Assignment: Diagnostic Analysis using Python

As a team of data analysts, the NHS has contracted us to better understand the reasons for people missing appointments as this is a significant cost that is potentially avoidable. To make a data-informed approach to solving this issue, two main questions have to be answered through analysing the NHS’ data:

1. Has there been adequate staff and capacity in the networks?
2. What was the actual utilisation of resources?

Firstly, it is important to understand the scale of the business problem by determining the number of locations in the data set as well as understanding how the differentiation of the appointment types through service settings, context types, national categories as well as whether the appointments were attended. The analyses were carried out using Python on Jupyter Notebook. The necessary libraries were imported and the data sets were loaded, ensuring that they were validated and checked for any missing values. The number of locations in this data set was determined to be 106 locations, with the 5 locations having the highest number of records being as follows:

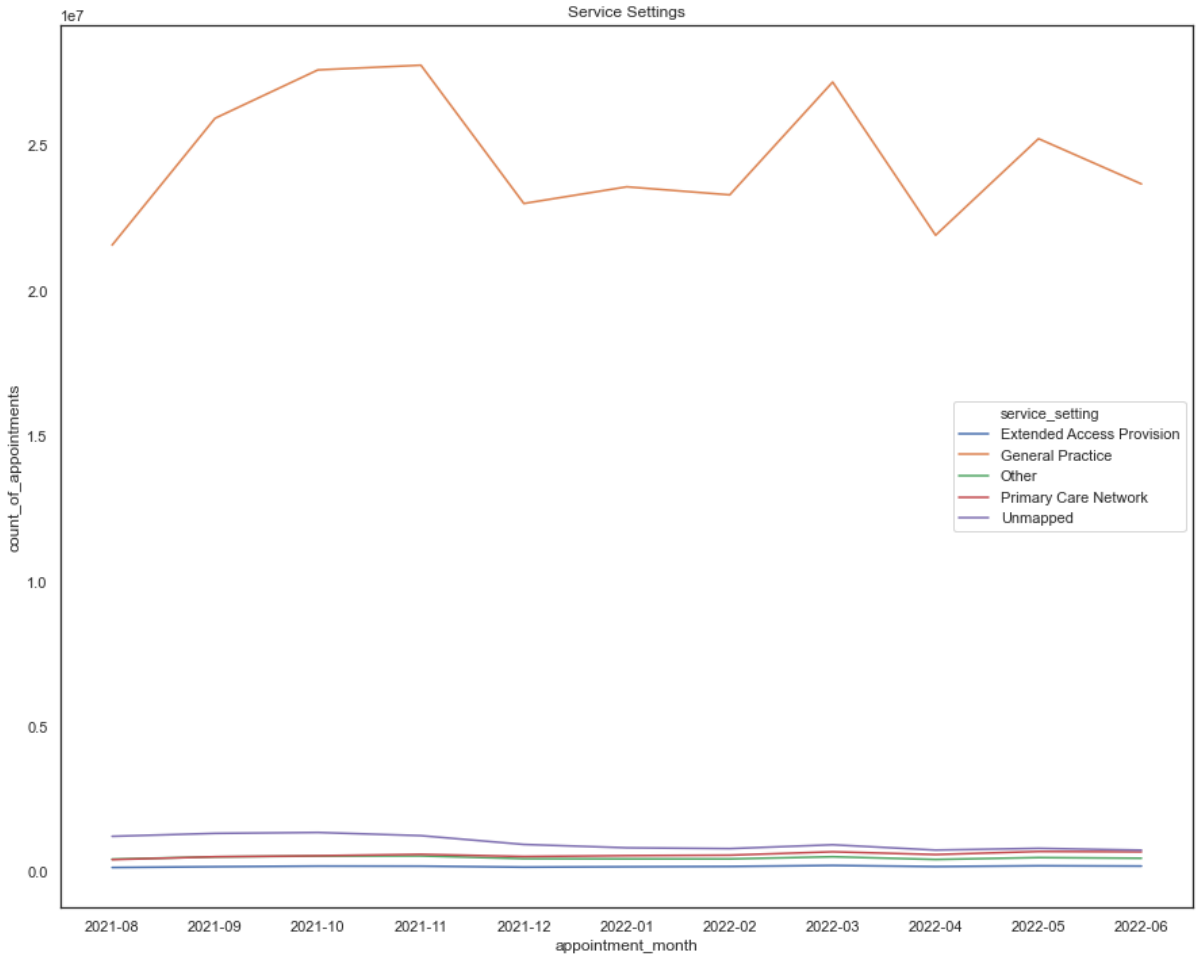
Location	**Number of records**
NHS North West London ICB - W2U3Z	13007
NHS Kent and Medway ICB - 91Q	12637
NHS Devon ICB - 15N	12526
NHS Hampshire and Isle Of Wight ICB - D9Y0V	12171
NHS North East London ICB - A3A8R	11837

Moving forward, it would be interesting to see over what period of time it took to accumulate this many records for these locations. Knowing how many types of differentiating factors there are for every appointment record can help us to break down and carry out analysis in segments to better identify trends within each factor which could contribute overall to the answering the main business question.

Moving on from the previous analysis, it is now appropriate to explore the time scale of the given the data sets. Firstly, the first 5 rows of the any columns pertaining to dates were viewed across all the data sets. I chose to normalise the date formats by changing all of the data sets’ ‘appointment_date’ columns to a date time format. The min and max functions were then used to determine the minimum and maximum dates for the data sets. It is observed that the two aforementioned data sets have different minimum dates but similar maximum dates (2022-06-01).

Now that we have a time scale, we can perform a more detailed analysis by looking at the most popular service setting for the location with the highest number of records (North West) within a 6 months period. This was done by firstly extracting just the relevant columns from the main data set to create a new data frame. Creating new data frames allows the user to visualise the output of each specific question in neater and more narrowed down tables.

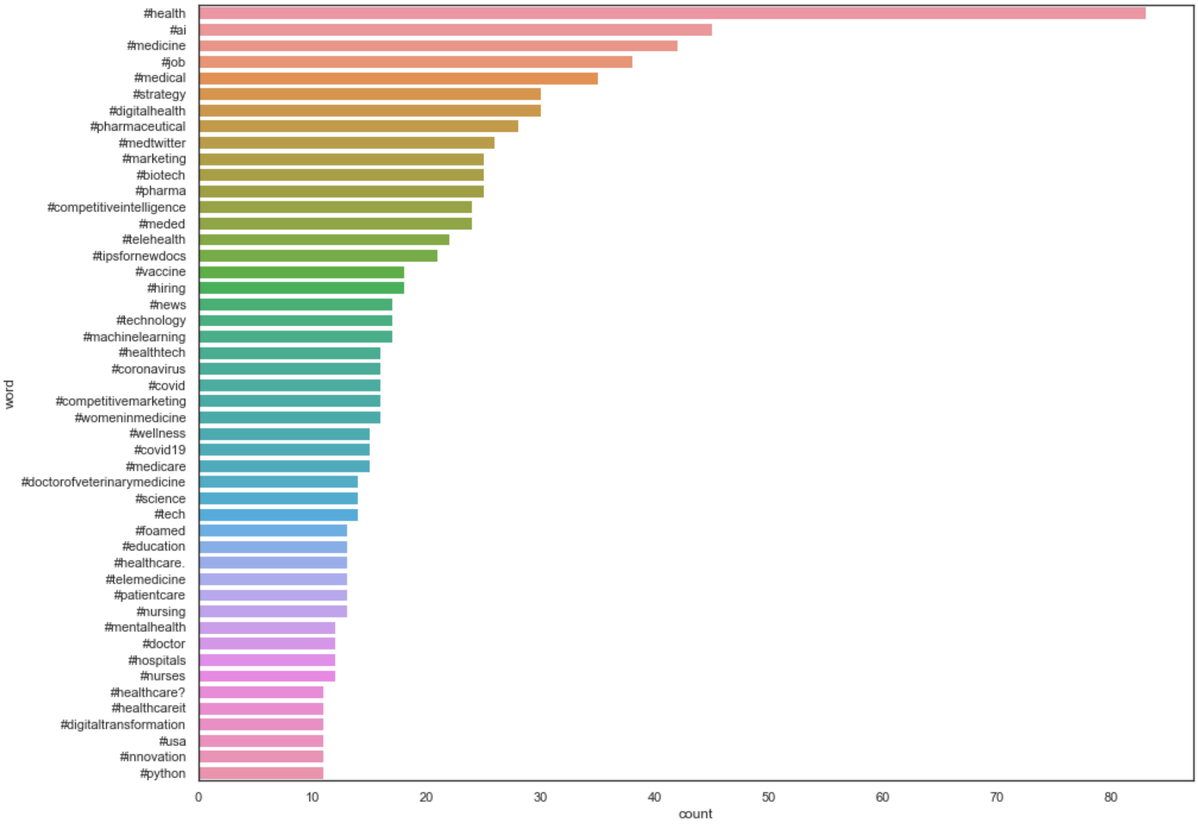
The Seaborn and Matplotlib libraries were imported to visualise the data. The groupby() function was used throughout all of the plots in order to group the data by month. Each variable is colour-coded on the line plot and the labels are shown on the legend to help the user identify what each line represents.



For service settings, it can be deduced that 'General Practice' is the most popular service setting by a very large margin. Additionally, the context type with the highest number of appointments is 'Care Related Encountered' by a similarly very large margin. The national categories, however, has the highest number of variables and a more varied spread with 'General Consultation Routine' having the highest number of appointments.

For a more granular look at the number of appointments, we can look at the breakdown of the count of appointments for each service setting by the 4 seasons (specified as 4 different months). The overall number of appointments stay consistent throughout the seasons with autumn having the highest amount by a small margin and summer having the least number of appointments. The sharp drops in the line plots are represented by the clinics not operating on weekends.

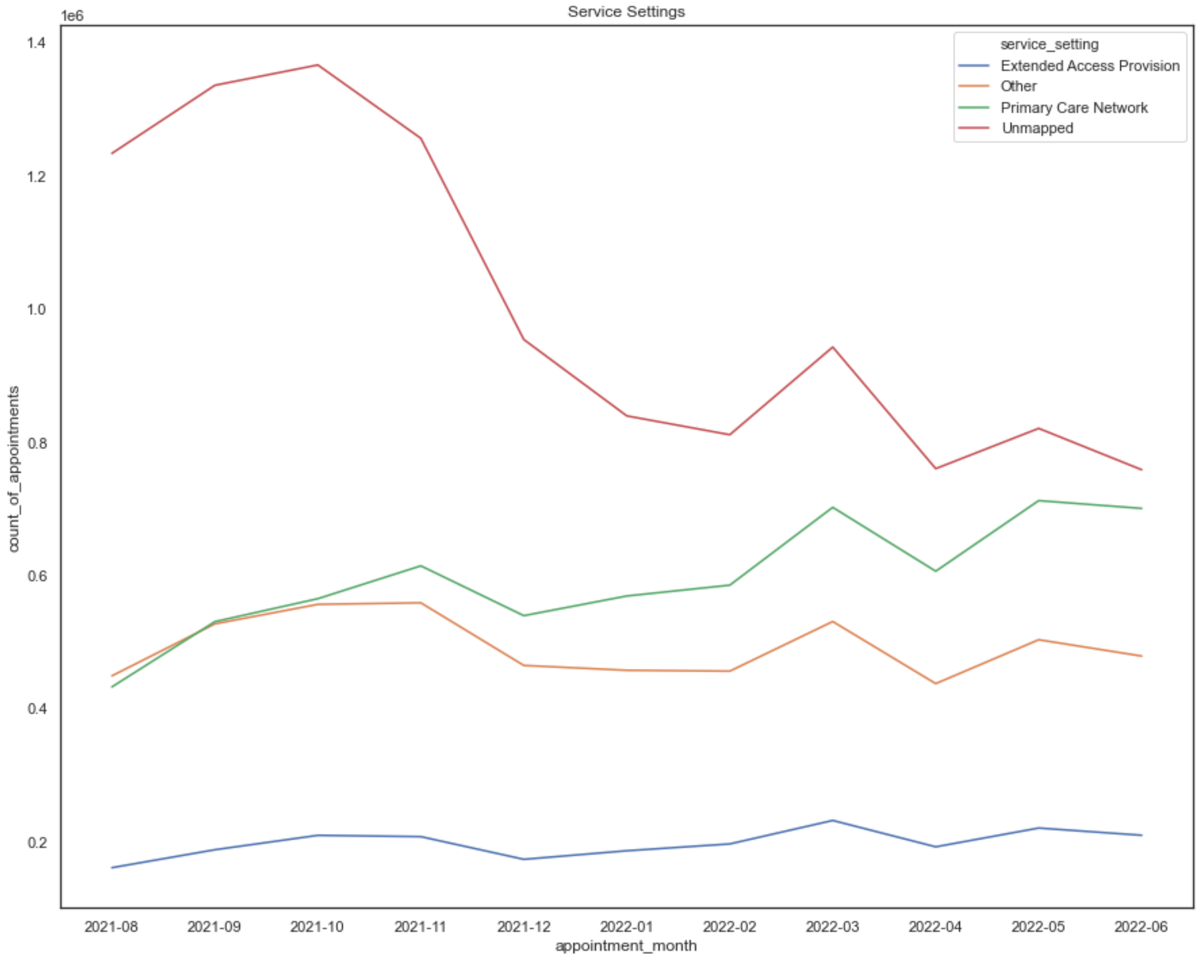
Analysing tweets from Twitter can help determine whether there has been a high volume of hashtags related to healthcare in the UK. Any hashtags with a count value of less than 10 was filtered out in the new data frame created from the data set as we are interested at looking at the popular hashtags. As there are a large number of hashtags, I chose to display a horizontal bar chart to make visualisation easier for the user. The most popular hashtag was removed as it was overrepresented by a large mar, which makes the new bar plot a lot more readable for the user which now shows that the most popular hashtag is '#health'. Moreover, it is noteworthy that the Twitter community web page has noted that the Twitter API sometimes pulls up duplicate tweets which could skew the data.



To answer whether the NHS staff should increase staff levels, we need to look at the utilisation rate of NHS' maximum capacity, which is 1,200,000 appointments a day. From the line plot, it can be deduced that utilisation levels are no where near the maximum with the highest utilisation rate being only 84.5% which shows us that the NHS staff do not need to increase their staff levels. However, this does not solve the issue where there is a significant portion of people missing their appointments.

Upon plotting additional line plots to investigate other factors, a notable observation is that there seems to be a large drop in the utilisation of services in the early months of 2020. This can likely be attributed to the lockdown that was brought upon by the COVID-19 pandemic and should be accounted for in the analysis or excluded from the line plot as that period would be deemed abnormal. Furthermore, it is noteworthy that same day appointments are the most popular, with more than 28 days between book and appointment being the least popular. It can be seen that as the time between book and appointment increases, the number of appointments decreases. For further analysis, it would be worth looking at the rate of missed appointments for the same day appointments.

Previously, we saw that the 'General Practice' service setting is extremely overrepresented. Removing this setting displays a clearer comparison between the other settings, and it is shown that the count of appointments for the 'Unmapped' service is decreasing over the time range whereas the 'Primary Care Network' is steadily increasing. If we were to extrapolate this line plot, it is likely the 'Primary Care Network' setting will eventually have a higher number of appointments than 'Unmapped'. The NHS should take this into account when planning their allocation of resources for the future.



To conclude, the NHS does not require more staff as utilisation is well within their capacity levels. Further analysis needs to be carried out to better determine the reasons why people are missing their appointments.