

Table of Contents

[INTRODUCTION 3](#_Toc23501073)

[DISCOVERY – Phase 1 5](#_Toc23501074)

[**Introduction and Background** 5](#_Toc23501075)

[**Project Aim and Objectives** 5](#_Toc23501076)

[**Scope** 6](#_Toc23501077)

[**Data Sources** 6](#_Toc23501078)

[**Risks** 6](#_Toc23501079)

[**Project Management** 8](#_Toc23501080)

[**Stakeholders** 9](#_Toc23501081)

[**Hypothesis** 9](#_Toc23501082)

[**Engagement with customer** 10](#_Toc23501083)

[ANALYSIS INTRODUCTION 12](#_Toc23501084)

[DATA PREPARATION – Phase 2 12](#_Toc23501085)

[**Collecting the Data** 12](#_Toc23501086)

[**Data Dictionary** 13](#_Toc23501087)

[**Data Conditioning** 14](#_Toc23501088)

[MODEL PLANNING – Phase 3 17](#_Toc23501089)

[**Data Exploration Analysis - Trend Charts** 17](#_Toc23501090)

[**Model Selection** 23](#_Toc23501091)

[MODEL BUILDING – Phase 4 25](#_Toc23501092)

[**ANOVA (Analysis of Variance) Model** 25](#_Toc23501093)

[**Further Analysis - Model Selection** 26](#_Toc23501094)

[COMMUNICATE RESULTS – Phase 5 34](#_Toc23501095)

[**Cluster Plot** 34](#_Toc23501096)

[**Time Series Plot** 37](#_Toc23501097)

[**Interactive Dashboard** 37](#_Toc23501098)

[**Summary and Conclusion of the Results** 40](#_Toc23501099)

[OPERATIONALIZE – Phase 6 41](#_Toc23501100)

[**Prepare for Production Environment** 41](#_Toc23501101)

[**Final Deliverables** 42](#_Toc23501102)

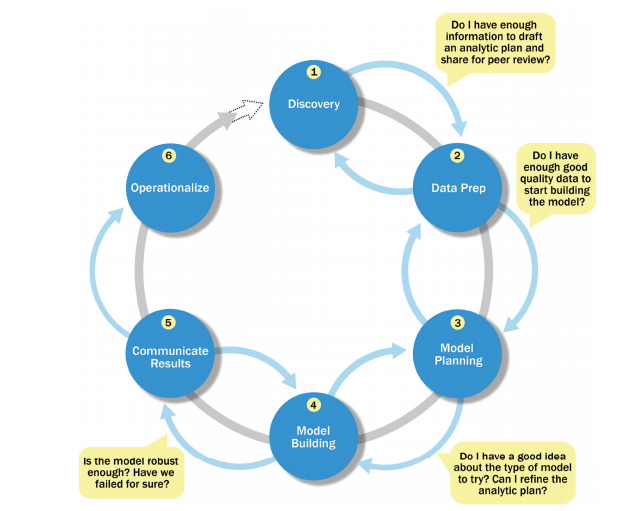
# INTRODUCTION

My name is Marlon Ducille and I works as an Application Developer at Trafford Council. My work at Trafford Council involves:

* Developing web sites, web-based applications, APIs (Application Programming Interface), GIS systems and integrated technologies to allow the council to conduct its business processes
* Developing and managing database systems.
* Supporting innovative and new ways of working across the council and with partners

This document is my Synopsis Project for the Data Analytics Level 4 Apprenticeship program. I have chosen Project B – NHS GP Prescription Analytics.

For this project, I have adopted the Data Analytics Lifecycle to conduct the analysis.



I have conducted a number of analysis for this project, each uses this lifecycle. The Discovery phase, though, is common for all the analysis in this project.

**Discovery – Phase 1**. In the discovery phase, I described the situationand context of the project. It shows my understanding of the business domain, the problem I need to solve and the timescale for delivery. I demonstrated the resources and technology I need to complete the project. I have demonstrated how I engaged with customers and used requirement elicitation process such as User Stories. I have shown how I planned the project and mitigated potential risks.

**Data Preparation – Phase 2.**  In the Data Preparation phase, I have shown how I have downloaded, cleansed and formatted the data, and getting it is a state ready for analysis.

**Model Planning – Phase 3.**. In the Model Planning phase, I determined the methods, techniques, and workflow I intended to follow for the subsequent model building phase, by data exploration. After exploring the data, I then decided on the model to select.

**Model Building – Phase 4**. In the Model Building phase, I have shown how I have used routine statistical analysis, and a range of analytical techniques such as data mining to build the model, and identify trends and patterns in the data.

**Communicate Results – Phase 5.** In the Communicate Results phase, I have translated analyses into insight, extracted meaningful pieces of information from the charts and tables to communicate to stakeholders.

**Operationalize – Phase 6.**  In the Operationalize phase, I have produced presentation, summarising and making recommendation to stakeholders

# DISCOVERY – Phase 1

## **Introduction and Background**

I works at Trafford Council with a growing IT Service which is looking to expand its portfolio.

The organisation is experiencing market saturation and would like to increase their service offering in order to maintain a competitive edge.

The Healthcare industry is a direction in which it has identified a possible route of expansion. Recently, Trafford council has merged with the NHS Trafford Clinical Commissioning Group (CCG).

Trafford CCG is a member organisation, consisting of Trafford’s 32 GP practices that holds the NHS budget for the borough and buys a range of healthcare services on behalf of residents.

The aim of the CCG is to ensure that the Trafford population has the best possible healthcare outcomes by buying high quality, equitable and integrated services.

It has three aims:

* To improve population health;
* To improve the care provided and the healthcare experience of individuals; and
* To lower the costs of providing the above.

As a result of this merge, Trafford Council have recently recruited a Chief Digital Officer whose role is to develop the IT strategy for Trafford ICT and CCG.

The Senior Management team have scheduled an initial meeting in a week’s time with NHS Commissioners. At the meeting, the intention is to hold an informed conversation demonstrating the area’s in which the organisation can provide cost savings; against GP prescriptions in particular.

## **Project Aim and Objectives**

The aims and objectives of this project is to:

* Produce an intelligent report which will assist the Senior Management team in their preparation for a meeting, which will demonstrate the areas in which savings can be made against GP prescriptions. The board believe that this initial report will throw open a wide door of opportunities in the NHS for the organisation and put them firmly on the map in the UK as a whole.
* Understand whether there is distinct difference between GP Practices with regards to prescription cost.
* To aid the commissioners in making a decision about potential cost savings and proper channelling of funding.
* Improve customer’s user experience and overall cost saving by helping commissioners in making a decision about potential cost savings and proper channelling of funding.

## **Scope**

Timescale

1 week (Monday to Friday)

Data to be taken from the London region between July 2018 and December 2018

Project deliverable

The report should include:

* A summary of the current cost of prescriptions.
* Analyses that demonstrate an understanding of GP prescription costs.
* A 6 months’ trend analysis.
* A summary of the ‘Area Team’, ‘PCO’ and ‘Practice’ views of the cost of prescriptions (Area Team, PCO and Practice are all metrics in the dataset).
* Be a ‘live’ report that can be updated when new data is released to allow the latest view of prescription cost is available.

Resources

The resources required for this project are: 1x Lead Data Analyst, R Studio, Microsoft Excel, Microsoft Word, Microsoft PowerPoint, Laptop/PC running Windows Operating system, Internet Connection, Project Management tool (Trello), GitHub, PowerBI

## **Data Sources**

The data source for this project is from the NHS Business Services Authority (NHSBSA)

<https://apps.nhsbsa.nhs.uk/infosystems/welcome>

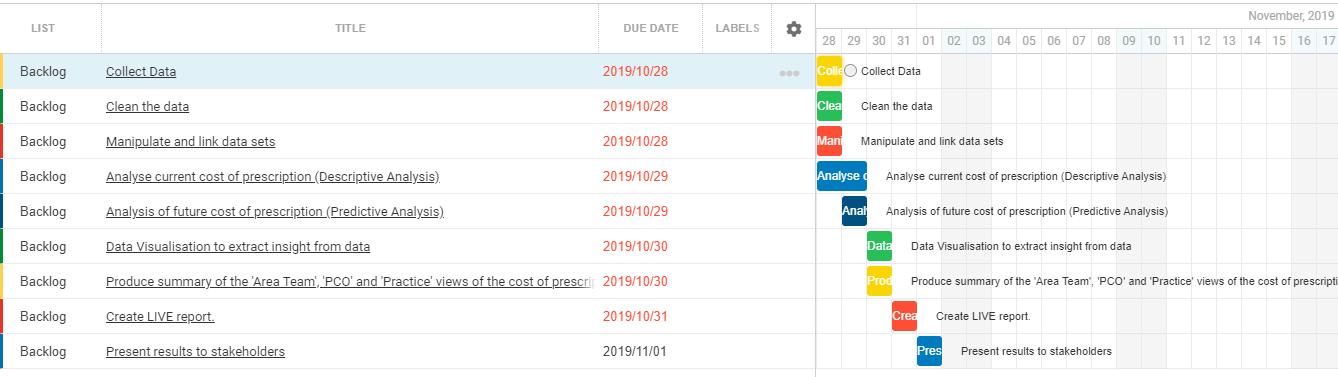
## **Risks**

Potential risks that may prevent or delay the delivery of this project are:

|  |  |  |
| --- | --- | --- |
| **Risks** | **How to mitigate the risks** | **Level (1 – 5)** |
| PC/laptop failure, which means I am unable to do the work | Have a another PC/laptop on standby | 4 |
| Network failure, which means unable to download data. | Take my work home or at another location where there is network available and download the data. | 4 |
| Losing my work | Regularly take a backup of my work so that work is not lost | 5 |
| Project is not delivered on time | Put a robust Time Management Plan in place and invest in more resources (i.e. put in more hours)  Beware of time wasters (unplanned meetings, distractions, unnecessary re-works) and works within the scope of the project | 4 |

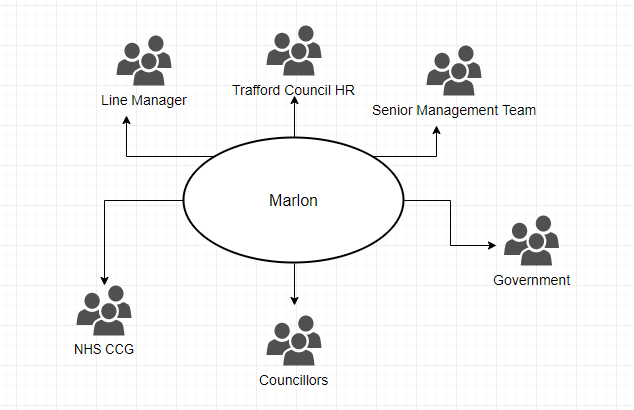
## **Project Management**

In order to manage the project effectively and to complete it within the desired timescale, I planned my work by creating Gantt chart in Trello



## **Stakeholders**

The stakeholders for this project is as follows:

**Assumption**

The assumptions are:

* I have already met with the customers to gather the requirements
* Although I work for Trafford Council, the report will be focused on London and a region (it being the capital). The assumption is that, this will be mirrored for other England regions after testing, along with a review of feedback from the target audience.

**Hypothesis**

Null Hypothesis: There is NO significant difference between GP Practices with regards to prescription cost.

Alternative Hypothesis: There is at least ONE GP Practices that is significant different with regards to prescription cost.

**Engagement with customer**

The organisation’s unique selling proposition is improving customer’s user experience and overall cost saving. In order to improve customer’s user experience, I arranged a series of meetings with the customer in order to gather their requirements. This way I can more effective deliver a solution with excellent user experience.

**Activity log**

|  |  |
| --- | --- |
| **Activity** | **Date** |
| I approached the Chief Digital Officer, for Trafford Council and CCG ICT department, to advise me as to where I can get more clarity on the CCG datasets.  He directed me to the Information Manager for CCG | July 17, 2019 |
| I had a meeting with the Information Manager for CCG where I gathered better understanding of the CCG datasets. | July 17, 2019 |

After meeting with the customer, I gathered the requirements and documented them in a user story.

**User Story – Requirement Elicitation**

Given I am business user

I should be able to view a report, showing:

* + A view of July – December, 2018 dataset in one table so to have the ‘Actual Cost’ metric in six columns in the final data set, differentiated by their months (Jul, Aug, Sep, Oct, Nov and Dec).
  + A trend chart showing GP prescription cost (Actual Cost) at Area Team level (Jul-18 to Dec-18).
  + A trend chart showing GP prescription cost (Actual Cost) at PCO level (Jul-18 to Dec18).
  + A trend chart showing GP prescription cost (Actual Cost) at GP Practice level (Jul-18 to Dec-18).
  + The top ten GP Practices with highest prescription cost.
  + The top ten GP Practices with highest quantity of prescription.
  + A comparison of the top ten GP Practice with highest cost and quantity of prescription

Given I am business user

When I am viewing a report

Then I should be able to compare prescription costs across GP practices

Given I am business user

When I am viewing a report

And comparing prescription costs across GP practices

Then I should be able to make a decision on the cheapest alternative to prescribe

Given I am business user

When I am viewing a report

Then I should be able to compare prescription costs for ‘Area Team’, ‘PCO’ and ‘Practice’

Given I am business user

Then I should be allowed to easily copy content for inclusion in other presentations.

Given I am business user

Then I should be able to view the report indicating which period the latest data covers.

# ANALYSIS INTRODUCTION

As I mentioned above, the purpose of NHS CCG is to procure the best quality health services to meet the needs of the residents. They are also responsible for setting a prescribing budget against each GP practices.

There are many medications that are prescribed, some more expensive than others. Since budget is a constraint. The purpose of this analysis is to determine if there are cheaper alternatives to some of the more expensive brands.

The purpose of this analysis is to compare prices and to test if there are any difference between GP Practices with regards to prescription cost.

# DATA PREPARATION – Phase 2

**Collecting the Data**

After my meeting with the customer and gathering the requirements, I then planned the steps in order to gather the necessary data for analysis. Since the customer need to view a 6 month trend analysis of prescription cost between July 2018 and December 2018, I approached the collection of data in the following way:

* I logged unto <https://apps.nhsbsa.nhs.uk/infosystems/welcome>
* I then selected ‘Prescribing Data’
* Then I selected the ‘Practices Prescribing Information – within Regional Office’ Data View
* The customer required that the report focuses on London as a region. For this reason, I selected’ London – Y56’
* Then I choose the period for July 2018
* I then repeated these steps for August, September, October, November and December 2018

After I have downloaded the data, I renamed the files accordingly, for easy readability and identification.

Jul 2018 Detailed\_Prescribing\_Information

Aug 2018 Detailed\_Prescribing\_Information

Sep 2018 Detailed\_Prescribing\_Information

Oct 2018 Detailed\_Prescribing\_Information

Nov 2018 Detailed\_Prescribing\_Information

Dec 2018 Detailed\_Prescribing\_Information

**Data Dictionary**

The fields contained in the **Detailed Prescribing Information** dataset are as follows:

* Regional Office Name – this is referring to east, west, north, south of the UK
* Regional Office Code
* Area Team Name – This is referring to the County
* Area Team Code
* PCO Name - The term Primary Care Organisations is used to define CCGs, Local Authorities, Commissioning Hubs and provider organisations or a combination of these organisations in relation to the information in reports supplied by NHS Prescription Services.
* PCO Code – This is the code for the Primary Care Organisations
* Practice Name - A GP practice is an organisation with one or more GP's acting in partnership. For the purposes of the NHS Prescription Services Information Systems, any organisation that shows information down to prescriber level can be set up as a practice. This includes Cost Centres for PCOs.
* Practice Code - This is the code for the Practice
* BNF Code – British National Formulary Code. The NHS Prescription Services uses additional pseudo BNF chapters 18-23 to classify: Preparations used in Diagnosis, Other Drugs and Preparations, Dressings and Appliances. These chapters are structured into sections, paragraphs, etc in the same way as the standard BNF chapters. This is a fifteen character code, based on the British National Formulary (BNF) classifications
* BNF Description - The name of the drug as stated in the British National Formulary (BNF)
* Items - Items is the number of times a product appears on a prescription form
* Quantity - Normally the quantity prescribed.
* ADQ Usage - The Average Daily Quantity value for a presentation an analytical unit used to compare prescribing activity
* NIC - Basic price of a drug
* Actual Cost - The actual cost is the basic price of the drug less discount based on the National Average Discount Percentage (NADP) from the previous month plus the container cost.

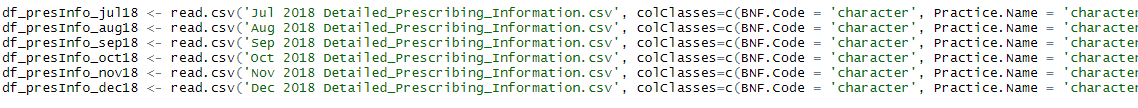
**Data Conditioning**

The data I collected consisted of some fields that are relevant for the analysis, and at the same time, fields that are irrelevant. There are also 6 different datasets which are not in the correct format for analysis.

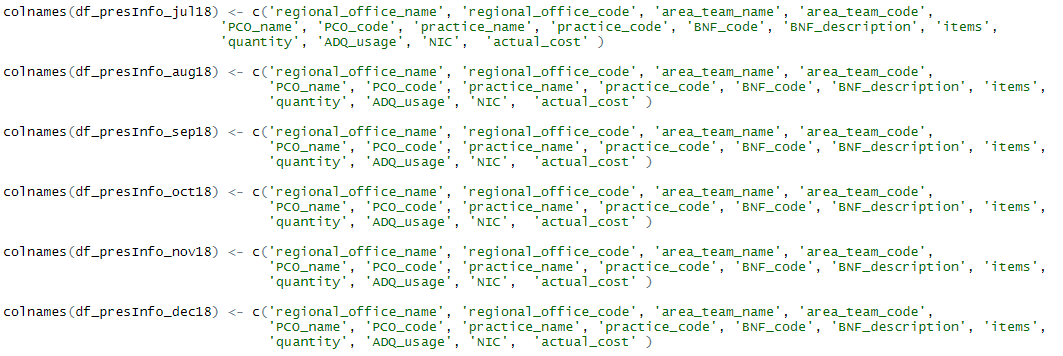
For this reason, I performed some data conditioning to clean and transform the data. I also merge datasets together to get the data in a state ready for analysis

The steps I have taken are as follows:

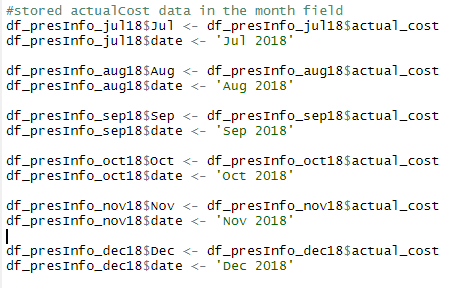
* Used R Studio to read the 6 datasets. I have used the parameter, ‘colClasses’, to convert BNF Code, Practice Name, PCO.Name and BNF Description into ‘character’ data types



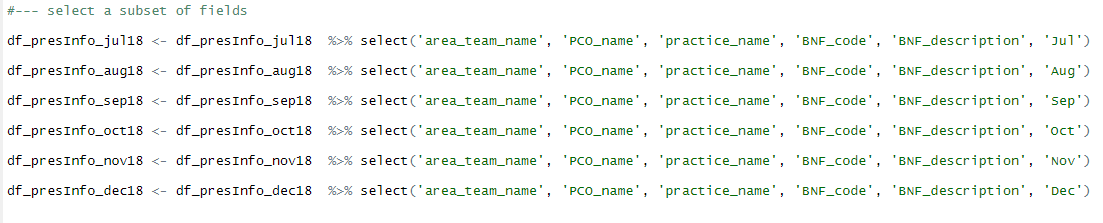
* I then renamed the columns for easy readability



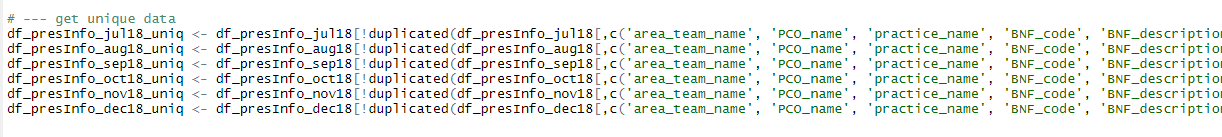
* In order to conduct a 6 month trend analysis, I then created the variables ‘Jul’, ‘Aug’, ‘Sep’, ‘Oct’, ‘Nov’, ‘Dec’ to the dataframes, df\_presInfo\_jul18, df\_presInfo\_aug18, df\_presInfo\_sep18, df\_presInfo\_oct18, df\_presInfo\_nov18, df\_presInfo\_dec1 respectively. I then assigned the ‘Actual Cost’ to these variables



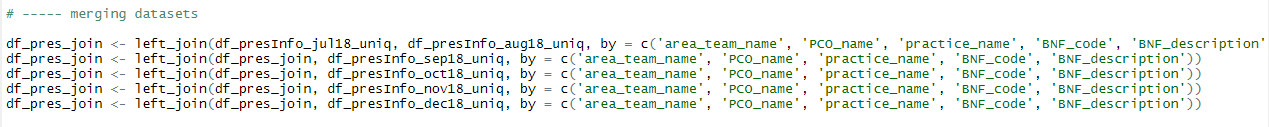
* Since only some of the fields are relevant for the analysis, I then selected a subset of fields for the analysis.



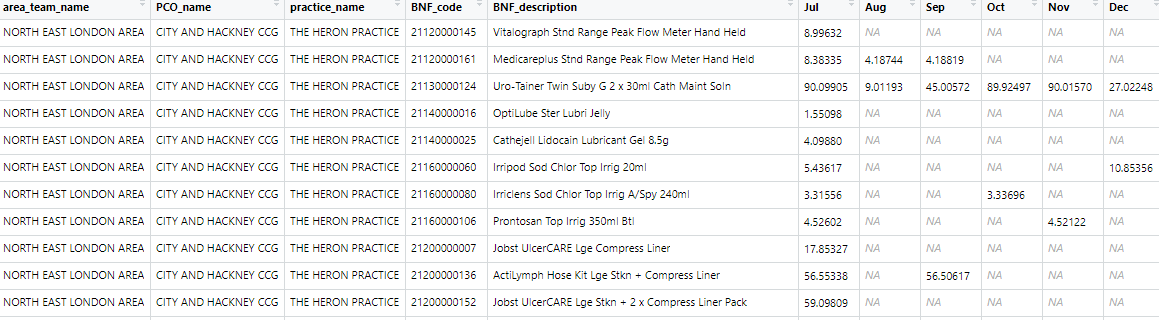
* I then did some data cleansing by getting rid of duplicate data



* I then merged the datasets using the following code



The result output after I ran the code is as follows.

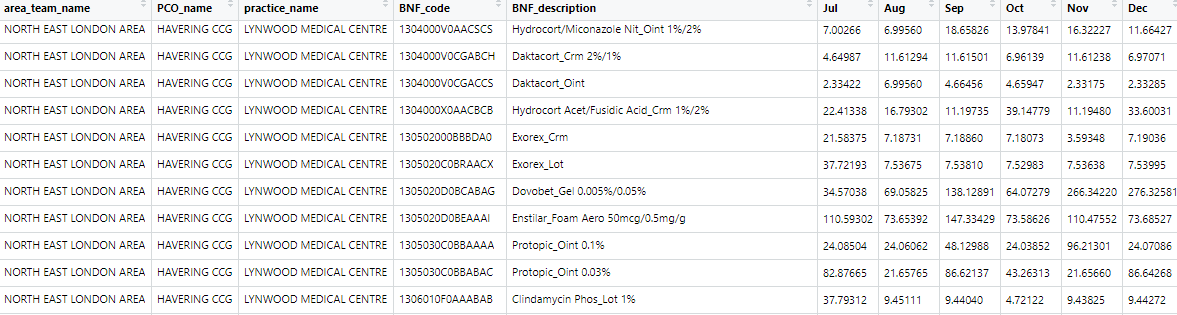


From this dataset, I observed lots of missing data (indicated by NA values), which require further cleansing.

* I then cleansed the data, using the code.



The result output after I ran the code is as follows.



The data is now formatted, cleansed and is now in a state to conduct the analysis.

# MODEL PLANNING – Phase 3

**Data Exploration Analysis - Trend Charts**

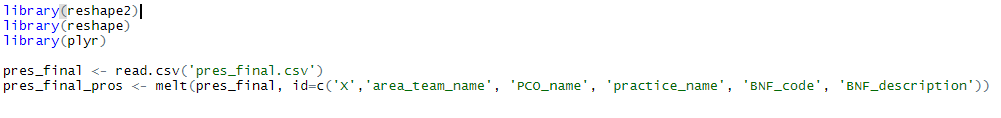
In order to better understand the pattern in the data between July and December 2018, I first needed to explore the data, using trend charts between July and December 2018.

First I transformed the data further in order to get the dates (from July to December) in one column, and their associate values in a separate column. Having the data in this format is easier display trends in the PowerBI tool. I did this using the following steps:

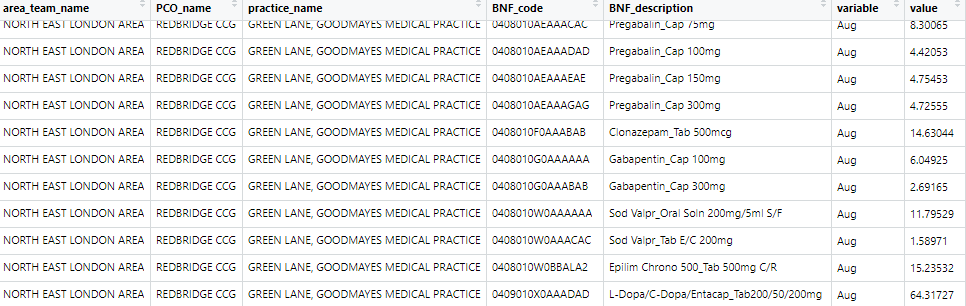
* I first saved the dataframe into a csv file.



* I then used the ‘melt’ function in R, from the ‘reshape’ library to transform the data.

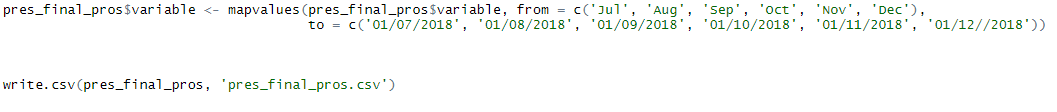


This function transforms the current dataframe. The result is as shown.



This now stores the months in a column called ‘variable’, and the actual costs in a column called ‘value’

* I then transform the month name to their corresponding dates, using the ‘mapvalues’ function in R, under the ‘plyr’ library. For example, I convert ‘Jul’ to ‘01/07/2018’



I then saved the resulting dataframe, ‘pres\_final\_pros’ into a csv file, which I then imported into PowerBI.

* After importing the file into PowerBI, I renamed the ‘variable’ column and ‘date’ and the ‘value’ column and ‘actual\_cost’

**Trend chart showing GP prescription cost (Actual Cost) at Area Team level (Jul-18 to Dec-18).**

In PowerBI, I created a Clustered Column chart, with the ‘area\_team\_name’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

I then created a Line chart, with the ‘date’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

The result is as shown below.



The visual is very interactive. Whenever I selected an Area Team Name on the column bar, the trend line changes accordingly. In this case, I selected on the North East London Area on the bar chart, and therefore the trend line, on the right, showing the trend for North East London Area CCG.

From this graph, I have observed that the actual cost for North East London Area in July was just under £8.5 million. It then decreases slightly in August, and drops further to its lowest in September, before rising again to its highest in October at just over £8.5 million, before dropping again in November and rises slightly in December.

**Trend chart showing GP prescription cost (Actual Cost) at PCO level (Jul-18 to Dec-18).**

In PowerBI, I created a Clustered Column chart, with the ‘PCO’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

I then created a Line chart, with the ‘date’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

The result is as shown below.



The visual is very interactive. Whenever I selected a PCO on the column bar, the trend line changes accordingly. In this case, I selected on the Havaring CCG on the bar chart, and therefore the trend line, on the right, showing the trend for Havaring CCG.

From this graph, I have observed that the actual cost for Havaring CCG in July was just under £850k. It then increases in August, before dropping to its lowest in September, before rising again to its highest in October at just under £875k, before dropping again in November and rises slightly in December.

**Trend chart showing GP prescription cost (Actual Cost) at Practice level (Jul-18 to Dec-18).**

In PowerBI, I created a Clustered Column chart, with the ‘practice\_name’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

I then created a Line chart, with the ‘date’ variable as Axis on the x-axis and ‘actual\_cost’ variable as value on the y-axis.

The result is as shown below.

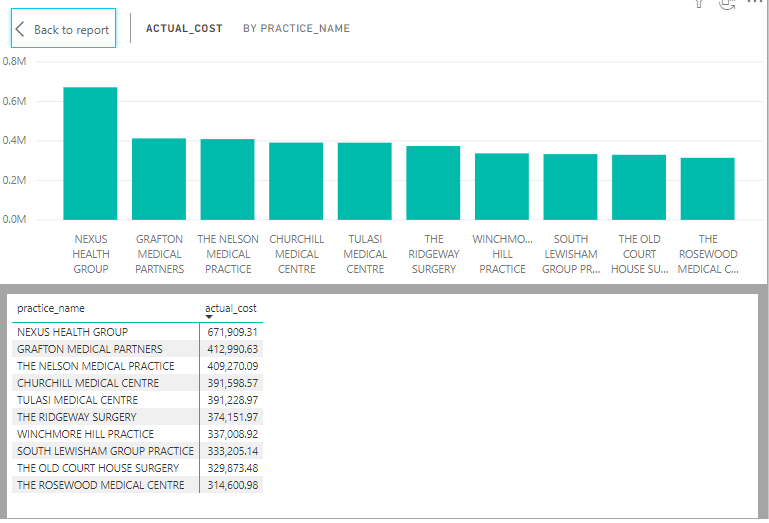


I have selected on Nexus Health Group practice on the bar chart, and therefore the trend line, on the right, showing the trend for Nexus Health Group practice

From this graph, I have observed that the actual cost for Nexus Health Group practice in July was just about £114k. It then decreases in August, before dropping slightly further to its lowest in September, before rising again in October and then to its highest in at just under £875k in November, before dropping again in December.

**Top 10 GP Practices with highest prescription cost.**

In PowerBI, I created a filter on the ‘practice\_name’ field in order to select the top 10 practices based on actual cost. The result is as shown below.

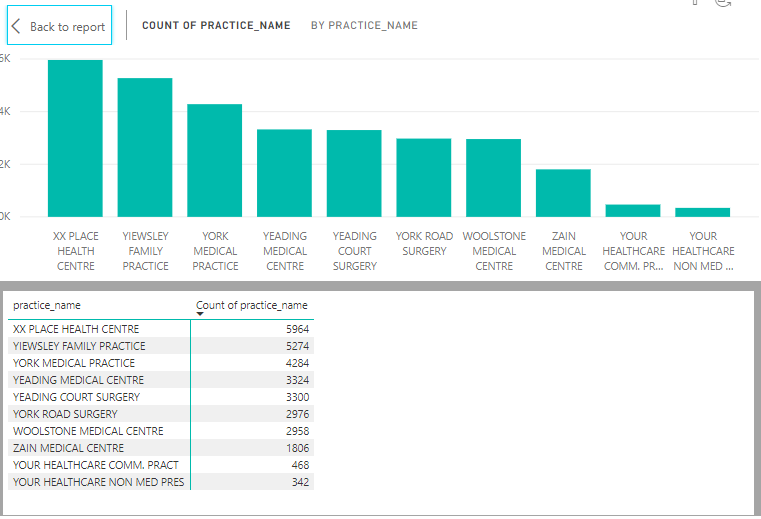


Based on the data, I have observed that the top 10 GP Practices based on costs are:

1. Nexus Health Group
2. Grafton Medical Partners
3. The Nelson Medical Practice
4. Churchill Medical Centre
5. Tulasi Medical Centre
6. The Ridgeway Surgery
7. Winchmore Hill Practice
8. South Lewisham Group Practice
9. The Old Court House Surgery
10. The Rosewood Medical Centre

**Top 10 GP Practices with highest quantity of prescriptions**

In PowerBI, I created a filter on the ‘practice\_name’ field in order to select the top 10 practices based on practice\_name. The result is as shown below.



Based on the data, I have observed that the top 10 GP Practices based on quantity of prescriptions are:

1. XX Place Health Centre
2. Yiewsley Family Practice
3. York Medical Practice
4. Yeading Medical Centre
5. Yealding Court Surgey
6. York Road Surgery
7. Woodstone Medical Centre
8. Zain Medical Centre
9. Your Healthcare Comm Practice
10. Your Healthcare Non Med Pres

**Model Selection**

Now that I have the top 10 practices according to their actual cost, my next task is to determine if there are any significant differences between these practices. I have chosen to ANOVA (Analysis of Variance) model. I have chosen this model because it’s a good model to use for testing differences between multiple groups.

Although visually, I can clearly see that the prescription cost for Nexus Health Group is different to the others, and the others appear reasonably similar, the ANOVA test will indicate whether or not they are significantly different.

Before I can use the ANOVA test though, I first need to test if the data is normally distributed because ANOVA test works on normally distributed dataset.

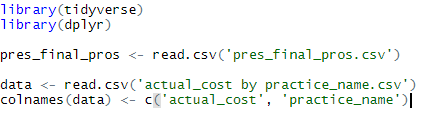
One of the best methods for testing for normality is the Shapiro-Wilk test. I set up my hypothesis in the following way.

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is NOT normally distributed

I tested for normality by using the Shapiro-Wilk test, and then checking whether or not the p-value is greater than 0.05. A p-value greater than 0.05 will indicate normality. I did this in the following way:

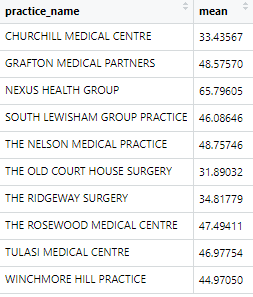
* I exported the top 10 practices according to actual cost, in PowerBI, into a csv file. This is called ‘actual\_cost by practice\_name.csv’
* I used R Studio to read from this file, as well as, the ‘pres\_final\_pros.csv’ file



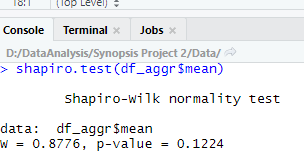
* I then filtered the pres\_final\_pros data to only include the top 10 practices by cost, then aggregate the data based on the mean.



The result after the after the aggregation is as follows:



* I then ran the Shapiro-Wilk test using the code below.



The result above shows a p-value of 0.1224 which is greater than 0.05. This means that I do not reject the null hypothesis, and therefor conclude that the data is normally distributed.

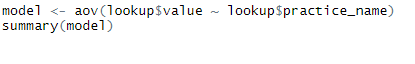
I can now proceed in building the model using ANOVA.

# MODEL BUILDING – Phase 4

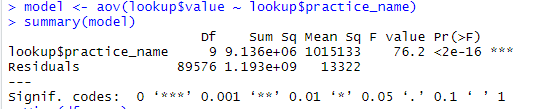
**ANOVA (Analysis of Variance) Model**

As mentioned above, in order to determine if there are significant difference between prescription prices across GP practices, I need to build an ANOVA model.

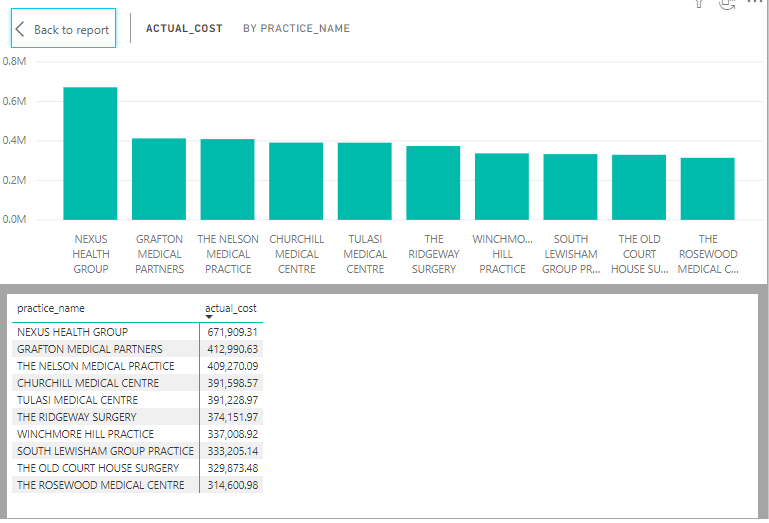
I build the ANOVA model using the ‘aov’ function in R. See code below



The result after building the model is as follows:



From this result, I have observed a very low p-value which is less than 0.05. This means that I reject the null hypothesis and accept the alternative hypothesis that there is at least ONE GP Practices that is significant different with regards to prescription cost.



Looking at the visualisation I created earlier, I can clearly see that Nexus Health Group has significant prescription cost compare to the other practices, and the ANOVA test has confirmed that.

**Further Analysis - Model Selection**

Since Nexus Health Group is significant in terms of prescription costs, I decided to conduct some further analysis on the data specific to Nexus Health Group.

To do this I needed to decide on cheaper alternative prescription I can recommend to stakeholders. In order to do this, I need to categorise these drugs in order to understand and determine which are cheap and which are expensive.

I decided to use the k-means clustering algorithm. I choose the k-means clustering because it is an ideal algorithm that is used for grouping objects. K-means clustering algorithm is also an unsupervised learning algorithm, which means it doesn’t rely on a target or labelled variable to train the dataset, as opposed to others like Naïve Bayes, or Random Forrest. Since my current dataset doesn’t have a labelled dataset, k-means clustering is the ideas algorithm.

K-Means Clustering

Now that I have decided to use k-mean clustering, I built the model using R Studio, in the following way:

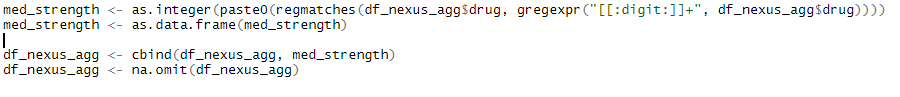
* I first read from the 'pres\_final\_pros.csv' dataset and filtered on ‘Nexus Health Group’



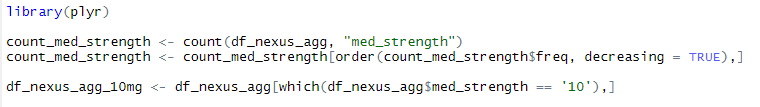
* I then aggregated the data to extract only the unique drugs. Then I sorted the data.

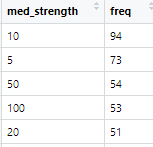


* The resulting dataframe returned was so huge that I decided to select only a subset of the data. My aim is to compare drug prices that belong to the same medication strength, i.e. 10mg are compared together, 20mg are compared together, etc. In order to perform an accurate comparisons of prescriptions, it makes sense to compare prescriptions with the same medication strength. For this reason, I have used R script to extract this information from the ‘drug’ field, using regular expression, and then cleaned the data (getting rid of NA). The code is as follows:



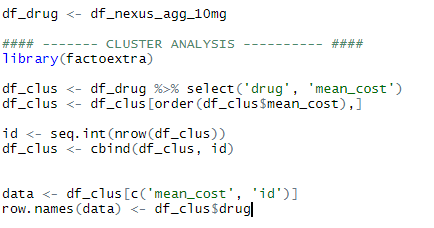
* I then wrote some script to determine the medication strength that occur the most.



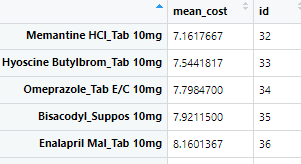


I have found that the 10mg prescriptions occur the most, so I have filtered the data based on 10mg.

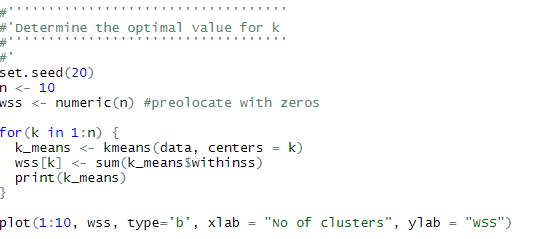
* I used the factoextra library to create the visualisation in R
* I then created an id column in the dataset. I did this in order to create a 2-D view to the map
* I then created the drug column as the index. I did this so that the drug’s name are displayed within the cluster they belong



The dataframe, again shows two columns, mean\_cost and id, with drugs being the index.



* Before applying the clustering algorithm, I need to first compute the number of clusters to use. To do this, I plotted the WSS (Within Sum of Squares) against the number of clusters. I choose 1 to 10 clusters to create the plot.





* I then used Elbow method to determine the number of clusters to use. Looking at the graph, I observed that the number of clusters where there is very minimal changes to the wss, is 4. Therefore, I have chosen 4 clusters to build my model.
* I then used the kmeans function in R to build the model.



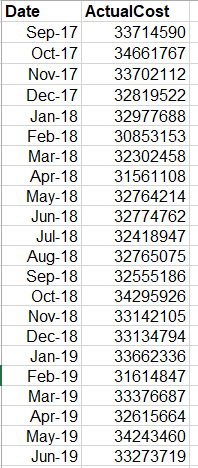
* I then used the fviz\_cluster function to visualise the data.



Time Series Forecasting Analysis

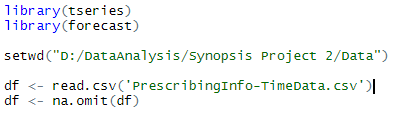
Given that I have acquired some time data between July 2018 and December 2018, I was interesting and curious to see what the prescription cost will be in the future. In order to do this, I needed to conduct some time series analysis in order to predict and forecast data in the future.

The current set of data that I collected earlier is too small a sample to do any real analysis. For this reason, I have collected additional data between September 2017 and July 2019, and used Excel to tally the total actual costs of prescriptions for each year. The data is as follows.



I then used R Studio to plot the data and forecast future prescription cost for the next 12 months. I have used the tseries and forecast libraries in R. The steps I have conducted the analysis is as follows:

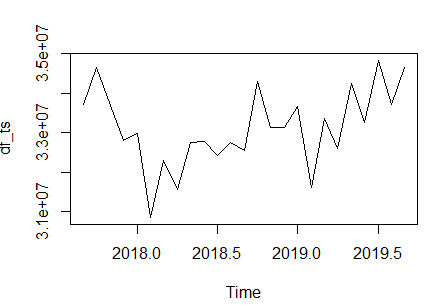
* Read the csv file into a dataframe and clean the data.



* Transformed the data into a time series starting from September 2017 to September 2019, with a frequency of 12 months.

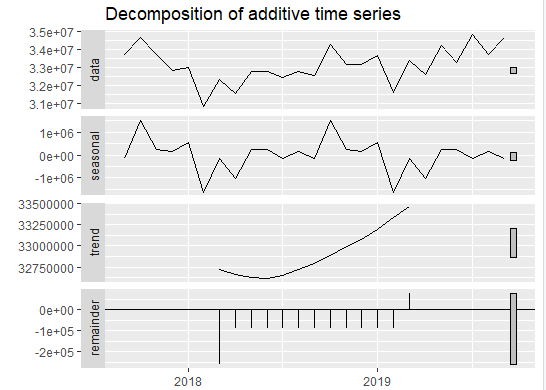


* After I plotted the data, the result is as follows.



* I then decompose the time series data to analyse the seasonality, trend and the residue from the data. The R code and resulting graph is shown below.





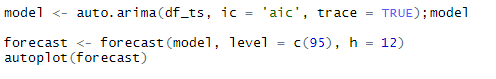
From the graph I can observe an exponential, increasing trend, from about the middle of 2018.

I have also observe a consistency in the seasonality, i.e. every year, the cost of prescription dips to its lowest in February. It fluctuates, while at the same time, increasing to its highest cost in October and decreases again in February. And the cycle continues.

The remainder or residue is what remains after the trend and seasonality is removed, and is used to make further predictions. It can also be used to detect anomalies and outliers.

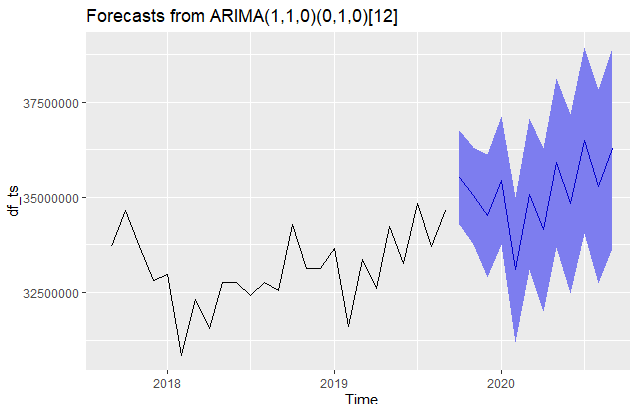
In order to forecast future data, the time series data need to be stationary. **Stationary time series** means that the time series data has a constant mean and variance. This occur when the trend and seasonality is removed from the data.

To achieve this, I used an **ARIMA** model, indicated by ARIMA(p,d,q) for non-seasonal data, or ARIMA(p,d,q)(P,D,Q)[m] for seasonal data.



I have made the prediction using a 95% confidence interval. This means that we are 95% confident that the predicted value will be within this range.

The resulting graph after forecasting is as follows.

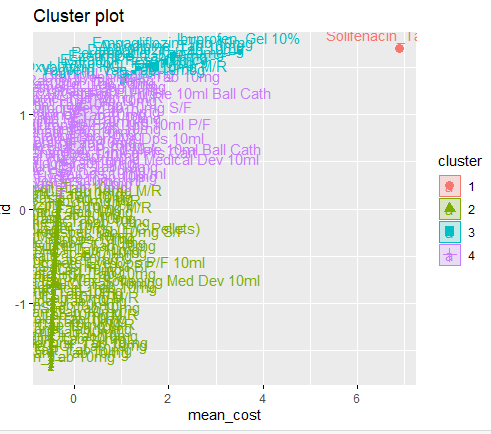


The blue trend line shows what has been predicted. The shaded blue area indicates the possible range of values, the predicted data could lie within, using a 95% confidence interval. This means that I am 95% confident that the predicted data could fall within this range.

# COMMUNICATE RESULTS – Phase 5

## **Cluster Plot**

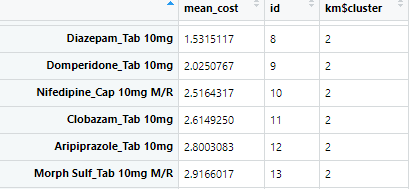
After I have built the model using k-mean clustering algorithm, I have created a visualisation of the clusters, with the drugs within each clusters.



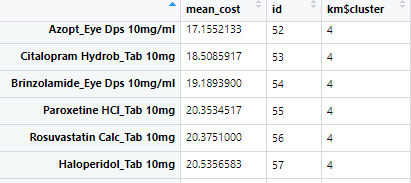
From the cluster plot, I can observe the 4 clusters the drugs are grouped under. The area in red is where the actual price is the highest and therefore is the group of the most expensive drugs. The area in green is the group of the cheapest drugs, with the lowest actual cost values. The area in purple is reasonably cheap as well. For this reason, I would recommend cheaper drugs within these groups as an alternative to the more expensive ones.

I have also created a dataframe to store the cluster number for the relevant drug.

For cluster number 2, highlighted by the green region, as subset of the dataset is shown below. This is the cheapest region.

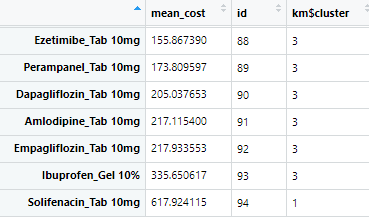


For cluster number 4, highlighted by the purple region, as subset of the dataset is shown below. This is the next cheapest region.



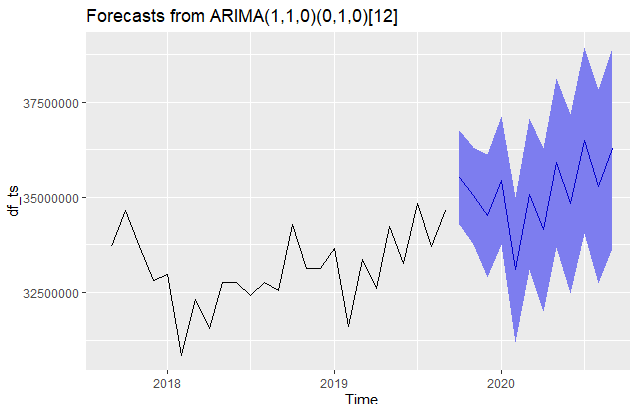
For cluster number 1, highlighted by the red region, as subset of the dataset is shown below. This is the most expensive region.

For cluster number 3, highlighted by the blue region, as subset of the dataset is shown below. This is the next most expensive region.



Having conducting my analysis and training the London dataset to group drugs into categories, I can therefore use this model to predict grouping for other drugs within London and other England regions.

**Time Series Plot**



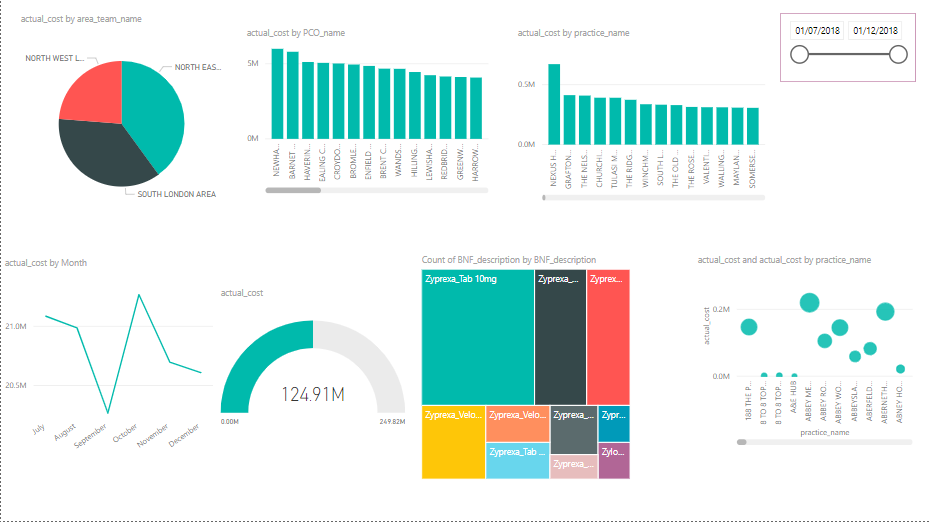
Looking at this visualisation, I have observed a rise in prescription cost in the next 12 months. This is a cause for concern, and I would flag this up to the Senior Management Team and NHS CCG to put things in place to mitigate the increase in and the cost it puts on the NHS.

**Interactive Dashboard**

In order to summarise my findings, I created an interactive dashboard that brings all the visualisation together. I have created the dashboard in PowerBI. It consisted of:

* Pie Chart showing the proportion of actual prescription cost at the Area Team level.
* Bar chart showing the distribution of actual prescription cost at the PCO level
* Bar chart showing the distribution of actual prescription cost at the Practice level
* Tree Map chart. This shows the hierarchal view of the number of times each drug is prescribed, starting with the largest value. It uses nested rectangles with the size of rectangle vary based on the number of times a drug is prescribed. So the more a drug is prescribed, the larger the rectangle.
* Scatter chart. This displays the practice name on the x-axis and the actual cost on the y-axis. I have also changed the size of the bubble so that it is proportional to the actual cost of prescription. So the larger the cost, the bigger the bubble on the chart.
* Line Graph showing the trend of prescription cost between July 2018 and December 2018. As I mentioned before, this is very interactive. So whenever I click on a particular Area Team on the pie chart, PCO name or Practice on the bar chart, drug on the Tree Map, or Practice on the scatter plot, the line graph changed accordingly, showing the trend for that specific data that I selected.
* Gauge. This also displays the actual cost. Like the line graph that displays changing time series, based on the data I have selected, this displayed the actual cost for that specific data that I selected
* Slicer. I have created a slicer to create an even more ‘live’ interactive which varies from July 2018 to December 2018. Each time I changed the slicer, the data on the graph is displayed for that specific date range. This will also indicate which period the latest data covers.

A screenshot of the dashboard is shown below.



**Summary and Conclusion of the Results**

After completed my analysis I concluded the following. During the period of July 2018 and December 2018, within the London region.

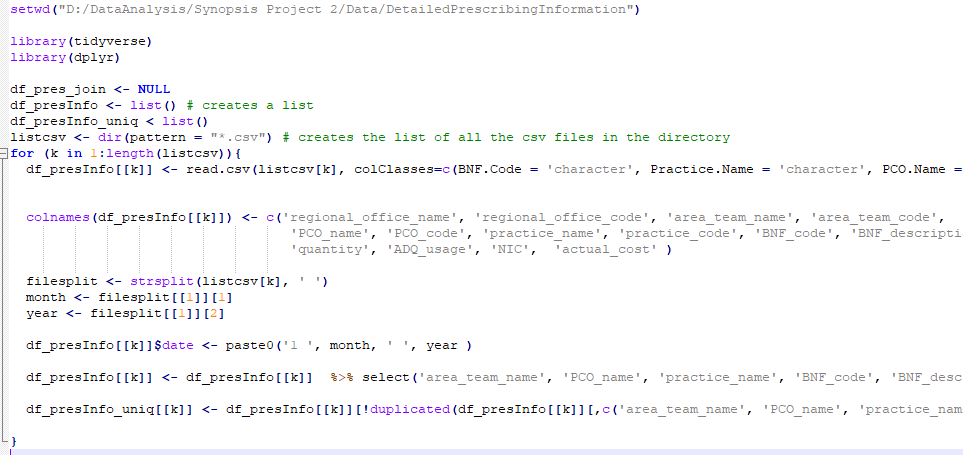
* North East London Area has occur the highest prescription cost of 40% , followed by South London Area of 36%, with lowest spending attributed to North West London Area of 24%.
* The PCO which cost the NHS the most is Newham CCG, with a cost of just under £6M.
* The GP prescription cost is at its highest in October and at its lowest in September.
* I have predicted and forecasted that there will be an exponential increase in prescription cost to the NHS within the next 12 months, if extra measures are not put in place.
* The general practice which cost the NHS the most is Nexus Health Group of £6.7k. I have also derived that there is at least ONE GP Practices that is significant different with regards to prescription cost. The practice though that prescribed the most drugs is XX Place Health Centre. When I compared the top ten GP Practice with highest cost and quantity of prescription, I observed that these practices are quite different. This means that those that those that made the most prescriptions not necessarily cost the NHS the most and likewise those that cost the NHS the most don’t necessarily prescribed the most drugs.
* The drug prescribed the most is called Zyprexa 10mg. This drug is used to treat certain mental/mood conditions (such as schizophrenia, bipolar disorder). From this analysis, I concluded that people with mental illness is of the greatest concern to the NHS since such medication is been prescribed the most. As I mentioned in my previous point, this does not necessarily mean that these mental illness cost the NHS the most, although that could also be the case. Further analysis would need to be taken to justify this.
* The total cost to the NHS in London between July 2018 and December 2018 is just under £125M

# OPERATIONALIZE – Phase 6

**Prepare for Production Environment**

After completing my analysis, my next step is to prepare the deliverables (reports and code) and getting the solution in the production environment. The code and data will be published on Github

As one of the requirement is to allow users to easily copy content for inclusion in other presentations, I need to do some amendments to my R code so that the process is automated. Instead of hardcoding my variables, I changed my code in such a way to read from a specific folder.

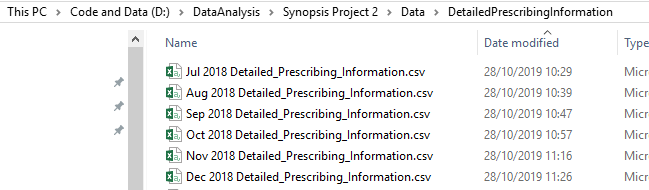


The file in the folder has the format

monthname year Detailed\_Prescribing\_Information.csv

e.g. Aug 2018 Detailed\_Prescribing\_Information

In this way I have easily extracted the month and year for that file. Whenever new data becomes available, e.g. Jul 2020 Detailed\_Prescribing\_Information.csv, my R code will capture that data, and merge it with the existing dataset.



Whenever I hit the ‘refresh’ button in PowerBI the most recent data is being picked up, and update the dashboard accordingly.

I then published the dashboard unto the web, for greater user experience. The link is

<https://app.powerbi.com/groups/me/reports/5f51d4bf-94f5-4efb-8110-a241765d24c1/ReportSection125227491c4378bcc950>

<https://app.powerbi.com/view?r=eyJrIjoiZmU2NzlkYTctNzkwOS00ZThlLWJlMGUtNzhjOThiODhlNzc0IiwidCI6Ijk3NjJmNDExLTQzOGQtNDQwYy04MzU3LTg5NmIzNWFlODcwNyJ9>

**Final Deliverables**

The deliverables for this project are as follows:

Documentation (it is this current document)

This include:

* How I’ve structured the problem and planning the work
* How I’ve engaged with the customer
* How I’ve gathered the data
* How I’ve conducted the analysis
* How I’ve developed insight and making recommendation
* How I’ve communicated my recommendation

It is a technical document showing how I’ve implemented the code.

R Code

This include how I’ve used R code to conduct my analysis. The files are as follows:

* SynopsisProject.R – This is the main code where I have conducted the analysis, primarily the data pre-possessing aspect of the analysis. In addition to this I have conducted further analysis.
* SynopsisProject-ANOVA.R. This is where I conducted some ANOVA hypothesis testing
* SynopsisProject-Adv.R. This is where I conducted some K-Means Clustering algorithm
* SynopsisProject-TimeSeries.R. This is where I conducted Time Series analysis to predict and forecast prescription costs in the future.

Reports and Dashboards

This is a PowerBI file, SynopsisProject\_Report.pbix, for displaying the summary visualisations. From this report, I have developed insights through an interactive user experience. There is also a web version of the reports and dashboard, <https://app.powerbi.com/groups/me/reports/5f51d4bf-94f5-4efb-8110-a241765d24c1/ReportSection125227491c4378bcc950>.

Presentation

This is a PowerPoint file, Synopsis-NHS\_Prescription.pptx, summarising everything in this documentation, at a higher level, for Senior Manager to takeaway. It contains summary visualisations and dashboards, and highlights how I’ve communicated recommendations.

Data

This contains some resulting datasets after conduction the analysis. This consists of the following files in the ‘Data’ folder:

* pres\_final.csv. This is a csv file with ‘Actual Cost’ metric in six columns in differentiated by their months (E.G. Jul, Aug, Sep, Oct, Nov and Dec). It shows the output of how I’ve gathered the data.
* pres\_final\_pros.csv. This is a csv file after some extra data pre-processing. This file is the final output, and it’s being using for the visualisation in PowerBI and for the other machine learning algorithms, used in the project.
* PrescribingInfo-TimeData.csv. This file is used by the Time Series algorithm to predict and forecast future prescription cost.