Report

Background/context of the business scenario:

The NHS incurs significant, potentially avoidable, costs when patients miss general practitioner (GP) appointments. The reasons for missed appointments need to be better understood.

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?

Additionally it will also be useful to consider.

- What is the number of locations, service settings, context types, national categories, and appointment statuses in the data sets?
- What is the date range of the provided data sets, and which service settings reported the most appointments for a specific period?
- What is the number of appointments and records per month?
- What monthly and seasonal trends are evident, based on the number of appointments for service settings, context types, and national categories?
- What are the top trending hashtags (#) on Twitter related to healthcare in the UK?

Analytical approach (350 words):

Data Quality , Import, Cleaning and Sense Check

Several key points were noted from the metadata file provided and have been taken into consideration during the analysis, please see Annex 1 for details.

The CSV files provided for analysis had already been cleaned and unnecessary columns had been dropped to reduce file size.

Pandas, Numpy, Seaborn and matplotlib libraries were imported to facilitate analysis and initial data exploration was used to validate data as follows:

Action	Code	Output	Conclusion
Import csv files	df = pd.read_csv('./*file_path*.csv')	Data imported as dataframe	Data successful imported
Import Excel file	df = pd.read_excel('./*file_path*.xlsx')	Data imported as dataframe	Data successful imported. But nc file is large. Use data subset for analysis to avoid lag time in notebook.

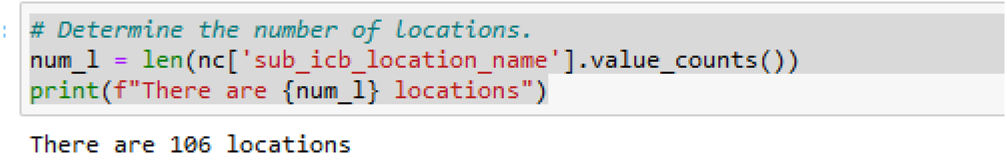
Check for missing values	df.isna().sum()	0	There were no null values for each dataframe.
Check column names	print(df.columns)	All column names listed	The key columns necessary to explore the business question were present
Check for metadata	df.info()	All data types listed	No obvious data discrepancies and the data types in the data set were suitable for the required analysis

Initial Exploratory Analysis

To get a sense of the data and verify it contained representative data the following questions were answered:

1. How many locations are there in the data set?
2. What are the five locations with the highest number of records?
3. How many service settings, context types, national categories, and appointment statuses are there?

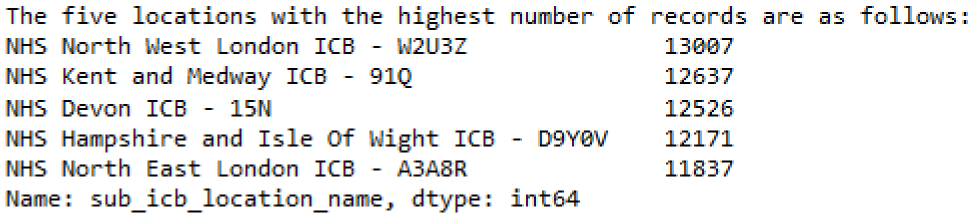
Length and value counts functions were used to answer these questions and f strings were used to print clear answers:



```
# Determine the number of locations.
num_l = len(nc['sub_icb_location_name'].value_counts())
print(f"There are {num_l} locations")

There are 106 locations
```

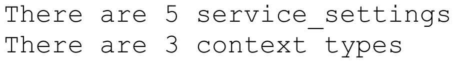
This was verified against the public information on https://digital.nhs.uk/services/organisation-data-service/integrated-care-boards/implementation-of-icbs-from-april-2022 to better understand how the data fits together.

In addition, a head function was used to summarise the data to answer question 2:



```
The five locations with the highest number of records are as follows:
NHS North West London ICB - W2U3Z               13007
NHS Kent and Medway ICB - 91Q                   12637
NHS Devon ICB - 15N                             12526
NHS Hampshire and Isle Of Wight ICB - D9Y0V     12171
NHS North East London ICB - A3A8R               11837
Name: sub_icb_location_name, dtype: int64
```

Answers:



```
There are 5 service_settings
There are 3 context types
There are 18 national categories
There are 3 appointment status
```

Further analysis

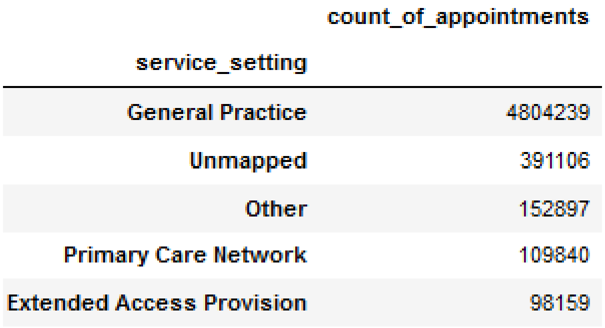
Additional questions considered to become familiar with appointment data and explore the business question relating to missed appointments:

1. **Between what dates were appointments scheduled?**

Using min and max functions along with strftime and f strings it was found that the dates were as follows:

- In the ad dataframe the earliest date is 01 December 2021 and the latest date is 30 June 2022.
- In the nc dataframe the earliest date is 01 August 2021 and the latest date is 30 June 2022

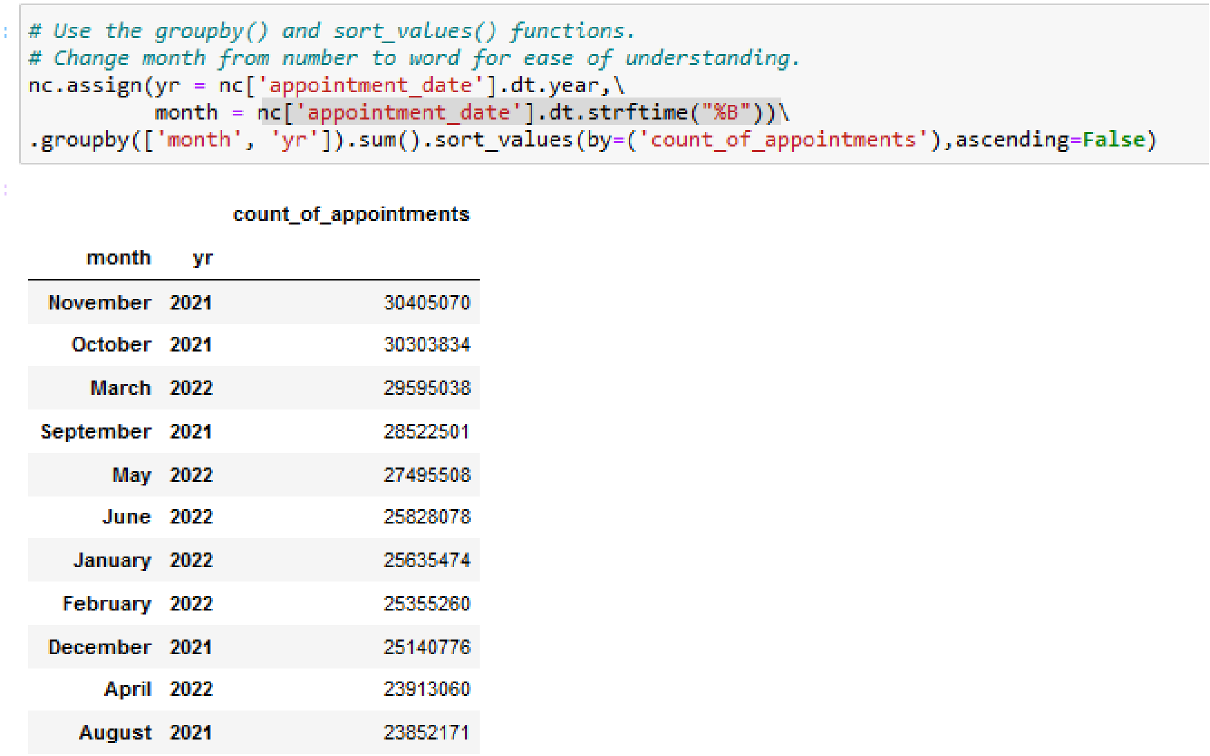
2. **Which service setting reported the most appointments in North West London from 1 January to 1 June 2022?**
Note in the previous step this location was identified as having the highest number of records.



	count_of_appointments
service_setting	
General Practice	4804239
Unmapped	391106
Other	152897
Primary Care Network	109840
Extended Access Provision	98159

General practice has by the far the highest number of appointments.

3. **Which month had the highest number of appointments?**
The following code as used to summarise the appointment count by month, showing November and October as the 2 months with the busiest months for appointments:

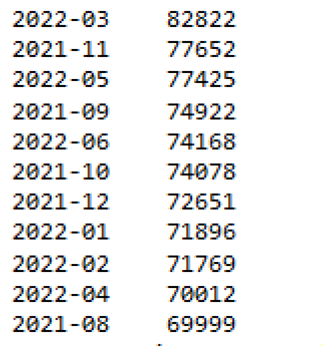


```
# Use the groupby() and sort_values() functions.
# Change month from number to word for ease of understanding.
nc.assign(yr = nc['appointment_date'].dt.year,\
          month = nc['appointment_date'].dt.strftime("%B"))\
.groupby(['month', 'yr']).sum().sort_values(by=('count_of_appointments'),ascending=False)
```

month	yr	count_of_appointments
November	2021	30405070
October	2021	30303834
March	2022	29595038
September	2021	28522501
May	2022	27495508
June	2022	25828078
January	2022	25635474
February	2022	25355260
December	2021	25140776
April	2022	23913060
August	2021	23852171

4. What was the total number of records per month?

A simple value count was used to count the number of records per month



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

It's unclear why there is such a variation across the months but note that it doesn't correspond to the number of appointments. It will be important for the rest of the analysis to focus on the sum of appointment count, rather than the count in order to answer the business question.

A visual analysis of the various factors that might affect appointments was conducted, in 3 main areas focussing on the following questions in order to answer the business problem:

1. Number of appointments per month for service settings, context types, and national categories.
2. Number of appointments for service setting per season
3. Whether there were adequate staff and capacity in the networks?
4. What the actual utilisation of resources was?
5. Should the NHS start looking at increasing staff levels?
6. How do the healthcare professional types differ over time?

7. Are there significant changes in whether or not visits are attended?
8. Are there changes in terms of appointment type and the busiest months?
9. Are there any trends in time between booking and appointment?
10. How do the various service settings compare?

Areas of exploration:

1. Service providers, context type and care type
    - Data was subset, pivoted and plotted as per example in Annex 3.
2. Seasons
    - Sample months were selected to represent the 4 seasons
    - Data was subset, filtered and plotted as per example in annex 4
3. Health care provider, booking duration and appointment mode.
    - Data was subset, filtered (to show days from August 2020 onward, to match the season data) and plotted as per example in annex 6

The key visuals and findings have been extracted as part of the insight section below.

In addition, a quick exploration of the Twitter data was performed and recommendations for future use are included in the concluding section of this report. Details of code used to extract hastags is included in Annex 5.

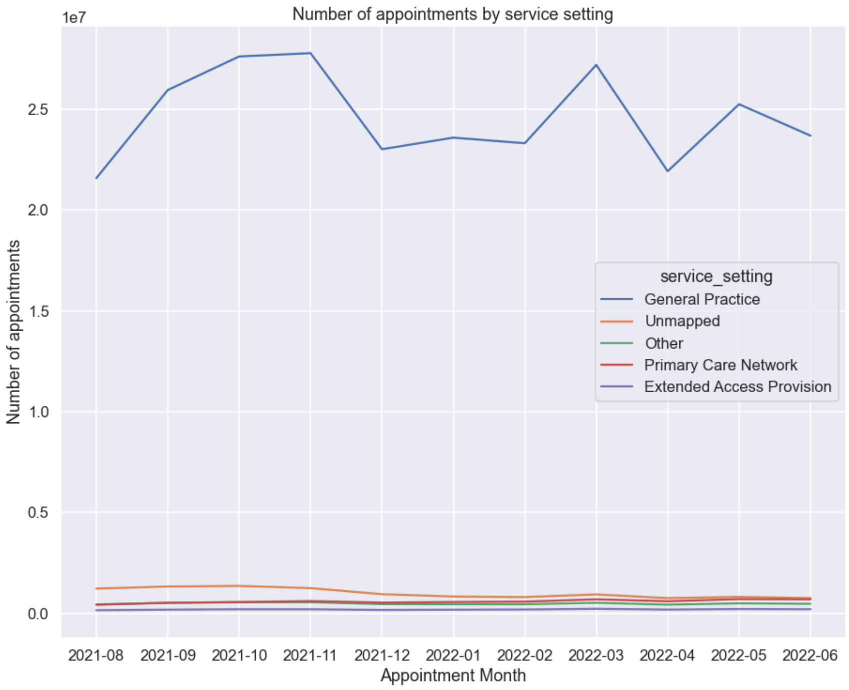
Full details of all code and step by step analysis can be found in the associated Jupyter notebook.

Visualisation and insights (350 words)

Information from visualisations indicating the number of appointments per month for service settings, context types, and national categories:

Line charts were selected to show changes across the months over the year.

- Service type shows that General Practice consistently provides over 90% of appointments



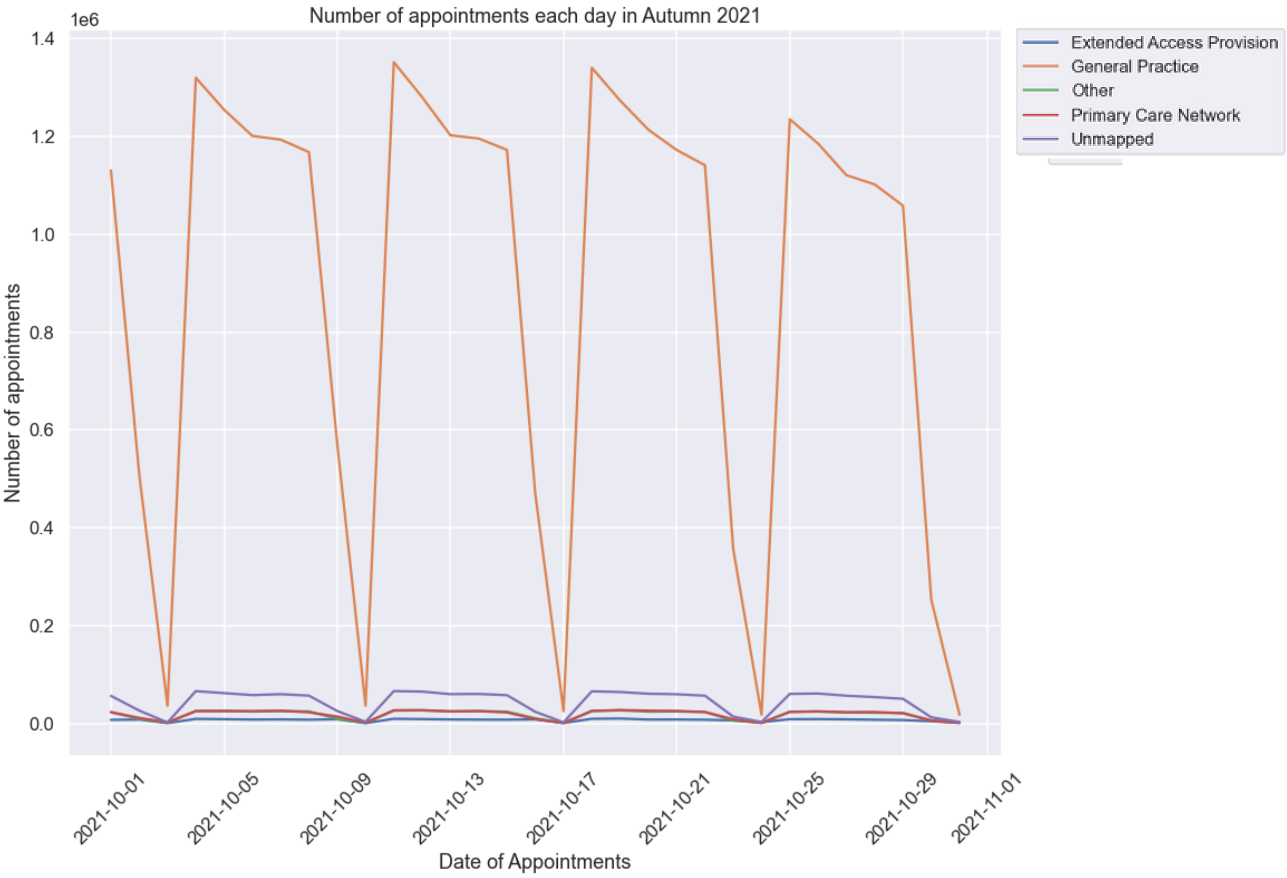
- Context type provides little valuable info as most appointments listed as care related
- National categories may provide useful info if it can be identified which appointments could be carried out by personnel other than GPs.

Seasonal information

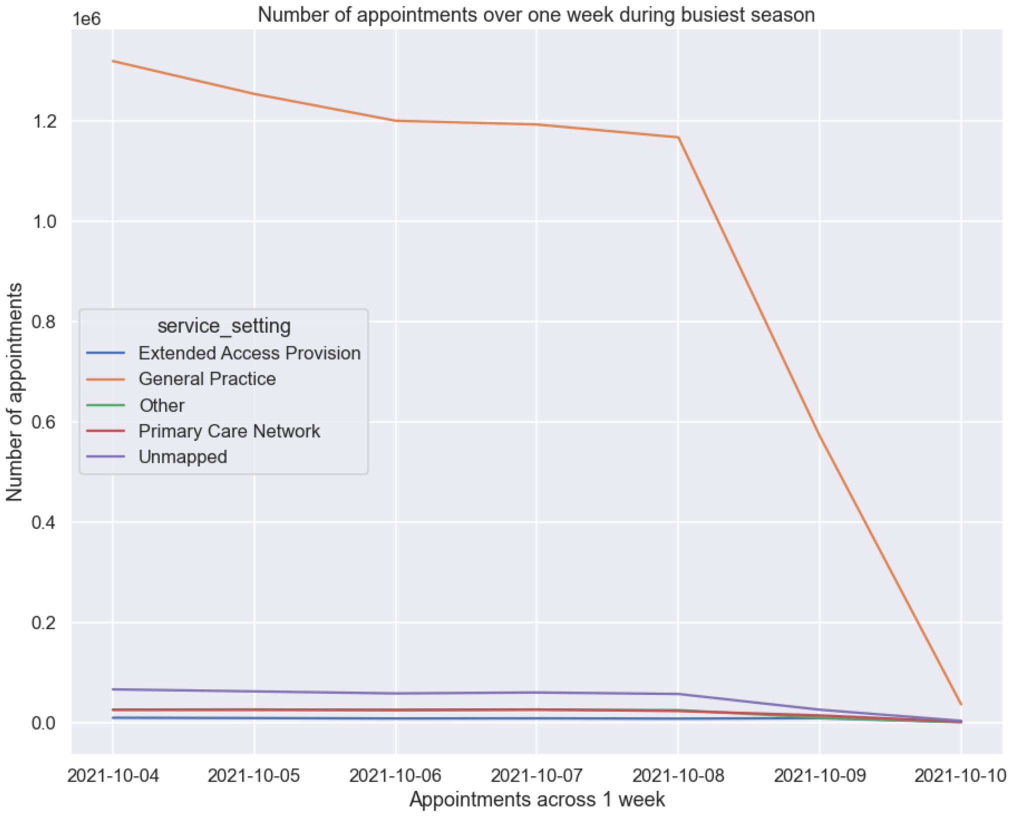
Sample months have been selected to represent the seasons as follows :

- Summer (August 2021),
- Autumn (October 2021)
- Winter (January 2022)
- Spring (April 2022)

As identified in the initial analysis there are season variation in appointment, with autumn being the busiest. Additionally, there are weekly variations.


Number of appointments each day in Autumn 2021
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Number of appointments
2021-10-01
2021-10-05
2021-10-09
2021-10-13
2021-10-17
2021-10-21
2021-10-25
2021-10-29
2021-11-01
Date of Appointments
Extended Access Provision
General Practice
Other
Primary Care Network
Unmapped

The highest number of appointments across the seasons is usually on Mondays, decreasing across the week:



Twitter

At this time the twitter data provided is of limited use to answer the business question. If hastags are to provide useful information on trending topics in future, decisions will have to be made about whether to group similar tags. For example:

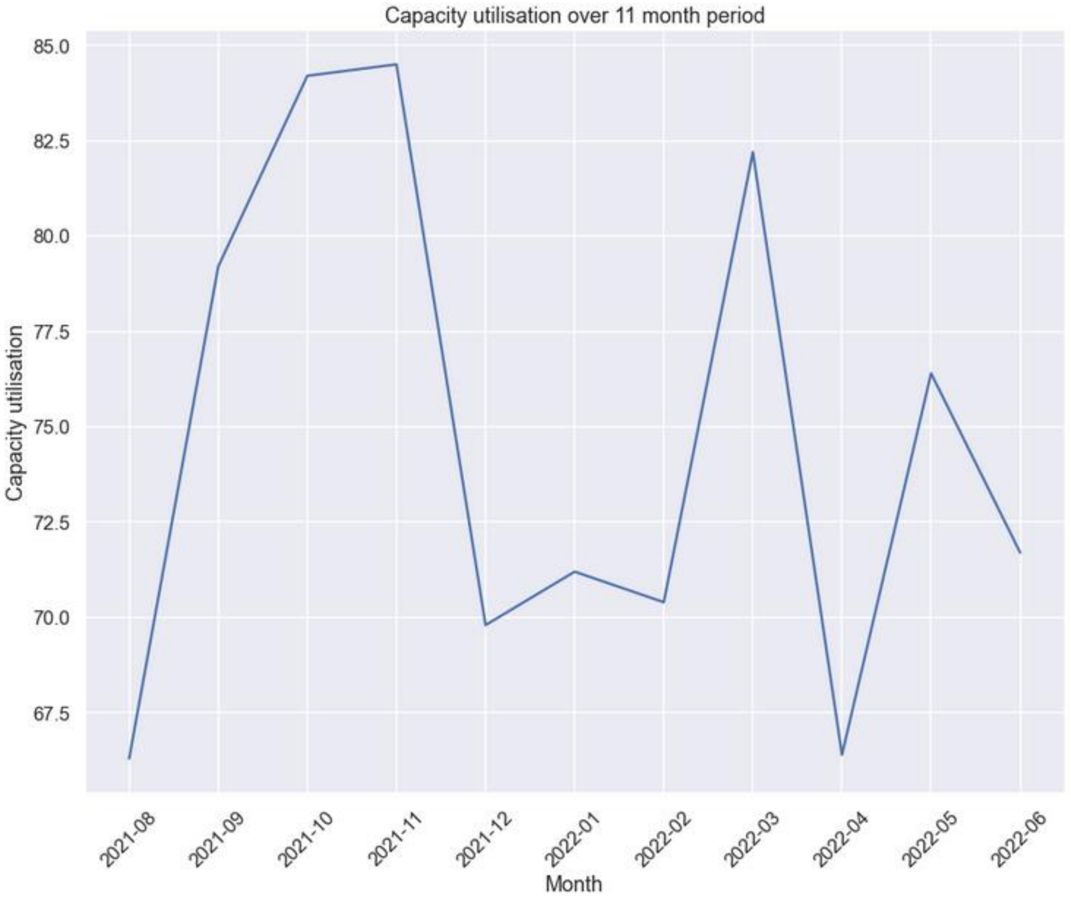
covid	16
coronavirus	16
covid19	14

If combined these tags come to 46, which would make it the 3rd most popular topic.

Utilisation, Health care provider, booking duration and appointment mode.

Data used was filtered to show months from August 2020 onward to match the season data analysis.

Utilisation

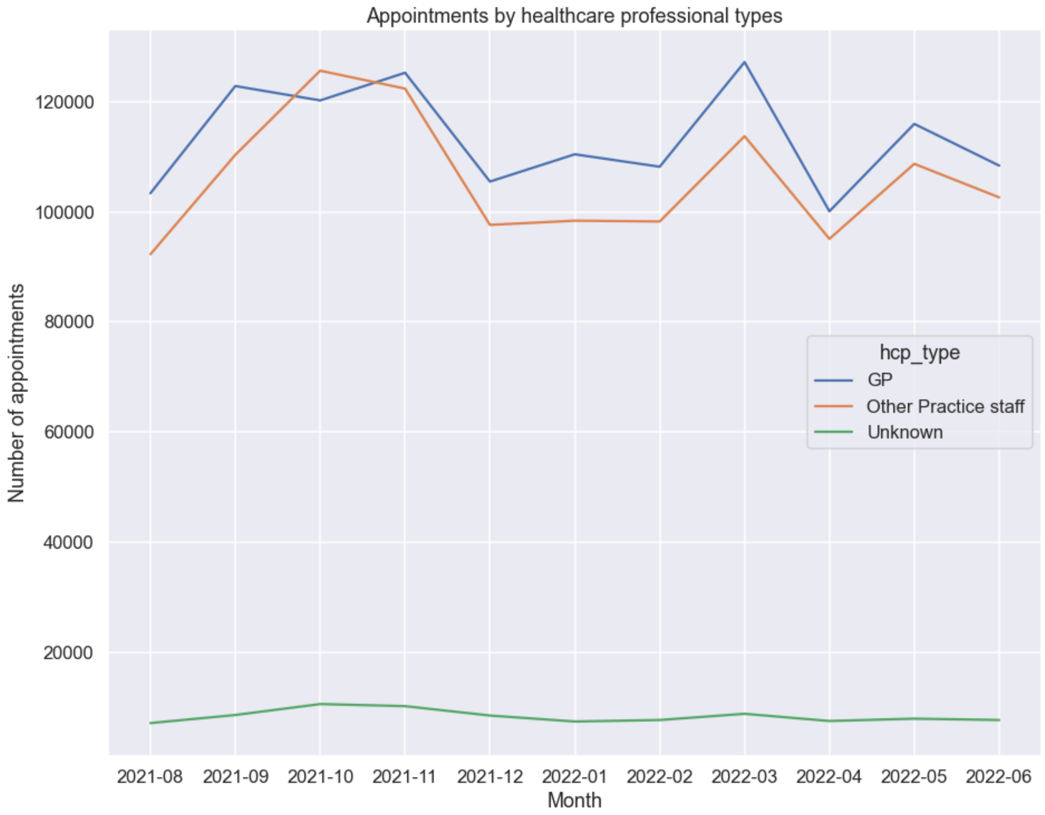


Maximum utilisation for busiest month, November 2021 is 84.5%. This suggests even in busiest period maximum utilisation is not reached. However it is not known what utilisation parameters are suitable for NHS to enable risk mitigation and keep services running. It is not known if utilisation varies in different regions and data for several years would be required to establish whether there is an increasing trend that may mean it is necessary to invest in further staff to avoid capacity problems in the coming years.

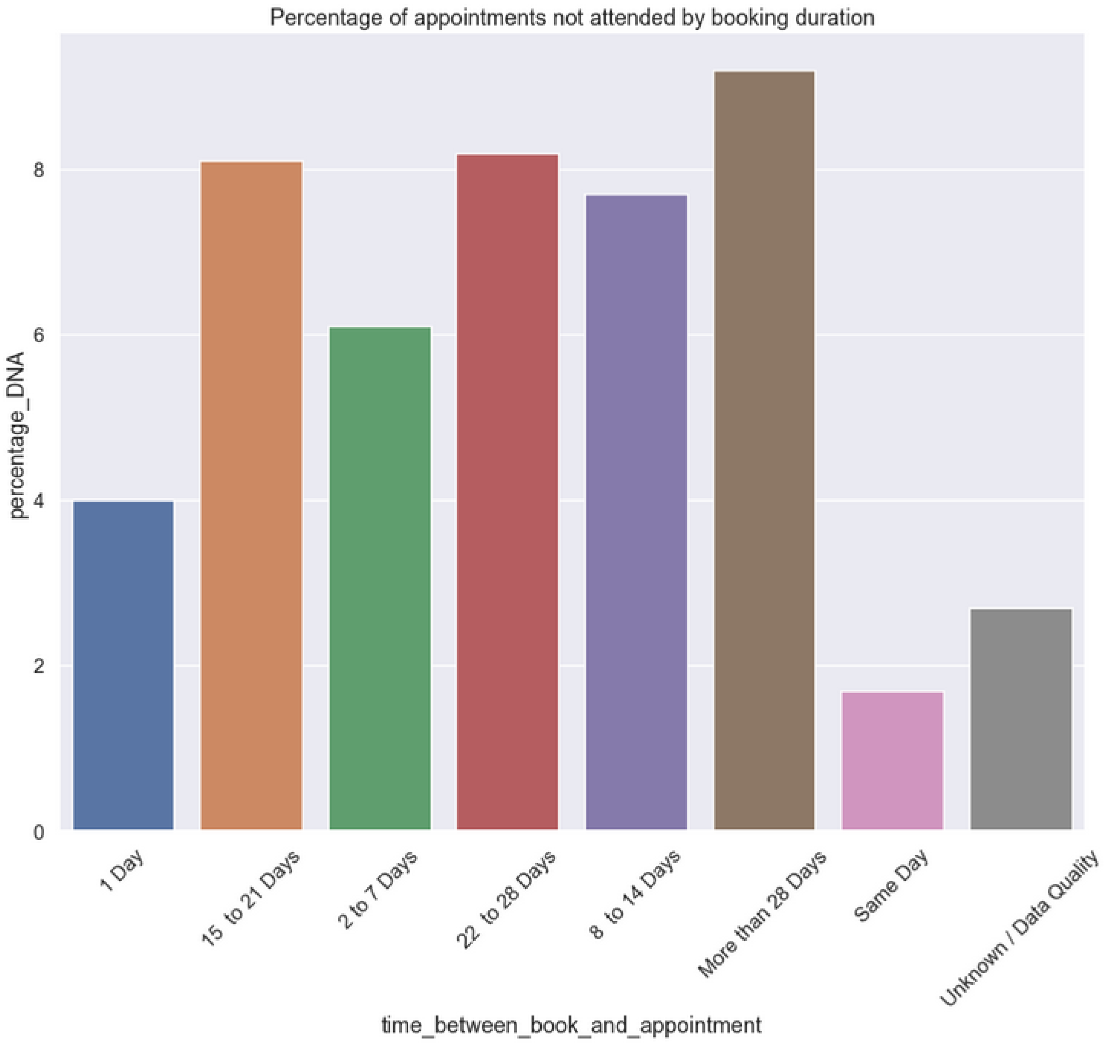
While it is possible to show the current utilisation it is not possible to say whether the NHS needs more staff.

Health care professional types

Most appointments are provided by GPs, while Other practice staff provided more appointments during busiest autumn phase, not clear why, but corresponds to slight drop in GP appointments.



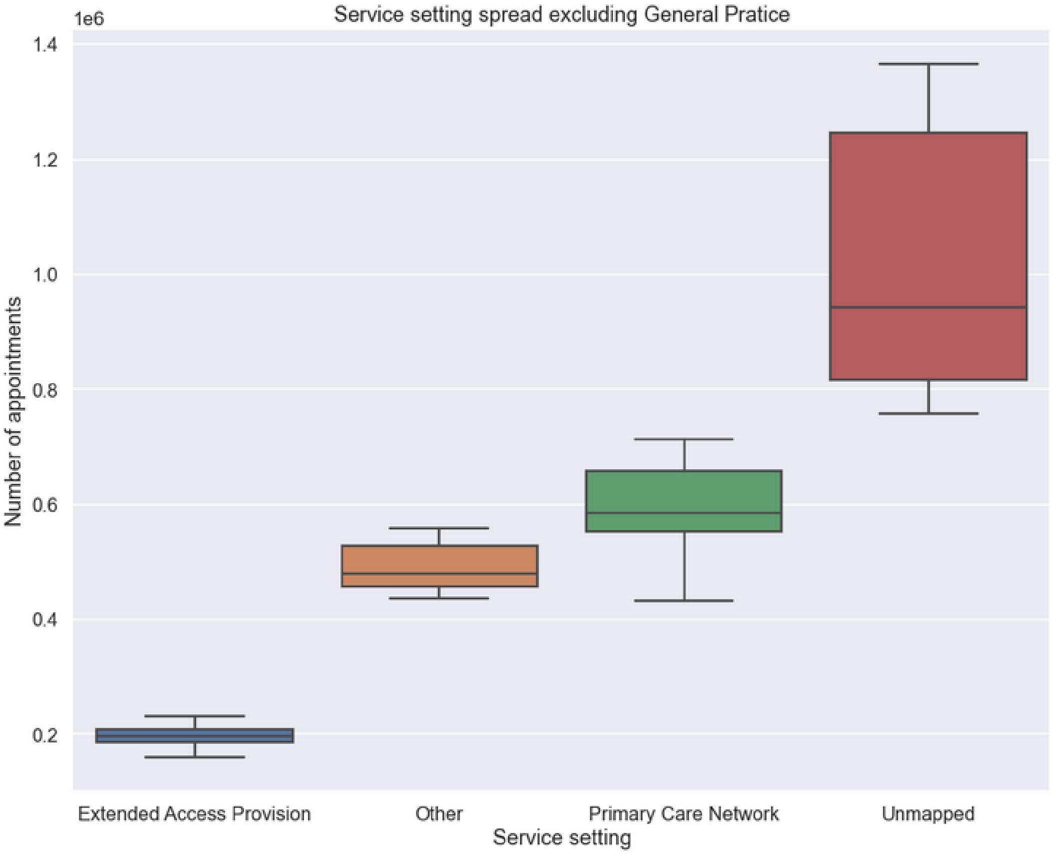
The longer the time between booking an appointment and attending it the higher the chance of a missed appointment.



Other points of note

- Most appointments are attended, with variations matching appointment number.
- Most appointments are face to face.
- Most appointments are booked and attended the same day.

Improved mapping of service types would provide more accurate data. Unmapped service setting showed the second highest appointments after General practices.



Patterns and predictions (200 words):

Summary of findings:

- General Practice consistently provides over 90% of appointments
- Autumn was the busiest season over the last year.
- Most appointments are provided by GPs.
- The highest number of appointments across the seasons is usually on Mondays, decreasing across the week.
- Maximum utilisation for busiest month, November 2021 is 84.5%.
- Most appointments are attended, with variations matching appointment number.
- Most appointments are face to face.
- Most appointments are booked and attended the same day.
- The longer the duration between booking and attending appointment the greater the chance of a missed appointment.

Conclusion

The reasons for missed appointments requires more information in order to be better understood.

At this stage of the project the two main questions posed by the NHS cannot fully be answered.

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

The data suggests that even in busiest period maximum utilisation reaches 84.5% and does not exceed full capacity. However it is not known what utilisation parameters are suitable for NHS to enable it mitigate its risks such as staff sickness and keep services running.

Recommendations

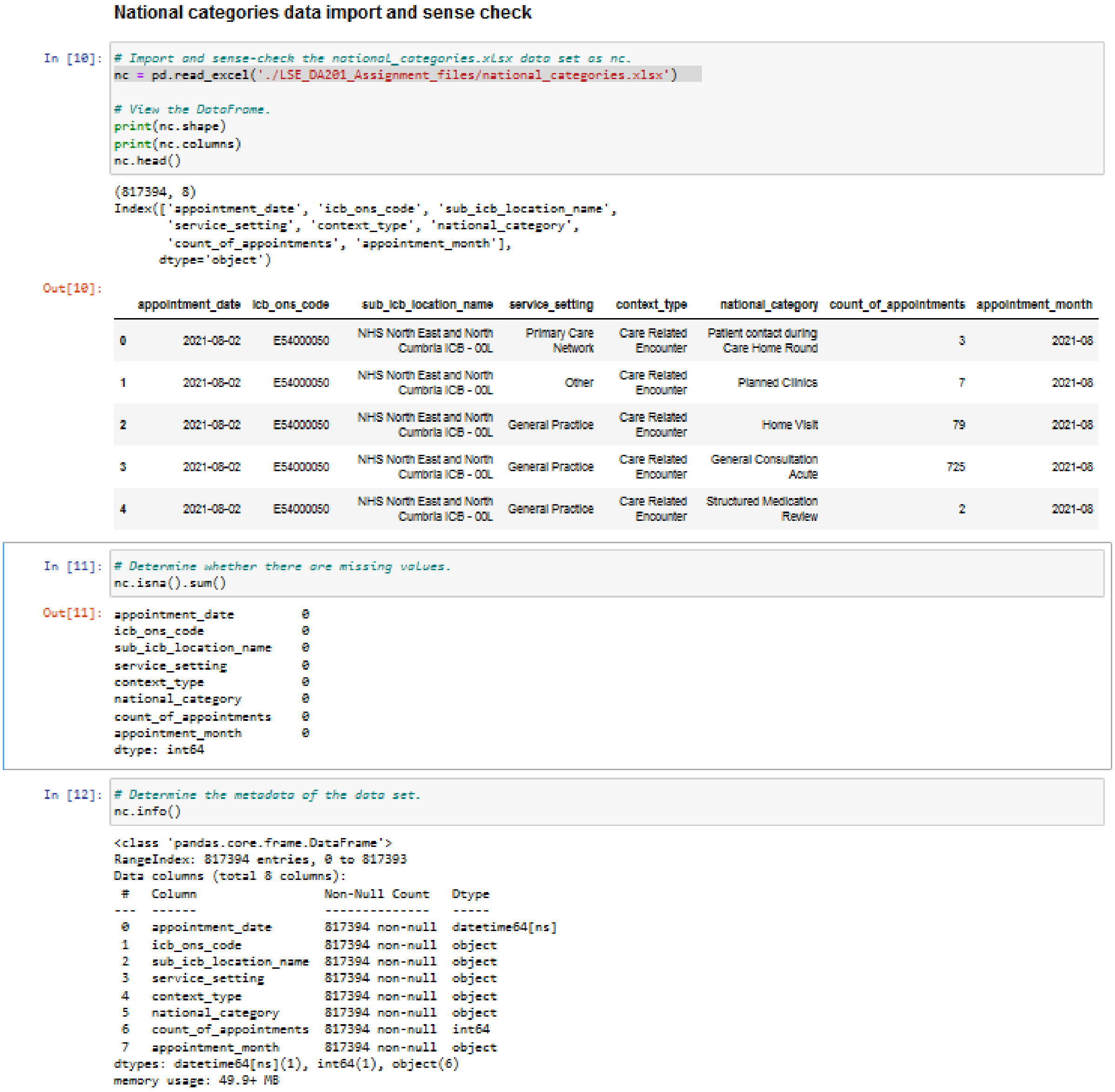
1. Source additional data on staffing levels and requirements over time and across different regions. Consider these 2 key questions
   a. What is the trend in appointment requirements over the years? Will more appointments be required in future?
   b. Do different regions vary in term of capacity utilisation and can regions support each other during busy times?
2. Consider implementing a reminder system for all appointments, not booked and attended of the same day. A pilot study could be undertaken for all bookings where the booking duration is more than 28 days.
3. Explore different appointment care types further, and establish if using other health care professional (hcp) types to provide certain appointment services instead of GPs could reduce number of GP provided appointments
4. Use Twitter to poll or host a Twitter live chat with NHS users and ask them why they miss appointments to identify further solutions.
5. Once Twitter poll data has been analysed and results of reminder trial are known, use Twitter to launch targeted and informed campaign to reduce missed appointments.

Annex 1 – Key notes from metadata file provided

The following points were noted from the metadata file provided and have been taken into consideration during the analysis:

- There are no national standards for data entry about activity, and widespread variation in approach to appointment management between practices.
- Appointment status changes over time and the reports include the final status of each appointment. For 3%–6% of monthly appointments the status is recorded as unknown because the final status was not updated and remained as 'booked'.
- Due to an issue with the data collection, DNA appointments were not captured correctly after June 2018 and are under-reported until and including November 2018 for all practices using the TPP SystmOne system.
- Appointment mode is set locally by the practices so it may not represent the actual
- care setting of the appointment.
- Many telephone triage and home visits appear as one long blocked period and are not booked to individual patients. Unless home visits and telephone triage are logged as individual appointments and booked to a patient, they will not appear in this publication.
- Practices using the Cegedim GP system are unable to supply appointment mode data. Consequently, the proportion of appointments with an 'Unknown' appointment mode is higher in releases from July 2019 onwards when Cegedim practices were included in the publication.
- Not all practices in England are included in this release (see Data Quality: Practice Coverage) meaning the total number of appointments is not known. An estimate of the total number of appointments in England has been provided. It does not include all GP activity or provide information about demand or capacity of appointments in general practice.

Annex 2 – Analysis exerts from Jupyter notebook



National categories data import and sense check

```
In [10]: # Import and sense-check the national_categories.xlsx data set as nc.
         nc = pd.read_excel('./LSE_DA201_Assignment_files/national_categories.xlsx')

         # View the DataFrame.
         print(nc.shape)
         print(nc.columns)
         nc.head()
```

```
(817394, 8)
Index(['appointment_date', 'icb_ons_code', 'sub_icb_location_name',
       'service_setting', 'context_type', 'national_category',
       'count_of_appointments', 'appointment_month'],
      dtype='object')
```

Out[10]:

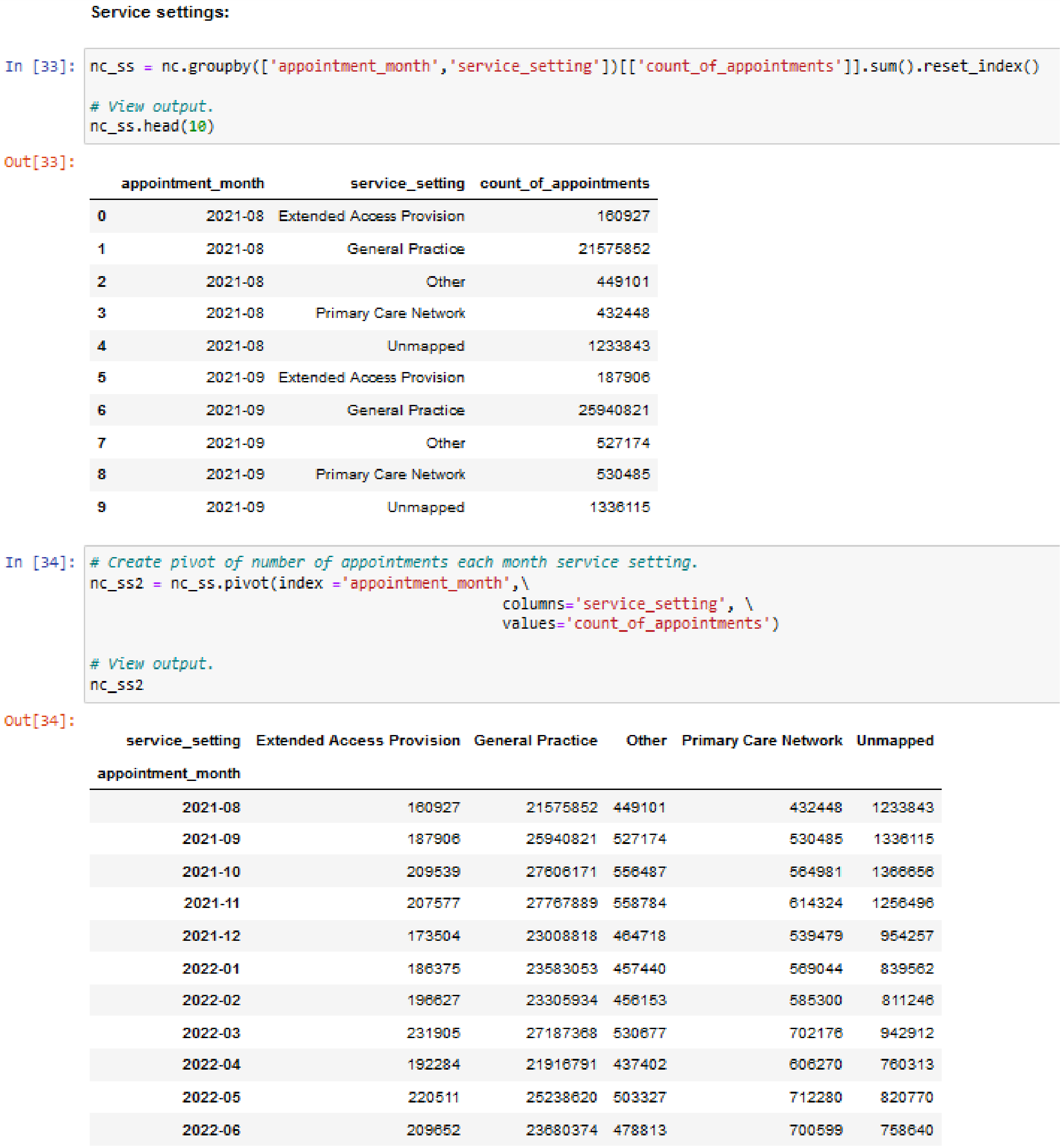
	appointment_date	icb_ons_code	sub_icb_location_name	service_setting	context_type	national_category	count_of_appointments	appointment_month
0	2021-08-02	E54000050	NHS North East and North Cumbria ICB - 00L	Primary Care Network	Care Related Encounter	Patient contact during Care Home Round	3	2021-08
1	2021-08-02	E54000050	NHS North East and North Cumbria ICB - 00L	Other	Care Related Encounter	Planned Clinics	7	2021-08
2	2021-08-02	E54000050	NHS North East and North Cumbria ICB - 00L	General Practice	Care Related Encounter	Home Visit	79	2021-08
3	2021-08-02	E54000050	NHS North East and North Cumbria ICB - 00L	General Practice	Care Related Encounter	General Consultation Acute	725	2021-08
4	2021-08-02	E54000050	NHS North East and North Cumbria ICB - 00L	General Practice	Care Related Encounter	Structured Medication Review	2	2021-08

```
In [11]: # Determine whether there are missing values.
         nc.isna().sum()

Out[11]: appointment_date         0
         icb_ons_code             0
         sub_icb_location_name    0
         service_setting          0
         context_type             0
         national_category        0
         count_of_appointments    0
         appointment_month        0
         dtype: int64
```

```
In [12]: # Determine the metadata of the data set.
         nc.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 817394 entries, 0 to 817393
         Data columns (total 8 columns):
          #   Column                 Non-Null Count   Dtype
         ---  ------                 --------------   -----
          0   appointment_date       817394 non-null  datetime64[ns]
          1   icb_ons_code           817394 non-null  object
          2   sub_icb_location_name  817394 non-null  object
          3   service_setting        817394 non-null  object
          4   context_type           817394 non-null  object
          5   national_category      817394 non-null  object
          6   count_of_appointments  817394 non-null  int64
          7   appointment_month      817394 non-null  object
         dtypes: datetime64[ns](1), int64(1), object(6)
         memory usage: 49.9+ MB
```

Annex 3 – Exert of code used for visuals analysis, part 4

Code for data frame grouping and pivoting



Service settings:

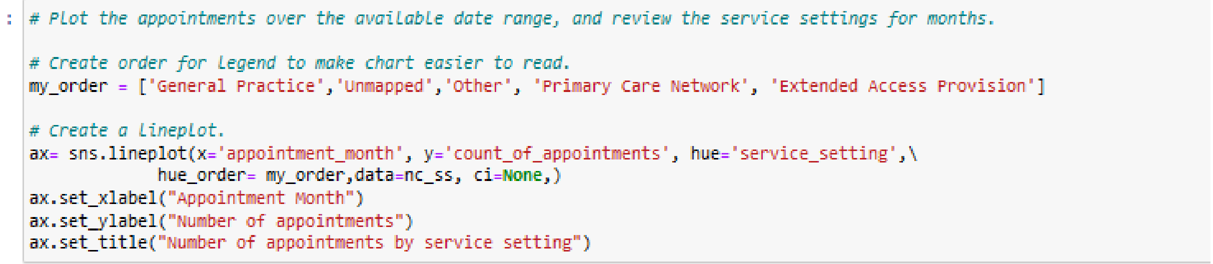
```
In [33]: nc_ss = nc.groupby(['appointment_month','service_setting'])[['count_of_appointments']].sum().reset_index()

         # View output.
         nc_ss.head(10)
```

```
Out[33]:
     appointment_month          service_setting  count_of_appointments
  0            2021-08  Extended Access Provision                 160927
  1            2021-08         General Practice               21575852
  2            2021-08                    Other                 449101
  3            2021-08     Primary Care Network                 432448
  4            2021-08                Unmapped                1233843
  5            2021-09  Extended Access Provision                 187906
  6            2021-09         General Practice               25940821
  7            2021-09                    Other                 527174
  8            2021-09     Primary Care Network                 530485
  9            2021-09                Unmapped                1336115
```

```
In [34]: # Create pivot of number of appointments each month service setting.
         nc_ss2 = nc_ss.pivot(index ='appointment_month',\
                                       columns='service_setting', \
                                       values='count_of_appointments')

         # View output.
         nc_ss2
```

```
Out[34]:
      service_setting  Extended Access Provision  General Practice   Other  Primary Care Network  Unmapped
    appointment_month
              2021-08                     160927          21575852  449101                432448   1233843
              2021-09                     187906          25940821  527174                530485   1336115
              2021-10                     209539          27606171  556487                564981   1366656
              2021-11                     207577          27767889  558784                614324   1256496
              2021-12                     173504          23008818  464718                539479    954257
              2022-01                     186375          23583053  457440                569044    839562
              2022-02                     196627          23305934  456153                585300    811246
              2022-03                     231905          27187368  530677                702176    942912
              2022-04                     192284          21916791  437402                606270    760313
              2022-05                     220511          25238620  503327                712280    820770
              2022-06                     209652          23680374  478813                700599    758640
```

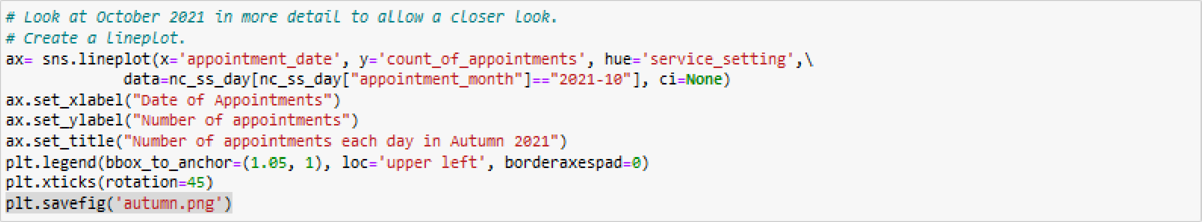
Code for chart



```
: # Plot the appointments over the available date range, and review the service settings for months.

  # Create order for legend to make chart easier to read.
  my_order = ['General Practice','Unmapped','Other', 'Primary Care Network', 'Extended Access Provision']

  # Create a lineplot.
  ax= sns.lineplot(x='appointment_month', y='count_of_appointments', hue='service_setting',\
               hue_order= my_order,data=nc_ss, ci=None,)
  ax.set_xlabel("Appointment Month")
  ax.set_ylabel("Number of appointments")
  ax.set_title("Number of appointments by service setting")
```

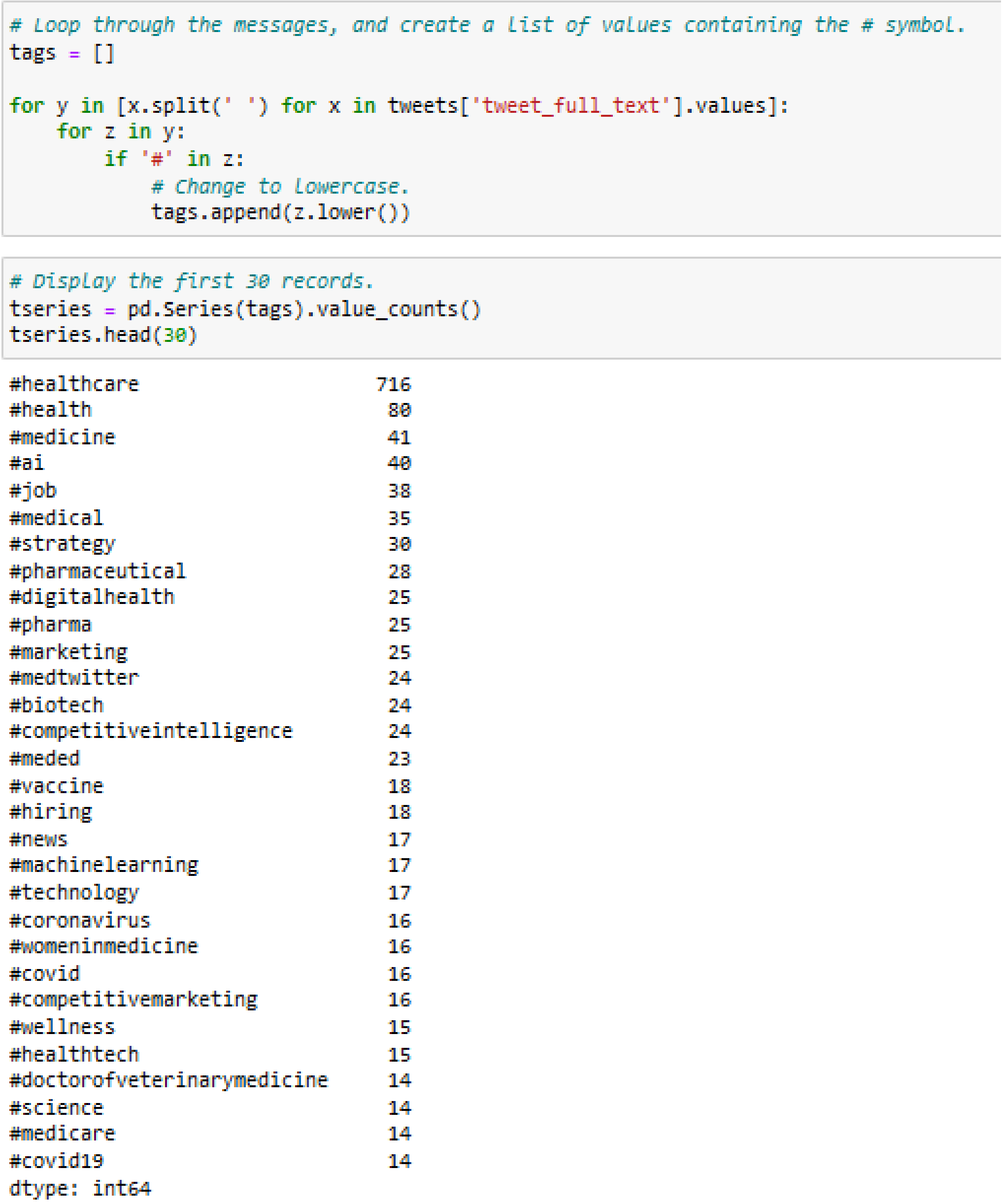
Annex 4 – Seasons

Code to create a line plot for autumn season using the sample month of October



```
# Look at October 2021 in more detail to allow a closer look.
# Create a lineplot.
ax= sns.lineplot(x='appointment_date', y='count_of_appointments', hue='service_setting',\
             data=nc_ss_day[nc_ss_day["appointment_month"]=="2021-10"], ci=None)
ax.set_xlabel("Date of Appointments")
ax.set_ylabel("Number of appointments")
ax.set_title("Number of appointments each day in Autumn 2021")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0)
plt.xticks(rotation=45)
plt.savefig('autumn.png')
```

Annex 5 – Twitter hastag analysis



```
# Loop through the messages, and create a list of values containing the # symbol.
tags = []

for y in [x.split(' ') for x in tweets['tweet_full_text'].values]:
    for z in y:
        if '#' in z:
            # Change to lowercase.
            tags.append(z.lower())
```

```
# Display the first 30 records.
tseries = pd.Series(tags).value_counts()
tseries.head(30)
```

```
#healthcare                   716
#health                        80
#medicine                      41
#ai                            40
#job                           38
#medical                       35
#strategy                      30
#pharmaceutical                28
#digitalhealth                 25
#pharma                        25
#marketing                     25
#medtwitter                    24
#biotech                       24
#competitiveintelligence       24
#meded                         23
#vaccine                       18
#hiring                        18
#news                          17
#machinelearning               17
#technology                    17
#coronavirus                   16
#womeninmedicine               16
#covid                         16
#competitivemarketing          16
#wellness                      15
#healthtech                    15
#doctorofveterinarymedicine    14
#science                       14
#medicare                      14
#covid19                       14
dtype: int64
```