

# Data Analytics using Python: Assignment



Analysing NHS real-world data

Rafa Abdelazim  
LSE career accelerator

## **Contents:**

- *Background.....page 3*
- *Analytical approach.....page 4*
- *Visualisations and Insights.....page 6*
- *Patterns and predictions .....page 12*

## ***Background:***

Due to an increase of population, rise of Covid in the UK and aging population the increase in demand in the NHS has been more than ever before therefore the NHS needs to focus on efficiently utilising their resources and have adequate staffing and capacity to improve efficiency in healthcare services, meet increasing demands and improve patient satisfaction.

Through our analysis we will investigate answering questions posed by the NHS using python to use a data-informed approach to gain insights and identify trends with-in the NHS using real-world data from 2020 to 2022, to present to stakeholders.

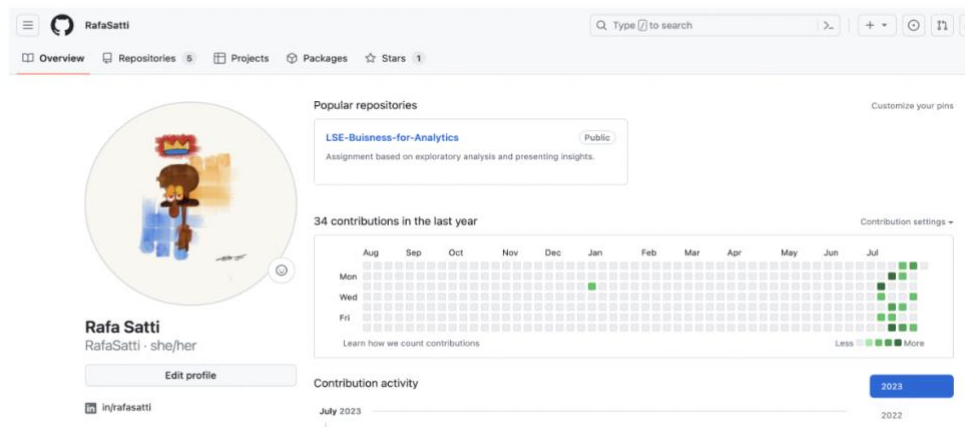
## Analytical approach:

The journey commenced by setting up the environment by importing crucial libraries (pandas, numpy, matplotlib, seaborn), and importing required datasets. Before delving into analysis, I conducted a thorough sense-checking of data. This included reviewing the first and last rows to understand the dataset's structure and ensure no metadata was erroneously stored. Additionally, I examined column names, data types, and the overall dataset integrity. With validated data, I explored metadata and statistical summaries to facilitate effective data pre-processing. This is to ensure data integrity and identify outliers to avoid skewing results.

## 1.0 Preparing the environment

### 1.1 Github link and account

<https://github.com/RafaSatti> - Github link



### 1.2 Import required libraries and datasets

```
In [4]: # Import all needed libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

print("Successfully imported!")
```

Successfully imported!

```
In [7]: # Upload all datasets
ad = pd.read_csv('actual_duration.csv')
ar = pd.read_csv('appointments_regional.csv')
nc = pd.read_excel('national_categories.xlsx')

print("Successfully imported!")
```

Successfully imported!

I then started my analysis. I identified the top 5 locations with the highest records using groupby and nlargest methods. Furthermore, I utilized groupby to compute total appointments and percentage distributions across diverse categories such as service settings, context types, national categories, and appointment statuses.

Visualization using seaborn was used to identify trends. The "appointment\_date" column was converted to datetime format to analyse scheduling trends. Data was grouped by appointment month, and total records per month were calculated.

The national category dataset was used to show appointments per month for various categories. Lineplots were created to visualize service settings for each season.

I then started exploring NHS-related Twitter for further data analysis which involved loading and exploring the dataset. Value\_counts() was used to calculate retweets and favourites. Hashtags related to the NHS were extracted and counted.

Recommendations were made after filtering data to August 2021 and onwards.

Appointments regional data was aggregated by appointment month, and utilization was calculated by dividing appointments per month by 30. Seaborn was used to visualize appointment numbers and utilization each month. Similar analysis was done for healthcare professionals, appointment statuses, and time between booking and appointment. Bar plots were used to compare service settings, excluding general practices for better understanding.

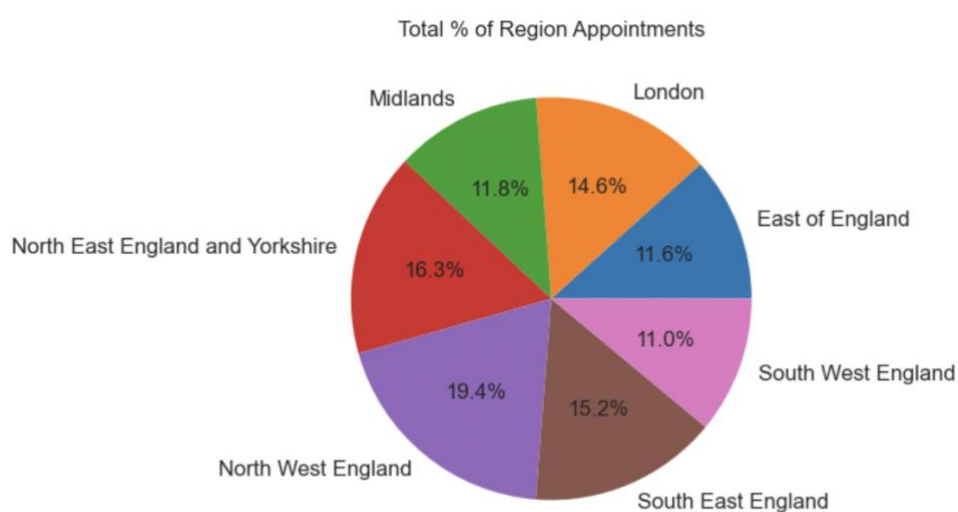
Markdown in Jupyter was utilized to document insights, headers, and summarize findings and recommendations.

The approach followed a structured sequence, starting with data preparation and sense-checking. Data pre-processing ensuring data quality, and data transformation enabled meaningful analysis. Visualization aided in identifying trends. Finally, the findings were presented through recommendations and summarized in Jupyter using Markdown.

## Visualisation and Insights:

### Locations:

Through my analysis the first insights gathered were there are 7 regions, further divided into 42 ICB-locations and further divided into 106 sub-ICB-locations. Based on the “actual\_duration” dataset, regions with the highest number of appointments from December 2021 to 30<sup>th</sup> of June 2022 are Northwest England with 32,574,555 appointments in 7 months (19.4% of total), followed by North East England and Yorkshire (16.3%) and South East England (15.2%).



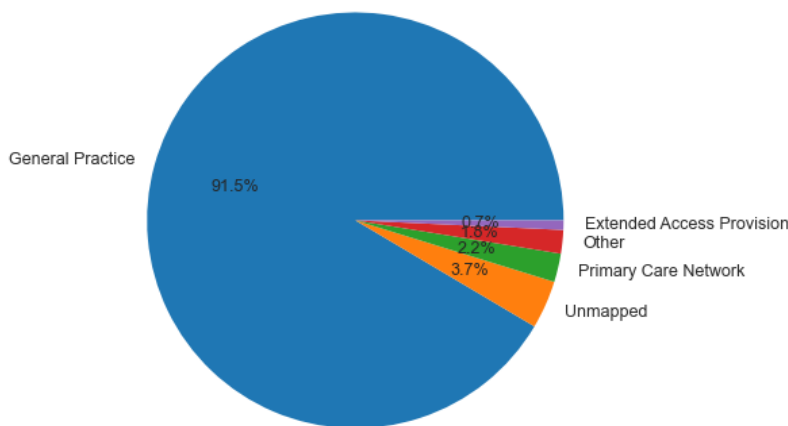
Based on national\_categories dataset sub-ICB locations with the highest number of appointments are Northwest London, Kent and Medway, Devon, Hampshire and Isle of Wight and North East London with appointment numbers ranging approximately from 7,000,000 to 12,000,000 appointments in 11 months.

Popular sub-divisions in various categories:

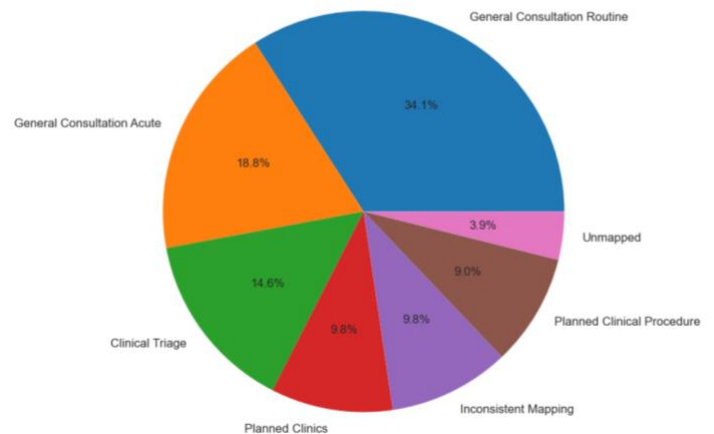
(Pie charts were used to show percentage ratio of each category)

From our analysis we've discovered that there are 5 different service settings, with "General practices" making up 91.5% of total number of appointments, 18 national categories but with 6 categories making up less than 3% of total appointments therefore were removed as outliers, with general consultations, routine (34.1%) and acute (18.8%) being most popular. 86.7% of three context types were care-related encounters and lastly with 91.3% of three appointment statuses fall under "Attended".

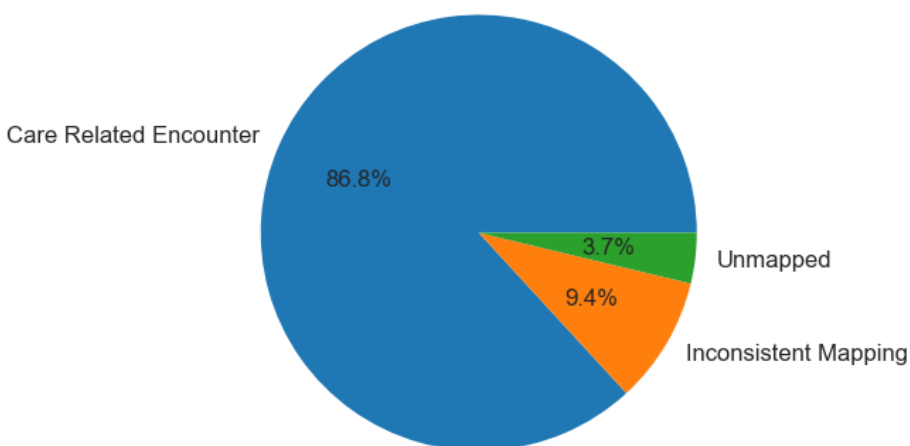
Total % of Appointments in Service Settings



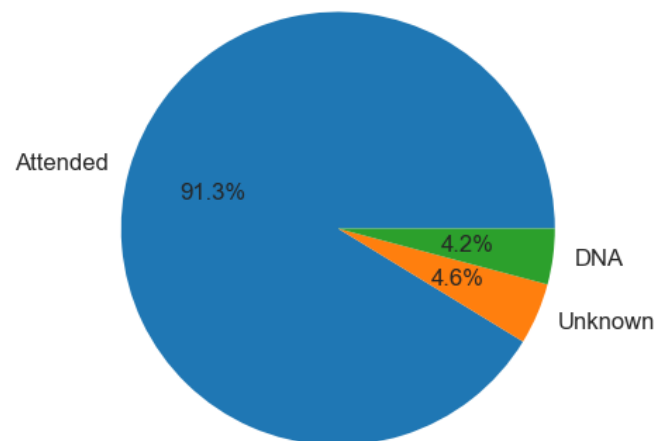
Total % of National Categories



Total % of Appointments Context Type



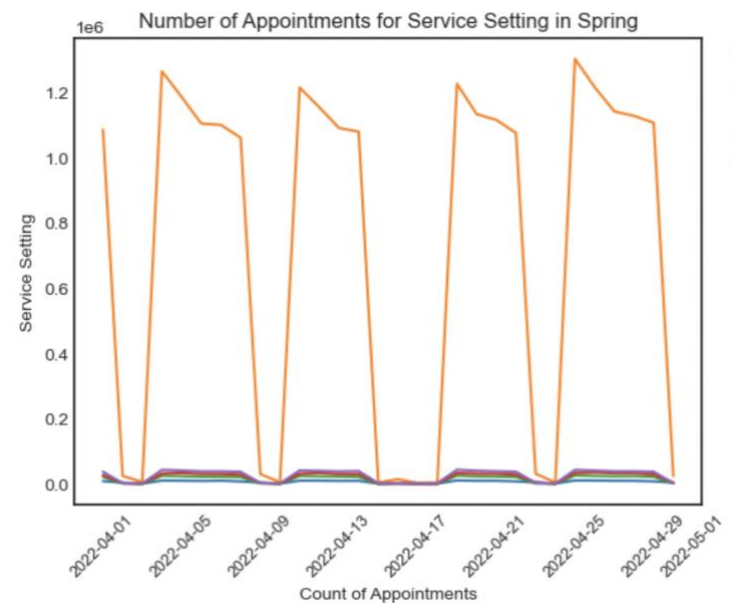
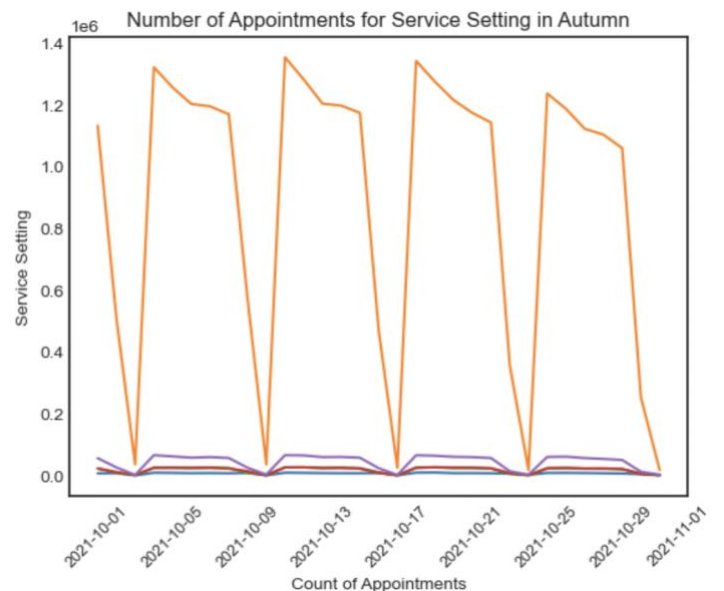
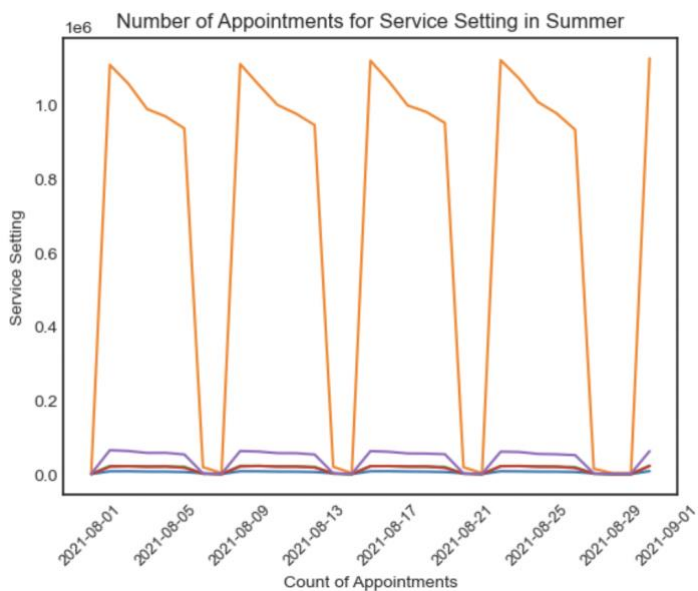
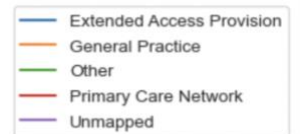
Total % of Appointment Status



Time-series data:

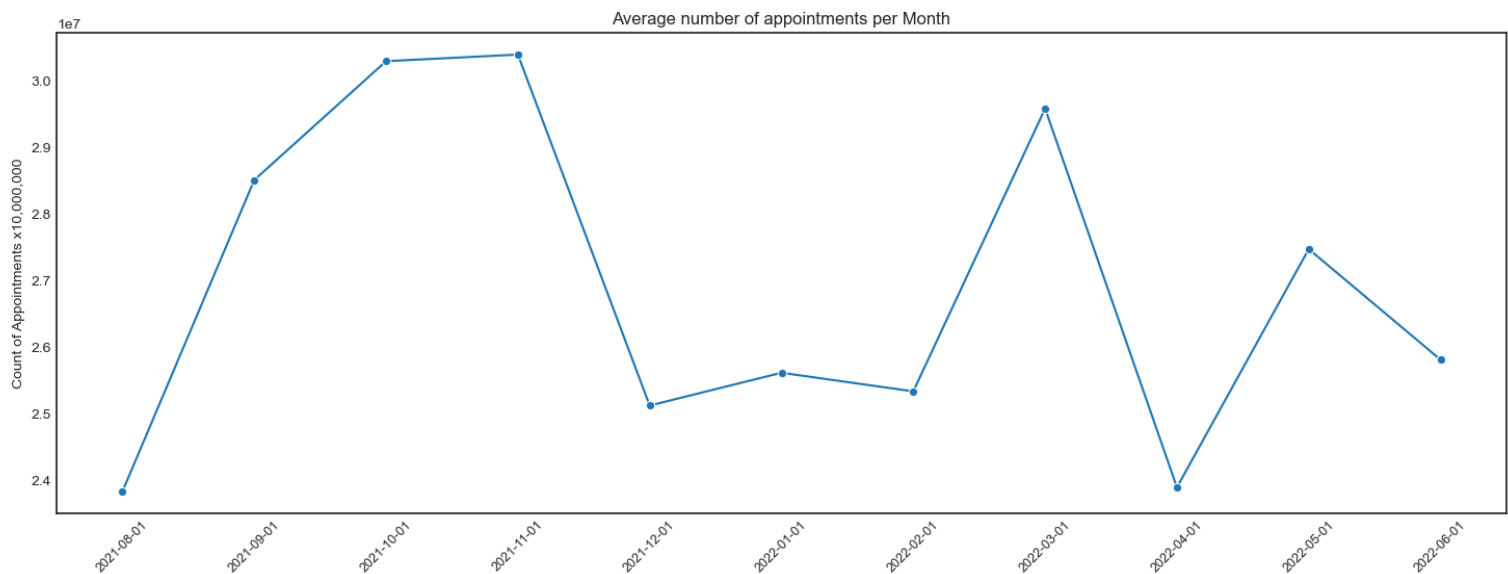
(Lineplots were used to show changes in appointments over time based on different criteria.)

Based on visualisations when looking at a 2021/2022 calendar we can see that days with the most GP appointments fall on Mondays, gradually decreasing through out the week. Appointments = 0 corresponds with the weekends. Autumn has the most GP appointments (sometimes almost up to 1.4 million which is way more than NHS capacity which explains why Sundays were the only days with no appointments on October to meet demands).



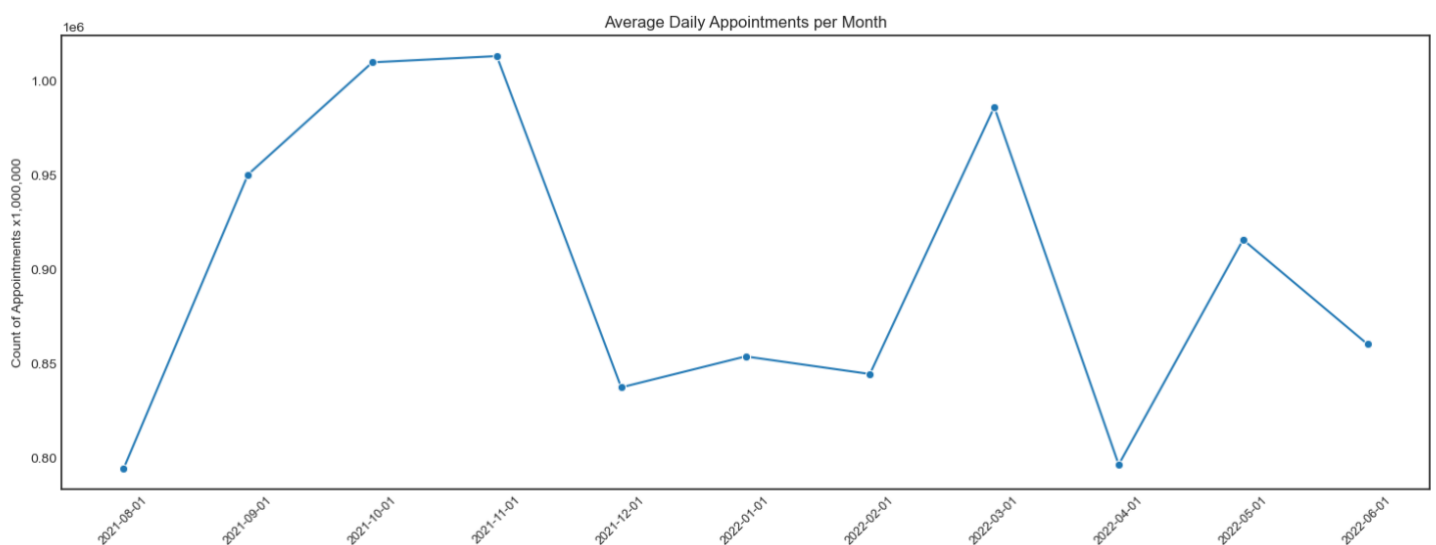


## Capacity and Utilisation:

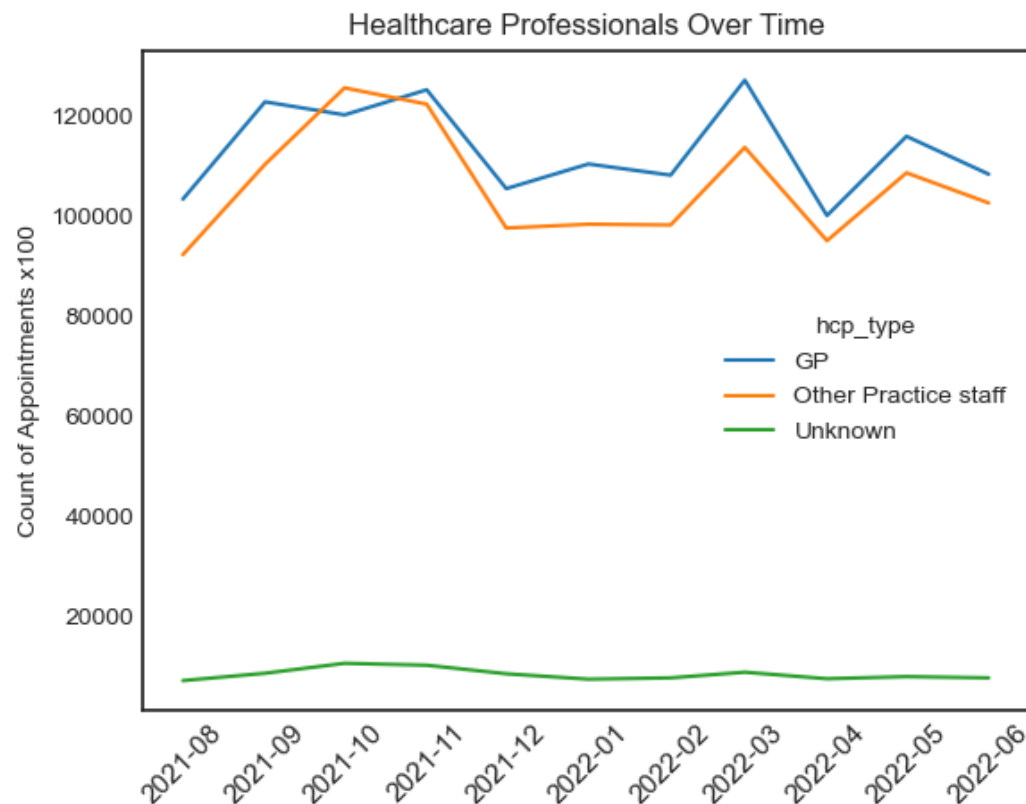


According to “actual\_duration” dataset we can determine that from August 2021 to June 2022, the months with the least appointments were August 2021 and April 2022 with approximately 24 million appointments. The months with the highest number of appointments were October and November 2022 with approximately 30 million appointments.

By dividing the count of appointments in previous graph by 30 we can calculate the average daily appointments to determine the actual utilisation of resources, based on the data the NHS has capacity as NHS can accommodate up to 1.2 million a day.

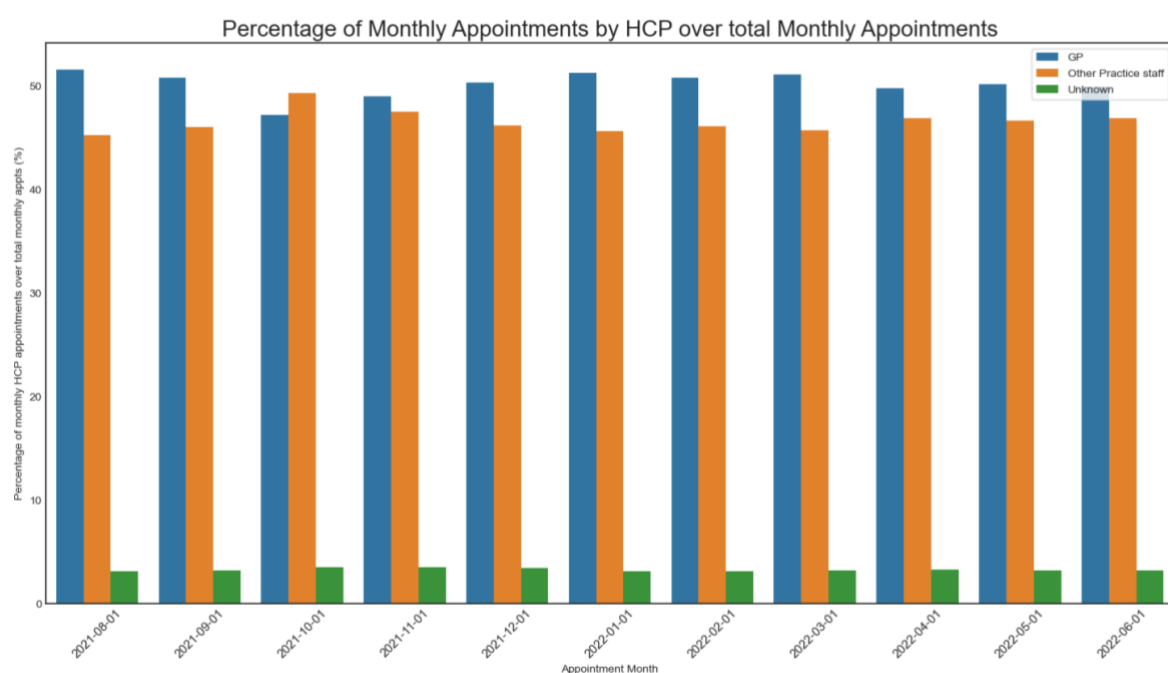


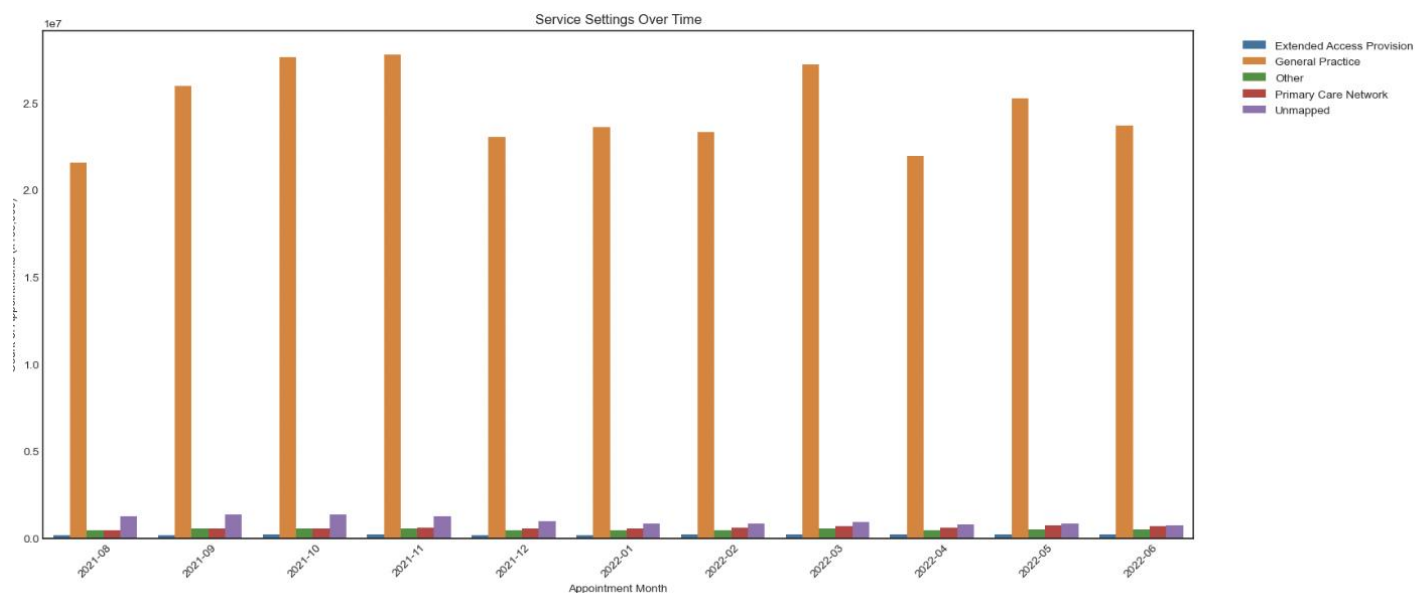
## Healthcare Professionals:



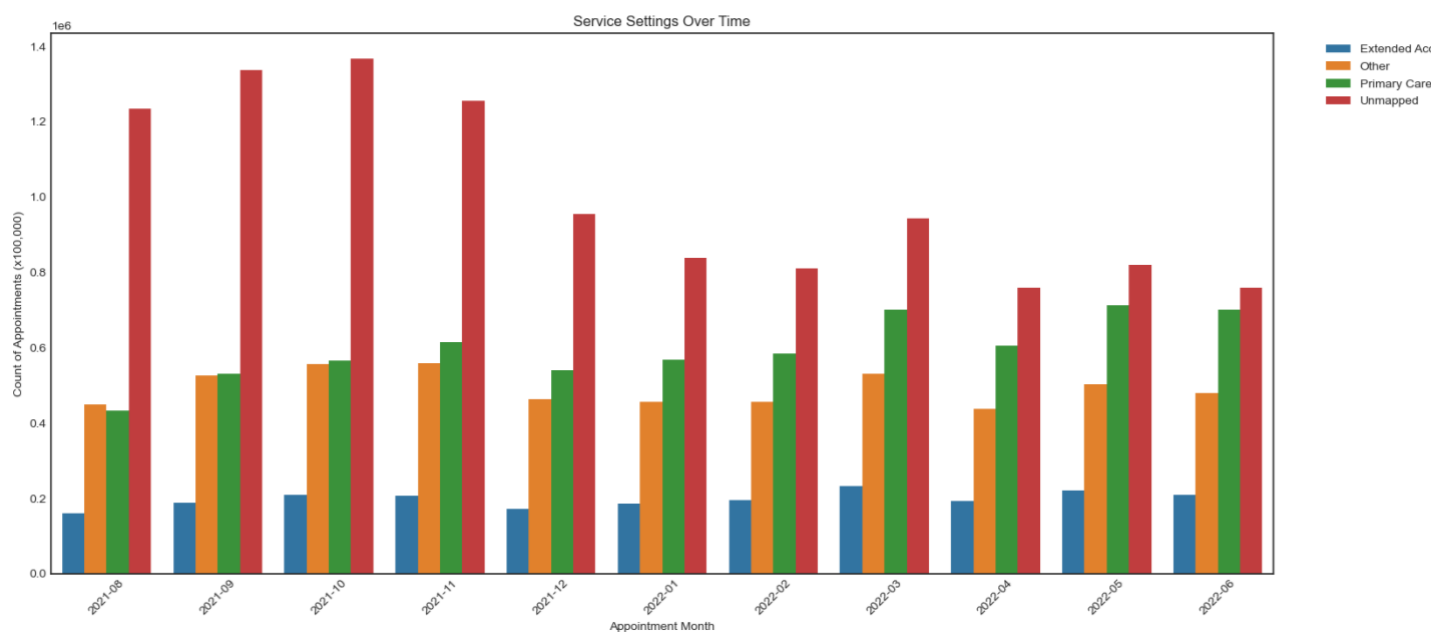
**GPs** make up most of the appointments ranging from around **46-51%** of total healthcare professionals, with **other practice staff** ranging from **45-49%** of total HCP and **Unknown** falling **under 4%** throughout therefore can be considered as **outliers**. **October 2021** is the only month where other practice staff (14.3 million appointments) were more than GPs (14.9 million appointments) by around 2%.

(Bar graphs were used to compare percentage of appointments with each hcp)





As established “General Practices” make up most of appointment settings, directly correlating to changes in number of appointments per month ranging from 2.3 to 2.8 million appointments per month.



Unmapped appointments range from around 730,000 to 1.37 million appointments per month, also corresponding to changes in total number of appointments per month. Number of Primary-care increasing over time.

### *Patterns and predictions discovered through analysis:*

- In "actual\_duration" appointments have been scheduled from 01/12/2021 and 30/06/2022 (7 months) with a total of 167,980,692 appointments.
- In "national\_categories" appointments have been scheduled from 01/08/2021 and 30/06/2022 (11 months), with a total of 296,046,770 appointments.
- In "appointments\_region" dataset where appointments have been scheduled from 01/01/2020 and 30/06/2022 (2 years 5 months), with a total 742,529,759 appointments.
- October 2021, November 2021, and March 2022 are the busiest months, with all making up around 4% of total appointments in "appointments\_region" dataset
- Number of appointments in general practices (service\_setting), care related encounters (context\_type) and general consultations, acute and routine (national\_category) are all directly proportional to changes in count of appointments per month.

### *Recommendations:*

- Ensure more staffing over start of the week to help meet increasing demands.
- Increase staffing over Autumn and Winter by offering temporary roles through locuming and overtime shifts over weekend to help meet demands.
- Ensure clear guidelines and policies on data documentation to decrease number of undocumented data.
- Increase telephone appointments to decrease strain on face-to-face appointments.
- Investigate fluctuations in appointments from February to June 2022.