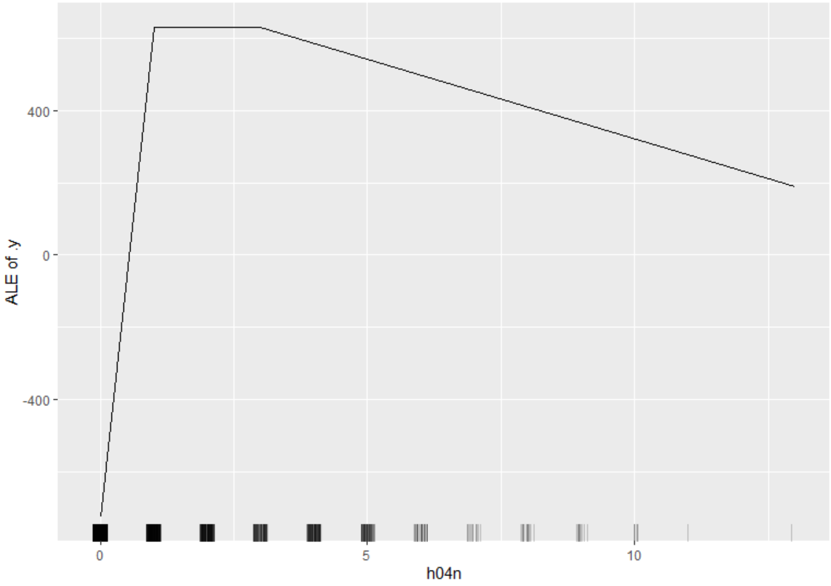

Figure 22: ALE for H04N on Dimension 2

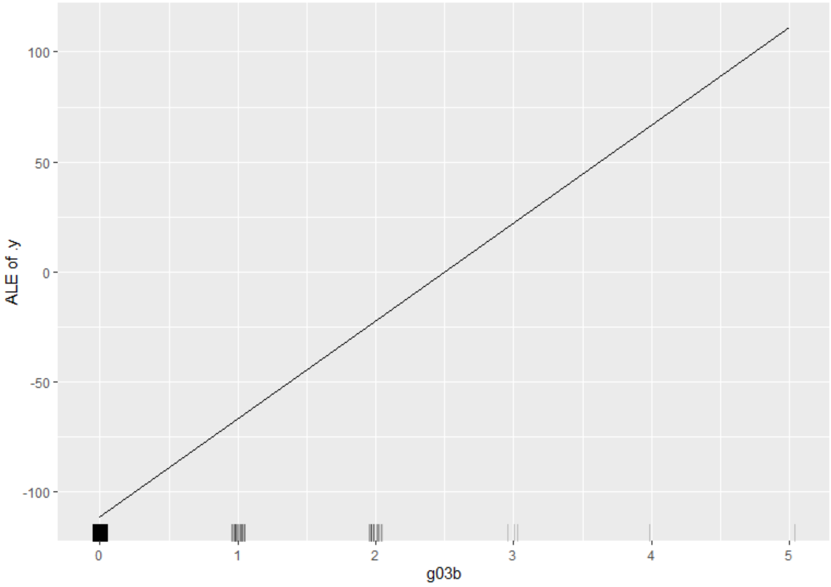

Figure 23: ALE for G03B on Dimension 2

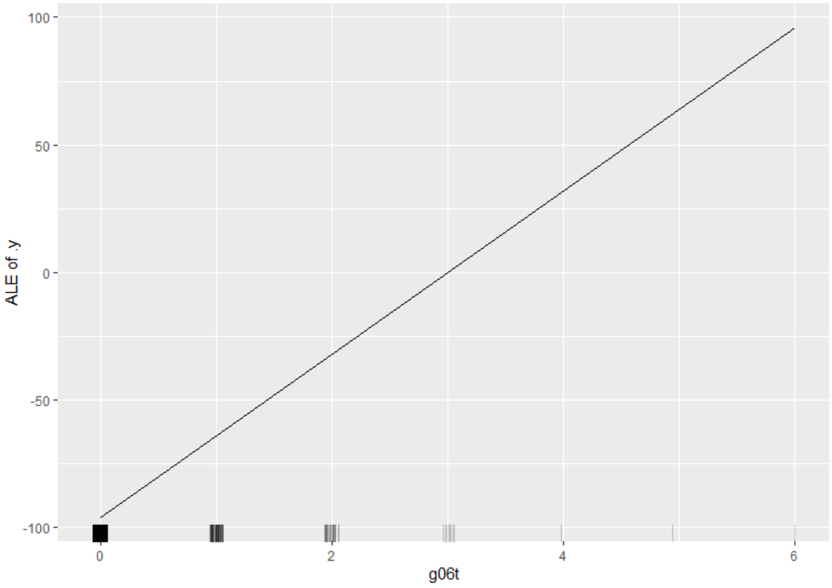

Figure 24: ALE for G06T on Dimension 2

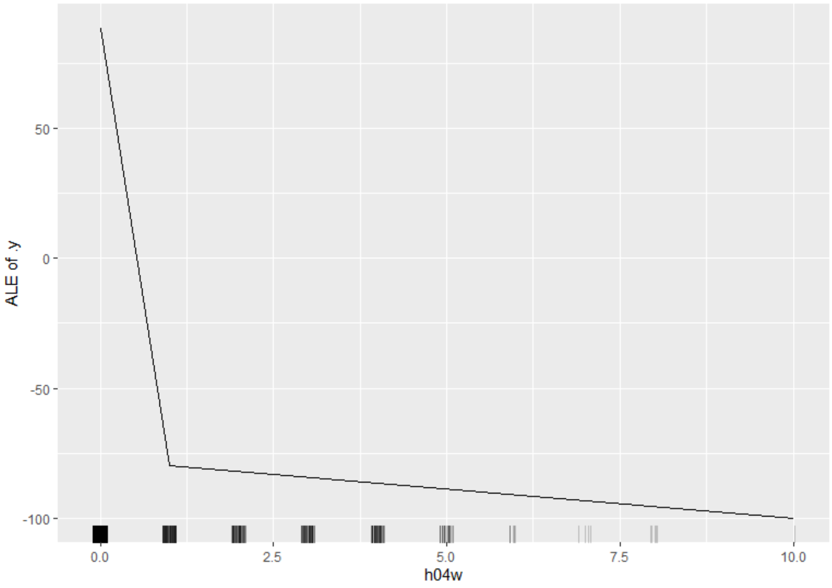

Figure 25: ALE for H04W on Dimension 2

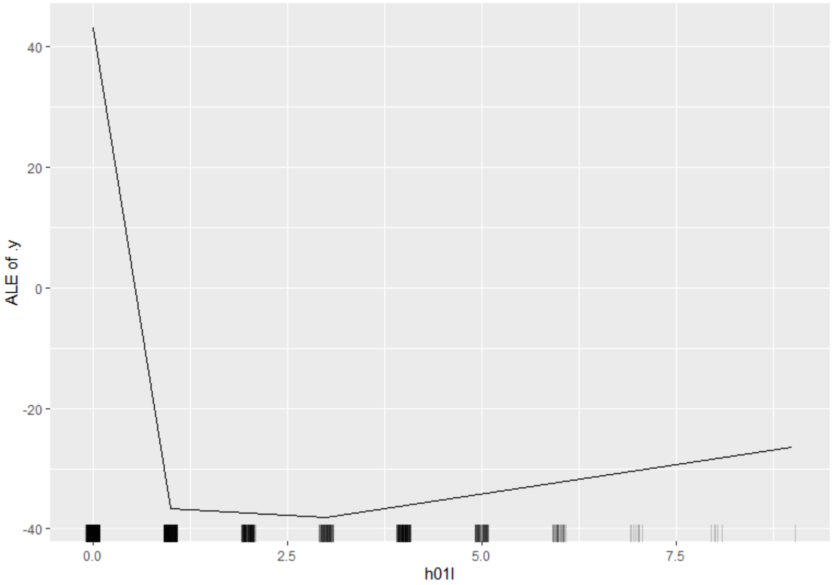

Figure 26: ALE for H01L on Dimension 2

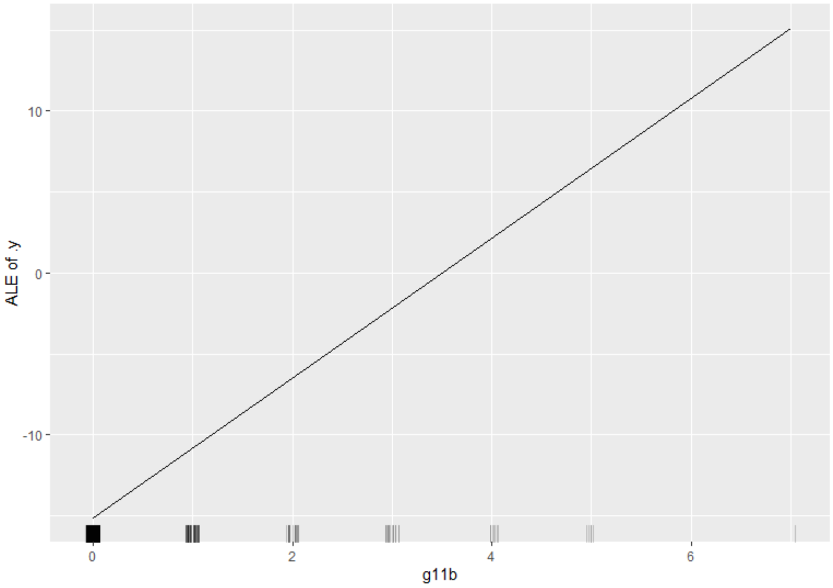

Figure 27: ALE for G11B on Dimension 2

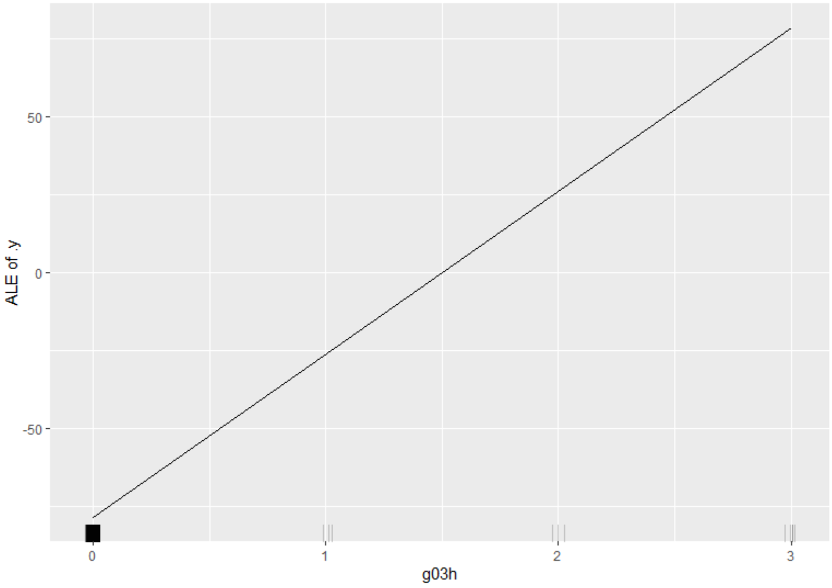

Figure 28: ALE for G03H on Dimension 2

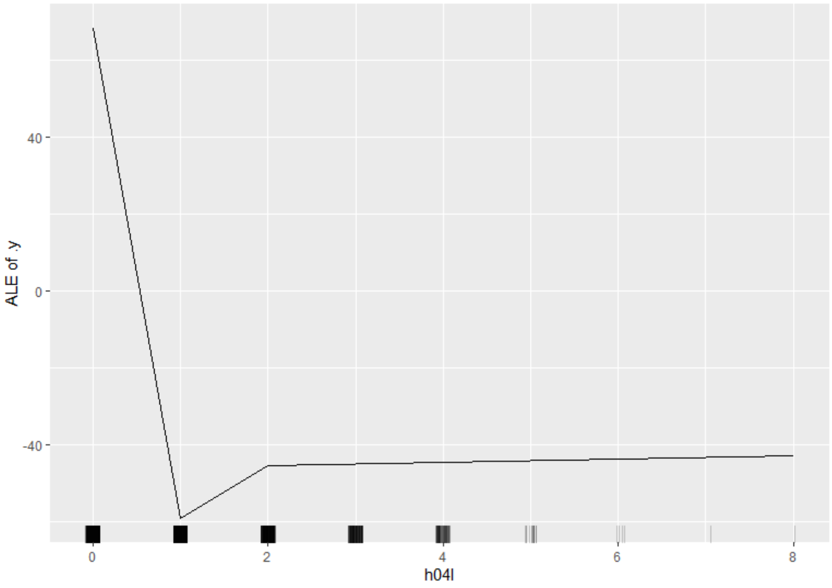

Figure 29: ALE for H04L on Dimension 2

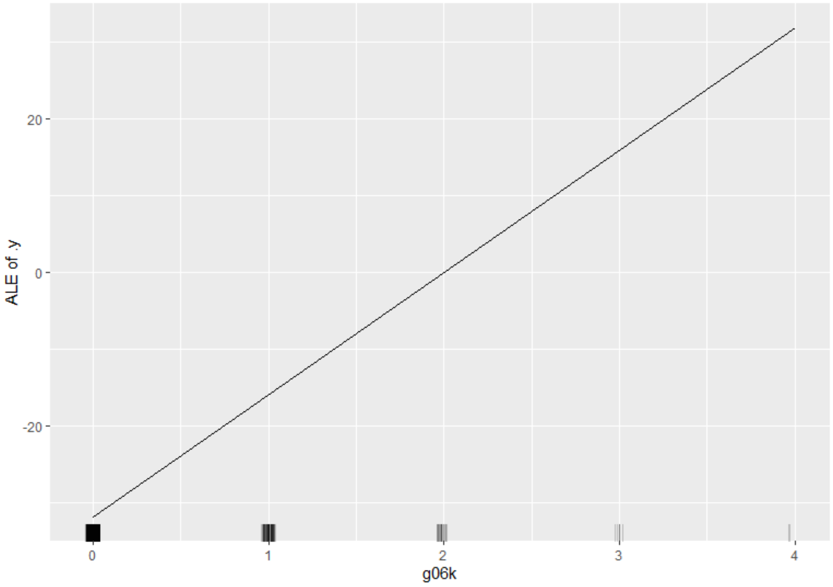

Figure 30: ALE for G06K on Dimension 2

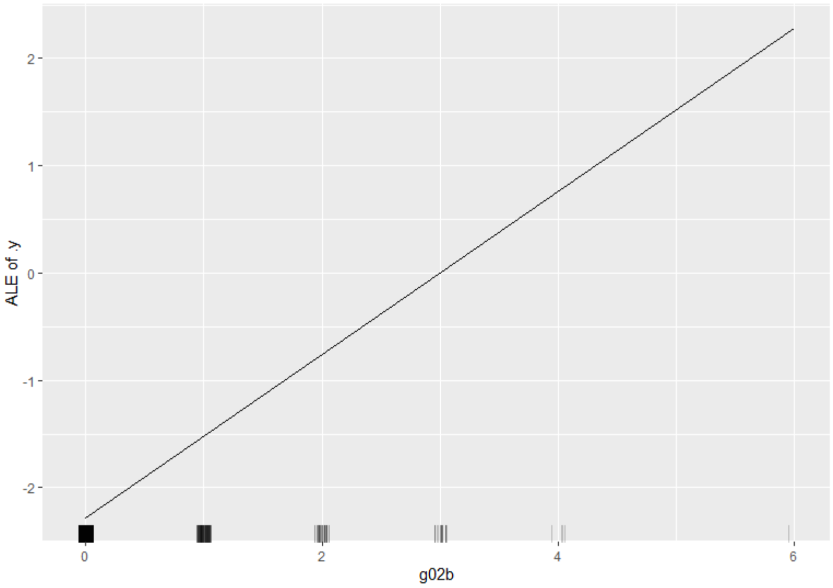

Figure 31: ALE for G02B on Dimension 2

Principal Component 3: Manufacturing

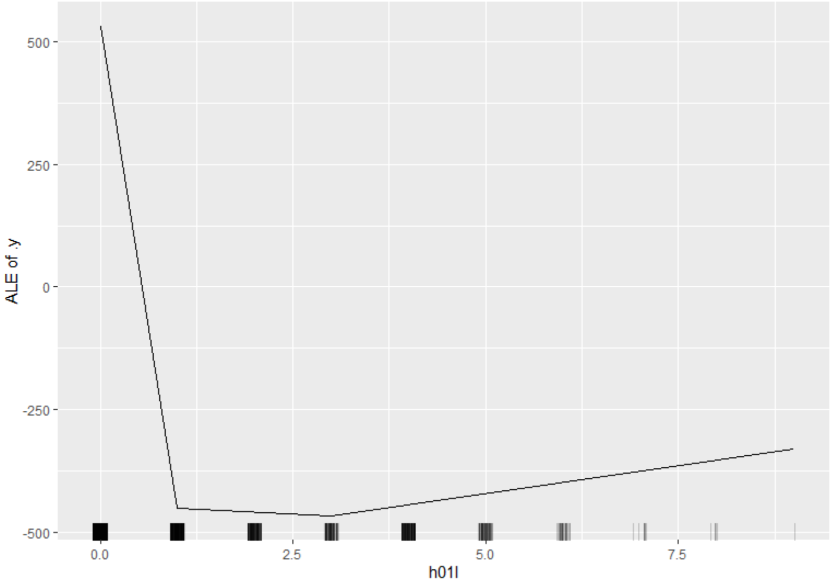

Figure 32: ALE for H01L on Dimension 3

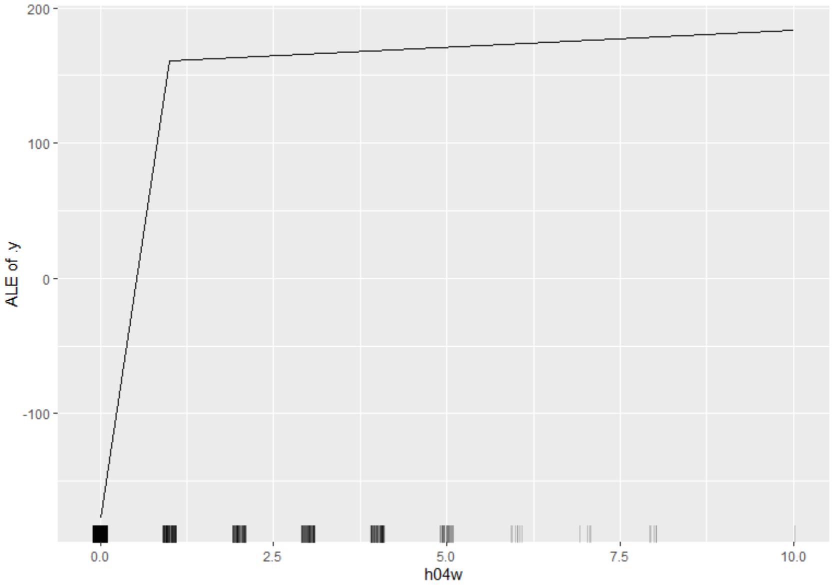

Figure 33: ALE for H04W on Dimension 3

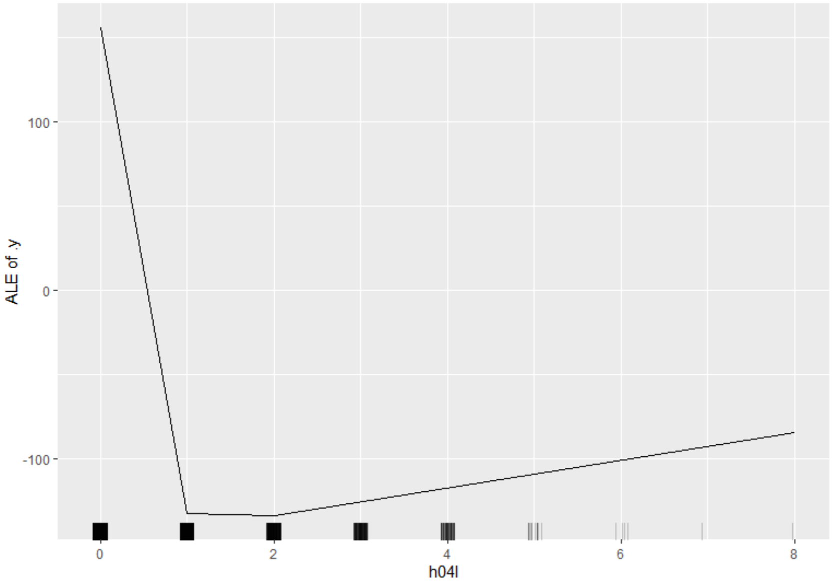

Figure 34: ALE for H04L on Dimension 3

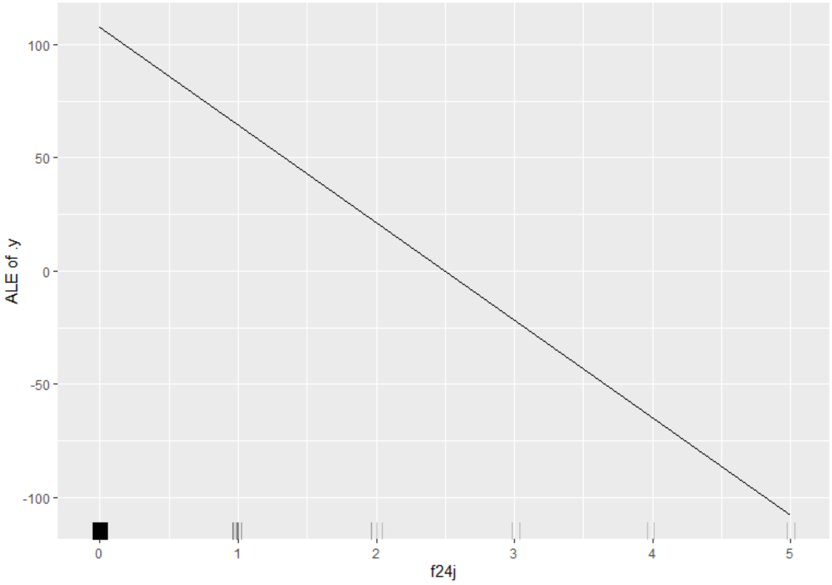

Figure 35: ALE for F24J on Dimension 3

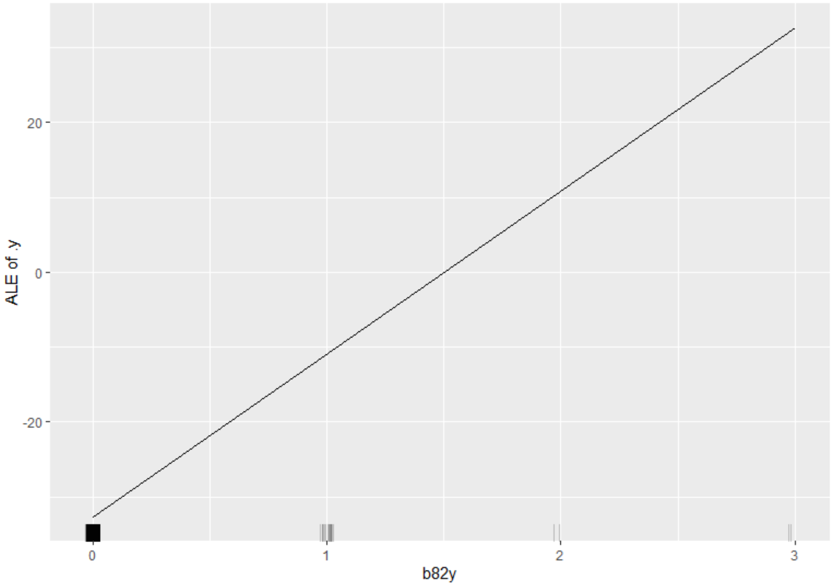

Figure 36: ALE for B82Y on Dimension 3

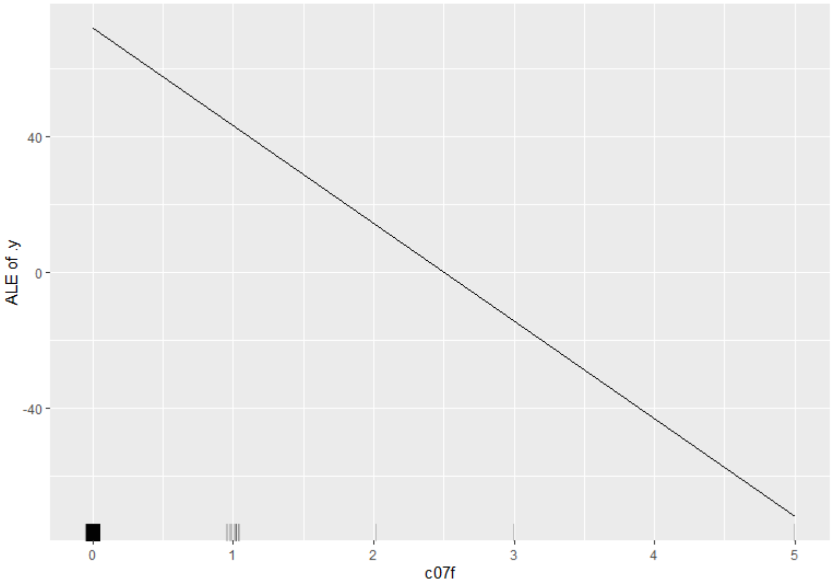

Figure 37: ALE for C07F on Dimension 3

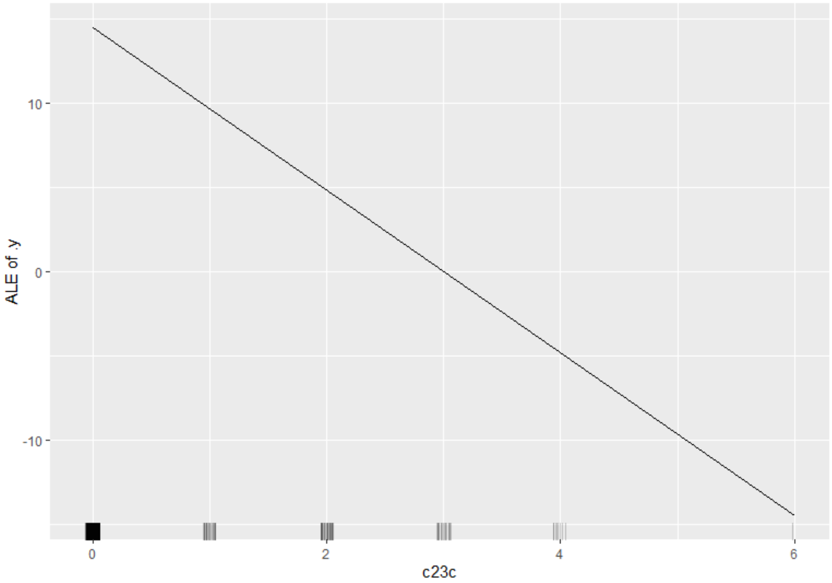

Figure 38: ALE for C23C on Dimension 3

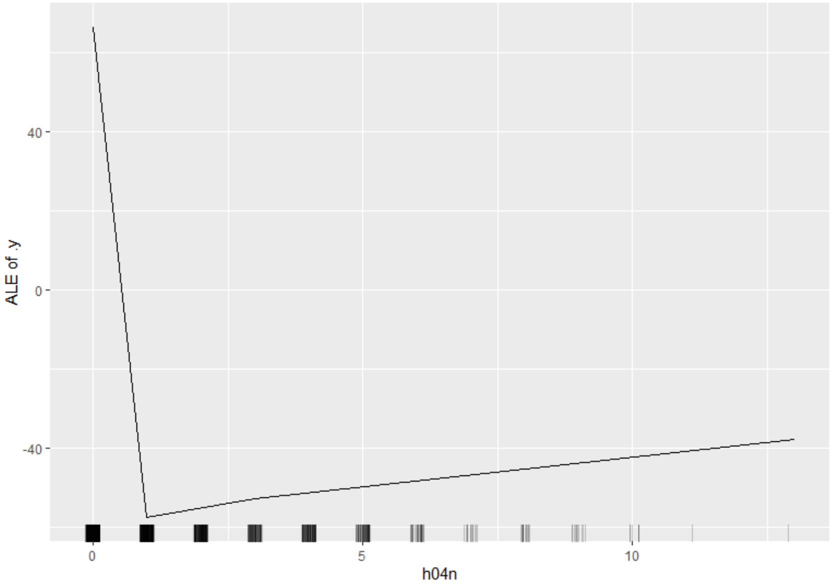

Figure 39: ALE for H04N on Dimension 3

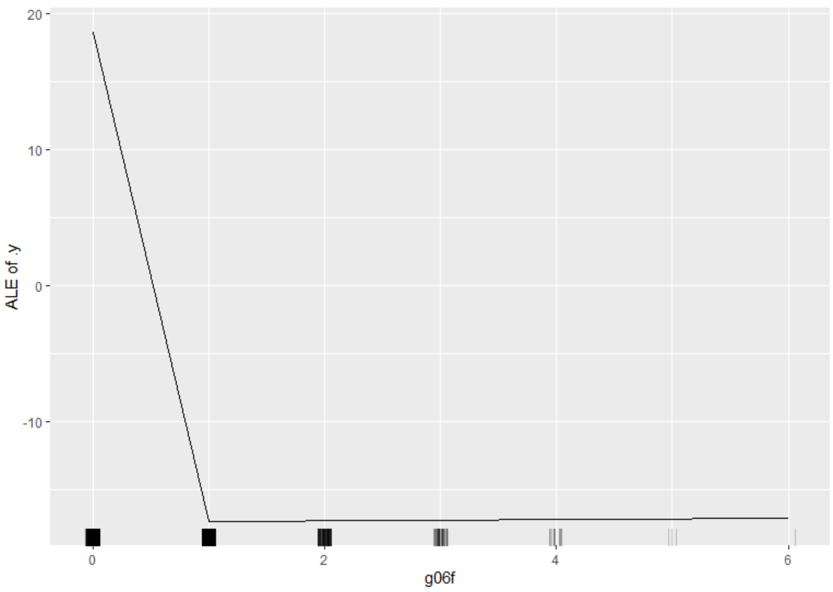

Figure 40: ALE for G06F on Dimension 3

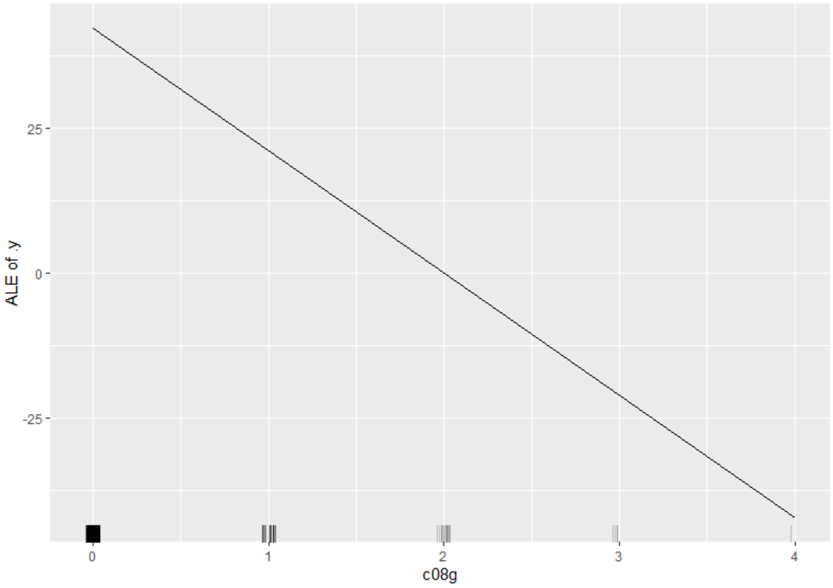

Figure 41: ALE for C08G on Dimension 3

Principal Component 4: Transmission

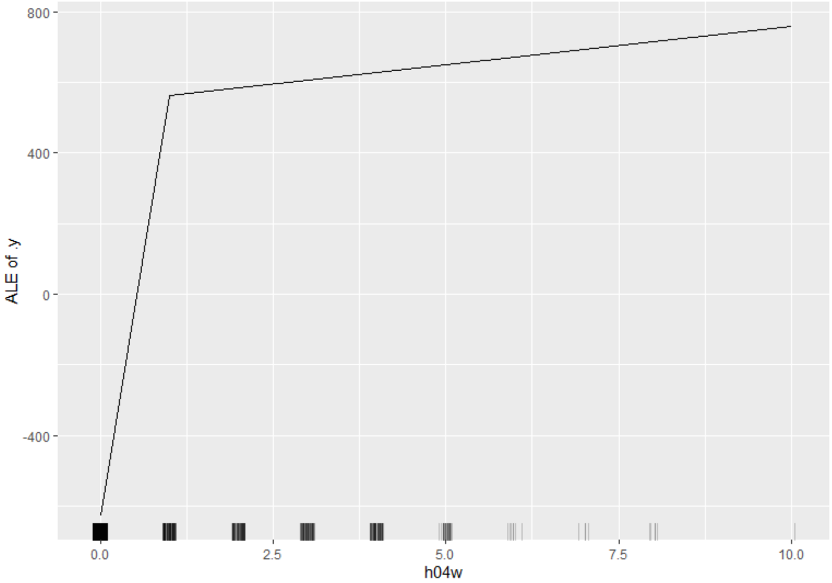

Figure 42: ALE for H04W on Dimension 4

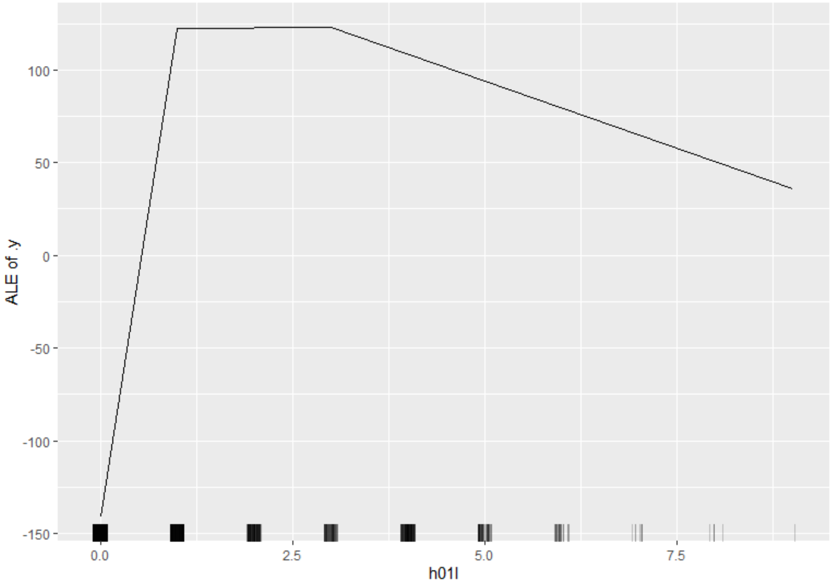

Figure 43: ALE for H01L on Dimension 4

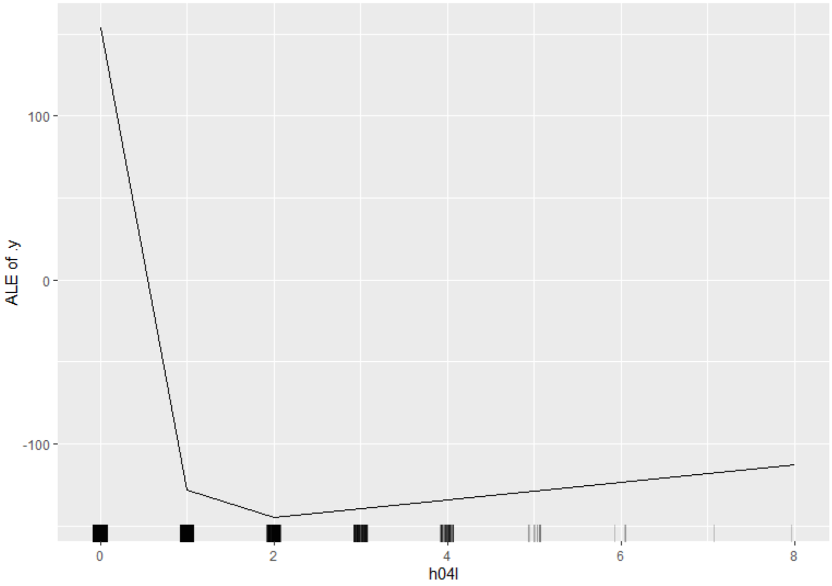

Figure 44: ALE for H04L on Dimension 4

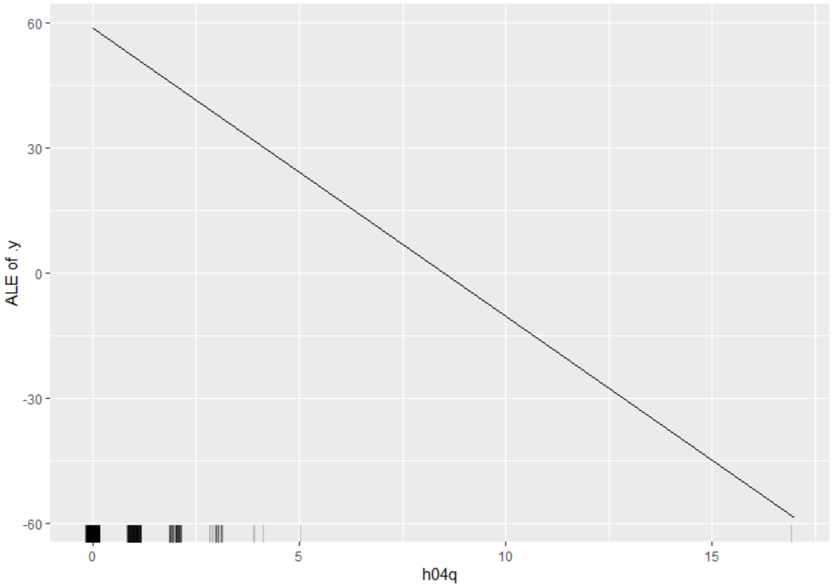

Figure 45: ALE for H04Q on Dimension 4

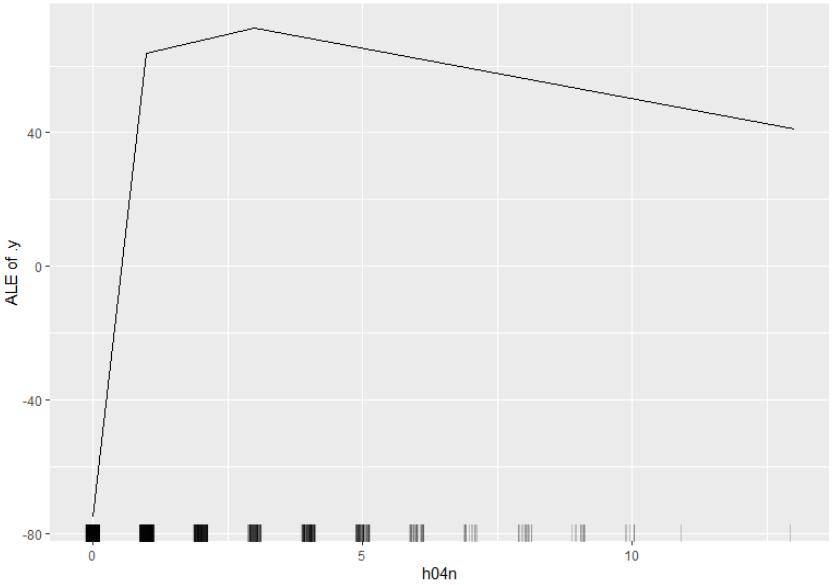

Figure 46: ALE for H04N on Dimension 4

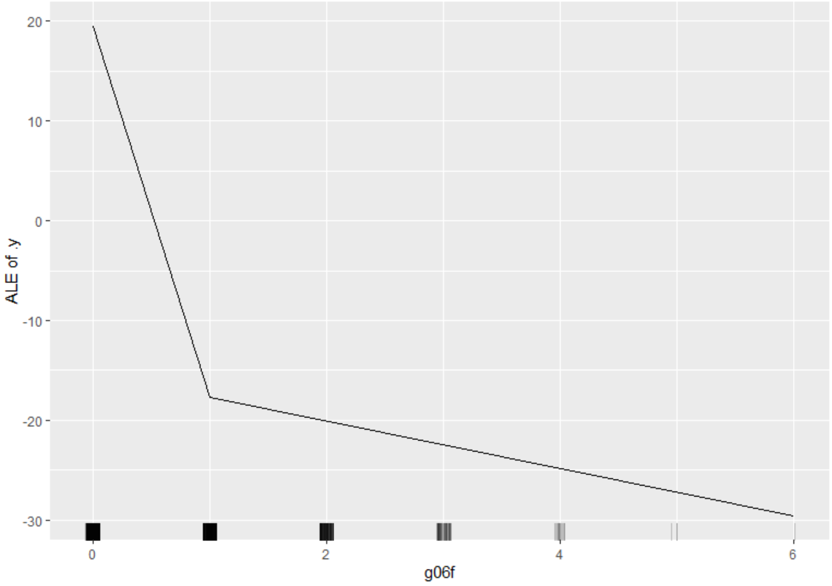

Figure 47: ALE for G06F on Dimension 4

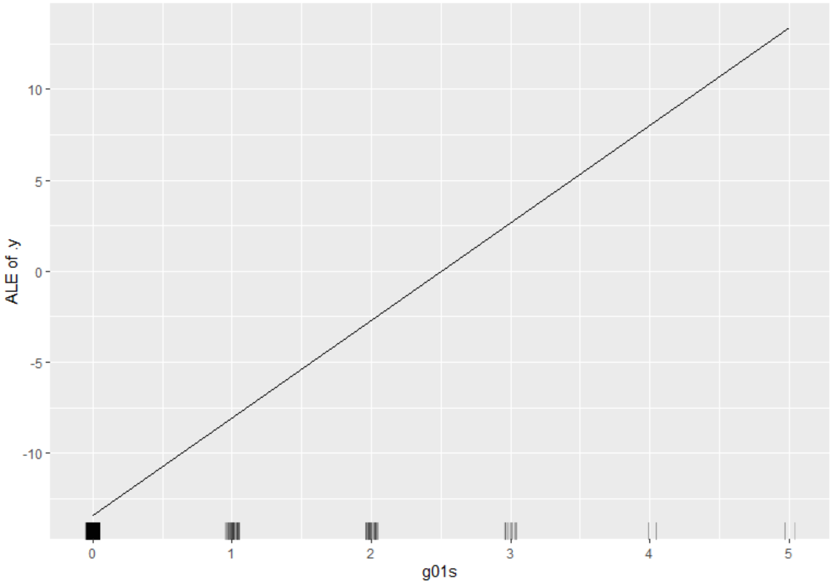

Figure 48: ALE for G01S on Dimension 4

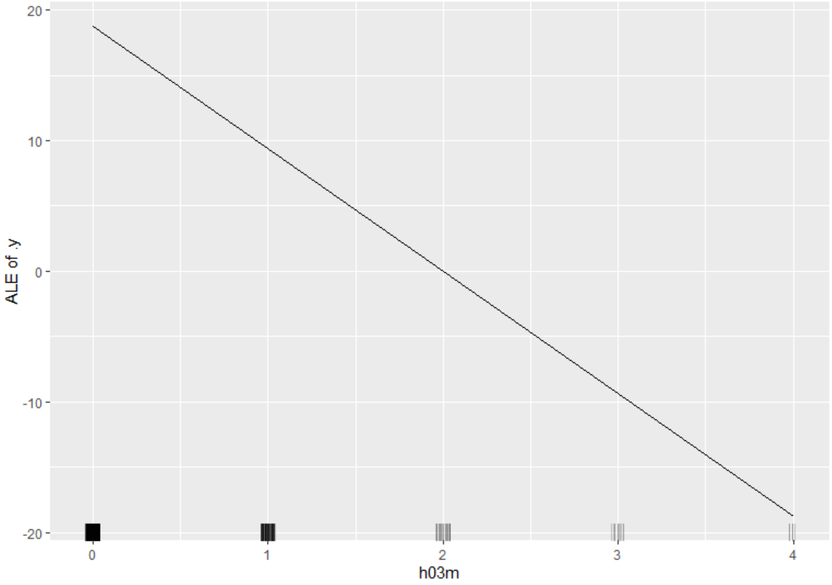

Figure 49: ALE for H03M on Dimension 4

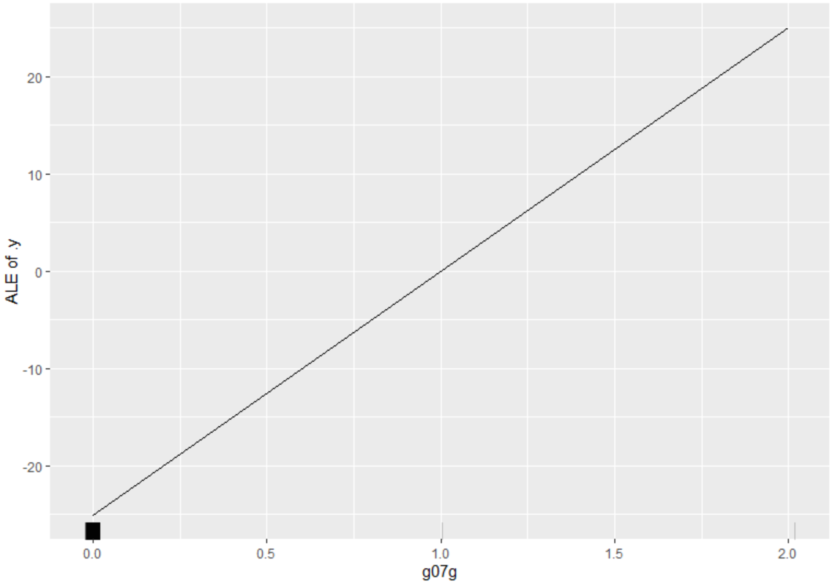

Figure 50: ALE for G07G on Dimension 4

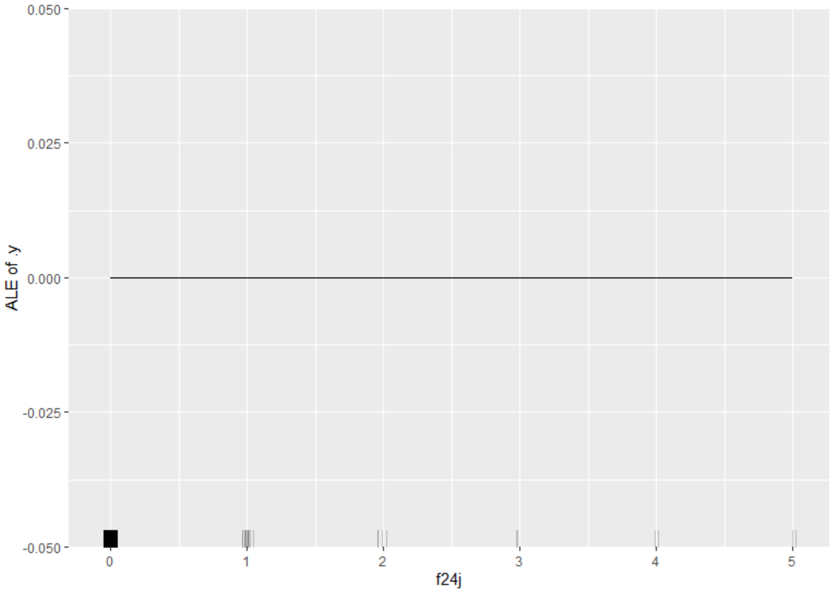

Figure 51: ALE for F24J on Dimension 4