

Diagnostic analysis using Python

Context.

It has been continuously reported in the media that NHS in the UK suffers from missed medical appointments by the patients which results in wasted time and financial resources. There have been numerous debates whether the patients should be financially penalized for non-attendance. However, the chair of the British Medical Association believes that financially penalizing the non-attendance might exacerbate the situation for the poorest and most vulnerable members of the society. Thus, this report outlines the diagnostic analysis that was undertaken to attempt looking into the problem and some recommendations are given at the end of this report.

Analytical approach.

The data sets provided include four files: 'actual_duration.csv', 'appointments_regional.csv', 'national_categories.xlsx', 'tweets.csv' and a text file with meta data. As indicated in the metadata, the data sets were cleaned, and some unnecessary columns were dropped to decrease the file size and for the ease of analysis.

First stage of the analysis includes importing the data sets and exploring the data. The libraries necessary for the analysis were imported first in Jupyter notebook, namely: pandas, matplotlib.pyplot and seaborn for visualisations. To explore the data, all the data sets were checked for types of data, column names and missing values.

```
# Create a DataFrame 'Actual_Duration'
ad = pd.read_csv('actual_duration.csv')

# Sense check the data frame
print(ad.shape)
print(ad.dtypes)
print(ad.columns)
missing_values = ad.isna().sum()
print(missing_values)
ad.head()
```

Moreover, as part of the exploration process, the unique ICB (integrated care board) locations were identified. There are 106 overall presented in the data sets with the following locations having the highest number of records: Birmingham and Solihull ICB, NHS Black Country ICB,

NHS Bath and North East Somerset Swindon, etc.

```
# What are the five locations with the highest number of records?
ad_usecols = pd.read_csv('actual_duration.csv', usecols = ['sub_icb_location_name',
                                                           'count_of_appointments'])

ad_usecols.head
print(ad_usecols.groupby('sub_icb_location_name').sum().head())
```

sub_icb_location_name	count_of_appointments
NHS Bath and North East Somerset Swindon and Wi...	3012568
NHS Bedfordshire Luton and Milton Keynes ICB - ...	2791385
NHS Birmingham and Solihull ICB - 15E	3600087
NHS Black Country ICB - D2P2L	3901431
NHS Bristol North Somerset and South Gloucester...	2756491

The data sets contain a large amount of data and records were taken during a considerable period, ranging from the first of December 2021 to the thirtieth of June 2022 in the 'actual_duration' data set and from the first of August 2021 to the thirtieth of June 2022 in 'national_categories' data set.

```
# Determine the minimum and maximum dates in the ad DataFrame.
```

```
print(ad['appointment_date'].min())
print(ad['appointment_date'].max())
```

```
2021-12-01 00:00:00
2022-06-30 00:00:00
```

```
# Determine the minimum and maximum dates in the nc DataFrame.
```

```
print(nc['appointment_date'].min())
print(nc['appointment_date'].max())
```

```
2021-08-01 00:00:00
2022-06-30 00:00:00
```

If we look at the number of appointments per month, it is evident that October, November, and March are the months with the highest number of appointments and thus, a heightened pressure on the healthcare system in general, while, starting from April and during the summer months, there is an evidence of decreased healthcare demand. Perhaps this is due to the seasonality of illnesses, warmer weather, holidays, etc.

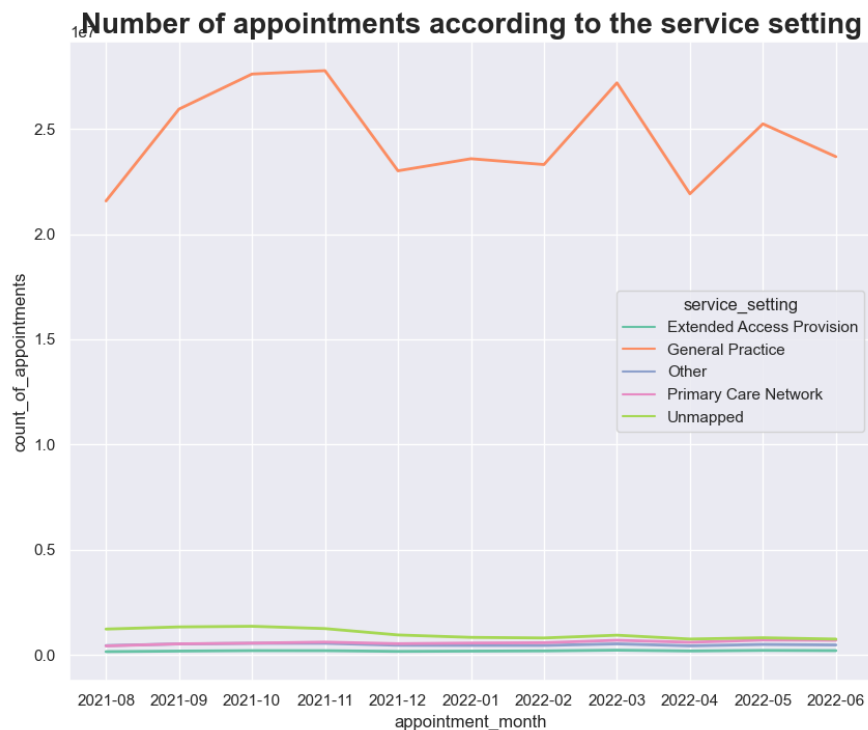
Q3. Which month had the highest number of appointments?

```
: print(nc.groupby(nc['appointment_date'].dt.strftime('%B')) \
      ['count_of_appointments'].sum().sort_values(ascending=False))
```

```
appointment_date
November      30405070
October       30303834
March         29595038
September     28522501
May           27495508
June          25828078
January       25635474
February      25355260
December      25140776
April         23913060
August        23852171
Name: count_of_appointments, dtype: int64
```

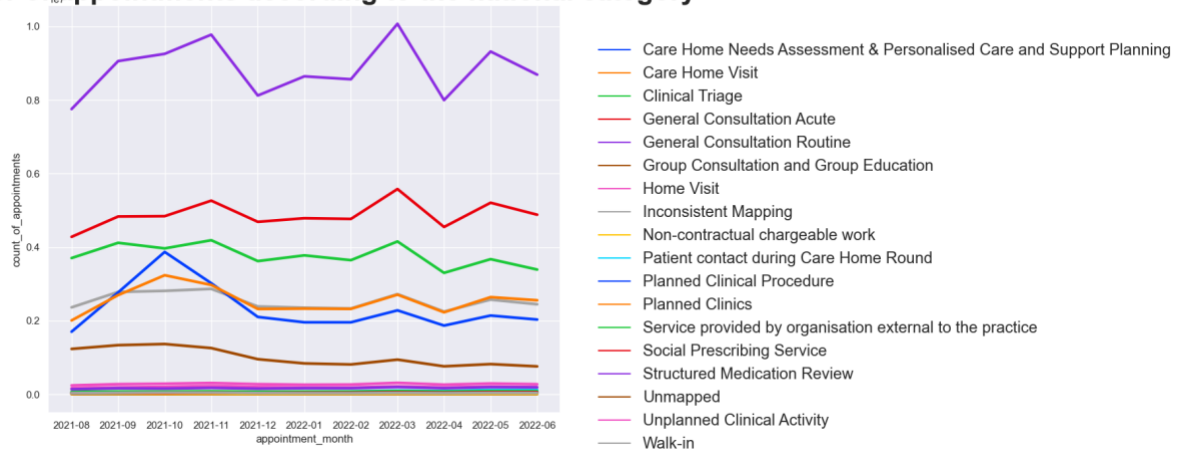
Visualisation and insights.

If we take the visualisations for the number of appointments according to the service settings, it's obvious that the highest number of appointments were scheduled for general practice with the peaks during the months of October, November 2021 and March 2022. The other service settings (extended access provision, primary care network and some other unmapped services) have been remaining relatively stable over the reported months.



Similarly, if we take context types, then the care related encounters prevail, which include several different categories, for example, general consultation acute, general consultation routine, planned clinical procedures, walk-ins, home visits, etc. Out of all these, routine general consultations have the highest number of visits, with general consultation acute and clinical triage taking second and third place respectively (see the fig. below). For this visualisation, a line graph was chosen and a colour palette 'bright' was chosen so that the categories stand out and are clearly visible.

Number of appointments according to the national category

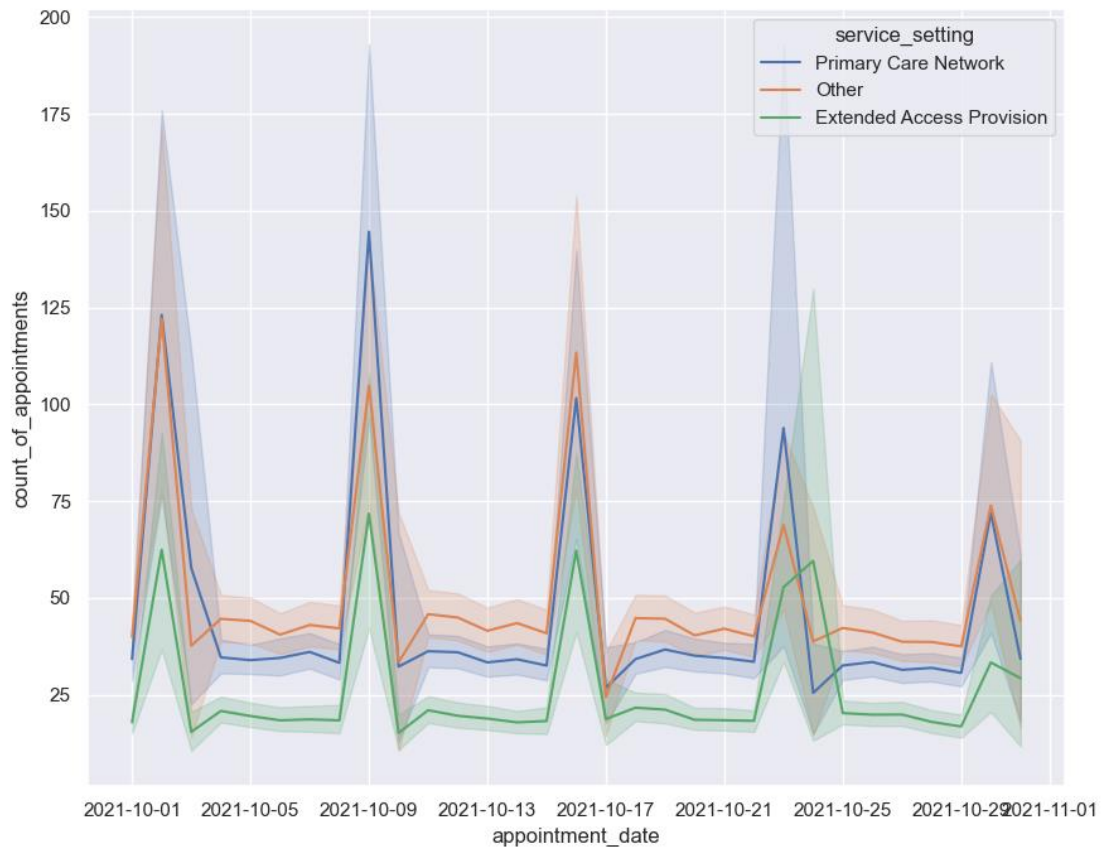


When it comes to service settings per season, according to the graphs, in all the seasons (summer, autumn, winter and spring), the general practice remains the most popular service setting. However, having a closer look at the least represented service settings, and excluding the 'general practice' and 'unmapped' categories, help us understand that primary care network remain consistently stable in the summer months. However, if we take the month of October, this is when the pressure on the primary care network and extended access provision increases. This is evident from the chart as well as from the fact that the increased number of appointments come on weekends. This was identified by checking the days with the highest numbers of appointments.

```
# To find out the day of the week of the dates presented
```

```
past_date = datetime.date(2021, 10, 16)
day_of_week = past_date.strftime('%A')
print(day_of_week)
```

Saturday



Twitter data.

Twitter data scraping was a very interesting part as I had to scrape the hashtags and, eventually, see what the interest of people is like in healthcare and health. To begin with, I have imported the necessary libraries, then adjusted the column width of the data frame to 200 to ensure that longer text values, such as tweets, are displayed completely.

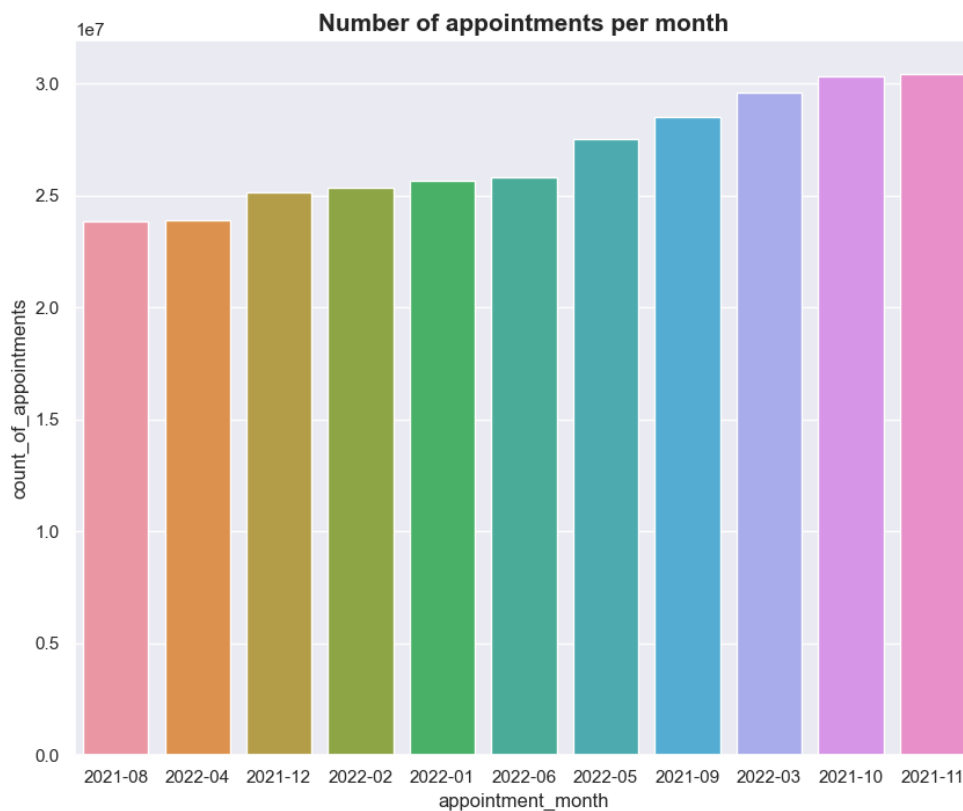
During the analysis of the hashtags, the new data frame called 'tweets_text' was created to focus only on the textual content of the tweets. This allowed to focus on the textual information of the tweets only. To analyse hashtags, an empty list named 'tags' is initialized.

By iterating over the tweets using a 'for' loop, hashtags were extracted and appended to the 'tags' list. I then determined the frequency of each hashtag, by creating a pandas Series and determining the top trending hashtags on Twitter. It was identified that the top three trending hashtags were #healthcare, #health and #medicine.

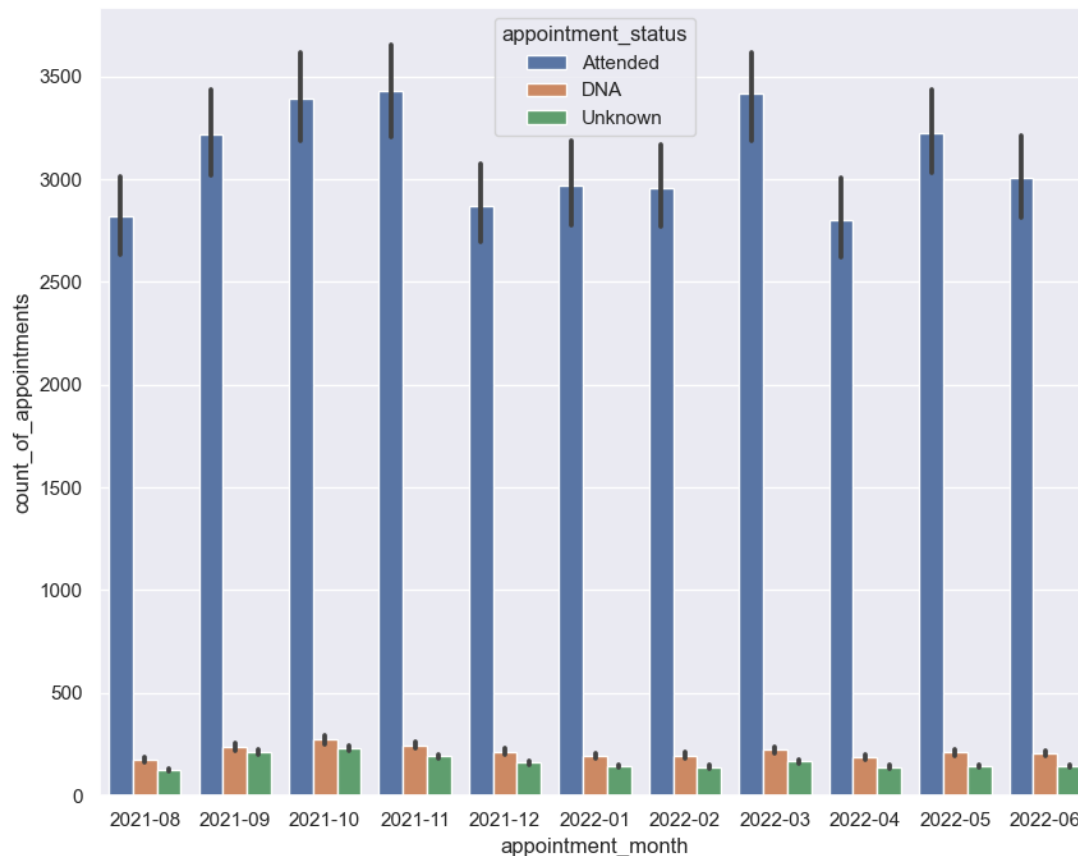
Recommendations

Generally, healthcare and medicine have been topics that are important to the British citizens; this can be supported by the twitter data. Furthermore, the data was collected during the time of COVID-19 pandemic and there was a heightened pressure on the healthcare and overall difficult mood among public.

It has been identified here during the diagnostic analysis that the busiest months for the NHS were October 2021, November 2021, and March 2022, and, most likely, this trend could be a regular situation considering seasonal illnesses.

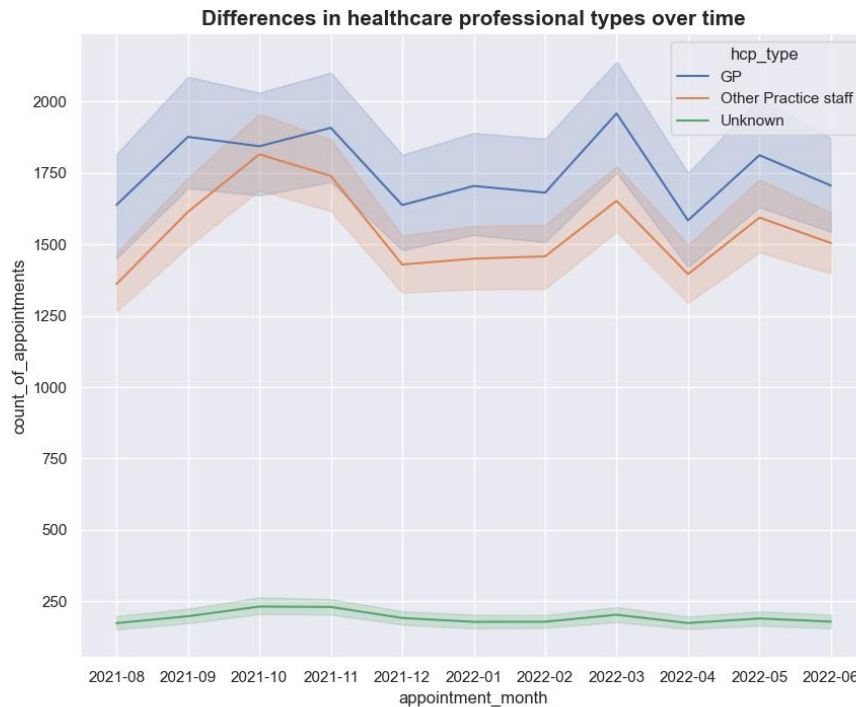


During the busiest months (October, November and March), there was a slight, but visible rise of the DNA (did not attend) status which could also be due to heightened pressure on the resources of the healthcare system and availability of medical staff.



Considering the above-mentioned data trends, there are three main recommendations that arose as a result of this diagnostic data analysis:

1. During the heightened seasons, namely, October, November and March, NHS should continue to **hire and engage people in the primary care network** and extended access provision service setting.
2. It is also evident that the **General practitioners** have the highest amount of workload, again, especially during the busiest times in the autumn and March. Therefore, hiring more GPs should be considered a top priority for the hospitals.



- In a research article from year 2005, about the topic of missed appointments in the UK, retrieved from this [web-site](#), it was reported that the main reasons for people missing the appointments were, among others: organisational factors (inability to cancel the appointment or reach the hospital to do so), too big of a time gap between the time of booking and the appointment, time management issues or just a matter of forgetfulness. Therefore, **as a third recommendation, I would suggest hiring more support staff** dealing with organizational and administrative issues, such as calling the patients and sending them reminders beforehand. There are numerous applications for booking and managing the bookings nowadays, such as NHS app, Patient Access or MyGP, however, for the older population, it might still be very beneficial to have more support and administrative staff in place to assist with booking, calls and cancellations.