DEPARTMENT OF MANAGEMENT SCIENCE

University of Strathclyde Business School

MSc Data Analytics Dissertation: Analysis of IT Service Desk Incidents for ScottishPower

Project Context/Scene Setting Document

Rebecca Charlotte Robertson

201766127

2018

Table of Contents

Organisational context

ScottishPower is one of the 6 top power companies present in the United Kingdom; they provide energy to over 5 million households across the country. ScottishPower are owned by Spanish energy company, Iberdrola, whom pride themselves on being one of the largest utilities companies in the world and a market leader in renewable – specifically wind-generated – energy (ScottishPower, 2018). This project is being held within the Corporate IT department of ScottishPower, who primarily focus on the monitoring and rectification of IT incidents for internal customers; i.e. those who are working with ScottishPower and have operational assets to undertake their role. The faults span across ScottishPower (SP Energy Networks, SP Retail and SP Renewables and SP Generation), IBM (IT company) and Vodafone (Communications company).

Background to project

Data has been gathered since 2012 which has been stored but never analysed. The client wished for this data to be combined, and mined for recurring themes that can be utilised to identify focus areas for further analysis, potential areas for incident reduction, and to improve incident resolution times. A combination of Historical Data Analysis, Live Data Visualisation Dashboard Creation and Forecasting has been requested to give the client greater insight into their data and an on-going, re-usable tool to track and monitor reported incident activity.

The Historical Analysis will include Exploratory and Structured analysis. Exploratory analysis should be conducted to gain a greater understanding of the data and what trends and relationships exist as well as some specific analysis. The client has requested the identification of the top 10 most frequently occurring incidents, and top 10 longest incidents to resolve. On top of this high level analysis, the client has requested that the top incidents be split by the three service categories (Infrastructure, Application and Network) as well as impact levels (Low, Medium, High and Critical) to understand how these influence the incident count and resolution times.

The Live Data Visualisation Dashboards will allow the client to track current incident reports, and prioritise incidents logged that have the greatest business impact.

Forecasting models will be updated automatically with live data to give the client a tool which can be used for forward planning of resources, and projects to implement self-service tools that could cut incident count and resolution times for incidents.

Why project is important

Fundamentally, the project will provide focus areas for the reduction in level of IT incidents, as well as reducing the average resolution time for incidents. This can reduce the amount of idle time experienced by the business, which could result in failure of ability to work efficiently, failure to reach targets, or potentially put customers or workers at risk (due to the business being an operational engineering company). The overall aim is to reduce the business impact of IT incidents occurring, through reduction of incident occurrence and reduction of resolution times. By analysing the trends in reported incidents, this can improve efficiency and reduce impact on business activity. The output from this analysis should pinpoint areas for further research and aid in reducing the level of incident count and time required to resolve said incidents.

Who is the immediate client/end user for the work

The immediate client for the work is the IT department of ScottishPower; being able to proactively manage the number of incidents logged and reduce the time required to resolve each incident will allow for less disruption to the business and greater mitigations to be in place for future potential faults. The combination of Historical Analysis, Live Data Visualisation Dashboards and Forecasting models will work effectively to provide the client with a means to understand historical trends, address business-critical incidents that should be prioritised to minimise business impact, and plan ahead for future predicted peaks in incident count.

Project Plan

Initial Plan

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13
02-Jul	09-Jul	16-Jul	23-Jul	30-Jul	06-Aug	13-Aug	20-Aug	27-Aug	03-Sep	10-Sep	17-Sep	24-Sep
Basic Plan, structure for 'project plan', introduction to data, request access to download R, Python, Tableau, begin research into literary review. Big ML?? Complete overview document (doc 1) Ethics form completed	Analysis of first month Team meeting Thursday Meet with Tim Tuesday	Methodology and begin Lit review Continue Analysis - present initial findings to team Team meeting Thursday	Continue Lit Review and revise methodology section Team meeting Thursday Moving home 26th July. Have ensured that weekends are free around this to complete work if necessary. Laptop provided by SP to work from home.	Continue analysis as per week 4. Team meeting Thursday Analysis - Begin formal analysis of data, and write up as findings arise. - Gain feedback throughout - Base style on feedback from week 3.	Continue analysis as per week 4/5. Team meeting Thursday	Re-draft Methodology and Literary Review. Team meeting Thursday	Conclusions and recommendations based on findings.	Write up conclusions and recommendations	Begin new full time role at ScottishPower	FT role at SP Draft copy of write-up	FT role at SP Re-draft week	FT role at SP Submit dissertation

Updated Plan

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13
02-Jul	09-Jul	16-Jul	23-Jul	30-Jul	06-Aug	13-Aug	20-Aug	27-Aug	03-Sep	10-Sep	17-Sep	24-Sep
Delay in beginning project due to HR/IT issue. Basic Plan, structure for 'project plan' Ethics form completed	Delay in beginning project due to HR/IT issue.	Begin with SP Request access to download R, Python, Tableau, Microsoft Power BI. Complete Overview Document.	Moving Home (no access to data) Begin research into literature review	Moving Home (no access to data) Continue research into literature review Begin modelling methodology	Access to data granted Create dissertation structure Develop literature review and methodology	Software request accepted for R, Tableau and Microsoft Power BI. Continue to model methodology, and choose technologies for analysis	Hip Injury	Hip Injury	Begin new full time role at ScottishPower - induction week out-of-office Unable to work on project	Exploratory analysis in R Issues with data processing: requested more up-to-date equipment (laptop repeatedly crashing)	Upskill in M Query Language and DAX (both used in Microsoft Power BI). Concatenate historical data	Upskill in M Query Language and DAX (both used in Microsoft Power BI). Meeting re Proactive/ Reactive split

Week 14	Week 15	Week 16	Week 17	Week 18	Week 19	Week 20	Week 21	Week 22	Week 23	Week 24	Week 25	Week 26
01-Oct	08-Oct	15-Oct	22-Oct	29-Oct	05-Nov	12-Nov	19-Nov	26-Nov	03-Dec	10-Dec	17-Dec	24-Dec
Data cleanse activity - meet with dept to understand key variables. New laptop issued due to hardware issues	No access to software or network drves with data due to laptop replacement	No access to software or network drves with data due to laptop replacement Software access granted late in week 16	Sandboxing/Quick and dirty analysis: Creation of interactive dashboards using sample of data and variety of visuals Meet with department for feedback	Exploratory Analysis in Microsoft Power BI and R Data Cleanse	Follow up with client on dashboard visuals, make changes as appropriate	Conduct analysis specific to project spec, using Microsoft Power BI	Forecasting analysis using Microsoft Power BI	Re-draft methodology and literature review	Conclusions and Recommendations	Re-draft	Christmas week/re-draft	Re-draft and send to printing for submission

References

Scottish Power (2018). About Scottish Power [online] Available at: https://www.scottishpower.co.uk/about-us/ [Accessed 20th December 2018]

DEPARTMENT OF MANAGEMENT SCIENCE

University of Strathclyde Business School

MSc Data Analytics Dissertation: Analysis of IT Service Desk Incidents for ScottishPower

Client Report

Rebecca Charlotte Robertson

201766127

2018

Word Count: 15,939

Signed statement

Except where explicitly stated, all the work in this dissertation – including any appendices – is my own and was carried out by me during my MSc course. It has not been submitted for assessment in any other context.

Signed: ……

Summary

The client (ScottishPower) had 2012-present day data, which contained information on all of their IT incident reports. The client wanted to identify focus areas for both incident count reduction and resolution time reduction. The client also wanted a view on the percentage of incidents reported that were planned, or unplanned (proactive or reactive). The client requested that analysis techniques were re-usable, and that data visualisation be used effectively to glean insight easily from the data.

The methodology, based on the literature review, shaped the format of the analysis. The methodology contained the following: Data Gathering, Data Pre-Processing, Exploratory Analysis, Sandboxing/Quick and Dirty Analysis, Full Data Analysis, Conclusions and Recommendations. The analysis was split into three sections: Historical, live (dashboard visualisations) and future forecasts.

From the literary review, appropriate analytical software were chosen to conduct analysis. R was chosen for analysis due to its statistical abilities and analytics packages, and its ease of use for replication (as per client's request). Microsoft Power BI was chosen for data visualisation due to its analytics packages, clean visuals and its use of M Query language (allowing for replication).

Exploratory Analysis showed that the split of planned:unplanned incidents was 40%:59%. It also highlighted that the Application category was the most common for incident count, and that there was a linear relationship between impact level and count of incident (low impact had greatest count). SAP has the greatest percentage of incidents associated with it overall, and August is the most common month for increased incident count. As expected, the majority of incidents are from the UK.

The client requested a 'top 10' list of the most common incident types for ScottishPower as a whole; these are as follows: Batch Process, Wintel Server, Unix Server, SAP Applications, Shared Mailbox, Application Service, Linux Server, Switch, IT Management Tool, and Other Support Service. The client requested that the 'Top 10' most common incident categories and resolution times be presented split by both impact and resolution category, to understand which impact levels and resolution categories occur most frequently, and which take the longest to resolve. The output from this was that SAP Application and Batch Process are the most common types of incidents to occur, and the longest to resolve.

Dashboards created for the client in Microsoft Power BI will present the level of incidents at any moment based on live data. The dashboards allow the client to track the level of incidents and resolution times occurring across all impact levels and categories, and to drill down into the data visualisations provided to gain greater insight into the data.

Forecasts were built for the 2019-2022 values for incident count and resolution time. These generate an accurate picture of the future incident count/resolution time values, based on live data.

Acknowledgements

I would like to thank ScottishPower for giving me the opportunity to work on this project, providing access to data as well as their guidance and support throughout the duration of this project. In particular, I would like to thank the ScottishPower Foundation for supporting me through my entire MSc – providing me with unforgettable experiences and opportunities that I appreciate immensely and have learned a huge amount from. I would also like to thank my advisor, Tim Bedford, for his direction and supervision through this piece of work.

Contents

1.0 Introduction/Problem Statement

The project undertaken was on behalf of ScottishPower's Corporate IT department, based in Glasgow. The client had a lake of data which had not had any analysis conducted on it; they wanted to glean insight from the gathered data to understand and address the reasons for incidents reported to the department for rectification. The client had some specific high level requests, as well as an element of forecasting to be conducted that would aid them in some way.

The client requested that the data was split by Service Category, that a 'top ten' incident list was created, that recurrent themes were identified and that focus areas for potential incident reduction/follow up could be identified. The client requested that analysis techniques that were implemented were re-usable, so that upon completion of the project the department could continue to monitor the incident levels without a data analyst in their team. At present, the data is extracted from a database into large Microsoft Excel documents which are difficult to cleanse, manipulate, perform analytics or gain useful information from.

At a high level, the client requested that data was extracted from ITSM (source), into Microsoft Excel, that re-usable analysis scripts were developed, the above analysis was conducted and presented, results formulated and conclusions presented.

The structure of the dissertation is as follows; an initial literature review was undertaken which formed and shaped the methodology which acts as a driver for the analysis, following the process of: data pre-processing, data cleanse, exploratory analysis, sandboxing, data analysis, and conclusion/write-up of findings. The second last section – data analysis – is split into a further three sections: Historical analysis, Predictive 'Future' modelling, and 'Live' analysis dashboards to meet client needs.

2.0 Literature Review

The Literature Review is split into 3 sections: General Data Analytics literature, IT Faults literature, and Data Visualisation literature. The Methodology was then created and written up in chapter 3: 'Methodology' reflecting the findings based on the literature reviewed.

2.1 General Data Analytics

2.1.1 Defining 'Data Analytics'

The client requested for data analytics to be performed on their data, to produce results that would aid the reduction of IT incidents logged, as well as aiding the reduction of resolution times for IT incidents that take the longest to resolve. Before commencing with the client project, it felt suitable to define what is meant by "Data Analytics". Informatica (2018) define data analytics as the "pursuit of extracting meaning from raw data using specialized computer systems". I would agree with this statement, and their extended definition which explains the underlying process of analytics being to "transform, organize and model the data to draw conclusions and identify patterns". Techopedia agreed with these definitions also, stating that data analytics is the name given to both the qualitative and quantitative methods used to improve how efficient and how profitable businesses are. They describe the process as extracting data to detect and analyse trends and activities in the data, changing the process slightly based on the needs of the business (Techopedia, 2018). These definitions back up the client's need for Data Analytics, as they wish to use the data gathered to identify patterns, draw conclusions and improve efficiency within the business.

2.1.2 Data Cleanse

The client outlined that data had been gathered for a long period of time. This data has not been formatted in any specific way, due to it not being used for any analysis. For this reason, the data must first be checked and cleansed as appropriate. Maletic and Marcus (2000) describe how the source of data is the most significant factor in terms of its integrity; they state that "data entry and acquisition is inherently prone to errors" and discuss the requirement for data cleanse to be conducted. Maletic and Marcus (2000) define data cleansing as the means by which "errors and inconsistencies" in data can be removed (Maletic, J. and Marcus, A., 2000). Edwin de Jonge. and Mark van der Loo (2018) describe a five-step process of cleansing and normalising data: 1. Checking and normalising the raw data, 2. Fixing the data to ensure that it is technically correct, before imputing it, 3. Ensuring data is consistent through estimation, analysis and derivations, 4. Forming results, whether that be in tabular form, or another visual plot, and finally, 5. Formatting an output for the client. This process of cleansing and 'fixing' the data must be completed before any analysis can take place, to ensure that the data – and therefore the results drawn from the data – are a high standard (de Jonge, E., and van der Loo, M., 2018).

2.1.2.1 Forming Concise Datasets

Narayanan (2016) explains the importance of only using data that is required, and not "using data for data's sake"; he highlights the warning signs associated with gathering excessive amounts of data, and – instead – highlights the significance of good quality, relevant data that is in-line with what the client is trying to achieve. Data should be able to tackle a business problem or have some sort of "actionable insights" that can be taken from it (Narayanan. R, 2016). A way of ensuring that only relevant data is being used for analysis when using coding languages such as R or Python, is to create dataframes or datasets that include only the required information for analysis. Once data has been cleansed, irrelevant data can be removed to allow analysis to commence. De Jonge and van der Loo (2018) give an overview of what a dataset is; they describe it as a collection of data that allows us to understand various elements of "real-world objects". If data is "technically correct", values can be recognised as belonging to a particular variable, and will be stored in a data type that represents the domain of said variable. At a high level – words will be stored as string values (text) and numerical values as numbers to ensure that formatting of data is consistent across the dataset (de Jonge, E. and van der Loo, M, 2018). Datasets are used to allow analysis to be conducted without processing all of the data that has been gathered by the client. This is beneficial for trying out different types of analysis quickly, without having to process excessive amounts of data that may be particularly time consuming, or require large processing power.

2.1.3 Sandboxing and Quick and Dirty Modeling

The action of conducting analysis on a small sample of data which can then be scaled up is called 'Sandboxing'. Narayanan (2016) discusses the benefits of the sandboxing technique; in order to design a model that would be suitable for a client; the sandboxing technique can be applied. 'Sandboxing' refers to using a small sample of data to create a model, ensuring data integrity and ensuring that the model is suitable for the full dataset before applying it on a large scale. Narayanan explains that this allows for manipulation, and greater understanding of the data but on a small scale. This means that data can be modelled, manipulated and tested using far less processing power, much faster than using the full dataset, allowing requirements to evolve in an incremental manner (Narayanan. R, 2016). Some elements of the project were defined explicitly by the client, however other elements were not. Applying a technique such as sandboxing would allow ideas to be evolved and presented to the client quickly and efficiently to gain feedback before scaling up to the full dataset The client requested that dashboards be created that present up-to-date information, however have not used any visualisation technologies in the past, and therefore could not give a precise description of what visuals they wished to see on a dashboard; this technique can be applied in this context to gain an understanding of client preferences. Similar to the sandboxing technique, 'quick and dirty' modeling is a means of conducting analysis in a fast and efficient way to try and test new techniques and present them to the client without investing an excessive amount of

time. Robertson (2013) explains the benefits of quick and dirty modeling in their book, 'Mastering the requirements process', where they describe how a quick and dirty prototype can provide a "rapid" explanation and can clarify areas of misunderstanding or where required elements may be missing (Robertson. S, and Robertson. J, 2013). This, combined with sandboxing would allow analysis to be conducted and presented to the client in a quick and efficient way, to present options to them and understand their preferences.

2.1.4 Exploratory Data Analysis

As mentioned, the client had some defined questions that they wanted answers to, but also wanted to understand what else their data could offer them. Mawer (2017) discusses the importance and significance of Exploratory Data Analysis; she describes it as "crucial" and highlights the benefits: it provides the context needed to start building data models, it provides certainty that the results produced by data analysis are valid, and it shows that the analyst themselves are not biasing the output with their own personal assumptions. Mawer states that Exploratory Data Analysis is used to "understand and summarise the contents of a dataset", relying on visualisations to highlight characteristics that the analyst would not otherwise know of. By conducting exploratory analysis, the client could gain trust that the findings presented on the defined questions were genuine, and would gain a greater understanding of the potential further analysis that could be conducted using the data that already existed. Peng and Matsui (2017) agree on the importance of exploratory data analysis, stating that it typically relates to looking at relationships and distributions of two or more variables, as well as the components of the dataset as a whole. They state that using visualisations for exploratory data analysis is the most important tool, due to how easy graphical information can be read and understood by the end user. Peng and Matsui state that there are three key goals of exploratory data analysis: "1. To determine if there are any problems with your dataset. 2. To determine whether the question you are asking can be answered by the data that you have. 3.To develop a sketch of the answer to your question" (Peng. R, and Matsui. E, 2017). In a book written by Tukey in 1977, he states that statistical hypothesis testing is used too often for conducting analysis, and instead that the data itself should be used to determine what should be tested in the first place – this links directly with the needs of the client; confirmatory data analysis would not be useful to the Scottish Power as they do not hold knowledge on the data; Exploratory Data Analysis would therefore be of greater benefit to develop an understanding of the data and to explore potential areas for further analysis.

2.1.5 General Data Analytics Software

A variety of Data Analytics tools are available on the market, reviewed on a regular basis by Data Scientists. In a review written by Gleeson (2017), the benefits of a variety of data analytics/data science languages are compared. The top three languages reviewed are R, Python, and SQL. Gleeson states that R, released in 1995, has become stronger over time and has a wide range of

quality open-source analytics packages; a key strength being its statistical and data visualisation capabilities. Gleeson points out the negative aspects of R also; it is not a fast means of processing data, and while it is good for analytics, it is not strong in general programming. For the purposes of this project, the chosen software package would be used for statistical analysis, and data visualisation; for this reason, R would fit the requirement for the analysis for this project. The second language to be reviewed was Python; this is a more general purpose programming language with a wide range of open-source modules also. It can, however, have some issues working with large datasets that are not fully cleansed due to it being a dynamically typed language – if type errors are present, this can cause significant problems. For this reason, Python may be a risky software package to use, given that the data gathered by the client has not undergone any pre-processing prior to this analysis. Gleeson states that for statistical and data analytical purposes, R is the preferred software between the two. SQL – the third-top language – is designed for querying databases; due to the nature of the analysis being conducted, this was not a requirement of the project and therefore this language would not be suitable.

User review website 'KD Nuggets' generates polls, for which data analysts/data scientists can submit reviews on their preferred top analytics/data science tools. The website ranks the top 3 tools as Python, R and SQL – in agreement with Gleeson (2017). Piatetsky's article for KD (2017) shows that from 2015-2017, Python has become increasingly more popular (increasing from 30%-55% of users utilising the tool). In 2017, R sits marginally behind Python as the preferred tool, however it has ranked higher overall, steadily increasing from 46%-54% of users preferring it as a tool (Piatetsky. G, 2017). Website 'Sharp Sight' wrote an article in 2018 discussing the current most appropriate language to learn for those who want to become data scientists; it was clear that R was the preferred tool over Python for data science activities for a number of reasons, including the built-in packages available in R, the syntax used, and its statistical roots (SharpSight.com, 2018). A range of data analytics tools designed with a focus on data visualisation are reviewed in section 2.2.2.

2.2 Data Visualisation

Due to the vast amount of data gathered, the client wanted to gain insight and identify 'unknown unknowns'. It became clear that the data being presented in tabular form was not beneficial for the client, as they found it difficult to extract meaningful insight. For this reason, data visualisation and creation of informative dashboards was presented as a solution. Thomas and Cook (2005) define data analytics visualisation as "the science of analytical reasoning facilitated by visual interactive interfaces". Visualisation as a tool is particularly important for data that is large in scale and contains multiple dimensions. Interactive visualisations, therefore, can allow for deeper exploration of data that will allow greater understanding of the dataset, in a way that would not otherwise be possible (Thomas. J, and Cook. K, 2005).

An important element of Data Visualisation is how visually pleasing each dashboard is. Often, dashboards contain bright colours to attract the attention of the viewer and draw their attention to specific points. Unfortunately, this can often be a major downfall for visualisations – colours that are too similar or too extreme (too bright/dark), or those which blend into the background are all risks when designing a dashboard. A paper by Christopher G Healey of The University of British Columbia explains how to effectively choose colours for visualisations. Healey states that three elements must be considered when choosing multiple colours: the distance between colours, the linear separation and the category of each colour. Healey's results suggest that the method that he formed can be used to select a group of colours that will provide easy-to-understand differentiation between elements. Healey measured the average response times to identifying the level of a target data type by first using 3 coloured lines, followed by 5, then 7 and finally 9 data types. For each test, all of the colours were used as targets to avoid any bias in terms of being able to see one colour more clearly than another. Findings showed that it was much easier to identify target points with 3 or 5 data-types, with timings increasing by almost 100% for the 7 and 9 data-type graphs. This shows that dashboards, while being effective in presenting a vast amount of information should not be overloaded or users will not be able to read them as effectively as we may hope (C.G. Healey, n.d.).

2.2.2 Data Visualisation in practice

Al-Hajj et al tackle the issue of the vast amount of multi-dimensional data generated in the health sector by – with the aid of experts – creating a visual analytics tool which could be used to aid the understanding of data by health professionals. The tool presents data in a clear, understandable way which can improve the level of knowledge and the justification of decisions made by these professionals. This is an example of how data visualisations can be used as a means of decision support mechanism. The unique element to this analysis tool is that it is capable of dealing with 'wicked' problems – ie those that require a "multidisciplinary approach to solve them and make informed decisions" (Al-Hajj et al, 2013). Another example of dashboards being used to solve problems and make informed decisions is by Selby (2005), a researcher who argues that measurement-driven dashboards are effective in identifying indicators for large scale systems. Selby investigates the importance and capability of dashboards, stating that "measurement-driven dashboards provide a unifying mechanism for understanding, evaluating, and predicting the development, management and economics of large-scale processes". Selby's paper focuses on real large-scale projects, as well as the relationships uncovered by the dashboard itself. Overall, Selby proved that measurement-driven dashboards are capable of uncovering information that allows for greater visibility and understanding of large-scale systems as well as being capable of feeding back information to companies or specific pieces of work. Selby believes that in managing IT infrastructure and all of the software used within companies through dashboards, they are able to identify areas of weakness/poor progress, and allows for people to learn from previous mistakes,

understand their current market, and plan ahead. A key weakness identified is that many companies fail to effectively manage the array of data types gathered and therefore fail to make correctly informed decisions; this could mean that incorrect findings can be drawn, and that it may take an exceptionally long time to reach conclusions, or that findings are very limited. Selby explains that different dashboards can be used for different purposes; the example he gives is that one may relate to developmental goals, and another management-oriented goals. Selby uses a number of detailed diagrams to allow the reader to understand the specifics of all of the dashboards used; a huge variety of elements are incorporated into each dashboard to visually communicate a number of things. Hyperlinks, colour-coding, drop-down menus and layered charts are among some of the methods used to communicate a vast amount of information in a simple, effective way (R. Selby, 2005).

2.2.3 Data Visualisation software

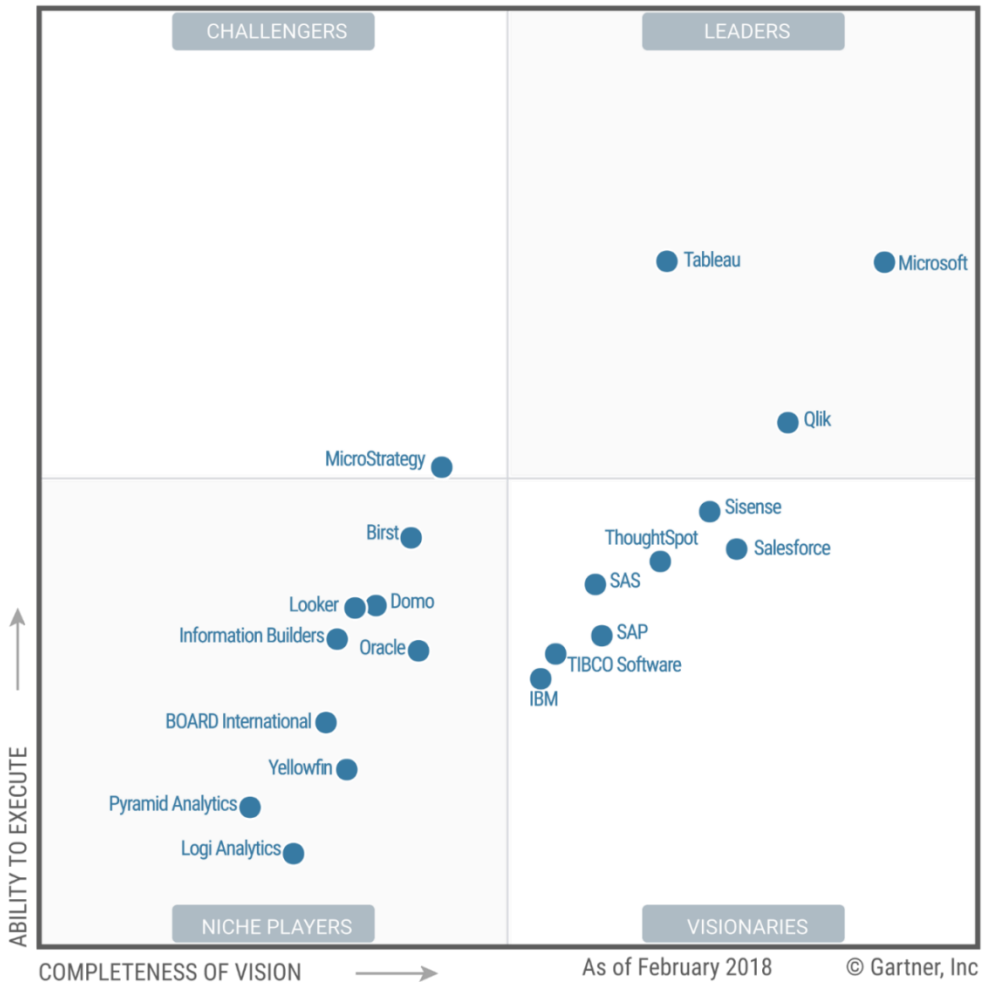
An article written by Zhang et al (2012) of Konstanz University, Germany, sets out to compare the market-leading visual analytics software available in 2012. In order to make a fair comparison, they sent a set of questions to each of the market leaders, and included the responses of those companies who returned their answers. The toolkits were judged on 3 key areas: visualisation functions, analysis capabilities and supported development environment. When considering toolkits that could work both on their own and integrated into other information infrastructures, some software stood out: Tableau, Spotfire, QlikView, Power Pivot (later to become Microsoft Power BI) and JMP (Zhang et al, 2012).

When designing a re-usable solution for the client, it was important to ensure that the chosen technology would be easily understood, as well as being easily adaptable should they require to make changes in the future. A key element of the analysis was to provide the client with a greater level of detail than what they can currently glean from their raw data. For this reason, a data visualisation tool with a straightforward interface had to be chosen. An article written by Sarah Anne Murphy of Ohio State University (2013) investigates the application of “rapid analytic and data visualisation software” for libraries. Murphy confirms that accessibility and low-cost solutions are available, and are ideal for making sense of large volumes of multi-faceted, messy data. One technology which is investigated is Tableau; the internal features of Tableau allow for irrelevant data, or obvious outliers, to be mitigated in a straightforward manner. Another notable benefit is that – similar to some popular applications such as Adobe – Tableau does not require users to buy the software if they simply want to view dashboards; this is an effective way of trialling the software to measure business response prior to committing to rolling out the software across the company. One of the best features of Tableau is that it can produce visualisations in real-time from databases. This removes the high risk associated with the likes of production of excel spreadsheets, where manual data entry is prominent and extremely risky. Tableau removes the human element (in terms of the

raw data) and can automatically update on a daily basis. Tableau also has online capabilities, so that it can be used on a number of platforms or embedded in website pages – this allows a wider range of access. While using Tableau is incredibly straightforward, there are a number of advanced capabilities available, should one wish to conduct more complex analysis or produce more advanced visualisations. Similar to many available software there is an easy-to-access level, with potential to advance and create unique analytic tools (S. Murphy, 2013).

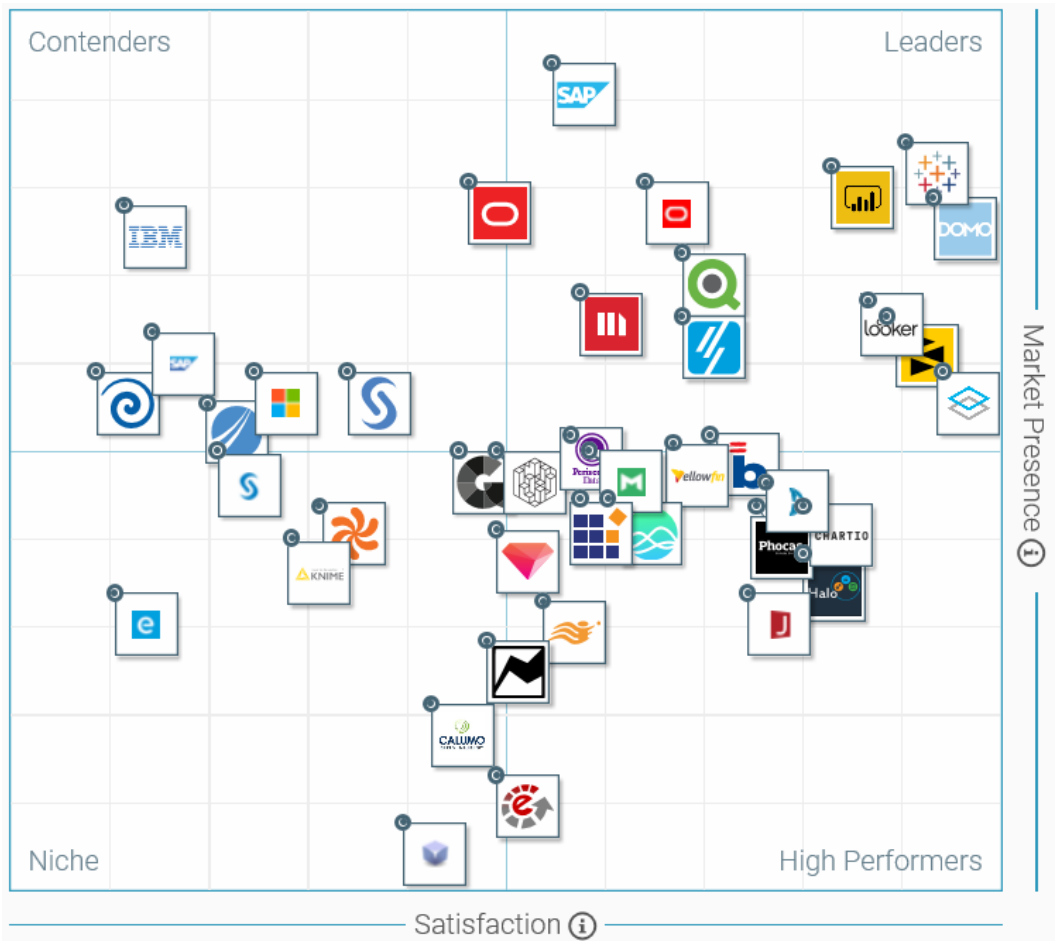
Gartner (2018) generate a report of the top Business Intelligence Platforms: "Magic Quadrant for Analytics and Business Intelligence Platforms". In the report, they define 15 product capabilities, as well as 15 critical capabilities (split into 5 sections: Infrastructure, Data Management, Analysis and Content Creation, Sharing of Findings and Overall platform capabilities). All of the market competitors were plotted on Gartner's 'Magic Quadrant' based on Completeness of vision and Ability to execute and subsequently split into 4 categories: niche players, challengers, visionaries and leaders. In the 'leaders' quadrant we see 3 key companies offering visualisation tools: Tableau, Microsoft and Qlik.

The report then goes on to give a critical review as well as strengths and cautions of each of the plotted companies/associated software. Tableau's strengths could be identified as 'Gold Standard' interactive visual exploration, focus on customer experience and success, expanding deployments and standardisation rates, as well as flexible deployment options. Tableau's cautions could be identified as market mainstreaming, their pricing and packaging and lack of complex data model support, and their product vision. Microsoft's strengths (for Power BI) could be identified as it being a low-priced incumbent, being particularly easy to use and having a strong visual appeal, the product vision, customer experience, and the azure cloud capabilities (in-line with the client's vision). Seeing Microsoft Power Pivot and Microsoft SQL Server BI merge to become Microsoft Power BI is encouraging as it combines the strongest elements of both successful Microsoft Excel Add-ons to become its own reliable data visualisation software. Key cautions of Microsoft's Power BI were the breadth of use, and their sales experience. The third top technology was 'Qlik', their strengths are identified as it being a scalable product, and they have differentiated marketing and partner networks. Their cautions are that it is a self-service product, with a high cost, and that there are migration challenges present (Gartner, 2018).



Leading BI Technologies Gartner (2018)

G2 Crowd produce a similar style of report to Gartner, however they are a “peer-to-peer business solutions review platform” showing unbiased ratings on user satisfaction for a number of platforms. The diagram below shows G2’s equivalent of Gartner’s magic quadrant; showing ‘contenders’, ‘leaders’, ‘niche’ and ‘high performers’. We see Tableau and Microsoft Power BI appearing in their leaders quadrant, alongside Domo, Looker, Qlik, Microstrategy and SAP.



Leading BI Technologies G2 Crowd (2018)

A review was undertaken of each of the top aggregated technologies, and these were raked based on their suitability for the client:

1. Microsoft Power BI
    - Offers data preparation, discovery, interactive dashboards and augmented analytics (Gartner, 2018)
    - Can operate in Azure cloud (fits with clients vision) or on its own report server (suitable for proof of concept)
    - Easy of use/integration with other Microsoft packages
    - Low-price licencing; no licence required to view reports
2. Tableau
    - On-par with Microsoft in terms of capability: offers interactive experience that allows business users and authors to access, prepare, analyse and present findings in their data (Gartner, 2018).
    - Offers data cleanse software as well as market-leading insight capabilities
    - Higher price-point than competitors

3. MicroStrategy
    - Strong integrated product: self-service data preparation, visual-based data discovery, and best-in-class enterprise reporting in one product. Considered ‘outstanding’ for planning, architecture and security, connectivity, and scalability (Gartner, 2018)
    - MicroStrategy's cloud solution lacks packaged domain and vertical content, and a robust content marketplace for customers and partners (Gartner, 2018)
    - Higher price-point than competitors and licences required to view reports
4. SAP
    - Several functional limitations noted; not suited to ‘citizen data scientists’
    - Limited interoperability between the two pieces of software available
5. Qlik
    - Lack of point-and-click capability; code-driven
    - Scored extremely low for migration between software and future development
6. Looker
    - Inaccessible for ‘citizen data scientists’ due to code-driven approach
    - Focussed on North America; minimal uptake globally

2.3 IT Faults

A key aspect of the analysis is surrounding the level of faults that the IT department can handle without putting pressure on the business. Fault tolerance of IT systems is of significant concern in terms of services and application access. Traditionally, check point/restart or duplication were among approaches taken for fault tolerance. It is also an option, however, to proactively monitor potential failures and mitigate the impact/likelihood of those on the system/application (G. Vallee et al, 2008). Järveläinen (2013) discusses the business impacts that arise as a result of IT incidents/failures, specifically focussing on the impact that data inaccessibility can have. Järveläinen recognises previous research with regard to data availability, disaster recovery and business continuity in terms of avoiding IT incidents though identifies a gap in terms of a “framework for information system continuity management”. Within the article, Järveläinen discusses the purpose and importance of business continuity management, describing it as a ‘socio-technical’ approach, with the aim of identifying and preventing risks before they occur and cause an impact (Järveläinen. J, 2013). An infographic produced by IBM on data analytics in 2018 gives a successful example of how implementing predictive analytics can benefit a business. The case study is on a US-based aircraft engine manufacturer, who utilised machine learning and predictive analytics for maintenance activities. The manufacturer became aware that, had the machine learning been implemented the year prior, they would have saved approximately $63 million (IBM Big Data & Analytics Hub, n.d.).

3.0 Methodology

3.1 Introduction

This chapter outlines the methodology followed to conduct analysis for the client. This was a 6-stage process which was developed to prepare the data to ensure its suitability for analysis, gain an understanding of the spread and nature of it, and then conduct analysis. The methodology was created based on the critical evaluation of data analytics literature, and was designed to ensure that the data could be used for each three elements that were requested from the client, as follows: Historical Data analysis, a 'Future' predictive model, and a 'Live' data visualisation model, as described in the proceeding sections below.

3.1.1 Historical data analysis

Historical data analysis of the full dataset was an effective means of tracking trends over time, using time-series analysis. Being able to perform monthly and annual comparisons was beneficial in presenting a graphical output that the business can use to gain a greater understanding of the data characteristics over time. Trend identification allows for solution implementation, whether that be greater levels of staffing to rectify IT faults, or the implementation of more self-service options for employees to resolve issues they are facing. The aim – as mentioned in the Problem Statement – is to reduce the level of incidents that have a high impact on the business. This analysis was completed in R as well as Microsoft Power BI, using DAX and M Query language.

3.1.2 Live data visualisation model

In order for the department to continue to monitor trends, and have an up-to-date view of their incidents, a dashboard was created that links directly to the raw incident reporting data, allowing users to drill down into the data to gain greater insight, should they wish or be required to do so. This means that at any point in time, the dashboard can be refreshed to give a snapshot of the current position for incident reports, and highlight any critical incidents that could have a strong impact on the business. This can be used to proactively manage faults, and highlight business critical incidents requiring immediate attention to avoid intensifying business impact. This was completed using Microsoft Power BI due to its interactive analytical capabilities.

3.1.3 Forecasting model

A forecasting model was built in Microsoft Power BI to be included as part of the client's live dashboard. The forecast will present the client with predicted incident counts based on each incident resolution category, and impact level. This prediction will allow for more efficient planning of resources to mitigate long waiting times, and will highlight areas where self-service solutions should be considered, to reduce business impact.

3.2 Methodology Model

The flowchart in figure 1 below shows the process that was followed to conduct analysis for the client. The first three steps of the methodology were applied to the full dataset as a whole, and the final three steps were repeated for each of the three types of analysis being conducted: Historical, live, and forecasting.

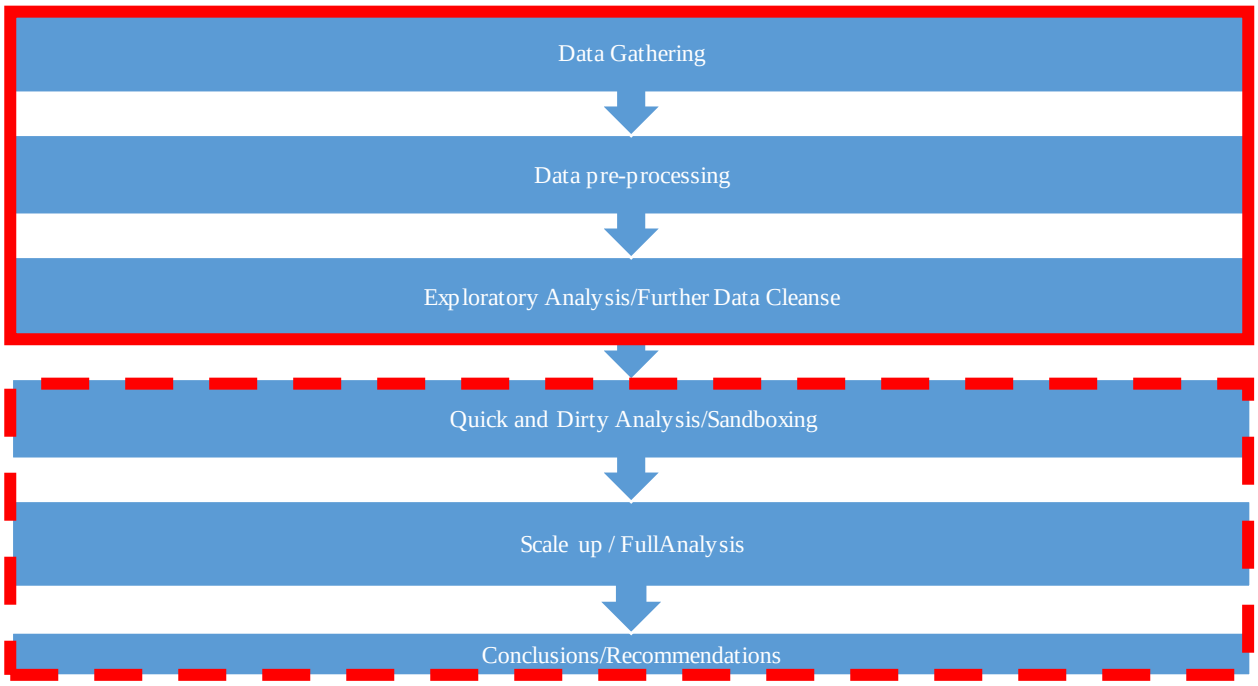


Figure 2 - Methodology Model

3.2.3 Analysis conducted to the full dataset

3.2.3.1 Data Gathering

The first step in the data analysis process is to gather data. Primary data was gathered from the company's source system – 'ITSM'. This data is predominantly quantitative, with some qualitative elements (for example, summary notes written up of the client's incident report). No secondary data was used for this analysis.

3.2.3.2 Data Pre-Processing

The data pre-processing step involves a data cleanse activity, ensuring that there are no anomalies present in the data, and – if the dataset is particularly large – removing columns/attributes that are not relevant to the analysis being conducted. Based on the literature and user preferences analysed in the literature review, R was chosen due to its data analytics packages and statistical background. Using R, the dimensions of the dataset can be presented, as well as the number of, and the type of different attributes present within the dataset. Another beneficial activity is to identify and remove blank/null values from the dataset; Narayanan (2016) highlights the importance of only using relevant data, as mentioned in the literature review. Given that the dataset being used had 55 columns (only 13 of which were relevant/suitable for analysis), removing those which were not

relevant resulted in the size of the data being analysed being reduced by over 75%. No information was removed that could have contributed towards the analysis; but several occurrences of repetition, null values, or call handler/employee-specific information was present that the client agreed should not be included in the final dataset. By removing this information, this allowed for the storage and processing of the data to be completed much more easily. A key limitation faced was the processing power of the laptop provided by the client, so this mitigated the risk of losing data due to computing hardware failure as a result of crashing due to overworking. The data cleanse can be conducted in R by creating data frames, or in Microsoft Power BI by removing irrelevant columns of data using M Query Language.

3.2.3.3 Exploratory Analysis and Further Data Cleanse

The third step in the process was to conduct exploratory analysis; this allows the analyst to gain a feel of the data's structure and characteristics. Mawer (2017) describes exploratory analysis as a "crucial step" in the data analytics process, as it mitigates the risk of analysts working on assumptions that they have of the data. Exploratory analysis can be done at a very high level in R, identifying attribute dimensions, highlighting correlations between attributes using correlograms and heatmaps, and visualising the spread and split of the data through histograms. Using Microsoft Power BI, a range of visuals can be adopted to give greater insight into the data; these include – but are not limited to – time series line graphs, split by various categorisations – whether that be product category, impact, year, or month; pie charts to show the split of the data, or mapping visuals to show the locational source of the data. Once irrelevant data can be identified as such, these can be removed as part of a second data cleanse activity.

3.2.4 Analysis conducted unique to analysis type (Historical, Live, or Forecasting)

3.2.4.1 Quick and Dirty Analysis/Sandboxing

The fourth step of the analysis was to apply the 'sandboxing' technique and 'quick and dirty' analysis. As previously mentioned, this technique is used when building models; taking a small sample of data which represents the variables present across the wider dataset; the analyst can test a model and check its suitability before scaling it up for use for the full dataset. This allows for changes to be made with ease, without requiring the processing power needed to analyse the full dataset or the time taken to produce a 'final' approach that the client may not like – this means that the client can request changes and they can be applied quickly and easily. On top of this, it also means that if any errors arise, it is much easier to understand the cause of the error on a small sample dataset, simply due to the lower volume of data.

Historical Analysis

Due to the client requesting specific analysis to be completed on historical data, the sandboxing approach was utilised to quickly design the visual output for this analysis and check that this would be suitable for the client prior to scaling up to the full dataset.

Live Dashboard

For the purposes of the 'live' model, the sandboxing technique and quick and dirty modeling was used as a means of back-and-forth with the client, gaining a greater understanding of their needs, and which visuals they preferred. Due to this being the first dashboard of its kind for the client, they were not able to articulate what exactly they wanted until they were presented with some examples.

Forecasting Model

For the forecasting model, sandboxing and quick and dirty modeling was utilised to determine which attributes should be predicted, and what visuals were preferred by the client.

3.2.4.2 Scale up/Full Analysis

The fifth step applied is to scale up the output from the Sandboxing applied previously. This involved importing the full dataset into Microsoft Power BI and R, and producing visuals for the client based on the full dataset, rather than the sample. Doing this gave a far more detailed picture when looking at time-series analysis, and allowed the client to 'drill down' into their data to a greater extent ('drill down' refers to investigating various levels of the data to gain greater insight/more information). This scaling up of the analysis was completed in each of the three analysis areas independently.

3.2.4.3 Produce Conclusions/Recommendations

The final, sixth, step of the analysis was to produce conclusions and recommendations for the client.

Historical Analysis

As requested, the output from the historical analysis was graphs showing the top 10 most common and top 10 longest to resolve incidents that were reported to IT, a) for the company overall, b) split by impact and c) split by resolution category. On top of this, informative output was presented from the exploratory analysis. These gave the client an insight into which areas required further analysis, and which areas should be focussed on for incident level, and resolution time reduction.

Live Dashboard

The output from the live dashboard was presented to the client, as a means of tracking incidents going forward.

Forecasting Model

The forecasting model presented to the client allows them to track the forecasted levels of incidents for the upcoming 4 years. This allows the client to plan ahead for forecasted peaks in incident count, and address areas that could be focussed on for incident count/incident resolution time reduction.

4.0 Data Pre-Processing

4.1 Introduction

This chapter outlines the data pre-processing techniques used to collect and cleanse the data prior to analysis being conducted. Data was collected as per the client's requirements and guidance, before being cleansed in R in preparation for exploratory analysis to be conducted. Following this, the data was cleansed in Microsoft Power BI, using M Query language to ensure that each time the client updated their source data, the same cleansing steps would be repeated, meeting the client requirement of 're-usable analysis scripts'.

4.2 Data Collection

Data was collected from a data lake in source location, 'ITSM (IT Service Management) store'. The ITSM data was then extracted into Microsoft Excel where it could be saved as a CSV file. No manual alterations were made in Microsoft Excel, due to the client's requirement that all scripts and analysis should be re-usable. Due to poor organisation of collected data, it took a significant period of time to collate all data, dating back to 2012. This historical collation of the data was a one-off and not something that the client will have to repeat, as the data models built will concatenate future data imports to one large data store which the client can filter on date as they feel appropriate.

4.3 Data Cleanse

The most extensive exercise conducted was that of data cleanse. This was done in two ways due to some analysis being conducted using R, and further analysis being conducted in Microsoft Power BI. Both data cleanse activities will be explained in detail. To make the initial data import easier, the mass of historical data was concatenated into one large excel file (as previously mentioned), to avoid importing over 80 files into R or Microsoft Power BI. As new data is gathered, this can be concatenated with the source file with ease in Microsoft Power BI.

The reports generated went into detail on a number of areas that were not required for analysis, so these were initially stripped out to ensure a concise dataset was used – this was necessary not only to keep the dataset as tidy as possible, but also to aid the processing issues mentioned previously.

Once the relevant fields were identified and highlighted, these were then checked for null values and anomalies. Where these existed, I went back to the business to gain a greater understanding of the data and its meaning. The initial guidelines given on how the data should be split varied from what was present in the reports, so this had to be agreed also. On top of this, many cells contained free-hand text where call handlers/customer support had written 'summary' notes, or similar. These notes could have been more informative if a more concise structure was followed; the free-hand element reduced the ability to glean insight from this data. After meeting with the business users, I was able to pinpoint key identifiers, clean the data and develop a suitable structure for analysis.

4.3.1 Analysis: R

In R, the data was read in, and the types of attribute in the dataset were checked, as well as the dimensions of the dataset (fig 2 below):

```
> #Check dimensions of the dataset
> dim(dataset)
[1] 1048575      55
```

Figure 3 - Dimensions of Dataset

The classes of data were a combination of factors, integers, and numeric values. Many columns were not required for analysis, as they contained information irrelevant to the client's requests. Rather than deleting data from the source, a dataframe was created which stripped out the data which was not required for analysis. Once the dataframe was created, values with multiple string values were replaced with numerical substitutes to allow for easier analysis. The total number of columns was significantly reduced to make the data more manageable (fig 3):

```
> dim(numeric.df)
[1] 1048575      13
```

Figure 4 - Dimensions of Dataframe

The next step in the data cleanse process was to remove rows which were incomplete or contained null values, renaming the dataframe appropriately (fig 4):

```
> #Remove NA values from dataset
> numeric2na.df <- na.omit(numeric.df)
```

Figure 5 - Removing NA Values

The dimensions of the new dataframe can be checked to ensure that null rows have been removed (fig 5):

```
> #Check dimensions of the dataset
> dim(numeric2na.df)
[1] 83867    13
```

Figure 6 - Dimensions of updated Dataframe

The class of the attributes in the dataframe can be checked for consistency (code shown in fig 6 and output shown in appendix 1):

```
> sapply(numeric2na.df, class)
```

Figure 7 - Attribute Class within Dataframe

The output from fig 6 (Appendix 1) shows us that that after formatting the data, all of the variables are integers other than the 'proactive.reactive' variable, which was added to differentiate whether the event was planned or unplanned.

A high level check of the data can be achieved through the 'head' function in R; this produces a sample of the data to gain understanding (code shown in fig 7 and output shown in appendix 2):

```
> #Check top 6 rows of the data
> head(numeric2na.df)
```

Figure 7 – Head Function: Overview of data

4.3.2 Analysis: Microsoft Power BI

The data cleansing abilities of Power BI were a contributing factor that was considered when choosing a 're-usable' method for analysis (as requested by the client). The query editor and data view of Microsoft Power BI works similarly to writing Macros, whereby all actions performed will be repeated each time the data is refreshed. This means that if you – for example – wished to desensitise data, you could do so in the data view of Microsoft Power BI, and this would be repeated in the background each time the report is refreshed, before the raw data is viewed by the end-user. This creates a semantic layer for the report-builder that does not affect the source data.

4.4 Difficulties

Throughout the analysis, various difficulties were encountered in each phase. The data collection was not as straightforward as expected, and the data cleanse had to be updated on several occasions as new issues arose. Visualising the data was relatively straightforward, but upskilling in M Query language and DAX within Microsoft Power BI required a large time-commitment.

4.4.1 Data Collection

Data collection presented difficulties due to the amount of data being analysed; the machine being used to process the data repeatedly crashed due to not being able to handle the file size. Initially, a sample was considered however when each year was plotted individually, trends could be identified that I did not want to smooth/overlook. To overcome these issues, a more powerful machine was required to process the data. This was achieved through desensitising the data, to allow for analysis to be conducted on a 64-bit virtual machine which could provide the processing power required.

4.4.2 Data Cleanse

The initial data cleanse activity converted all of the variables into numerical values for ease of analysis, and desensitised the data. As each step of exploratory analysis was conducted, steps had to be put in place to rectify issues encountered. These were addressed through removing values, and using dataframes in R. Once issues were identified and rectified in R, they could be written in M Query language in Microsoft Power BI to automatically be applied to all new data being uploaded to the Microsoft Power BI report (both for the analysis and the dashboard that was created).

5.0 Exploratory Analysis

5.1 Introduction

Exploratory analysis was conducted in R and Microsoft Power BI to gain a greater understanding of the data, before a further data cleanse and full analysis was conducted.

5.2 Analysis: R

Initial exploratory analysis was conducted in R to investigate the correlations between variables, as well as the spread of data. This was conducted using a table of numerical correlations, a correlogram and heatmaps.

5.2.1 Attribute Dimensions

Using R, a summary of the data could be produced to give greater insight into the spread of the data; this was a 5-figure summary, as well as the average, for each attribute (code in fig 8, and output in Appendix 3):

```
> #Summarise attribute distributions
> summary(numeric2na1.df)
```

Figure 8 - Summary of Attribute Distributions

5.2.2 Correlations

To calculate the correlation between variables, all variables must be numeric/integer values. The non-numeric (factor) values must be removed to allow this analysis to be conducted (fig 9):

```
> #Correlation of cleaned data
> correlation.numeric2na1.df <- cor(numeric2na1.df)
> correlation.numeric2na1.df
```

Figure 9 - Correlation of Cleansed Data

5.2.2.1 Class Distribution

The split of data across the three resolution categories can be seen below in fig 10; '0' represents blank values, '1' represents Network, '2' represents Infrastructure and '3' Application.

```
> #Summarise class distribution of Resolution Category
> percentage <- prop.table (table(numeric2na1.df$Reso$
> cbind(freq=table(numeric2na1.df$Resolution.Product.$
   freq percentage
0    99   0.118044
1 13480  16.073068
2  9864  11.761479
3 60424  72.047408
```

Figure 10 - Summary of Class Distribution of Resolution Category

We see that the highest proportion of incidents comes from the Application category (72%), followed by the Network category (16%), then the Infrastructure category (11.8%), and finally those uncategorised (0.1%).

When analysing the split of proactive (planned) and reactive (unplanned) events, fig 11 shows us that 40.6% of events are planned, and 59.4% are unplanned.

```
> #Summarise class distribution of Proactive and Reac$
> percentage <- prop.table (table(numeric2na1.df$Proa$
> cbind(freq=table(numeric2na1.df$Proactive1.Reactive$
   freq percentage
1 34049   40.59881
2 49818   59.40119
```

Figure 11 - Class Distribution of Planned/Unplanned Incidents

After the initial exploratory analysis, R visuals were utilised to gain a more in-depth overview of the data and develop the exploratory analysis in a more visual way, presenting the data parameters and spread. Visuals were produced using R code to build a picture of the data and allow for an understanding of the data dimensions. R visuals were limited to the early analysis due to the visual output produced in R, compared to that of Microsoft Power BI. R can give a very informative view of the data, and can be powerful in terms of analytics however is not as sleek and malleable in terms of interactivity and data drill-down when compared to Microsoft Power BI. Microsoft Power BI was chosen as the desired data visualisation software due to its user-friendly interface and interactive abilities to manipulate and visualise data in different ways. It was therefore chosen for this analysis due to the client company utilising the tool going forwards, so it allowed them to have first-hand access to the capabilities of the tool. Various visuals were chosen to present data in the best possible way, as discussed later in this section. Gartner (2018) and G2 (2018) both place Microsoft Power BI in their leaders quadrant when comparing data visualisation/business insight technologies.

5.2.2.2 Correlogram

Correlograms were utilised to visualise the correlations between attributes; boxes shaded in pink have a negative correlation, and boxes shaded in blue have a positive correlation. The stronger the pigment of the colour, the stronger the correlation is between the variables. The first correlogram completed contained all of the variables present after the initial data cleanse (fig 12):

```
> corrgram(numeric2na1.df, order=NULL, panel=panel.shade, text.panel=panel.txt,
+            main="Correlogram")
```

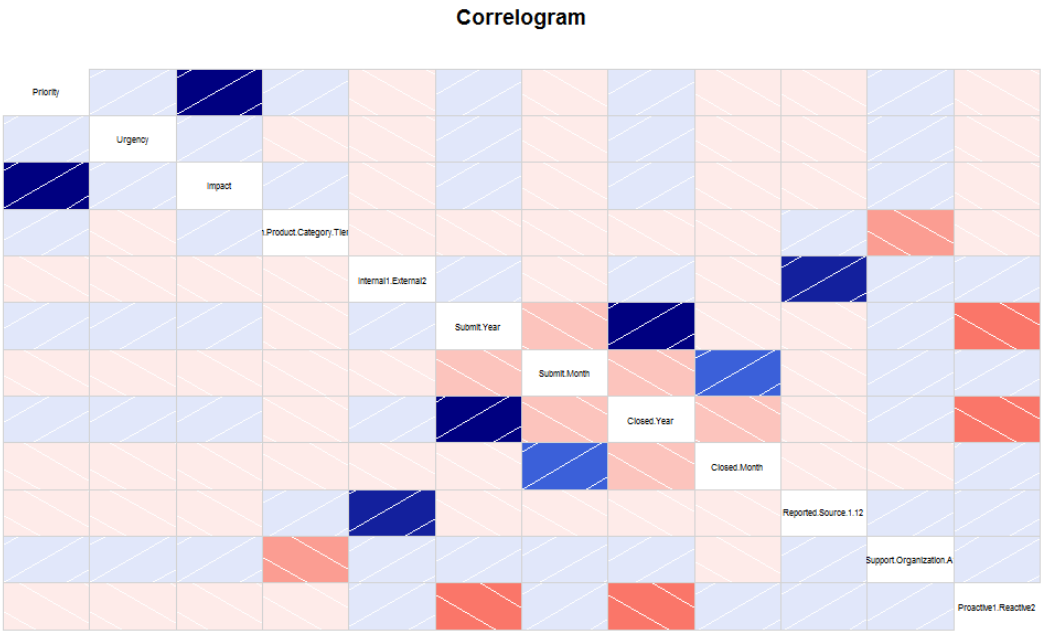


Figure 12 - Detailed Correlogram

After an initial model was build containing all of the attributes, a more concise model was built containing only those attributes that held a positive correlation. This can be seen below in fig 13:

```
> #More concise dataframe
> df3 <- numeric2na1.df[,c("Closed.Month", "Reported.Source.1.12", "Resolution.Product.Category.T$
> corrgram(df3, order=NULL, panel=panel.shade, text.panel=panel.txt,
+            main="Correlogram")
```

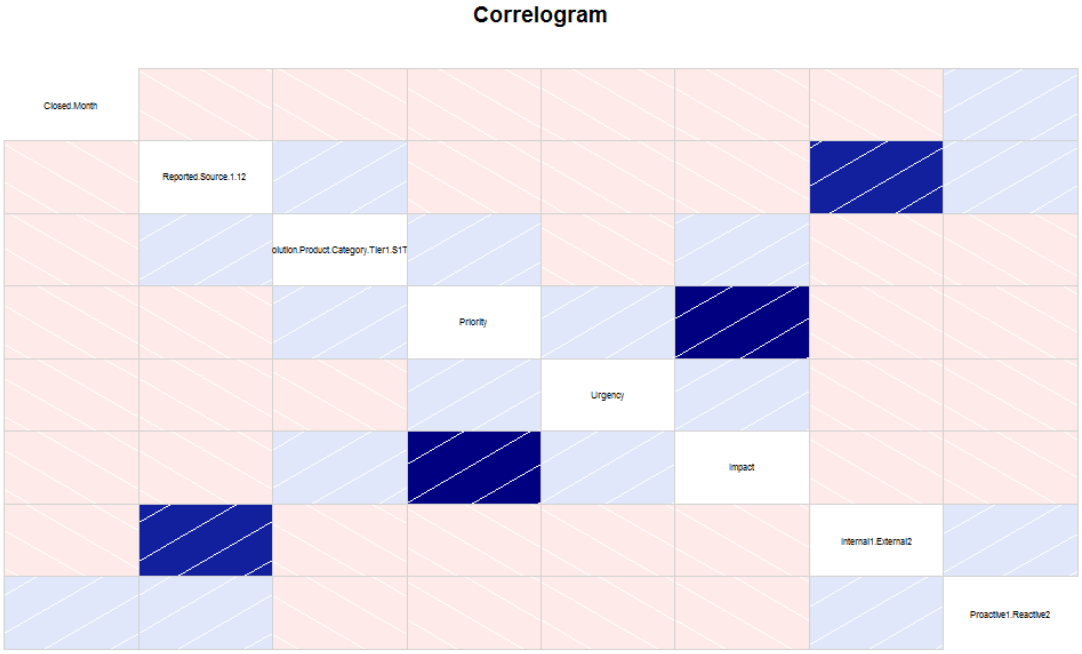


Figure 13 - Concise Correlogram

The above visual gives a clear, concise view of the strongest correlations present within the data; we see Impact:Priority, and Internal/External:Reported Source holding the strongest correlations, followed by Urgency:Priority, Impact :Urgency, Impact:Resolution Product Category, Proactive/Reactive:Reported Source, Proactive/Reactive:Closed Month, Priority:Resolution Product Category, and Resolution Product category:Reported source.

5.2.3 Heatmaps

A number of heatmaps were created to visualise the correlations between specific variables in a different way. This process was repeated for 'Impact, Urgency and Priority'; 'Closed Month, Resolution Category and Impact'; 'Closed Month, Reported Source and Impact'; and finally 'Proactive/Reactive, Reported Source and Impact'. The code for all of the heatmaps can be found in Appendix 4.

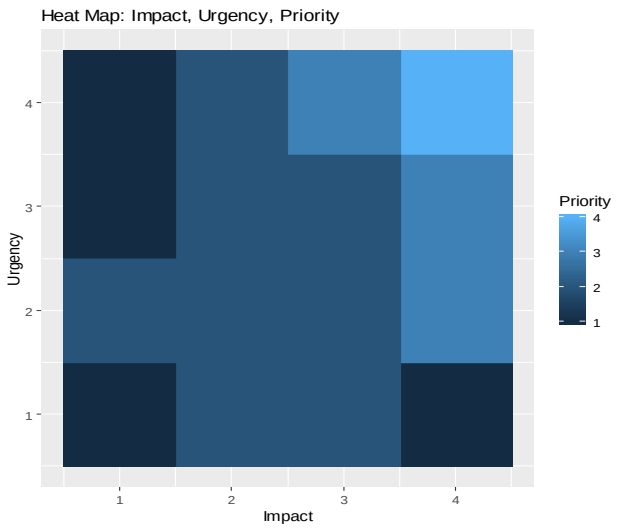


Figure 14 - Heatmap: Impact, Urgency, Priority

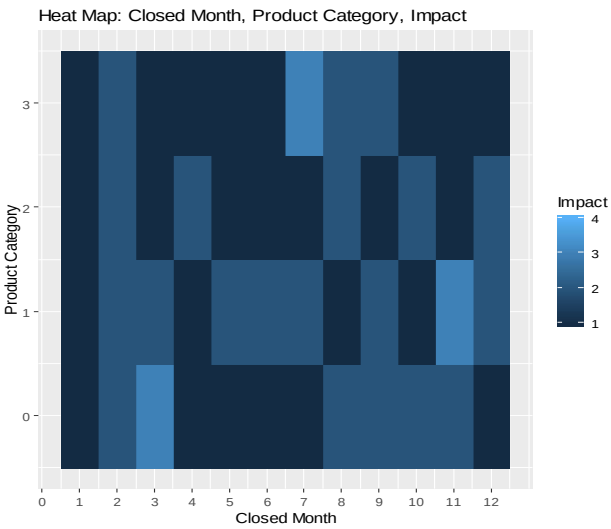


Figure 15 - Heatmap: Closed Month, Product Category, Impact

The first heat map (fig 14) is designed to show the priority level based on the impact and urgency levels. The paler blue, the higher the priority level. This correlates, as we see an impact and urgency of 4 giving a priority of 4. Interestingly, something with an impact level of 4 but urgency of 1, and something with an impact of 1 and urgency of 4 both return a priority level of 1.

The second heatmap (fig 15) looks at the level of impact based on the month the incident was closed, and the associated product category. We see uncategorised March events, Application category July events and Network category November events having the highest impact ('0' represents blank values, '1' represents Network, '2' represents Infrastructure and '3' Application).

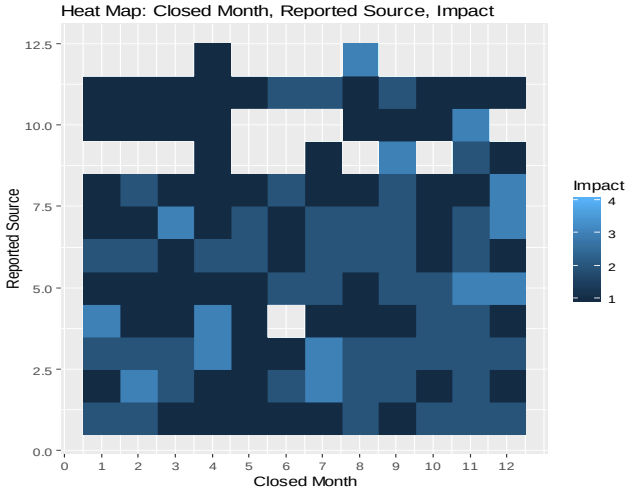


Figure 16 – Heatmap: Closed Month, Reported Source, Impact

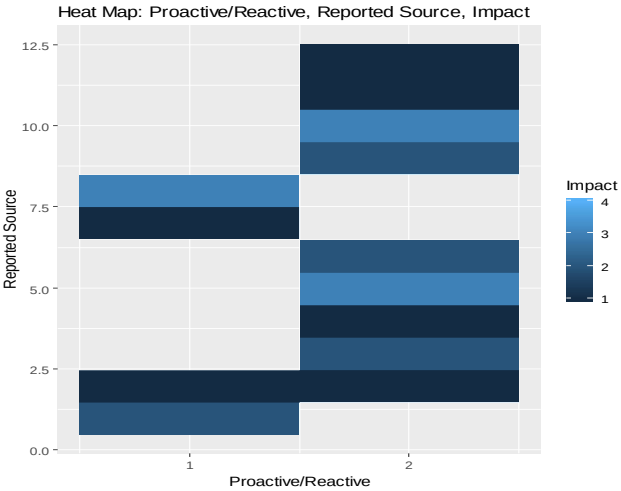


Figure 17 – Heatmap: Proactive/Reactive, Reported Source, Impact

Figure 16 shows the correlation between closed month and reported source. The blanks in the heatmap relate to instances where in a particular month, no incidents were reported by a specific source. We see January: 'External Escalation', February: 'Direct Input', March: 'Self Service', April: 'Email' and 'External

Escalation', July: 'Direct Input' and 'Email', September: 'Voice Mail', November: 'Other', and December: 'Other', 'Self Service' and 'Systems Management' having the highest impact levels (The list of correlated Sources to the numerical values can be found in Appendix 5).

Figure 17 shows the Impact level based on whether the incident was proactive or reactive, and what the reported source was. We see a lower range of sources for proactive events, compared to reactive. The highest impact source for proactive events is 'System Management', and the highest impact events for reactive incidents are 'Other' and 'Walk in'.

5.2.4 Histograms

Histograms were an effective means of identifying the split of the data. These were produced to understand the split and spread of each category. Multiple were produced to identify a number of aspects. The code for all of the histograms can be found in Appendix 6.

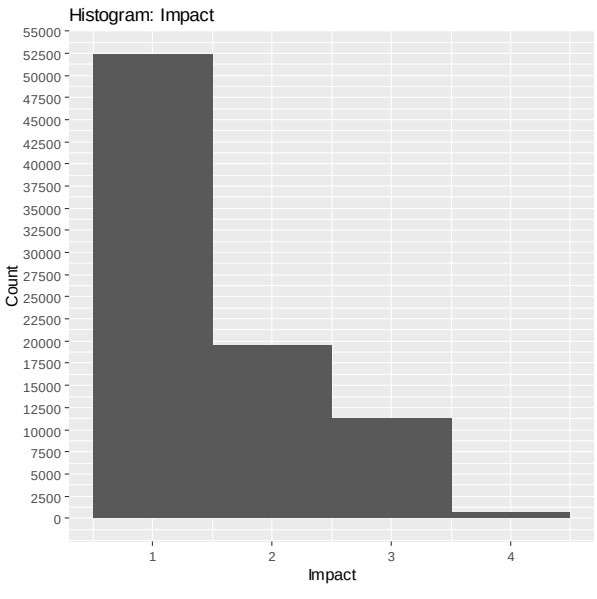


Figure 18 - Histogram: Impact

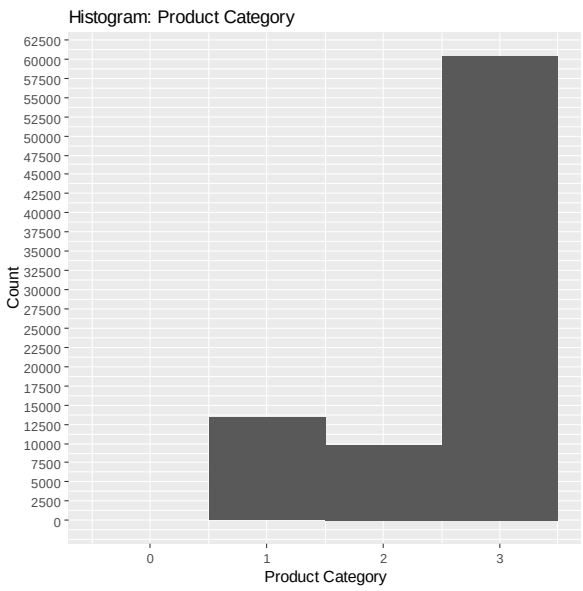


Figure 19 - Histogram: Product Category

Figure 19 highlights a significantly larger proportion of events falling into the lowest impact category, and following a linear pattern, with the count of instances reducing as the impact level increases.

Figure 20 provides a count of instances per product category; we see the largest proportion falling into the Application category, followed by the Network category, and finally the Infrastructure category.

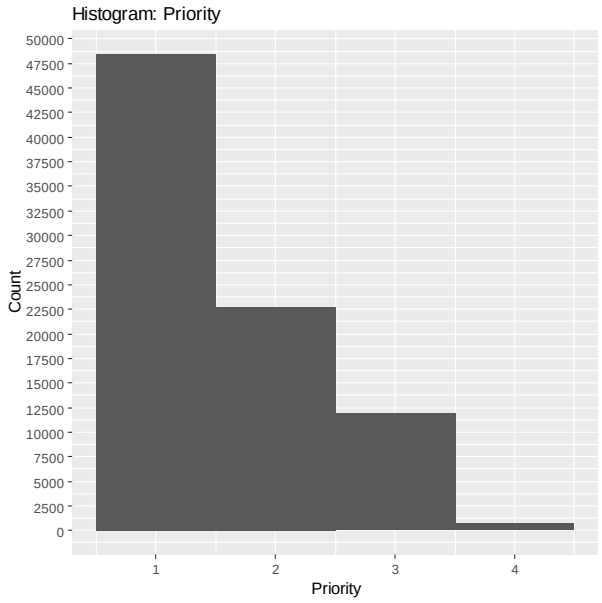


Figure 20 - Histogram: Priority

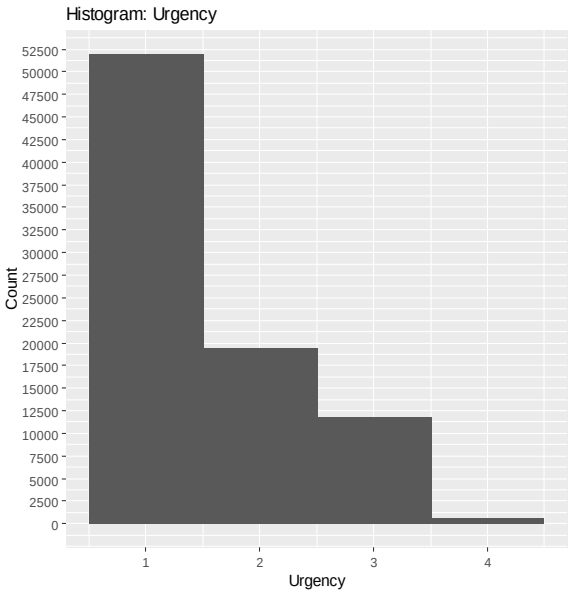


Figure 21 - Histogram: Urgency

The third and fourth histograms (Figures 20 and 21) follow the same linear pattern as figure 19; we see the count of instances reduce as the priority and the urgency reduce, respectively.

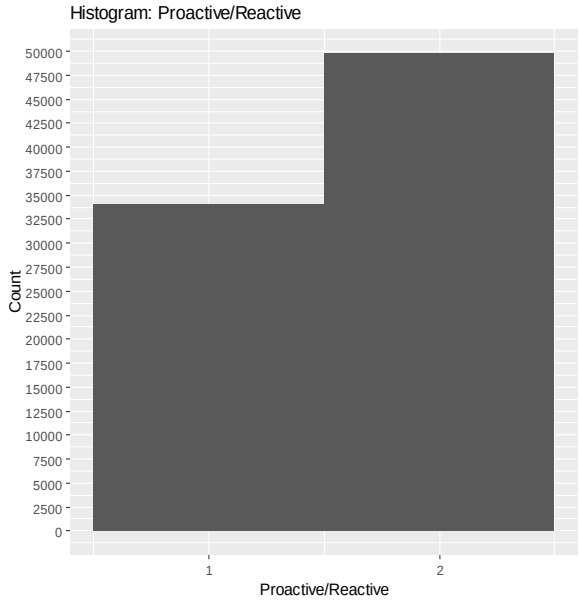


Figure 22 - Histogram: Proactive/Reactive

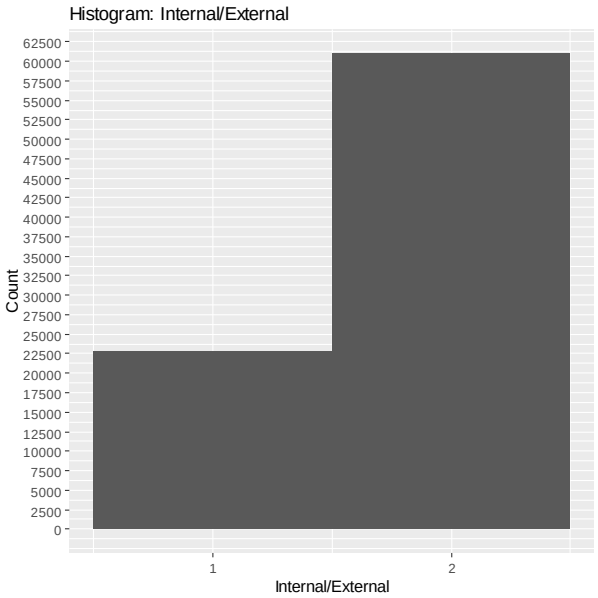


Figure 23 - Histogram: Internal/External

In terms of the split between proactive and reactive incidents (fig 22) (planned and unplanned events), we see that the there is approximately one third more reactive incidents reported than proactive. Figure 23 shows the split between internal and externally reported incidents; this shows us that a far larger proportion of events are reported externally.

5.3 Analysis: Microsoft Power BI

Microsoft Power BI was utilised to produce exploratory visualisations after the initial analysis in R; this included bar charts, funnel graphs, line graphs, pie charts, doughnut charts and data tables, as well as time-series graphs. This section highlights the key findings from the historical analysis conducted on data gathered from 2012-2018. A variety of comparisons were made, and high level analysis drilled down to give greater insights on findings.

The combinations of graphics were chosen to give a variety of visually appealing infographics that can be drilled down for further analytics or to view source data. Nika Aleksejeva (2015) states that a key risk of using the likes of pie charts is that, as humans, we fail to correctly identify the difference in sizes between segments; often we subconsciously form a bias towards a specific segment, perhaps due to the colours used. By including data labels on the graphics, this eliminates the risk of bias in reading results.

5.3.1 Time-Series Analysis

5.3.1.1 Count of Incident ID by Year

Figure 24 below shows the count of incidents that occurred over a 6-year period. We see that overall there is an increasing trend as time has passed, with a dip between 2015 and 2017. We see the number of incidents in 2018 being the greatest, approaching 20,000.

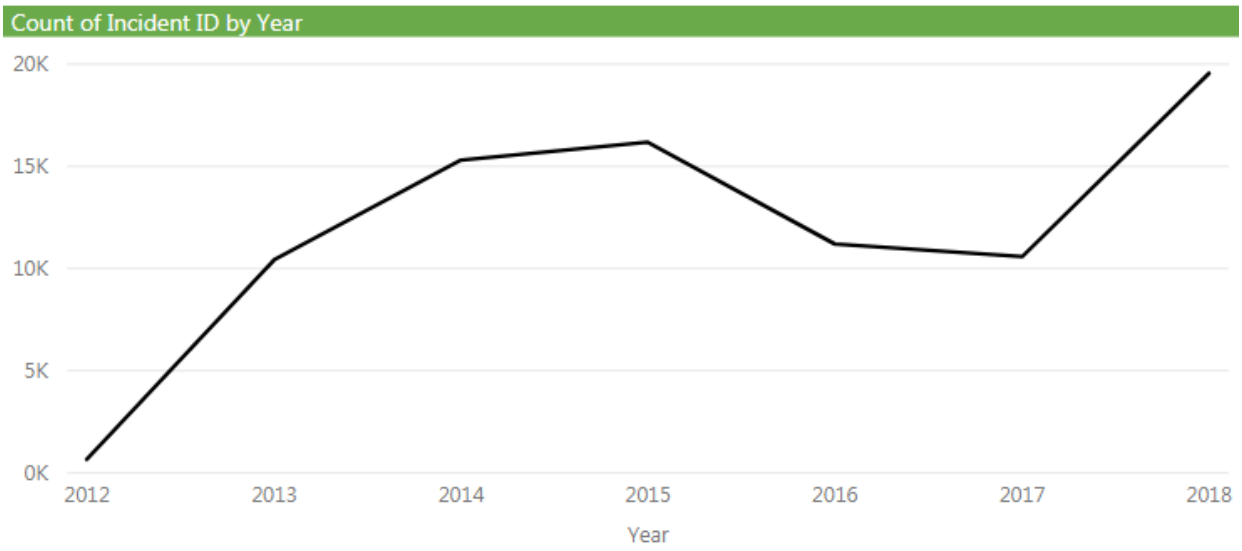


Figure 24 - Count of Incident ID by Year

5.3.1.2 Count of Incident ID by Year and Impact

Adding the 'impact' element (Figure 25 below) identifies the key reason for the increased level of incidents occurring – Minor/Localised events.

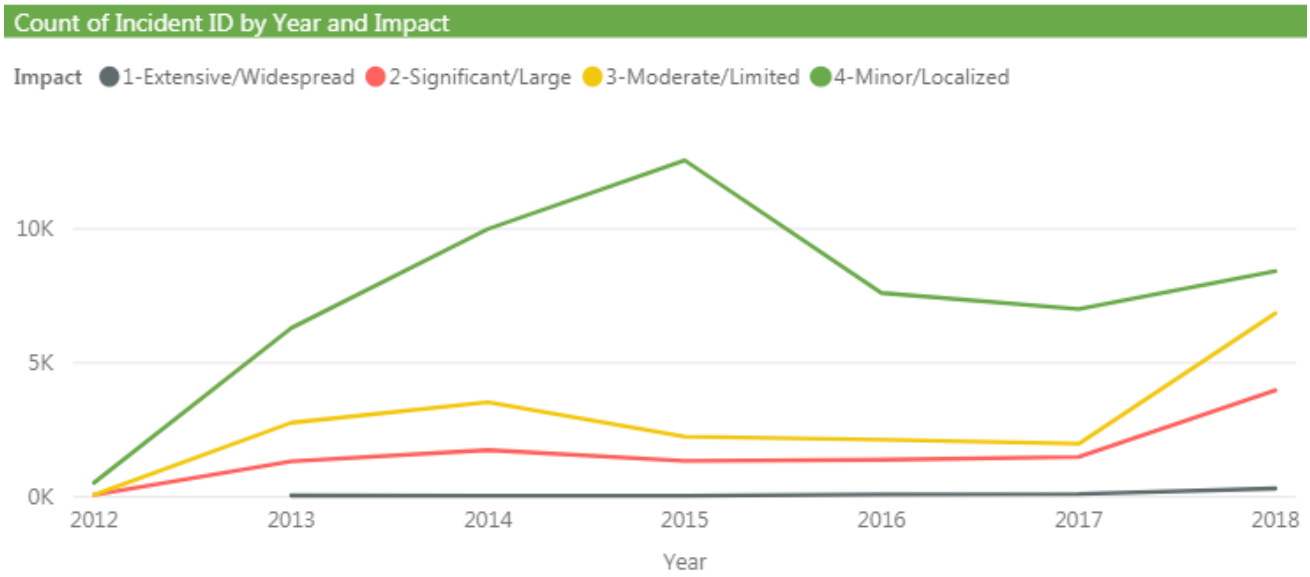


Figure 25 - Count of Incident ID by Year and Impact

In the above (Figure 25) we see 'self-service' incidents increased by 4303, from 80 the previous year; this indicates that a new system was most likely implemented that required a high volume of password resets. Drilling down further, it is confirmed that a CRM system was launched in 2015, which had the highest impact on SP Energy Retail business. Of the three categories – Applications, Infrastructure and Network – 'Applications' saw the greatest increase in incidents. Looking at the raw data, it is clear that SAP Applications have the greatest increase – linking directly to the launch of the CRM system.

5.3.1.3 Count of Incident ID by Year and Product Categorisation Tier 1

In figure 26 we see the number of incidents come to a peak at the end of 2013/beginning of 2014, before remaining relatively consistent and beginning a downward trend from 2015 until 2018. When we drill down into the 2014 data, we see that SP Energy Retail have an increase in incidents of 76%, with an increase of 2940 (153%) incidents relating to CRM Applications. There is an overall 981% increase in SAP related incidents indicates from 2013 to 2014, with similar numbers in 2015.

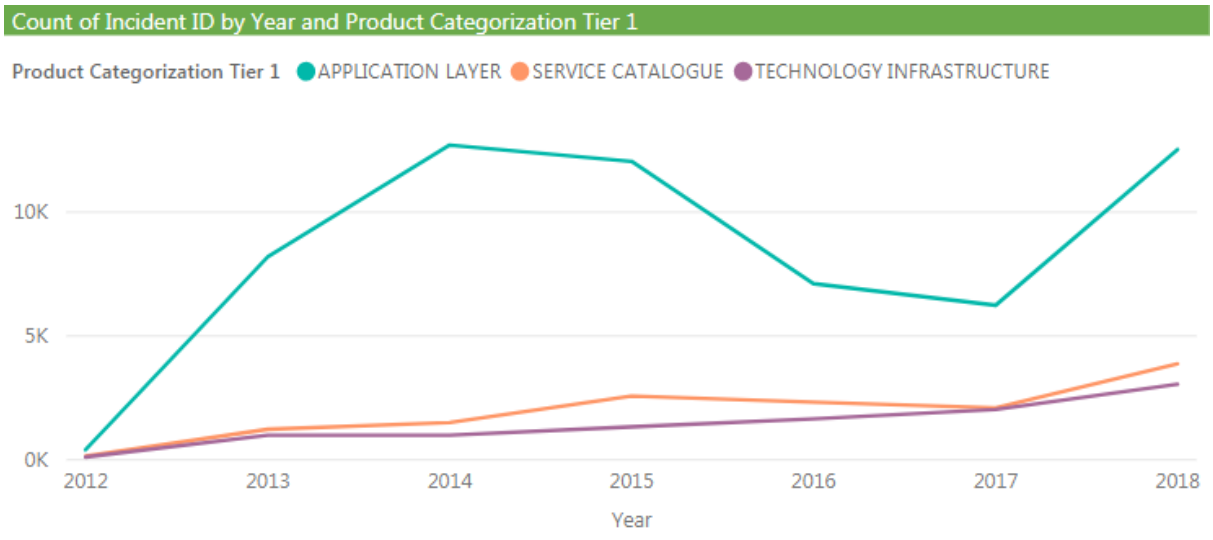


Figure 26 - Count of Incident ID by Year and Product Categorisation Tier 1

When we drill down into the 2018 data, we see a significant increase in incidents reported from SP Energy Networks (1,225% increase from 2017). The majority of these incidents come from self-service activity, from the 'Other Application' category. Here, a key issue is highlighted – when incidents are logged, the person logging the incident has the option to leave elements of the report blank, or choose 'other' rather than a specific categorisation. Where this may be necessary in some unique situations, it is clear that this is used as a means of avoiding searching for the correct category. When the data is drilled down here, the greatest detail that can be gleaned is that the incidents fall in the 'Application' category – there is no detail as to which application is the reason for this increase and therefore no further work can be done to rectify the specific issue. This will be highlighted as a key recommendation.

5.3.1.4 Count of Incident by Month and Year

In figure 27 below, the count of incident ID is split by month on the x-axis, and each year differentiated by a different colour plotted on the graph, with count on the y-axis. We see the level of incidents being logged increase year on year, with 2016 and 2017 being exceptions – dropping below the 2014 level. We see 2012 being particularly low due to the data gathering beginning here, and see fault levels being most similar across all years – with the smallest range –in the month of September. The most significant peak that can be identified is in August of 2018, although 2018 as a whole is notably higher than previous years.

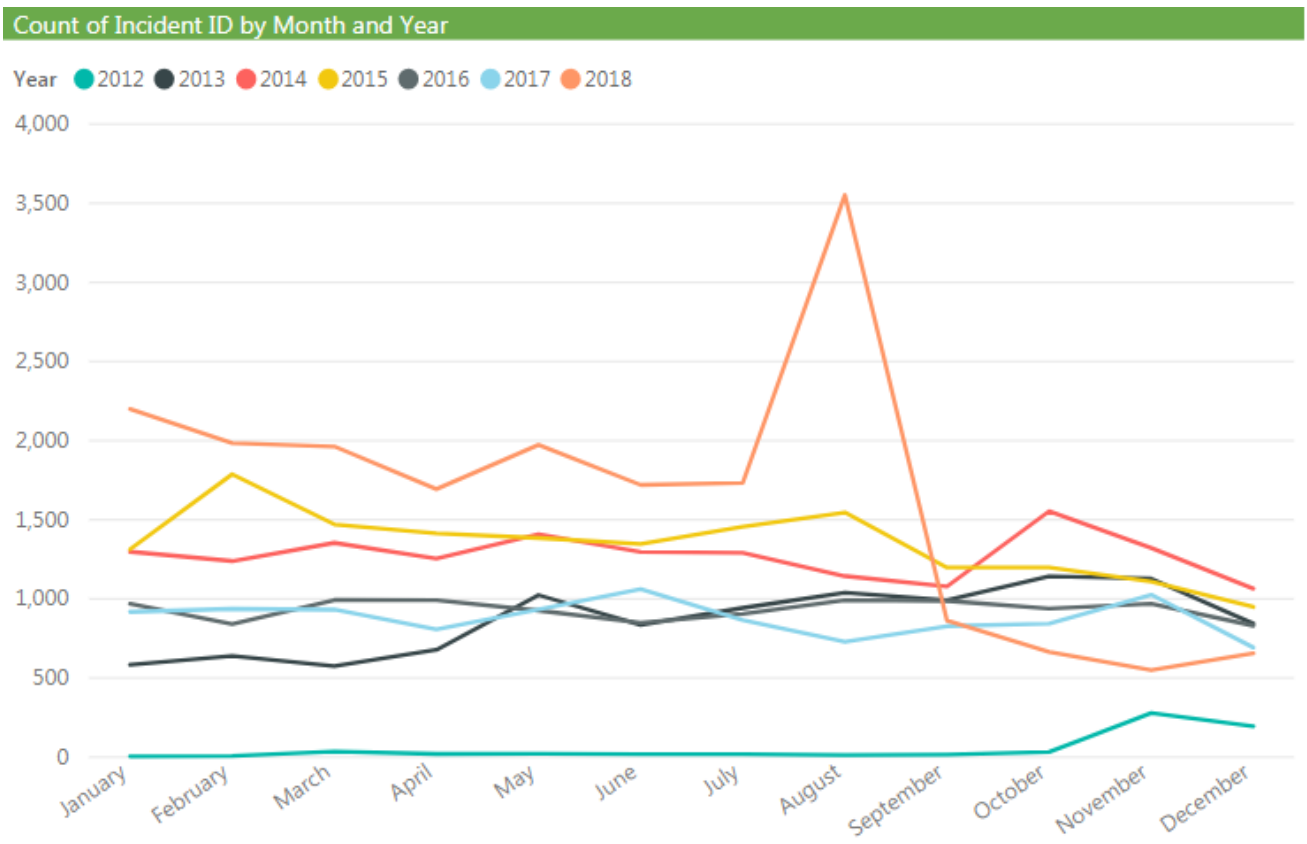


Figure 27 - Count of Incident ID by Month and Year

The drill down can be seen in figure 28 below. We see the significant increase of 1816 incidents being split across all three of the resolution product categorisations, with just over half of these being present in Application layer. Approximately half of the Application incidents come from batch process, with around a quarter coming from SAP. Of these incidents, these are split almost equally between employees in the UK and in Spain.

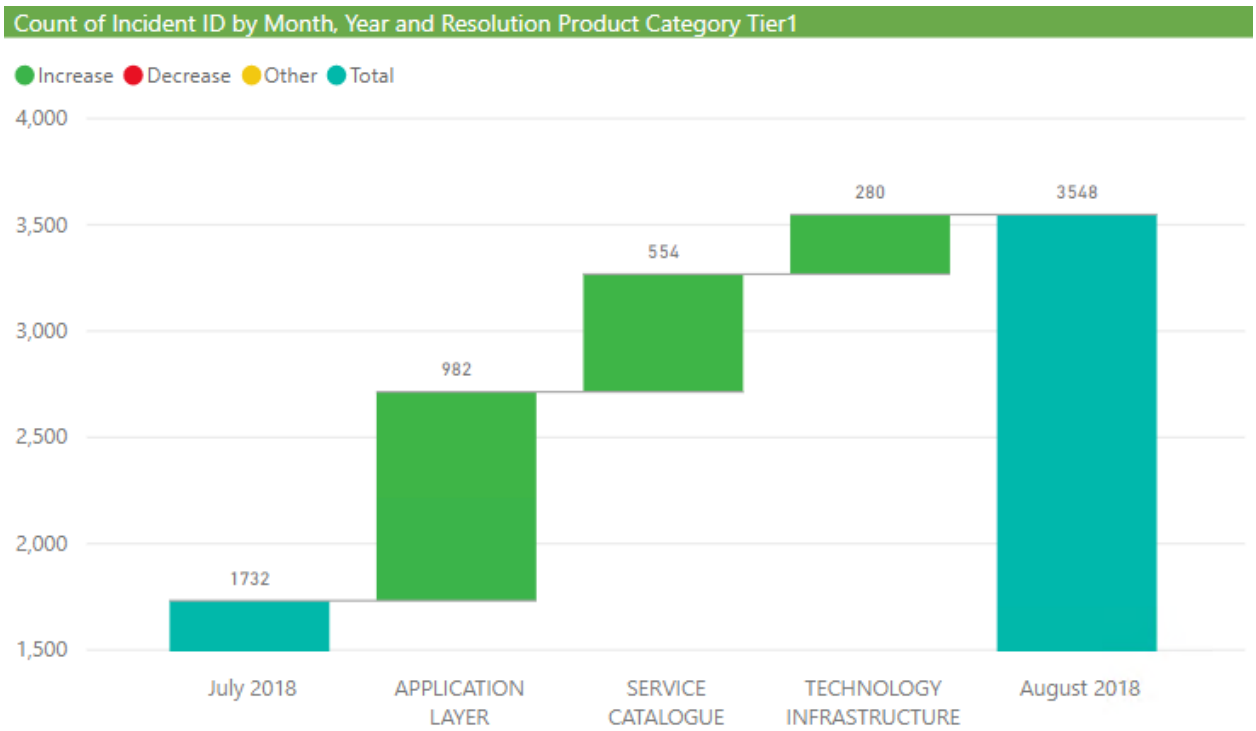


Figure 28 - Drill down: Incident ID by Month, Year and Resolution Product Categorisation Tier 1

5.3.1.5 Count of Incident ID by Month and Impact

When we divide the incidents by impact level, there is a clear split (Fig29). We see a chronological pattern in terms of number of incidents based on impact: the least number of incidents come from the extensive impact level, and the most incidents come from the minor impact level. The difference between the first three categories – extensive, significant and moderate – is relatively equal with a far larger increase in incident count present between moderate and minor. This is reassuring, as this shows us that the highest proportion of incidents have a very low impact on the business. This can be identified as a key opportunity; given that a high quantity of low impact incidents are raised, this indicates that there is room for improvement. Due to the significantly higher level of incidents, there is most likely going to be a higher level of repeated faults that could be addressed. The incidents raised are less likely to be unique, due to the volume.

Across all impact levels, we see a peak occurring in August; we see a dramatic increase in the number of self-service incidents, the majority of which were reported from the Energy Networks business. A large proportion of the incidents come from Applications (predominantly SAP).

Following the same trend as the overall incident count, we see the gradient of each August peak flattening as the impact level increases.

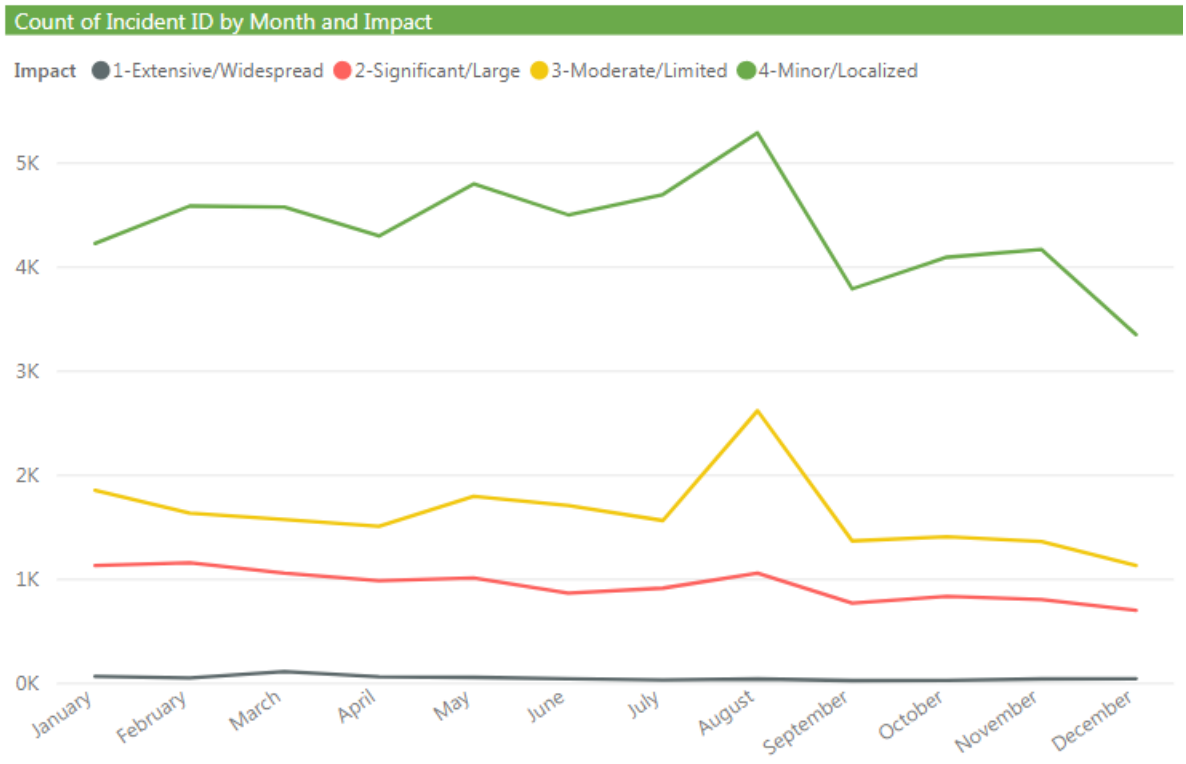


Figure 29 - Count of Incident ID by Month and Impact

5.3.2 Further Data Visualisation

5.3.2.1 Count of Incident ID by Impact

A high level count of incident ID by impact is presented in figure 30 within a pie chart (Data can be found in Appendix7). This effectively illustrates how the incidents are split, emphasising the amount of minor impact incidents that are reported (62.66%). The pie chart also emphasises how few incidents reported fall into the extensive impact area (0.75%).

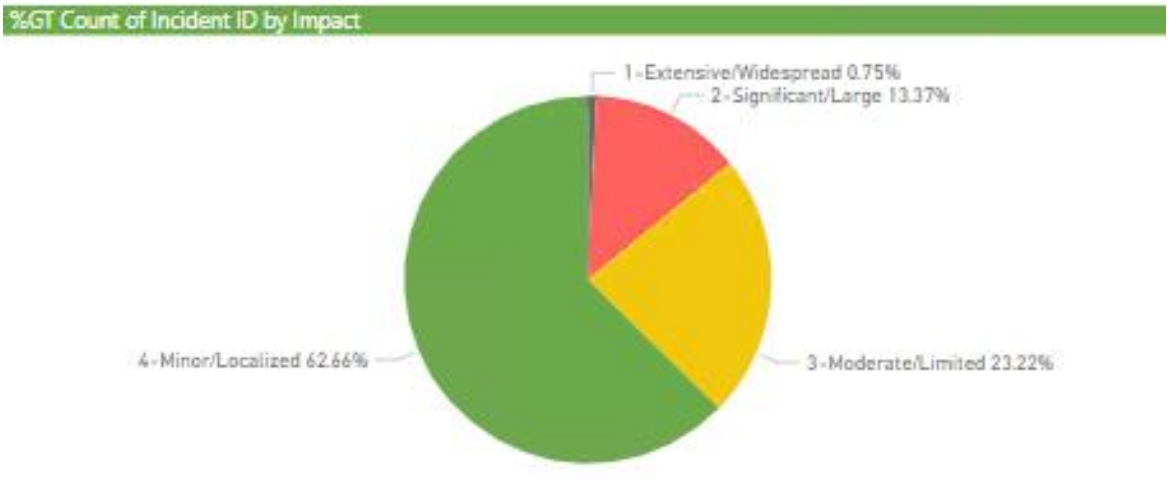


Figure 30 - Pie Chart: Count of Incident ID by Impact

5.3.2.2 Count of Incident ID by Product Categorisation Tier 1

Figure 31 shows us that if we count all incidents 2012-2018 and split them into the 3 high-level categorisations, we see that the majority of incidents fall into the 'Application Layer' category (71.25%).

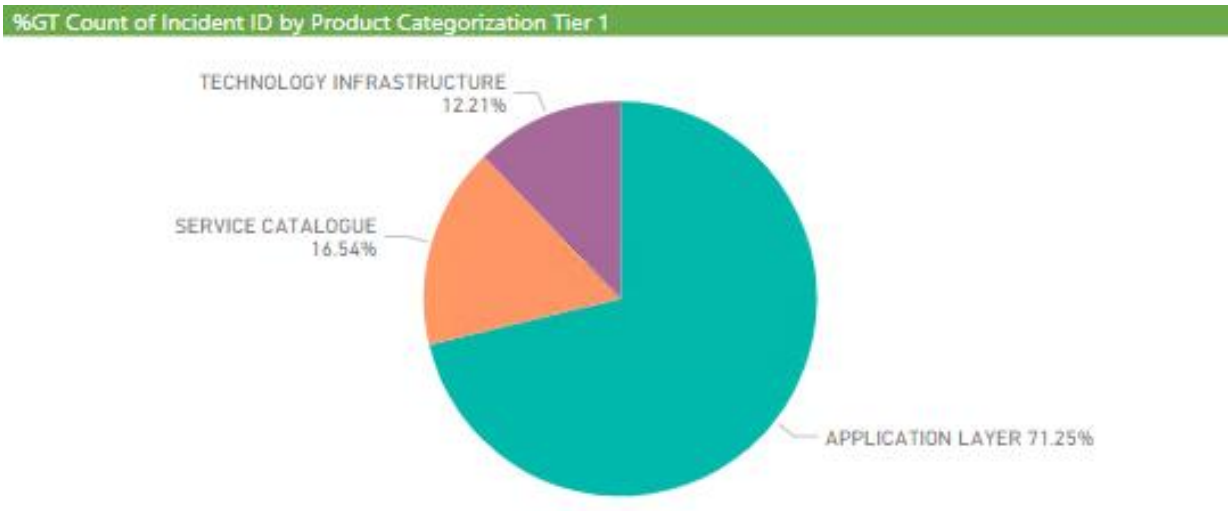


Figure 31 - Pie Chart: Count of Incident ID by Product Categorisation Tier 1

5.3.2.3 Key Application Incidents

When we drill down further, we can see greater detail of the key application incidents.

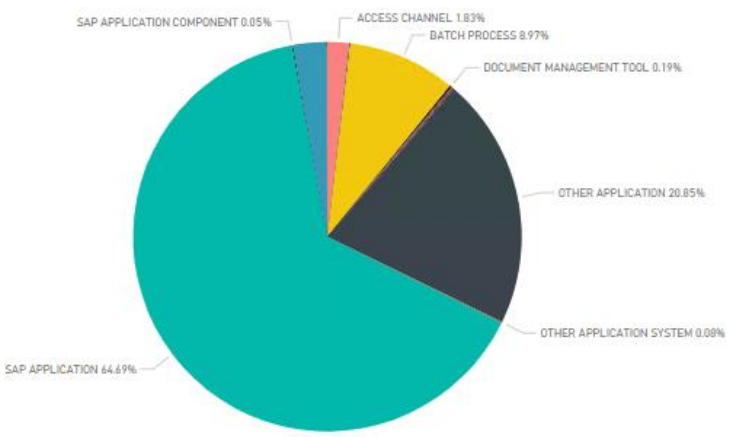


Figure 32 - Pie Chart: Drill down to Application Incidents

We see the greatest number of application faults coming from SAP Applications, followed by 'Other Applications' and Batch process. As previously mentioned, the ability to select 'Other' is a limiting analytical factor as it provides an opportunity not to effectively categorise incidents and encourages this. This will be highlighted as a key issue.

The data in figure 33 shows the count of all incident categories.

5.3.2.4 Count of Incident ID by Product Category Tier 3

The unfiltered tier 3 pie chart in fig 33 shows all of the incidents grouped to give more detailed insights. We see that approximately half of the incidents are SAP related, with around 17% being identified as 'other'. The full data for this pie chart can be found in Appendix 10.

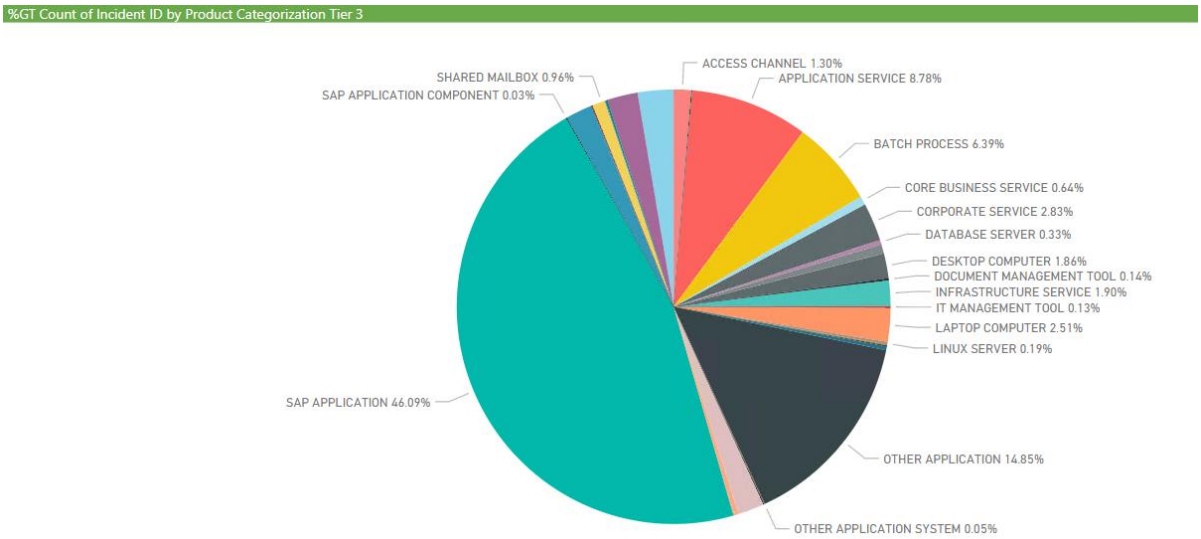


Figure 33 - Pie Chart: Incident ID by Product Categorisation Tier 3

5.3.2.5 Count of Incident ID per Country

A definitive feature of Microsoft power BI is its ability to integrate with mapping tools to present data. In figure 34, the data is presented on a good-bad (green-red) scale to show the count of incident ID per country. We can see from the map that the UK has significantly more incidents than the rest of the countries where ScottishPower operate. This is understandable given the number of ScottishPower employees in the UK. Following the UK, Spain have the next most significant number of incidents, followed by India. The United States of America, Germany and France, Korea and the Netherlands all have associated incident reports, however are so few over a 6 year period that these are more likely to be anomalies than common events. The data can be found in Appendix 11.

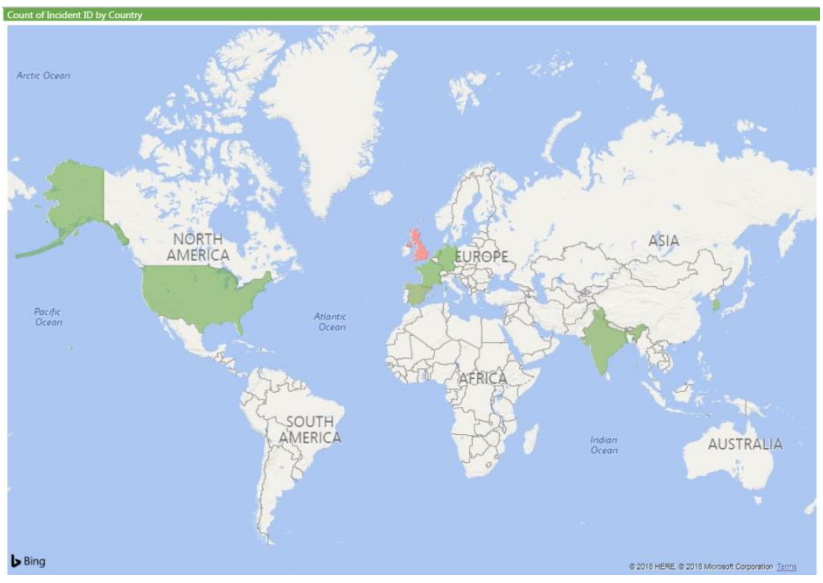


Figure 34 - Map of Incident Reported Geographical Source

6.0 Analysis and Discussion of Findings

6.1 Introduction

This section contains the analysis conducted, split into the three sections mentioned previously: Historical analysis, 'Live dashboard' analysis and Forecasting model. Each of the three sub-sections investigates the data limitations, gathering, and analysis techniques specific to that section.

6.2 Historical Analysis

This section will focus on the historical quantitative data analysis conducted to identify trends in the historical incident reports.

6.2.1 Data Limitations

Key limitations were present when conducting analysis. Firstly, all of the data was processed remotely on a daily basis, and compiled into monthly 'data pull' reports that were generated in Microsoft Excel. Due to the scale of the data being gathered, Excel was not an option for time series analysis due to the processing power required. No alternative analysis tool had been adopted by the company when the project began. Secondly, due to the company being in the early stages of its big data strategy roadmap, the hardware available was not capable of processing large amounts of data without crashing on a regular basis. To try and combat the above issues, I created a business case to trial using Microsoft Power BI – a market-leading data analytics and visualisation tool – to process the data. This software was chosen both to its stance in the Data Analytics field, with the likes of Gartner placing it in their 'leader' quadrant, comparing top Data Visualisation and Business Insight technologies (Gartner, 2018). Unfortunately, due to security risks, data could not be stored remotely (on cloud servers), so the hardware issue remained. Due to the high security around IT infrastructure, various access requests had to be requested throughout the duration of the project, resulting in the most significant limitation being time. Many of the requests required several levels of authentication, and often small errors occurred during the validation process which significantly slowed (and in some cases halted) progression.

6.2.2 Data Gathering

Data gathering was relatively straightforward, with incidents being logged live. The key challenge related to the extraction of said data, and version control; due to the server (ITSM) where the report is generated being held in Spain (where ScottishPower's parent company reside), security around the server is extremely high and therefore data extracts have to be used by the company to protect data privacy. This added an element of risk, due to being unable to conduct data quality assurance on the spreadsheets. On top of this, each extract from ITSM was saved individually; given that several years of data was required to capture ongoing trends, this made the data gathering and acquisition process relatively time consuming.

6.2.3 Analysis: Microsoft Power BI

The chosen technology was Microsoft Power BI, due to its adaptability and versatile data preparation options. As mentioned previously, Microsoft Power BI is one of the market leaders in data visualisation software, and a large focus of the software is to create informative dashboards for users to see various visuals on one page to be as informative and bold as possible. ScottishPower wanted to move away from the standard excel reporting format and try something new.

A key requirement of the project was to build something that would be re-usable for the future and Microsoft Power BI opened that door: being able to build a dashboard/reporting pack which would automatically update and provide active insights on a daily basis.

As previously mentioned, ScottishPower are in the process of building their Data Analytics Roadmap. For this reason, ScottishPower allowed a Proof of Concept to be trialled for Microsoft Power BI; this was the preferred option due to being able to purchase the software as an add-on to their current Microsoft Professional package, and to present those working at ScottishPower with a relatable interface with a coding language (DAX) which has basic functionality that aligns closely with Excel and VBA. The alternative software was Tableau, however this sat at a much higher price point and was a less familiar interface for users.

Using a tool which could be integrated with the current licencing for the company future-proofed the project and the analysis, as it ensured that everyone who would require access to the data/reports could gain that without any unreasonable cost for the company. A key barrier in many companies relates to inconsistent systems being implemented and used by different departments; by unifying this process and choosing a software that employees would feel comfortable with makes data analysis far easier and ensures greater data integrity.

Gartner describes Microsoft Power BI as a product that encompasses various elements of analytics, from preparation, to discovery and interactive dashboards (Gartner, 2018). They mention the cloud functionality – Azure cloud – which aligns directly with the cloud preference of ScottishPower. Microsoft Power BI can also be used on-premise at a much lower cost; this is what was used for the proof of concept, along with an in-house report server.

After choosing Microsoft Power BI as an appropriate software, a further data cleanse took place and new columns added to the aggregated dataset to allow quicker uploads, and only relevant information to be included in the active data source for the report.

As mentioned in the ‘Problem Structuring’ chapter, the client requested some key information from the analysis being conducted; they requested the top ten incident count and resolution times for ScottishPower, split by Impact (Low, Medium, High and Critical), and by Service Category (Infrastructure, Application, and Network), and to identify recurring themes.

6.2.3.1 Top ten incidents categories for ScottishPower

Figure 36 - Top ten incident Categories for SP

From fig 36 above, we can see that 'batch process' is the top occurring incident for ScottishPower. Within the top ten, we see three servers appear: Wintel (2nd), Unix (3rd) and Linux (7th) - this highlights that there are ongoing, frequent problems with not just one but several of the servers used within the company. SAP is the 4th highest cause of incidents within the company – this was something that the client highlighted in advance of analysis and most likely is the contributing factor to 'Application Service' (6th) and 'Other Support Service' (10th) which will both contain uncategorised SAP incidents (as per client's advice). The remaining categories are 'Shared Mailbox' (5th), 'Switch' (7th) and 'IT Management Tool' (8th).

6.2.3.2 Top Incidents per impact

The following funnel graphs show the top 10 incident counts based on impact level (Low, Medium, High and Critical).

Low Impact

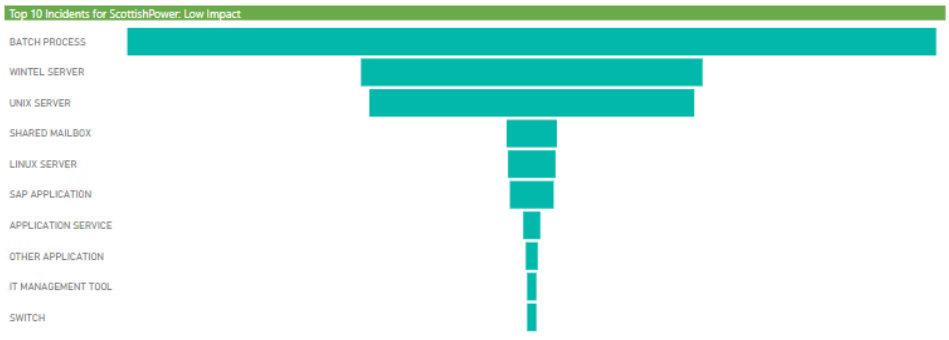


Figure 37 - Incidents per impact: Low

Due to low impact events dominating the incident occurrence count, it is not surprising that the categories that appear in the overall top 10 match those appearing in fig 37 above.

Medium Impact

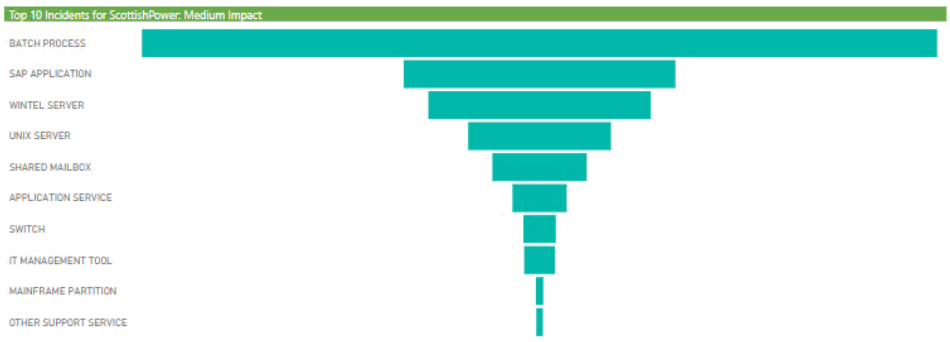


Figure 38 - Incidents per impact: Medium

As the impact decreases (and the count of incident events), we see the top 10 incident categories changing slightly; in fig 38 above, we see the Linux server disappear from the top 10, and the introduction of 'Mainframe partition'.

High Impact

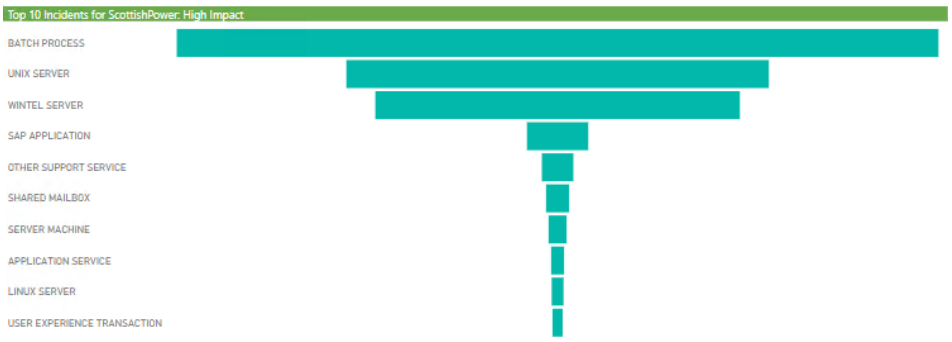


Figure 39- Incidents per impact: High

Following on from Medium impact, we see the introduction of 'Server Machine' and 'User Experience Transaction' appear in the top 10, and the removal of 'Mainframe Partition' and 'Switch'.

Critical Impact

Figure 40 - Incidents per impact: Critical

There are very few critical impact events that occur, in comparison with the other impact categories. For this reason, there is a reduced list of service categories for 'Critical Impact'; we see all three servers present (Wintel, Linux and Unix), as well as Batch Process, SAP and Shared Mailbox. All of the critical impact incidents align with those reported as the top 10 for ScottishPower as a company.

6.2.3.3 Incidents per service category (Top Ten)

The client requested that a 'top 10' list rank could be created based on the service category; Infrastructure, Application and Network. Due the nature of each of these categories, we see varying incident types occurring in each.

Infrastructure Category

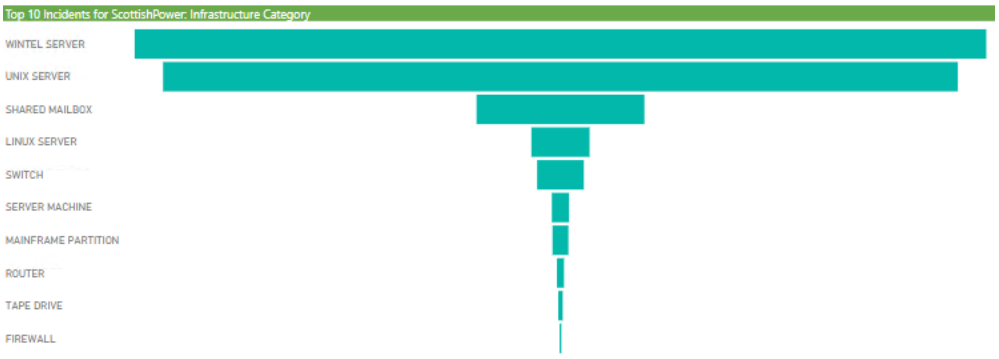


Figure 41- Incidents per service category: Infrastructure

For Infrastructure, we see Wintel and Unix servers dominating the incident occurrences (1st and 2nd). Shared mailbox is in 3rd place and Linux server and Switch are present in 4th and 5th place, but with significantly fewer associated incident reports. Following this, we see Server machine, Mainframe partition, Router, Tape drive and Firewall, however in comparison to the top 2 causes of incident reports, these figures are nominal.

Application Category

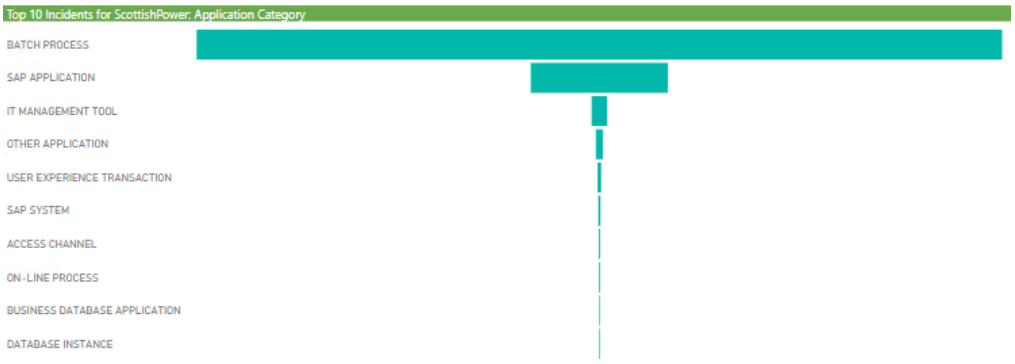


Figure 42- Incidents per service category: Application

For the Application category, 'Batch Process' dominates the incident occurrence, followed by SAP application. Thereafter, the level of incidents are extremely small by comparison: 'IT Management tool' (3rd), and 'Other application' (4th) being the most significant to follow.

Network Category

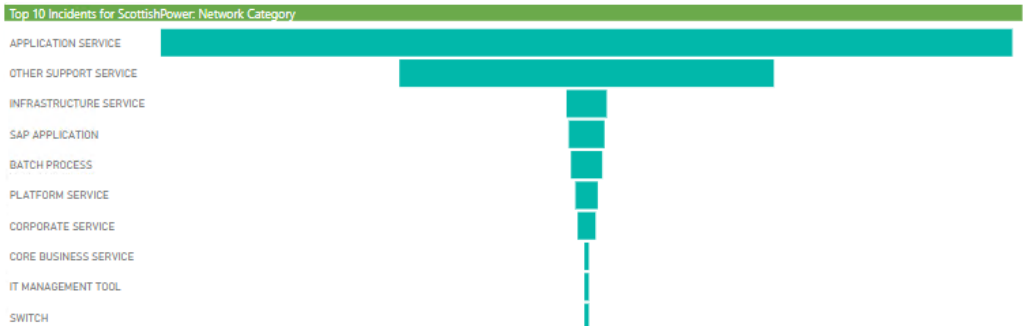


Figure 43- Incidents per service category: Network

For the Network category, there is a slightly better spread of data; we see 'Application service' (1st) and 'Other Support Service' (2nd) dominate the incident count, followed by relatively even counts between 'Infrastructure Service' (3rd), 'SAP Application' (4th) and 'Batch Process' (5th). Roughly even splits can then be seen between 'Platform Service' (6th) and 'Corporate Service' (7th). Finally, roughly even splits can be seen between the final four incident categories: 'Core Business Service', 'IT Management Tool' and 'Switch' (8th-10th respectively).

6.2.3.4 Resolution time (average) per Impact

The following funnel graphs present the top 10 resolution times for incidents, based on each impact level (Low, Medium, High, and Critical).

Low Impact

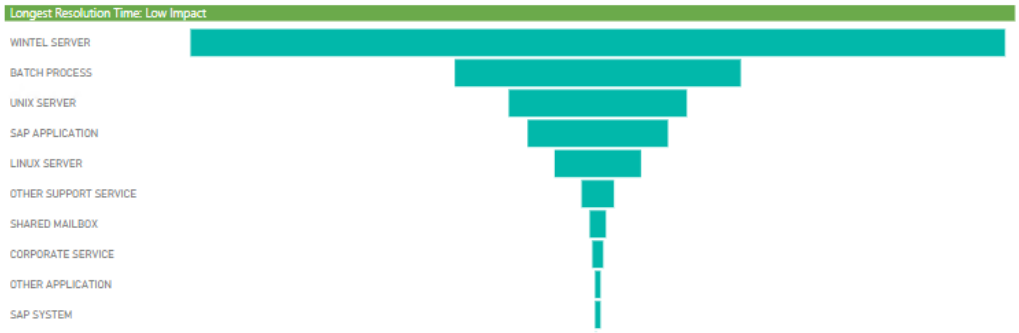


Figure 44- Average resolution time per Impact: Low

For low impact events, we see Wintel server issues taking the longest to resolve, and then a stepped decrease in resolution times for the categories that follow; Batch Process coming 2nd, Unix server 3rd, followed by SAP application and Linux Server (4th and 5th), then Other Support Service, Shared Mailbox, Corporate Service, Other Application and SAP System (6th-10th respectively).

Medium Impact

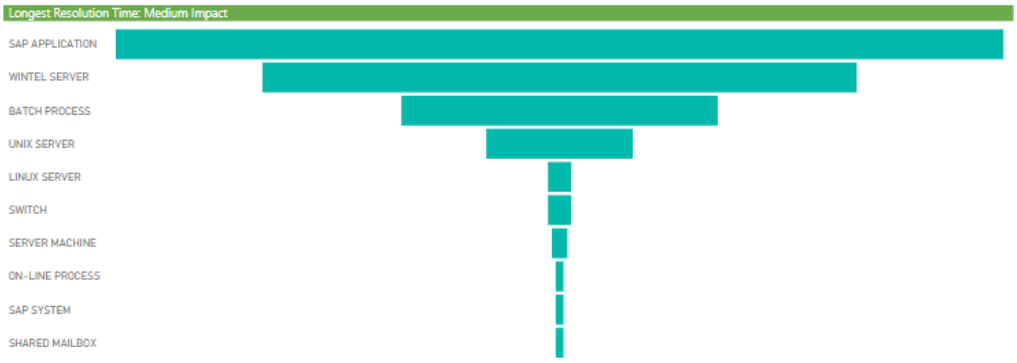


Figure 45- Average resolution time per Impact: Medium

For medium impact, we see SAP Applications and Wintel Server issues being the longest to resolve (1st and 2nd), followed by Batch Process and Unix Servers (3rd and 4th) which all step down by even amounts. Following this, we see Linux Server and Switch taking approximately the same amount of time to resolve (5th and 6th), with Server Machine taking slightly less time. The final 3 (8th-10th): On-line process, SAP System and Shared Mailbox take roughly the same amount of time to resolve; a fraction of the time taken compared to the top 4.

High Impact

Figure 46- Average resolution time per Impact: High

The high impact incidents are split roughly into three categories; Wintel and Unix Server incidents take the longest to resolve (1st and 2nd), followed by SAP application, Batch Process and User Experience Transaction (3rd-5th), and finally Linux Server, Load Balancer, Other Application and Other System Software (7th-10th) are grouped at taking roughly the same amount of time to resolve.

Critical Impact

Figure 47- Average resolution time per Impact: Critical

6.2.3.5 Resolution time (average) per Service Category

The final set of funnel graphs show the average time to resolve incidents based on their Service Category (tier 1): Infrastructure, Application and Network.

Infrastructure Category

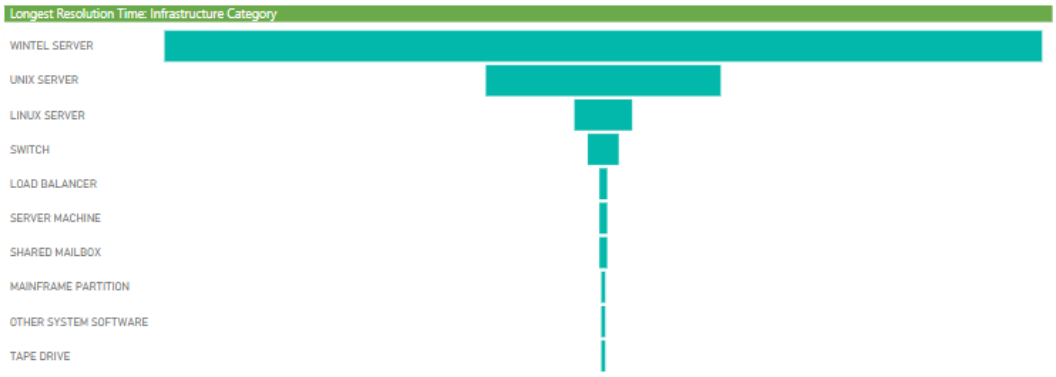


Figure 48- Average resolution time per Category: Infrastructure

For Infrastructure incidents, the servers appear to cause the greatest problems; Wintel Servers taking significantly longer than any other incident to resolve, followed by Unix and Linux (1st-3rd). Following this, Switch is the 4th longest incident to resolve, with all remaining categories taking significantly less time: Load Balancer, Server Machine, Shared Mailbox, Mainframe partition, Other system software and Tape drive all taking approximately the same amount of time (5th-10th).

Application Category

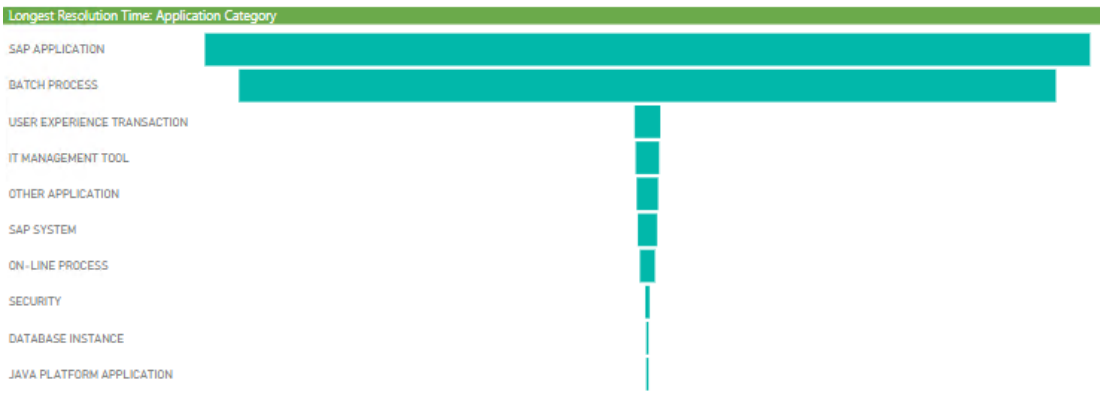


Figure 49- Average resolution time per Impact: Application

For the Application category, SAP Application Process and Batch Process incidents are close in resolution time (1st and 2nd) before the time drops considerably for the 3rd-7th longest to resolve: User Experience Transaction, IT Management Tool, Other Application, SAP System, and On-line Process. The bottom 3 incident categories take significantly less time (8th-10th): Security, Database Instance and Java Platform Application.

Network Category

Figure 50 - Average resolution time per Impact: Network

For the Network Category, fewer categories were present (6 in total). The most significant was Application Service, taking the longest to resolve, followed by Other Support Service (1st and 2nd). The 3rd-6th categories took significantly less time: Infrastructure Service, Platform Service, Corporate Service and finally, Core Business Service.

6.3 Forecast Modelling: Microsoft Power BI

Due to the time-series analysis being conducted in Microsoft Power BI, the predictive analysis for future incident count was completed in Microsoft Power BI also – for consistency. The models have been created with 95% accuracy, and the shaded area shows the upper and lower bounds (uncertainty) for each prediction. Predictions for each have been made from 2018 to 2022. The 2018-2019 prediction has greatest accuracy for most of the line graphs.

6.3.1 Incident count 2012-2018 and prediction of incident count 2018-2022.

Figure 51- Forecasting Model: Overall for ScottishPower

When we look at the overall incident count, we see a pattern forming whereby every 2 years (approximately) we see a change in direction in terms of the number of incidents reported. Despite the dips in incident counts, an upward trend can be observed 2012-2018, and therefore the forecast is for the level of incidents to continue to rise until 2020. Should the department make changes and automate processes for the most common incident causes, this could mean that the number of incidents reported would reduce over time. Should the department choose not to do this, I would expect the same pattern from previous years to continue, with a continuing upward trend.

6.3.2 Incidents per impact – historical incident count 2012-2018 and predictions of incident count 2018-2022

6.3.2.1 Low Impact

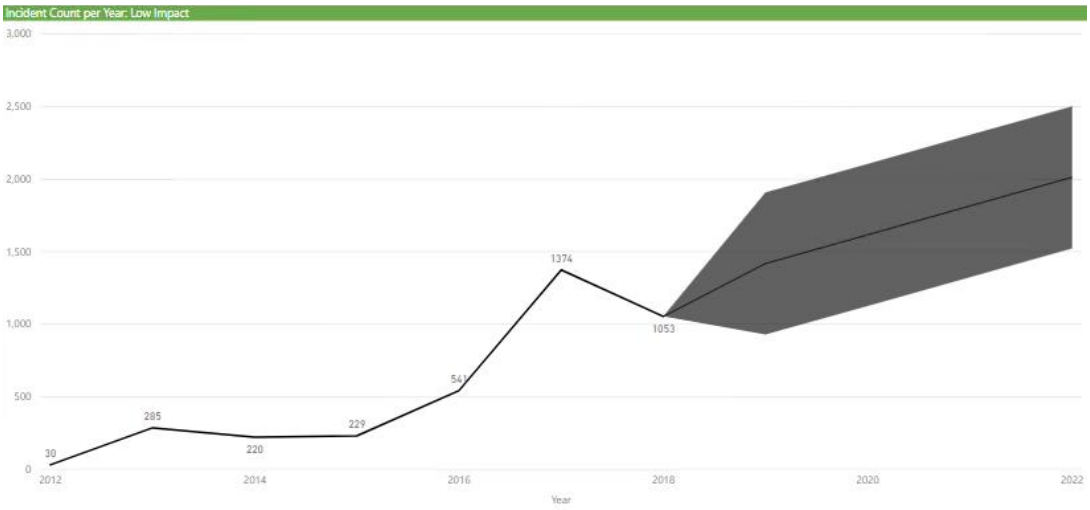


Figure 52- Forecasting Model, Incidents per impact: Low

Observing the low impact events, we see a significant peak appearing in 2017. This reduces in 2018, and could potentially reduce further in 2019. Despite these possible reductions, we see a steady upward trend, in-line with the overall incident count. The low impact events were the most common, and therefore if automation of resolution could be adopted (e.g. self-service for a common category) then this could have a significant impact on the count of low impact events.

6.3.2.2 Medium Impact

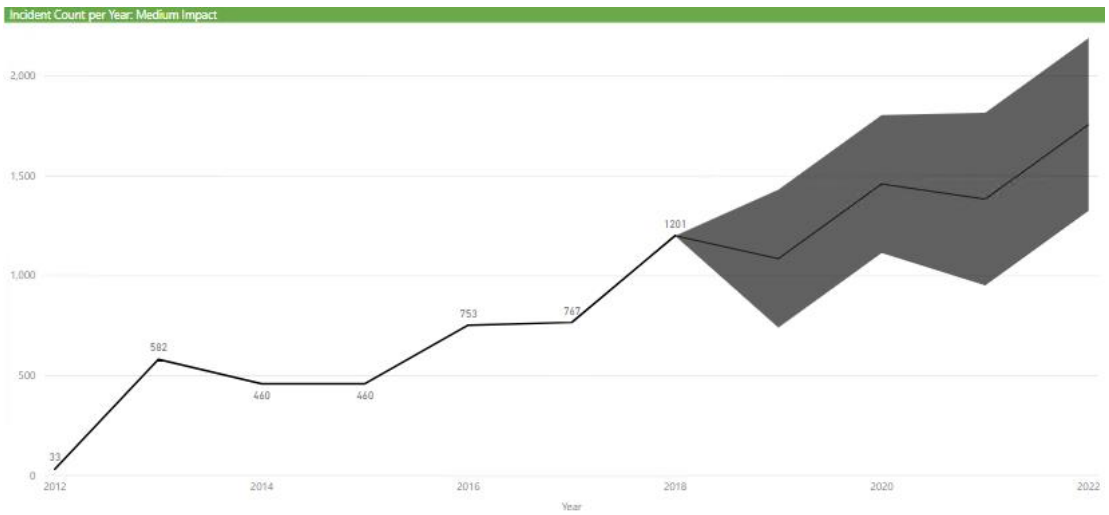


Figure 53- Forecasting Model, Incidents per impact: Medium

Medium impact events show a greater level of constant fluctuation to other impact events, though still with a steady upward trend. The forecast replicates this fluctuation, with peaks in 2020 and 2022.

6.3.2.3 High Impact

Figure 54- Forecasting Model, Incidents per impact: High

High impact events follow the same trend of low impact, in relation to the peak in 2017. We see low variation in incident count from 2013-2015, with incident levels growing to a peak from 2016-2017. This drops in 2018; however the forecast follows the upward trend, predicting another peak in 2022.

6.3.2.4 Critical Impact

Due to low levels of critical impact events, there was not a suitable amount of data to produce a reliable forecast. Of the data available, an upward trend can be observed.

6.3.3 Incidents per service category – historical incident count 2012-2018 and predictions of incident count 2018-2022

6.3.3.1 Infrastructure Category

Figure 55- Forecasting Model, Incidents per service category: Infrastructure

For the Infrastructure category, we see peaks in 2013 and 2017 and drops in incident count in 2014 and 2018. Due to the dramatic peaks/troughs, the forecast predicts a steady level of incidents at approximately the 600 level, though with a large margin of error, due to the lack of a consistent pattern or upward/downward trend.

6.3.3.2 Application Category

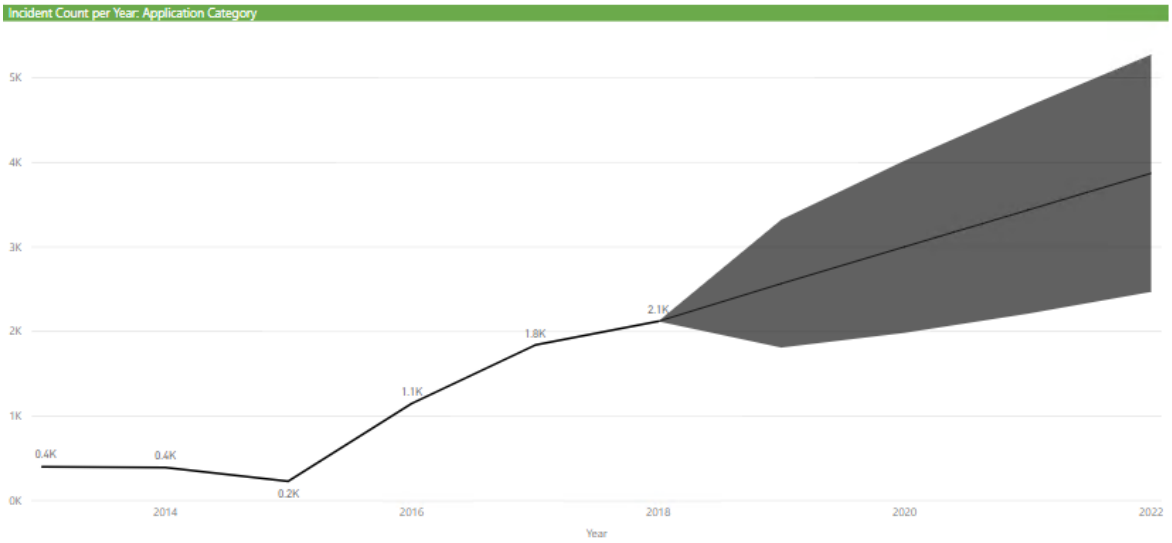


Figure 56- Forecasting Model, Incidents per service category: Application

The Application category has a relatively steady upwards trend, with one dip in 2015. The forecast therefore predicts the continuation of this upward trend, but with a margin of error should the dip (or resultant peak following the dip) be repeated.

6.3.3.3 Network Category

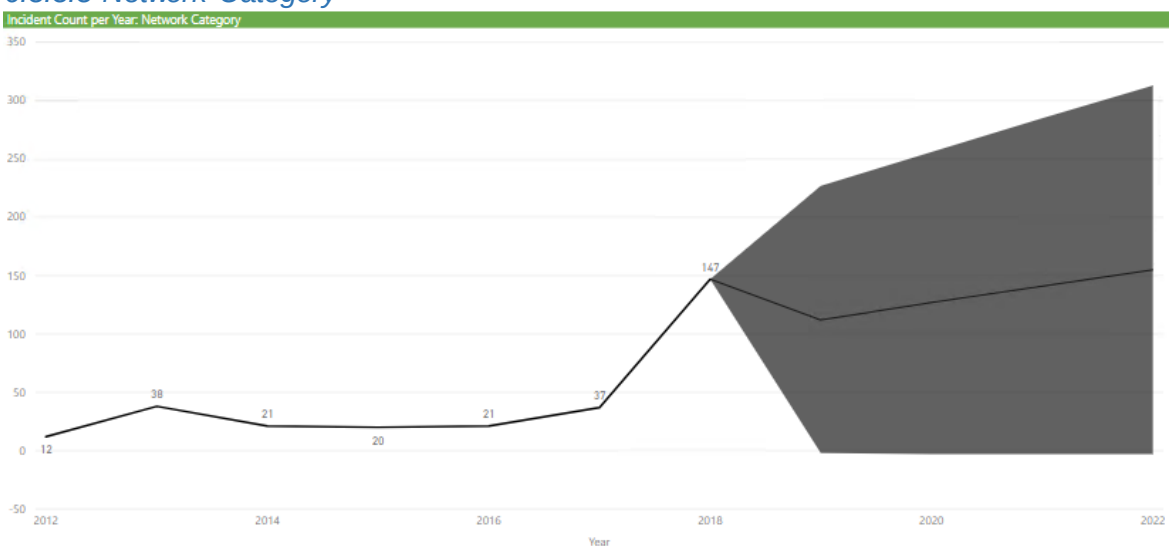


Figure 57- Forecasting Model, Incidents per service category: Network

The Network Category shows a relatively consistent level of incidents reported, until a peak formed from 2017-2018. Due to this being quite a significant increase in reported incidents, the forecast predicts a slight reduction 2018-2019, before a shallow but upwards trend until 2022.

6.3.4 Resolution time (average) per Impact – historical average resolution time 2012-2018 and predictions of average resolution time 2018-2022

6.3.4.1 Low

Figure 58- Forecasting Model, Average resolution time per impact: Low

The low impact incidents show dramatic fluctuation in resolution times, with peaks in 2012 and 2015, and troughs in 2013 and 2018. For this reason, the forecast predicts a slight peak in 2019, before levelling out until 2022, with high margins for error.

6.3.4.2 Medium

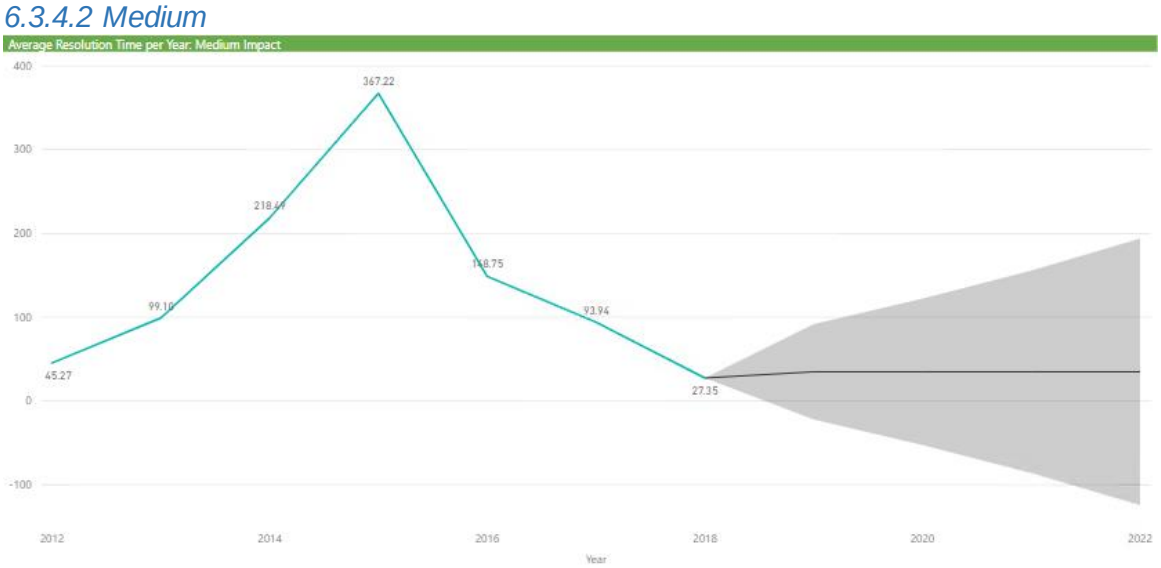


Figure 59- Forecasting Model, Average resolution time per impact: Medium

The Medium impact incidents category shows a steady increase in resolution times until a peak in 2015, before reducing steadily until 2018. The forecast predicts a steady level of resolution times until 2022.

6.3.4.3 High

Figure 60- Forecasting Model, Average resolution time per impact: High

The high impact incidents have extremely fluctuating resolution times, with peaks in 2012, 2014, and 2017, and troughs in 2013 and 2016. The forecast therefore predicts relatively steady resolution times until 2022, with a wide confidence level.

6.3.4.5 Critical

Due to low levels of critical impact events, there was not a suitable amount of data to produce a reliable forecast. Of the data available, a strong downward trend can be observed.

6.3.5 Resolution time (average) per Service Category – historical average resolution time 2012-2018 and predictions of average resolution time 2018-2022

6.3.5.1 Infrastructure

Figure 61 - Forecasting Model, Average resolution time per service category: Infrastructure

The Infrastructure category has peaks in 2012 and 2015, with quite a wide range in resolution times. The forecast predicts a relatively steady, medium resolution time for 2018-2022, with a wide confidence level should the trend of peaks and troughs continue.

6.3.5.2 Application

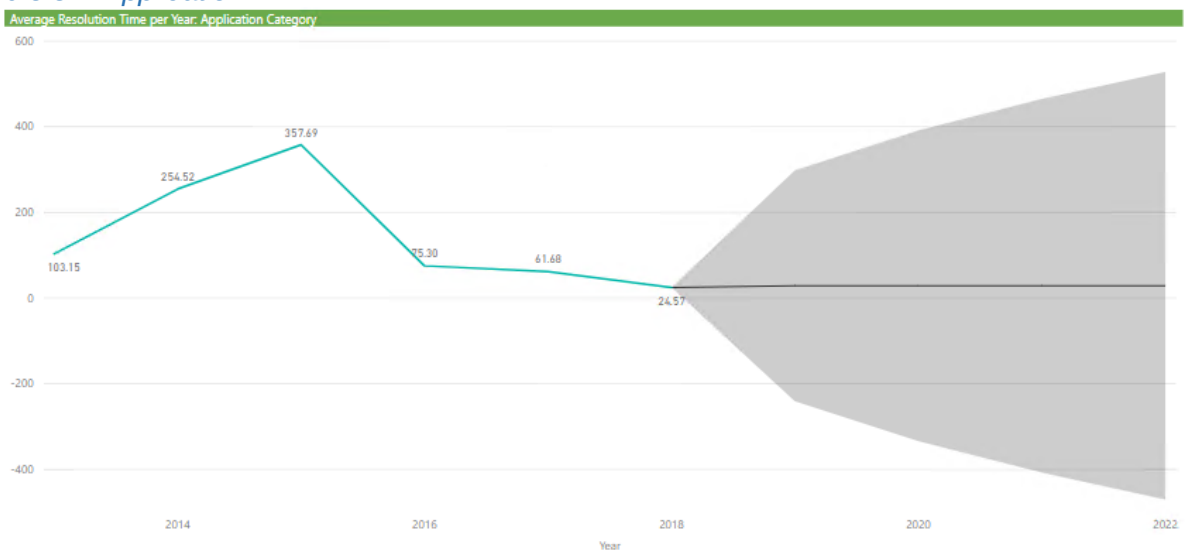


Figure 62 - Forecasting Model, Average resolution time per service category: Application

The Application category shows resolution times growing to a peak from 2012-2015, reducing and steadying from 2015-2016 and onwards. The forecast predicts a low average resolution time, with a growing margin for error over time, should the 2015 peak repeat itself.

6.3.5.3 Network

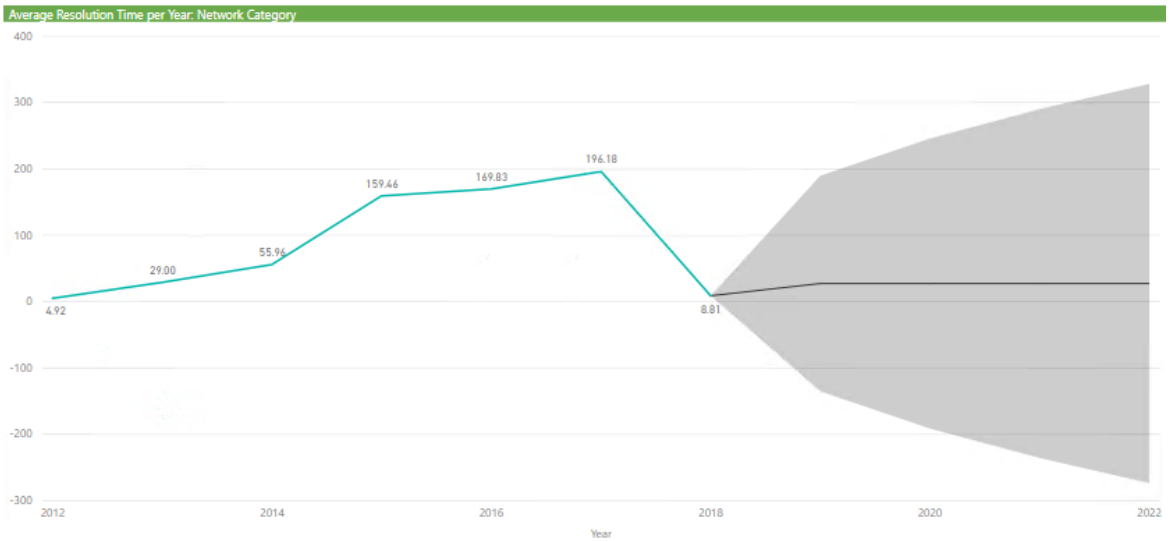


Figure 63 - Forecasting Model, Average resolution time per service category: Network

Finally, the Network category shows a steady increase in average resolution time from 2012-2014, with an expedited increase between 2014-2017, before a drop between 2017-2018. The forecast predicts a low average resolution rate for 2018-2022, with a margin of error should the peak seen in 2017 occur again.

6.4 Live Dashboard Analysis

This section will focus on the analysis applied to create a live report, which allows the client to monitor progress and current position of incident count and resolution times.

6.4.1 Data Limitations

A key limitation was – once again – the data gathering process and the way by which ScottishPower stored and processed their data. By altering this to a daily update, a far more transparent view would be possible, that would allow for active benchmarking and performance monitoring/improvement. Doing this, however, would require a change in the way that reports are generated in the UK due to the source data being stored and processed in Spain. Until this change can be made, a weekly view can be produced that can then be updated with appropriate data when the infrastructure is in place. This is useful both from a continuity perspective as well as in terms of predictive analytics – where trends can be identified, it may be possible to automate processes for recurring issues, or staff appropriately to tackle peaks in incident events. To design and test the dashboard, the sandbox approach was used; this, according to Narayanan (2016), is the process of using a small dataset to build and test a model, or to manipulate data to produce results rather than using a full dataset which contains large volumes and will take longer to process. This helps to ensure that a suitable tool has been chosen that fits the requirements of the work required. If the

chosen technology is successful, and the output is sufficient in meeting the client's needs, then this can be up-scaled to the full dataset.

6.4.2 Data Gathering

The data gathering process was the same as previous, however only focussing on the most recent report generated. The key issue of this – especially when wanting a live view – is that there is a lag in the data processing that means that the data available may not reflect the true current position of the incident reports.

6.4.3 Dashboard creation

Three dashboards were created for the client: Overview, Impact and Service Category. Each of the dashboards contain all of the historical data in the background, with a date slider on each which allows the client to view both the current, live dashboard as well as historical dashboards. This allows the client to monitor how they are performing at present, and how that compares to past performance. This creates a benchmark that the incidents can be compared to and improved against. Voss et al (1997) discuss the background of benchmarking, originally initiated by Xerox, and what the purpose of benchmarking is: "…the way of identifying how superior companies organise their processes, a company can seek to adopt and adapt these practices". Voss et al also discuss how benchmarking can be used to set goals, as doing so often encourages objectives to be stretched further than their initial scope (Voss et al, 1997). With this in mind, it is clear that the action of benchmarking is beneficial and would encourage continuous performance improvement throughout the company (one of ScottishPower's key performance indicators). Having greater transparency of incident levels will also remove the ability to hide areas of poor performance, and instead it will force improvements to be made.

6.4.3.1 Overview Dashboard

The overview dashboard can be seen in figure 64 below; the dashboard allows the user to gain a high level overview of the current position of IT incident reports. On the top right of the report, the user has the ability to change the date range of the data they want to view, if they are interested in a specific time period. Below this, the user can check that the dashboard they are viewing is associated with the correct data, and when that source data was last updated. Drop down slicers were created to allow the user to filter down the data on-screen. The user can see what percentage of incidents reported were done so by managers – and therefore had no direct business impact. The data is split into three distinct 'tiers', namely 1, 2, and 3. Tier 1 is high level, and splits the data based on service categorisation (Application, Network or Infrastructure); Tier 2 is more detailed and splits the data based on an overview of the specific problem (e.g. 'Business Service', 'Processing System'); Tier 3 splits the data on specific incident categories (e.g. 'SAP application', 'Wintel server').

As seen below, an overview is given of key data: A graph showing the incidents reported split by impact level, and colour-coded by their tier 1 service categorisation; a pie chart showing the percentage split of incidents per tier 1 service categorisation and a donut chart showing how the incidents were split by tier 2 category. At the bottom of the dashboard, a tornado graph is used to highlight the top 10 tier 3 incidents that occurred during the period. Finally, on the bottom right, data has been extracted and presented to direct the attention of the user to the critical impact incidents reported during the period (period defined in the top right using the date slider). The drop down slicers can be selected to filter down data – these include 'company' (due to ScottishPower being owned by Iberdrola), and 'country' (due to the UK and Spanish ScottishPower/Iberdrola offices, and the outsourced IT work in India).

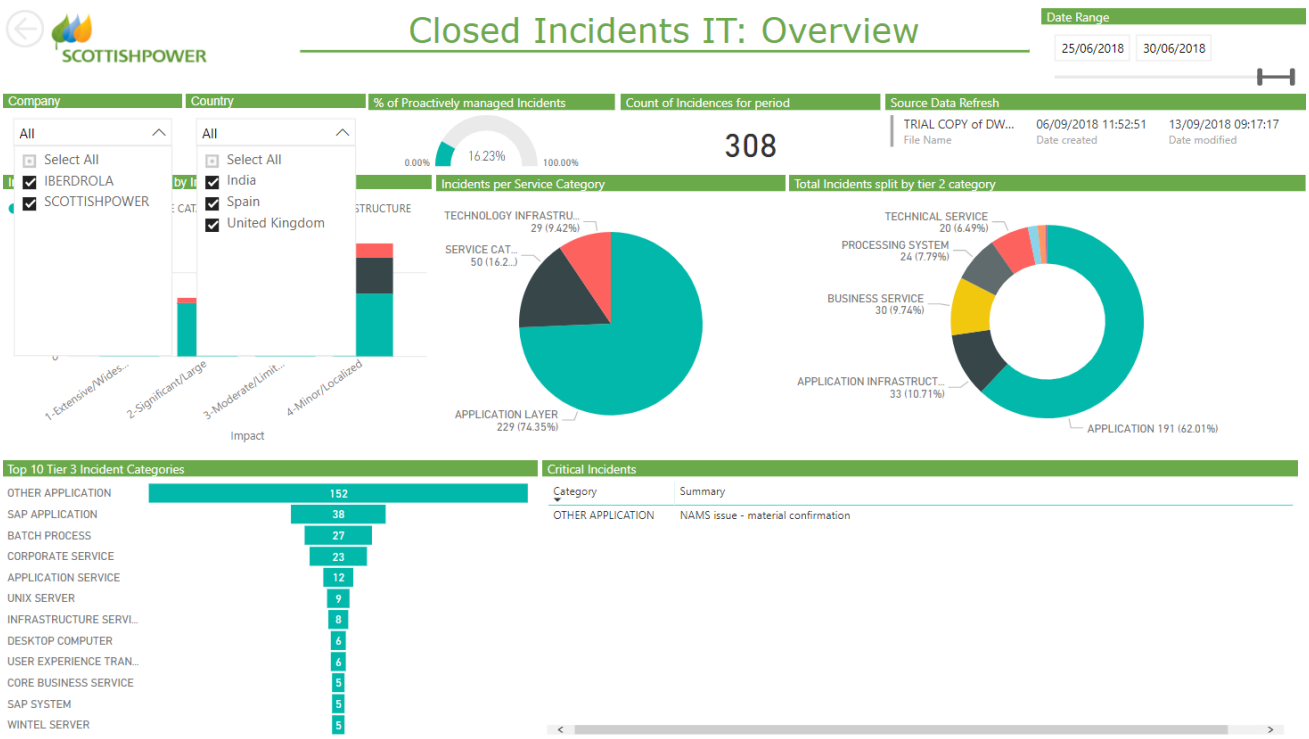


Figure 64 - Live Dashboard: Overview of key data

The design of the dashboard is very 'point and shoot' in terms of its ease of use; if the user wishes to filter on a specific element, they can do so by selecting their filters from the drop-down slicers or by simply clicking on said element and the report will automatically update. This can be seen in figure 65 below, where 'Application Layer' was selected from the centre pie chart. We see Power BI fading out the non-relevant data, and updating the numerical values and critical incident data extract as appropriate.

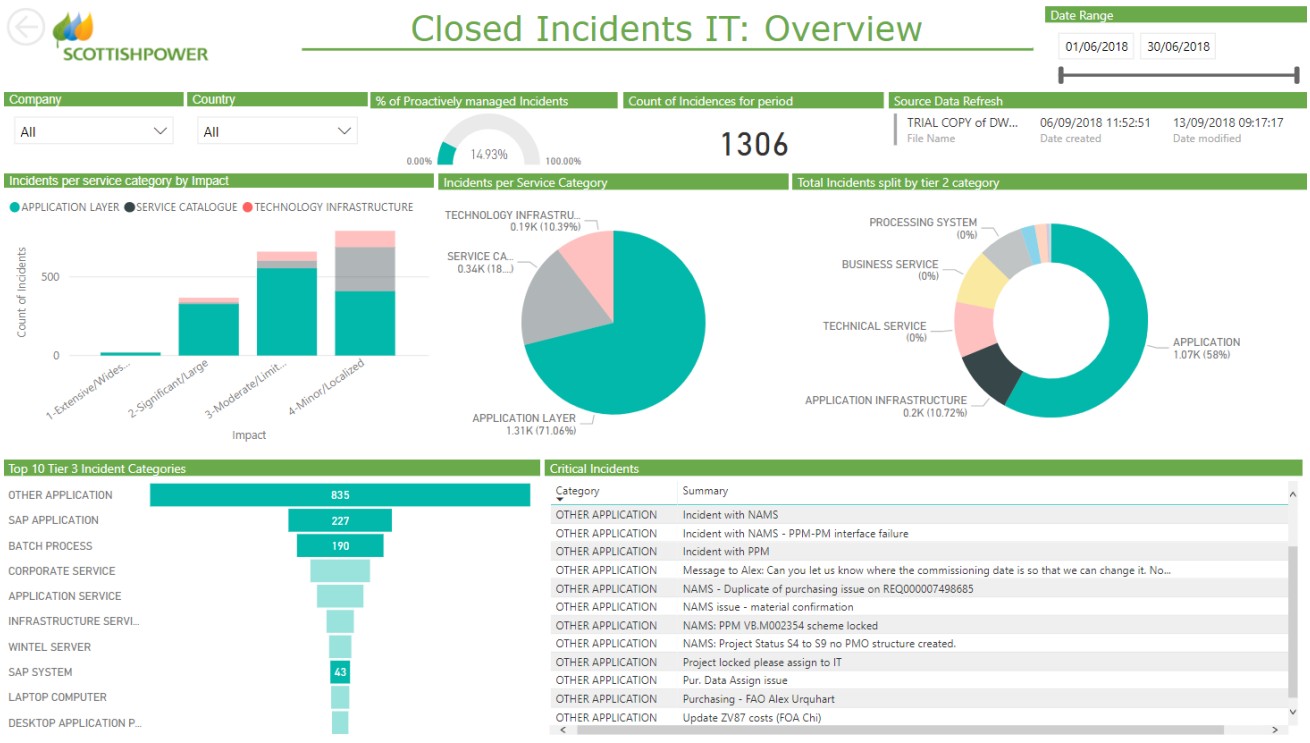


Figure 65 – Live dashboard: filtered overview of data

6.4.3.2 Impact Dashboard

The impact dashboard can be seen in figure 66 below. This dashboard is simple, containing four tornado graphs which show the top 10 average resolution times split by the impact level (minor, moderate, significant and extensive). This is key in highlighting what the most time-consuming incidents are to deal with and how business-critical these incidents are.

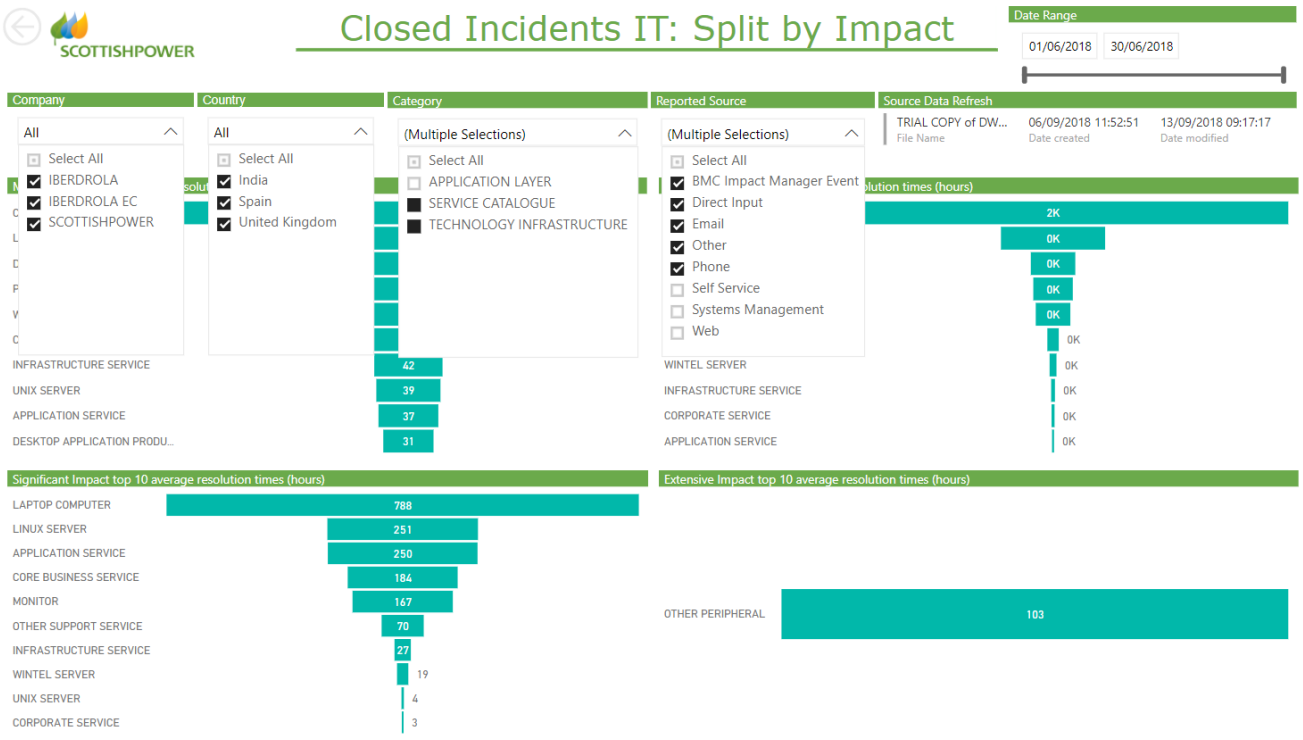
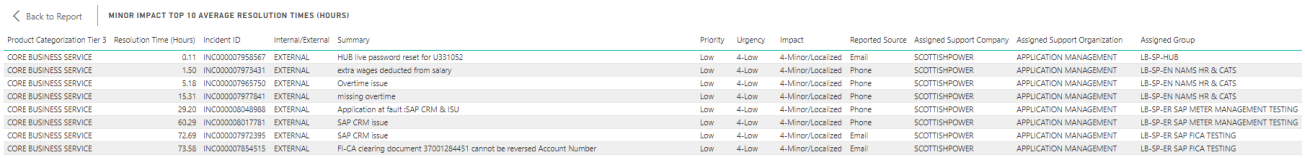


Figure 66 - Impact Dashboard

Once again, as shown above, drop-down slicers are available to filter the data on company and country as well as category, and reported source (e.g. BMC Manager (proactive), email, phone, self-service etc). The date sliders can also be used to focus on a particular time period, and the source data refresh remains available to view when the data was last updated.

Figure 67, below, shows the 'drill down' data available to the user; if the user wants more information on a specific element, they can right-click and select 'see records' – this allows the user to look at each individual instance. Below shows the output when 'Core Business Service' is selected.


Back to Report | MINOR IMPACT TOP 10 AVERAGE RESOLUTION TIMES (HOURS)

Product Categorization Tier 3	Resolution Time (Hours)	Incident ID	Internal/External	Summary	Priority	Urgency	Impact	Reported Source	Assigned Support Company	Assigned Support Organization	Assigned Group
CORE BUSINESS SERVICE	0.11	INC000007958567	EXTERNAL	HUB live password reset for U331052	Low	4-Low	4-Minor/Localized	Email	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-HUB
CORE BUSINESS SERVICE	1.50	INC000007973431	EXTERNAL	extra wages deducted from salary	Low	4-Low	4-Minor/Localized	Phone	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-EN NAMS HR & CATS
CORE BUSINESS SERVICE	5.18	INC000007965750	EXTERNAL	Overtime issue	Low	4-Low	4-Minor/Localized	Phone	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-EN NAMS HR & CATS
CORE BUSINESS SERVICE	15.31	INC000007977841	EXTERNAL	missing overtime	Low	4-Low	4-Minor/Localized	Phone	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-EN NAMS HR & CATS
CORE BUSINESS SERVICE	29.20	INC000008048988	EXTERNAL	Application at fault :SAP CRM & ISU	Low	4-Low	4-Minor/Localized	Phone	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-ER SAP METER MANAGEMENT TESTING
CORE BUSINESS SERVICE	60.29	INC000008017781	EXTERNAL	SAP CRM issue	Low	4-Low	4-Minor/Localized	Phone	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-ER SAP METER MANAGEMENT TESTING
CORE BUSINESS SERVICE	72.69	INC000007972395	EXTERNAL	SAP CRM issue	Low	4-Low	4-Minor/Localized	Email	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-ER SAP FICA TESTING
CORE BUSINESS SERVICE	73.58	INC000007854515	EXTERNAL	FI-CA clearing document 3700128445 1 cannot be reversed Account Number	Low	4-Low	4-Minor/Localized	Email	SCOTTISHPOWER	APPLICATION MANAGEMENT	LB-SP-ER SAP FICA TESTING

Figure 67 - Drill down data: Impact dashboard

6.4.3.3 Service Category Dashboard

Figure 68 below shows the incident reports split by service category. The structure and filters available are the same as those present in the impact dashboard (figure 67). The dashboard contains three tornado graphs and one donut chart. The tornado graphs focus on the three service categories: Service (Network), Application and Technology Infrastructure, showing the top incidents associated with each.

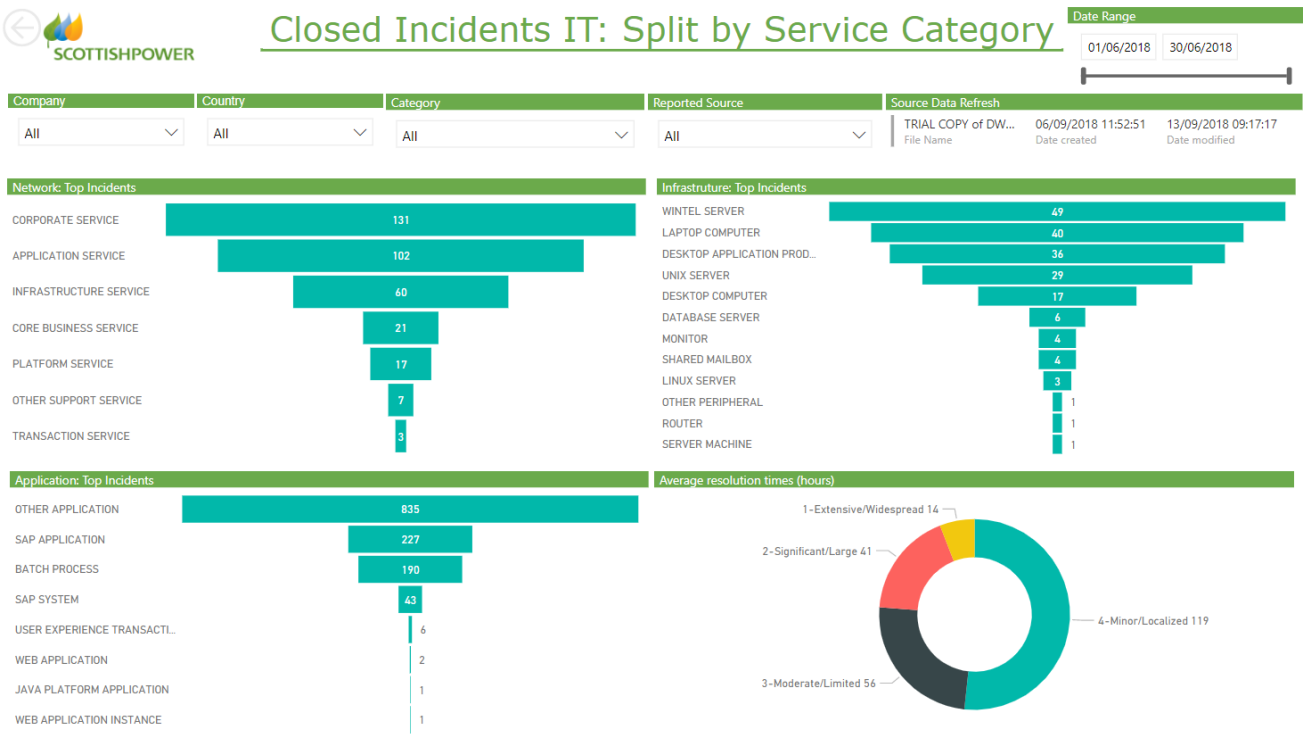


Figure 68 - Live dashboard: Service categories

The donut chart in the bottom right hand side plots the average resolution times of each incident in hours based on the level of business impact.

Figure 69 below shows the output when one individual category is selected: Application. We see the top two areas being blank, with the Application tornado graph showing the top incidents and the donut chart of average resolution times being specific to Application incidents.

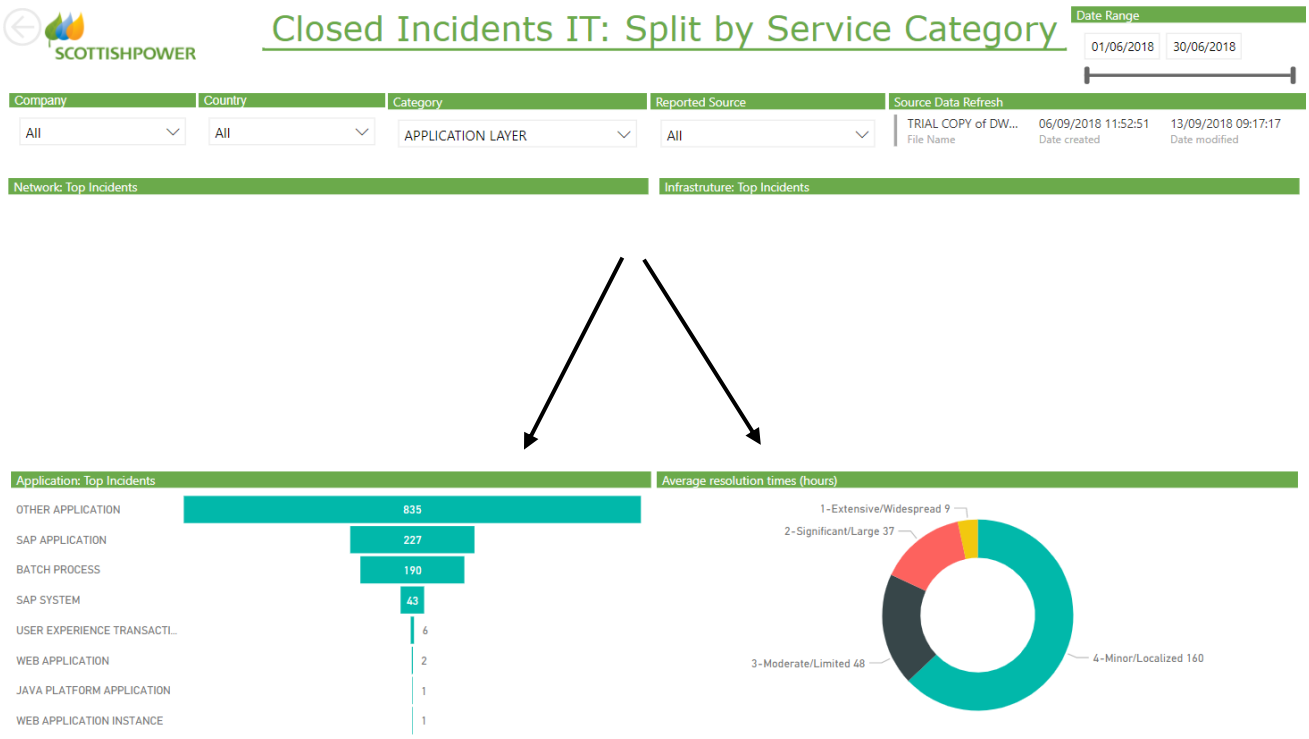
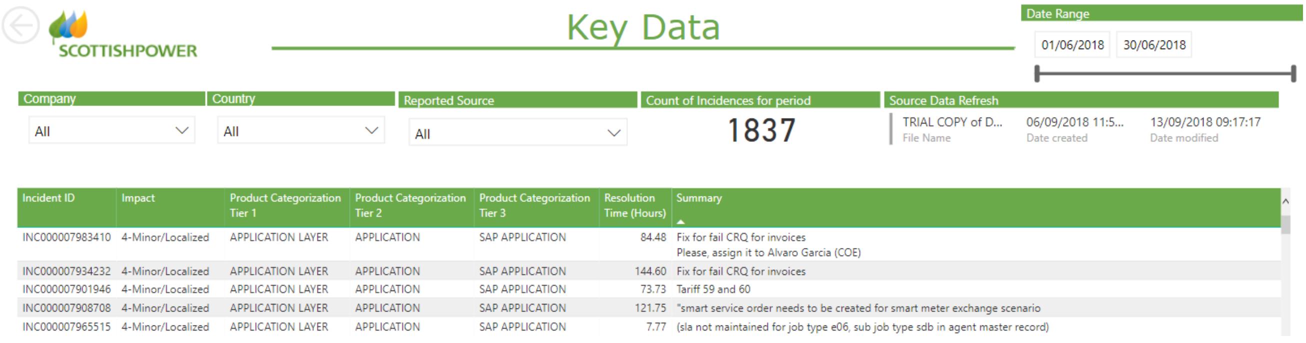


Figure 69 - Live Dashboard: Interactive filters

In order to ensure that all required detail is still available to users, all elements from the data extract were also requested to be contained within the report. Due to the vast amount of elements present – many of which do not add value – it was decided that a 'key data' report would be generated, as well as an 'all data' report for anyone who required it. By doing this, the Microsoft Power BI reporting pack can become the new 'go-to' for the full IT Incident Report.

The Key Data report can be seen in figure 70 below; it pulls data from the columns deemed most important: Incident ID, Impact level, Product categorisation tiers 1, 2, and 3, the resolution time (in hours) and the free-hand summary written by the person processing the incident. The report maintains the same level of filtering capabilities as the reports previously mentioned: Company,

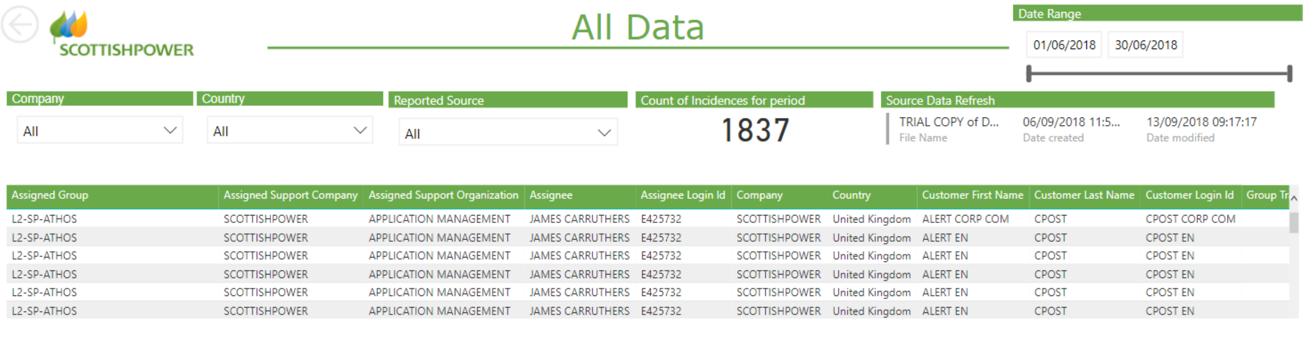
Country, Reported Source and Date Range.



Incident ID	Impact	Product Categorization Tier 1	Product Categorization Tier 2	Product Categorization Tier 3	Resolution Time (Hours)	Summary
INC000007983410	4-Minor/Localized	APPLICATION LAYER	APPLICATION	SAP APPLICATION	84.48	Fix for fail CRQ for invoices Please, assign it to Alvaro Garcia (COE)
INC000007934232	4-Minor/Localized	APPLICATION LAYER	APPLICATION	SAP APPLICATION	144.60	Fix for fail CRQ for invoices
INC000007901946	4-Minor/Localized	APPLICATION LAYER	APPLICATION	SAP APPLICATION	73.73	Tariff 59 and 60
INC000007908708	4-Minor/Localized	APPLICATION LAYER	APPLICATION	SAP APPLICATION	121.75	"smart service order needs to be created for smart meter exchange scenario
INC000007965515	4-Minor/Localized	APPLICATION LAYER	APPLICATION	SAP APPLICATION	7.77	(sla not maintained for job type e06, sub job type sdb in agent master record)

Figure 70 – Live dashboard: Key data table

As seen in figure 71 below, the ‘Full Data’ report again maintains the same level of filtering capability and holds the same structure as the Key Data report, but containing all extracted data. This area acts as a ‘Data lake’ in terms of how the data is being used; it can be stored here and retrieved when required or used for exploratory analysis.



Assigned Group	Assigned Support Company	Assigned Support Organization	Assignee	Assignee Login Id	Company	Country	Customer First Name	Customer Last Name	Customer Login Id	Group Tr
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT CORP COM	CPOST	CPOST CORP COM	
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT EN	CPOST	CPOST EN	
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT EN	CPOST	CPOST EN	
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT EN	CPOST	CPOST EN	
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT EN	CPOST	CPOST EN	
L2-SP-ATHOS	SCOTTISHPOWER	APPLICATION MANAGEMENT	JAMES CARRUTHERS	E425732	SCOTTISHPOWER	United Kingdom	ALERT EN	CPOST	CPOST EN	

Figure 71 - Live dashboard: All data table

7.0 Conclusions and Recommendations

The Conclusions and Recommendations have been split into the three sections of analysis. This has been done so that findings from each area can be highlighted independently.

7.1 Conclusions

7.1.1 Historical Analysis

The analysis of historical data was done both in an exploratory sense, and a structured sense. These unveiled answers to pre-defined questions that the client had, and also provided the client with answers to 'unknowns'.

7.1.1.1 Exploratory

Mawer's research (2017) backed up the need for exploratory analysis, as the visualisations created uncovered elements of the data that would otherwise have remained unknown. The exploratory analysis revealed that 60% of incidents reported to IT were due to unplanned events; meaning that these were unexpected and had to be dealt with reactively rather than proactively. It also revealed that over 71% of the incidents reported to IT fell into the 'Application' category; of which, nearly 65% were due to SAP Applications. SAP Applications were responsible for 46% of the incidents reported to IT overall. Peng and Matsui's research (2018) also strengthened the analysis, as exploratory analysis was conducted to identify correlations within the data; we saw impact and priority, and internal/external reporting and reported source having the closest relationship. Using the heat mapping tool, we were able to identify the highest impact category for proactive incidents as being 'System Management' and the highest impact category for reactive incidents being 'Other' and 'Walk in'.

Across all years, there was a consistent peak in the time series analysis in incident count in the month of August. A linear relationship could be identified between the number of incidents logged and the level of impact (low impact having the greatest incident count, and critical impact having the lowest).

7.1.1.2 Structured

The structured analysis shows the top tier 3 incident categories from the incident count and the average resolution times. Only the most significant categories have been included, as in many cases 2-5 categories would dominate the incident count/top resolution times before reducing to an insignificant level.

Overall top incident count

All three of the server types appeared in the top 10 incident count, with the top 5 incident occurrences as follows: Batch process, Wintel and Unix servers, SAP Applications, and Shared Mailbox.

Split by Impact

a) Top Incident Count

All of the 6 critical-impact categories (Wintel Server, Batch Process, Linux Server, SAP Application, Shared Mailbox and Unix Server) can be found across each of the 'top 10' lists per impact level. These 6 should therefore be focussed upon in order to reduce business impact and quantity of incidents logged.

b) Top Average Resolution Time

All 3 servers appear in almost all of the impact levels, as well as Batch Process and SAP Application; this indicates that regardless of the impact level, these categories take the longest time to resolve. These 5 should therefore be focussed on to reduce resolution times.

Split by Tier 1 Category

Analysis showed that the Application category holds 71% of incidents, the key elements for the client to focus on and conduct further analysis into would be Batch Process and SAP Application due to the volume of incidents that will be associated with these categories.

a) Top 10 Incident Count

There is greater variation in incident cause when the data is split by category, rather than impact. There is a link between the most frequently logged incidents and those that take the longest to resolve for each category.

Application	Batch Process and SAP Application are the most significant.
Infrastructure	Wintel and Unix Servers are the most significant, before Shared Mailbox, Linux Server and Switch.
Network	Application Service holds the highest incident count, followed by Other Support Service, Infrastructure Service, SAP Application and Batch Process.

Table 1 - Top 10 Incident Count split by Tier 1 Category

b) Top 10 Average Resolution Time

Once again, there is greater variation in the tier 3 categories taking the longest when split by each tier 1 category, compared to when the data was split by impact level. As previously mentioned, with the Application category holding 71% of the incident reports, these should be the focus for the client.

Application	SAP Application and Batch Process take significantly longer to resolve than any other category.
Infrastructure	Wintel Server takes the longest average time, followed by Unix and Linux servers and Switch.
Network	Application Service and Other Support Service take the longest time to resolve, followed by Infrastructure Service.

Table 2 - Top 10 Resolution Time split by Tier 1 Category

7.1.2 Dashboards

Robertson and Robertson (2013) explored the benefits of quick and dirty modelling, and Narayan (2016) researched the benefits of the sandboxing technique; adopting both of these methods, suitable dashboards were created that met the needs of the client. There are three in total: Overview, split by Impact, and split by Tier 1 Category. On top of this, two tabular data dashboards were created: one which contains 'key data' that was used for analysis, and one that contains 100% of the data present in the source file, should it be required.

The dashboard can provide the client with an in-the-moment snapshot, reliant on how frequently the source data can be refreshed. Once a live stream of data can be provided to the report, this will rectify any data lag issues. The dashboards can be used to proactively monitor and track current levels of incident reports, flagging critical incidents that may have strong business impact. .This is in-line with research conducted by Selby (2005) who states that dashboards allow users to understand, evaluate and predict development, and that they can be used to uncover areas of poor performance or weakness. Benchmarking can also be done, using the date slider to view older data. Should the client wish for the historical time series analysis to be present in their dashboard pack in the future, the code can simply be copied into a new dashboard and refreshed with up-to-date data.

7.1.3 Forecasts

Through quick and dirty modelling and sandboxing, suitable forecasts were created for the client to predict their future incident counts and average resolution times. For consistency, it was decided that these would match the structured historical analysis. The overall finding was that there is an upward trend across all of the forecasts; even those that can be seen fluctuating annually.

7.2 Recommendations

7.2.1 Historical Analysis

If the department want to effectively monitor the split of proactive/reactive incidents, this must be added to the source document when incidents are logged; this way there will be greater data integrity and forecasting and analysis can be performed with confidence. A large proportion of incidents were logged as 'other' at various levels of categorisation; it is clear by looking at the free-hand notes that these could be more specifically categorised. By removing or limiting the use of this category, the quality of the analysis would improve overall.

Tukey's research (1977) was particularly relevant, as the use of exploratory analysis rather than purely confirmatory produced suggestions for further research to the client SAP and Batch Process events should have further analysis conducted into them to understand how incident count and resolution times can be reduced. The application category should also be analysed further, as it holds over 71% of the incidents logged.

The low impact incidents have the highest frequency, which suggest that a self-service tool may be suitable for the resolution of a proportion of these incidents.

7.2.2 Dashboards

Dashboards should be used with an automatic data refresh to proactively monitor live incident reports, and respond to critical impact events occurring that could cause detrimental business impact. Benchmarking activities should also be undertaken to compare current position with previous years; Voss et al (1997) explain that goals set by benchmarking encourage companies to perform to a higher standard. All of the visuals presented in the dashboard are fully interactive and can be 'drilled down' to gain greater insight, or to view the raw data. Microsoft Power BI was chosen based on literature, and approved as a successful proof of concept project within the client's company.

7.2.3 Forecasts

Forecasts should be used to plan ahead for future peaks in incident count, or extended incident resolution times. They should also be used to identify areas where self-service tools or proactive work should be undertaken to reduce potential business impact and speed up resolution times.

9.0 References

Aleksejeva, N. (2015). The Infamous Pie Chart: History, Pros, Cons and Best Practices, [online] Available at: https://infogram.com/blog/the-infamous-pie-chart-history-pros-cons-and-best-practices/[Accessed 23 Nov. 2018].

Al-Hajj, S., Pike, I. and Fisher, B. (2013). Interactive Dashboards: Using Visual Analytics for knowledge Transfer and Decision Support. [online] Available at: https://www.cc.gatech.edu/gvu/ii/PublicHealthVis/Papers/Dashboards%20-%20Decision-Making%20Support%20for%20Health%20Informatics_F.pdf [Accessed 26 Oct. 2018].

Burgelman, L. (2015). *The Rise of the Citizen Data Scientist – NGDATA*. [online] NGDATA. Available at: https://www.ngdata.com/the-rise-of-the-citizen-data-scientist/ [Accessed 26 Oct. 2018].

de Jonge, E. and van der Loo, M. (2018). *An introduction to data cleaning with R*. [online] Cran.r-project.org. Available at: https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf [Accessed 16 Dec. 2018].

Energy Networks Australia (2017). *Electricity Network Transformation Roadmap: Final Report*. [online] Available at: https://www.energynetworks.com.au/sites/default/files/entr_final_report_web.pdf [Accessed 14 Dec. 2018].

Gartner (2018). Magic Quadrant for Analytics and Business Intelligence Platforms [online] Available at: https://www.gartner.com/doc/3861464/magic-quadrant-analytics-business-intelligence [Accessed 14 Dec. 2018].

Gleeson, P. (2017). Which Languages Should You Learn For Data Science? [online] Available at: https://medium.freecodecamp.org/which-languages-should-you-learn-for-data-science-e806ba55a81f [Accessed 14 Dec. 2018].

G2 Crowd (2018). Best Data Visualisation Software [online] Available at: https://www.g2crowd.com/categories/data-visualization [Accessed 18 Dec 2018].

Harpham, B. (2016). How Data Science is Changing the Energy Industry. [online] Available at: https://www.cio.com/article/3052934/big-data/how-data-science-is-changing-the-energy-industry.html [Accessed 15 Dec. 2018].

Healey, C. (n.d.). Choosing Effective Colours for Data Visualization (Christopher G Healey, undated). [online] Available at: https://www.csc2.ncsu.edu/faculty/healey/download/viz.96.pdf [Accessed 26 Oct. 2018].

IBM Big Data & Analytics Hub. (n.d.). *Extracting business value from the 4 V's of big data*. [online] Available at: https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data [Accessed 4 Dec. 2018].

Informatica.com. (2018). *What is Data Analytics: Definition | Informatica US*. [online] Available at: https://www.informatica.com/services-and-training/glossary-of-terms/data-analytics-definition.html#fbid=xeNnN1W-iAp [Accessed 28 Dec. 2018].

Järveläinen, J. (2013). IT incidents and business impacts: Validating a framework for continuity management in information systems. *International Journal of Information Management*, 33(3), pp.583-590.

Mawer, C. (2017). The Value of Exploratory Data Analysis. [online] Available at: https://www.svds.com/value-exploratory-data-analysis/ [Accessed 28 Nov. 2018].

Maletic, C. and Marcus, A. (2000) Data Cleansing: Beyond Integrity Analysis. [online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.5212&rep=rep1&type=pdf [Accessed 27 Dec. 2018].

Murphy, S. (2013). Data Visualization and Rapid Analytics: Applying Tableau Desktop to Support Library Decision-Making. *Journal of Web Librarianship*, 7(4), pp.465-476.

Narayanan, R. (2016). A Roadmap to Implementing Big Data Projects. [online] Available at: https://re-magazine.ireb.org/articles/a-roadmap-to-implementing-big-data-projects [Accessed 28 Nov. 2018].

Peng, R. and Matsui, E. (2017). The Art of Data Science. [online] Available at: https://bookdown.org/rdpeng/artofdatascience/exploratory-data-analysis.html [Accessed 20 Dec 2018].

Piatetsky, G. (2017). New Leader, Trends and Surprises in Analytics, Data Science, Machine Learning Software Poll. [online] Available at: https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html [Accessed 15 Nov 2018].

Robertson, S. and Robertson, J. (2013). Mastering the requirements process. Upper Saddle River, NJ: Addison-Wesley.

Selby, R. (2005). Analytics-Driven Dashboards Enable Leading Indicators for Requirements and Designs of Large-Scale Systems. *IEEE Software*, 26(1), pp.41-49.

SharpSight.com (2018) R vs Python: Which to learn for Data Science. [online] Available at: https://www.sharpsightlabs.com/blog/r-vs-python/ [Accessed 28 Dec. 2018].

Techopedia.com. (2018). *What is Data Analytics? - Definition from Techopedia*. [online] Available at: https://www.techopedia.com/definition/26418/data-analytics [Accessed 28 Dec. 2018].

The Four V's of Big Data. [online] Available at: https://www.ibmbigdatahub.com/infographic/four-vs-big-data [Accessed 28 Dec. 2018].

Thomas, J & Cook, K & Electrical, Institute & Engineers, Electronics. (2005). Illuminating the path: The research and development agenda for visual analytics. California: IEEE Computer Society Press.

Tukey, J.W. (1977). Exploratory Data Analysis. Pearson.

Vallee, G., Engelmann, C., Tikotekar, A., Naughton, T., Charoenpornwattana, K., Leangsuksun, C., & Scott, S. (2008). A framework for proactive fault tolerance. In *Third International Conference on Availability, Reliability and Security*, Barcelona, Spain.

Voss, C., Åhlström, P. and Blackmon, K. (1997). Benchmarking and operational performance: some empirical results. International Journal of Operations & Production Management, 17(10), pp.1046-1058.

Zhang, L., Stoffel, A. and Behrisch, M. (2012). Visual Analytics for the Big Data Era – A Comparative Review of. *IEEE Conference on Visual Analytics Science & Technology*, [online] pp.173-182. Available at: http://kops.uni-konstanz.de/bitstream/handle/123456789/22540/Zhang_225405.pdf?sequence=2 [Accessed 26 Oct. 2018].

10.0 Appendices

Appendix 1: Output from fig 6, Attribute Class of Dataframe

```
> sapply(numeric2na.df, class)
                                    Priority
                                   "integer"
                                     Urgency
                                   "integer"
                                      Impact
                                   "integer"
 Resolution.Product.Category.Tier1.S1T2A3
                                   "integer"
                         Internal1.External2
                                   "integer"
                                Submit.Year
                                   "integer"
                               Submit.Month
                                   "integer"
                                Closed.Year
                                   "integer"
                               Closed.Month
                                   "integer"
                        Reported.Source.1.12
                                   "integer"
Assigned.Support.Organization.A1.S2.I3.O4
                                   "integer"
                        Proactive1.Reactive2
                                   "integer"
                          Proactive.Reactive
                                    "factor"
```

Appendix 2: Output from fig 7, Overview of Data: Head Function

```
> #Check top 6 rows of the data
> head(numeric2na.df)
  Priority Urgency Impact Resolution.Product.Category.Tier1.S1T2A3
1        2       2      2                                        3
2        3       3      3                                        3
3        2       2      2                                        3
4        2       2      2                                        3
5        1       1      1                                        2
6        2       2      2                                        3
  Internal1.External2 Submit.Year Submit.Month Closed.Year Closed.Month
1                   1        2018            8        2018            9
2                   1        2018            8        2018            9
3                   1        2018            8        2018            9
4                   1        2018            8        2018            9
5                   1        2018            8        2018            9
6                   1        2018            8        2018            9
  Reported.Source.1.12 Assigned.Support.Organization.A1.S2.I3.O4
1                    1                                        1
2                    1                                        1
3                    1                                        1
4                    1                                        1
5                    1                                        1
6                    1                                        1
  Proactive1.Reactive2 Proactive.Reactive
1                    1          Proactive
2                    1          Proactive
3                    1          Proactive
4                    1          Proactive
5                    1          Proactive
6                    1          Proactive
```

Appendix 3: Summary of Attribute Distributions

```
> #Summarise attribute distributions
> summary(numeric2na1.df)
    Priority       Urgency          Impact      Resolution.Product.Category.Tier1.S1T2A3
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :0.000
 1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
 Median :1.00   Median :1.000   Median :1.000   Median :3.000
 Mean   :1.58   Mean   :1.535   Mean   :1.525   Mean   :2.557
 3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000
 Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :3.000
 Internal1.External2  Submit.Year    Submit.Month     Closed.Year    Closed.Month
 Min.   :1.000       Min.   :2012   Min.   : 1.000   Min.   :2012   Min.   : 1.000
 1st Qu.:1.000       1st Qu.:2014   1st Qu.: 3.000   1st Qu.:2014   1st Qu.: 3.000
 Median :2.000       Median :2015   Median : 6.000   Median :2016   Median : 7.000
 Mean   :1.728       Mean   :2016   Mean   : 6.169   Mean   :2016   Mean   : 6.444
 3rd Qu.:2.000       3rd Qu.:2017   3rd Qu.: 9.000   3rd Qu.:2017   3rd Qu.: 9.000
 Max.   :2.000       Max.   :2018   Max.   :12.000   Max.   :2018   Max.   :12.000
 Reported.Source.1.12 Assigned.Support.Organization.A1.S2.I3.O4 Proactive1.Reactive2
 Min.   : 1.000       Min.   :1.000                             Min.   :1.000
 1st Qu.: 3.000       1st Qu.:1.000                             1st Qu.:1.000
 Median : 6.000       Median :1.000                             Median :2.000
 Mean   : 5.393       Mean   :1.102                             Mean   :1.594
 3rd Qu.: 7.000       3rd Qu.:1.000                             3rd Qu.:2.000
 Max.   :12.000       Max.   :4.000                             Max.   :2.000
```

Appendix 4: Heatmap Code

```
> ggplot(numeric2na1.df, aes(Impact, Urgency))+
+   geom_raster(aes(fill = Priority))+
+   labs(title ="Heat Map: Impact, Urgency, Priority", x = "Impact", y = "Urgency")+
+   scale_fill_continuous(name = "Priority")
>
> ggplot(numeric2na1.df, aes(Closed.Month, Resolution.Product.Category.Tier1.S1T2A3))+
+   geom_raster(aes(fill = Impact))+
+ labs(title ="Heat Map: Closed Month, Product Category, Impact", x = "Month", y = "Product Categ$
+ scale_x_continuous("Closed Month", breaks = seq(0,12,by = 1))
>
> ggplot(numeric2na1.df, aes(Closed.Month, Reported.Source.1.12))+
+   geom_raster(aes(fill = Impact))+
+ labs(title ="Heat Map: Closed Month, Reported Source, Impact", x = "Month", y = "Reported.Sourc$
+ scale_x_continuous("Closed Month", breaks = seq(0,12,by = 1))
>
> ggplot(numeric2na1.df, aes(Proactive1.Reactive2, Reported.Source.1.12))+
+   geom_raster(aes(fill = Impact))+
+ labs(title ="Heat Map: Proactive/Reactive, Reported Source, Impact", x = "Proactive1.Reactive2"$
+ scale_x_continuous("Proactive/Reactive", breaks = seq(0,12,by = 1))
```

Appendix 5: Correlated Reported Sources to Numerical Values

1: BMC Impact Manager Event

2: Direct Input

3: Email

4: External Escalation

5: Other

6: Phone

7: Self Service

8: Systems Management

9: Voice Mail

10: Walk In

11: Web

12: Blank

Appendix 6: Histogram Code

```
> ggplot(numeric2na1.df, aes(Resolution.Product.Category.Tier1.S1T2A3)) + geom_histogram(binwidth$
+   scale_x_continuous("Product Category", breaks = seq(0,6,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Product Category")
>
> # Histogram Impact count
> ggplot(numeric2na1.df, aes(Impact)) + geom_histogram(binwidth = 1)+
+   scale_x_continuous("Impact", breaks = seq(0,4,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Impact")
>
> # Histogram Priority count
> ggplot(numeric2na1.df, aes(Priority)) + geom_histogram(binwidth = 1)+
+   scale_x_continuous("Priority", breaks = seq(0,4,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Priority")
>
> # Histogram Urgency count
> ggplot(numeric2na1.df, aes(Urgency)) + geom_histogram(binwidth = 1)+
+   scale_x_continuous("Urgency", breaks = seq(0,4,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Urgency")
>
> # Histogram Internal/External count
> ggplot(numeric2na1.df, aes(Internal1.External2)) + geom_histogram(binwidth = 1)+
+   scale_x_continuous("Internal/External", breaks = seq(0,4,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Internal/External")
>
>
> # Histogram Proactive/Reactive count
> ggplot(numeric2na1.df, aes(Proactive1.Reactive2)) + geom_histogram(binwidth = 1)+
+   scale_x_continuous("Proactive/Reactive", breaks = seq(0,4,by = 1))+
+   scale_y_continuous("Count", breaks = seq(0,85000,by = 2500))+
+   labs(title = "Histogram: Proactive/Reactive")
```

Appendix 7: Data for fig 30

Impact	Count of Incident ID
4-Minor/Localized	52087
3-Moderate/Limited	19301
2-Significant/Large	11113
1-Extensive/Widespread	626

Appendix 8: Data for fig 31

Product Categorization Tier 1	Count of Incident ID
APPLICATION LAYER	59230
SERVICE CATALOGUE	13750
TECHNOLOGY INFRASTRUCTURE	10147

Appendix 9: Data for all Incident Categories fig 32

Product Categorization Tier 3	Count of Incident ID
SAP APPLICATION	38315
OTHER APPLICATION	12347
BATCH PROCESS	5312
SAP SYSTEM	1654
ACCESS CHANNEL	1084
DOCUMENT MANAGEMENT TOOL	113
IT MANAGEMENT TOOL	106
MOBILITY APPLICATION	51
OTHER APPLICATION SYSTEM	45
SECURITY	42
APPLICATION ROLE	32
SAP APPLICATION COMPONENT	27
USER EXPERIENCE TRANSACTION	22
BUSINESS DATABASE APPLICATION	10
ON-LINE PROCESS	9
WEB APPLICATION	8
DOCUMENT MANAGEMENT SYSTEM	7
JAVA PLATFORM APPLICATION	7
DATABASE INSTANCE	6
BUSINESS INTELLIGENCE TOOL	5
WEB APPLICATION INSTANCE	5
	4
DATABASE	4
LOTUS NOTES APPLICATION	4
SAP APPLICATION INSTANCE	4
DESKTOP APPLICATION	2
IT MANAGEMENT SYSTEM	2
APPLICATION SUBROLE	1
JAVA PLATFORM SYSTEM	1
MOBILITY APPLICATION SYSTEM	1

Appendix 10: Count of Incident ID for Product Categorisation Tier 3 fig 33

Product Categorization Tier 3	Count of Incident ID
SAP APPLICATION	38315
OTHER APPLICATION	12347
APPLICATION SERVICE	7297
BATCH PROCESS	5312
CORPORATE SERVICE	2352
WINTEL SERVER	2212
LAPTOP COMPUTER	2089
UNIX SERVER	1891
SAP SYSTEM	1654
OTHER SUPPORT SERVICE	1627
INFRASTRUCTURE SERVICE	1579
DESKTOP COMPUTER	1546
ACCESS CHANNEL	1084
SHARED MAILBOX	799
DESKTOP APPLICATION PRODUCT	575
CORE BUSINESS SERVICE	529
PLATFORM SERVICE	307
DATABASE SERVER	271
MONITOR	232
LINUX SERVER	154

Appendix 11: Count of Incident ID per Country fig 34

Country	Count of Incident ID
United Kingdom	71216
Spain	9312
India	2846
United States	8
Germany	5
France	2
Korea, Republic Of	1
Netherlands	1