Background/context of the business scenario

The NHS has access to an exhaustive database that ranges through several key performance indicators that can be used to evaluate trends. The purpose of utizing this data is to ensure to have measurable benchmarks and actionable recommendations, as often these trends can ensure to increase and improve productivity, effectiveness, resource allocation, and more. The NHS datasets include information regarding the appointments, the different types of national categories, regional appointments, and a twitter sample.Each of these datasets can be utilized to explore different insights into improving the overall performance. Additionally, the NHS has implemented several questions that they are looking to answer with data-driven decision making, which will be addressed and discussed in this report.

Analytical approach

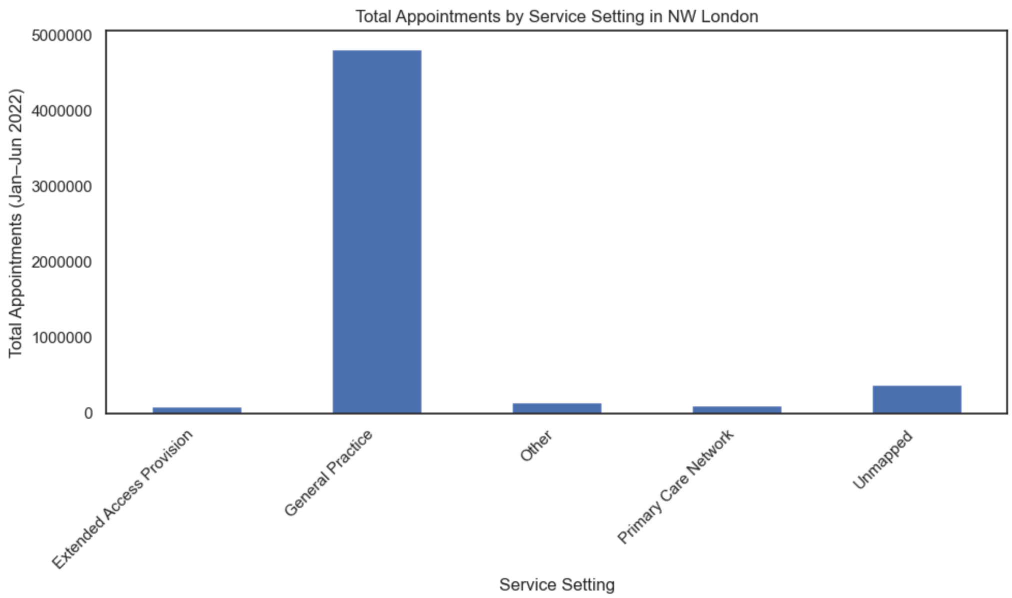
The first step in any data-exploration begins before importing the data, with actually understanding the metadata. Several columns within the dataset use specific terminology related to the NHS, which needs to be studied to further understand what information the column is provided. One important step that I always take when conducting data-exploration is to skim over the excel or csv files and get a sense of several characteristics of the data. For example, I try to grasp the size of the data, number or rows, the way certain variables are being inputted, and an overall intuitive sense of the dataset. This allows for further intuitive stops when doing analytics, as it is more likely to be able to evaluate that something does not seem right when the result massively mismatches the amount of rows on the original data file. For the NHS, I firstly noticed the format of the dates being different, the durations being estimated, and the percentage of missing data that is accounted for, that is data that is unknown but does not appear as a null value, but is defined as “unknown” in some datasets.

To begin the analysis, I imported the essential libraries and set global plotting styles for consistency. Then, I loaded the three primary data sources: actual_duration.csv, appointments_regional.csv, and national_categories.xlsx into separate DataFrames named ad, ar, and nc respectively. Utilizing the functions of df.info(), df.head(), and df.describe(), I was able to get a grasp of the datasets, and compared it to the previous skim of the data to ensure that the datasets were correctly loaded.

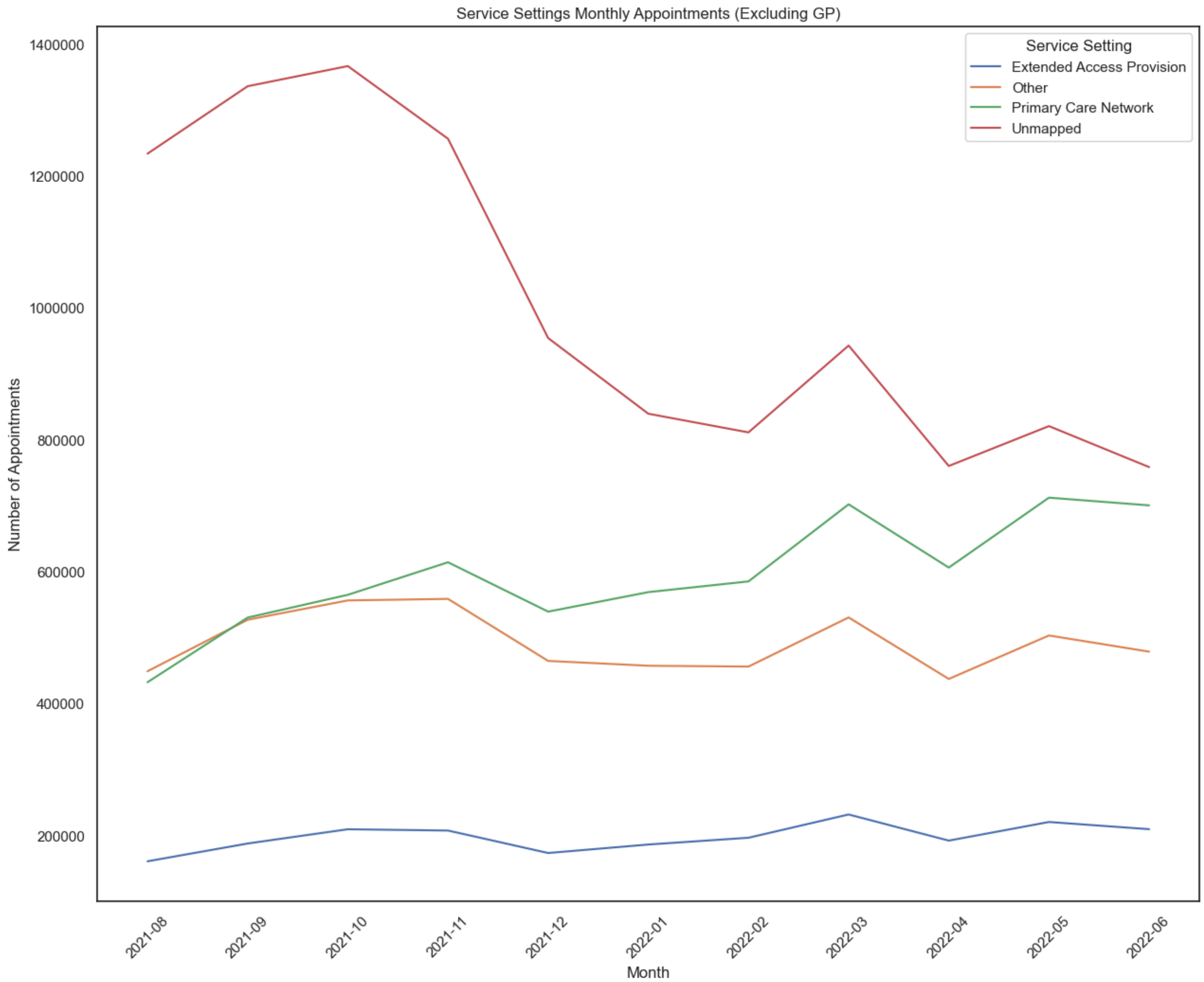
Starting with the analysis, I decided to do an exploration of the different service settings, context types, national categories, and appointment statuses, so that I can become more familiar with the dataset. Certain patterns that I considered for the analytics was certain variables having a large proportion of the data belonging to a specific option. I kept this in mind previous to the visualization stage, as I understood how in most cases, when a variable has most of the data, it can make the other variables differences visually seem as less than they actually are.

Visualisation and insights

The problem previously mentioned in the analytics section became a key consideration in the visualization stage. The example mentioned can be applied to service settings, where General Practice includes most of the data.



As it can be seen, this makes for a graph that makes the comparison of the other service settings to be unseen, and so segregating this variable, which occurred several times throughout the dataset, was a priority. For example, the following graph evaluates the number of appointments by service setting, excluding the general practice. As it can be seen, the comparison is much clearer, so presenting two distinct graphs ultimately allows for more clarity.

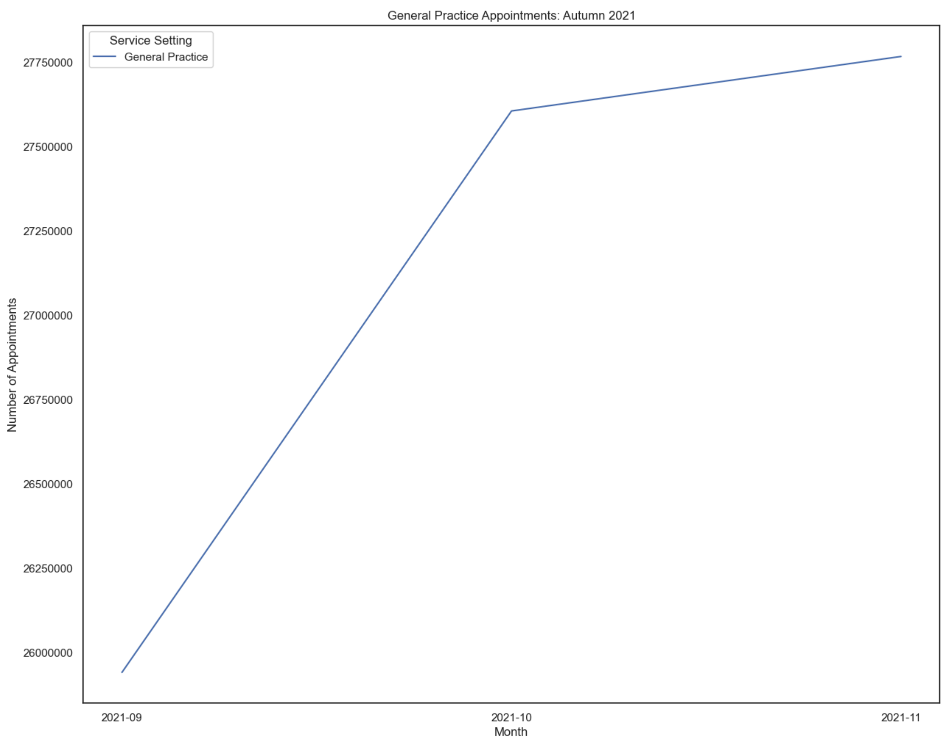


For the usage of lineplots, one important concept that I had in mind is to have clear legends with colors. By setting the legends to be in the corner, it creates a visible distinction for the user, where the usage of lineplots allow for a glance of the movement in the trend of appointment popularity throughout the months. Additionally, the x and y labels where rotatted and edited depending on the space of the data.

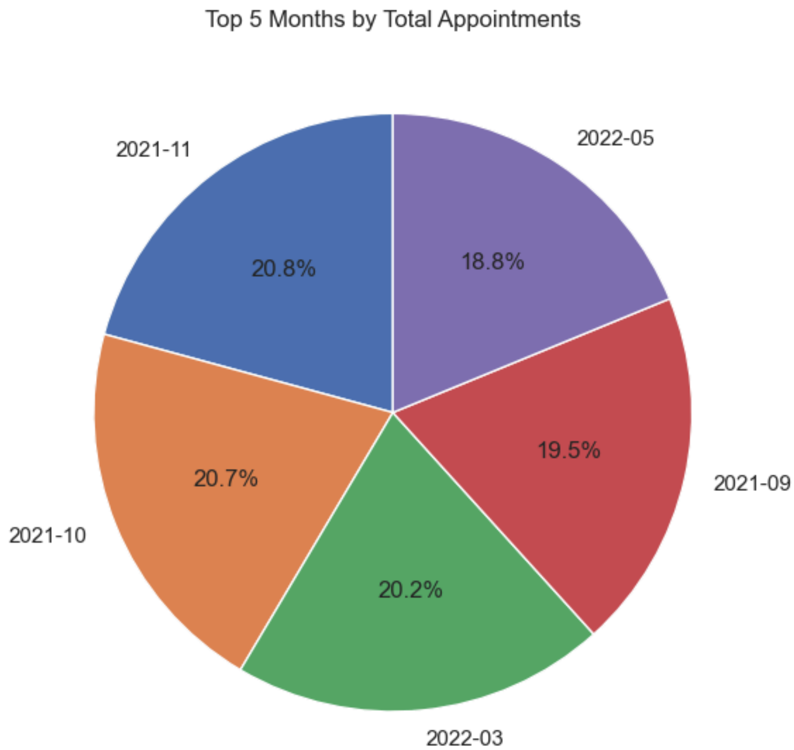
One problem that I saw consistent in the NHS data was the missing data. While not a significant percentage of the overall sample, omitting unknown data can create a bias if the data is not random, and follows a pattern. As it is unclear how precise the recollection of data is conducted by the NHS, it is better to call attention to the missing data, which can be visualized through pie charts, where the comparisons are clearer and show a snapshot of the overall proportion of appointments in general rather than through months. The reason for this choice is because the actual amount of missing data is not very clear in a linegraph. This is replicated again for other cases in the report.

Patterns and predictions

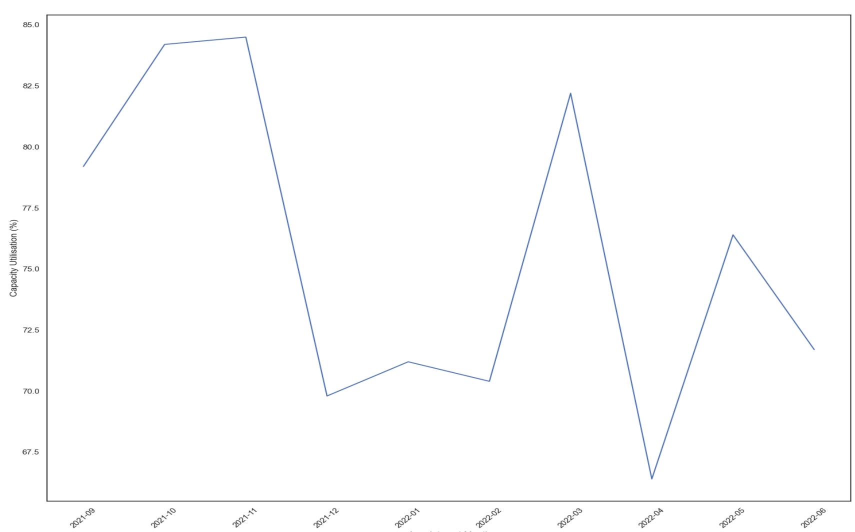
Through data analysis, we are able to visualize information based on the NHS data to understand the potential implementations that could be done.More specifically, the recommendations are based on appointment management. Firstly, through exploring the appointment volumes through seasons, we can see how majority of the general practice appointments increase during Autumn (September to November).



In fact, the season of Autumn has the three months with the largest amount of total appointments, with only March performing similar to these months, and May still having a large volume of appointments but not as much as the previously mentioned months. For this reason, the NHS should consider anticipating the volume increase for the season of Autumn.



Additionally, the utilisation percentage is calculated through a formula provided by the NHS, and during the months in Autumn, the capacity utilisation seems to be at around 85%. It is important that the staff in the NHS is not overworked, to ensure an efficient management of the patients concerns because of the high-stakes being faced.



Finally, as mentioned, the NHS needs to handle missing data more effectively, as the inability as deeming the missing data as missing randomly complicated the process of analytics.

Proportion of Appointments by Healthcare Professional Type