PROJECT REPORT – National Health Service

We were given a task by NHS to find out whether there is enough capacity and subsequent staffing to use the capacity and the resources available, and to see if those resources were used to the full extent or there were gaps or shortfalls. This was done in the view of escalating costs of missed appointments at GP clinics and the Government ruling to bring the costs down.

I was given the task of preparing the data files for analysis by making sure they were loaded correctly and that the necessary codes were applied to make it compact enough to be used for creation of data tables and visualisations that will help NHS make decisions or understand where the problems are and direct the thought process on finding solutions.

I made sure the data files needed for analysis were imported in to **Jupyter Notebook** through python and its associated libraries such as **Pandas, Numpy** and for visualisations libraries such as **Seaborn, MatPlotLib, Requests** and were prepared for analysis by sense-checking using functions like **.dtypes(), .info(), .shape() and .columns**. These functions give information such as names of columns, the data types of each column, the number of rows and columns, and the number of rows with values. Using the code line **DataFrame.head()** gave a view of the data file, thus ensuring it has been imported correctly. This helps prepare the data for calculations and manipulations.

This is the result of the **dataframe.info()** code:-

```
class 'pandas.core.frame.DataFrame'>
RangeIndex: 817394 entries, 0 to 817393
Data columns (total 8 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   appointment_date      817394 non-null  object
 1   icb_ons_code          817394 non-null  object
 2   sub_icb_location_name 817394 non-null  object
 3   service_setting       817394 non-null  object
 4   context_type          817394 non-null  object
 5   national_category     817394 non-null  object
 6   count_of_appointments 817394 non-null  int64
 7   appointment_month     817394 non-null  object
dtypes: int64(1), object(7)
```

I then used functions such as **.value_counts()** which returns the count of occurrences of a particular data item within the column, **.sum()** to find the Totals, to start answering the various questions NHS have asked such as:

Q1 Should the NHS start looking at increasing staff levels?

Q2 which *service settings* reported the most appointments for a specific period?

Q3 Are there any trends in the time between booking and the actual date of appointment ?

Q4 Are there any significant changes in attendances of patients for appointments ?

I made use of functions like **.loc()** to extract the columns needed for analysis, **to_datetime()** function to convert the column into date data type for time-based calculations, **.groupby()** function to extract the necessary columns from a larger data set, sort them and extract specific date from those columns.

Examples are:

1) nc['appointment_month'].value_counts(ascending=True)

2) ac2 = nc_subset.sort_values(['count_of_appointments'], ascending=False).groupby([nc_subset['appointment_date'].dt.year, nc_subset['appointment_date'].dt.month]).head()

A quick analysis of the data files shows the sharp contrast between East and South-East of England and the rest of the country. The 4 NHS trusts with the highest number of appointments recorded are in the East and South-East of the country whilst London could only manage one trust. But a substantial chunk of the records of appointments shows the patient booked but did not attend the appointment and a smaller number were recorded as **Unknown** as it is not clear whether the patient attended or not. These errors can be attributed to diverse ways of recording appointments and statuses, which when collated can give errors like Unknown, making it difficult for analysis and thus arrive at any conclusions. I wanted to find out how many unknown in question so used the **value_counts()** function and it threw some surprising results.(see **Appendix A**).

The visualisation techniques offered by Python and used on the NHS database threw up a few observations, which may not be many but are still important as they shed light on the utilisation of services of NHS and show where the expenditure of funds will be focused and where cutbacks can be made.

I proceeded to setup the code for visualisation that can give a bigger picture of the data and perhaps some insights and recommendations. I set the display size and then using **groupby()** function selected the columns that were relevant to the analysis. I then used codes from Seaborn starting with "**sns**"(which is an alias for Seaborn) to create lineplot and bar plots.(see **Appendix B**)

As can be seen from the visualisations, the number of people attending GP surgeries are the highest and that is the main focus of primary healthcare in UK, and it has been noted that GP surgeries can offer more than just consultation and can act as a community healthcare point. I created many visualisations for each question, splitting them further for analysis and make it easier on the eye for the stakeholders who can then understand and take some inferences from the plots. Some of the plots are given in **Appendix C.**

As social media becomes the communication norm and more and more information about products and services can be had, I decided to scrap Twitter for any hashtags that are trending about NHS and in general about any particular service. The data scrap revealed that **#healthcare** is the most prominent and that shows many a views

were put forward under this hashtag. A visual is in the below **Appendix D.** I then checked for other hashtags by removing the most popular ones and that revealed other hashtags that are trending but not as much as #healthcare. I used the “remove outlier method” which eliminated some hashtags and displayed the rest.

Summary: We were given semi-cleaned data files and I made them ready for analysis by using various functions and then got together columns that were needed for plotting graphs and to scrutinise the data, especially the data from Twitter which was acquired using data scraping techniques This gave an in-depth look at the data and was used to draw inferences and understand the trend and gather insights.

All visualisations and data show majority of appointments were made on the same day and were face-to-face, illustrating that patients are going directly to GP or even A&E for treatment and that means pressure on staff to cope with the influx of patients thus increasing waiting times. The NHS was used the most during winter and summer times especially as the pollen season sets in. The visualisations highlight an interesting point where more GP surgeries are needed, and they be enhanced to offer services such as blood tests which are still done at many hospitals. But as is seen Hospitals still offer the most type of services and adequate levels of staffing is needed at all times. The dataset also throws light on the fact that a few NHS trusts are recording large numbers of patients whilst others show less numbers which can either be interpreted as some trusts are overwhelmed whilst some are under utilised OR the data is not being recorded properly hence does not give a true picture of utilisation of resources and capacity. No cuts should be done to staffing levels till a standard system of reporting is put in place which will allow the understanding of current capacity and utilisation of resources and services

Appendices:

Appendix A

```
ar['appointment_status'].value_counts()
```

Care Related Encounter	700481
Inconsistent Mapping	*89494*
Unmapped	*27419*

NHS Norfolk and Waveney ICB - 26A',
NHS Kent and Medway ICB - 91Q',
NHS North West London ICB - W2U3Z',
NHS Bedfordshire Luton and Milton Keynes ICB - M1J4Y',
NHS Greater Manchester ICB - 14L

```
nc['national_category'].value_counts()
```

inconsistent Mapping	89494
General Consultation Routine	89329

General Consultation Acute	84874
Planned Clinics	76429
Clinical Triage	74539
Planned Clinical Procedure	59631
Structured Medication Review	44467
Service provided by organisation external to the practice	43095
Home Visit	41850
Unplanned Clinical Activity	40415
Patient contact during Care Home Round	28795
Unmapped	27419
Care Home Visit	26644
Social Prescribing Service	26492
Care Home Needs Assessment & Personalised Care and Support Planning	23505
Non-contractual chargeable work	20896
Walk-in	14179
Group Consultation and Group Education	5341

Appendix B

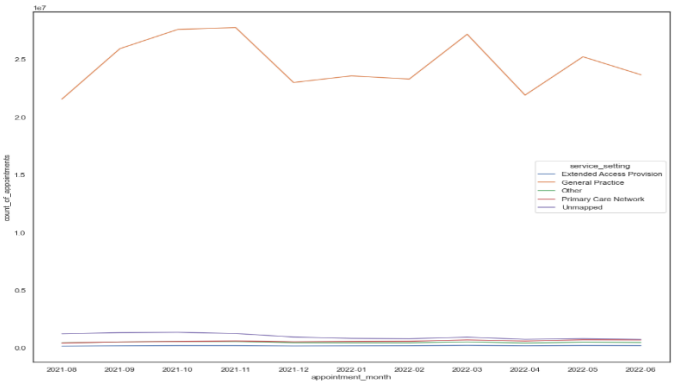
To set the size of the plot:--

```
sns.set(rc={'figure.figsize':(15, 12)})
sns.set_style('white')
```

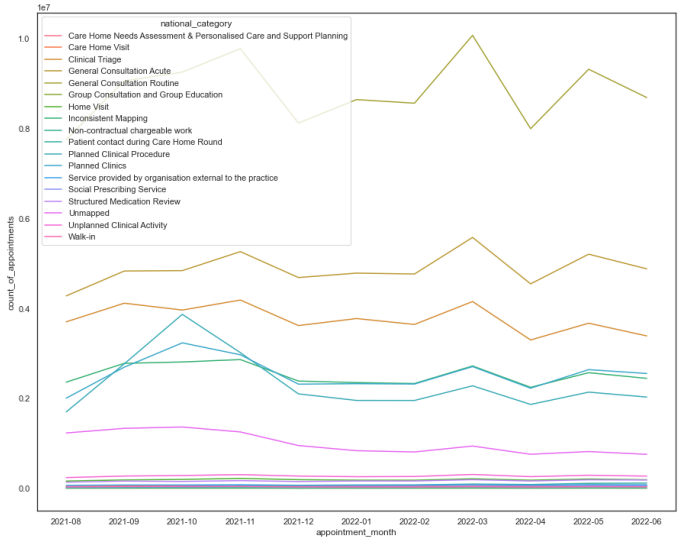
The following code line is an example of a code to create a lineplot wher e “x” and “y” are columns names, nc_ss is the name of the dataframe:--

```
sns.lineplot(x='service_setting', y='count_of_appointments', hue='appoint ment_month'=='2021-08', data=nc_ss, ci=None)
```

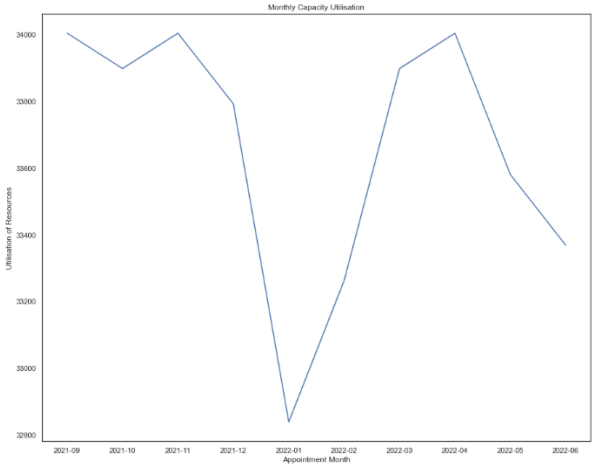
Appendix C



The above visualisation shows the different type of service settings

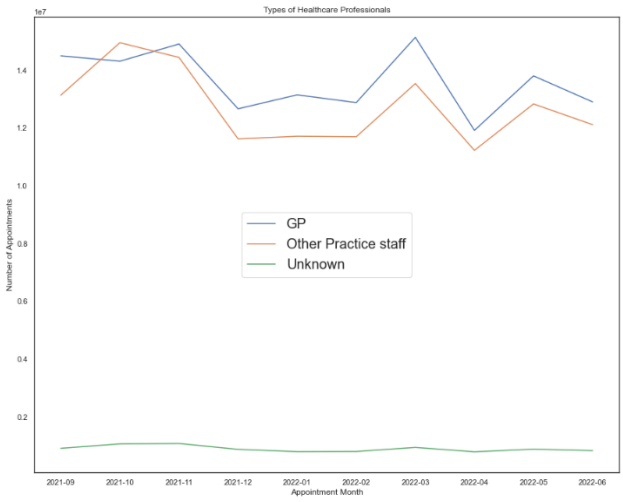


The above visualisation shows different type of services



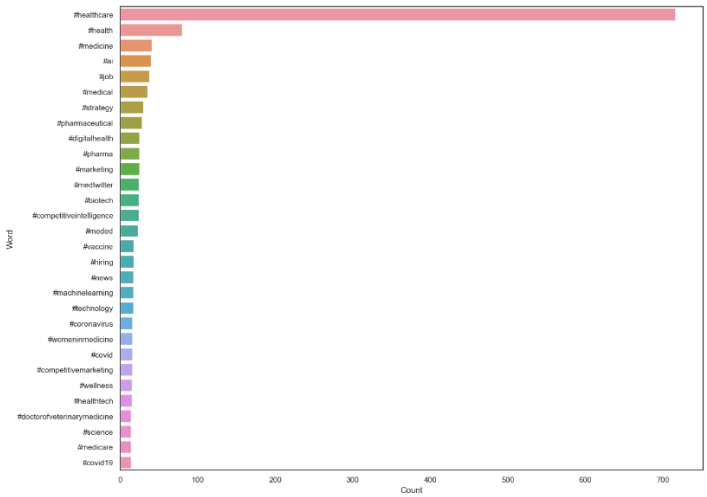
The above graph shows that NHS was used more in the winter and summer months.

This visual shows where most patients go to for treatment or consultation.

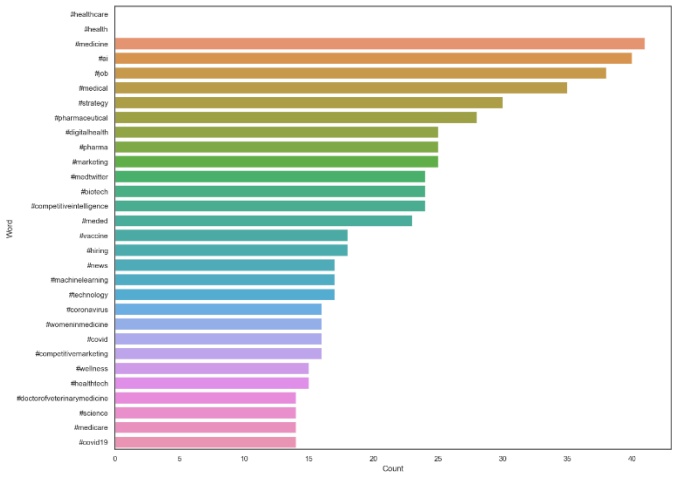


The above visual gives an idea of the type of healthcare professional most consulted by patients

Appendix D



The Above visual is with #healthcare



The Above visual is without #healtcare to give a better understanding of other trending hashtags.