ANALYSIS OF ENGLISH PRESCRIBING DATASET (EPD)

NAME: OBISANYA, IDOWU ISAIAH

STUDENT ID: C2861872

BIG DATA AND BUSINESS INTELLIGENCE

[COURSE CODE: CIS4008-N-BF1]

SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES (SCEDT)

COURSE OF STUDY: M.SC COMPUTER SCIENCE (WITH ADVANCED PRACTICE)

JANUARY 2024

Table of Contents

SECTION 1

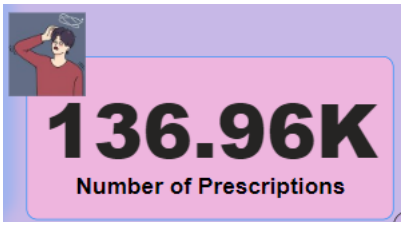
BUSINESS INTELLIGENCE (BI) STAKEHOLDERS' REPORT

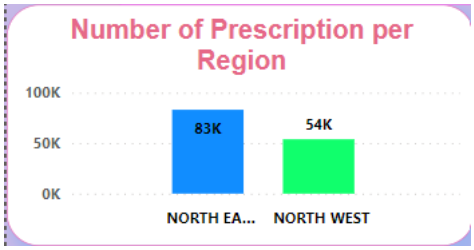
CHAPTER ONE

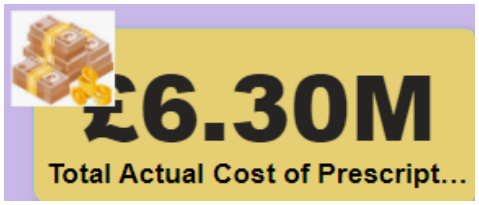
Executive Summary

This project accessed and processed the English Prescribing Dataset (EPD) for two months: August and September, 2023 for two (2) selected regions and four (4) cities to analyze it and develop dashboards based on the business questions proposed. Various charts and visuals and other external features of Microsoft Power BI were used including Infographic Designer, Artificial Intelligence as well as usage of **DAta eXpression**, **M Language** were demonstrated.

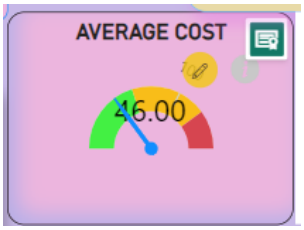
From the visualization, **136.96K prescriptions** were recorded between the two months where **North East and Yorkshire** was **83K** and **North West** was **54K**. While the **total quantity** of drugs prescribed was ***75 million***, the **Total Actual cost** of the prescriptions was ***£6.30 million***.

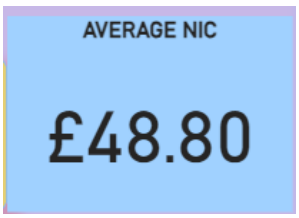






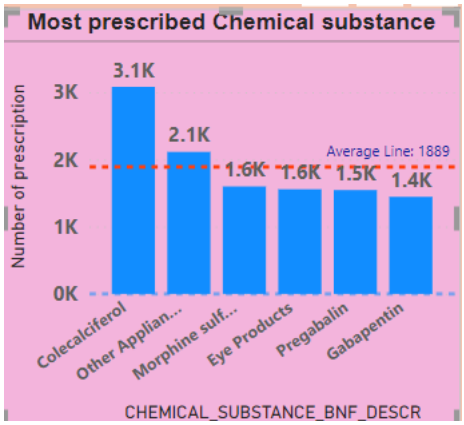
It was also established that the **average Actual cost** of prescription for drug was **£46** while ***Net Ingredient Cost (NIC) was £48.80*** being the amount charged by the drug manufacturers or value contained in the Drug Tariff. This shows that a prescribed drug in the UK is being subsidized by the Government by an average amount of **£2.80** per prescribed drug (difference between ***NIC and Actual Cost***).



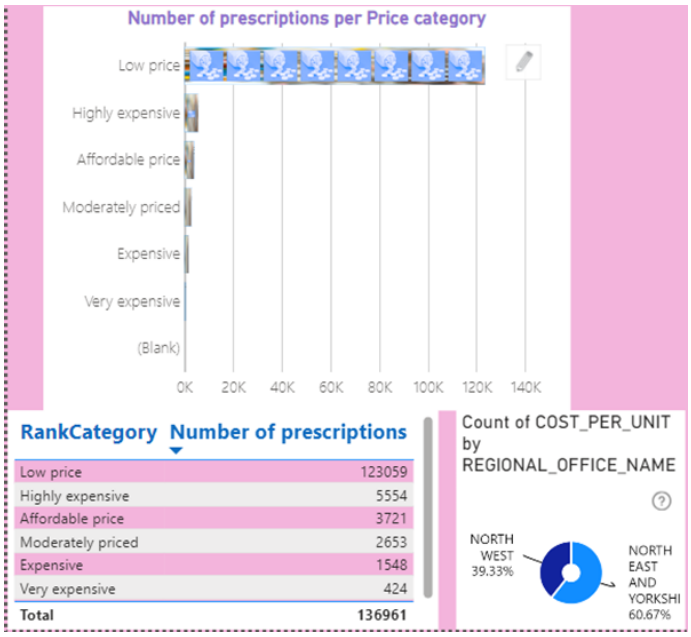


It was also discovered that ***Colecalciferol;*** a drug used for Vitamin D deficiency or Kidney disease is the most prescribed drug with a value of **3,080** within the 2 months for the sample locations in the

UK. Hence, Government needs to focus more on more effective **preventive measures** for Vitamin D deficiency or Kidney disease rather than a corrective prescription.



In the same vein, it was also established that drugs that are mostly prescribed in the UK are categorized as **low priced** with a count of **123,059** (**89.8%)** out of a total count of **136,961** prescriptions given within **North East and Yorkshire** and **North West regions** for the **selected 4 cities**.



Introduction

One of the major needs in any human society is access to quality health care. The United Kingdom, through the National Health Service (NHS), is committed to providing quality health care as enshrined in ***The Values*** of **NHS** which are: working together for patients, respect and dignity, commitment to quality of care, compassion, improving lives, everyone counts ((NHS), 2023). To achieve the values above, it is important for all stakeholders to have insight on the prescriptions that have been delivered in the past, affordability of the cost of drugs prescribed, identifying major illnesses or diseases prevalent in the country and across specific regions, cost of providing health care, how much is required to provide health care and to identify areas of improvement. This will enable Government and other agencies to prioritize and allocate fund appropriately. In this Project, we reviewed the English Prescribing Dataset (EPD) for two months for selected Regions and Cities in order to provide insight on the healthcare delivery in the UK and to identify areas of improvement. At the end of this project, skills on Data pre-processing, Data Modelling, Visualization, use of Artificial Intelligence (AI) for insight, DAta eXpression (DAX) and M Language would have been demonstrated.

Dataset Description

The English Prescribing Dataset (EPD) for two months (August and September, 2023) were selected as data source from the National Health Service (NHS) official digital platform, known as NHS Open Data Portal Business Services Authority. This can be accessed from the link: EPD_August2023 and EPD_September2023. Due to the hardware requirement, time constraint to deliver this project and the huge volume of the entire monthly EPD, which is in excess of **17.7million** records each, the EPD dataset used in this project were limited to 2 sample regions: ***North West*** and ***North East and Yorkshire*** and also limited to 4 sample cities: ***Liverpool***, ***Middlesbrough***, ***Sunderland*** and ***Newcastle upon Tyne***. The two filtered dataset contain **70,956** records (August 2023) and **66,005** records (September 2023) respectively and **26** columns each.

Overview of Business Questions

This BI project would be able to provide all stakeholders with insights on the healthcare delivery with regards to the drugs prescription to patients, comparison of the prescription with different regions and cities and specific Hospitals or Practices in a community. It will also help to provide insights on the cost implication of different prescription by comparing the actual cost with the basic price of the drug known as Net Ingredient Cost (NIC). The Net Ingredient Cost is the original cost of the active ingredient of the drug excluding any discount, or rebate or cost of dispensing the drug. The analysis in this project is originally concerned with the following questions:

- What is the total count, cost and quantity of prescription in certain region, cities and practices in the country?
- What are the most prescribed chemical substances? (i.e. KPI: top chemical substance vs number of occurrence)
- What are the major organs of the body been treated by the prescription in specific region?
- Has there been any improvement in the ailment based on the trend of diagnosis?
- What are the factors that can influence the average price of the Net Ingredient Cost (NIC) of prescription?
- What drug has the highest cost of prescription?
- What Drug has the least total variance between Net ingredient Cost (NIC) and Actual Cost between the two months.

- How affordable are the prescribed drugs and what factors can influence the cost of these drugs?

Aims

This BI project is aimed to achieve the following:

- analyse the English Prescribing Dataset to identify patterns and trends in prescription in the healthcare sector.
- develop business intelligence technique to produce actionable insight from the dataset.
- create visualisations and dashboards to present the derived insights in effective ways.
- Identify areas of opportunities to reduce cost of prescription.

Objectives

- To collect and clean English Prescribing Dataset (EPD) within the first 2 weeks of the project.
- To develop dimension tables such as: Address, Practice, Region, Chemical substance etc
- To develop visualisation dashboard that will give actionable insights that can improve the effectiveness of the health care sector in UK.

Stakeholders

The stakeholders that can make use of the output in this project include the following:

- Healthcare practitioners
- Regulatory Bodies
- Pharmaceutical Companies
- Healthcare Researchers

Data Dictionary

This specifies all the column names in the fact table, their data type and description.

This is helpful to gain deeper understanding and to know what each of them represents.

Below is the table of the columns in the dataset and their descriptions:

S/NO.	Column	Title	Type	Description
1	YEAR_MONTH	Year and Month as YYYYMM	number	Example: 201401
2	REGIONAL_OFFICE_NAME	Regional Office Name	string	The name given to a geographical region by NHS England. Each region supports local systems to provide more joined up and care for patients.
3	REGIONAL_OFFICE_CODE	Regional Office Code	string	The unique code used to refer to a Regional Office.
4	ICB_NAME	Integrated Care Board (ICB) Name	string	The name given to a geographical statutory organisation by NHS England that is a smaller division of a Region.
5	ICB_CODE	Integrated Care Board (ICB) Code	string	The unique code used to refer to an ICB.

6	PCO_NAME	Primary Care Organisation Name	string	An NHS organisation that commissions or provides care services involving prescriptions that are dispensed in the community. For example: a Sub Integrated Care Board Location (SICBL), an NHS Trust.
7	PCO_CODE	Primary Care Organisation Code	string	The unique code used to refer to a Primary Care Organisation.
8	PRACTICE_NAME	Practice Name	string	The name of an organisation that employs one or more prescribers who issue prescriptions that may be dispensed in the community. For example: a GP Practice, an Out-of-Hours service, a hospital department within an NHS Trust.
9	PRACTICE_CODE	Practice Code	string	The unique code used to refer to a Practice.
10	ADDRESS_1	Address Field 1	string	The Address used by a Practice. This data is supplied by Primary Care Support England (PSCE), NHS England ICBs or the SICBL, whenever a new practice is opened or if a change of details is required.
11	ADDRESS_2	Address Field 2	string	The Address used by a Practice. This data is supplied by Primary Care Support England (PSCE), NHS England ICBs or the SICBL, whenever a new practice is opened or if a change of details is required.
12	ADDRESS_3	Address Field 3	string	The Address used by a Practice. This data is supplied by Primary Care Support England (PSCE), NHS England ICBs or the SICBL, whenever a new practice is opened or if a change of details is required.
13	ADDRESS_4	Address Field 4	string	The Address used by a Practice. This data is supplied by Primary Care Support England (PSCE), NHS England ICBs or the SICBL, whenever a new practice is opened or if a change of details is required.
14	POSTCODE	Post Code	string	The Address used by a Practice. This data is supplied by Primary Care Support England (PSCE), NHS England ICBs or the SICBL, whenever a new practice is opened or if a change of details is required.
15	BNF_CHEMICAL_SUBSTANCE	British National Formulary (BNF) Chemical Substance Code	string	A unique code used to refer to a BNF Chemical Substance. For example, 0501013B0
16	CHEMICAL_SUBSTANCE_BNF_DESCR	British National Formulary (BNF) Chemical Substance Description	string	The name of the main active ingredient in a drug or the type of an appliance. Determined by the British National Formulatory (BNF) for drugs, or the NHS BSA for appliances. For example, Amoxicillin
17	BNF_CODE	British National Formulary (BNF) Code	string	The unique code used to refer to a BNF Presentation. For example, 0501013B0AAABAB
18	BNF_DESCRIPTION	British National Formulary (BNF) Description	string	The name given to the specific type, strength, and formulation of a drug; or, the specific type of an appliance. For example, Amoxicillin 500mg capsules

19	BNF_CHAPTER_PLUS_CODE	British National Formulary (BNF) Description	string	The name given to a British National Formulatory (BNF) Chapter that includes the prescribed product. Includes the numerical code used to refer to the chapter. For example, 05: Infections
20	QUANTITY	Quantity	number	The quantity of a medicine, dressing or appliance for which an individual item was prescribed and dispensed, for each BNF Presentation. This represents a pseudo pack size, to illustrate the typical range of prescribed quantities of a given presentation. For example, a quantity of 28 for Amoxicillin 500mg capsules means that the pack size dispensed was 28 capsules.
21	ITEMS	Items	number	The number of times a product appears on a prescription form. Prescription forms include both paper prescriptions and electronic messages.
22	TOTAL_QUANTITY	Total Quantity	number	The total quantity of a drug or appliance that was prescribed. This is calculated by multiplying Quantity by Items. For example, if 2 items of Amoxicillin 500mg capsules with a quantity of 28 were prescribed, total quantity will be 56.
23	ADQUSAGE	Average Daily Quantity (ADQ)	number	Average Daily Quantity (ADQ) is the typical daily dose of a medication, prescribed to adult patients by GP Practices. This field shows the quantity prescribed multiplied by the strength, which is then divided by the Average Daily Quantity value.
24	NIC	Net Ingredient Cost (NIC)	number	In GBP. The amount that would be paid using the basic price of the prescribed drug or appliance and the quantity prescribed. Sometimes called the "Net Ingredient Cost" (NIC).The basic price is given either in the Drug Tariff or is determined from prices published by manufacturers, wholesalers or suppliers. Basic price is set out in Parts 8 and 9 of the Drug Tariff. For any drugs or appliances not in Part 8, the price is usually taken from the manufacturer, wholesaler or supplier of the product.
25	ACTUAL_COST	Actual Cost	number	In GBP. The basic cost after adjustment for the national average discount and some payments to the dispenser. See notes for details of the calculation. The calculation is: Net Ingredient Cost - National Average Discount Percentage + (payment for consumables + out of pocket expenses + payment for containers)
26	UNIDENTIFIED	Unidentified	string	This field shows data from prescription forms that could not be allocated to a Practice. Please see the guidance notes for a more detailed explanation of this field, and its uses

Table 1: Data Dictionary for English Prescribing Dataset (EPD)

As more dataset is being released on a monthly basis from the National Health Service Open Data Portal, the dataset used in the visualization would have to be updated to reflect the current data.

CHAPTER TWO

Key Findings

This section analyses the various visualization created on the dashboard and it highlights major observations which will form the basis of the conclusions drawn in this project.

1. Prescription by Region and Practice

   From the visualization, the dataset helps us to answer one of the questions of interest that:

What is the total number, quantity and cost of prescription and in certain region in the country?

It was observed that about ***137,000*** count of prescriptions were given in Both North East and North West Regions in the UK between August and September, 2023. Similarly, North East accounts for more prescription than North West where the former was **83,000** and the latter was **54,000**. (***Figure 1***)

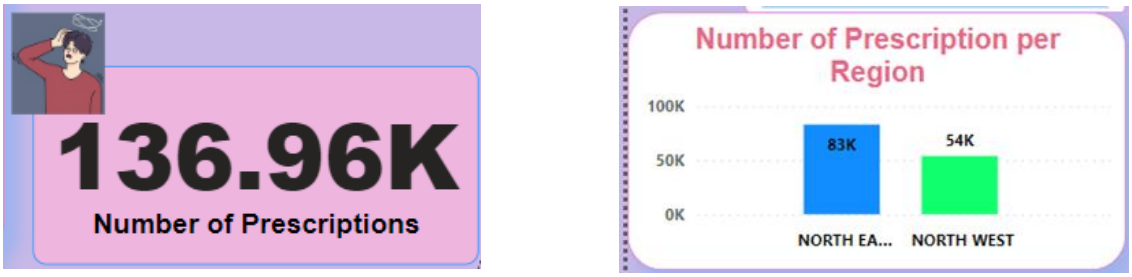


Figure 1: Total number of prescriptions given and breakdown by region

In terms of quantity of drugs prescribed, the visual shows that ***75 million*** drugs were prescribed to patients across various hospitals and medical practices in the two regions where North East was 15 million higher. That is, all the hospitals in North East region in the UK prescribed 45 million drugs (i.e. 60.2% of total prescription) while North West was **30 million**. (39.8%) (***Figure 2***)

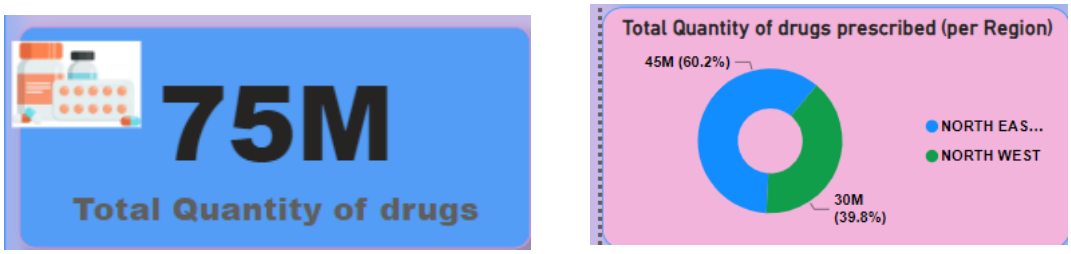


Figure 2: Total quantity of drugs prescribed and breakdown by region

The total cost incurred on all drugs and appliances prescribed in the regions was **£6.3 million.** North West region incurred a total cost of **£2.3 million** in prescription to deliver health care services in the area whereas it costs a total of **£4.0 million** for all prescriptions made to patients by all practitioners in North East. (***Figure 3***)

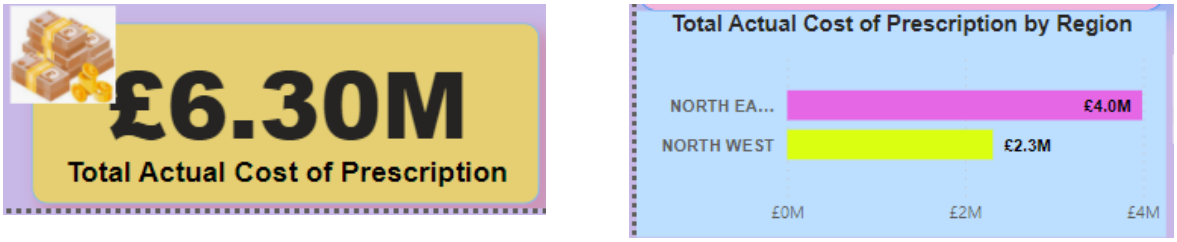


Figure 3: Total Actual cost of prescriptions and breakdown by regions

From the dashboard, one can also view the top practice name that issue the most prescription. Due to the large number of practice names in the dataset, we have limited the visual to show only the top 4 practice name with the most prescription out of the 29 practice names. From the ***Animated Bar Chart Race*** visual, ***Millfield Medical Group, Thornfield Medical Group, Cruddas Park Surgery*** and ***Westerhope Medical Group*** are the ***top 4 practice names*** with count of prescriptions: **4 ,336; 4,278; 4,068 and 4,001** respectively. (***Figure 3***).

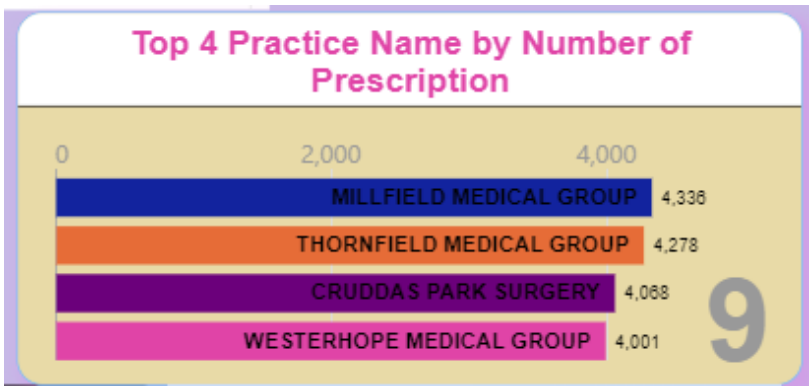


Figure 4: *Top* 4 Practice Name with highest number of prescriptions

We can view the top practice name with the most significant cost. From the ***funnel*** visual, practice name with practice code A89017 (***Millfield Medical Group***) incurred the highest cost on prescription within the period under review. Note, the description of each practice code can be viewed from the ***Table*** visual (Figure 5)

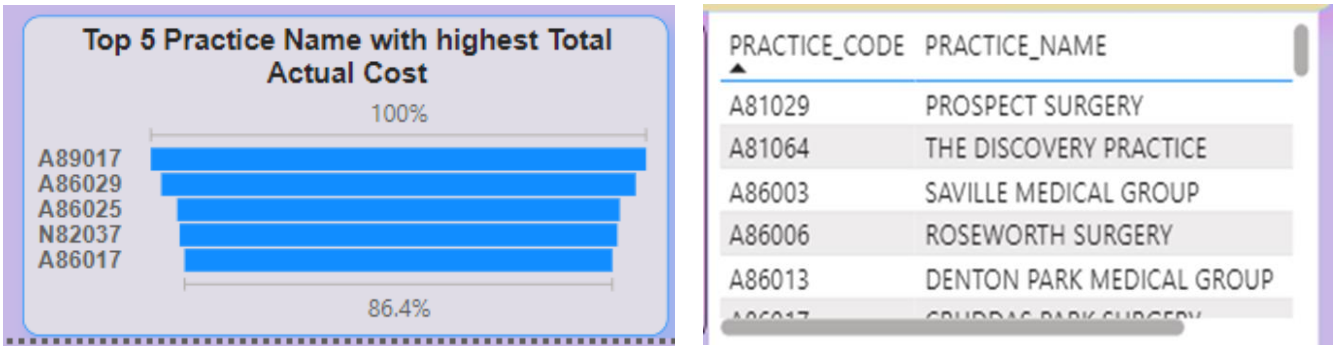


Figure 5: Top 5 Practice Name by total cost

We also used Advanced Gauge visual (Figure 6) to monitor what is the average cost of all the prescriptions. This measure helps to know at what period is the average cost of drug beyond the barest minimum targeted by NHS. From Figure 6, Average Actual Cost of a prescribed drug was **£46** while, average NIC (Net Ingredient Cost) was **£48.80.** The latter represents the basic cost of the active chemical from the manufacturer or as defined in the drug Tariff at it excludes any discount, or cost of dispensing the drug which is included in the ***Actual Cost***. Based on prior medical research

conducted, the variation in prescribing costs mostly reflects demand (Morton-Jones, T. and Pringle, M., 1993).

Figure 6: Average Actual cost and average Net Ingredient Cost (NIC)

2. Diagnosis and Drug prescribed

(a) What are the most prescribed chemical substances? (i.e. KPI: top chemical substance vs number of occurrence)

Another business question that is of interest is to know the drug that is mostly prescribed in all the medical facilities covered by the dataset. ***Colecalciferol*** occurs the most in the dataset with a total count of **3,100 prescriptions** in the two months (***Figure 7***) i.e. August (**1,601**) and September 2023 (**1,479**) . This shows most patients who visited the GP practice and other prescribing locations were likely diagnosed with **Vitamin D deficiency** or **Kidney disease**.

From this project visual (Figure 7), the average number of prescription is indicated by the average Line (colour red) which was **1, 889**.

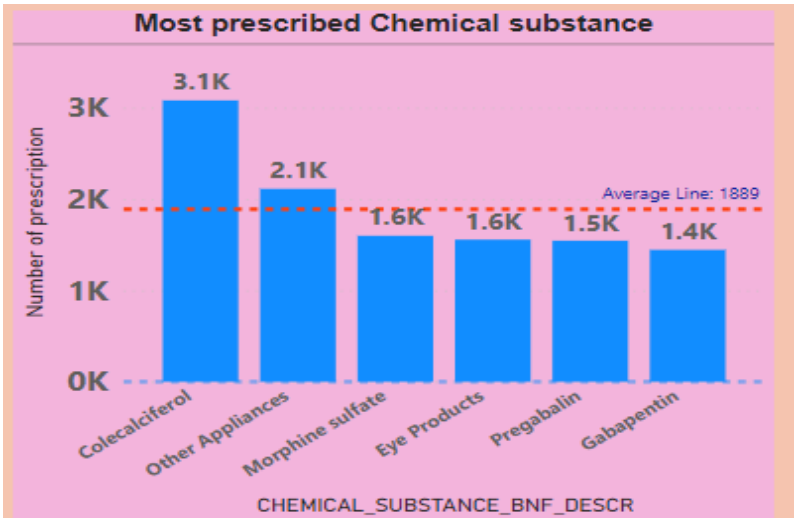


Figure 7: Average Actual cost and average Net Ingredient Cost (NIC)

This finding corroborates the result of recent research on cost on Colecalciferol within Northumbria Healthcare NHS Trust from 2012 to 2016 which shows a sharp increase between 2014 and 2016 (Woodford, H.J., Barrett, S. and Pattman, S., 2018) (***Figure 8***). This could be useful for NHS and other stakeholders within the health sector to focus more on preventive measure to reverse the ugly trend of the rising Vitamin D or Kidney disease rather than spending huge amount to treat the disease.

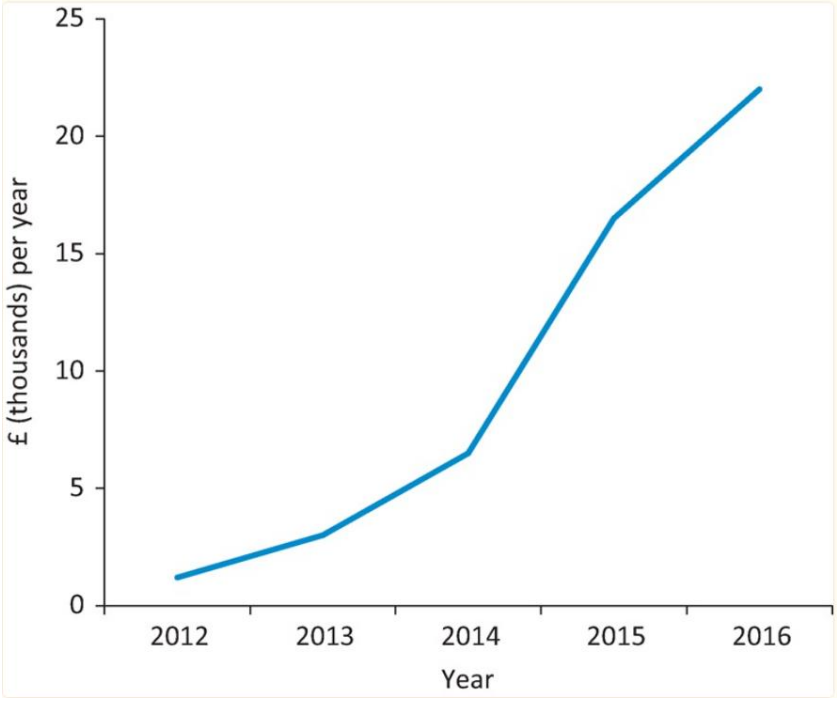


Figure 8: Spending on colecalciferol 20,000 IU within the Northumbria Healthcare NHS Trust from 2012 to 2016. IU = international units

(b) What are the major organs of the body been treated by the prescription in specific region?

An ***infographic Designer visual*** is used to depict the organ of the human body mostly treated (Figure 9). It shows that Central Nervous System is the topmost human system to be treated with **36.8k** number of cases recorded by the prescription recommended in all the medical facilities while, Diagnosis concerning Cardiovascular System is the second highest case with **20.63K cases.**

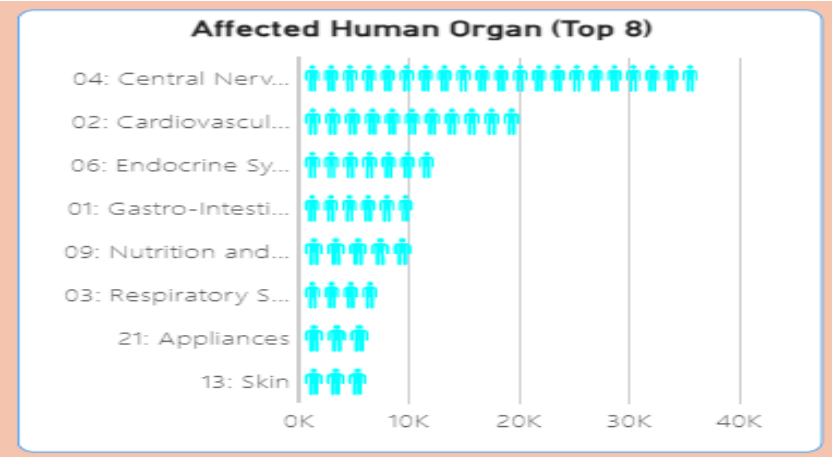


Figure 9: Affected human organ diagnosed

(c) Has there been any improvement in the ailment based on the trend of diagnosis?

The waterfall chart (**Figure 10)** shows the variance in the number of cases based in terms of the affected body organ between August 2023 and September 2023. For instance, diagnosed cases relating to Central Nervous System dropped significantly by **1,400** while diagnosis relating to

Cardiovascular System reduced by about 700 between the two months. We can infer from this data that there are improvements in the two cases between August and September, 2023

We can equally drill down by Region, by Practice name or by Drug type (BNF Description). For instance, Figure 9 (b) shows analysis of the variance by Region. There are **47 more cases** relating to ***Nutrition and Blood diseases*** while diagnosis on ***Gastrointestinal system*** dropped significantly by **35 cases** in the North West Region of UK between August and September 2023.

Figure 10: variance in cases by month and by affected body organ (b) by Region (North West)

3. Net Ingredient Cost (NIC) Analysis

(a) What are the factors that can influence the average price of the Net Ingredient Cost (NIC) of prescription?

The Net Ingredient Cost is an important KPI that Health care practitioners monitor. It is the basic price of drug as specified by the manufacturer or the value contained in part 8 or 9 of the Drug Tariff. The Key Influencer visual (***Figure 11***) uses artificial Intelligence feature in Microsoft Power BI. The visual shows that the average of NIC is more likely to ***increase*** by ***£34.10*** when the average Daily Quantity Usage ***(ADQUsage***) is zero (0). Similarly, when the ***number of items*** prescribed is between ***5 and 22***, average value of NIC is more likely to increase by ***£30.75***

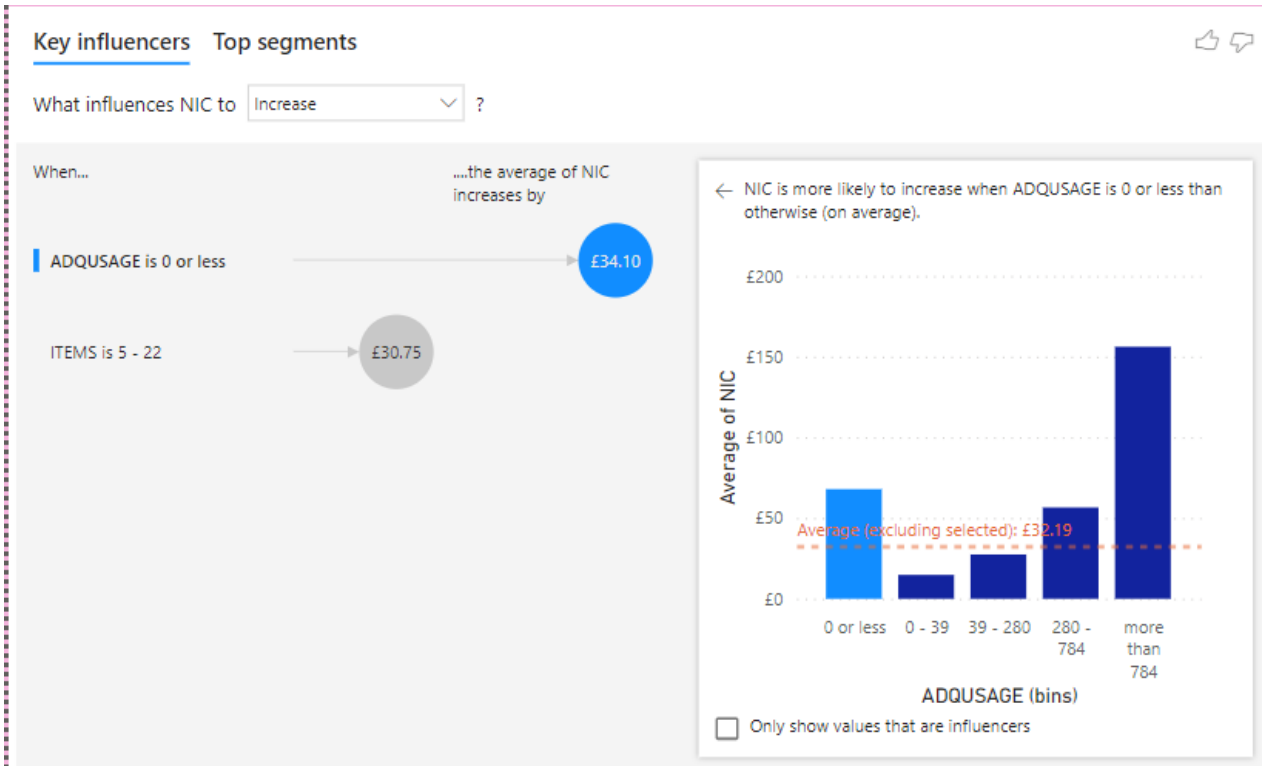


Figure 11: Key Influencers to increase Average Net Ingredient Cost (NIC)

Conversely, the average Net Ingredient Cost (NIC) is more likely to reduce by ***£33.96*** if the number of item prescribed is 1 or less. Also, ***if items prescribed is between 1 and 5***, ***Net Ingredient Cost (NIC)*** is ***likely to decrease*** by ***£17.20*** (Figure 12)

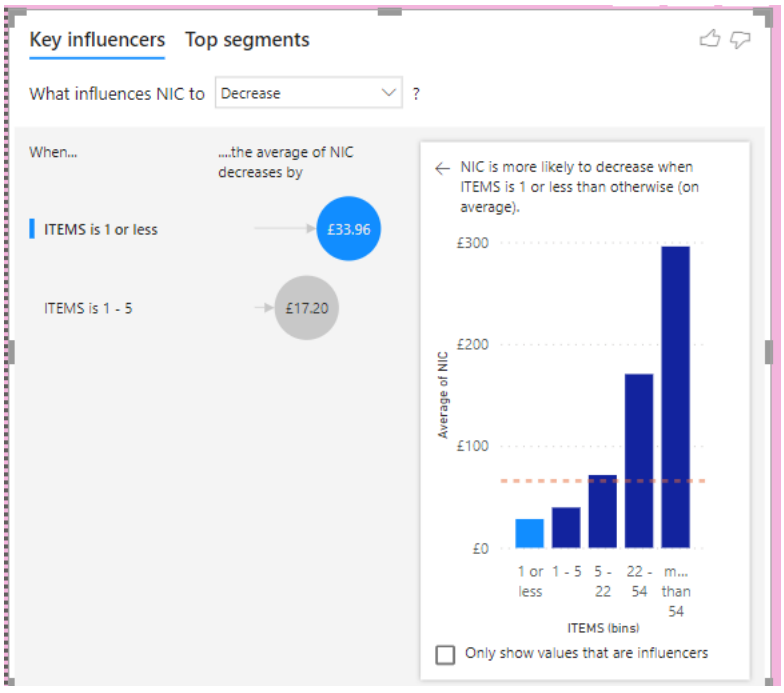


Figure 12: Key Influencers to decrease Average Net Ingredient Cost (NIC)

An analysis of the top Segments on the NIC. The Key influencer visual which uses Artificial Intelligence (AI) shows 5 top segments when the average of NIC is low (***Figure 13***).

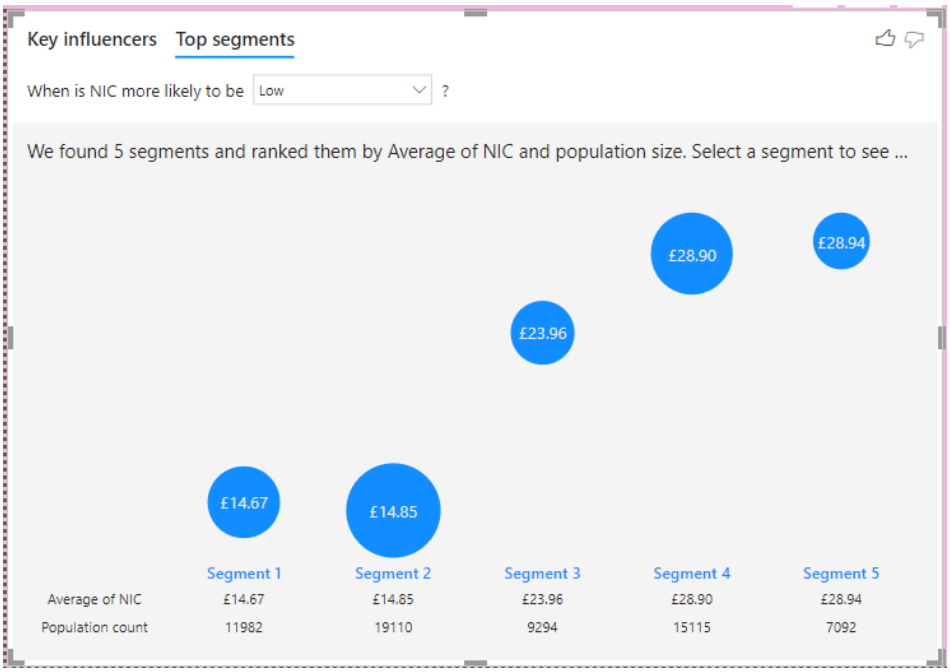


Figure 13: Key Influencers of Average Net Ingredient Cost (NIC): Top Segments

By analysing the segment with the least average NIC, it shows that average NIC is **£14.67** when ADQUSAGE is ***not more than 39*** or when, at least more than **1 item** is prescribed (Figure 14)

This is significantly lower than the overall average of NIC at **£48.80.** Healthcare policy makers can utilize this information by exploring ways to keep ADQUSAGE within this range.

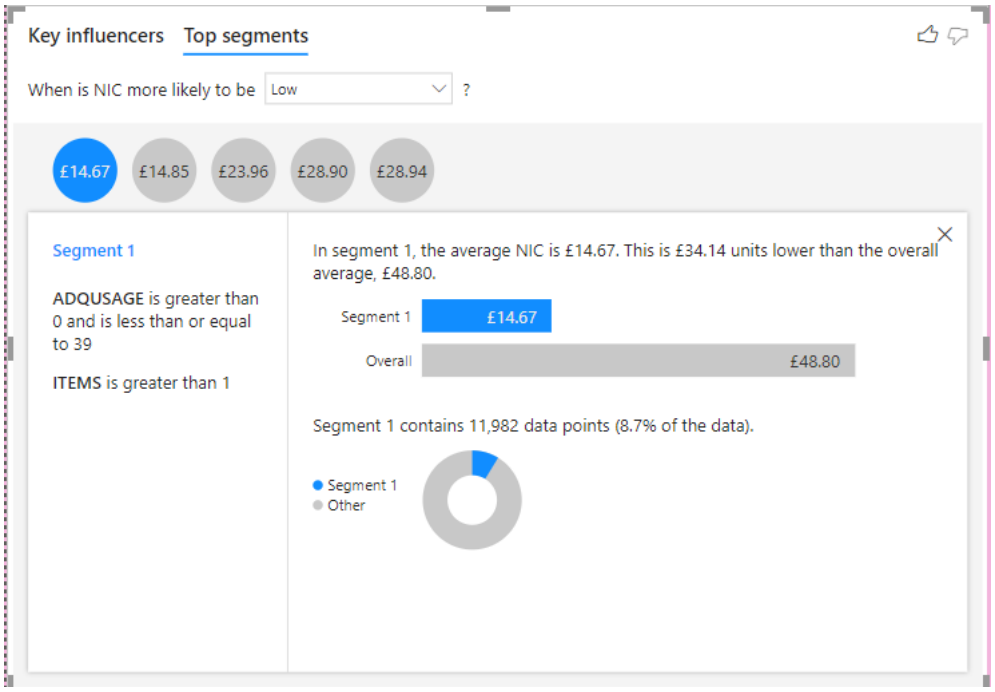


Figure 14: Key Influencers of Average NIC: Segment with the least cost

(b) What drug has the highest cost of prescription?

The drug with the highest cost of prescription is ***Apixaban*** with a total ***Net Ingredient Cost*** of ***£213,557.50.*** This is shown using the Decomposition tree visual (***Figure 15***) This agrees with the finding on NHSBSA publication on **The Prescription Cost Analysis (PCA)** where Apixaban was identified as the drug with the highest cost in England (Wilson, 2021).

The Decomposition Tree Visual also shows that North East and Yorkshire Region accounted for ***£***132.7K of the cost spent on Apixaban while North West's was ***£80.8K.***

Based on the dataset analysed, out of all the cities in North East and Yorkshire, Newcastle recorded the highest cost on Apixaban with a cost of ***£77,380.34*** followed by Middlesbrough (***£28,986.03***) and Sunderland (***£26,413.33)*** (***Figure 15***)

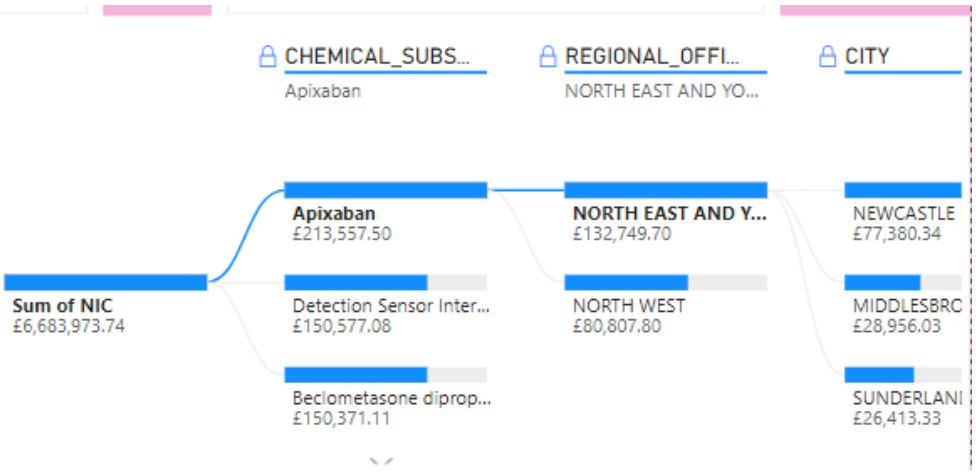


Figure 15: Decomposition Tree: Drug with the highest cost

In the same vein, Decomposition Tree in ***Figure 16*** shows the breakdown of the total NIC by ***BNF Chapter Plus Code*** and ***ICB Name*** i.e. the British National Formulatory (BNF) chapter that includes the prescribed product and the numerical code used to refer to the product. From the breakdown by the BNF Chapter plus code, cost incurred on **Cardiovascular System** was the highest with a total amount of ***£1.1 million*** while, breakdown by **ICB Name** shows that ***NHS North East and North Cumbria Integra*** tops the list of Integrated Care Board (ICB) with the highest amount of cost on Cardiovascular System. The total amount of NIC on it was ***£616,067.21***

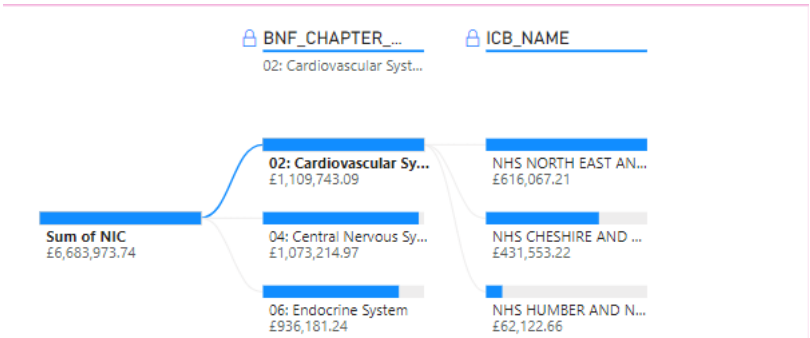


Figure 15: Decomposition Tree: Analysis of total NIC by BNF Chapter Plus Code and ICB

4. NIC vs Actual Cost Margin Analysis

(a) What Drug has the least total variance between *Net ingredient Cost (NIC)* and *Actual Cost* between the two months.

From the waterfall chart (Figure 16), ***Beclometasone dipropionate*** has the least variance between the basic price of drug (Net Ingredient Cost) and the retail price the drug is sold told patients (Actual Cost) at the aggregate level. This shows a total variance of **£300.** This means that there is significant variance between NIC and actual cost.

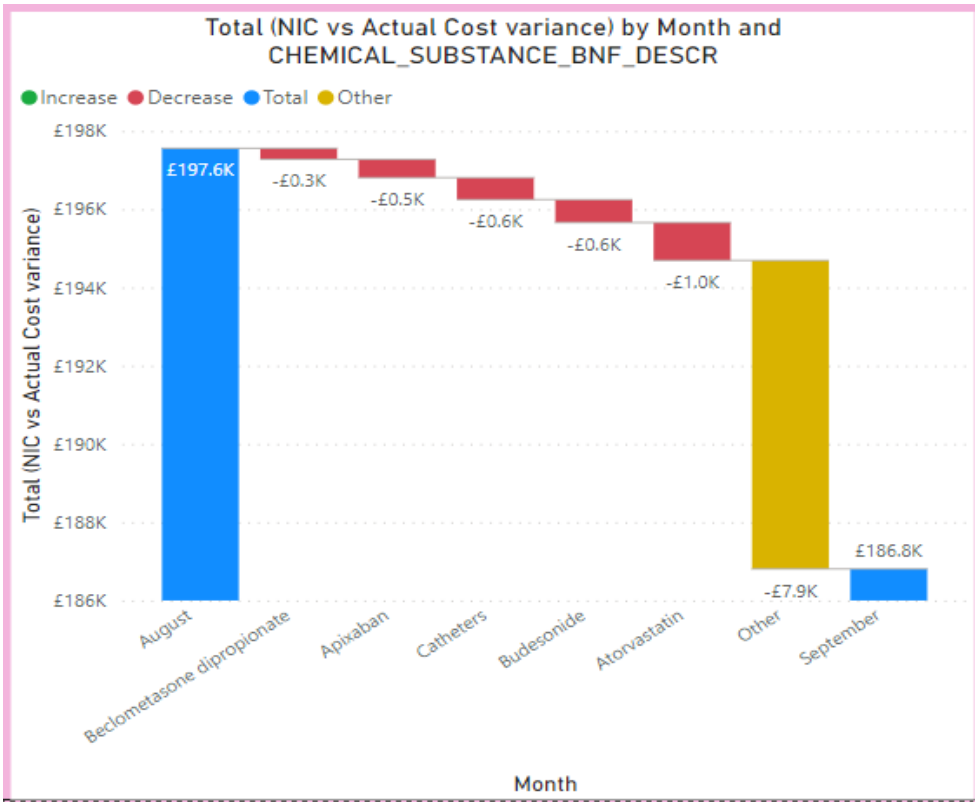
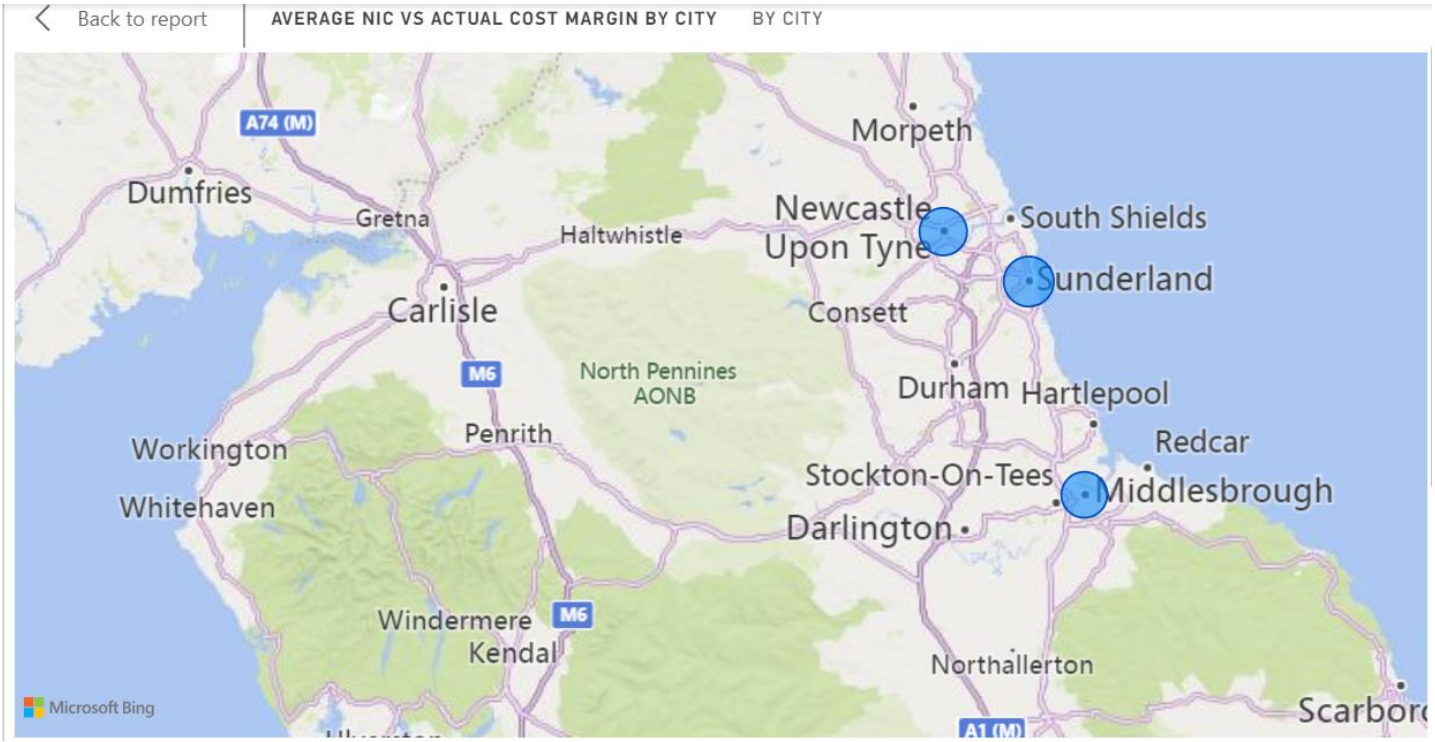


Figure 16: Waterfall chart showing variance of NIC and Actual Cost

(b) What city has the highest gap between NIC and Actual cost?

To identify the city with the highest gap between the two prices, this can be seen in the Map visual (***Figure 17***). Sunderland, with a gap of **£3.16,** has the highest margin. From the dataset, it was observed that Prices in Sunderland for the NIC (basic price quoted by manufacturer/drug tariff is higher than Actual Cost of the prescription. This means most prescriptions in **Sunderland** were either discounted by Government or other variables that affect actual cost such as dispensing cost, transportation etc, were reduced as a result of Government effort to improve Healthcare in the city compared to other cities that were reviewed.



CITY	Average NIC vs Actual Cost Margin
SUNDERLAND	3.16
NEWCASTLE UPON TYNE	2.82
LIVERPOOL	2.76
MIDDLESBROUGH	2.61

Figure 17: City and Average Margin between NIC and Actual Cost

5. Categorization of prescription by price

How affordable are the prescribed drugs and what factors can influence the cost of these drugs?

Firstly, we group the prices into different categories to represent their level of price affordability (***Figure 18***). Finding from **Figure 19** shows that ***123,059*** prescriptions are categorized as ***low price*** i.e. each of the prescriptions in this category is less than **£2.** Similarly, **5,554** prescriptions are classified as ***highly expensive*** i.e. each prescription **is at least £10.**

Category	Lower_level	Upper_Level
Highly expensive	£10.00	£1,000,000.00
Very expensive	£8.00	£10.00
Expensive	£6.00	£8.00
Moderately priced	£4.00	£6.00
Affordable price	£2.00	£4.00
Low price	£0.00	£2.00

Figure 18: categorization of prescription prices

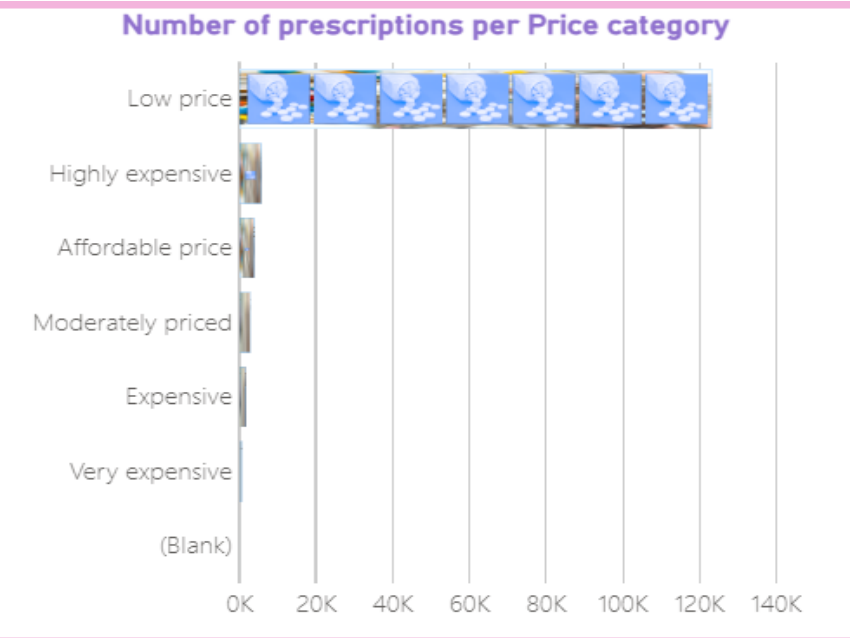


Figure 19: Infographic Designer visual: Number of Prescription and level of affordability

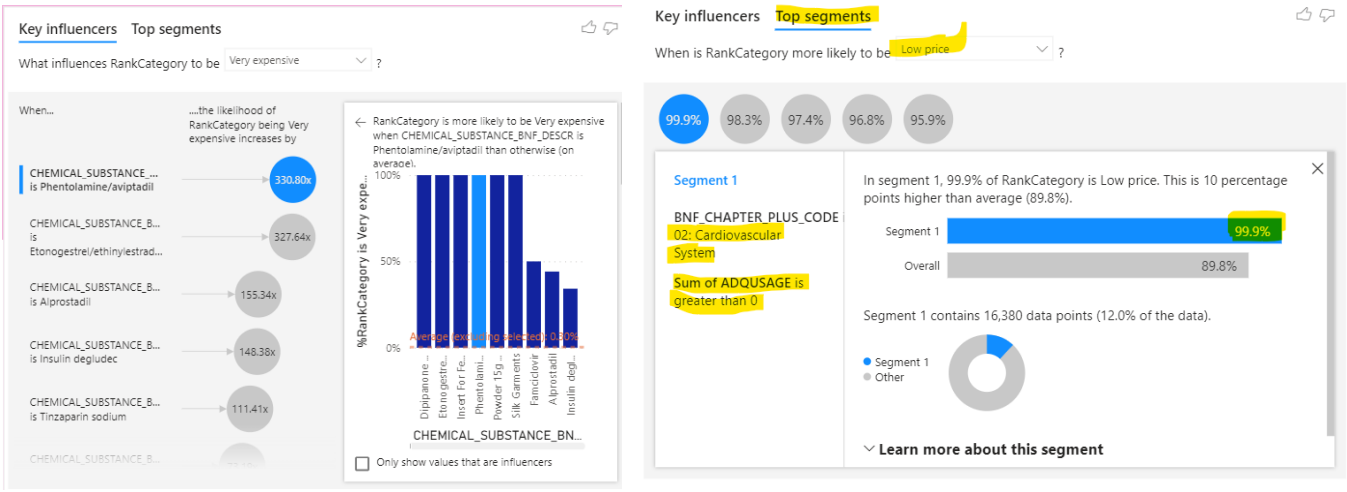
A table visual in ***Figure 20*** also displays the count of prescriptions in each cost category. From this visual, it is evident that majority of the presciptions analysed (**123,059 number of prescriptions**) is termed as "**low price**" i.e. the cost of the prescribed drug is not more than **£2** (that is, it is within the band **£** 0 and **£** 2).

RankCategory	Number of prescriptions
Low price	123059
Highly expensive	5554
Affordable price	3721
Moderately priced	2653
Expensive	1548
Very expensive	424
	2
Total	**136961**

Figure 20: Number of prescriptions per rank category

An advanced feature of Power BI, Artificial Intelligence, was used to predict factors that are likely to affect the affordability of a prescribed drug (Figure 21). From this visual, the likelihood of ***Phentolamine/aviptadil*** drug to be very expensive increases by ***330.80 times*** than any other drug.

In the same vein, 5 top segments were found for predicting the likelihood of ***Low-price*** drug. Figure 21(b) shows that in **segment 1**, 99.9% of prescription that are related to ***Cardiovascular system*** with ***ADQUSAGE*** above 0, is ***low price*** (i.e. the cost is not more than **£2**)



(a) (b)

Figure 21: Key influencers for cost prediction

SECTION 2

TECHNICAL REPORT

CHAPTER ONE

Introduction

Before any meaningful insight can be drawn from raw data, the data must be transformed, and certain pre-processing activities would be performed on the data. These pre-processing activities which were carried out in this project include the following: renaming column name, changing data type, removing duplicate records, merging columns containing similar contents, promoting first row as header (if column names are missing), removing null value or NA from the cells etc. Also, due to the huge volume of the dataset which is in excess of 17.7million for each month, the dataset has been filtered to **4 locations**: Liverpool, Middlesbrough, Sunderland and Newcastle upon Tyne and 2 Regions: ***North East and Yorkshire*** and ***North West.*** This filter was applied to the data source which was accessed online from NHS open Data portal to arrive at **70,956** records and **66,005** records for August 2023 and September 2023 respectively.

Other steps performed on the dataset include the following: changing the format of the file downloaded from zip file to csv file by following the steps: right click on the downloaded zip file, then, click on "***extract all…"*** option, loading the data into power BI, using power query editor for data transformation and writing of DAX and M language, merging the dataset of August with September 2023 to create better Business Intelligence visuals and for making comparison in terms of the KPI, creating dimension tables from the fact table. A star schema was created where the dimension tables are linked by ***Foreign Keys*** with the fact table that has the ***Primary key***. This creates ***Parent-child table relationship*** with ***cardinality*** that is usually ***one-to-many*** in the database design. The advantage of this **star-schema** is that it helps in **efficient processing of the query** and for better performance when visualizing the data in the dashboard. It also helps to maintain **data integrity** and **easy identification and correction of errors**.

All the steps carried out are discussed further below:

Data Source

The Dataset used is the English Prescription Data for the month of **August** and **September 2023** which was shared by the National Health Service (NHS), UK on her official website. For the month of August 2023, it can be accessed here: EPD_Aug2023_Dataset . It is important to note that the report is generated monthly and it contains millions of records with significant storage implication. Due to hardware limitation and time constraint to deliver this project, the dataset was filtered to focus on only two regions and 4 cities in the United Kingdom (UK) (***Figure 22***). The count of records after applying the filter is **70,956**.

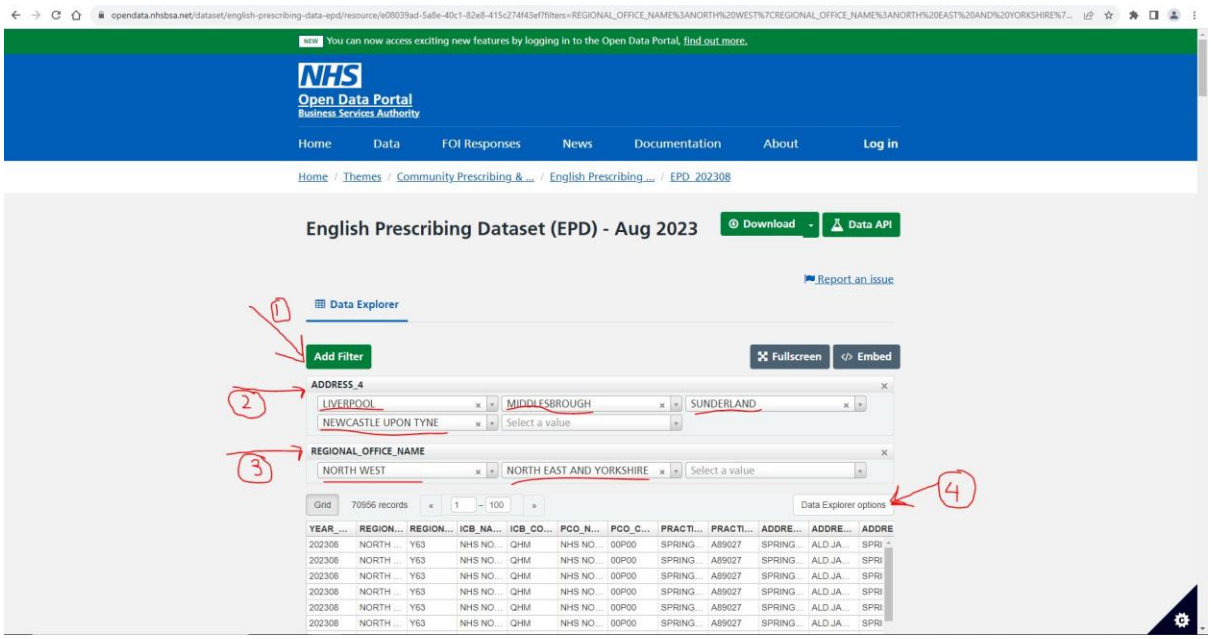


Figure 22: NHS EPD Dataset -August 2023 (filtered by 4 cities and 2 regions)

After the dataset has been downloaded from the NHS website, the file is automatically in a zip format in a folder on the local computer system.

The downloaded file needs to be unzipped before loading it in power BI. To do this, right-click on the zipped file, choose "Extract All…" option (***Figure 23***). This will convert the zip file into a csv file usable for loading and data visualization in Microsoft Power BI.

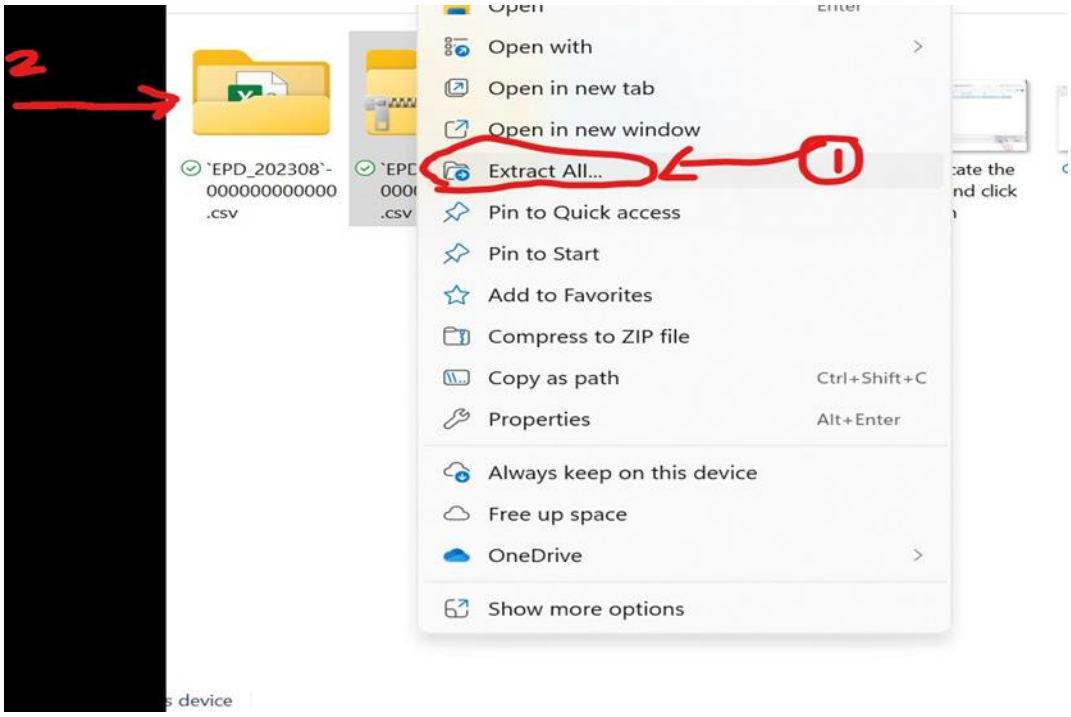


Figure 23: How to unzip the *downloaded* file

CHAPTER TWO

Loading the dataset into Power BI

After launching power BI application on the computer system, follow the steps below:

Click Home menu → Get data → Text/csv (***Figure 24***)

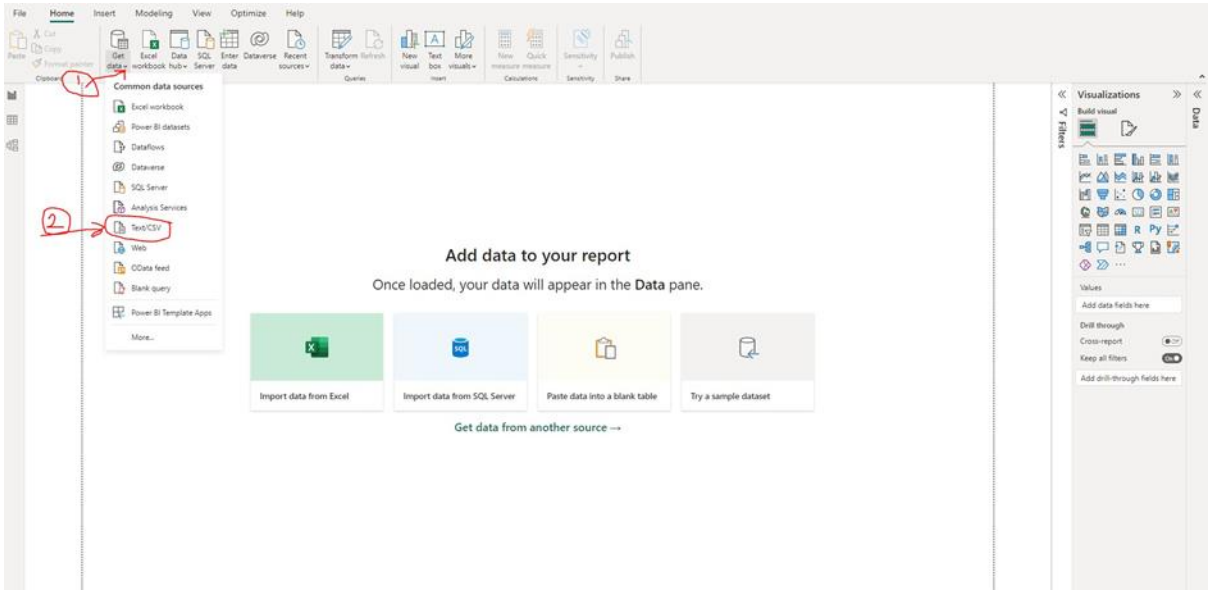


Figure 24: How to load dataset into Power BI

Alternatively, the data can be imported into power BI by clicking the "***Get data***" option from the first page after launching power BI application (Figure 25)

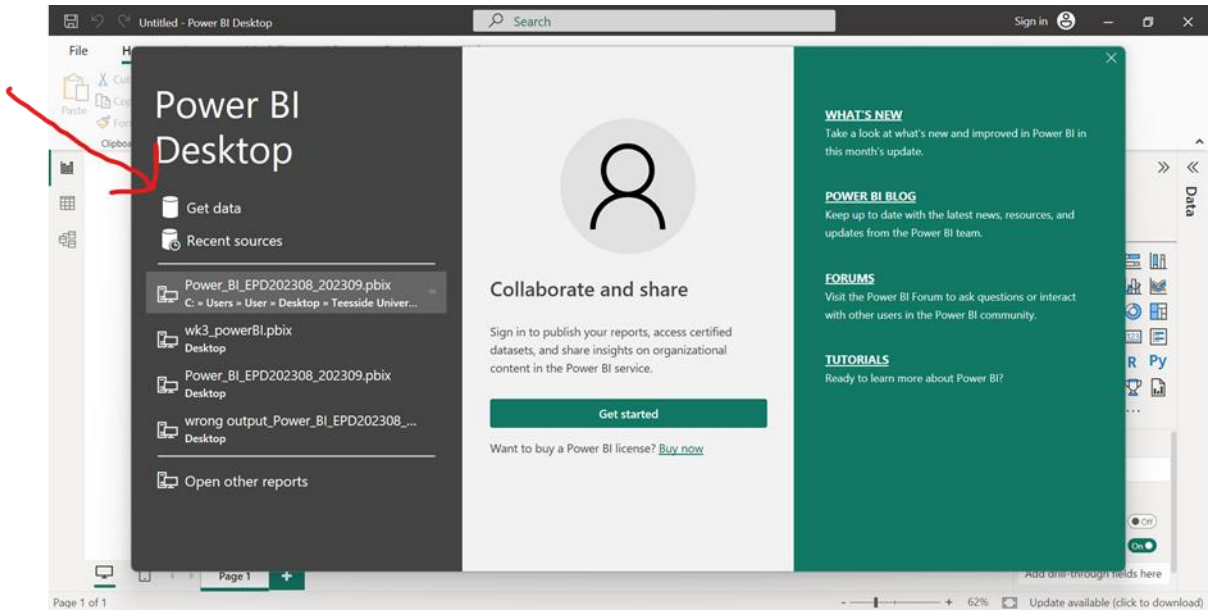


Figure 25: Option 2 for loading the dataset into power BI

Then, locate the folder where the dataset containing the csv file is stored on the computer.

Select the filename and click "***open***" (Figure 26)

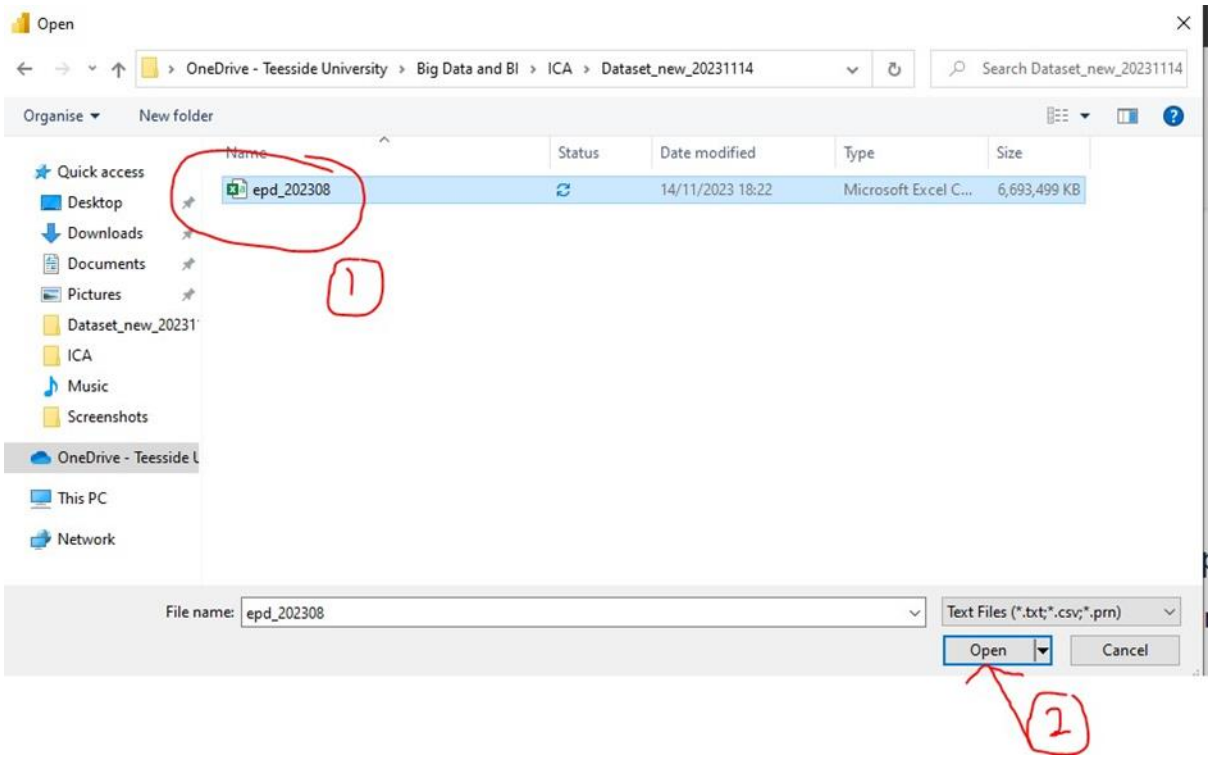


Figure 26: Locating the file for the dataset and opening it in Power BI

A preview of the file is displayed, then, click "***Load***" (**Figure 27**).

Repeat the same steps above to load the csv file for ***September 2023*** into Microsoft Power BI.

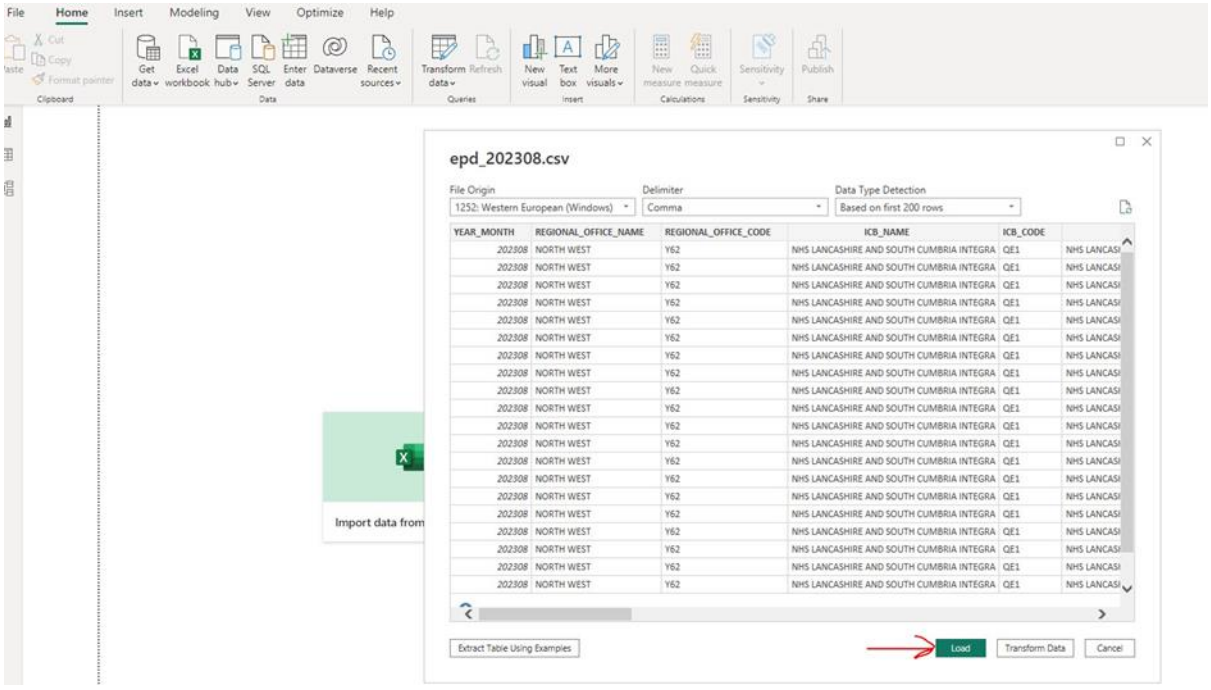


Figure 27: Loading of the selected file to power BI

Combining Two Datasets (August 2023 and September 2023)

In order to make the insight in the visualization richer, both the data i.e. csv files for ***August 2023*** and ***September 2023*** were combined.

The EPD for September 2023 was downloaded as well and the same filter used for the August 2023 was applied. Record count for this is **66,005.** This dataset for ***September 2023*** is appended to ***August 2023*** to form **one big dataset** containing data for both months (***Figure 28***). To do this, click on Transform menu, select the file name for August 2023 (on the left hand side), click 'Append Queries', (from the ***Append*** dialogue box, option "***Two tables***" is selected by default, click the drop down box and choose the second file to be appended, in this case, the file is the csv file for EPD September 2023.

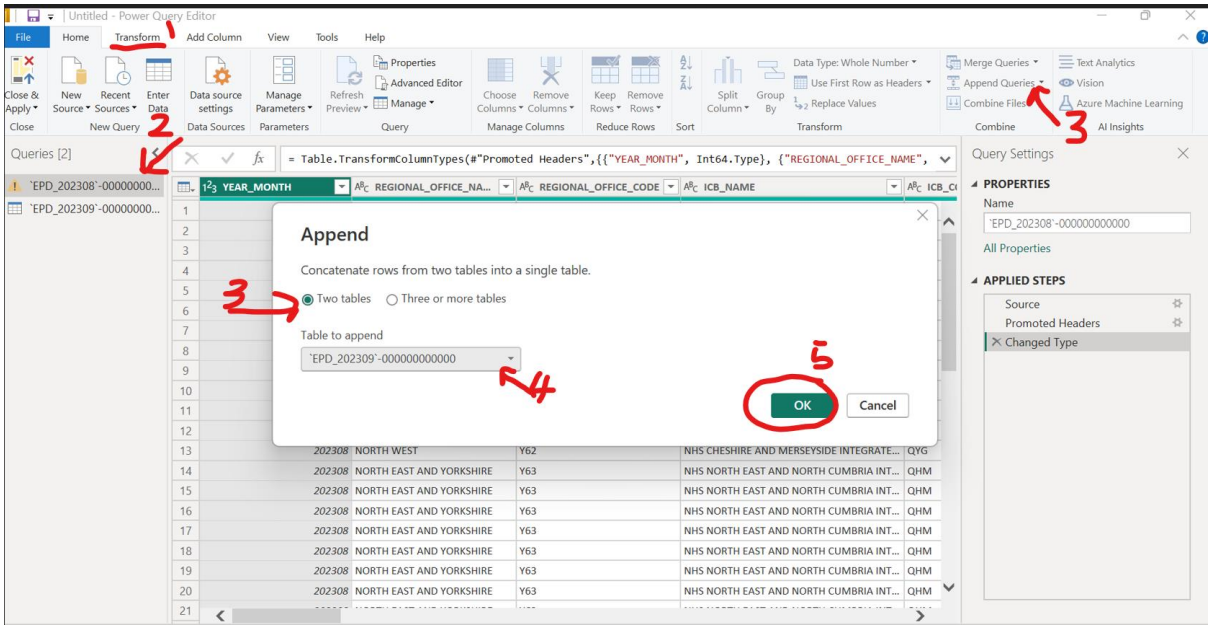


Figure 28: Appending September 2023 NHS EPD Dataset with August 2023

The total count of the combined records is ***136,961*** with the same number of columns for both months and the newly combined table i.e. **26 columns**.

CHAPTER THREE

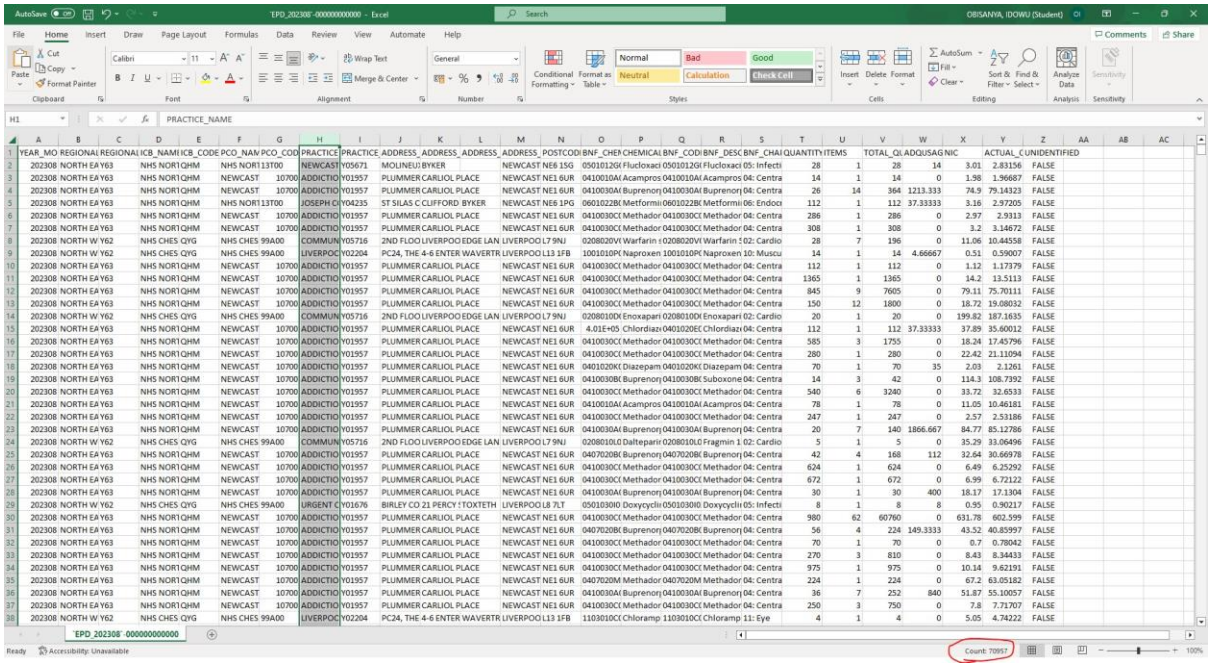
Data Pre-processing and cleansing

The data pre-processing and cleansing is the most crucial stage when analysing dataset.

After the dataset of interest has been identified and loaded into power BI, some activities were performed in order to produce data visualization with reliable insight. Some of these tasks include confirming data completeness, removing duplicate records, removing records containing null values or 'NA', promoting first row as header (if no header is available), changing the datatype as appropriate (e.g. if numbers are stored as text, then, this is corrected at this stage), renaming header with non-descriptive information ("Address 4" containing names of Town can be renamed as "**Cities**") etc.

Confirmation of data completeness

There is need to confirm the completeness of the dataset. To do this, select a ***non-empty column*** from the csv file and take note of the record count shown (***Figure 29***). This is done by comparing the record count in the file with the record count displayed online in the NHS website before the csv file was downloaded.



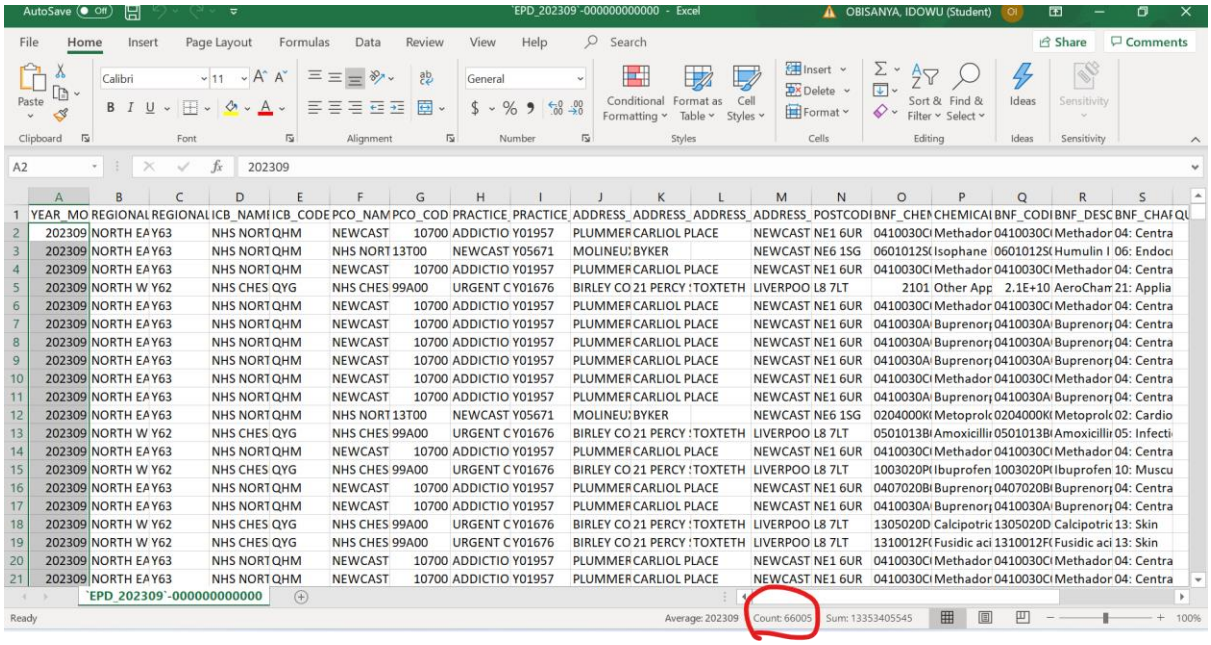


Figure 29: checking data completeness of the downloaded dataset (August and September, 2023)

Removing duplicate records

One of the pre-processing activities carried out in this project is to check for duplicate records. To do this, click ***Home*** menu, then, click ***Transform Data***. Select the file that needs to be cleaned of duplicate records, then right click on the file and select: '***Remove Duplicates'*** (***Figure 30***)

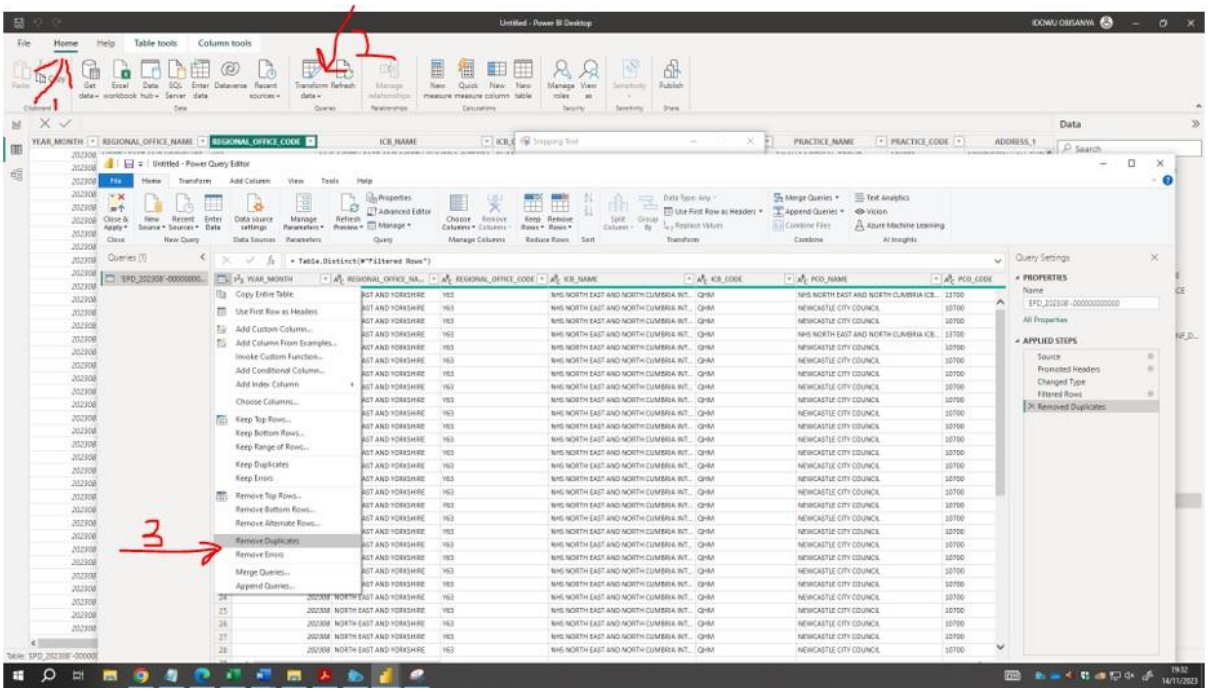


Figure 30: Removing Duplicate Record from the Dataset

Resolving column with missing values by merging similar columns

It was observed that there are some cells under '***Address 3'*** columns that are blank while some contain values (***Figure 31***).

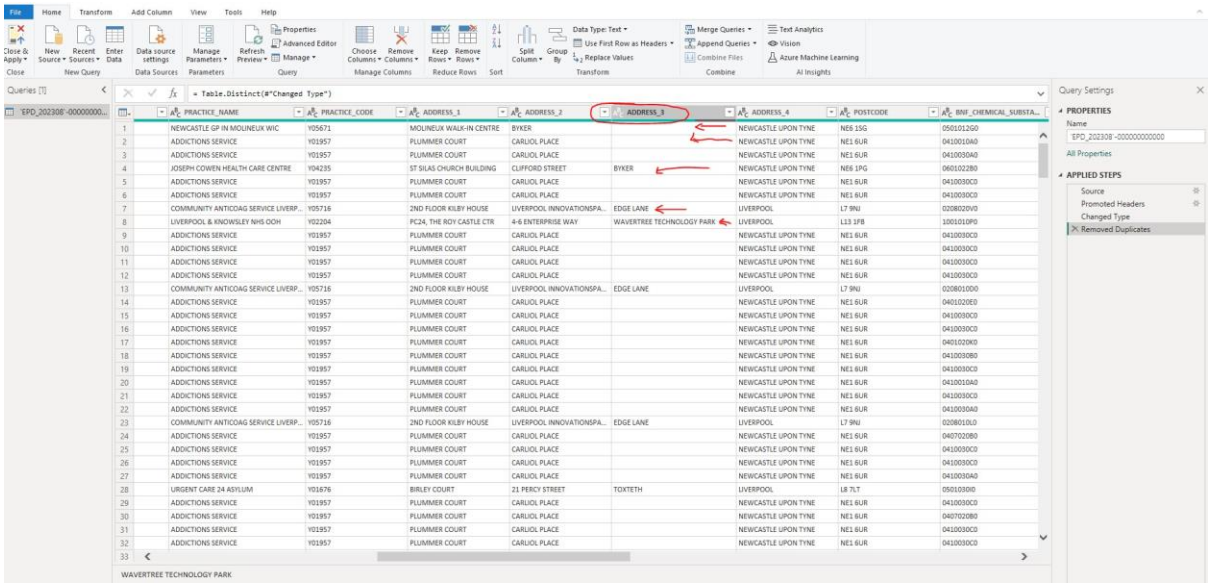


Figure 31: Address column with some blank cells

On cursory examination, it was observed that Address 3 gives further details about the description of address 2. Hence, both Address 2 and 3 were merged to form a new column. To merge these two columns, from the 'Transform Data' menu, select the columns from the header, click on 'Add column' menu, then, click, 'Merge columns', choose the desired separator (e.g. space) then, specify the name of the new column and click ok (***Figure 32***).

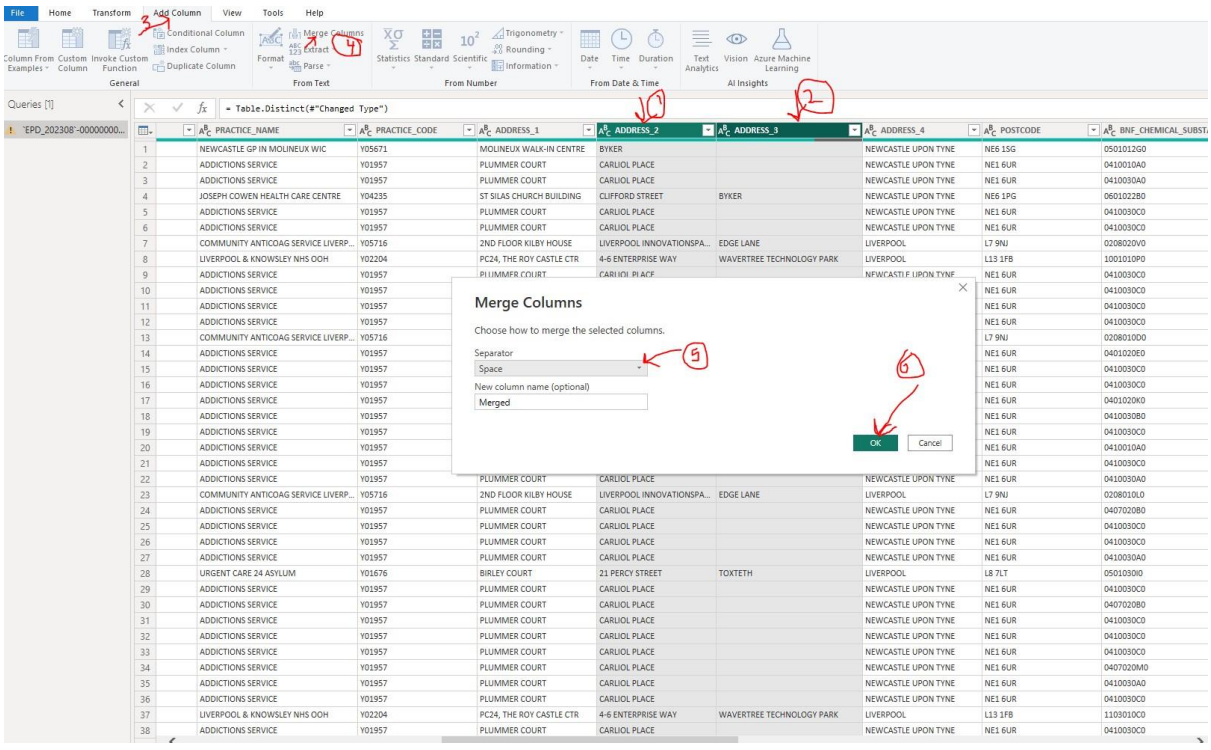


Figure 32: Merging two columns together

M language to merge the two columns together is:

```
= Table.AddColumn(#"Renamed Columns", "Merged", each Text.Combine({[ADDRESS_2], [ADDRESS_3]}, " "),
    type text)
```

Removing columns

Some columns are to be removed from this table. For instance, the two columns for Address 2 and Address 3 are to be removed (since a new column has been created where both columns were merged). To do this, select both columns, ***right click*** and click '***Remove column'*** option (***Figure 33***). The new column now replaces ***Address 2*** and ***Address 3***.

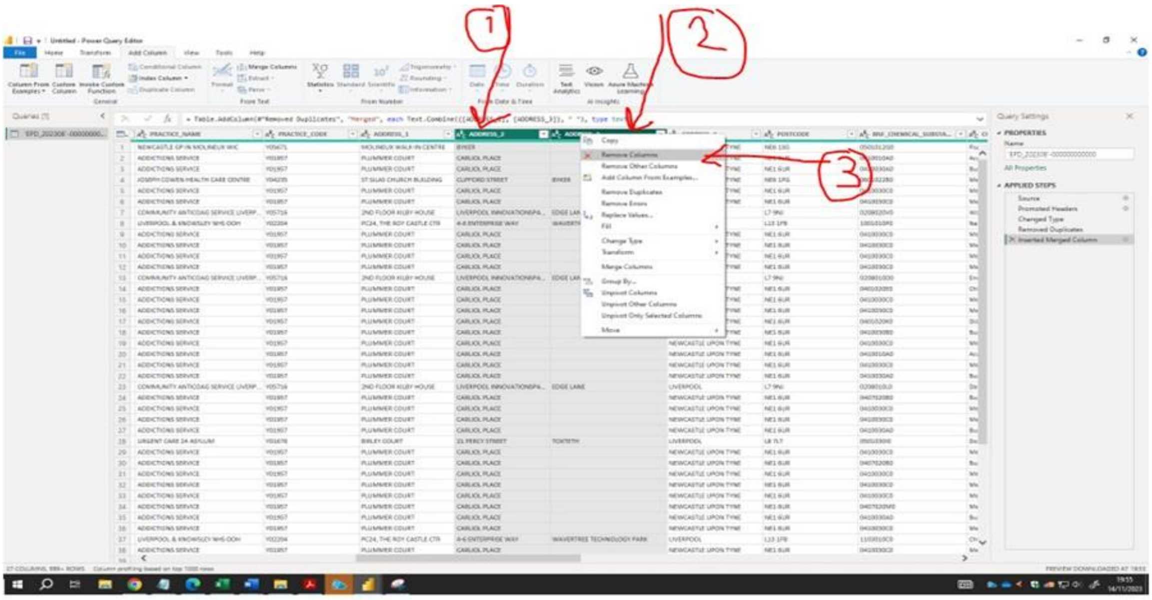


Figure 33: Removing existing columns that have some blank cells

Renaming columns

The columns with header ***Address_1***, ***Address_2*** and ***Address_4*** were renamed to reflect the contents in these columns. Hence, ***Address_1*** renamed as ***Location*** and ***Address_2*** renamed as ***Street*** and ***Address_4*** renamed as ***City*** (***Figure 34***). This is to give clearer information about what the content is and to make interpretation of the data easier.

RACTICE_NAME	PRACTICE_CODE	ADDRESS_1	ADDRESS_2	ADDRESS_3	ADDRESS_4	POSTCODE
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
E MEDICAL GROUP	A86003	NEWBIGGIN HALL SURGERY	TREVELYAN DRIVE	NEWBIGGIN HALL	NEWCASTLE UPON TYNE	NE5 4BS
NORTH SURGERY	A86006	ROSEWORTH SURGERY	27-29 ROSEWORTH AVENUE	GOSFORTH	NEWCASTLE UPON TYNE	NE3 1NB
NORTH SURGERY	A86006	ROSEWORTH SURGERY	27-29 ROSEWORTH AVENUE	GOSFORTH	NEWCASTLE UPON TYNE	NE3 1NB

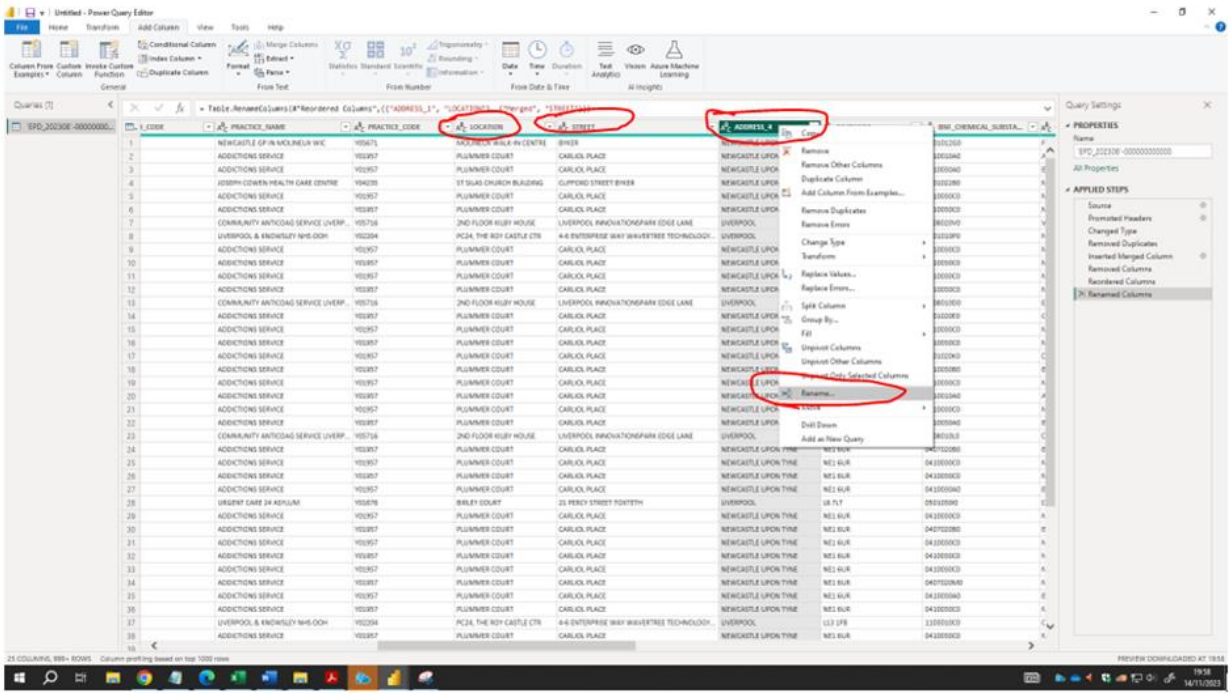


*Figure 34: **Renaming columns with non-descriptive header***

Data Formatting: Changing Data Type

Some columns in the dataset such as ***NIC*** and ***Actual_cost*** did not contain the appropriate data type. i.e. no currency symbol indicated for ***NIC*** while ***Actual_cost*** column has its values to ***5 decimal places*** (instead of 2 decimal places for currency) as well as ***missing currency symbol.*** The data types were changed from ***General*** data type to ***Currency*** and well formatted (***Figure 35***). To change the data type for the column name: ***NIC*** (***Net Ingredient Cost***) and **Actual_cost**, click ***Column tools***, select the column to be formatted, under ***Format*** option, select ***Currency***, click on the drop down highlighted as 4 below to select the currency symbol (i.e. £).

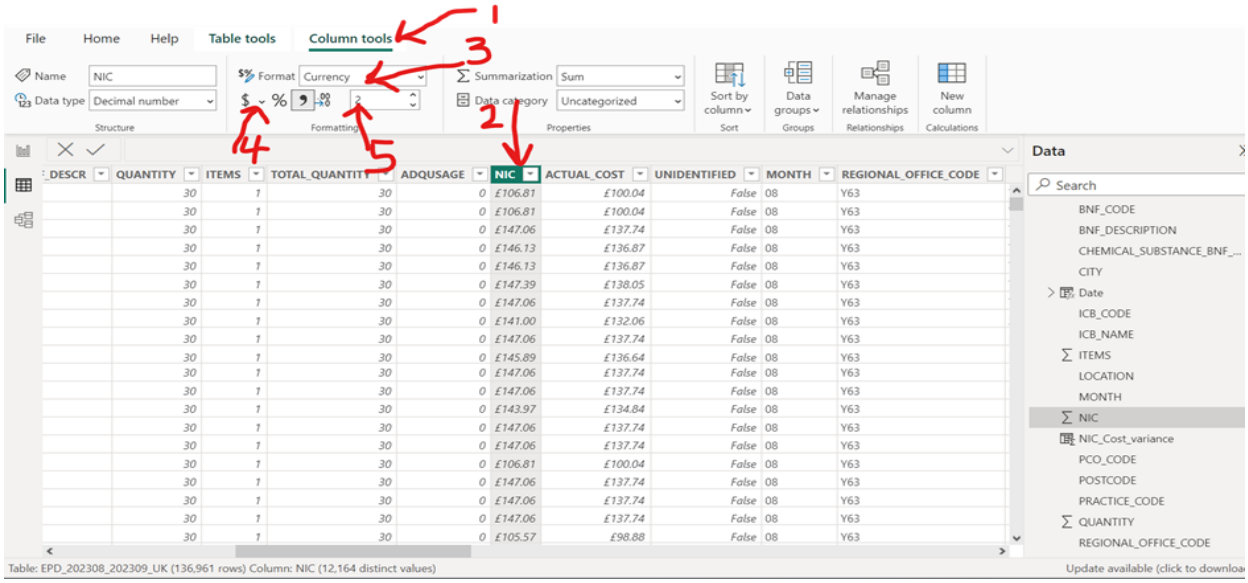


Figure 35: Data formatting (changing data type)

Splitting column

The column name ***Year_Month*** contains both the Year and month together. In order to enhance the analytics, we may need to create visual such as slicer on the Year only or month only. Hence, it would be good to split the column into two and format it as date accordingly (***Figure 36***). To achieve this, click Transform menu, select the table from the left, click on the column header to be split, click the drop down icon beside the Split Column option (labelled as step 2 in ***Figure 36***), type 4 inside the box for "Number of Characters" (since the value for ***Year*** is 4 characters e.g. 2023),choose first option under Split, then, under split into: choose "***columns***", then click ***ok***.

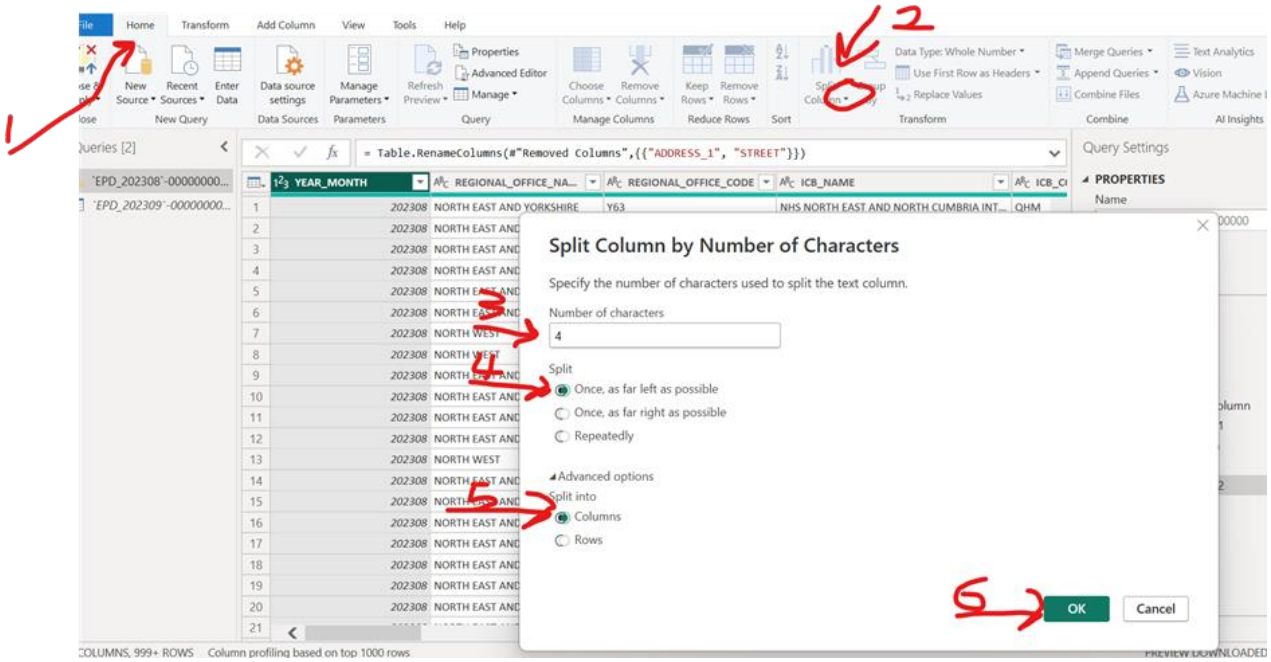


Figure 36: Splitting column into Year and Month columns

The M language to perform the column split is as shown below:

```
= Table.SplitColumn(Table.TransformColumnTypes(#"Renamed Columns2", {{"YEAR_MONTH", type text}},
    "en-US"), "YEAR_MONTH", Splitter.SplitTextByPositions({0, 4}, false), {"YEAR_MONTH.1",
    "YEAR_MONTH.2"})
```

Applied steps

Figure 37 shows all the applied steps at the data cleansing stage. This can be seen by clicking on the Transform menu, and from the Properties pane in Query settings on the right-hand side. From the listed steps, one can easily navigate to any previously applied steps and undo or remove any unwanted steps by clicking on the cross sign that shows when the step is selected by the user.

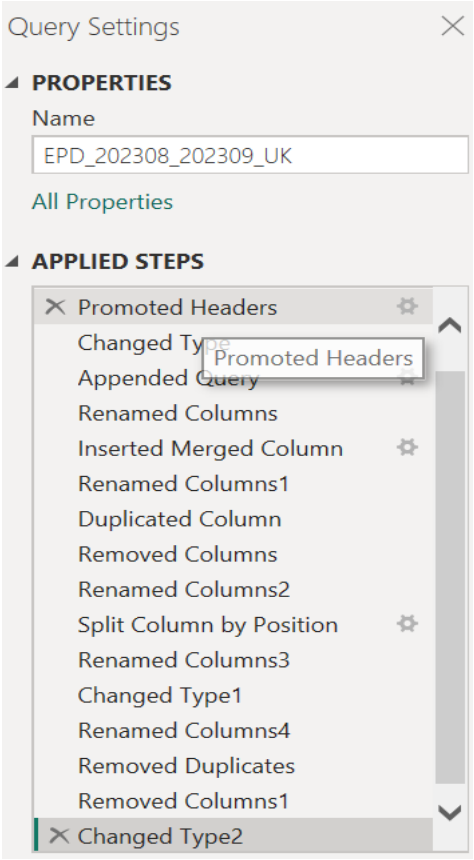


Figure 37: Applied steps during data cleansing

CHAPTER FOUR

BI Data Modelling

Data modelling is the process of creating a simplified diagram of a system and the data elements it contains. It uses text and symbols to indicate how data flows. It serves as the blueprint for designing a new database or reengineering a new application system (Stedman, 2005).

In this project, we have a fact table and from it, multiple dimension tables are created which are linked to the fact table through a ***foreign key***. The dimension table helps to answer the question: ***where, why, when*** and ***how is*** the data represented? The dimension table measures certain aspect of the entire data.

Using Star schema for Data Modelling

We use star-schema for modelling the data. The fact table and dimension tables are linked together through a ***one-to-many*** relationship (***Figure 38***).

The Fact table is named:

- ***EPD_202308_202309_UK***: This is the Parent table from where the dimension tables are created.

The dimension tables that were created from the fact table are:

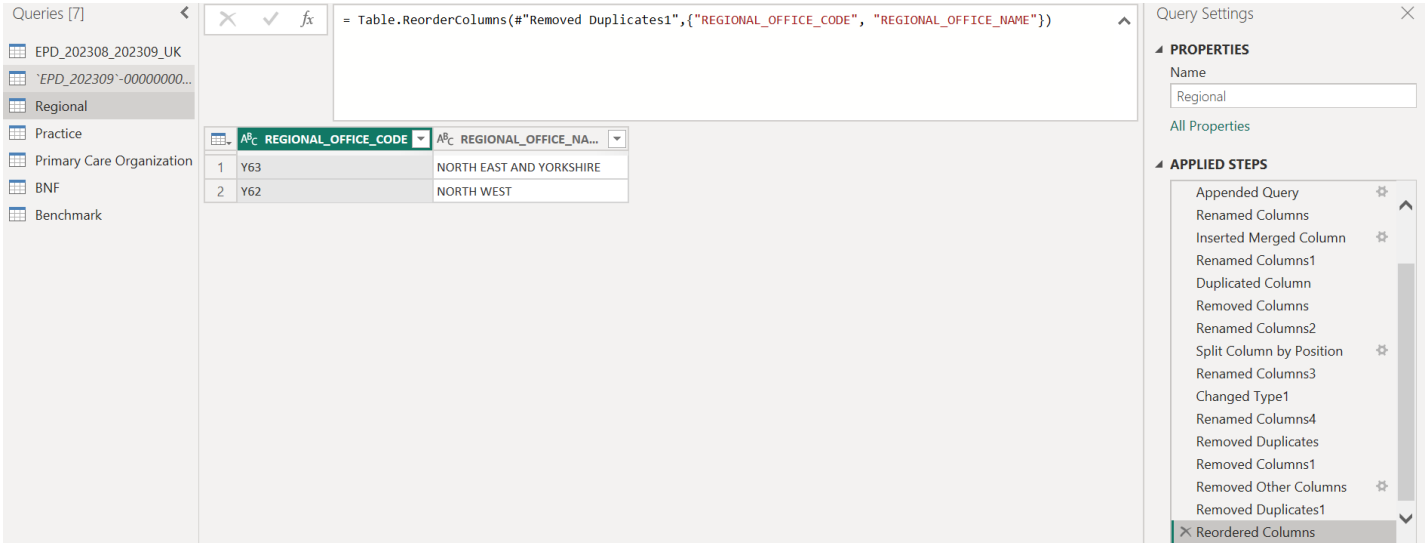
- ***Regional***: This table contains the Regional data. One can generate the analytics by Region using slicer visual. This dimension table contains only two columns: ***REGIONAL_OFFICE_CODE*** and ***REGIONAL_OFFICE_NAME***. This dimension table is created by duplicating the Fact table (EPD_202308_202309). Then select the two columns mentioned above, right click and choose "***remove other columns",*** then right click and choose remove duplicate. This will then leave only 2 Regions in the table that are uniquely identified by the ***Regional_office_code***.

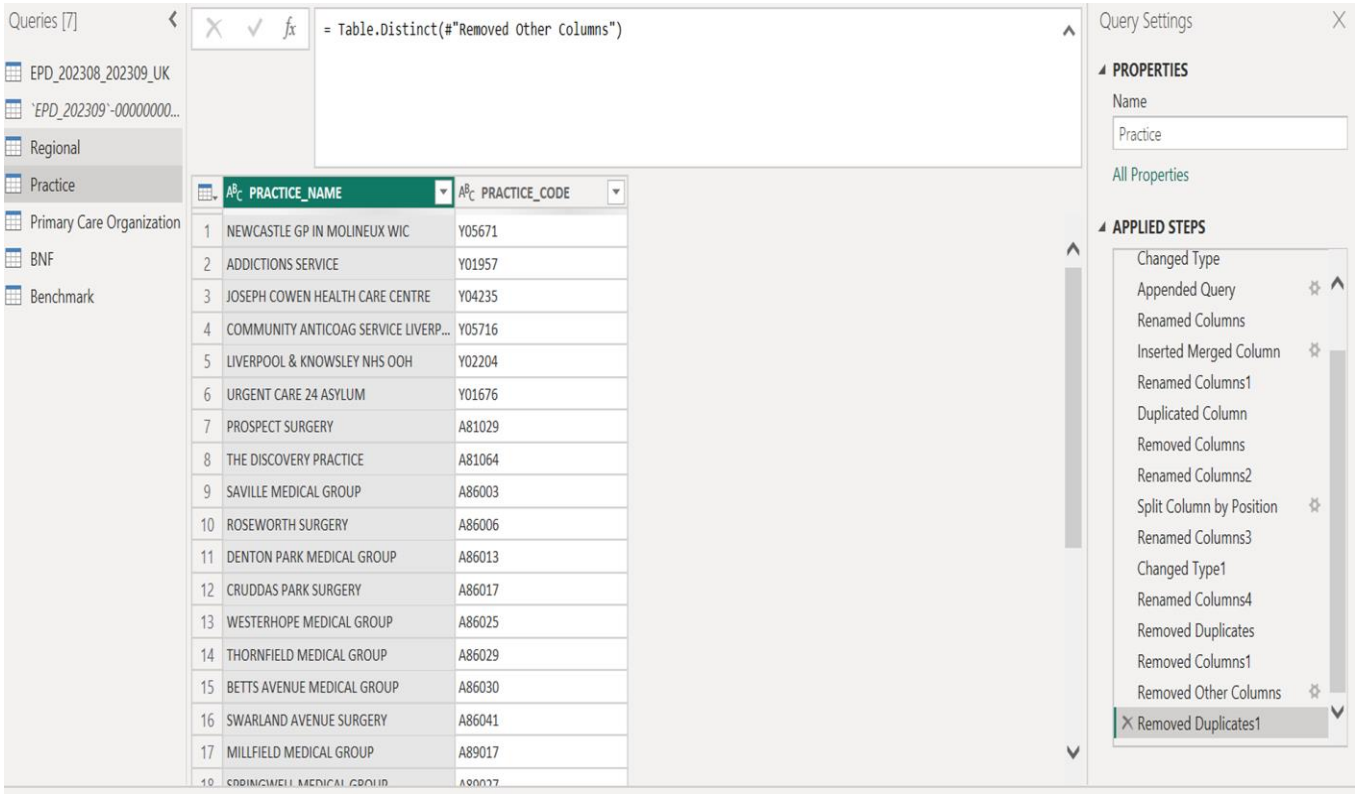
- **Practice:** This dimension table gives information about organization that employs one or more prescribers that issue prescription that may be dispensed in the community. For instance, GP Practice, Out-of-Hour services, a hospital department within an NHS Trust. Creating this dimension table is similar to the Regional table above. Firstly, the fact table is duplicated, choose all the columns for the Practice table, right click on the column header and click remove other columns.

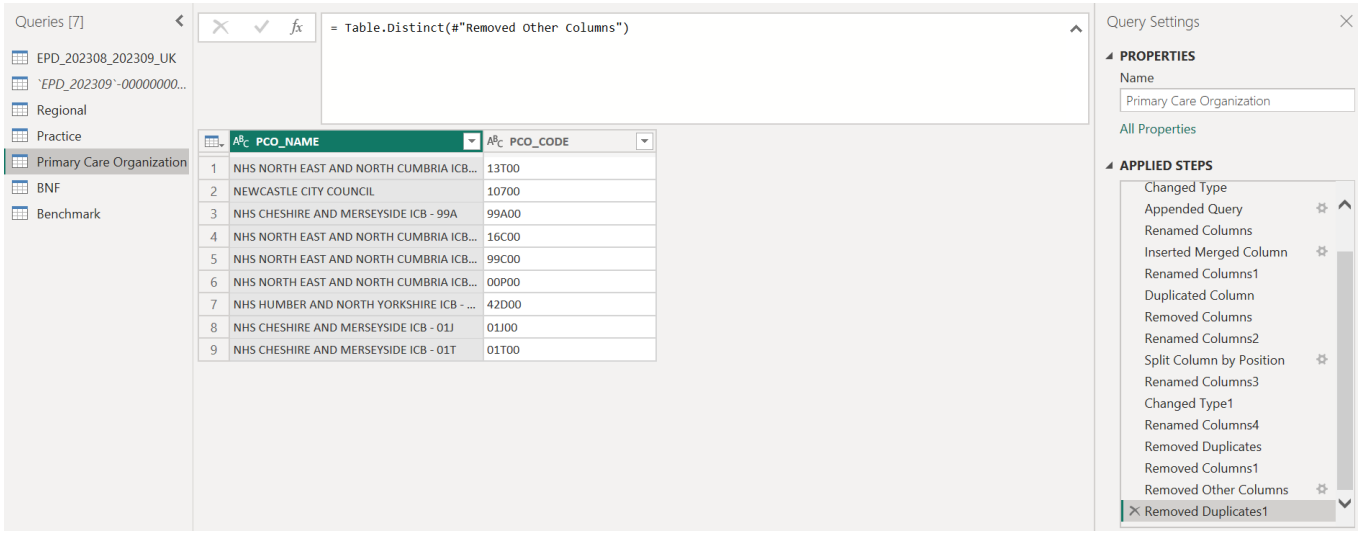
- **Primary Care Organization:** This shows information about NHS organization that either commissions or provide care services involving prescription that are dispensed in the community. For instance, SICBL (Sub Integrated Care Board Location), an NHS Trust

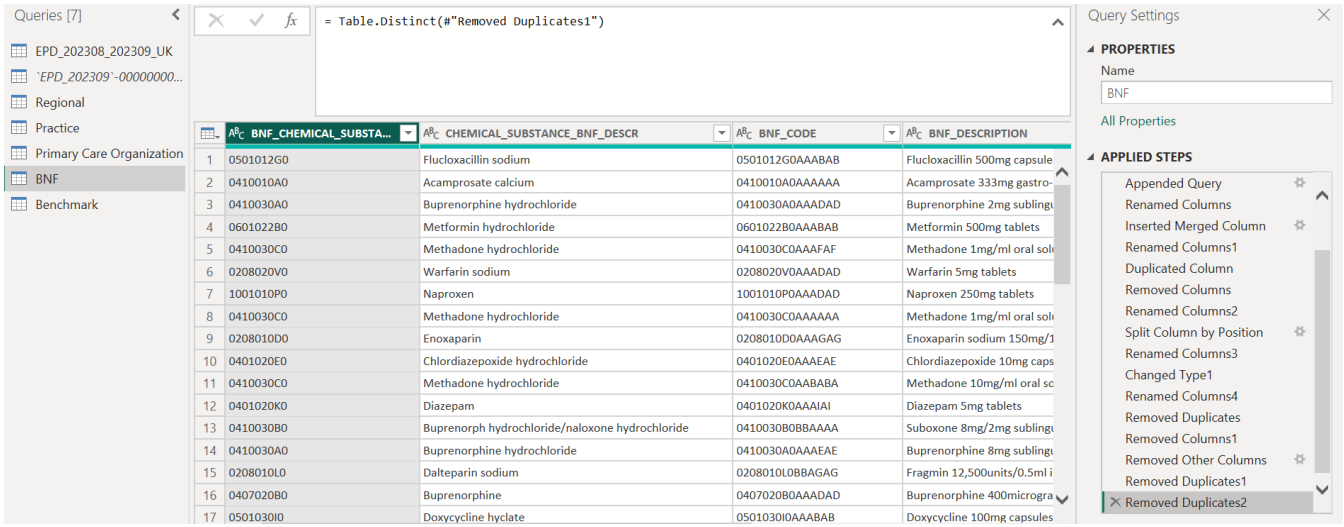
- **BNF:** This dimension table gives information about the British National Formulary (BNF) that describes the name of the main or active chemical substance in the drug, its description, the category of disease or organ of the body that it is used for such as Central Nervous System, Cardiovascular, Respiratory System etc.

Below are the steps carried out to create each of the dimension tables:


Queries [7]
EPD_202308_202309_UK
`EPD_202309`-00000000...
Regional
Practice
Primary Care Organization
BNF
Benchmark
= Table.ReorderColumns(#"Removed Duplicates1",{"REGIONAL_OFFICE_CODE", "REGIONAL_OFFICE_NAME"})
REGIONAL_OFFICE_CODE
REGIONAL_OFFICE_NA...
1 Y63 NORTH EAST AND YORKSHIRE
2 Y62 NORTH WEST
Query Settings
PROPERTIES
Name
Regional
All Properties
APPLIED STEPS
Appended Query
Renamed Columns
Inserted Merged Column
Renamed Columns1
Duplicated Column
Removed Columns
Renamed Columns2
Split Column by Position
Renamed Columns3
Changed Type1
Renamed Columns4
Removed Duplicates
Removed Columns1
Removed Other Columns
Removed Duplicates1
Reordered Columns


Queries [7]
EPD_202308_202309_UK
`EPD_202309`-00000000...
Regional
Practice
Primary Care Organization
BNF
Benchmark
= Table.Distinct(#"Removed Other Columns")
PRACTICE_NAME
PRACTICE_CODE
1 NEWCASTLE GP IN MOLINEUX WIC Y05671
2 ADDICTIONS SERVICE Y01957
3 JOSEPH COWEN HEALTH CARE CENTRE Y04235
4 COMMUNITY ANTICOAG SERVICE LIVERP... Y05716
5 LIVERPOOL & KNOWSLEY NHS OOH Y02204
6 URGENT CARE 24 ASYLUM Y01676
7 PROSPECT SURGERY A81029
8 THE DISCOVERY PRACTICE A81064
9 SAVILLE MEDICAL GROUP A86003
10 ROSEWORTH SURGERY A86006
11 DENTON PARK MEDICAL GROUP A86013
12 CRUDDAS PARK SURGERY A86017
13 WESTERHOPE MEDICAL GROUP A86025
14 THORNFIELD MEDICAL GROUP A86029
15 BETTS AVENUE MEDICAL GROUP A86030
16 SWARLAND AVENUE SURGERY A86041
17 MILLFIELD MEDICAL GROUP A89017
Query Settings
PROPERTIES
Name
Practice
All Properties
APPLIED STEPS
Changed Type
Appended Query
Renamed Columns
Inserted Merged Column
Renamed Columns1
Duplicated Column
Removed Columns
Renamed Columns2
Split Column by Position
Renamed Columns3
Changed Type1
Renamed Columns4
Removed Duplicates
Removed Columns1
Removed Other Columns
Removed Duplicates1


Queries [7]
EPD_202308_202309_UK
`EPD_202309`-00000000...
Regional
Practice
Primary Care Organization
BNF
Benchmark
= Table.Distinct(#"Removed Other Columns")
PCO_NAME
PCO_CODE
1 NHS NORTH EAST AND NORTH CUMBRIA ICB... 13T00
2 NEWCASTLE CITY COUNCIL 10700
3 NHS CHESHIRE AND MERSEYSIDE ICB - 99A 99A00
4 NHS NORTH EAST AND NORTH CUMBRIA ICB... 16C00
5 NHS NORTH EAST AND NORTH CUMBRIA ICB... 99C00
6 NHS NORTH EAST AND NORTH CUMBRIA ICB... 00P00
7 NHS HUMBER AND NORTH YORKSHIRE ICB - ... 42D00
8 NHS CHESHIRE AND MERSEYSIDE ICB - 01J 01J00
9 NHS CHESHIRE AND MERSEYSIDE ICB - 01T 01T00
Query Settings
PROPERTIES
Name
Primary Care Organization
All Properties
APPLIED STEPS
Changed Type
Appended Query
Renamed Columns
Inserted Merged Column
Renamed Columns1
Duplicated Column
Removed Columns
Renamed Columns2
Split Column by Position
Renamed Columns3
Changed Type1
Renamed Columns4
Removed Duplicates
Removed Columns1
Removed Other Columns
Removed Duplicates1


Queries [7]
EPD_202308_202309_UK
`EPD_202309`-00000000...
Regional
Practice
Primary Care Organization
BNF
Benchmark
= Table.Distinct(#"Removed Duplicates1")
BNF_CHEMICAL_SUBSTA...
CHEMICAL_SUBSTANCE_BNF_DESCR
BNF_CODE
BNF_DESCRIPTION
1 0501012G0 Flucloxacillin sodium 0501012G0AAABAB Flucloxacillin 500mg capsule
2 0410010A0 Acamprosate calcium 0410010A0AAAAAA Acamprosate 333mg gastro-
3 0410030A0 Buprenorphine hydrochloride 0410030A0AAADAD Buprenorphine 2mg sublingu
4 0601022B0 Metformin hydrochloride 0601022B0AAABAB Metformin 500mg tablets
5 0410030C0 Methadone hydrochloride 0410030C0AAAFAF Methadone 1mg/ml oral sol
6 0208020V0 Warfarin sodium 0208020V0AAADAD Warfarin 5mg tablets
7 1001010P0 Naproxen 1001010P0AAADAD Naproxen 250mg tablets
8 0410030C0 Methadone hydrochloride 0410030C0AAAAAA Methadone 1mg/ml oral sol
9 0208010D0 Enoxaparin 0208010D0AAAGAG Enoxaparin sodium 150mg/1
10 0401020E0 Chlordiazepoxide hydrochloride 0401020E0AAAEAE Chlordiazepoxide 10mg caps
11 0410030C0 Methadone hydrochloride 0410030C0AABABA Methadone 10mg/ml oral sc
12 0401020K0 Diazepam 0401020K0AAAIAI Diazepam 5mg tablets
13 0410030B0 Buprenorph hydrochloride/naloxone hydrochloride 0410030B0BBAAAA Suboxone 8mg/2mg sublingu
14 0410030A0 Buprenorphine hydrochloride 0410030A0AAAEAE Buprenorphine 8mg sublingu
15 0208010L0 Dalteparin sodium 0208010L0BBAGAG Fragmin 12,500units/0.5ml i
16 0407020B0 Buprenorphine 0407020B0AAADAD Buprenorphine 400microgra
17 050103OI0 Doxycycline hyclate 050103OI0AAABAB Doxycycline 100mg capsules
Query Settings
PROPERTIES
Name
BNF
All Properties
APPLIED STEPS
Appended Query
Renamed Columns
Inserted Merged Column
Renamed Columns1
Duplicated Column
Removed Columns
Renamed Columns2
Split Column by Position
Renamed Columns3
Changed Type1
Renamed Columns4
Removed Duplicates
Removed Columns1
Removed Other Columns
Removed Duplicates1
Removed Duplicates2

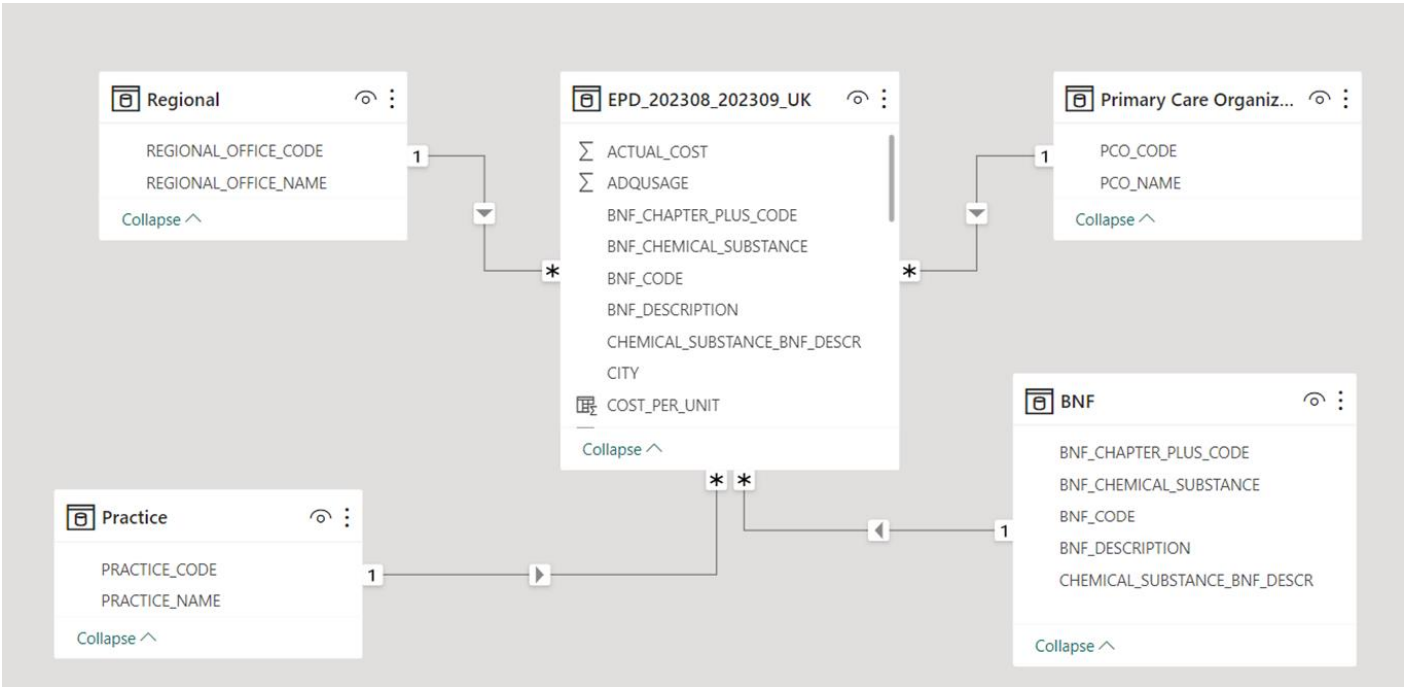


Figure 38: Star-schema

DAX and M Language

There are certain measures written when developing the dashboard. These were created use DAX (***DAta eXpression***) and M Languages.

AVERAGE_COST (New Measure)

This is used to calculate the average of all the actual cost of prescriptions given to patients in a certain period.

This derived clicking on New measure option in power BI and type the DAX formula below:

```
AVERAGE_COST = AVERAGE(EPD_202308_202309_UK[ACTUAL_COST])
```

AVERAGE_NIC (New measure)

This is used to calculate the average of all the NIC (Net Ingredient Cost) of prescriptions given to patients in a certain period.

```
AVERAGE_NIC = AVERAGE(EPD_202308_202309_UK[NIC])
```

NIC_Cost_variance (New column)

We want to know the gap between the two costs (***NIC*** and ***Actual_Cost***). A simplified way is to create a new column and type the formula below:

```
NIC_Cost_variance = [NIC]-[ACTUAL_COST]
```

AverageNIC_costMargin (New measure)

This measure is used to calculate the average of variance between NIC and Actual Cost. This gives a general indication of how wide apart the two costs are:

AverageNIC_costMargin = AVERAGE(EPD_202308_202309_UK[NIC_Cost_variance])

COST_PER_UNIT (New column)

This is used to establish how much a unit quantity of the drug cost. For instance, if 5 units quantity of a drug costs £10, then COST_PER_UNIT is: £10/5 i.e. £2.

The formular used is:

COST_PER_UNIT = 'EPD_202308_202309_UK'[ACTUAL_COST]/'EPD_202308_202309_UK'[TOTAL_QUANTITY]

RankCategory (New column)

Here, we used **M Language** to create a column for the cost category.

1. RankCategory =
2. VAR CatVar=[COST_PER_UNIT]
3. RETURN
4. CALCULATE(VALUES(Benchmark[Category]),CatVar>Benchmark[Lower_level]&&CatVar<=Benchmark[Upper_Level])

Concatenate Year and Month column to Date column

DAX formula for Date function was used to concatenate ***Year*** and ***Month*** columns. DAX functions are useful for creating Business intelligence in Microsoft Power BI (Brett, 2017)

```
1 Date = DATE([YEAR],[MONTH],30)
```

M Language used for creating category of Actual_cost of prescription into :

low price: £0 – £ 2

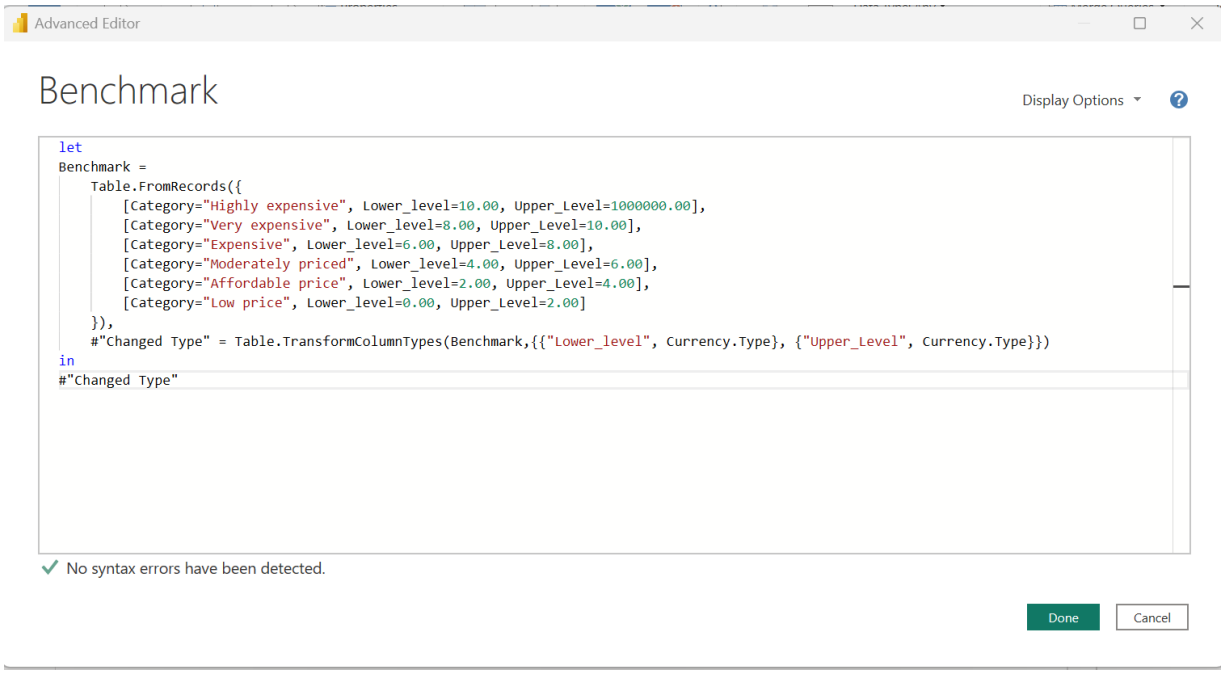
affordable price: £2 – £ 4

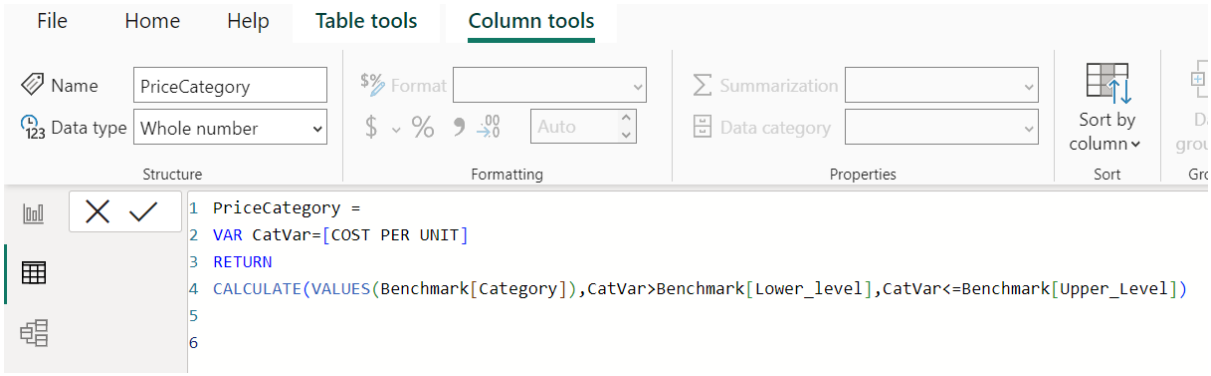
moderately priced: £4 – £6

expensive: £6 – £8

very expensive: £8 – £10

Highly expensive: £10-£1,000,000





Error encountered

1. **Circular dependency error when writing M Language**
   An error called circular dependency was encountered when calculating the measure cost per unit. The input should be column: Actual_cost divided by column: Total Quantity, but, the measure Total_Qty was used erroneously. This has been corrected and it is now OK.

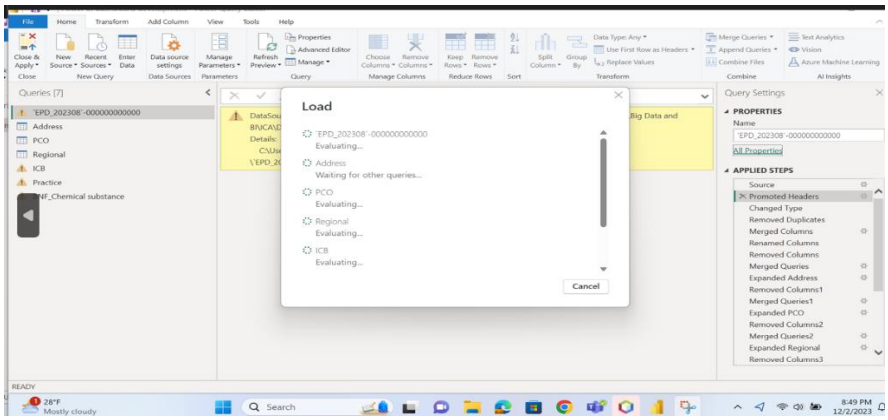
Correct formula:

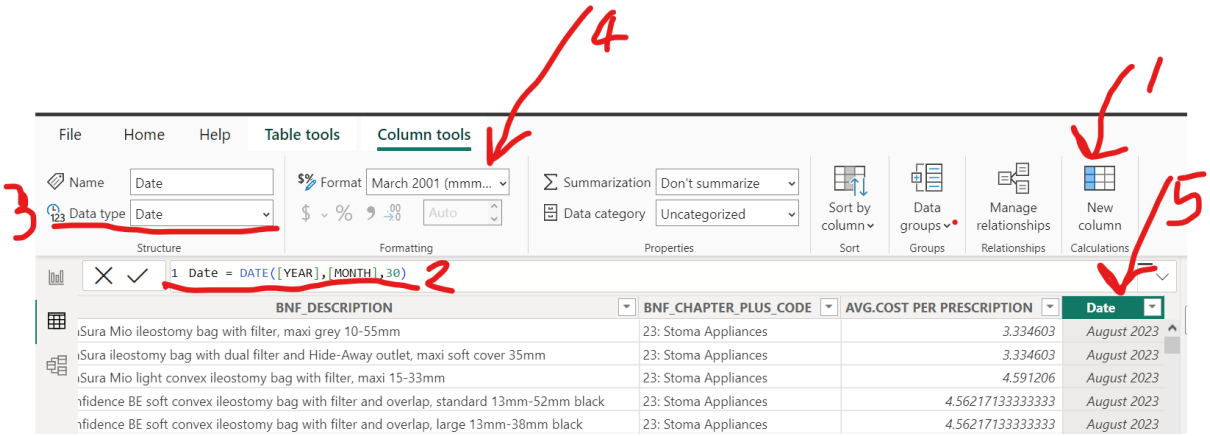
```
COST_PER_UNIT = 'EPD_202308_202309_UK'[ACTUAL_COST]/'EPD_202308_202309_UK'[TOTAL_QUANTITY]
```

2. **Data Loading to power BI**

   Another error encountered was the network error when attempting to load the source file from the work folder created on OneDrive when there was connectivity problem.

   The error was corrected by moving the downloaded csv file to the local system used and load into power BI from there.



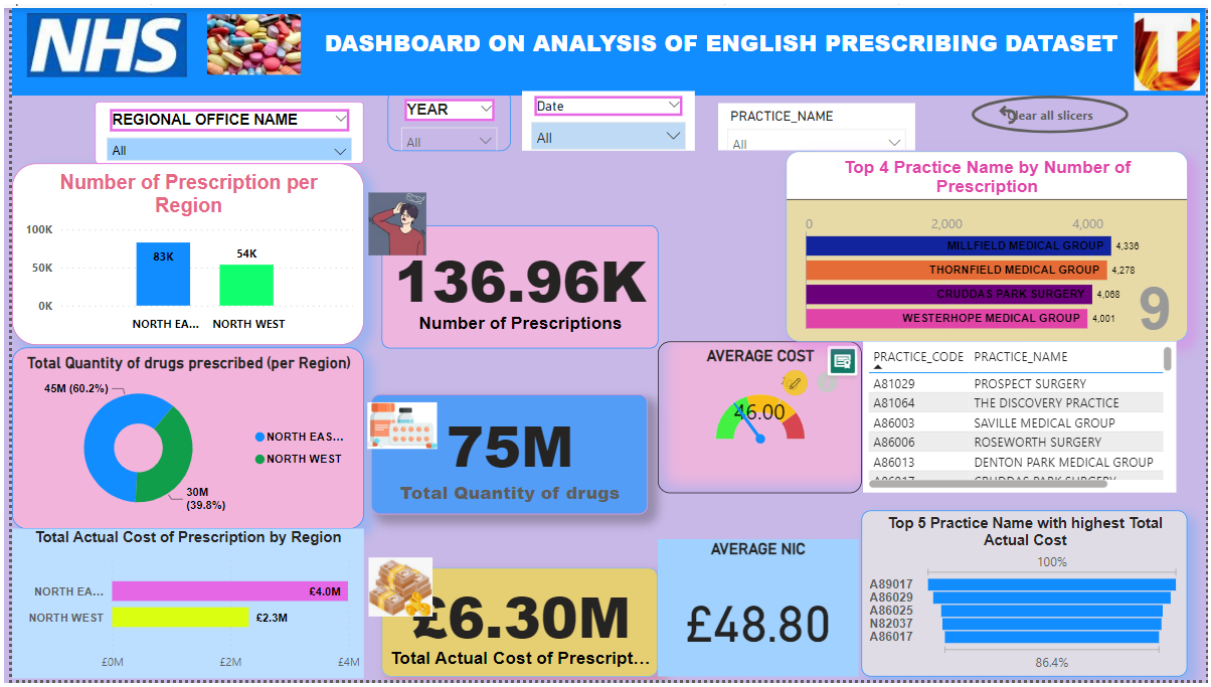


CHAPTER FIVE

Overview Dashboard

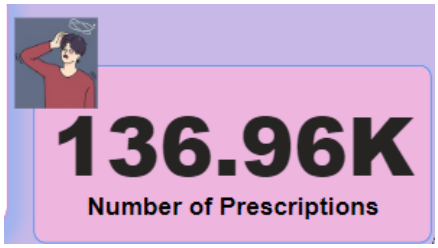
1. Prescription by Region and Practice

The first dashboard displays some related analysis about prescription by Region and by Practice. Different visuals in Microsoft power BI were used to present the information such as Stacked column chat, Donut chat, stacked bar chat, card, Advanced Gauge, Animated Bar chart race, Table and Funnel. At the top of the page are various slice visuals which user can choose to view the report based on different dimensions such as Regional office name, Year, Date and Practice Name. There is also a reset button at the top right-hand corner to clear all filters.



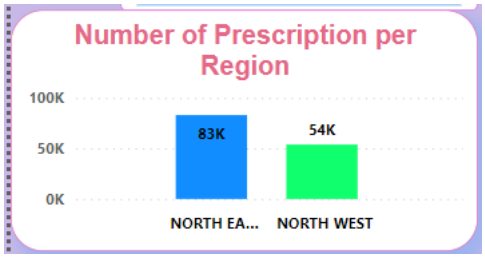
Number of Prescriptions

This uses a card visual to display the total count of prescription for all the dataset. Total count of prescription given is 136.96K for both months (August and September,2023) including all the regions.



Number of Prescriptions per Region

This Is presented using a stacked column chart. North West Region has **54K** while North East and Yorkshire region has **83K**.

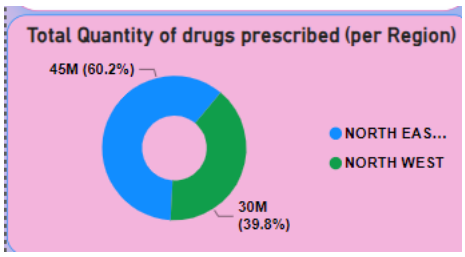


Total Quantity of drugs

We used a card visual to send the total quantity of drugs prescribed. It shows that throughout the period covered in the dataset, **75 million** quantities of drugs were prescribed to patients.

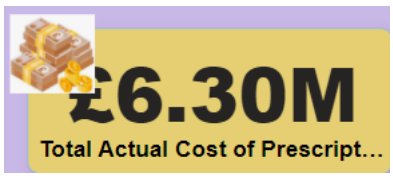
Total quantity of drugs prescribed per Region

From the donut visual, quantity of drugs prescribed per region was 45 million and 30 million for North East and Yorkshire and North West region respectively.



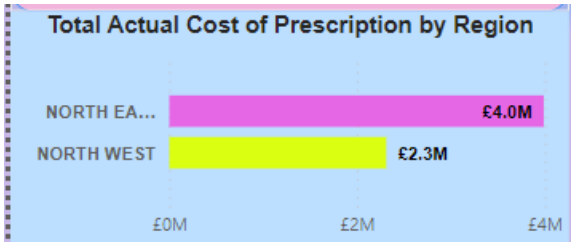
Total Actual cost of prescription

It costs a total sum of **£6.30 million** being the Actual cost of prescription to patients in the united Kingdom across the **4 selected cities** in **North East** and **Yorkshire and North West** regions.



Total Actual cost of prescription per Region

Based on the Stacked bar chart visual below, Actual cost of prescription in North East and Yorkshire **£4.0 million** and North West region is **£2.3 million**



Top 4 Practice Name by number of prescription

We also show the top 4 Practice names based on the total count of prescription they give within the two months. This is illustrated using Animated Bar Chart Race. From the below visual, the leading practice names are: MILLFIELD MEDICAL GROUP (4,336), THORNFIELD MEDICAL GROUP (4,278), CRUDDAS PARK SURGERY (4,068), WESTERHOPE MEDICAL GROUP (4,001)

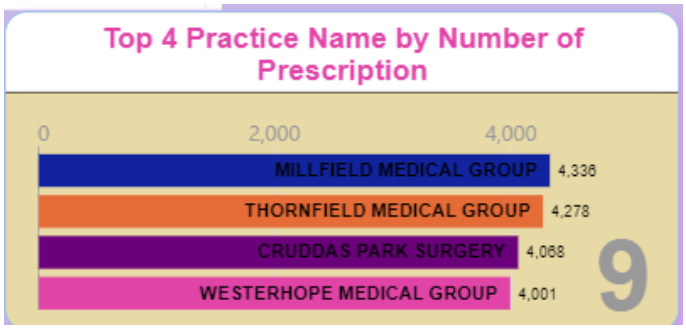
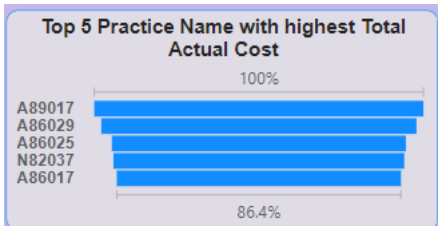


Table for Parctice Code Lookup

PRACTICE_CODE	PRACTICE_NAME
Y01957	ADDICTIONS SERVICE
A86030	BETTS AVENUE MEDICAL GROUP
N84020	BLUNDELLSANDS SURGERY
Y05716	COMMUNITY ANTICOAG SERVICE
A86017	CRUDDAS PARK SURGERY

Top 5 Practice Name with Highest Total Actual Cost

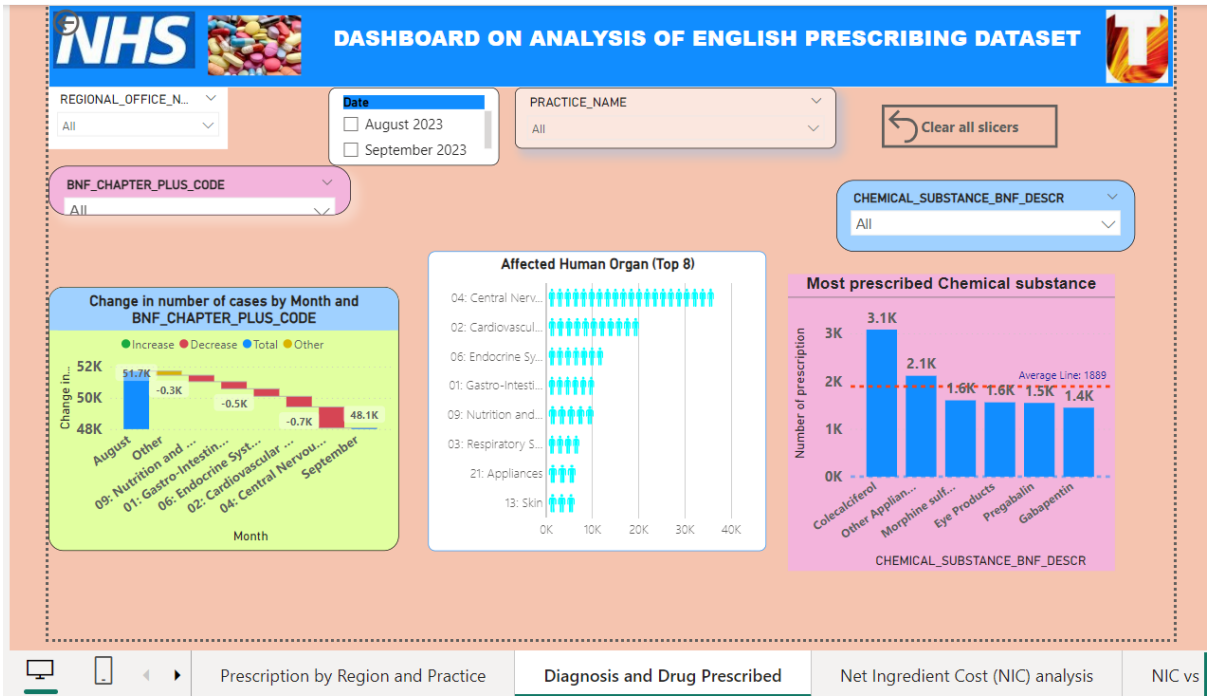
Listed in the funnel visual below are the top 5 practice name that have the highest total actual cost.



User can also filter based on any of the slicers at the top of the page to view

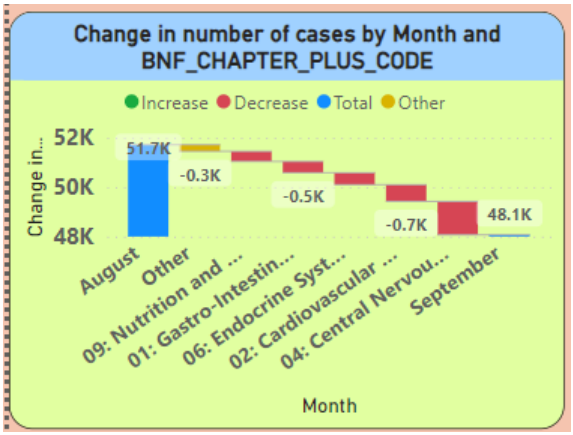
2. Diagnosis and Drug Prescribed

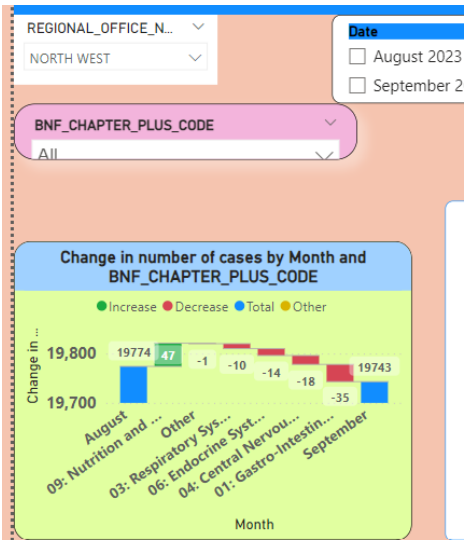
This second dashboard gives insight on the Diagnosis and Drug prescribed across the 4 cities.



Change in Number of Cases by month and BNF Chapter Plus code

A waterfall chart is used to show the level of change in the cases recorded between August, 2023 and September, 2023. From the visualization here, it shows that Cardiovascular related cases dropped by 700 cases based on the prescription given in August compared to the month of September, 2023.
One can also choose a region to view its report. From the image on the right, cases of diagnosis that relate to Nutrition and Blood increased by 47.

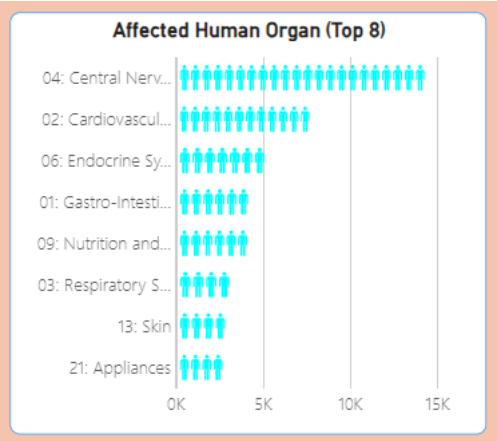




Top cases of human organ affected

We use an infographic designer visual to present the organs of the body that require medication.

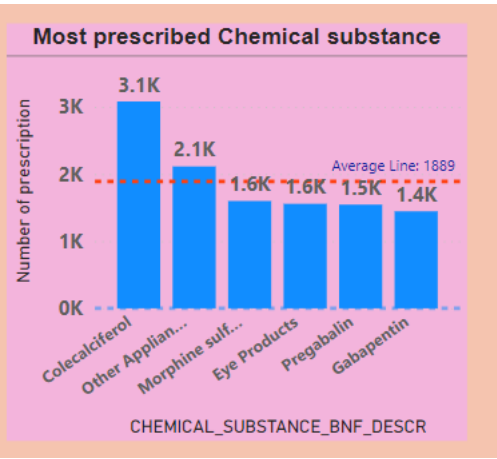
Listed below are the top **8 organs** that were affected. Top on the list is the **Central Nervous System 36.8K**, followed by Cardiovascular System having **20.63K** cases diagnosed. While, diseases and sickness affecting the Endocrine System was **12.88K**



Mostly Prescribed Chemical substance

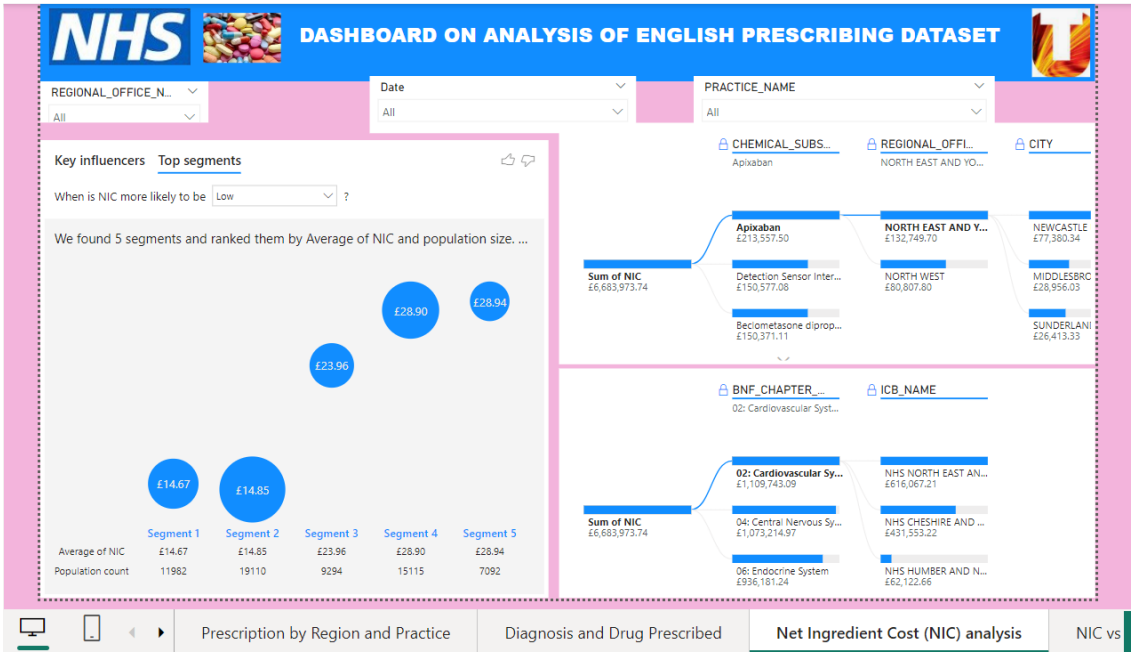
A clustered column chart was used to display the number of times each of the chemical substance was prescribed. From the visual, the top chemical substance prescribed is: Colecalciferol with a count of prescription being **3.1K**. It is a major active chemical used for the treatment of vitamin D deficiency or kidney disease. From this visual, it could be concluded that most patients who visited all the practice name within the two months require vitamin D or they have kidney related diseases.

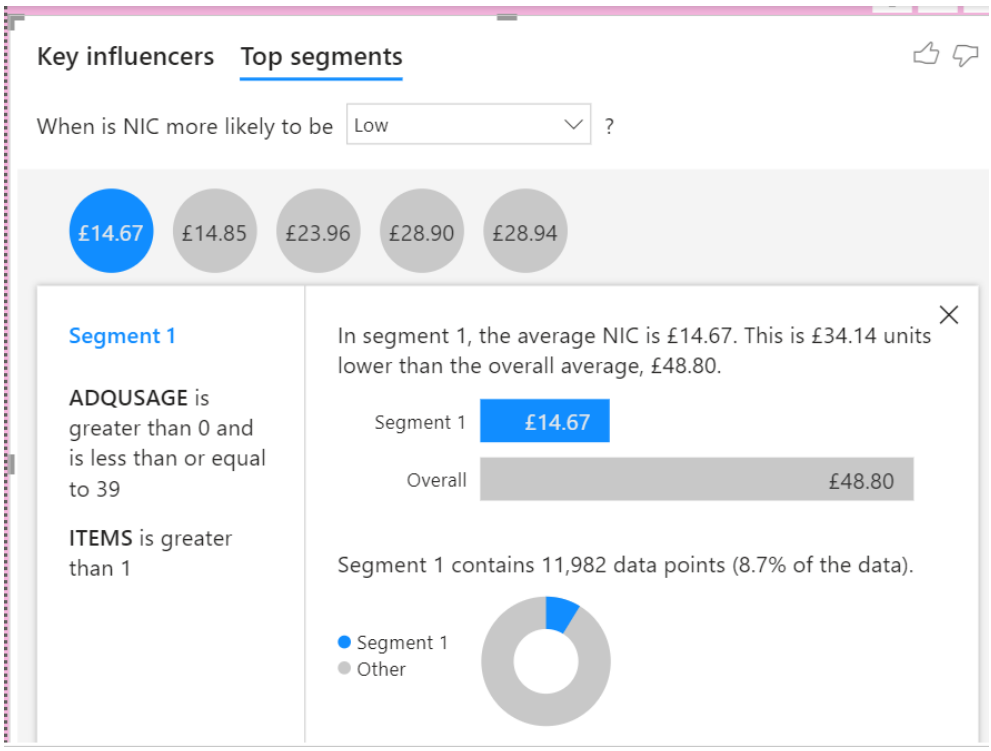
An average Line was also inserted to give insight on the average count of all the prescriptions in the period. The average count of prescription is **1,889**



3. Net Ingredient Cost (NIC) Analysis

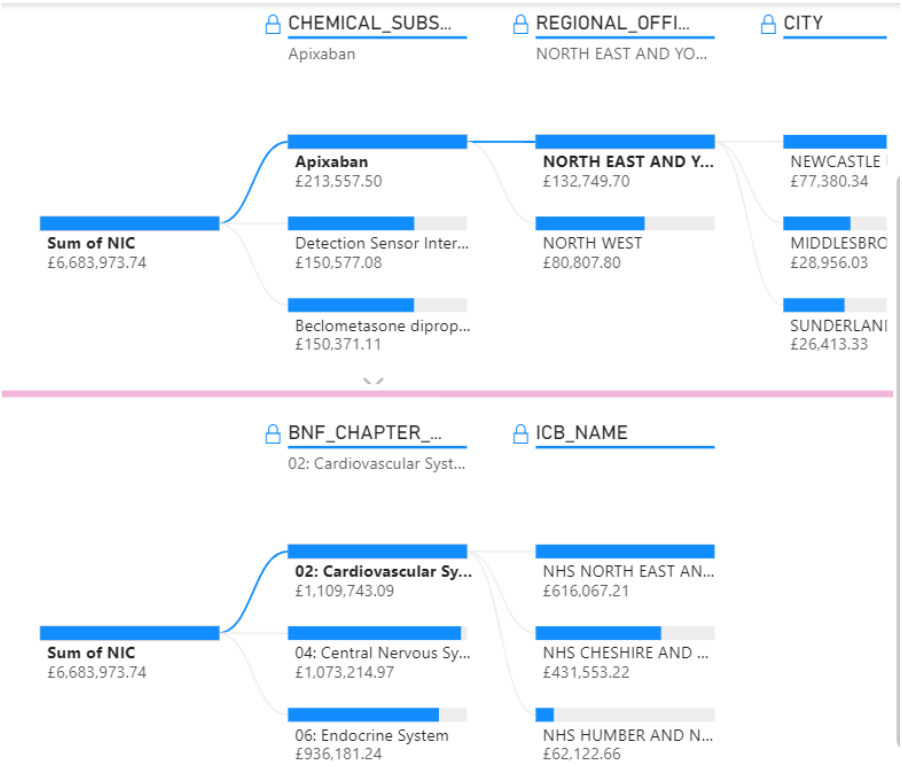
This visual use artificial intelligence (AI) feature in Microsoft Power BI to predict specified KPI of interest in the dataset.
For instance, the key influencers visual on the left, predicts when NIC is likely to be low and segment the data into different group. As seen from the second screenshot, in segment 1, if ADQUSAGE is greater than 0 and less or equal to 39 and item is greater than 1, the average NIC is £14.67 which is £34.14 below the overall average of £48.80





Analysis on Total NIC

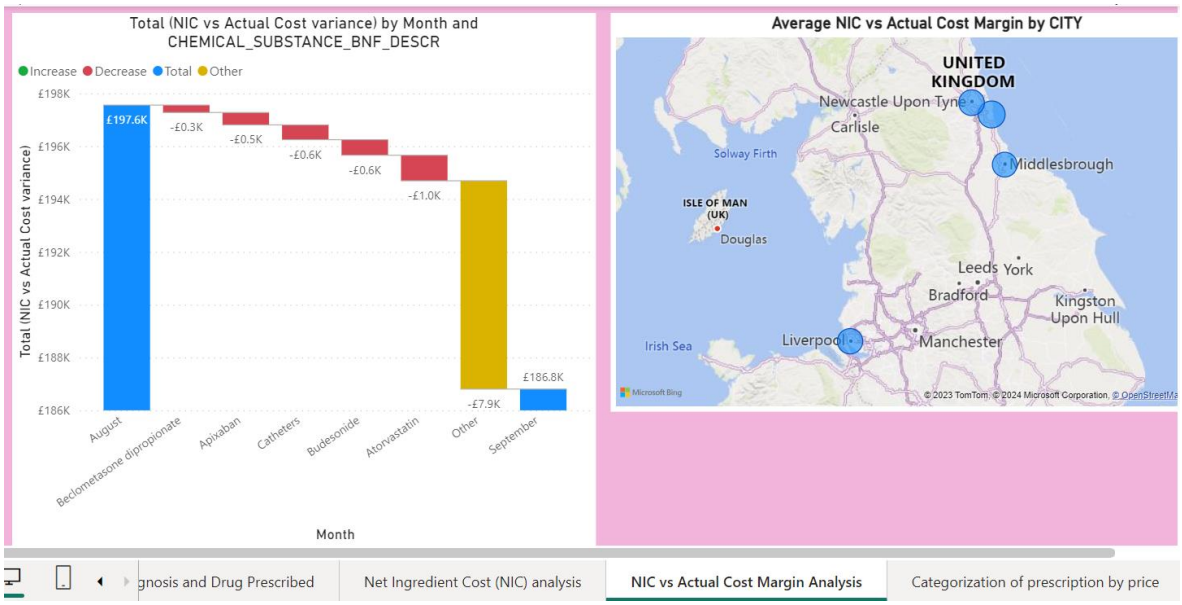
This also uses artificial intelligence to analyse the sum of the NIC. It shows that out of the total £6.68 million on prescribed drug, Highest NIC cost of £213,557.50 was spent on Apixaban. Breakdown by region and city for the Apixaban is also as shown and highest cost was incurred by the practice names in Newcastle. Also, Breakdown of the **total NIC** by **BNF_Chapter_Plus_Code** shows that highest total amount of **£1.1million** was spent on Cardiovascular system and in ICB named **NHS North East and North Cumbria Integra**



4. NIC vs Actual Cost Margin Analysis

Below is a water fall visual used to illustrate the cost margin. That is, difference between the NIC and Actual Cost by month for each of the Chemical substance. For instance, Beclometasone dipropionate, variance of two costs: **NIC and Actual cost** in August £4,914.39 and in September it is £4,632.96. This shows a drop in the gap for the two months to £281.43. This KPI can be used to determine if factors that contribute to increased cost of drug (such as logistics, dispensing costs, discounts given to patients)are being worked on month-on-month when compared to the NIC cost from the drug manufacturer or as defined in the Drug Tariff.

Similarly, we also show the average variance between NIC and Actual cost of all the prescriptions for all the cities using a **Map visual**. The visual shows that **Middlesbrough has the least variance of £2.6, while, Sunderland** has the **highest average variance £3.16.**



5. Categorization of Prescription by prices

An infographic designer visual is used to represent the count of various drugs accordingly to how cheap or expensive they are. We categorize the prices of all the prescriptions based on the following:

low price: £0 – £ 2

affordable price: £2 – £ 4

moderately priced: £4 – £6

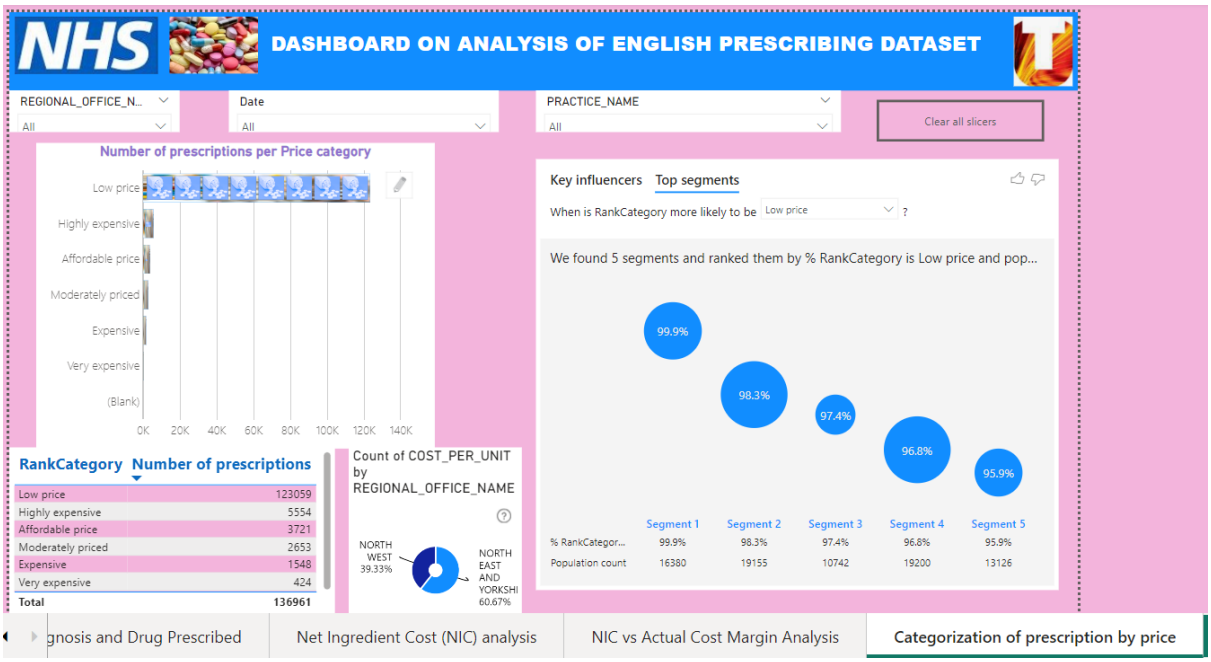
expensive: £6 – £8

very expensive: £8 – £10

Highly expensive: £10-£1,000,000

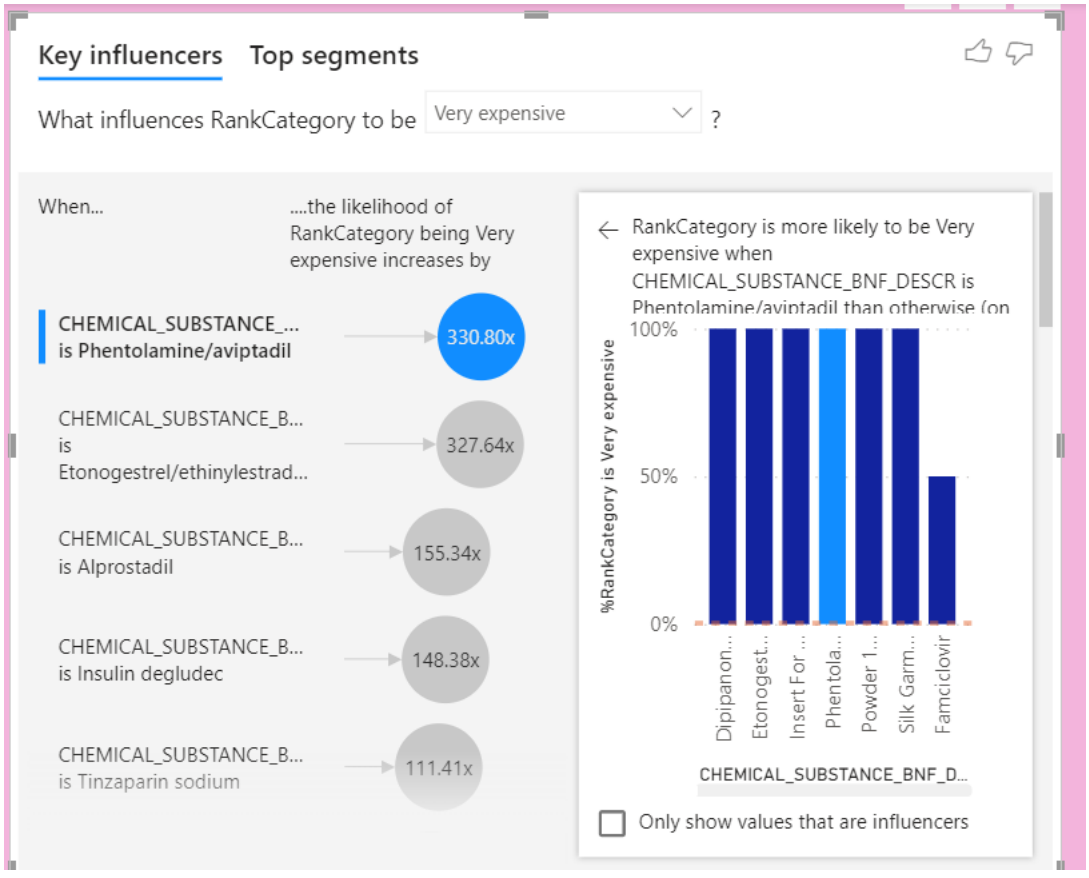
This was done by creating a benchmark table

The infographic visual shows that **123.06K** number of the drug is classed as **low price** i.e. it costs between £0 and £ 2 while**, 5.55K** number of drugs prescribed are classified as **Highly expensive** i.e. each costs between **£10 and £1,000,000**



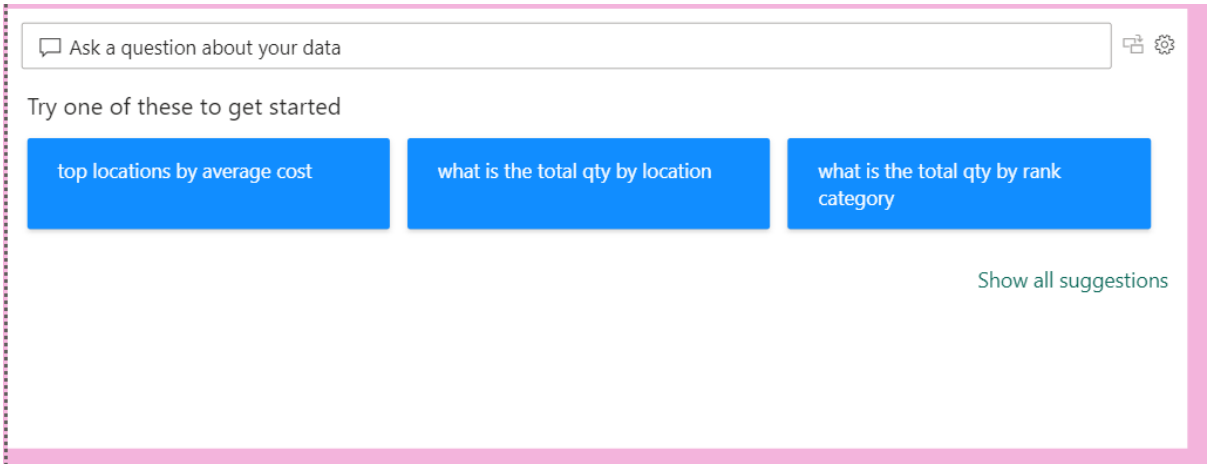
The key influencers visual below also predicts the likelihood of a chemical substance according to price categorization. For instance, when, the chemical substance is Phentolamine/aviptadil, the likelihood of the Rank category being ***very expensive increases by 330.80X***

There is also a segment tab to view the segment and the count of prescription in that bucket.



6. Questions & Answers

This Q & A visual allows user to interact with the dashboard for any insight they require about the data at a glance.



CHAPTER 6

Conclusions

Based on the insight identified above, it can be concluded that the average **Actual Cost** of drug in UK i.e. **£46** is lower by **£2.80** compared to the average **Net Ingredient Cost (NIC) £48.80** used by the drug manufacturer or value in the Drug Tariff. This suggests that amount paid by the patients for the prescribed drug has been subsidized through Government effort to make healthcare service affordable to all residents in the country.

In the same vein, from the categorization of prescription by price, it was discovered that **123,059 (89.8%)** count of prescription of drugs out of **136,961 total prescription** count were categorized as low price i.e. the price paid by most patients who visited medical practice centres is not more than **£2**.

Recommendation

NHS in the UK can focus more on preventive measures such as increased health awareness campaign rather than corrective drugs of certain diseases whose trends are increasing monthly such as ***Kidney disease*** and ***Vitamin D*** deficiency where ***Colecalciferol*** is being prescribed as the highest drug. Also, Government can invest more in Health Care Research, Data Analytics and Technology in order to promptly identify prescription pattern, monitor various health KPI across all the locations in the UK or detect deviation from t***he NHS values.***

References

(NHS), N. H. S., 2023. *The NHS Values.* [Online]
Available at: https://www.healthcareers.nhs.uk/working-health/working-nhs/nhs-constitution
[Accessed 23 December 2023].

Brett, P., 2017. *Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of Analytical Data Models, Reports, and Dashboards.* First ed. Birmingham: Packt Publishing Ltd..

Digital, N., 2023. *https://digital.nhs.uk/.* [Online]
Available at: https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing/practice-level-prescribing-in-england-a-summary/practice-level-prescribing-glossary-of-terms
[Accessed 20 12 2023].

Morton-Jones, T. and Pringle, M., 1993. Explaining variations in prescribing costs across England. BMJ. *British Medical Journal,,* Volume 306(6894), p. 1731.

Stedman, C., 2005. *TechTarget.* [Online]
Available at: https://www.techtarget.com/searchdatamanagement/definition/data-modeling
[Accessed 05 January 2024].

Wilson, M., 2021. *NHSBSA Statistics and Data Science.* [Online]
Available at: https://www.nhsbsa.nhs.uk/statistical-collections/prescription-cost-analysis-england/prescription-cost-analysis-england-2020-21
[Accessed 01 January 2024].

Woodford, H.J., Barrett, S. and Pattman, S., 2018. Vitamin D: too much testing and treating?. *Clinical Medicine,* Volume 18(3), p. 196.

Title: PowerBI - Self-Assessment and Skills Development Reflection Form

Form:

Learning Outcome 1: Reflect on Your Performance and Skills Development

Please take a moment to reflect on your experiences and skills development during the PowerBI project. Your feedback is valuable in helping us understand your progress and areas for improvement.

Name: IDOWU **OBISANYA**

Module: BIG DATA AND BUSINESS INTELLIGENCE

Project Title: ANALYSIS OF ENGLISH PRESCRIBING DATASET (EPD)

PowerBI Tutor: AMITHA RAJ (LAB TUTOR), NAWAZ, MANSHA (COURSE INSTRUCTOR)

Self-Assessment Grading

Notes for Self-Assessment

- **Score Range**: Assign yourself a grade based on your overall score.
- **Lab Proficiency**: Consider your comfort and skill level in the labs mentioned for each grade. Reflect on your ability to understand and apply the concepts and techniques covered.
- **Team Role and Skills**: Evaluate how well you can perform in a team setting based on your PowerBI skills. Consider your ability to contribute, support, or lead in a PowerBI team environment.

Self Assessment Grade (and Mark is optional (see table below and note the supporting evidence is the symetery to your PowerBI Lab work.	Grade [F, P, M, D	Mark [0-100]
ICA1: PowerBI .Pbix file	D	99
ICA2: PowerBI Technical Report – Each KPI provide DAX overview	D	99

ICA3: PowerBI Stakeholder Report – Each BI Dashboard presentation overview (3-6 Dashboards)	D	99

PowerBI Skills Self-Assessment Table

Grade	Score Range	Description	Lab Proficiency	Team Role and Skills
PASS	50-59	Can work as a Junior Member of an Industry PowerBI Team. Possesses fundamental PowerBI skills necessary for basic tasks and analysis.	Proficient in Labs 1-6 (Basic Data Handling, Simple Visualizations, Foundational Reporting). Competent in Labs 7-8 (Intermediate Data Modeling, DAX, and Visualizations).	Capable of supporting senior team members, handling basic PowerBI tasks, and contributing to simple projects. Requires guidance and support for more complex tasks.
MERIT	60-69	Can work as an effective, productive member of a PowerBI Team with little or no support. Demonstrates a broad range of PowerBI skills.	Proficient in Labs 1-8 (covering Basic to Intermediate PowerBI Skills). Attempted at a competent level some of Labs 10-12 (Advanced Data Analytics and Visualizations).	Able to independently handle a variety of PowerBI tasks, contribute significantly to team projects, and provide support to less experienced team members. Requires minimal guidance.
DISTINCTIVE	70-100	Can work independently and as a PowerBI Team Leader. Exhibits a high level of proficiency and competency in PowerBI.	Proficient and competent in Labs 1-12 (covering Basic, Intermediate, and Advanced PowerBI Skills).	Capable of leading PowerBI projects, making strategic decisions, mentoring team members, and handling complex PowerBI challenges. Exhibits leadership and advanced analytical skills.

This table is designed to help students evaluate their PowerBI skills in relation to industry standards and team roles. It encourages self-awareness of their abilities and identifies areas for further development and growth in their PowerBI learning journey.

Self-Assessment Categories

Instructions for Self-Assessment Categories

- **Complete each category**: Reflect on your performance and understanding in each category based on your experience in the corresponding labs.
- **Choose a grade**: Select P [pass 50-59], M [merit 60-69] or D [distinctive [70-100] based on your overall performance in each category.
- **Self-assessment**: Indicate whether you 'Need Improving', 'Met Expectations', or 'Exceeded Expectations' for each category.
- **Rate yourself**: Use the 1-5 scale to rate your proficiency and understanding.

Assessment Categories

Category	Lab Content	Grade (F/P/M/D)	Needs Improving	Met Expectations	Exceeded Expectations
Data Handling and Basic DAX (Labs 1-6)	Basic data import, cleaning, simple DAX operations	D			Exceeded Expectations
Simple Visualizations (Labs 1-6)	Creation of basic charts and tables	D			Exceeded Expectations
Foundational Reporting (Labs 1-6)	Basic report assembly, use of titles and descriptions	D			Exceeded Expectations
Advanced Data Modeling (Labs 7-9)	Complex data models, star schema design	D			Exceeded Expectations
Intermediate DAX and	Intermediate DAX functions,	D			Exceeded Expectations

Category	Lab Content	Grade (F/P/M/D)	Needs Improving	Met Expectations	Exceeded Expectations
Visualizations (Labs 7-9)	interactive reports				
Expert Data Analytics and Visualizations (Labs 10-12)	Advanced DAX, AI features, custom visuals	D			Exceeded Expectations
OVERALL POWERBI SKILL COMPETENCY	Project leadership, strategic decision-making, mentoring	D			Exceeded Expectations