Diagnostic Analysis Using Python

Course 2 Final Project Report

OCTOBER 22, 2022

NAME: ABHISHEK PILLAI

Contents

INTRODUCTION

The National Health Services (NHS) is a nationally funded medical management organisation, situated in England. When patients fail to attend appointments, the NHS is subjected to substantial and unnecessary costs. Whilst it is exasperating when patients don't show up at appointments, the explanation behind why this occurs should be explored to remove omitted appointments as this would be commercially favourable. Therefore, this work is undertaken as the government wants to answer whether there is necessary staff in the networks as well what the substantiative implementation of resources using a data – driven procedure via python.

ANALATYICAL APPROACH

The initial work revolved around analysing the actual_duration.csv, appointments_regional.csv and national_categories.xlsx files, with an emphasis on the use of pandas data frames. Therefore, pandas and numpy were imported initially before the 3 files were imported and read using the read_csv() and read_excel() methods as the ad, ar and nc data frames respectively. Each of the imported data frames was then sense checked using the df.head(), df.shape, df.columns and df.dtypes (where df = name of the data frame) in order to check that the data frames had been imported into python correctly.

Once it was confirmed that all the data frames had been imported correctly, an initial exploration into the locations was made. This included initially calculating and obtaining the top 5 locations using the nc dataframe. The code used for this initial step used the .value_counts() and .count() functions, ensuring that the initial list of unique locations was displayed in descending order of records so as to obtain the top 5 locations easily with the .head() function. Moreover, whilst the nc data frame was preferred to be used due to its large data set, the ad data frame was used to confirm that the number of unique locations obtained from nc was correct.

Following this, a more detailed investigation into the nature of the appointments was made. The areas that were explored were the earliest and latest recorded dates, the service setting with the most appointments in northwest London and the month with the most appointments. The .max() and .min() functions were used on the ad and nc data frames once the dates had been changed to date time format to obtain the earliest/latest dates whilst the most popular service setting was obtained by filtering nc for the correct location and time frame before using the .value_counts() function. On the other hand, the month with the greatest number of appointments was obtained with the nc data frame using groupby('appointment_month') as well as summing and sorting values by the appointment counts by using the .sum() and .sort_values() functions.

This was then followed by the creation of line plots for the monthly appointments per service setting, context type and national category before line plots illustrating the number of appointments per service setting on a seasonal basis were made, ensuring that the necessary data frames incorporating the required information was created initially in addition to importing the seaborn and matplotlib.pyplot libraries so as to create the visualisations. Note that in order to create the line plots, sns.lineplot(x=…,y=…,data=…,hue=…,ci = none), with ci = none being essential to enhance the plots.

The next step was to analyse twitter hashtags after importing and reading the tweets.csv file using the read_csv() method. The first item to be analysed here was the retweeted tweets count and the favourite tweets count. Afterwards, the individual hashtags associated with the tweets was

extracted with the help of a for loop appending each separate hashtag in addition to .value_counts() function being used to obtain the number of occurrences for each hashtag. This was then visulaised via a bar plot by using sns.barplot(x=…,y=…,data=…). Outlier values (healthcare: 716; health: 80; medicine: 41) were then removed from the original dataframe after checking via the use of a boxplot, which was created using sns.boxplot(y =…, data =…) and a further bar plot was then created to illustrate the effects of this.

Finally, a detailed analysis into whether the NHS should increase staffing capacity, the variation of healthcare professionals, significant changes in whether visits are attended, variations in the booking times, variations in the appointment type and how the various service settings compare. Conclusions were then drawn after plotting various visualisations once the appropriate data sets had been created, considering that the NHS can accommodate up to 1.2 million appointments a day.

Visualisations And Insights

I. What is the number of locations, service settings, context types, national categories, and appointment statuses in the data sets?

- Number of unique locations: 106
- Number of Service Settings: 5
- Number of Context Types: 3
- Number of National Categories: 18
- Number of Appointment Statuses: 3

II. What is the date range of the provided data sets and Which service settings reported the most appointments for a specific period?

- "actual_duration": 1st December 2021 - 30th June 2022
- "national_categories": 1st August 2021 – 30th June 2022
- "appointments_regional": January 2020 – June 2022

Moving on to the appointments per service setting the following graph illustrates the trend for the number of appointments per service setting:

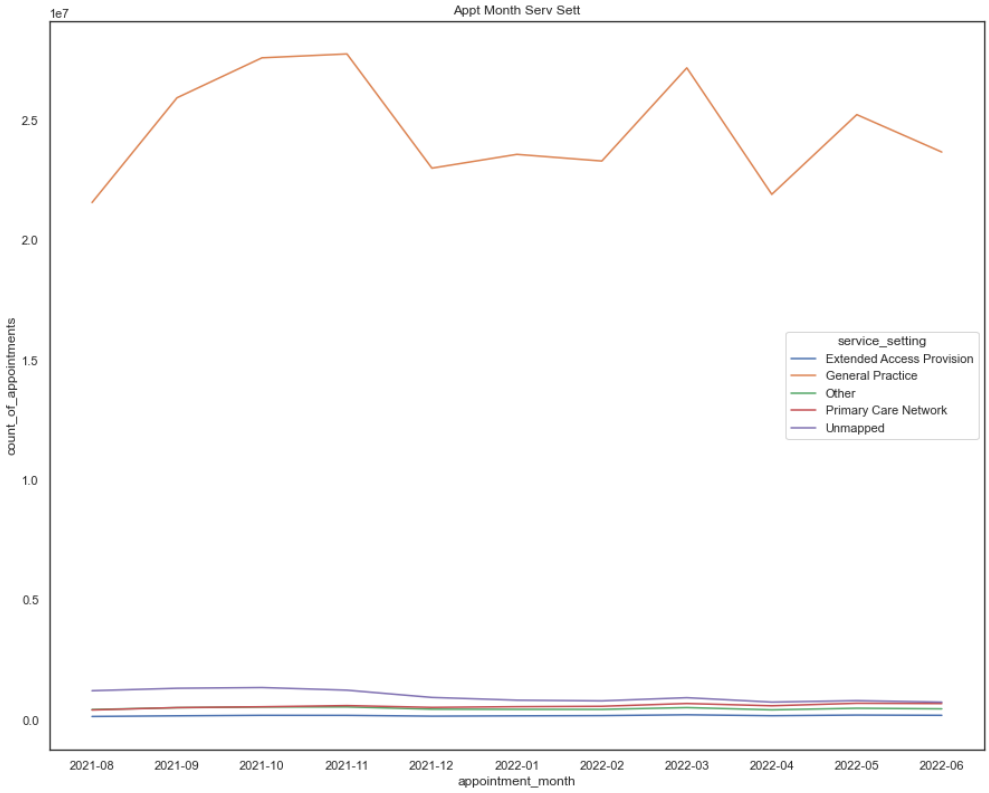


Figure 1. Monthly Appointments grouped by service setting from August 2021 to June 2022

As can be observed from figure 1, the service setting that reported the most appointment from August 2021 – June 2022 was General Practice, with the volume of appointments for the other service settings being negligible compared to this.

III. What is the number of appointments and records per month?

Month	Records	Appointments
Aug - 21	69,999	23,852,171
Sept - 21	74,922	28,522,501
Oct - 21	74,078	30,303,834
Nov - 21	77,652	30,405,070
Dec - 21	72,651	25,140,776
Jan - 22	71,896	25,635,474
Feb - 22	71,769	25,355,260
Mar - 22	82,822	29,595,038
Apr - 22	70,012	23,913,060
May - 22	77,425	27,495,508
June - 22	74,168	25,828,078

IV. What monthly and seasonal trends are evident, based on the number of appointments for service settings, context types, and national categories?

The monthly trends for national categories and context types are shown in the following figures:

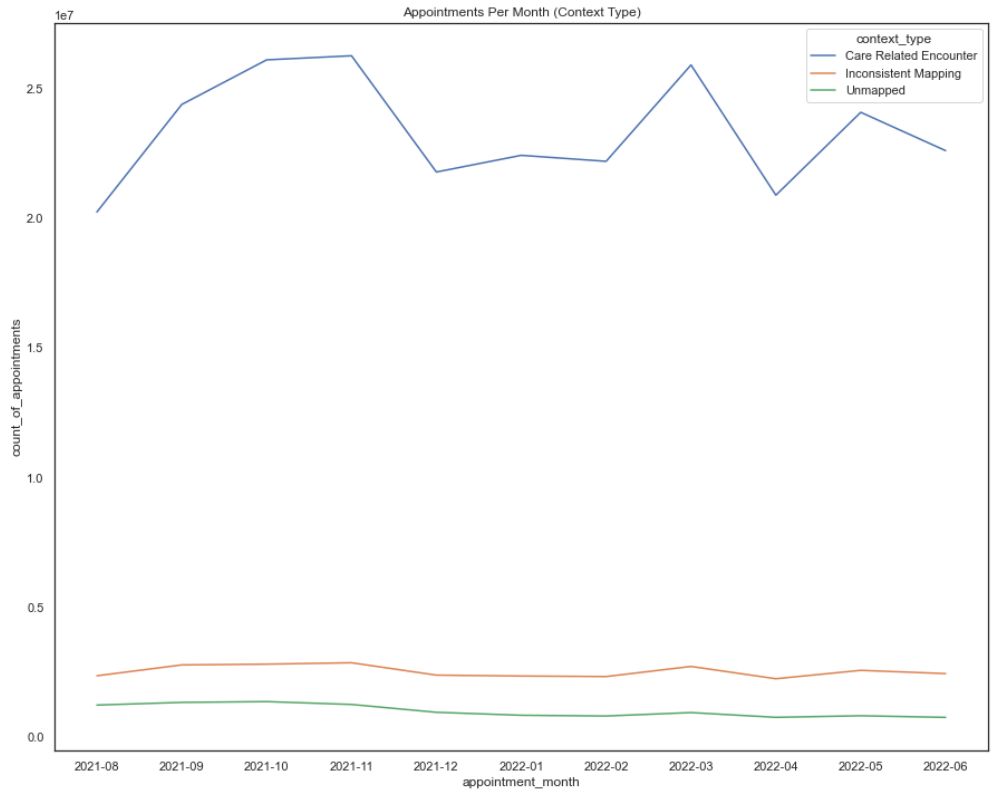


Figure 2. Monthly Appointments grouped by context type from August 2021 to June 2022

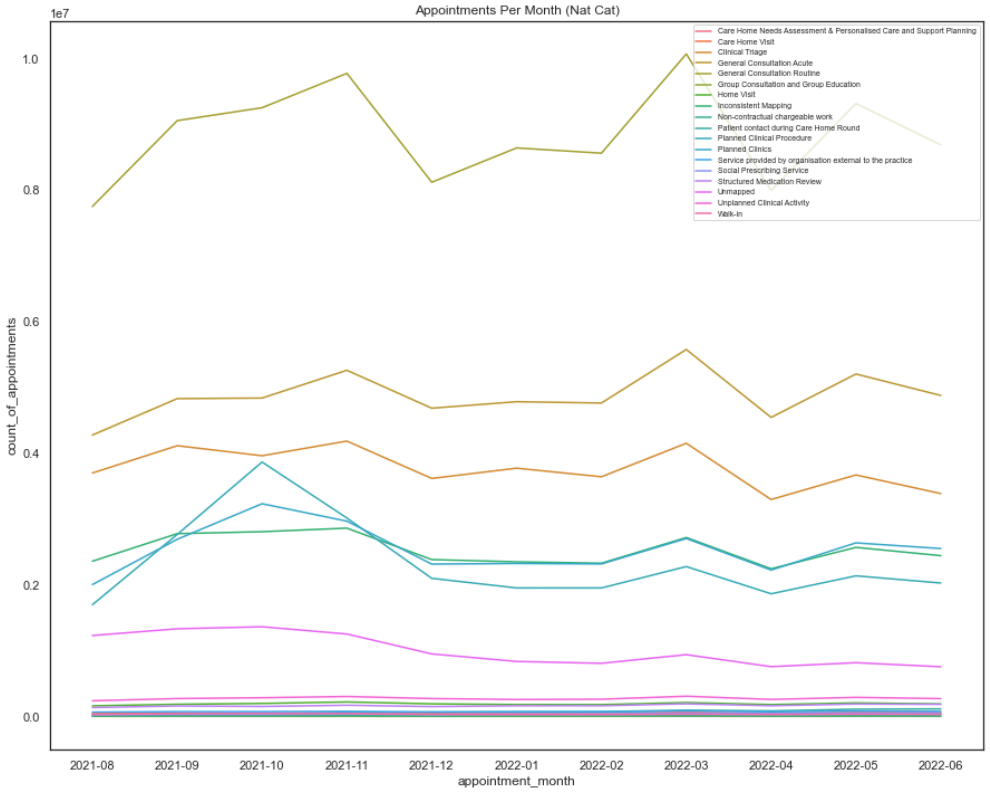


Figure 3. Monthly Appointments grouped by national categories from August 2021 to June 2022

For The Most Common Group: The 3 monthly appointment line plots show that the demand increases from August to November before rapidly declining in December. Then the demand is either stagnant or slightly increases between December and February, before the demands alternates between increasing and decreasing starting with an increase from February to March

All Other Groups: The trend for all other groups can either mirror the most common groups or they remain flat/constant.

The seasonal variations for the appointments per service setting is also shown below as follows:

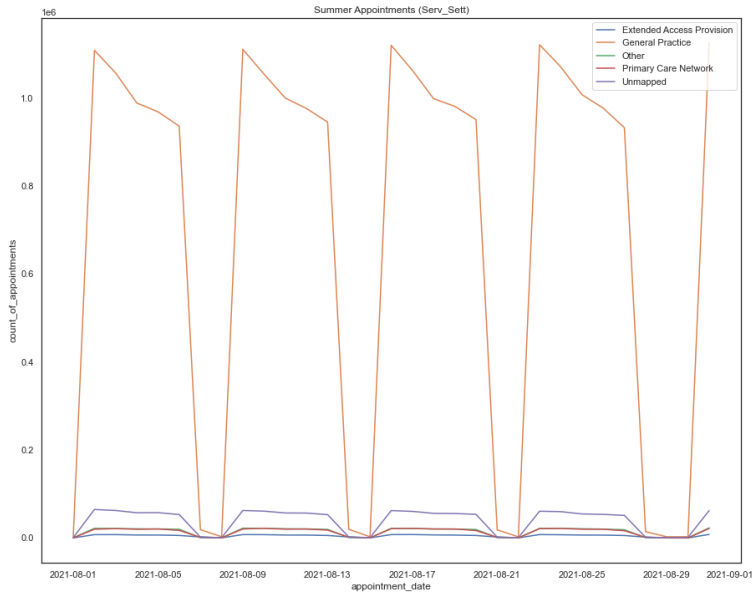
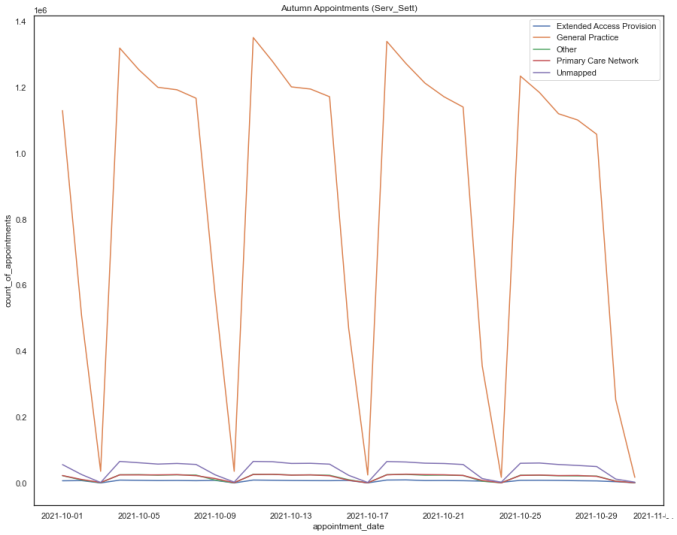
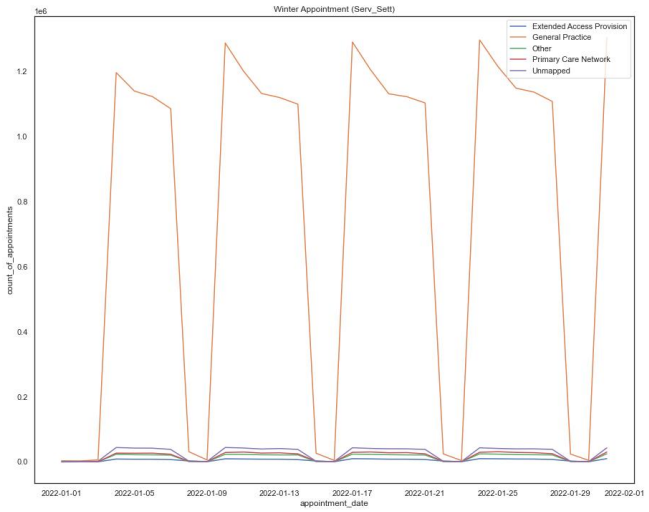


Figure 4. Service Setting Variation for Summer





Figures 5 & 6. Service Setting Variation for Autumn & Winter

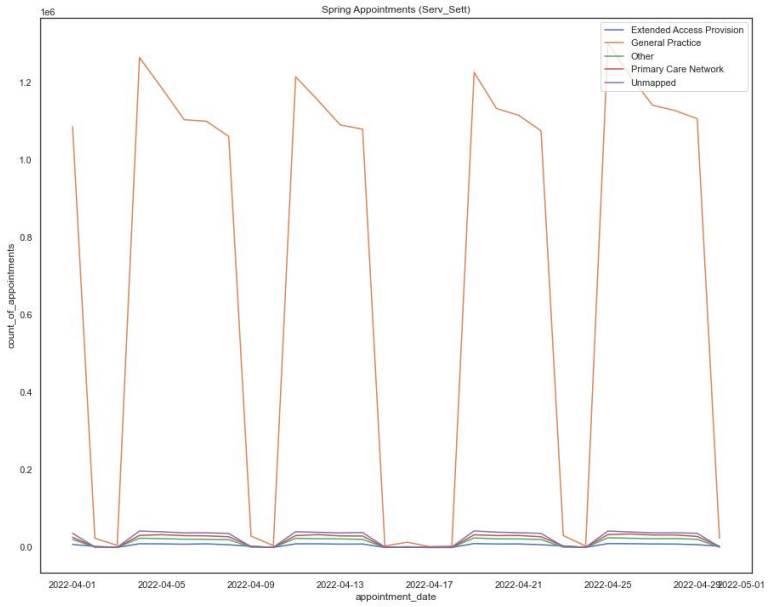


Figure 7. Service Setting Variation for Spring

For all 4 months, the Mondays are the busiest, with the number of appointments decreasing throughout the week, with Saturdays and Sunday having essentially no appointments (possibly because practitioners' hours are more limited on a weekend). The number of appointments tend to increase on a daily basis as you move from summer to autumn, although by the end of autumn, the number of appointments notably decreases compared to the rest of the season. It is only until halfway into winter that appointments begin to increase with no clear trend as to an increase/decrease in the number of appointments for spring.

V. What Are the Top Trending Hashtags On Twitter Related To UK Healthcare?

The following bar plot shows the overall spread of the hashtags analysed:

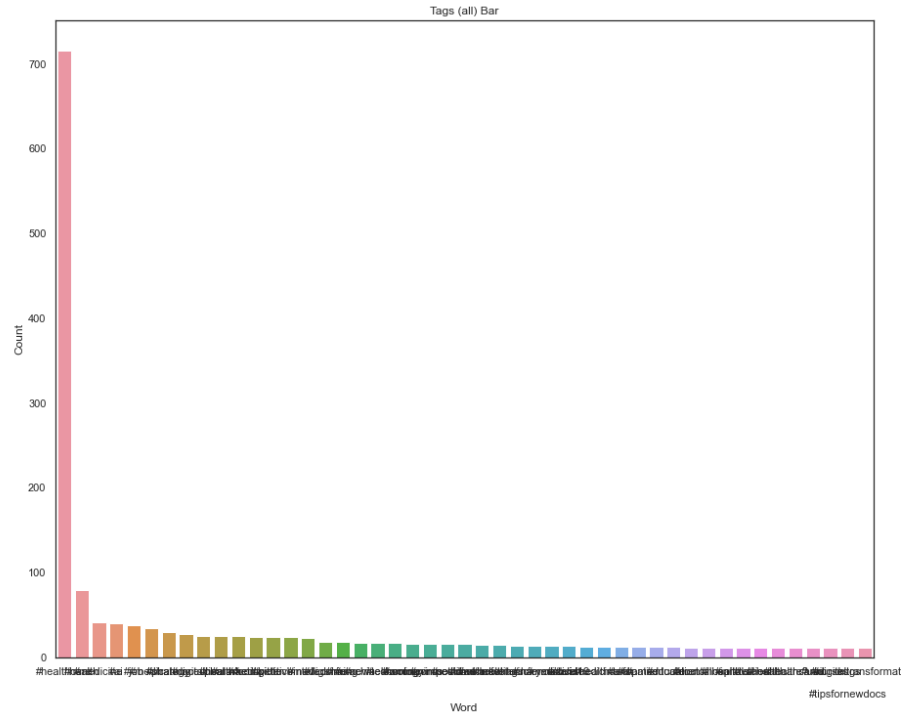


Figure 8. Frequency of Each Tag Bar Plot

Looking at figure 8, it can be observed that healthcare is the most frequently used hashtag. However, compared to other tags this is anomalously high as well as the fact that considering we are looking at healthcare tweets it doesn't offer many insights. Therefore, a second bar plot shown below in figure 9, which removes the outliers (healthcare, medicine, and health) is shown below as follows:

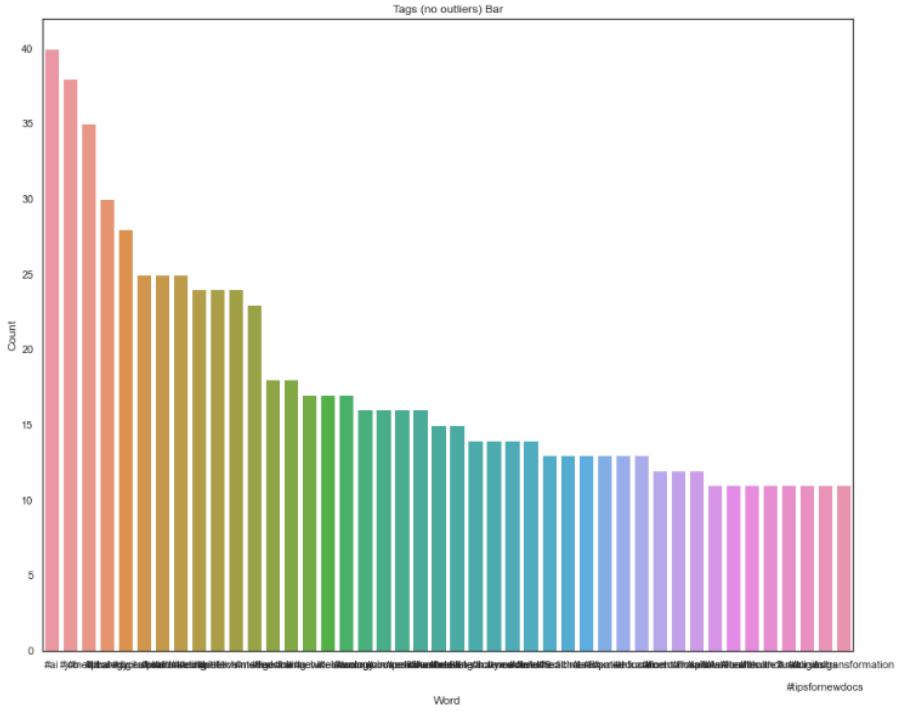


Figure 9. Frequency of Each Tag Bar Plot, With Outliers Removed

Therefore, after removing all outliers, it was found that the top trending tweets were ai, medical and strategy.

VI. Utlisation Of Resources

The following plot shows the average daily utilsation of resources per month

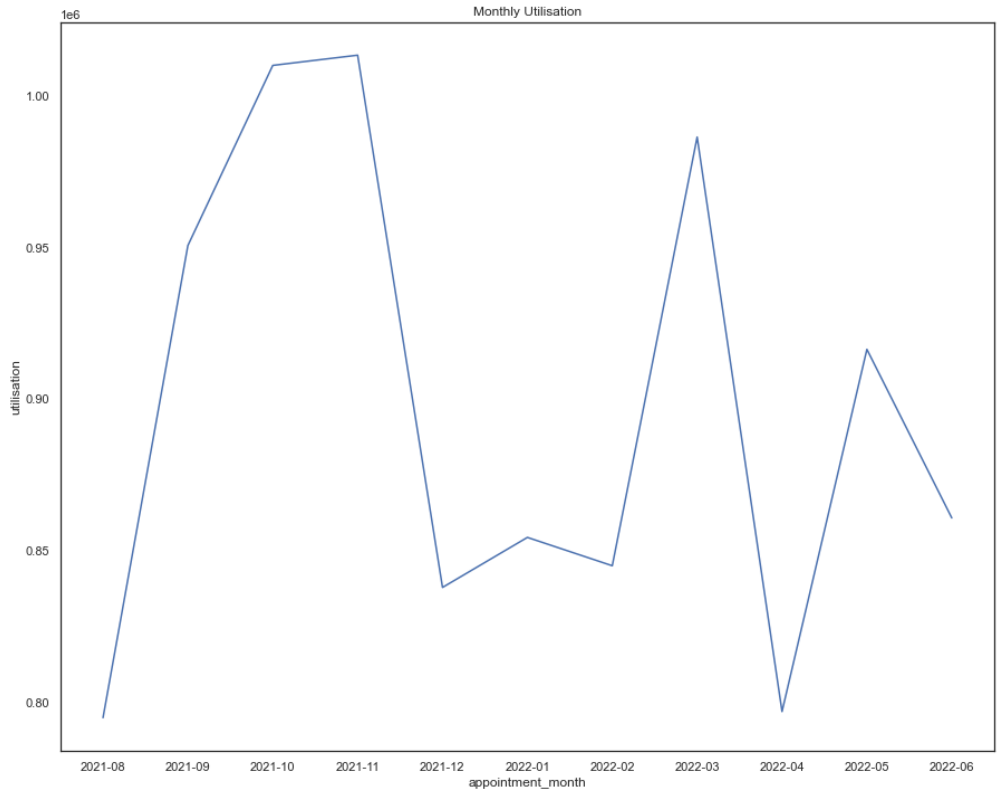


Figure 10. Line Plot Depicting Average Daily Utilisation

- Utilisation max = 1.014 x 10^6
- Utilisation min = 0.795 x 10^6
- NHS Max Capacity = 1,200,000 = 1.2 x 10^6

Final Recommendations

Firstly, whilst the utilization value never exceeds the NHS max capacity (meaning that based on the limited data, the NHS doesn't need to raise staff levels), it would be wise to increase the staff levels around the months of September – November as the utilsation levels here are far greater than the typical value observed elsewhere. Then looking at the trending hashtags, considering that ai is the top tag, consumers believe that the use of technology is key for medical strategies in current times so it may be good to start looking for ways in which technology can complete tasks that are manually done to free up staff that are otherwise occupied. Then, considering there are many national categories and service settings that have negligible demand in addition to the fact that two of the three context types also fall into this category, it might be useful to cut down on or remove these services all together and only focus on delivering those that are in high demand. Other conclusions that can be arrived at is that people prefer face to face appointments or telephone appointments as opposed to other modes (figure 11) and that a significant proportion of patients book within 14 days, with most booking on the same day (figure 12) so it might be beneficial to only offer face to face or telephone appointments, with a focus on implementing a 14-day window between bookings and appointments. One final point to be made is that whilst the number of people attending appointments does fluctuate, the number of 'unknown' or 'did not attend' appointments remain relatively constant (figure 13) so this fluctuation could be a natural consequence of the respective months having fewer bookings.

Appendix

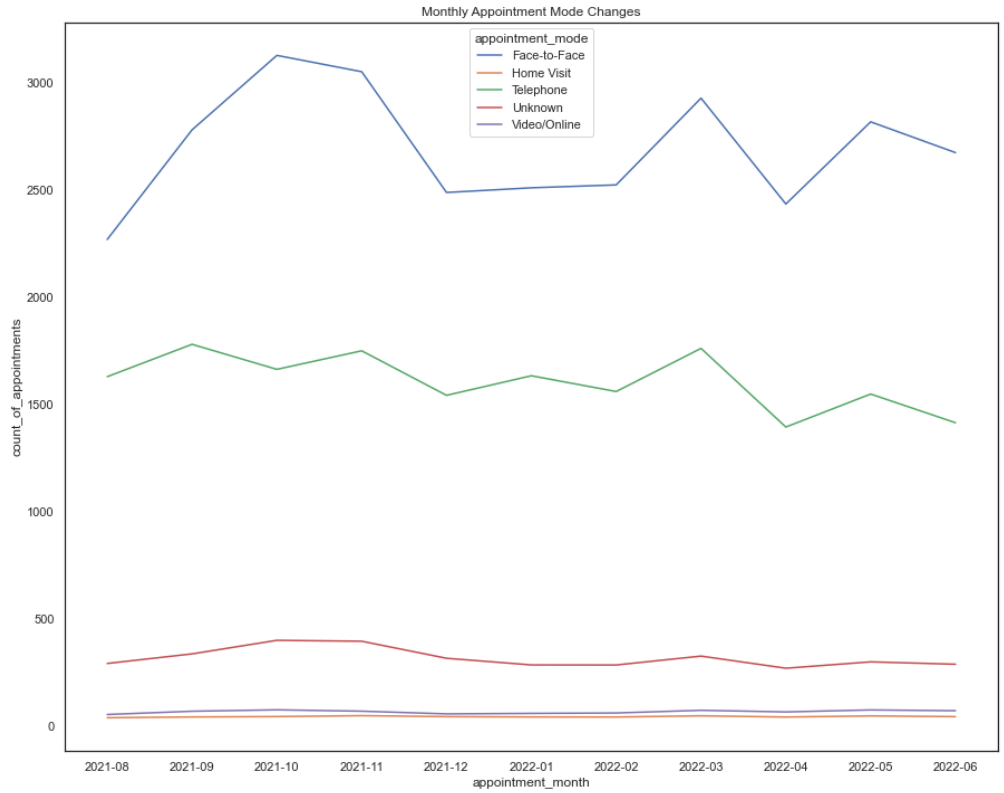


Figure 11. Variation in Appointment Modes

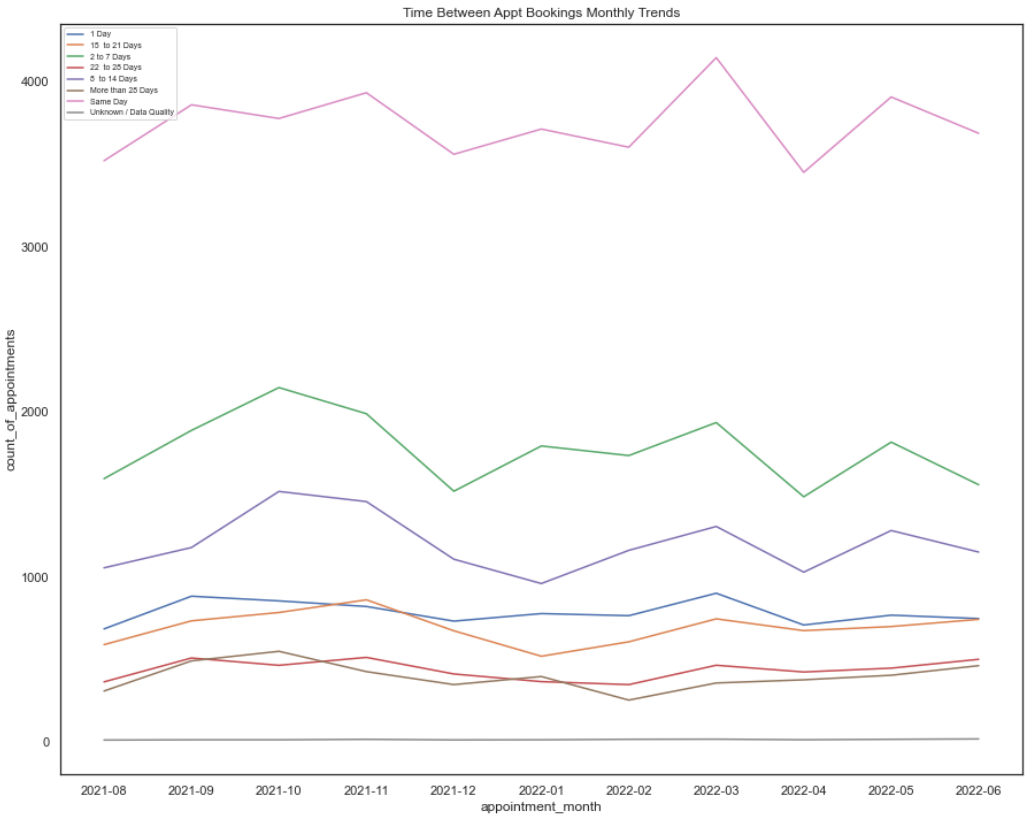


Figure 12. Variation in Appointment Booking Times

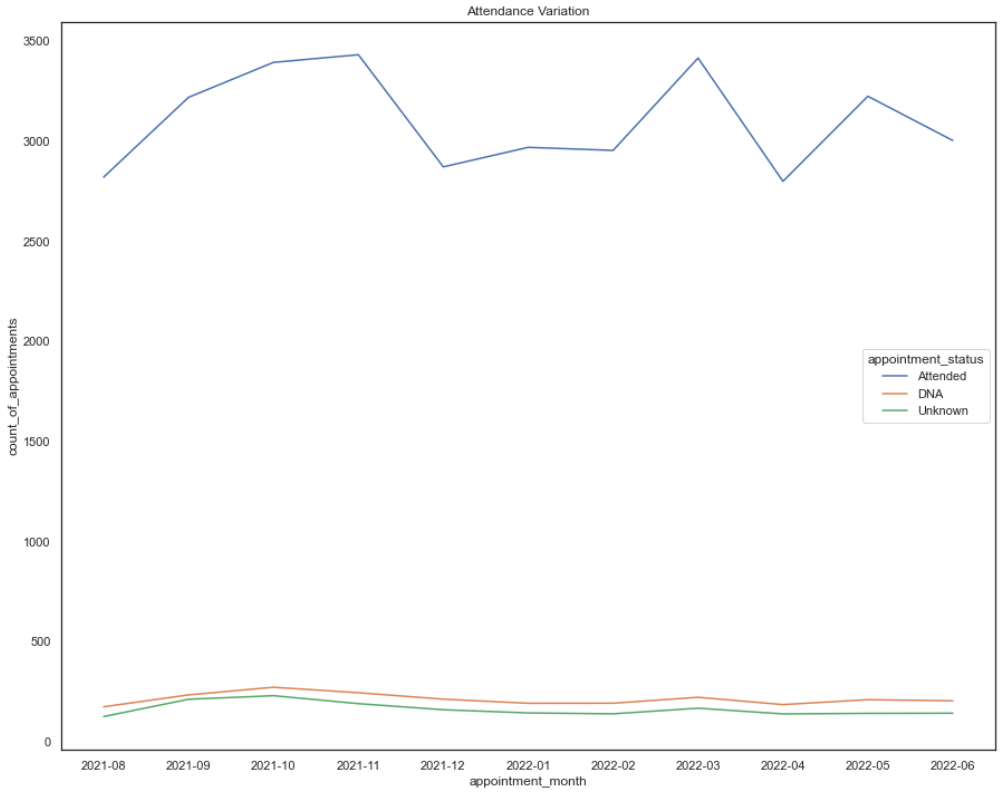


Figure 13. Variation in Appointment Attendance