LSE_DA201 Python for Data Analysis: Final Assignment

NHS - Appointments in General Practice

Ryan Lin

Word Count: 1164

Introduction & Background

This report and the resulting findings are derived from three data sets published by NHS England includes information about scheduled activity and usage of general practice (GP) appointments in the UK within primary care.

Dataset	Dataset Contents
Table AD actual_duration.csv	**Details of appointments made by patients**: • Regional information, date, duration, and number of appointments pertaining to a certain class.
Table AR appointments_regional.csv	**Details on the type of appointments made by patients:** • Regional information, the month of appointment, appointment status, healthcare professional (HVP Type), appointment mode, the time between booking and appointment, as well as the number of appointments pertaining to a certain class.
Table NC national_categories.csv	**Details of the national categories of appointments made by patients:** • Regional information, date of appointment, service setting, context type, national category, and the number of appointments pertaining to a certain class.
Tweets.csv	Details of twitter data (tweets) related to healthcare in the UK scraped from Twitter.

Note: See meta data file titled metadata_nhs.txt for more information

There is a significant financial and social cost when patients miss their GP appointments. Due to this, The NHS are keen to develop a data-driven approach to determine how to reduce or eliminate missed appointments (business problem) by exploring two main areas:

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

A variety of questions needs to be asked to better understand the project and the data provided. Additional areas to explore outside the scope of this project are also listed (next page).

Questions Related to This Project

- Who are the key decision makers and main internal & external stakeholders involved? Department of Health and Social Care, Secretary of State, ICB Chairs, other involved MP's?
- Who will the finding and insights be presented and shared with?
- What does success look like from what we can determine regarding this analysis?
- Who will be implementing the recommendations based of the resulting finding?
- What is the reason for trying to establish a relationship between missed appointments and capacity or resource utilisation?

Questions Related to The Data

- Why are we using this data to determine capacity and resource utilisation when you (the NHS has stated) "the data does not give a complete view of GP activity so should not be used to infer a view of workload" and "No information is included on capacity (the proportion of available appointments that are used) of appointments in general practice."
- Why are there only 1174 tweets in the tweets.csv file. Specifically, how was this tweet data collected and in what time period was it collected?

Additional Areas to Explore

- Exploring capacity properly by measuring and analysing demand (volume of people attempting to book appointments) and utilisation (proportion of available appointments that are used).

Analytical Approach

Note: The subheadings contained in the **Analytical Approach** and **Visualisations and Insights** chapters below correspond to those found in the associated Jupyter Notebook.

1. Setting up GitHub Repository

A GitHub Repository was set up to host the Jupyter Notebook files and related documents used to complete the analysis.

The various libraries utilised for this analysis are outlined below:

Library	Function	Benefits
Pandas	Powerful Python Library for data manipulation and analysis which provides data structures and functions to work efficiently with structured data. Widely used due to its versatility, simplicity and performance.	Data Import, Data Cleaning, Data Manipulation, Data preparation for Visualization, Statistical Analysis, Time Series Analysis, Integration with Numpy, Performance.
Numpy (Numerical Python)	Essential Python Library for numerical computations and handling multi dimensional arrays.	Used alongside Pandas and other libraries to perform numerical computations, manipulate data and handle multi-dimensional arras
Natural Language Toolkit (NLKT)	Designed for natural language processing (NLP) and text analysis. It provides a wide range of tools and resources for working with human language data.	Useful to process and analyse text from tweet data related to healthcare, extract meaningful information and insights from textual aspects of the data and perform sentiment analysis.
Matplotlib & Seaborn	Used for data visualisation. Each provide a wide range of static, interactive high quality plots, charts and graphs	Used to create informative and visually appealing visualisations to help convey trends and insights and communicate finding effectively.

2. Importing and Exploring Data

The provided semi-wrangled data was then imported and sense-checked, where the column names df.columns(), data types df.dtypes(), number of missing values df.isna(), metadata df.info() and descriptive statistics df.describe() were determined for each data set (See Section 2):

actual_duration.csv as ad

sub_icb_location_code	object
sub_icb_location_ons_code	object
sub_icb_location_name	object
icb_ons_code	object
region_ons_code	object
appointment_date	object
actual_duration	object
count_of_appointments	int64

dtype: object

- 137793 rows × 8 columns
- no rows with NaN in 8 columns

appointments_regional.csv as ar

icb_ons_code	object
appointment_month	object
appointment_status	object
hcp_type	object
appointment_mode	object
time_between_book_and_appointment	object
count_of_appointments	int64

dtype: object

- 596821 rows × 7 columns
- no rows with NaN in 7 columns

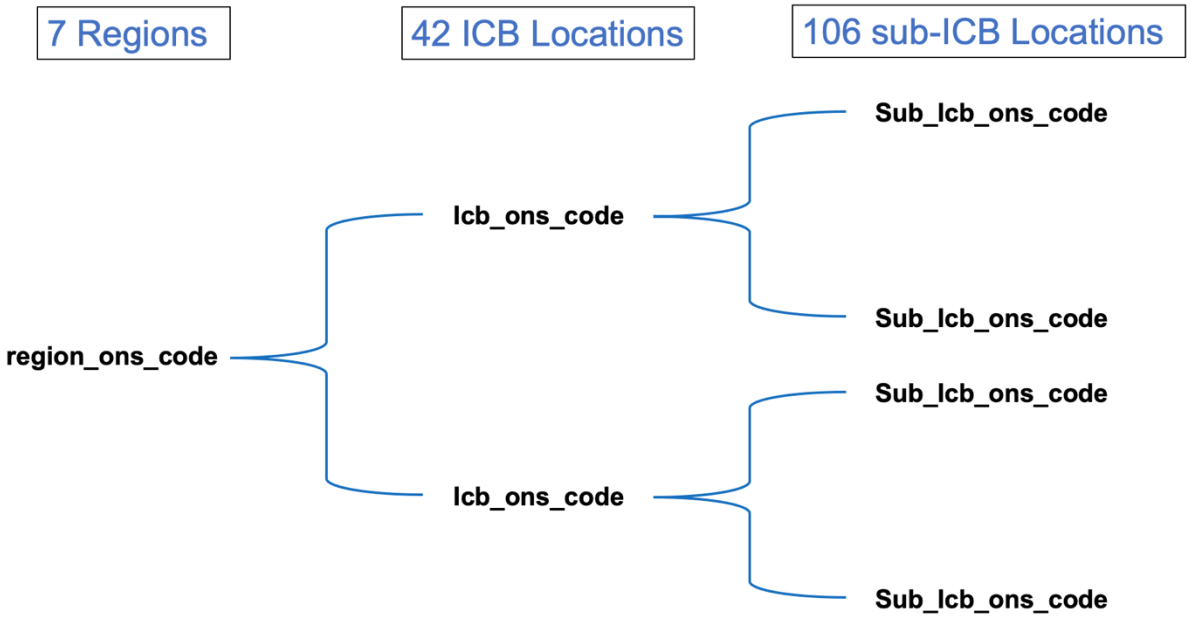
national_categories.xlsx as nc

appointment_date	datetime64[ns]
icb_ons_code	object
sub_icb_location_name	object
service_setting	object
context_type	object
national_category	object
count_of_appointments	int64
appointment_month	object

dtype: object

- 817394 rows × 8 columns
- no rows with NaN in 8 columns

From the results, none of the data sets contained missing values. The presence of potential duplicates was also checked with df.duplicated() and outlined in section 2.1.2. Additionally, through isolating and checking for duplicates in the region_ons_code, icb_ons_code, and sub_icb_location_ons_code columns of the ad DataFrame, the taxonomy of the data classification codes was confirmed to be:



No duplicate entries were found in the ad and nc DataFrames when checking for all columns and when excluding the 'count_of_appointments' column. For the ar DataFrame, 21,604 (3.6%) rows were shown to be duplicates. However, upon further inspection of the original Appointment_GP_Metadata file provided by the NHS, it seems the sub_icb classification code columns were removed from the ar dataset. As a result, one can assume that the duplicate rows are datapoints at the sub_icb level and were not removed (See 2.2.2 for detailed explanation).

Continuing the initial phase of data exploration (Section 2.4), various characteristics of the data were determined utilising pd.unique():

Service Settings 5	Context types 3	National Categories 18	Appointments Status 3
General Practice	Care Related Encounter	General Consultation Routine	Attended
Primary Care Network	Unmapped	General Consultation Acute	Did Not Attend (DNA)
Other	Inconsistent Mapping	Planned Clinics	Unknown
Unmapped		Home Visit	
Extended Access Provision		Unplanned Clinical Activity …	

Number of Sub ICB Locations	Note
106	The number of Sub-ICB locations determined in section 2.1.2 above is validated here by determining the count of unique values in the 'sub_icb_location_name' column of the nc Data Frame.

Outlier analysis was disregarded because when the categorical values were plotted with boxplots, it became evident that the data naturally leads to outliers with high appointment counts with low numbers. This is likely a result of the low number of appointments from the weekend as well as the disparity between the number of appointments in urban cities compared to remote placed in the UK (see 2.5 for more information).

3. Analysing the Data

To better understand the timeline of the data sets and how they relate to one another, the minimum and maximum appointment date was determined for each dataset:

Data Set	Minimum Date	Maximum Date
ad: Actual Duration	2021 – 12 – 01	2022 – 06 – 30
ar: Appointment Regional	2020 – 01 – 01	2022 – 06 – 01
nc: National Categories	2021 – 08 – 01	2022 – 06 – 30

The date format in each of the datasets were formatted to a datetime format to facilitate data manipulation, sorting, time based aggregations, plotting and data filtering.

Across all 3 data sets the overlapping date range is between the latest minimum date and the earliest maximum date:

- **2021-12-01 - 2022-06-30**

*As the dataset ar only contains appointment month, assume latest date as 2022-06-30

As the nc dataset contained most of the parameters explored, the timeframe of 2021-08-01 to 2022-06-30 was followed as it allowed for the most overlap between the datasets, accepting that only had data for actual duration (from the ad dataset) from 2021-12-01 onwards.

Much of the analysis from here is done using df.groupby() to group the data by month for time-series analysis to determine trends and patterns. The table outputs

below show the months with the highest number of appointments and the total number of records per months.

Number of Appointments per Month

appointment_month	count_of_appointments
2021-11	30405070
2021-10	30303834
2022-03	29595038
2021-09	28522501
2022-05	27495508
2022-06	25828078
2022-01	25635474
2022-02	25355260
2021-12	25140776
2022-04	23913060
2021-08	23852171

Average Number of Appointment a Month = **26,913,343 appointments**

Number of Records per Month

appointment_month	number_of_records
2021-08	69999
2021-09	74922
2021-10	74078
2021-11	77652
2021-12	72651
2022-01	71896
2022-02	71769
2022-03	82822
2022-04	70012
2022-05	77425
2022-06	74168

Average Number of records a Month = **74,309 records**

4. Sentiment Analysis: NHS-related Twitter Data

The proposed objective to analyse the top trending hashtags in the UK related to healthcare was deemed of little value to the overall business objective (See 4.1 for more information). Instead, a sentiment analysis was conducted using the NLTK Library using VADER (See 4.3), which provides a sentiment score between -1 (negative) and 1 (positive) for each tweet.

	tweet_full_text	sentiment_score
0	As Arkansas' first Comprehensive Stroke Certified Center, UAMS provides Arkansans with access to the most advanced stoke care. Join us in our mission to make a difference in the health and well-be...	0.8384
1	RT @AndreaGrammer: Work-life balance is at the foundation of how decisions are made and where #PremiseHealth is headed. We're #hiring for…	0.0000
2	RT @OntarioGreens: $10 billion can go a long way to fixing our broken #Healthcare system.\n\nYet Doug Ford would rather spend it ALL on a hig…	-0.4767
3	RT @modrnhealthcr: 🚨#NEW:🚨 Insurance companies are figuring out the best ways to collect information about members' race and ethnicity data…	0.6369
4	ICYMI: Our recent blogs on Cybersecurity in Accounting https://t.co/4nnK0FiVVL and Digital Transformation in Healthcare Finance https://t.co/jlqn52lHD3 are a great read, take a look!\n\n#blogs #di...	0.6588
...	...	...
1169	RT @PotomacPhotonic: Potomac #Innovation Report: #precisionFabrication techniques Optimize #Microfluidic Mixing of Viscous Fluids \n\n#manuf…	0.4939
1170	Not a cent towards workers who would like to advance their training, especially those already employed by SHA or who for various reasons cannot obtain a student loan. Half of our department applie...	0.3612
1171	The @hfmaorg Region 9 presents "The Value of ESG to the Healthcare Industry" and our own Kris Russell and Ron Present will be the key speakers. This #webinar will be taking place 9/13 and will exp...	0.4939
1172	Happy physiotherapy 🩺 day 🎉..\n#bpt #physiotherapy \n#HealthyNation #healthcare \n#medicalcare \n#csjmu .\n@WHO \n@MoHFW_INDIA \n@nitish_0210 https://t.co/NQHdloYymC	0.5719
1173	RT @KimcoStaffing: Apply now to work for #MediQuestStaffing as EVS - #Hospital - 1st #shift - Interviewing Now!! (#NewportBeach) #job http…	0.0000

1174 rows × 2 columns

```
# Calculate the average sentiment score
average_sentiment = tweets_text['sentiment_score'].mean()

# Print the average sentiment score
print("Average Sentiment Score:", average_sentiment.round(2))
```

```
Average Sentiment Score: 0.25
```

The overall sentiment score (average) was calculated to be 0.25, indicating a slight overall positive sentiment, but still relatively close to neutral.

Visualisation and Insights

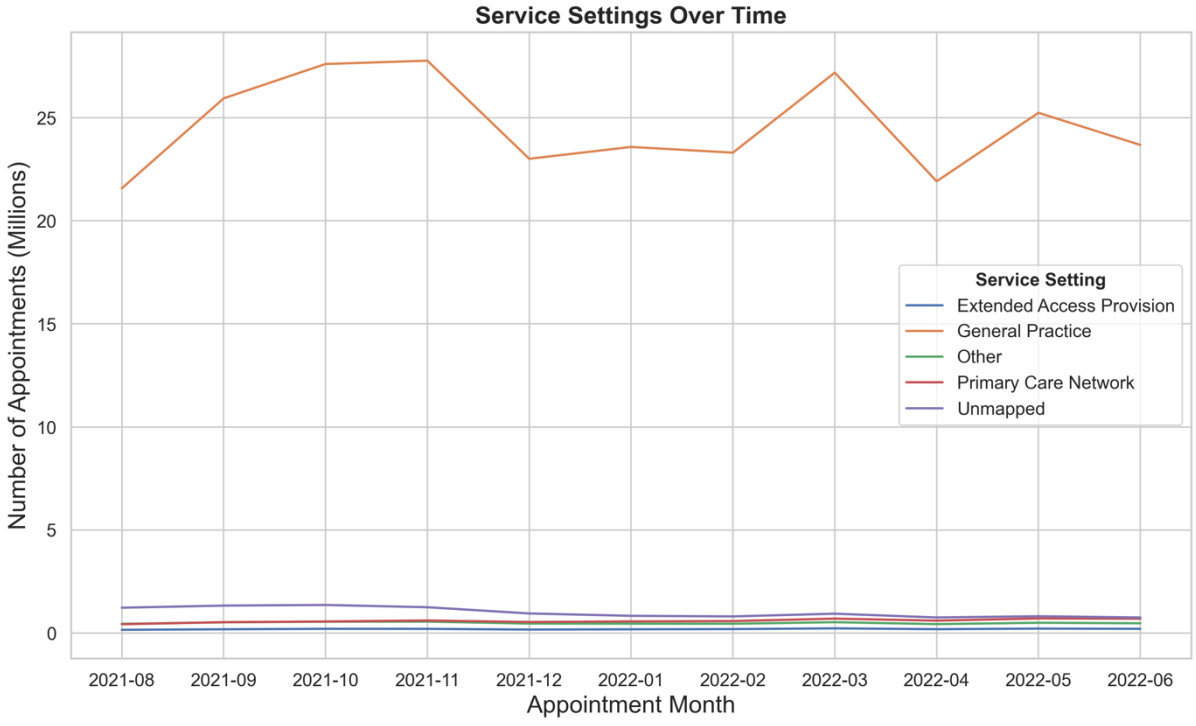
5. Visualising and Identifying Initial Trends

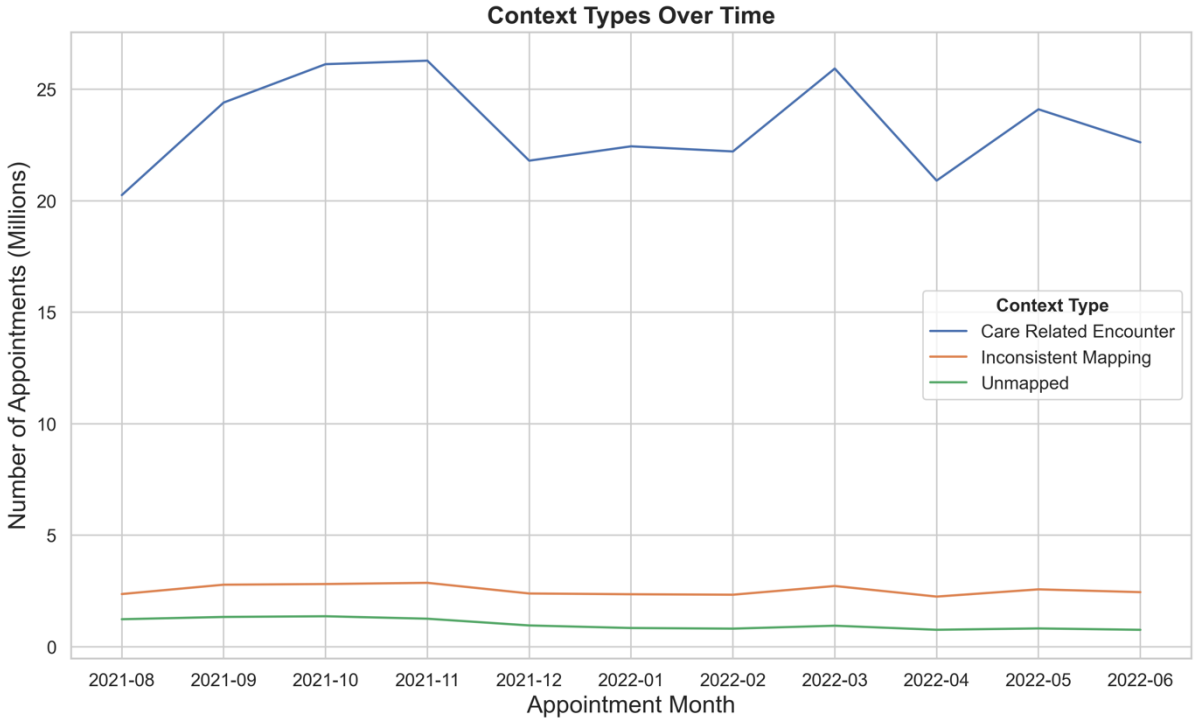
Initial visualisations were created using Matplotlib and Seaborn to determine monthly and seasonal trends of the number of appointments, service settings, context types and national categories. The nc dataset was aggregated on a monthly level to determine the number of appointments per month (section 5.1). Line plots were primarily used as they best represent sequential time-based data clearly, making it easy to identify and compare trends.

For three categories, the data is heavily biased to the highest occurring values:

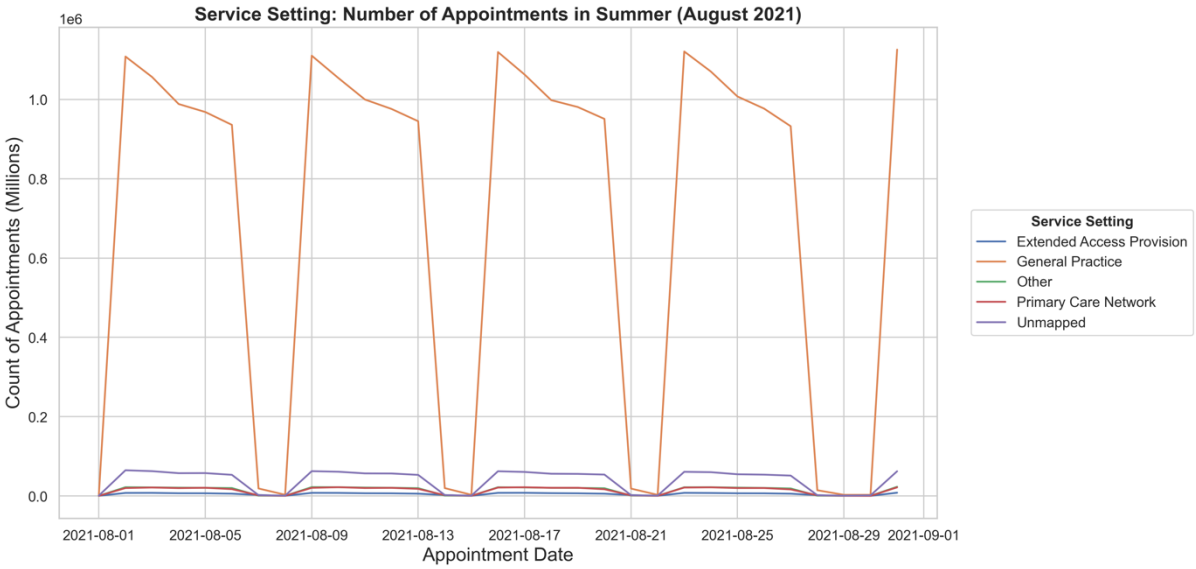
Service Settings	%	Context types	%	National Categories 18	%	Appointments Statuses	%
General Practice	91.5	Care Related Encounter	86.8	General Consultation Routine	32.9	Attended	91.2
Unmapped	3.7	Inconsistent Mapping	9.4	General Consultation Acute	18.1	Unknown	4.6
Primary Care Network	2.8	Unmapped	3.7	Clinical Triage	14	DNA	4.2
Other	1.8			Planned Clinics	9.5		
Extended Access Provision	0.7			Planned Clinical Procedure …	8.7		

This it to be expected as the NHS primarily delivers care through General Practice appointments and direct patient care. All three (generally the highest weighted values) follow a similar trend of an increase in appointments in the winter months building up towards the end of the year with a sharp decline in January which then plateaus.





The seasonal variation in the number of appointments for service settings is in line with above. The seasonal figures in section 5.2 show a weekly cycle in the number of appointments for the different Service Settings, with only a small number of appointments on weekends and lower appointment number on public holidays (Summer Example below, See 5.2.4 for more details)



Patterns and Predictions

6. Findings and Recommendations

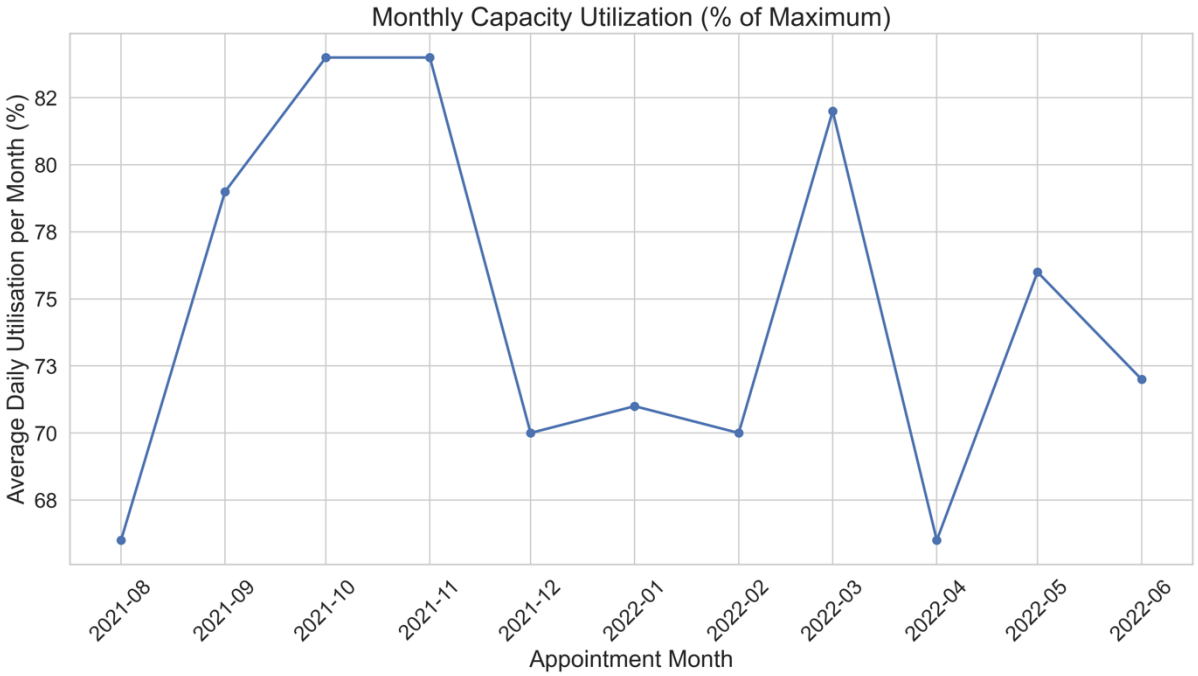
Here, we dive deeper into changes relating to the number of appointments per month to determine attendance and utilisation for HCP Types, Appointment Attendance, Appointment Type and Time between Booking an Appointment. An

attempt is also made to explore the answers to the questions below, relevant findings are discussed.

1. Should the NHS start looking at increasing staff levels?
2. How do healthcare professional differ over time?
3. Are there significant changes in whether or not visits are attended?
4. Are there changes in terms of appointment type and the busiest months?
5. Are there any trends in time between booking and appointment?
6. How do the various service settings compare?

Note: even in consideration of the limitations mentioned previously, to try an gauge capacity, it is assumed that the NHS can accommodate a maximum of 1,200,000 appointments per day (source unverified).

The average utilisation rate for the time period 2021-08 – 2022-06 is shown below (see 6.1 for details), the aggregated sum of the count of appointments per month was divided by 30 to get the average daily value, which was divided by 1,200,000 to get the average utilisation rate for the time period in question.

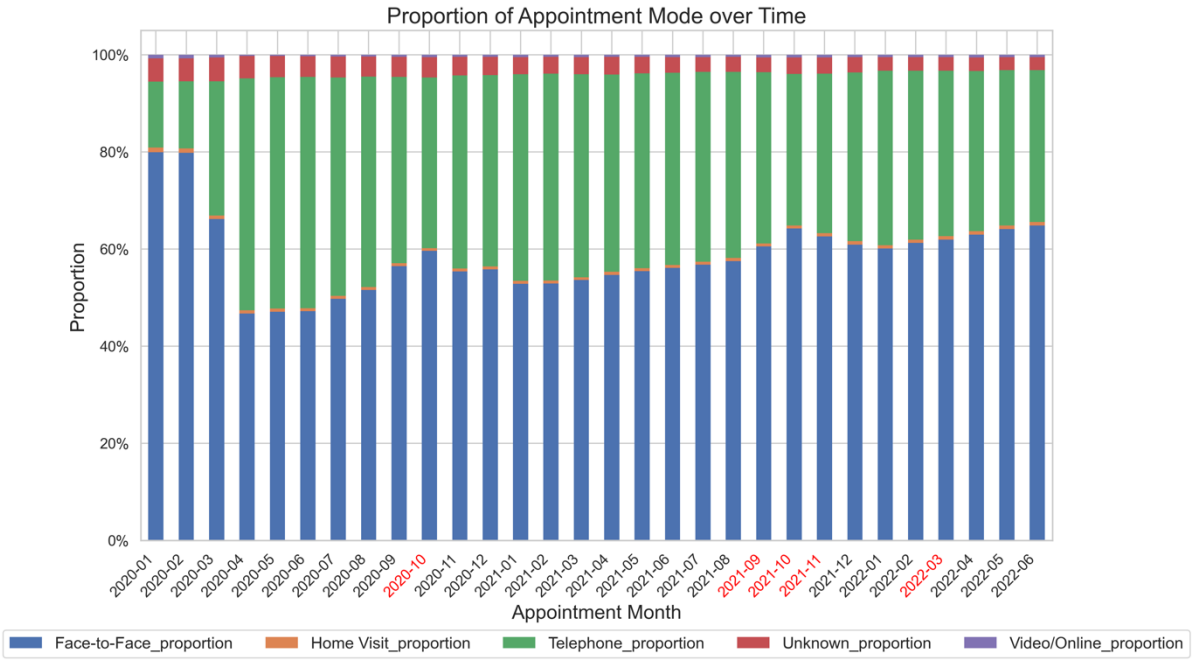


As seen in the figure the monthly utilisation capacity increases substantially from 66% in August 2021 to 84% in November 2021 and then drops and hovers around 70% until February 2022. This is in line with previous findings due to increased rates of flu and end-of-year winter bugs appointments (NHS, 2022).

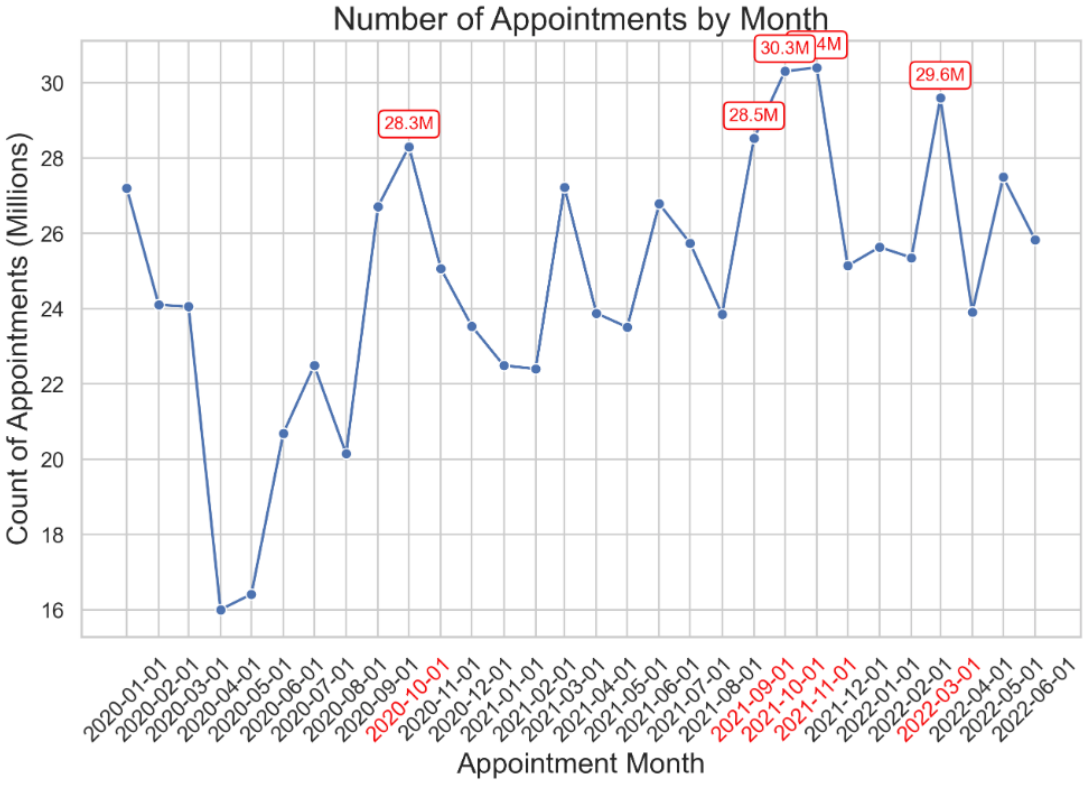
As we are primarily investigating the change in the number of appointments over time of different variables, their values will be affected by the count of appointments so where possible a percentage or a proportion is used with corresponding stacked bar charts.

Changes in attendance follow a similar trend to the total number of appointments, with attendance increasing following Covid (NHS D. , 2022). There is also a trend

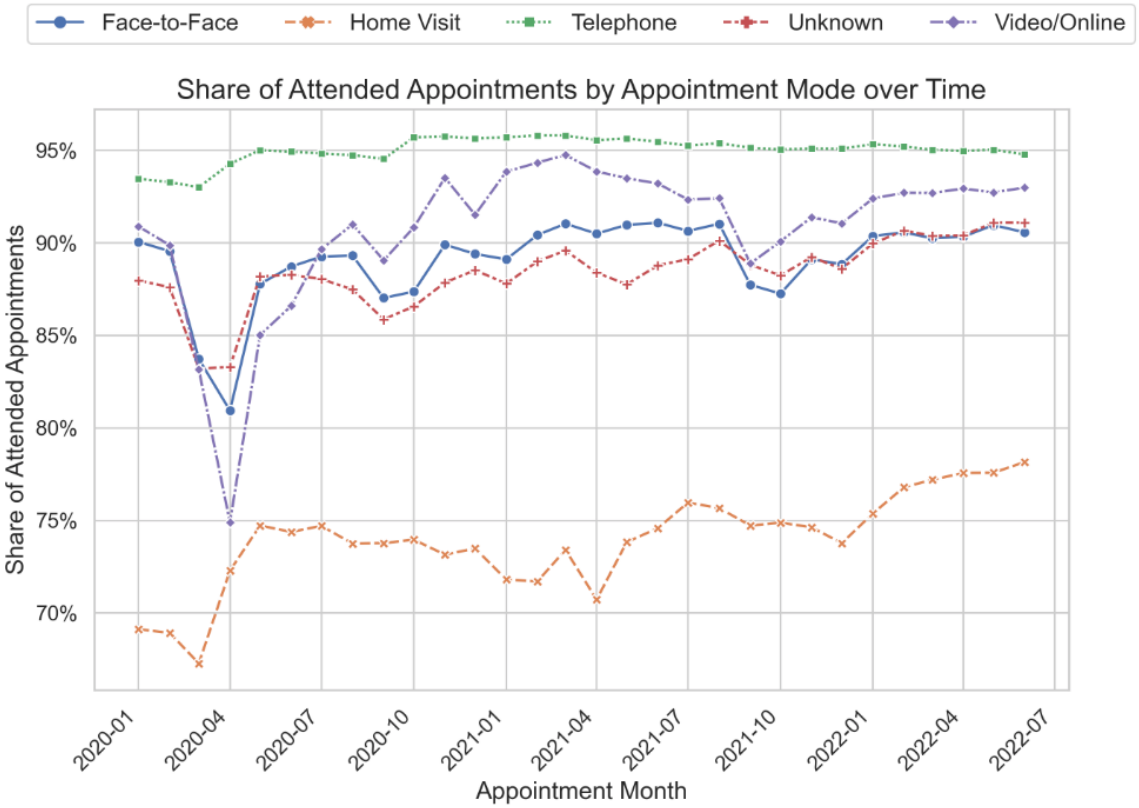
seen in a decrease in the proportion of appointments attended from July to October each year.

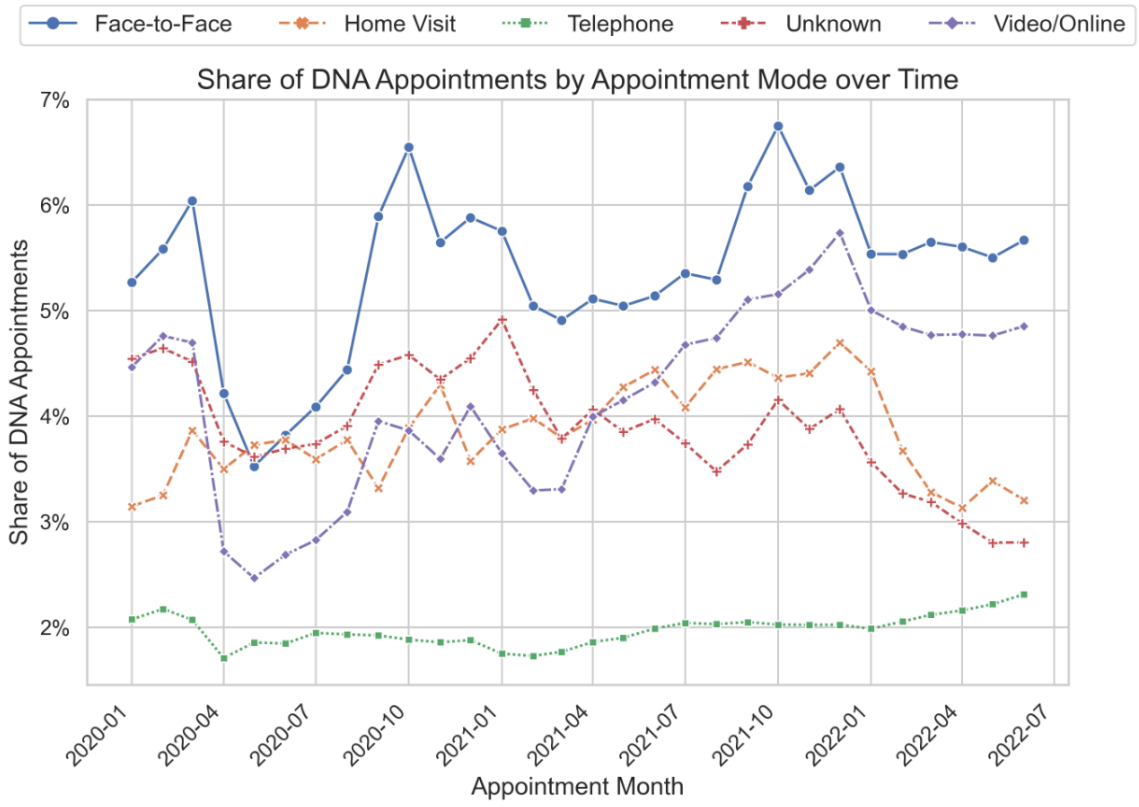


The stacked bar chart above indicates that before the pandemic face-to-face appointment generally accounted for 80% which then dropped to 50% until July. Although they have gradually increased since, it's unlikely to return to pre-pandemic values due to the adoption of more telehealth and digital health options (NHS, Listening to digital health innovators report 2021, 2021). Compared to the busiest months (highlighted in red), little difference is seen in the proportion of appointment modes.

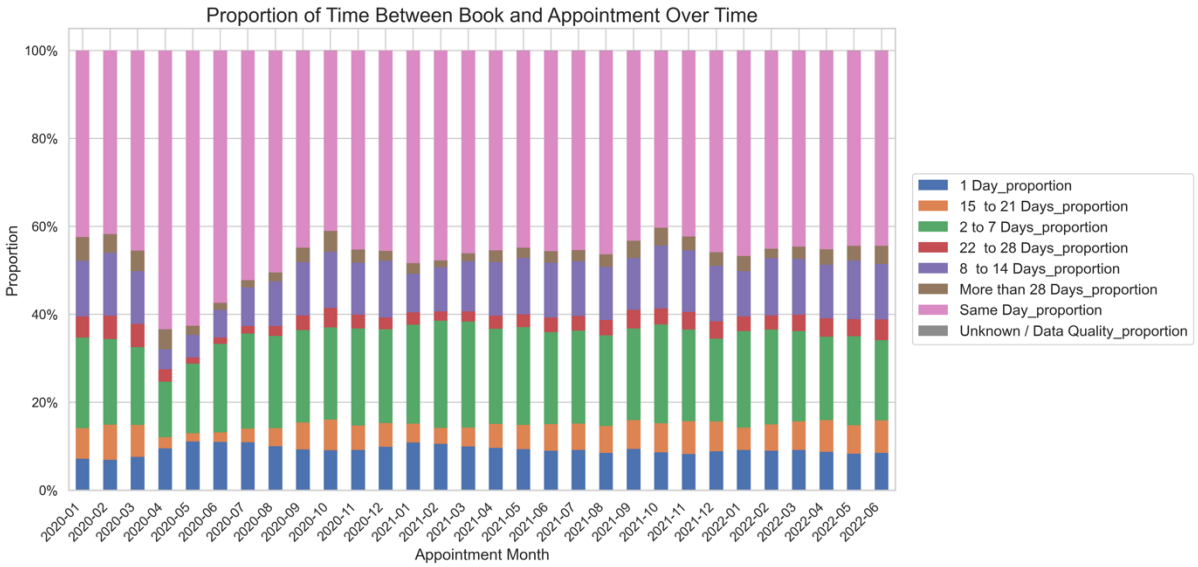


In terms of mode, the figure indicates the highest share of attended appointments as telephone, followed by video/ online. Interestingly, face-to-face had the highest share of DNA appointments.

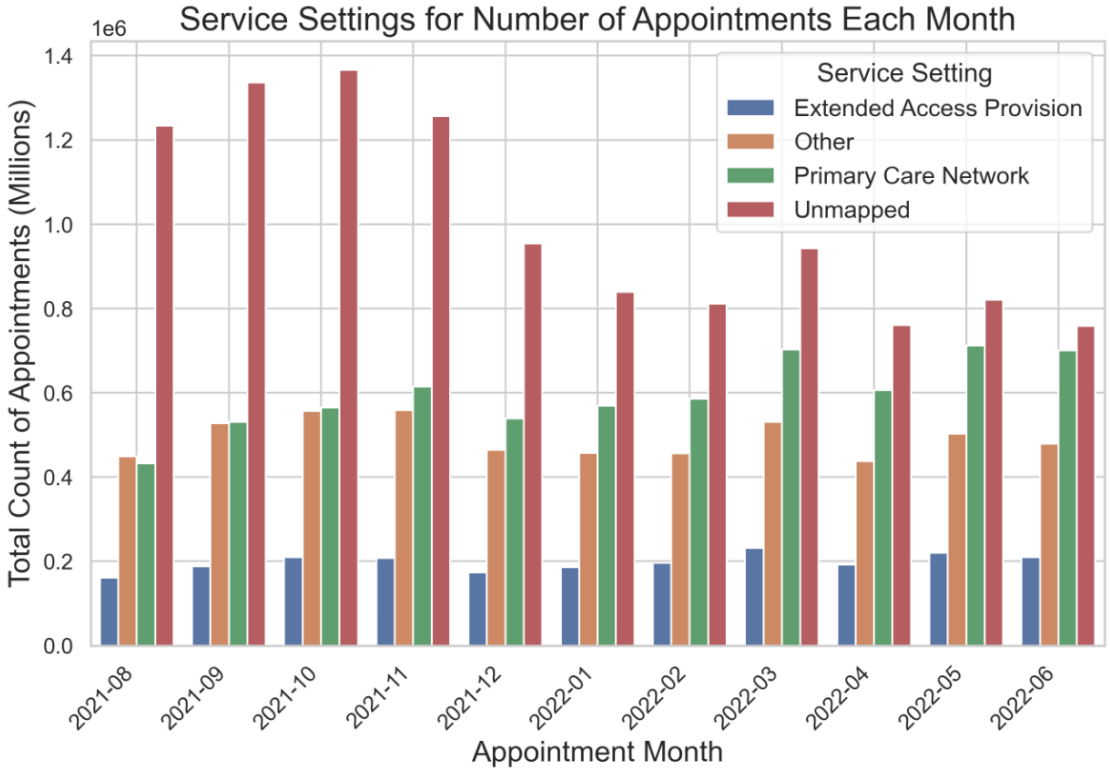




Trends in time between booking an appointment show a slight increase in same day appointments, and a decrease of 2 – 7 days and 8 to 14 days wait times from 2021-10 onwards. After the initial drop of appointments at the start of Covid, an increase in proportion of same day appointments is seen by around 20% which returned to pre-pandemic level relatively quickly from 2020-04 to 2020-10. In the more recent months, the longer wait time (see 22 – 28 days and 28+ days) have been increasing in line with NHS Data (NHS, 2021).



The service settings for GP follows similar trend to count of appointments. In the figure shown excluding GP, there is a decline of about 500,000 unmapped appointments, indicating success in the NHS's initiative to establish more consistent recording principles (BMA, 2020).



Finally, limitations of the existing data as well as recommendations for further areas to explore are listed below.

Data Limitations

- Unclear duplicate rows in ar database, and differences in date formats in each database makes it difficult to join the tables for deeper exploration if required.
- Without including more information and measurements that relate to capacity (such as number of GPs available, number of patients and the proportion of available appointments that are actually used), resource utilisation cannot be accurately measured.
- Same goes for demand (the volume of people attempting to book appointments), which will also be heavily affected by the demographics of patients registered at different Sub-ICBS/ practices which in turn will impact the nature of appointments required.
- Lists and untimed appointment vary between sub ICBs systems. These may appear in the appointment system as one appointment whereas other practices may record these as individual appointments throughout the day.

The NHS Specifically States:

- There is ”widespread variation in approach to appointment data management” which were made worse with COVID-19. Much of the data was “did not have universal and reporting standard and consultation activity is not reflected in GP appointment statistics collected and reported during this time.”
- The data “covers scheduled and planned activity recorded on the GP practice and PCN appointment systems only, rather than the totality of interactions or activity/workload.”
- “Not all GP activity, consultations or encounters are presented in this publication as only appointment slots captured in the practice or PCN appointment systems are collected.”
- “This information does not give a complete view of GP activity so should not be used to infer a view of workload.”
- **HCP Types:** The only HCP type currently collected with high enough consistency for publication is GP
- **Appointment Mode:** set locally by the practices so may not represent the actual care setting of the appointment. E.g. some telephone and video appointments are logged as face-to-face.
- **Time between booking and appointment:** unable to verify when contact was first made to request an appointment and “the data does not measure how long an individual may have waited to book an appointment (such as time spent telephoning the GP practice). As such this measure can not be used as a measure of demand by itself.”
- **Actual Duration:** This field is recorded differently depending on the practice’s system supplier. Any appointments with a null duration or a duration of less than one minute or greater than 60 minutes have been grouped into an ‘Unknown / Data Quality Issue’.

Recommended Future Actions and Areas to Explore

- As mentioned previously, more reliable metrics to measure demand and capacity.
- As one of the key problems identified are missed appointments, new measurements such as reason for missed appointment or whether a mail, text, telephone reminder was given could be implemented into sub ICB data management systems.
- If twitter data is analysed again in the future, a more useful and advanced analysis would be to use contextual semantic search (CSS) to derive more actionable insights from textual data.

References

BMA. (2020). *More accurate general practice appointment data – guidance.* London: NHS.

NHS. (2021, 08 20). *Listening to digital health innovators report 2021*. Retrieved from Transform England: https://transform.england.nhs.uk/about-us/get-involved/listening-to-digital-health-innovators-report-2021/

NHS. (2021, 08 12). *NHS cuts waiting times despite busy summer period*. Retrieved from NHS 75 England: https://www.england.nhs.uk/2021/08/nhs-cuts-waiting-times-despite-busy-summer-period/

NHS, D. (2022, 09 29). *Summary of outpatient appointments and attendances, 2011-12 - 2021-22*. Retrieved from Digital NHS: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2021-22/summary-report---attendances#summary-of-outpatient-appointments-and-attendances-2011-12-2021-22

NHS, L. (2022, 11 16). *Improving access to primary care this winter*. Retrieved from Sussex Health&Care: https://www.sussex.ics.nhs.uk/improving-access-to-primary-care-this-winter/)