The NHS data, which has been provided to us contains four main files:
- actual_duration.csv – containing details of appointment made by patients (ad)
- appointments_regional.csv – containing regional appointment details data (ar)
- national_categories.xlsx – containing national appointment details data (nc)
- tweets.csv – containing data mined from Twitter in relation to healthcare in the UK

Initial analysis of data:

All the appointment data has been initial analysed for properties and data quality. Data was loaded and python functions like "shape", "types", "describe()" have been employed to initially familiarise ourselves with the set. Data has been found to be of a good quality with no missing values (although in some categories entries like "Unknown" are present).

Data exploration:

National data has been analysed for number of records for each location (based on location codes and names) by employing "groupby" and "sort_values" functions. Number of records ranges between 1013 to 1484 for each location. For ease of further analysis, unique values for "appointment status", "national categories" and "context type" have been listed using "unique" Python function, e.g., below:

```
# Determine the number of appointment status.

print(ar.appointment_status.unique())
ar['appointment_status'].nunique()
```

```
['Attended' 'DNA' 'Unknown']
```

Further data exploration and initial NHS questions:

We have answered several NHS questions about observation data range, most popular service settings (for NHS North West London) and the months with highest numbers of appointments. To answer the questions several data wrangling techniques were employed. In particular:

- We had to be very careful with data format (either as datetime or object) and swap them as necessary.

```
# Creare new column displaying appointment month in text format
ad['appointment_date_month']=ad['appointment_date'].dt.strftime('%B %Y')
```

- Group by/ reset index and sorting functions were employed for aggregation purposes:

```
# Group by to display month with highest number of appointment
ad_month=ad.groupby('appointment_date_month')[['count_of_appointments']]
ad_month=ad_month.sum().sort_values(by=['count_of_appointments'], ascending=False).reset_index()
```

- Filtering was employed as well as renaming and within column number manipulation (for neat plots display)

- Option 1 for Filtering

```
# Filter the nc dataframe for North West London only
nc_NWLondon=nc[nc['sub_icb_location_name'].str.contains("NHS North West London")]
nc_NWLondon
```

- Option 2 for Filtering

```
# Create dataframe only for August 2021
nc_s_summer=nc_s_new[nc_s_new['appointment_month']=='2021-08']
```

- Data calculations within cells

```
nc_servsett_NW['count_of_appointments']=nc_servsett_NW['count_of_appointments']/1000000
nc_servsett_NW.rename(columns={'count_of_appointments':'count_of_appointments (mln)'}, inplace= True)
```

Analysing service settings, national categories context across the dates

Several techniques were employed to efficiently analysed 3 settings/categories across different months (and within months):

- We have used a line plot as the analysis was across various months/dates.
- We made sure x-axis labels are legible by limiting range or changing angle of the display.

```
# Set the limits for x - axis as without it - it wasn't legible
plt.xlim('2022-04-01', '2022-04-30')
```

```
plt.xticks(rotation=30)
```

- All graphs have been titled, legend formatted and put in a right place.

Twitter trends analysis

Twitter raw file was uploaded to Jupyter notebook and the hashtag extracted into list using loop through function. Subsequently occurrence of each hashtag was counted using Counter function from the collection module. This has created a dictionary which was subsequently converted into a DataFrame.

Hashtags were analysed for duplication and similar hashtags were grouped together. This included both different variation of the same word, e.g., “healthcare” and “healthcare,” as well summing different names for the same term, e.g., “covid” and “coronavirus”

Example of functions utilised:

```
#Use replace function for names describing the same term
data_rpl=data_rpl.replace(['#machinelearning','#ai','#digitalhealth','#tech','#biotech','#healthtech','#bigdata',\
                           '#ehealth','#innovation'],'#technology')
data_rpl=data_rpl.replace('#meded','#education')
data_rpl=data_rpl.replace(['#covid19','#coronavirus','#pandemic','#corona'],'#covid')
```

```
#Use lambda function to capture variation of the same term
data_rpl['word']=data_rpl['word'].apply(lambda x: '#technology' if 'digital' in x.lower() else x)
data_rpl['word']=data_rpl['word'].apply(lambda x: '#technology' if 'tech' in x.lower() else x)
data_rpl['word']=data_rpl['word'].apply(lambda x: '#marketing' if 'marketing' in x.lower() else x)
data_rpl['word']=data_rpl['word'].apply(lambda x: '#covid' if 'covid' in x.lower() else x)
```

Order was important in the replacement process. For example, both words ‘monkeypox’ and ‘vaccine’ occur twice in the same hashtag (see below). In this case it has been decided, that monkeypox as a key word is more important, and the replacement process for monkeypox should run first.

word	count
#vaccine	18
#vaccines	10
#coronavirusvaccines	6
#monkeypoxvaccine	4
#monkeypoxvaccine,Ä¶	1
#hpvvaccine	1

As word “healthcare” is used in multiple contexts, e.g., present in hashtags like #healthcarejobs, the aggregation by this word was run last to avoid cases being mis-categorised into generic “healthcare” category.

As a result of aggregation, some of the hashtags appeared more prominently, e.g., #patientcare, #covid, #monkeypox, #job (see below for before/after comparison)

- Before

	word	count
0	#healthcare	716
124	#health	80
90	#medicine	41
201	#ai	40
29	#job	38
91	#medical	35
95	#strategy	30

- After

	word	count
568	#healthcare	905
1223	#technology	250
561	#health	80
707	#job	79
346	#covid	76
764	#marketing	53
952	#pharmacy	43

Further data wrangling and plotting – addressing NHS questions

Final data work included analysis of utilisation of capacity, healthcare professional types over time, missed appointments and times between bookings and appointments. To perform the analysis, function used were similar to the described already in the document (group by, filtering, plotting with neat legend and legible labels etc).

From technical point of view some of the highlights have been creating two grouped data frames and then merging them to calculate appointment status as % of all appointments for given months. Some of the code snippet below:

```
# Create a new grouped data set with aggregation by 'hcp_type' and appointment month
ar_subset_visits=ar_subset.groupby(['appointment_month','appointment_status'])[['count_of_appointments']]\
.sum().reset_index()

# Create a new grouped data set with aggregation by appointment month only
ar_subset_visits2=ar_subset.groupby(['appointment_month'])[['count_of_appointments']].sum().reset_index()

# Merge 2 dataframes
ar_subset_visits=ar_subset_visits.merge(ar_subset_visits2, on='appointment_month')
```

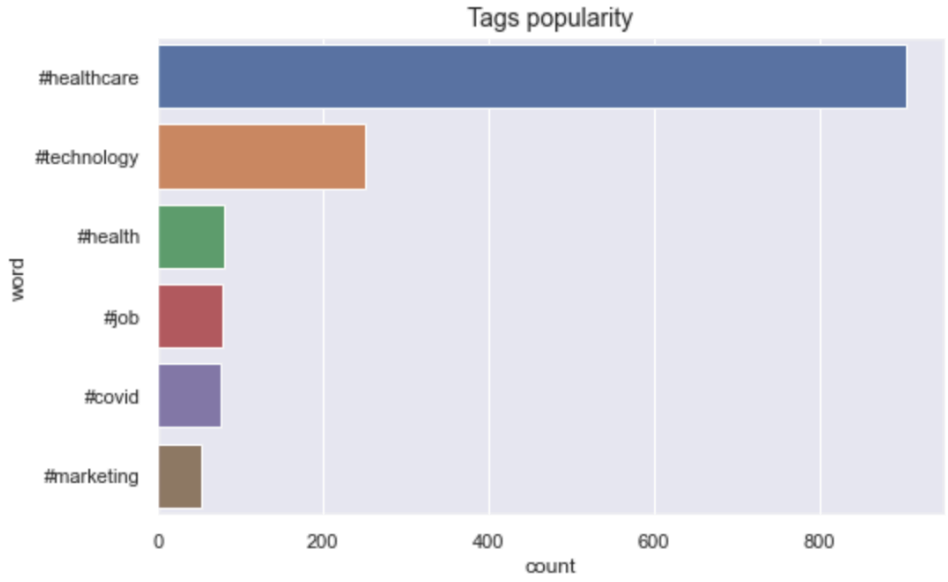
```
# Create new column called attendance (showing % by appointment status for given month)
# by employing lambda function

ar_subset_visits=ar_subset_visits.assign(attendance=lambda x:
                                         x.count_of_appointments_x / x.count_of_appointments_y)
```

General data conclusions:

1) Twitter data

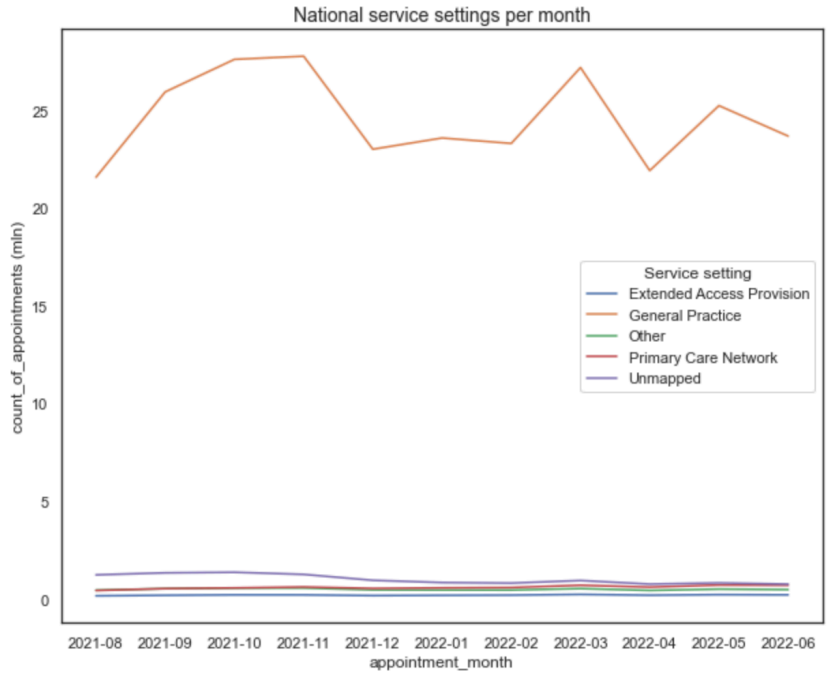
Healthcare and technology have been found as most popular hashtags within the data provided. The presence of technology could indicate that people are, either worried about technological progress or hopeful that technology might alleviate some of the current NHS woes. As Covid pandemic subsided, it became less of a concern, ranking 5th among hashtags.



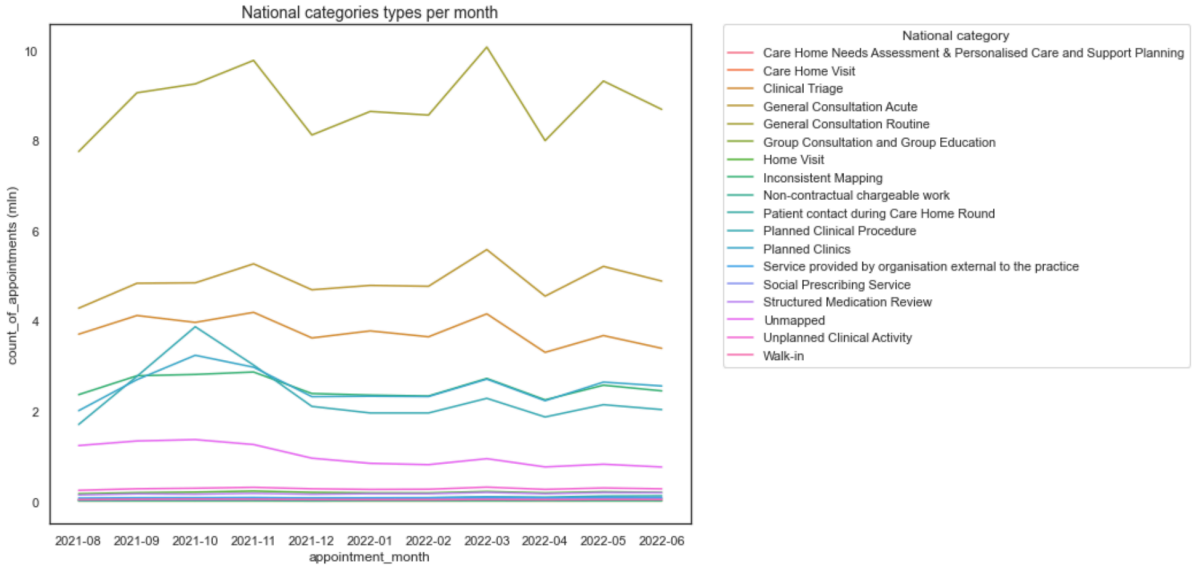
2) National data

a. Service settings and categories

GP is by far the most popular setting for the appointment at national levels. Appointments per month are relatively stable, surprisingly dipping around winter in period, where flu normally dominates and increases pressure on NHS. This can be possible attributed to presence of Covid which made latest trends harder to analyse (Covid waves did not follow normal flu seasons and Covid illnesses did not necessary occur just in winter)



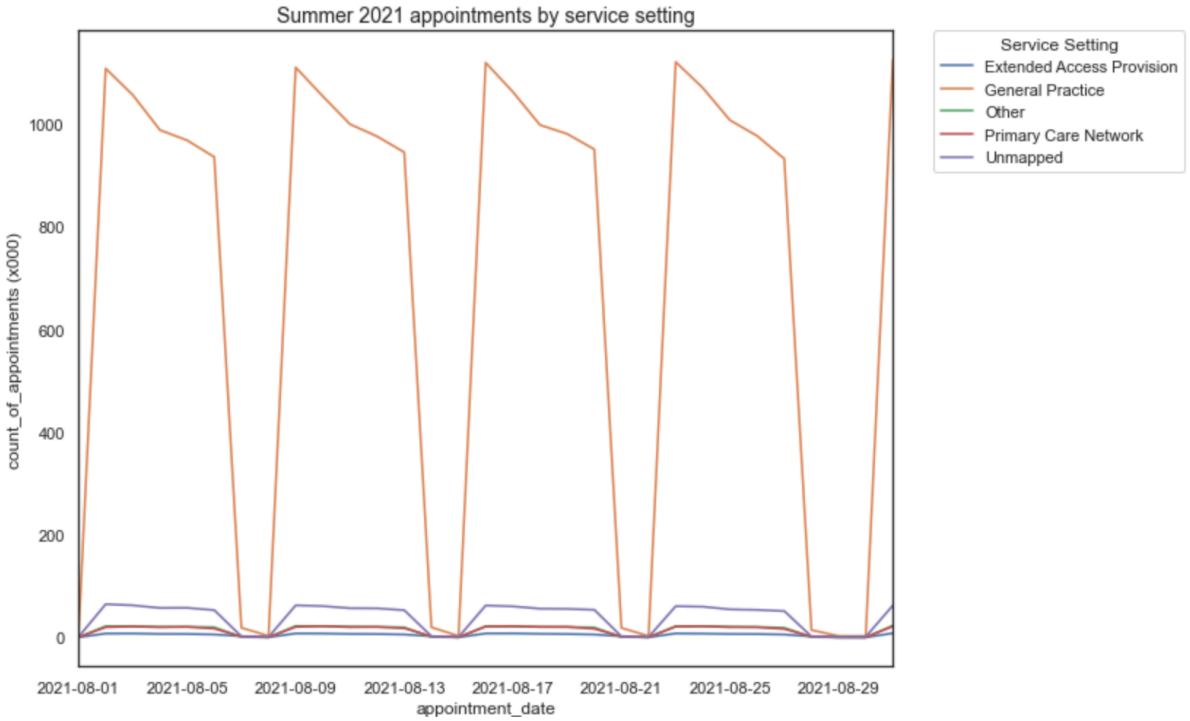
Further splitting the data, routine general consultation dominates the group



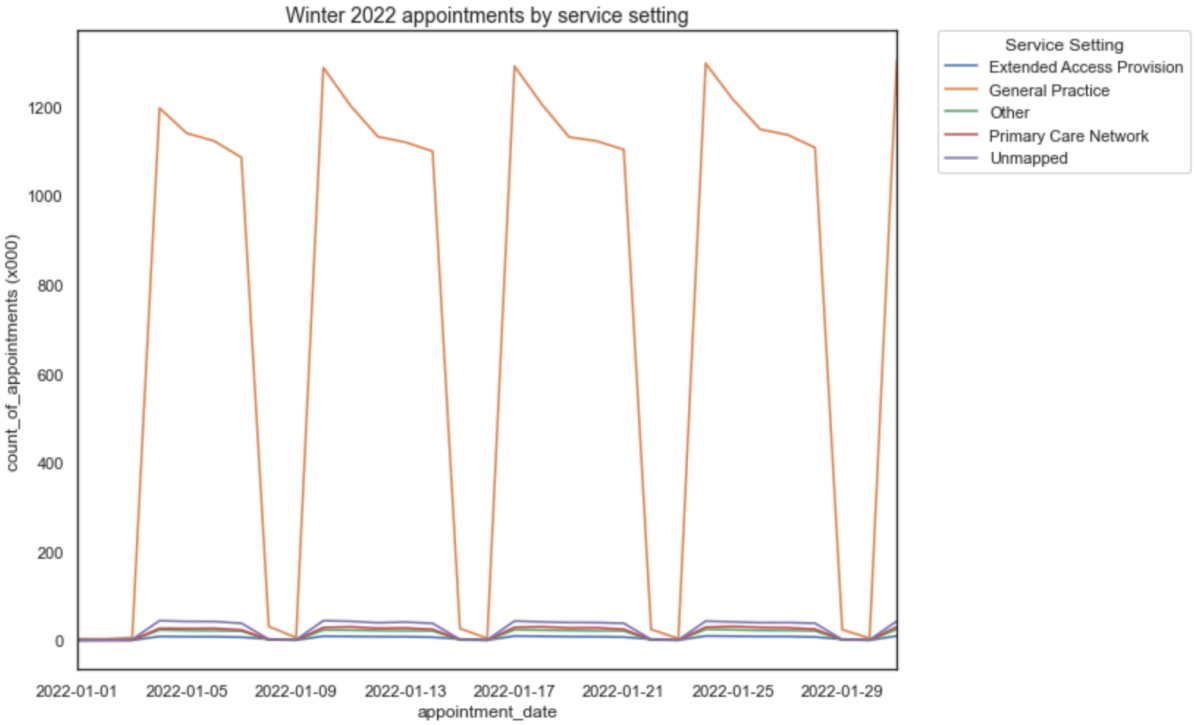
b. Appointments trends within each month

Moving on to appointments per month, we see clear trends of almost no appointments over weekends, highest appointments in the beginning of the weeks with number subsequently dipping as the week progresses. The drop is most pronounced for the GP setting.

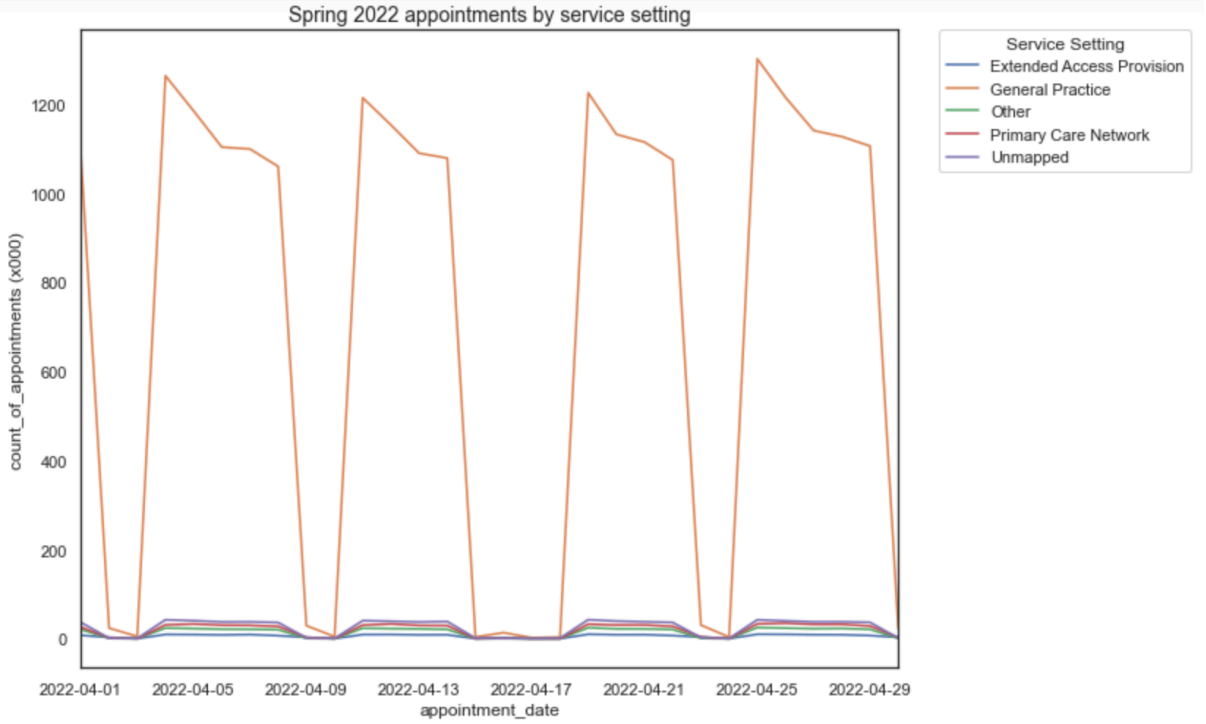
Data for August 2021



Data for January 2022



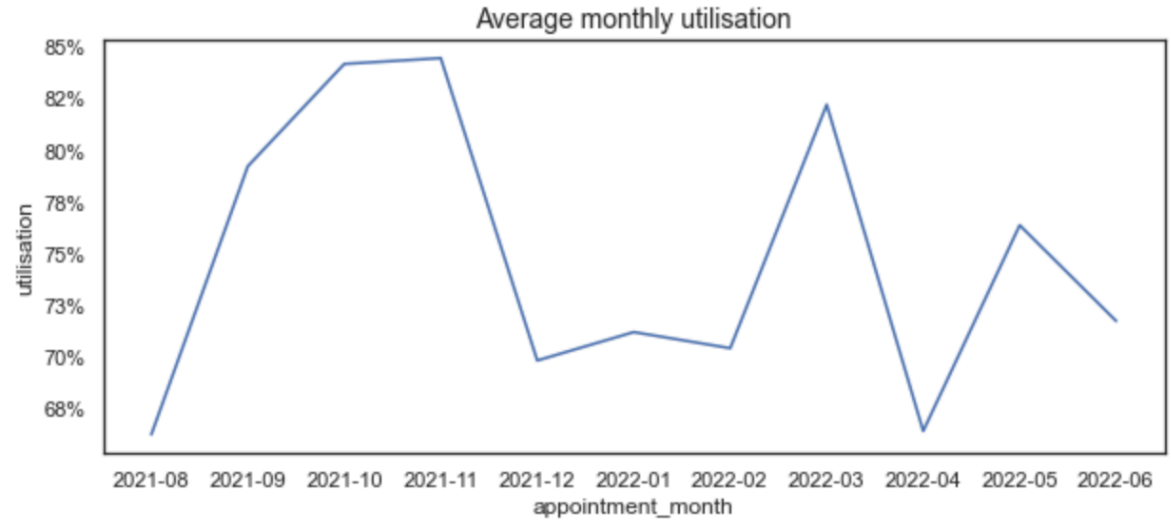
Data April 2022



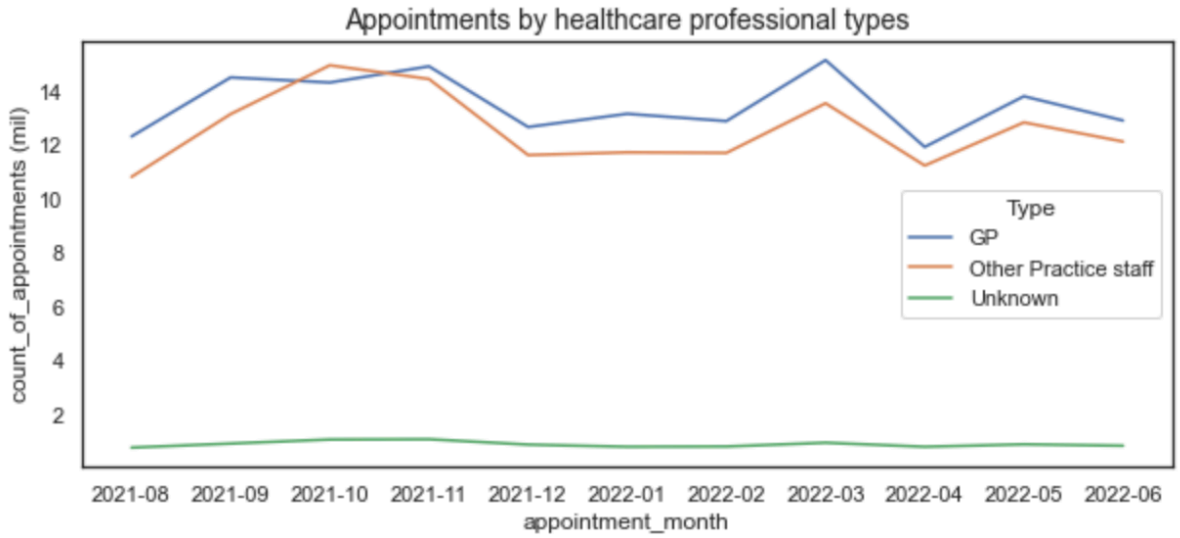
Note, how number of appointments takes time to recover after New Year in January and how it goes down around Easter 2022 (mid-April). Weekdays around Easter see smaller number of appointments than equivalent weekdays in January and April.

3) Regional data – Staffing level & utilisation

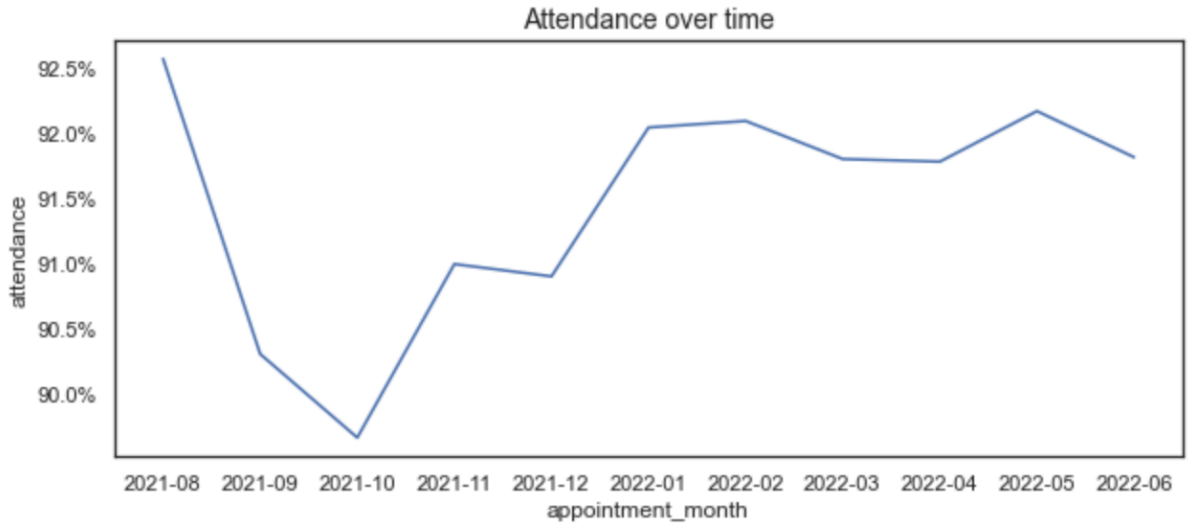
Staff level seems to be appropriate assuming 1.2mil daily capacity of appointments. Average monthly utilisation varies significantly. This is purely driven by number of appointments as we have been provided with assumption of fixed capacity over the period analysed.



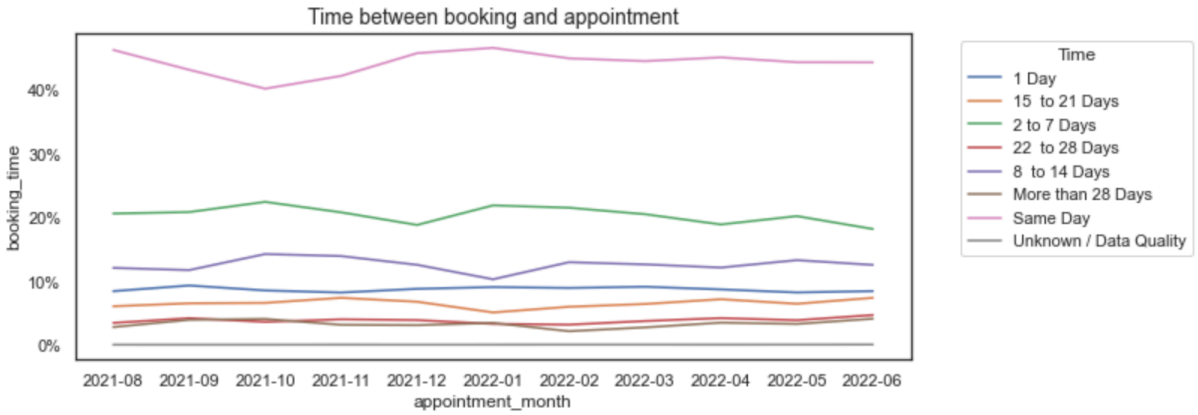
There is also small variation of appointments by healthcare professional type over the period. For yet unknown reasons, “Other practice staff” appointments seem to overtake GP appointments in October 2021. Yet in general they mirror each other quite closely.



Attendance of appointments over time is relatively constant.



People tend to book appointments on the same day (more than 40% of appointments). Interestingly the same day appointments as % of all appointments dropped during September-November 2021 (which coincides with the months of highest numbers of appointments). This might indicate that, despite being within total utilisation, there were localised shortages and people struggled to get appointments on the same day in certain areas/ GP practices. More granular data would need to be provided to explore this further.



appointment_month	time_between_book_and_appointment	booking_time
2021-08	Same Day	46.33%
2021-09	Same Day	43.22%
2021-10	Same Day	40.25%
2021-11	Same Day	42.25%
2021-12	Same Day	45.82%
2022-01	Same Day	46.66%
2022-02	Same Day	45.01%
2022-03	Same Day	44.58%
2022-04	Same Day	45.16%
2022-05	Same Day	44.40%
2022-06	Same Day	44.38%

Key recommendations:

- People indicated interest in Technology within healthcare context. Reasons for that should be investigated further and NHS should consider more investment into technology.
- Staffing level should be varied based on the weekdays as early days of the week seem to be more popular in terms of appointments booked. This could be done for example by limiting holidays on Monday/Tuesday.
- Staffing level should be investigated further in terms of division between GP and other practice staff. The data provided is not granular enough to judge if the split is correct.
- Data quality is not perfect with presence of unknowns and missing categories. NHS should create more categories or/and work on highlighting the importance of correct categorisation.