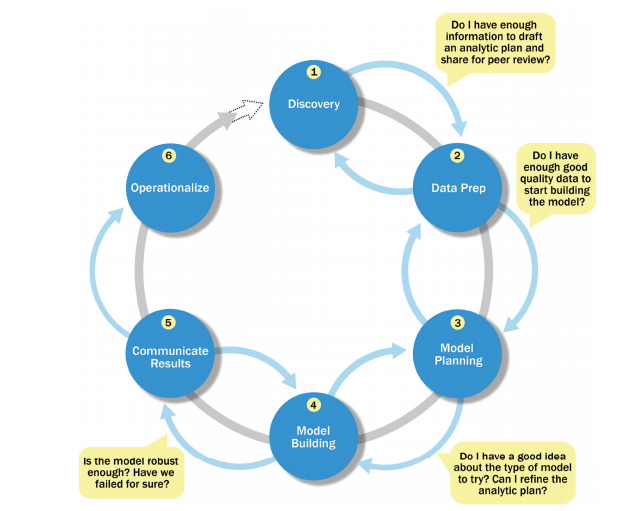
# INTRODUCTION

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**. In the discovery phase, I described the **situation** and 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 mitigate potential risks.

**Data Preparation.**  In the Data Preparation phase, I have shown how I have downloaded the data, cleans and formatted the data, ready for analysis.

**Model Planning**. In the Model Planning phase, I determined the methods, techniques, and workflow I intended to follow for the subsequent model building phase.

**Model Building**. In the Model Building phase, I have shown how I have used data mining techniques, time series and geo-spatial analysis to build the model

**Communicate Results.** 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. I have produced report showing cost analysis and comparisons between Area Teams, PCO and Practices.

**Operationalize.**  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 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 your company can provide cost savings; against GP prescriptions in particular.

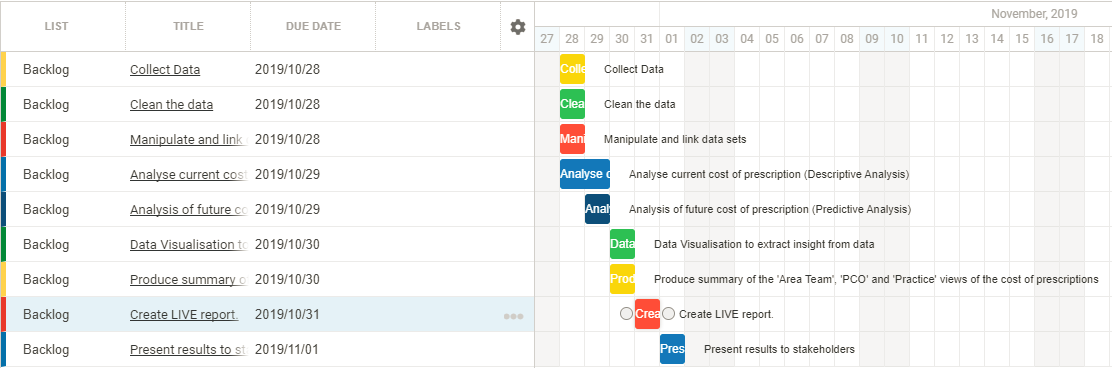
Pg. 363 on Big Book

**Project Aim and Objectives**

The objective 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

**Project Management**

I created this Gantt chart using Trello



**Scope** (https://www.projectmanager.com/blog/project-scope)

Timescale

1 week

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 |

**Stakeholders**

*do similar diagram from protfolio*

The stakeholders are as follows: Senior Management Team, NHS CCG, Trafford Council (HR), and Line Manager

**Assumption**

The assumptions are:

* I have already met with the customers to gather the requirements

**Engagement with customer**

**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 |

**User Story**

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

When I am viewing a report

Then I should be able to visualise on a map, where these GP practices are located

Use 1.3c, 1.3e, 1.2 in portfolio

Say how I engage with customers. Use examples from Portfolio

**Identify, collect and migrate data to / from a range of internal and external systems**

# ANALYSIS – Case Study 1

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.

**DATA PREPARATION**

**Collecting the Data**

…..

The data collected for this analysis are:

* Detailed Prescribing Information – I collected this dataset since it contains the data of interest to conduct the analysis
* BNF Code Information – I collected this dataset since contains additional data, not contained in the Detailed Prescribing Information dataset, that are of interest.
* Trafford General Practices – I collected this dataset since it contains spatial information which I can display on a map.

**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.

In order do a more accurate cost comparison, I have created an additional field, NIC\_per\_Total, where

NIC\_per\_Total = NIC / (Items X Quantity)

The fields contained in the **BNF Code Information** dataset are as follows:

* BNF Chapter
* BNF Chapter Code
* BNF Section
* BNF Section Code
* BNF Paragraph
* BNF Paragraph Code
* BNF Subparagraph
* BNF Subparagraph Code
* BNF Chemical Substance
* BNF Chemical Substance Code
* BNF Product
* BNF Product Code
* BNF Presentation
* BNF Presentation Code

BNF Presentation – The BNF Presentation is similar to the BNF Description in the Detailed Prescribing Information.

The other fields are hierarchical. So BNF Presentation is subset of BNF Product, which is a subset of BNF Chemical Substance, which is a subset of BNF Subparagraph Code, etc..

The fields contained in the **Trafford General Practices** dataset are as follows:

Name – Name of the GP Practice

Address – Address of the GP Practice

Postcode – Postcode of the GP Practice

Telephone – Telephone of the GP Practice

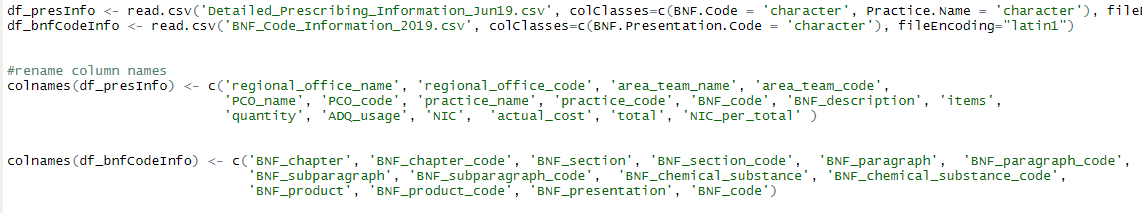
Lon – Longitude coordinate of the GP Practice

Lat - Latitude coordinate of the GP Practice

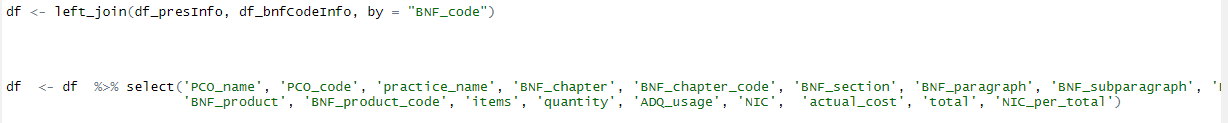
**Data Conditioning**

Before conducting my analysis, there are certain data conditioning I had to perform to get the data in a state for analysis. The steps I have taken are as follows:

* I have used R Studio to read the ‘Detailed Prescribing Information’ and BNF Code Information’ datasets for June 2019.
* I then rename the columns for easer readability.



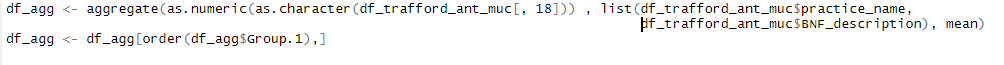
* I then merged the two datasets together on the BNF\_Code, which is a common field that links the datasets together
* I then selected a subset of fields, that are of interest to conduct the analysis



* I then filtered the dataset to only contain the records belonging to Trafford Council. Trafford Council has a PCO (Primary Care Organisation) code of 02A00.
* Then I choose a specific BNF section to do the analysis on. I have chosen Antisecretory Drugs+Mucosal Protectants.



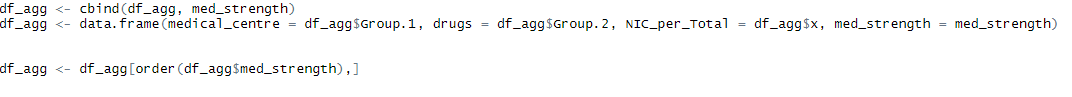
* Because some of data is replicated, i.e. a similar medication is prescribed for numerous patients, I did some data aggregation, taking into account the mean of the NIC/Total field, indicated by the 18th column in the table. I have also limited the result to only return the Practice Centre Name, BNF Description and NIC/Total, and re-order the data by the Practice Centre Name.



* In order to perform an accurate comparisons of prescriptions, it makes sense to compare prescriptions with the same medication strength, i.e. prescriptions of 10mg are compared together, those of 20mg are compared together, etc. For this reason, I have used R script to extract this information from the BNF Description field, using regular expression. The code is as follows:



I then merged the med\_strength dataframe with the already created dataframe, df\_agg, and renamed the fields to make them more readable.

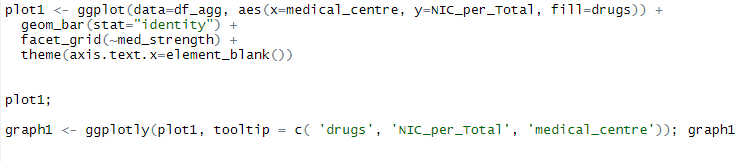


The data is now clean, formatted and is now in a state to conduct the analysis.

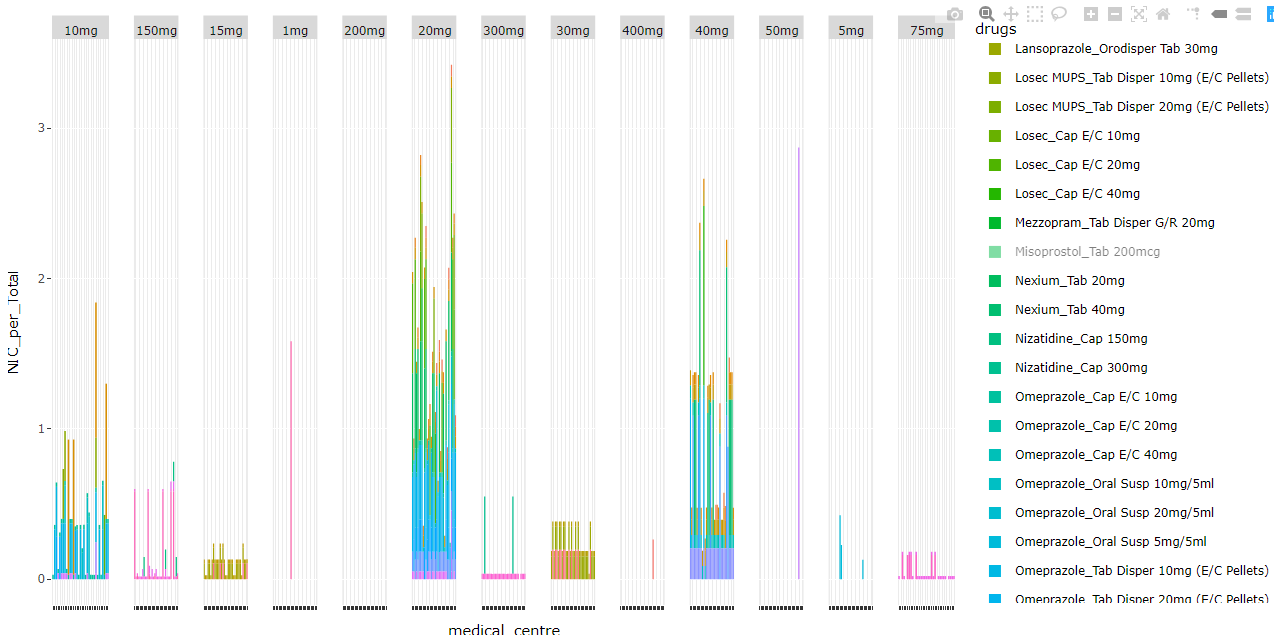
**MODEL PLANNING**

**Data Exploration – Bar Chart Plot**

After I prepared the data, I then plot the data to explore to better understand the data. I created a bar plot in R, using ggplot and leaflet. See code below.



The result of the plot is as follows.

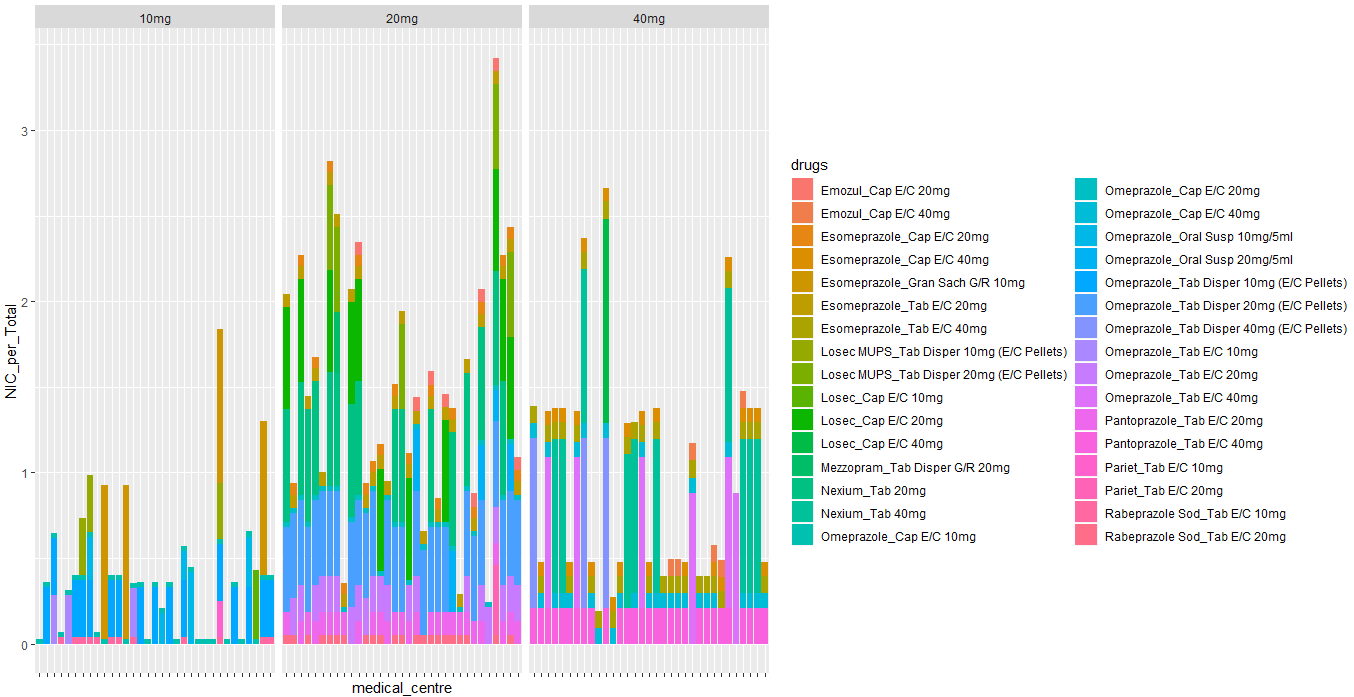


I have plotted NIC\_per\_Total on the y-axis and the medical centre on the x-axis. I have also grouped the plot based on their medical strength, as shown at the top of the plot.

Looking at the visualisation, I have observed that there are more drugs prescribed for 10mg, 20mg and 40mg for this BNF section (Antisecretory Drugs+Mucosal Protectants).

Because of the vast amount of data, the graph fails to show the visual clearly. For this reason, I have limited the visualisation to just showing 10mg, 20mg and 40mg areas which seems to contain most of the data.

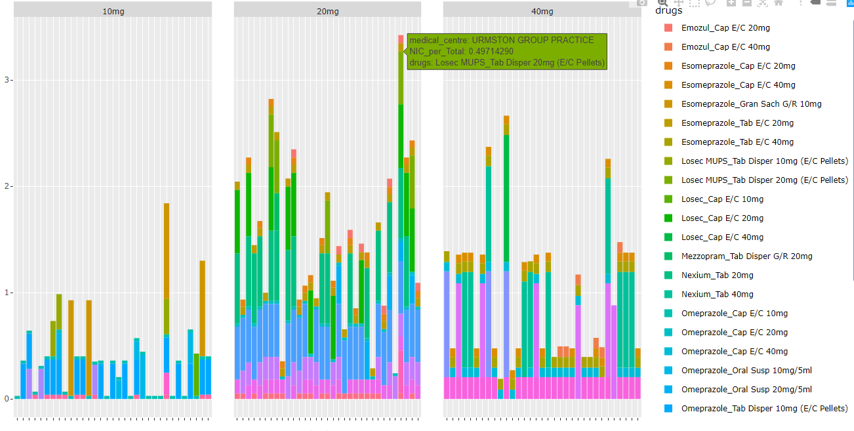
The updated visualisation is as follows



I have interpreted the graph as follows:

* There are 3 groups that represent the medical strength of the drugs, 10mg, 20mg and 40mg
* Within each group, there are 32 bars. Each bar represent a medical centre or GP practice centre within Trafford
* The colour on the bars represent the type of drugs prescribed at that centre, which can be derived from the legend on the right.
* The length of that colour on the bar indicates how expensive that particular drugs is. So the longer the colour bar, the more expensive it is.

I have created the graph using the R plotly library, so whenever I roll the mouse over a bar, it show more information about that bar. Here are a few screenshots of that in operation

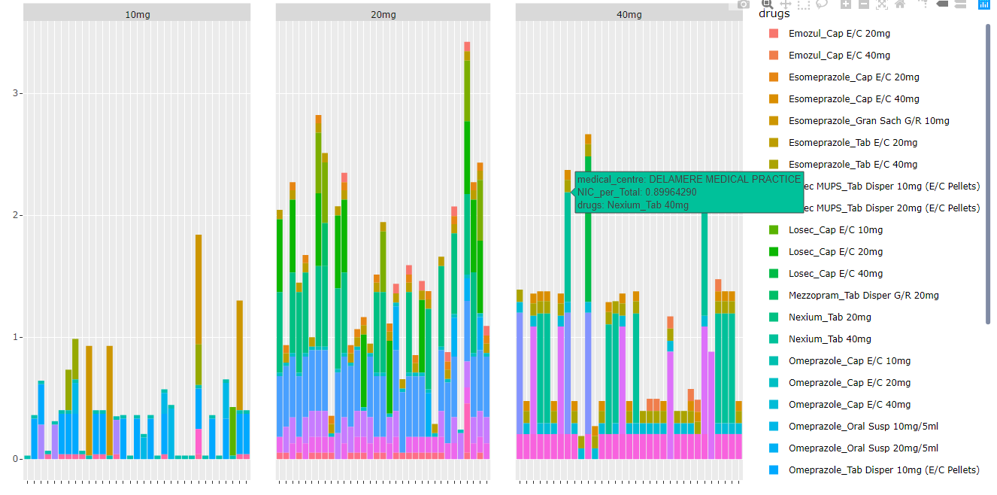


When I rolled my mouse over a green section of the bar under to 20mg group, as shown above, I am present with a pop-up of:

Medical centre: Urmston Group Practice

NIC per Total: 0.49714290

Drugs: Losec MUPS\_Tab Dispenser 20mg (E/C Pellets)



When I rolled my mouse over a tiel section of the bar under to 40mg group, as shown above, I am present with a pop-up of:

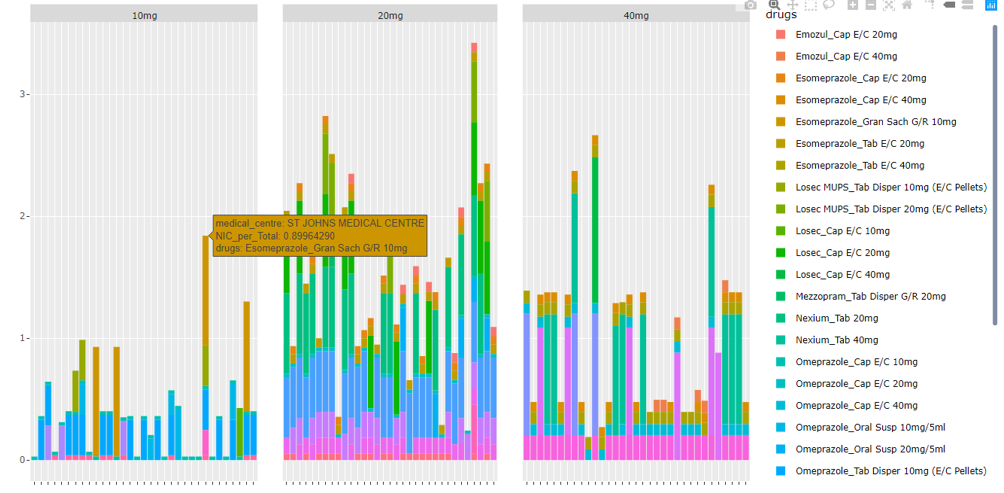
Medical centre: Delamere Medical Practice

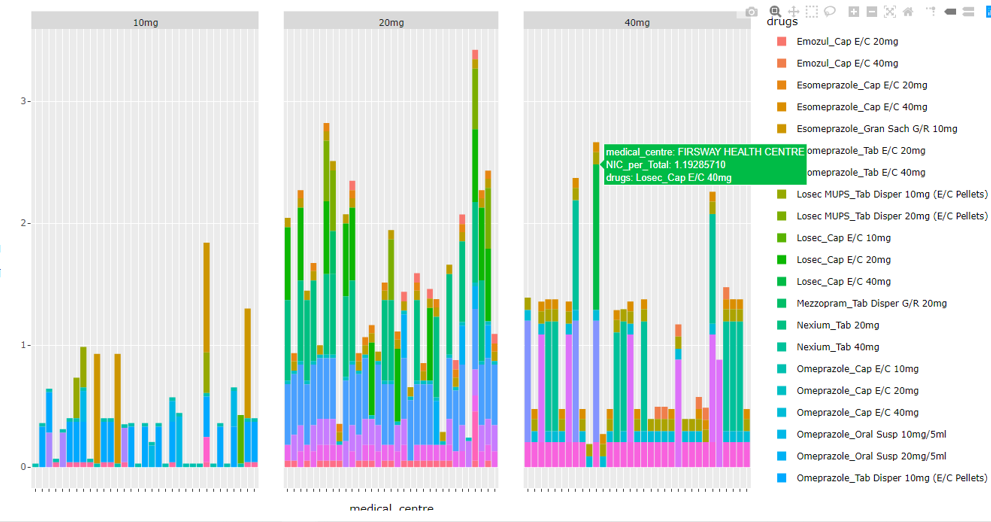
NIC per Total: 0.89964290

Drugs: Nexium\_Tab 40mg

I have repeated this process for the bars that are more significant, i.e. those with the highest peak, and I have derived the following:

* St John Medical Centre have a higher spending cost for 10mg prescription drugs, Esomeprazole\_Gran Sach G/R 10mg being the most expensive with an NIC/Total of 0.8996. This is a cause for concern and I would flag this up to suggest a cheaper alternative prescriptions



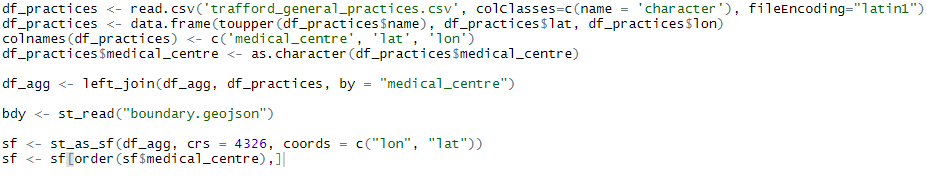
* Urmston Group Practice Centre have a higher spending cost for 20mg prescription drugs, followed by Davyhulme Medical Centre.
* Firsway Health Centre have a higher spending cost for 40mg prescription drugs, followed by Dalamere Medical Centre. Two such expensive drugs prescribed by Firsway are Losec\_Cap E/C 40mg and Omeprazole\_Tab Disper 40mg with NIC/Total of 1.192 and 0.994 respectively. These are causes for concern and I would flag these up to suggest cheaper alternatives.

**Data Exploration – Spatial Data Plot**

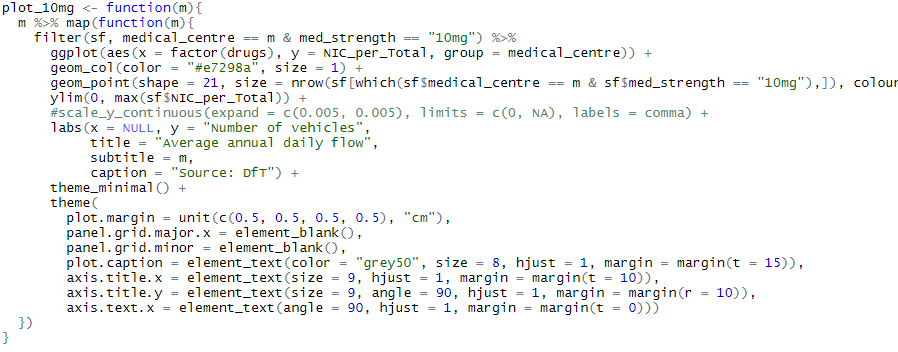
The bar plot I created easier gives a quick overview of the data, but to understand better, and to compare prescription costs across GP practices geographically, I have decided to explore the data on a map. The task is to give me the ability to drill down to specific geographical area, and select on a point on the map and present a bar chart popup of cost comparisons of various prescriptions.

Again, I have limited the search to just 10mg, 20mg and 40mg medication strength, since these groups contain more data to conduct the analysis.

I have merge the current dataset with Trafford General Practice dataset to capture the spatial data (lat, lon), and the convert the lat/lon values into geometry coordinates using the st\_as\_sf function in R. See code below.

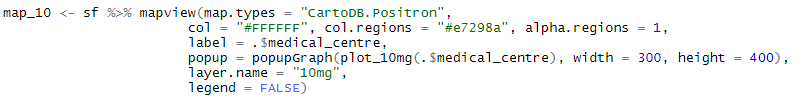


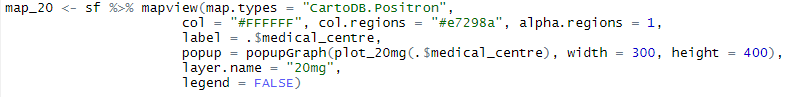
I then created geospatial map using the mapview, leafpop and ggplot libraries in R.

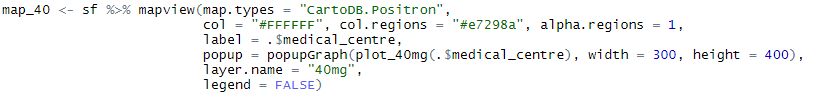


I repeated this code for 20mg and 40mg, passing in these values into the filter function.

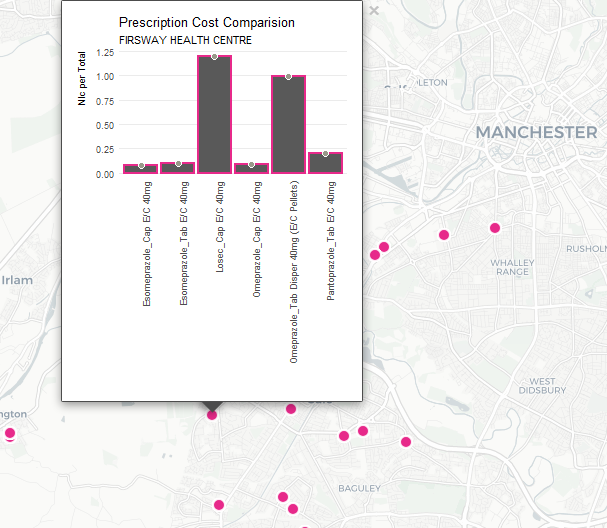
I then created a layer for each group







The result is as follows



The screenshot above shows a popup bar plot when I selected the 40mg layer and clicked on Firsway Health Centre on the map. As mentioned above, it highlights 2 expensive drugs, a Losec\_Cap E/C and Omeprazole\_Tab Disper 40mg that could be swapped for a cheaper alternative.

**Model Selection**

After exploring the data using exploratory analysis, 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.

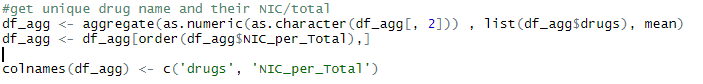
I decided to use the k-means clustering algorithm. I choose the k-means clustering because it is an ideal algorithm that is used for grouping objects. K-means clustering algorithm is also an unsupervised learning algorithm, which means it doesn’t rely on a target or labelled variable to train the dataset, 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.

**MODEL BUILDING**

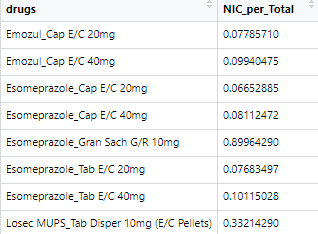
**K-Means Clustering**

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

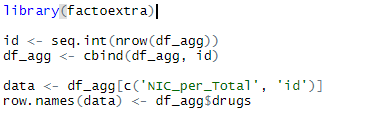
* I had to do some more formatting of the data in order to retrieve the unique drugs from the dataset. I did this using the aggregate function in R



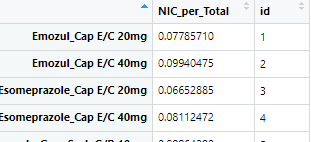
This dataframe now contains two columns, drugs and NIC\_per\_Total



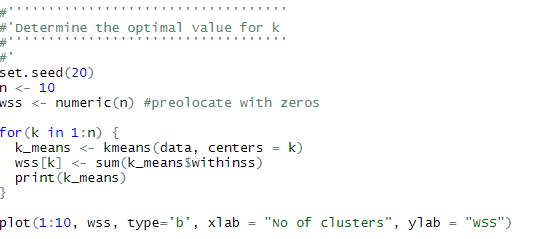
* I used the factoextra library in R
* I then created an id column in the dataset. I did this in order to create a 2-D view to the map
* I then created the drug column as the index. I did this so that the drug’s name are displayed within the cluster they belong

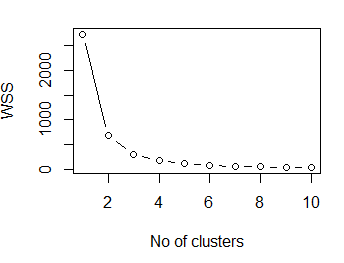


The dataframe, again shows two columns, NIC\_per\_Total and id, with drugs being the index.



* 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.

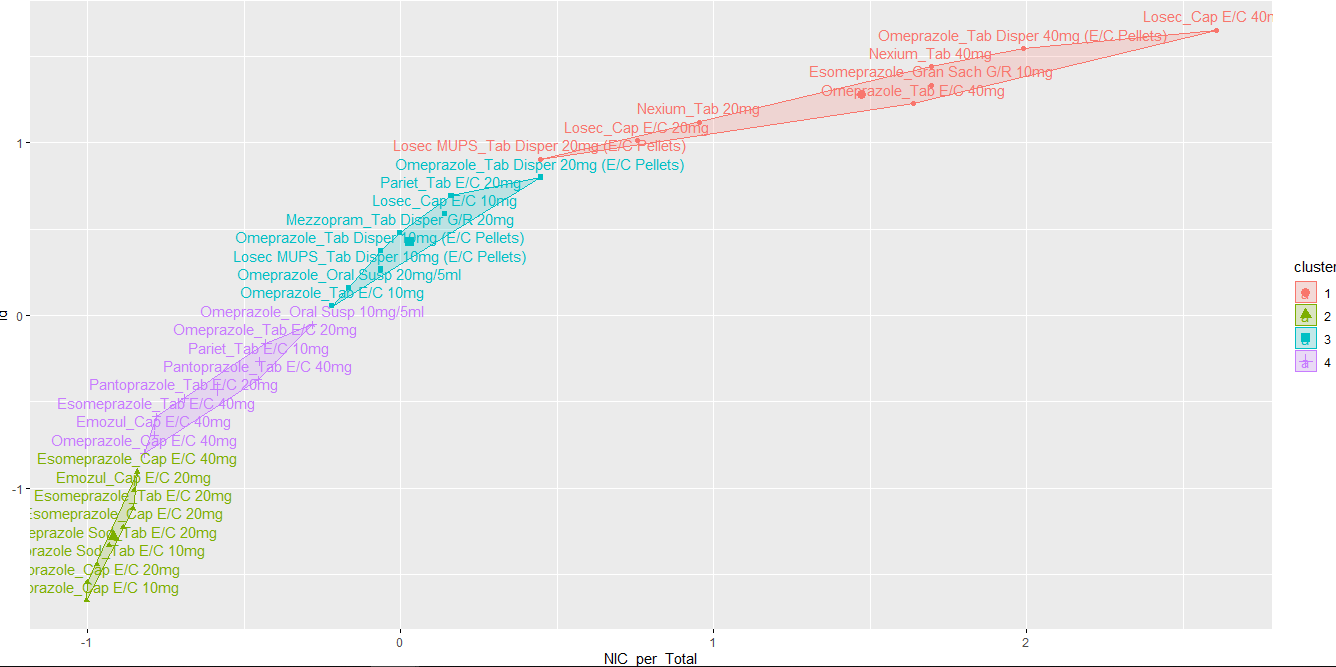


* I then used the fviz\_cluster function to visualise the data.



**COMMUNICATE RESULTS**

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 NIC/Total 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 NIC/Total 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 can also use this to predict which category a new drug will be grouped under, given their NICperTotal value.

**Dashboard**

See dashboard.csv and dashboard.pbix

LIVE

I used slicer in PowerBI to create live …..

Also, I have created the dashboard and reports in PowerBI in such a way such that, each time a file is available in the folder, I easily refresh PowerBI to get the new dataset

**OPERATIONALIZE**

**Action**

The actions I took to conduct this analysis is as follows:

Use 1.1c to explain migrating data into powerbi for dashboard

1b in Portfolio

Use PowerBI for live stream, and use slicer for dynamic data. Use appendix 1 in portfolio for live stream.

Use STAR technique to answer questions and add Introduction and Content

**Manipulate and link different data sets as required**

2.1a, 2.1b in portfolio

**Analysis**

Q. Analyses that demonstrate an understanding of GP prescription costs.

A. Say how did they prescribe the drug. Is it funded by government ?

<https://www.pharmaceutical-journal.com/download?ac=1064834&firstPass=false>

Do cost per city and doctor and compare over months

Plot real cost vs sale , and gather insight

**Use a range of analytical techniques such as data mining, time series and modelling techniques to identify and predict trends and patterns in data**

7.1 in portfolio

**Apply the tools and techniques for data analysis, data visualisation and presentation**

9.1b in portfolio

**Summarise and Present the results of data analysis to a range of stakeholders making recommendations.**

Q. 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)..

Can I Compare using mean and Standard deviation, like on my Project management project

A. Do dashboard, and gather insight from the dashboard.

In dashboard, do 6 – 8 charts, include bar chart, pie chart, GIS, etc

11 in portfolio

**Collecting Data**

Download: Practice Prescription Information – within Area Team data collected for June 2019 (period April 2019 – June 2019), within Greater Manchester Area Team. (Request number: **REQ0059741)**

Download: BNF Code Information

**Analysis**

I then calculated the Total by multiplying Item and Quantity, and calculated the NIC per Total

Search “synopsis” in my googlemail and see links with examples of reports, etc.