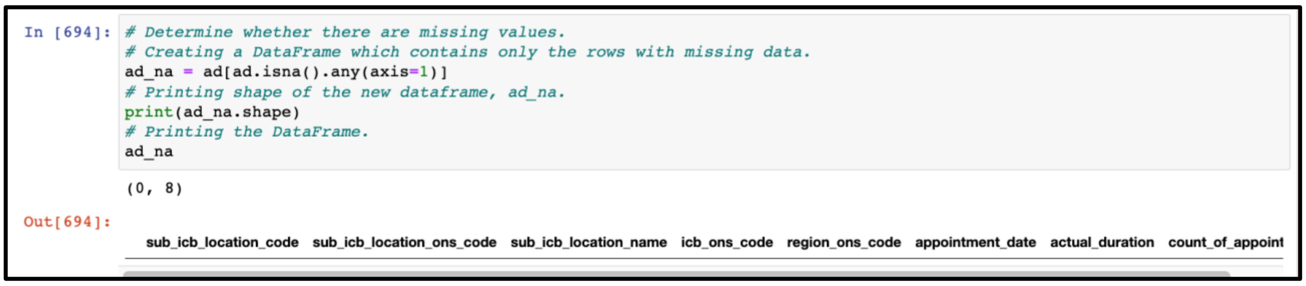
The NHS has been the backbone of the UK since 1948, providing free healthcare based on people's needs and not their financial status. However, the free nature of the system has been taken for granted in recent years, with the NHS incurring significant costs whenever patients miss GP appointments. The NHS thus want to understand why this happens so that they can reduce missed appointments, which would be financially beneficial to them.

I began my analysis by importing the necessary libraries, Pandas and Numpy, before I imported the files 'actual_duration.csv', 'appointments_regional.csv' and 'national_categories.xlsx' into Python and which were respectively named 'ad, 'ar' and 'nc. For each of these three files I studied their metadata and shape of data, as well as checking for any missing data. None of the files had any missing values. The below table uses the 'ad' data set as an example to show the output when running code to check missing data: 8 columns but 0 rows, indicating no missing values.



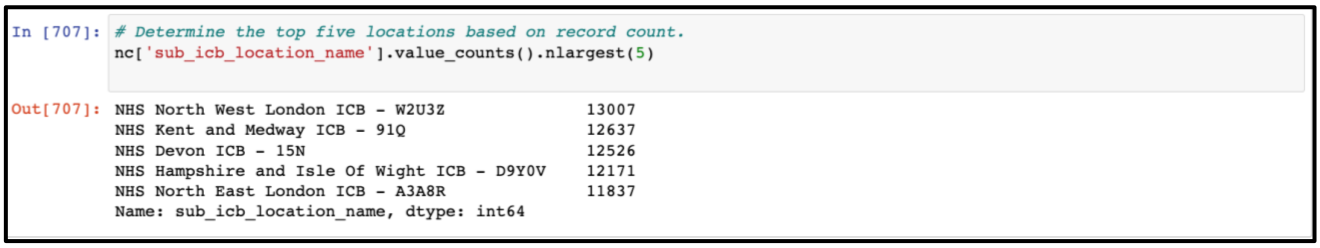
```
In [694]: # Determine whether there are missing values.
          # Creating a DataFrame which contains only the rows with missing data.
          ad_na = ad[ad.isna().any(axis=1)]
          # Printing shape of the new dataframe, ad_na.
          print(ad_na.shape)
          # Printing the DataFrame.
          ad_na

          (0, 8)

Out[694]:
```

sub_icb_location_code	sub_icb_location_ons_code	sub_icb_location_name	icb_ons_code	region_ons_code	appointment_date	actual_duration	count_of_appoint

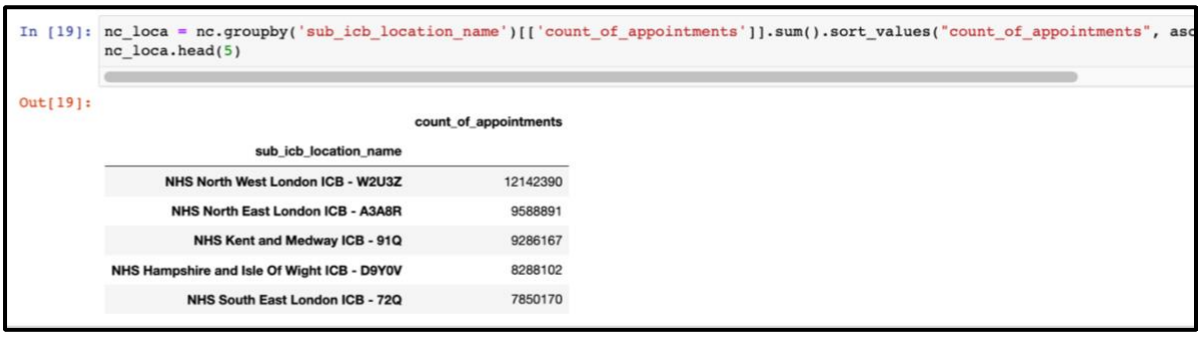
Using a value.counts() function, I studied the 'nc' dataset to determine which locations were most frequent in terms of records.



```
In [707]: # Determine the top five locations based on record count.
          nc['sub_icb_location_name'].value_counts().nlargest(5)

Out[707]: 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
```

I then ran a groupby function on the 'nc' dataset to group the different locations and the number of appointments to find the location with the most appointments. The output I received is shown in the table below.



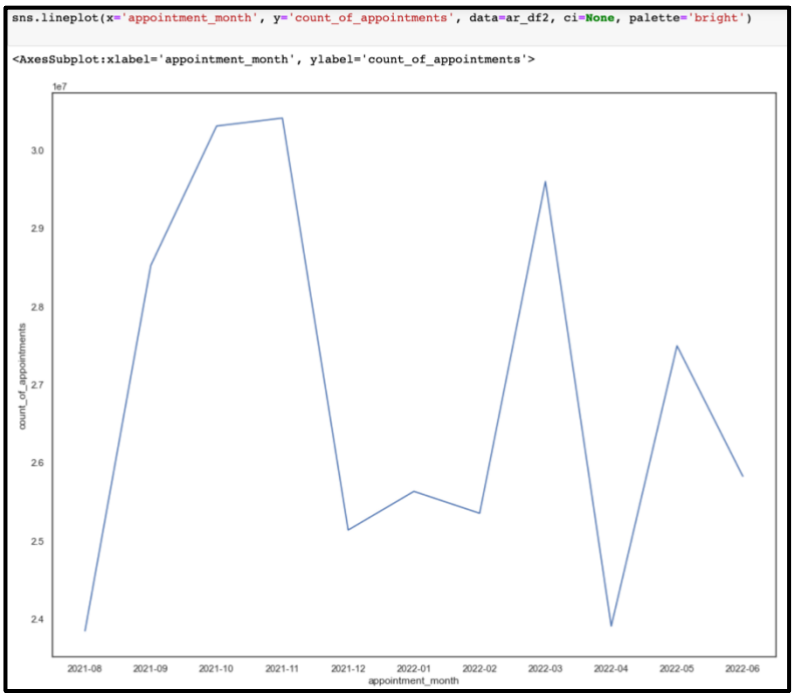
```
In [19]: nc_loca = nc.groupby('sub_icb_location_name')[['count_of_appointments']].sum().sort_values("count_of_appointments", asc
         nc_loca.head(5)

Out[19]:
```

	count_of_appointments
sub_icb_location_name	
NHS North West London ICB - W2U3Z	12142390
NHS North East London ICB - A3A8R	9588891
NHS Kent and Medway ICB - 91Q	9286167
NHS Hampshire and Isle Of Wight ICB - D9Y0V	8288102
NHS South East London ICB - 72Q	7850170

Comparing both outputs, we can see that ‘NHS North West London ICB - W2U3Z’ is the most frequent in terms of records and also has highest total number of appointments. However, looking at the first table we can see that ‘NHS North East London ICB - A3A8R’ is the fifth most frequent in terms of records but the second highest in terms of total appoints, whereas for example ‘NHS Devon ICB – 15N’ is third most frequent in terms of records but doesn’t even appear in the top 5 for most appointments. What we can infer from this is that there must have been certain months in the timeframe in which there were an increased number of appointments booked at certain locations compared to others. I attempted to merge the ‘nc’ DataFrame with the ‘ar’ DataFrame to determine the number of appointments attended vs the number not attended but was repeatedly faced with dead kernels.

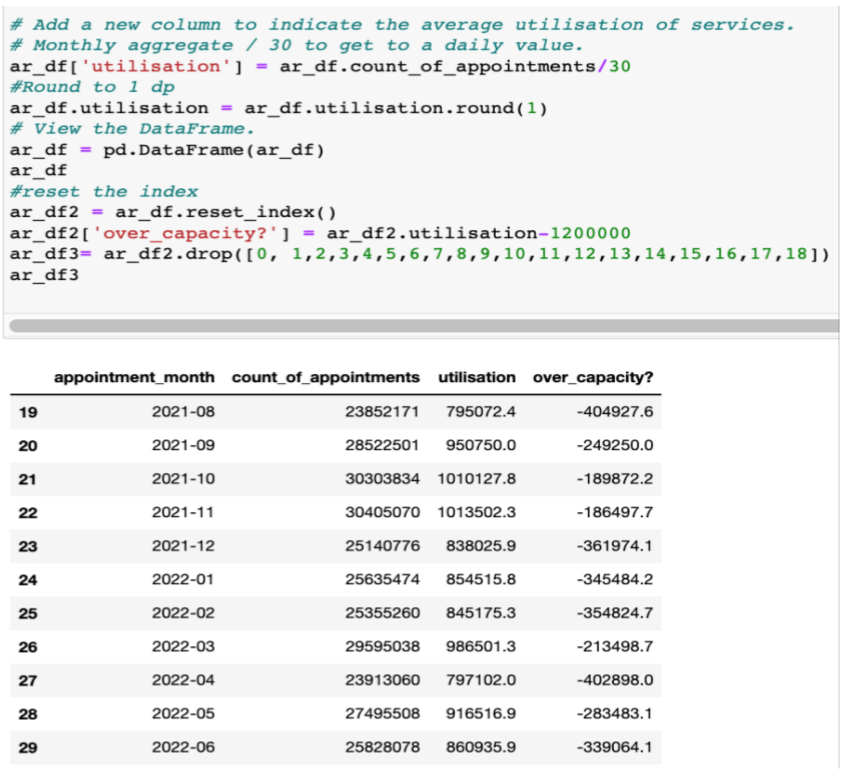
The graph below shows the count of appointments across the period 08/21 to 06/22.



```
sns.lineplot(x='appointment_month', y='count_of_appointments', data=ar_df2, ci=None, palette='bright')
```

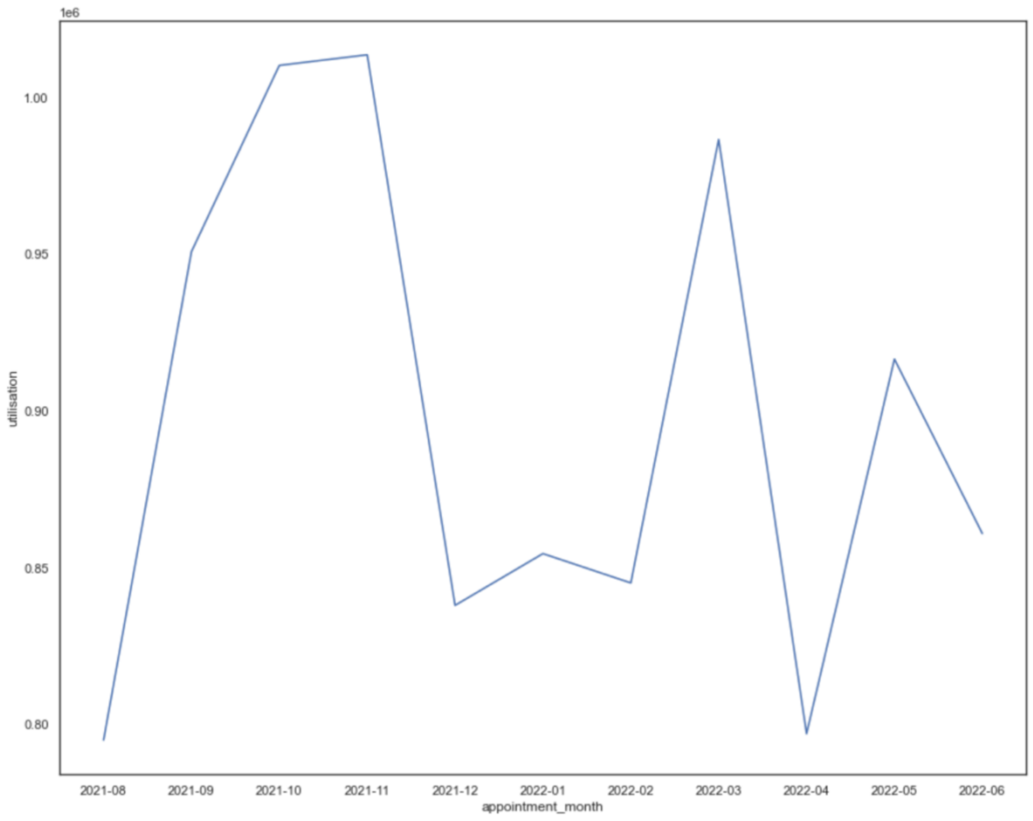
```
<AxesSubplot:xlabel='appointment_month', ylabel='count_of_appointments'>
```

Studying the graph, we can see that most appointments were made between October 2021 and November 2021. This rise can be seen as potentially being caused by rises in cases of Omicron, a variant of corona virus, with the virus being more lethal in Winter. In April 2022, there is a trough of the graph which can be accredited to the NHS delivering antivirals to patients at home, lowering the need for appointments to be booked.



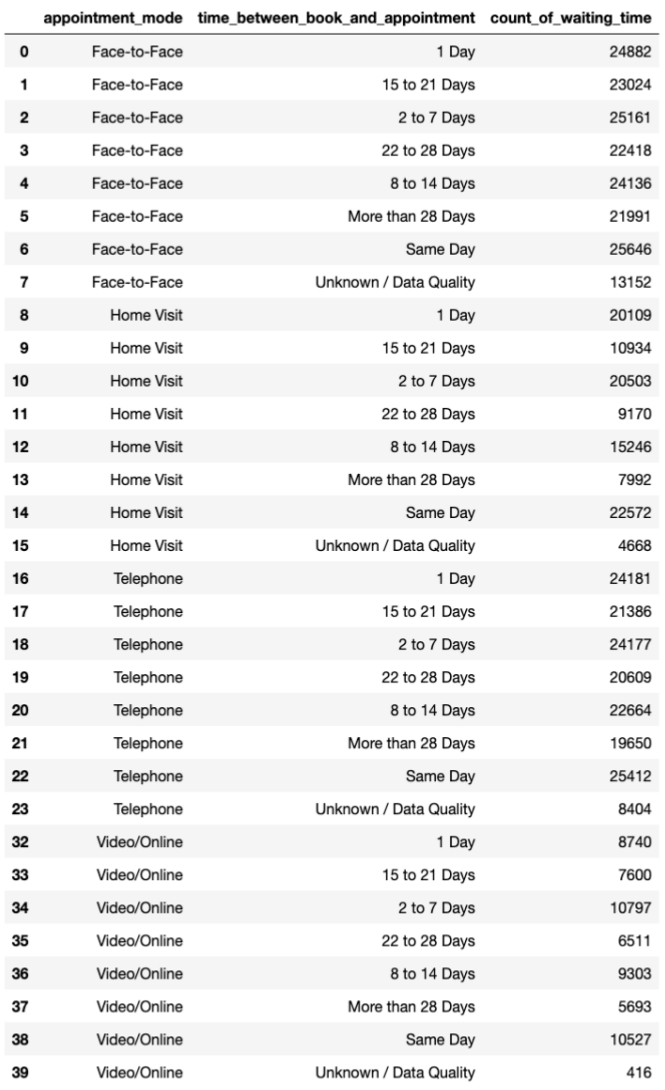
```python
# Add a new column to indicate the average utilisation of services.
# Monthly aggregate / 30 to get to a daily value.
ar_df['utilisation'] = ar_df.count_of_appointments/30
#Round to 1 dp
ar_df.utilisation = ar_df.utilisation.round(1)
# View the DataFrame.
ar_df = pd.DataFrame(ar_df)
ar_df
#reset the index
ar_df2 = ar_df.reset_index()
ar_df2['over_capacity?'] = ar_df2.utilisation-1200000
ar_df3= ar_df2.drop([0, 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18])
ar_df3
```

	appointment_month	count_of_appointments	utilisation	over_capacity?
19	2021-08	23852171	795072.4	-404927.6
20	2021-09	28522501	950750.0	-249250.0
21	2021-10	30303834	1010127.8	-189872.2
22	2021-11	30405070	1013502.3	-186497.7
23	2021-12	25140776	838025.9	-361974.1
24	2022-01	25635474	854515.8	-345484.2
25	2022-02	25355260	845175.3	-354824.7
26	2022-03	29595038	986501.3	-213498.7
27	2022-04	23913060	797102.0	-402898.0
28	2022-05	27495508	916516.9	-283483.1
29	2022-06	25828078	860935.9	-339064.1

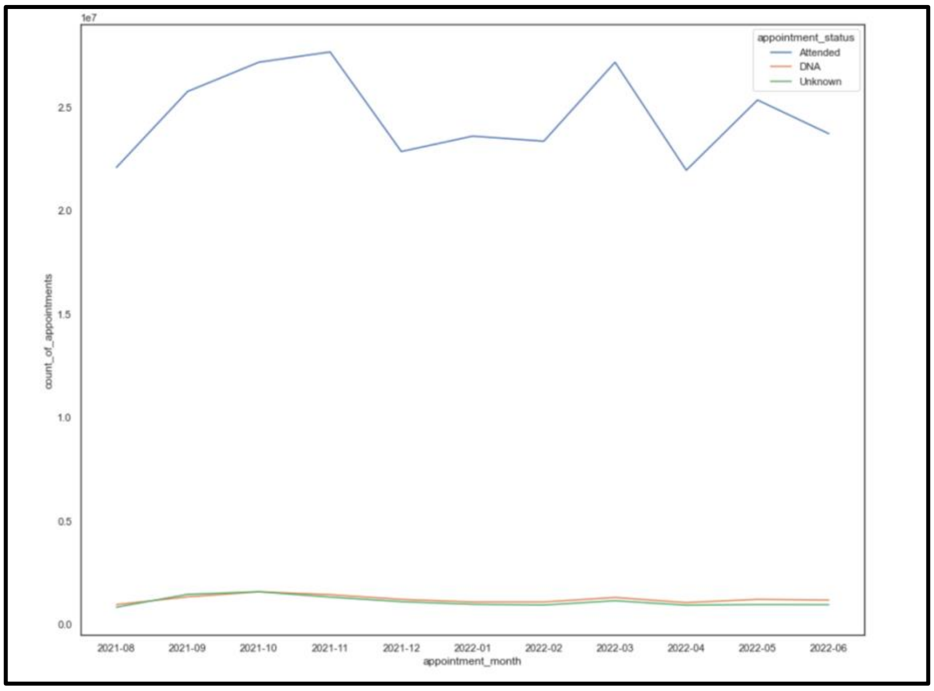


The graphs on the previous page exhibit the average utilisation of resources. The column 'utilisation' was created by dividing the 'count of appointments' column by 30 to get a daily average of utilisation. I then created another column called 'Over Capacity?' by taking away 1,200,000 from each of the values in the utilisation column. The reason we use the value 1,200,000 is because it is the maximum number of appointments that the NHS can accommodate nationwide in one day. The closer the values in the 'Over Capacity?' column to zero, the greater the utilisation of resources. For example, if we look at November 2021, the utilisation graph is at its peak whilst the value in the 'Over Capacity?' column is also the lowest. The graph clearly shows that the NHS was able to accommodate for all appointments across the months.

I conducted further analysis by grouping together the different types of appointments and the different waiting times for each appointment. The output table is shown below. Across all types of appointments, the most frequent time between book and appointment is 'Same day' (aside from Video/Online). However, this is closely followed by '2 to 7 days' being the second most frequent waiting time. This could be a possible explanation for the increased number of missed appointments as the patient could overcome the illness/injury themselves in that waiting period, thus rendering no need to attend the appointment.

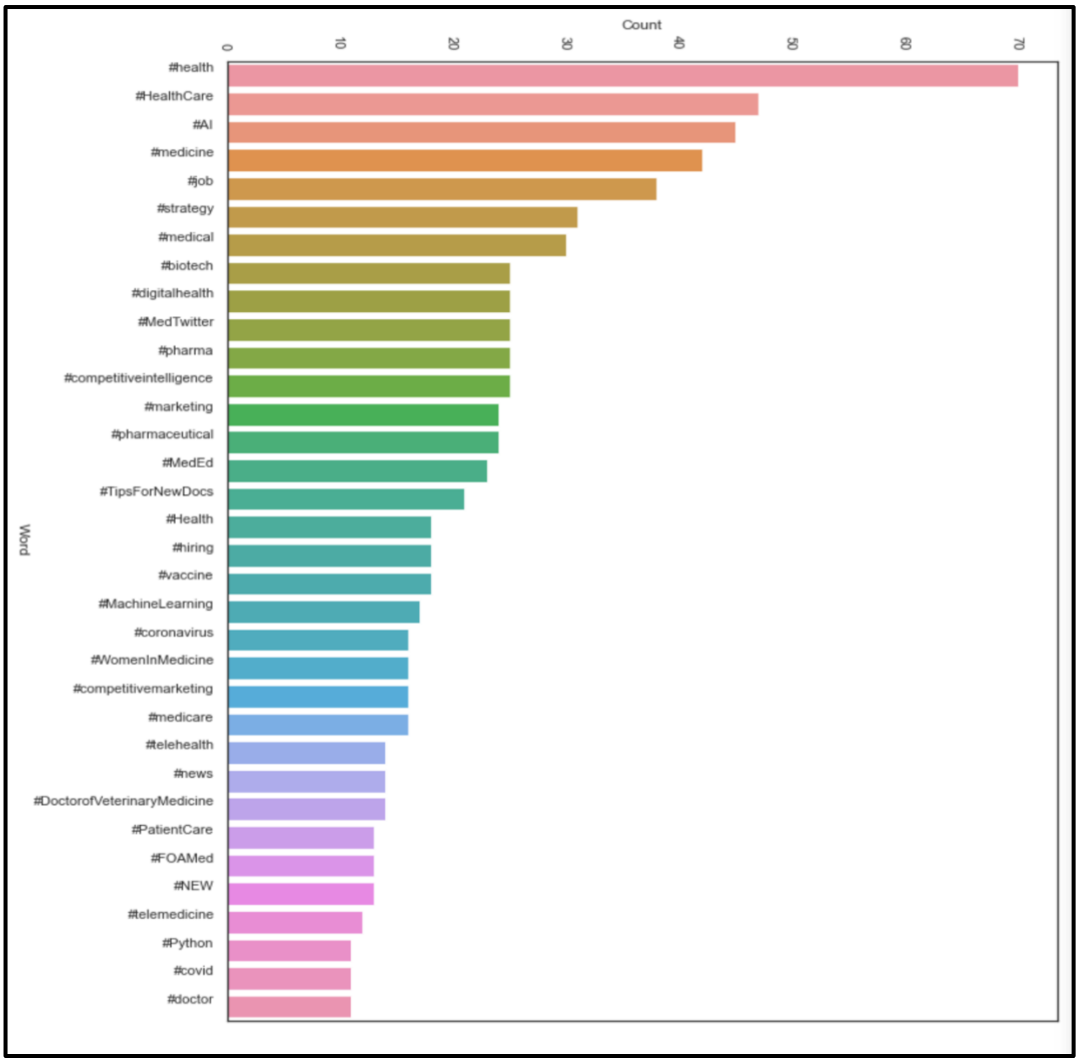


	appointment_mode	time_between_book_and_appointment	count_of_waiting_time
0	Face-to-Face	1 Day	24882
1	Face-to-Face	15 to 21 Days	23024
2	Face-to-Face	2 to 7 Days	25161
3	Face-to-Face	22 to 28 Days	22418
4	Face-to-Face	8 to 14 Days	24136
5	Face-to-Face	More than 28 Days	21991
6	Face-to-Face	Same Day	25646
7	Face-to-Face	Unknown / Data Quality	13152
8	Home Visit	1 Day	20109
9	Home Visit	15 to 21 Days	10934
10	Home Visit	2 to 7 Days	20503
11	Home Visit	22 to 28 Days	9170
12	Home Visit	8 to 14 Days	15246
13	Home Visit	More than 28 Days	7992
14	Home Visit	Same Day	22572
15	Home Visit	Unknown / Data Quality	4668
16	Telephone	1 Day	24181
17	Telephone	15 to 21 Days	21386
18	Telephone	2 to 7 Days	24177
19	Telephone	22 to 28 Days	20609
20	Telephone	8 to 14 Days	22664
21	Telephone	More than 28 Days	19650
22	Telephone	Same Day	25412
23	Telephone	Unknown / Data Quality	8404
32	Video/Online	1 Day	8740
33	Video/Online	15 to 21 Days	7600
34	Video/Online	2 to 7 Days	10797
35	Video/Online	22 to 28 Days	6511
36	Video/Online	8 to 14 Days	9303
37	Video/Online	More than 28 Days	5693
38	Video/Online	Same Day	10527
39	Video/Online	Unknown / Data Quality	416



Looking at lineplot above, it is instantly clear that the number of attended appointments is much higher than the number of attended. The count of appointments is actually quite volatile across the months. However, these were the stages in which the UK was coming out of the covid phase which could be a plausible explanation. If we look at November 2021, where total appointments were highest at 30,405,070, the attendance rate was at 90.1% whilst non-attendance was at 4.7% and 5.2% being unknown. This is actually a high attendance rate, but there are other factors to consider such as seriousness of the cause for appointment. People are more likely to attend appointments for more serious illnesses such as covid compared to an appointment booked for common cold symptoms.

The horizontal barchart below shows the top trending hashtags on twitter that have been used in tweets related to the NHS and healthcare. The most common hashtag is '#health' followed by '#HealthCare' and '#AI'. The NHS could incorporate these hashtags into their tweets to post cost statistics of people missing appointments and their implications on the future of the NHS. Another interesting perspective would be to incorporate live data as trending hashtags relating to the NHS may have changed, and the marketing team in charge of marketing campaigns could change their hashtags to s adapt to changing trends on social media.



To conclude, the data shows that the NHS were never over full capacity. This is due to number of appointments daily always being lower than the NHS' maximum threshold. However, the data does also point at waiting times stretching from 2 to 7 days quite frequently across multiple forms of appointments. Reducing timeframes could be a start for the NHS, however in recent times it is well known that the staffing levels have been stretched from the lingering effects of the pandemic.

A suggestion would be to analyse the most popular locations, in terms of appointments booked, and check their attendance rate. Finding the location with the lowest attendance rate would allow the NHS to investigate further reasons behind, for example weak transport links that hinder patients from making it to their appointment.

Based on the most popular hashtags, the NHS' marketing team could utilise the hashtag #health to post statistics explaining the implications of missed appointments and their costs on the NHS.