

# **NHS Diagnostic Analysis: Executive Summary**

**Prepared for Willis Hurds, Head of Specialised Team at NHS England**

**Prepared by Alejandra Anguil Vanier, Data Analyst at dAtaAnalysisAV**

**February 2023**

## **Contents:**

<b>Background and Context .....</b>	<b>3</b>
<b>Analytical Approach .....</b>	<b>4</b>
Data Ingestion.....	4
Describe the Data .....	4
Data Wrangling .....	5
Exploratory Analysis.....	6
<b>Visualisation and Insights .....</b>	<b>6</b>
<b>Patterns and Predictions .....</b>	<b>10</b>
Recommendations and next steps.....	11

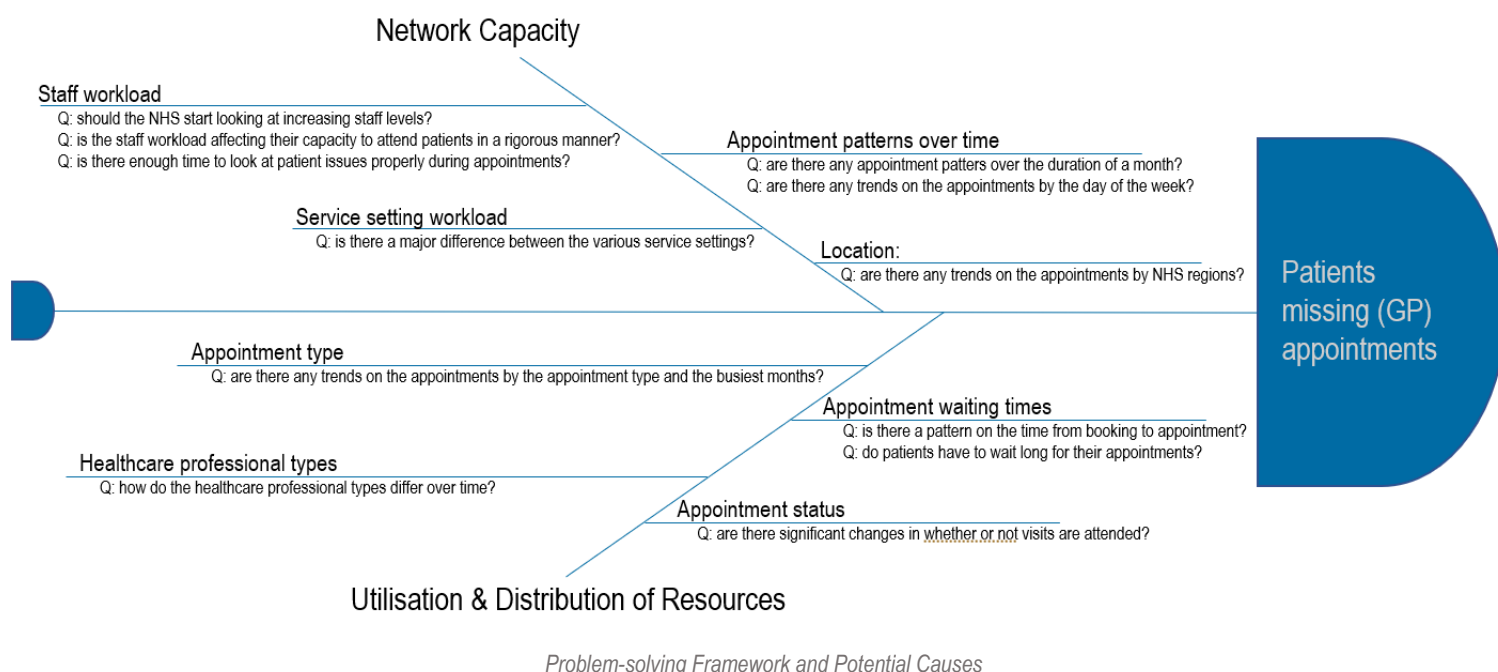
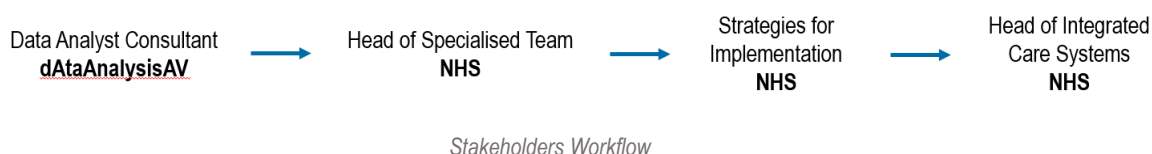
## Background and context

The NHS, a publicly funded healthcare system in England, incurs in significant, potentially avoidable, costs when patients miss general practitioner (GP) appointments. The reasons for this need to be better understood.

Reducing missed appointments would be beneficial financially as well as socially. At this stage of the project the two main questions posed by the NHS are:

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

We have been asked to look into the information about scheduled activity and usage of appointments that NHS Digital produces monthly (containing information about the previous 30 months, where available), to help inform the government how best to handle this problem.



## Analytical Approach

### 1. Data Ingestion

We have first proceeded to import the data into Python using the Pandas library and the correct format and names (descriptive names will make the next steps much easier to follow).

### 2. Describe the Data

To make sure the data has been loaded correctly and familiarise ourselves with it, we performed a descriptive statistical analysis (`info()`, `describe()`, `isnull().sum()`), which helps us to describe the characteristics of the datasets in a meaningful way, and potentially spotting errors, unwanted columns, missing values, etc.

Actual Duration (ad)	Appointments Regional (ar)	National Categories (nc)
sub_icb_location_code	icb_ons_code	appointment_date
sub_icb_location_ons_code	appointment_month	icb_ons_code
sub_icb_location_name	appointment_status	sub_icb_location_name
icb_ons_code	hcp_type	service_setting
region_ons_code	appointment_mode	context_type
appointment_date	time_between_book_and_a ppointment	national_category
actual_duration	count_of_appointments	count_of_appointments
count_of_appointments		appointment_month

*Original Column Distribution for all Datasets*

We can see already the columns we'll use later on to perform exploratory analysis, but also some columns on the ad data set that will probably not tell much for this analysis. However, having the region names would be more beneficial for future presentations to the stakeholders.

### 3. Data Wrangling

We have proceeded to clean the data in Python by:

- Identifying and removing unwanted columns with the `drop()` function and including columns which could add value to the analysis using the `merge()` function.
- Identifying duplicates in Excel on the basis of duplications in all columns. However, we have only found duplicates in the ar dataset and think they could be genuine values as there are so many locations that two records could be the same. Could we please request to confirm this assumption is correct?
- Fixing inconsistencies:
  - Date columns converted to date data types by using `to_datetime()` function.
- Identifying and dealing with outliers using scatter plots: after doing this exercise in Python we can see that by removing the outliers found in each dataset we lose a considerable amount of records that would be translated into millions of appointments. In this particular case, we don't feel this approach is appropriate as it won't be a true reflection of the NHS workload reality. Therefore, we are going to proceed with the analysis on the original datasets.
- Assumption: each row on a dataset corresponds to a record.

Actual Duration (ad)
sub_icb_location_name
icb_ons_code
appointment_date
actual_duration
count_of_appointments
Region Name
appointment_month
weekday

Appointments Regional (ar)
icb_ons_code
appointment_month
appointment_status
hcp_type
appointment_mode
time_between_book_and_a ppointment
count_of_appointments
Region Name

National Categories (nc)
appointment_date
icb_ons_code
sub_icb_location_name
service_setting
context_type
national_category
count_of_appointments
appointment_month
Region Name
weekday

*Final Column Distribution after performing cleaning with Python for all Datasets*

#### 4. Exploratory Analysis

Using the same software, we then performed an exploratory analysis to understand the data by:

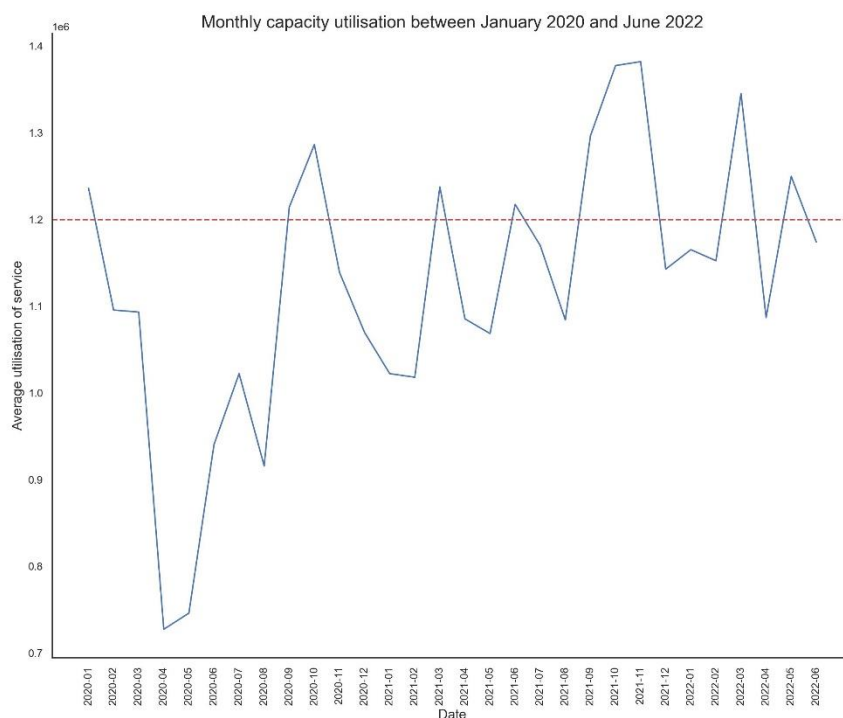
- The time period covered by the datasets
- Evaluate the number of different types for each variable (locations, service settings, context types, etc.)
- Understand the distribution of these variables by count of appointments and over time

### Visualisation and insights

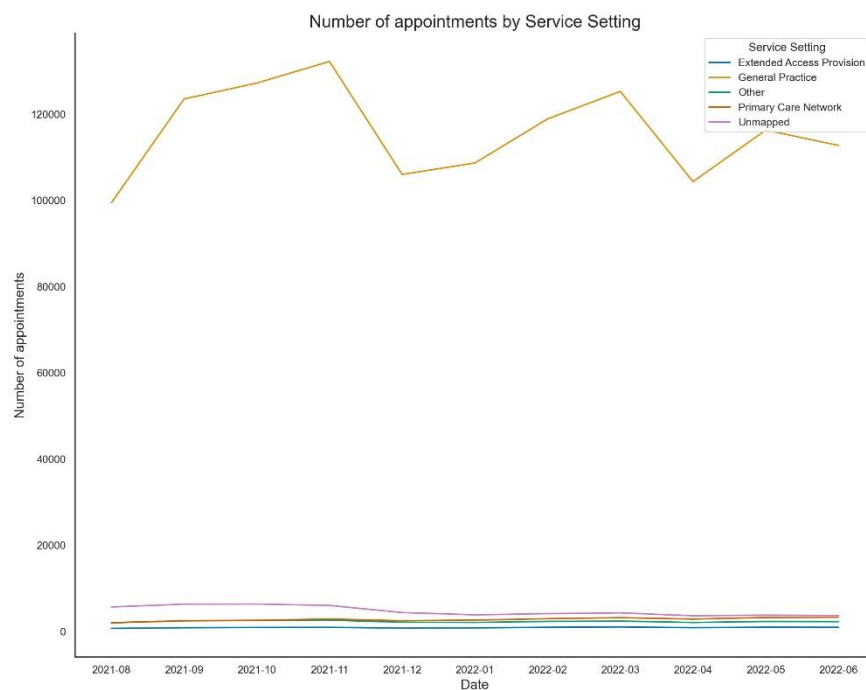
This step aims to perform an explanatory analysis of the data, in this case using Python, and the creation of a series of visualisations to present the insights discovered to the stakeholders. The visualisations have been chosen carefully to be simple to understand while conveying the message intended: mostly bar and line plots, which the vast majority of people is familiar with. The colour palette is consistent and avoids any colours that could be challenging for people with a colour vision deficiency.

These visualisations try to respond the initial queries posed by NHS:

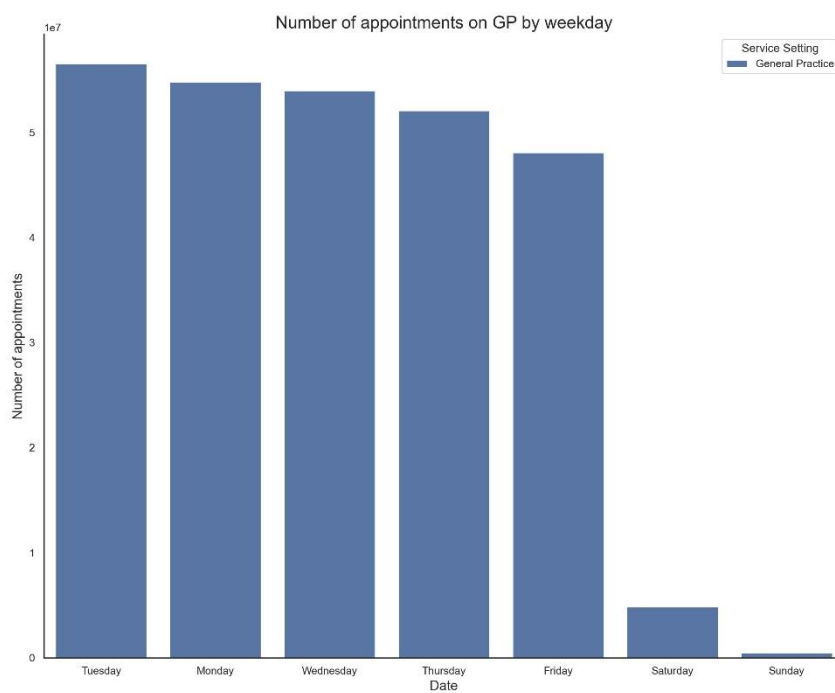
- Has there been adequate staff and capacity in the networks?
  - Staff workload



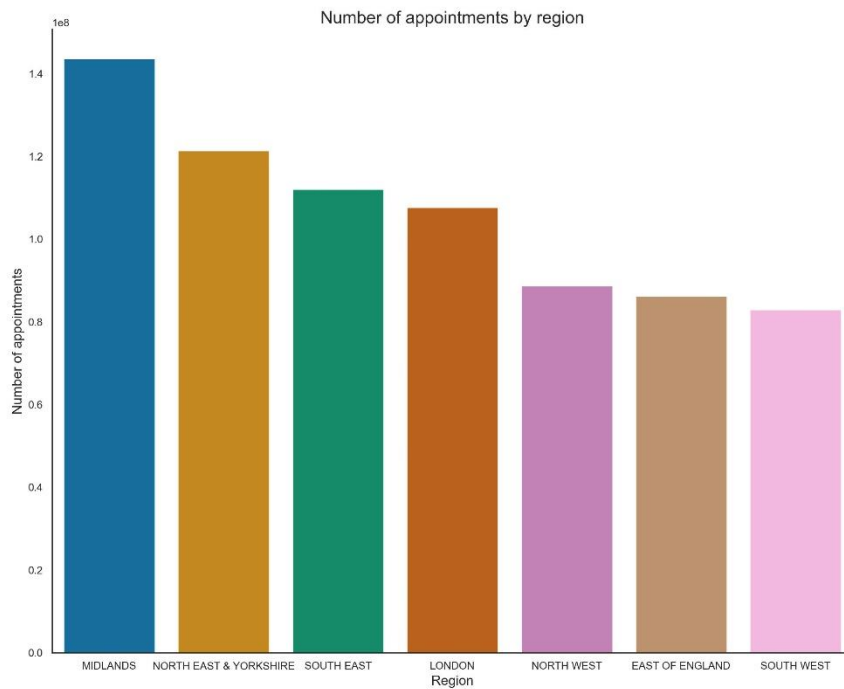
- Service setting workload



- Appointments by weekday

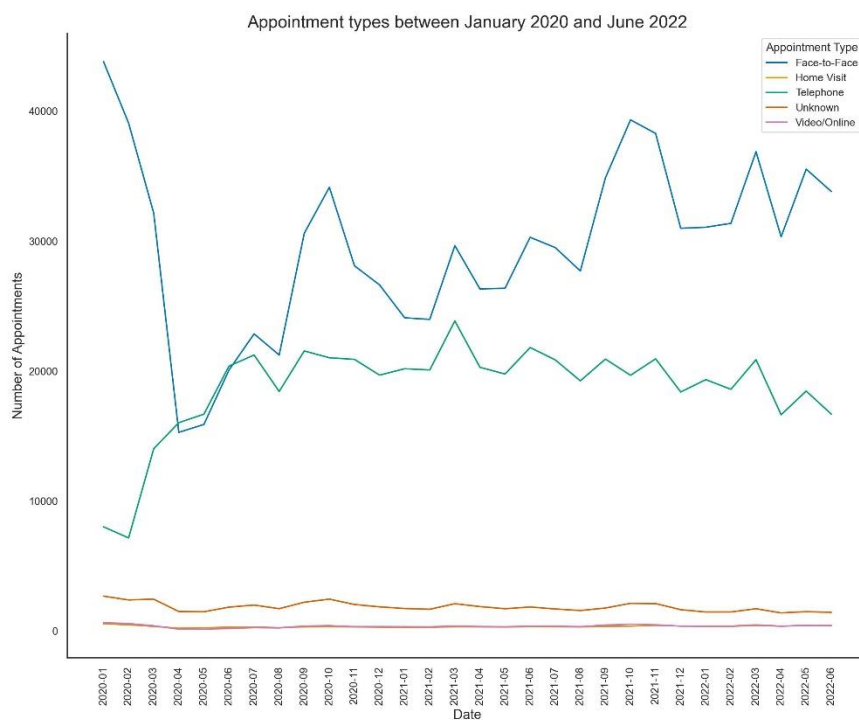


- Workload by regions



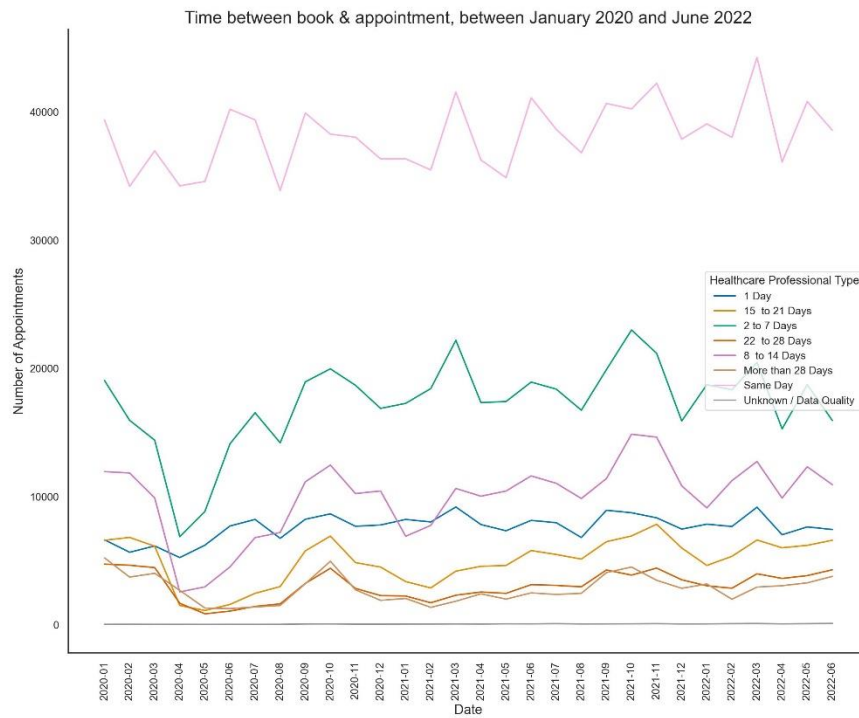
- What was the actual utilisation of resources?

- Appointment types

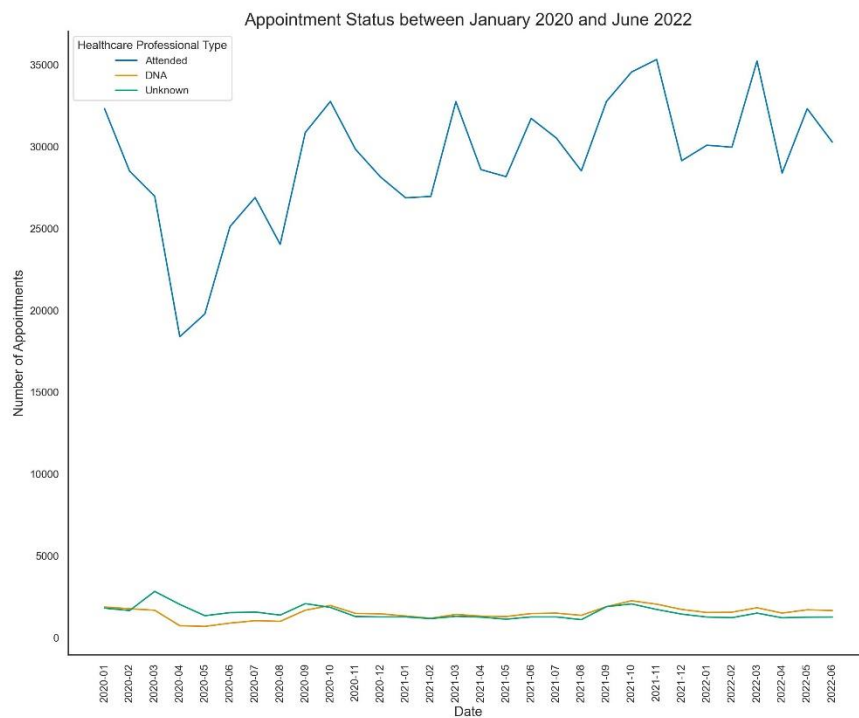




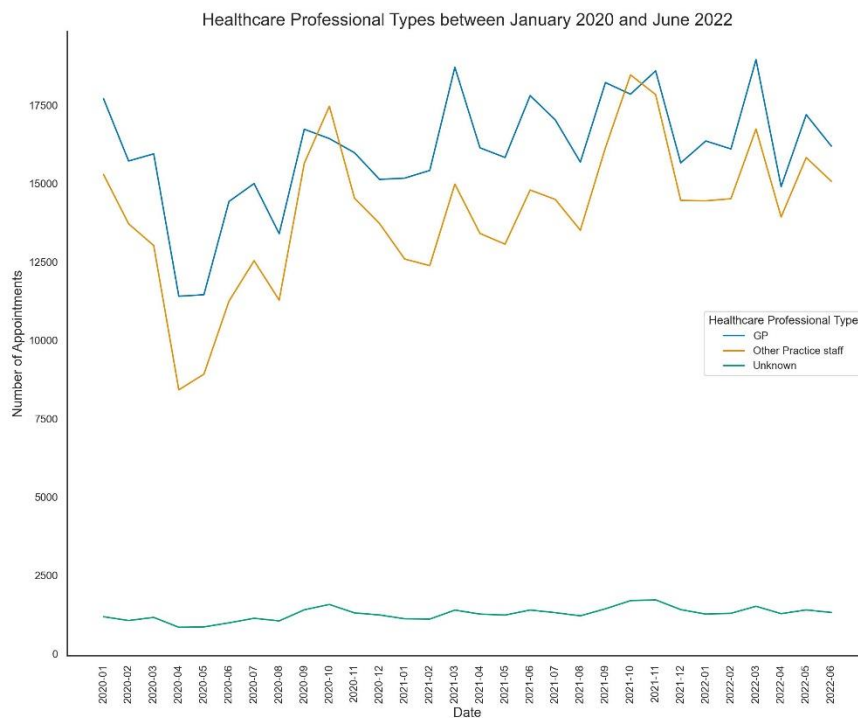
- Appointment waiting times



- Appointment status



- Healthcare professional types



## Patterns and predictions

Based on the analysis described above, we have found the following limitations:

- The quality of the data is not accurate, therefore these recommendations are not conclusive and further investigations are recommended.
- Too many unknowns and unmapped values makes the analysis weak.
- This historical data only covers from January 2020 to June 2022, which is a limitation for the analysis for two reasons:
  - Limited period of time covered
  - Coincides with the Covid19 pandemic, and therefore the data could differ vastly from a normal period of time

However, taking into account these quality considerations, we can conclude that:

1. Network Capacity
  - a. GPs are under pressure in terms of number of appointments. Resources need to be better distributed to balance these demands between the service settings and professionals.
  - b. Staff is at full capacity or beyond in certain periods of the year and needs to be reinforced (we have used 22 days per month to get the daily capacity utilisation due to discounting the weekends). These pressures could be affecting at:

- i. their capacity to attend patients in a rigorous manner
    - ii. having enough time to look at patient issues properly during appointments: we have discovered most of the appointments lasted less than 10 minutes, with a percentage of them even less than 5 minutes. This can be another sign of potential staff shortage.
  - c. The busiest NHS commissioning region is Midlands.
  - d. Weekdays are more busy than weekends, while Mondays and Tuesdays are busier than Wednesdays, Thursdays and Fridays.
2. Utilisation & Distribution of Resources
- a. Face-to-face and telephone appointments are the most used types by patients.
  - b. Too many appointments have a period bigger than 8 days between the booking and the appointment, and this could be another sign of the pressure GPs are under, as well as potential staff shortage.
  - c. October and November are the busiest months of the last two years, and it is when the most not attended appointments happen. More granular data is required to investigate why this is happening.

#### **Recommendations and next steps:**

- NHS should start looking at hiring more staff.
- The current staff could be reorganised:
  - o by region and
  - o by day of the month
- Standardisation of digital resources across the NHS network and staff training recommended to unify data collection
- Collect patient data related to religion, childcare necessity, etc. (while respecting their privacy) to perform a sentiment analysis to look at potential personal reasons for missing appointments
- Schedule a session to discuss data assumptions and potential further requirements

#### **DECLARATION**

Names, companies and logos used on this assignment (report, notebook and presentation) are my own invention (except NHS) for the purpose of this exercise and should not be used outside the LSE Data Analytics Career Accelerator programme