My GitHub repositories.
'https://github.com/oroy40051/NHS-ANALYTICS-PROJECT'
'https://github.com/oroy40051/mypythonactivity'

Data set initial insights

Pandas was used for the import

All three data sets have no outliers or missing values.Data is clean. Descriptive statistics are only available for 'count_of_appointments' as the other variables are all categorical. There are 18 different national categories.
3 different context types:'Care Related Encounter' 'Unmapped' and 'Inconsistent Mapping'.

3 different appointment statuses: 'Attended' 'DNA'(did not attend) and 'Unknown'.

'nunique' function was used to determine number of unique 'categories' for the categorical variables.
'unique' function gave the names of the different
categories.

value_counts function was used to give the count of
the different locations and thus records.

The descriptive statistics were visualised for each data set using seaborn and a for loop.

Further exploration and insights

For the first question in order to find the maximum and minimum dates two tables were stacked using the concat function (ad and nc) as they contain the largest and correct range of date values.
The appointment date of the ad dataframe was then converted to the same format as the nc data frame
(YYYY-MM-DD) and was also converted into a datetime
object using the datetime module allowing for better analyis.
For the second question we filtered the nc dataframe to get the values from 1st of January to 1st of June 2022. We then used a lambda function to find locations containing the specific string of 'NHS North West London ICB - W2U3Z'. We then used the groupby function to determine the

number of records for each service setting with the given location and time period. We then sorted by the largest number of records first to get the correct data needed.

For the third question, we used the groupby function with the appointment months to find the month with the highest number of appointments. We see that November 2021 had the highest number of appointments.

For the fourth question we filtered the ar dataset to show the appointment month. We then used the value_counts function to find the number of records for each month.

Visualizations and trends

We then employed seaborn to visualise the total number of appointments per month by creating a time-series with three variables, where in each case the x-axis was the appointment month, the y-axis was the total number of appointments and a comparison between the different categories was done by a different colour for each category (i.e different colours for the different service settings).

For the service settings we see that the most appointments are with the General Practice by far with peaks at Nov-2021 and Mar-2022.

For the context types we see that the most common context type is care related encounters. Inconsistent and unmapped encounters are much lower which is a positive in regards to error control.

Care related encounters also have peaks at Nov-2021 and Mar-2022.

For the national categories General Consultation, Acute and group education are the top three respectively following a similar peak pattern as the service settings and context types.

We then dived deeper into the service settings by looking at the number of appointments per season.
We used the group by function to group by appointment date and appointment aggregated on the sum of the count of appointments. Clearly General Practice is the most common once again. It would be better to do an analysis on the rest of the service settings separately from the general practice for better insights.

We create time series plots filtering out the general practice setting type.

Now we have some interesting insights. All the seasons are very similar except Autumn of 2021. Here we see a rapidly fluctuating time series plot and an overall decreasing trend as the month progresses.
We can see that the NHS hits high levels of appointments during this time, which could be due to the increase in COVID-19 cases in this season.

Unmapped appointments is the 'best of the rest' and once again this is something that needs to be address with regards to error control and data management.

Analyzing NHS related tweets

We import the tweets data using panda and explore the metadata using the BeautifulSoup module by converting it to a beautiful soup object and then applying the prettify function.

We explore the number of retweets and favorites column individually. Once again using the robust value_counts function we see the word healthcare has the most retweets and favourites, however the word health is second and closer than in other catgories in the retweets column.

There is not much use in exploring these columns individually as they are a metric of popularity and display the same information as the count of tweet hashtags which we explore anyway. Also without the corresponding word there is no use of the columns individually.

We then create a dataframe consisting only of string values and use a for loop to find all the word with the '#' symbol. We then convert the resulting series into a dataframe using pandas, rename the columns and prepare for analysis.

We create a barplot using seaborn to find the hashtags with more than 30 counts. The resulting bar plot is very cluttered and hard to intrepret.
We thus decompose the bar plot.
We create a barplot for the last 5 values. Interestingly we see drugs and nurses are present, this could indicate the lack of awareness of health problems due to recreational drug use aswell as the under-recognition of the efforts of nurses in the NHS.

We see that after the 'meded' hashtag the counts are very similar so we explore the 'middle values' to be able to obtain insights of any value.

We then create a plot for the hashtags with more than 30 counts and see healthcare is by far the greatest.
This is a positive outlier thus we create a top 5 plot excluding this. Health is a expected synonym of the word healthcare. Interesting insights we find are the words 'ai' and 'job' which

could indicate the boom in machine learning and artificial intelligence use in the medical industry. The word 'job' could indicate the high demand for work in the healthcare or medical industry.

Recommendations and summary

We see quite clearly that same day appointments are significantly more common than other ones. As the time between booking and appointment increases the number of appointments decreases. This indicates a negative correlation but not necessarily a causation. We however mention correlation with caution as the number of days is treated as a categorical variable instead of a numerical one. In order to reduce workload during busy months we can reduce or stop same day appointments in the given busy time-period.

.

Looking at the time-series plot for the number of appointments we see a generally strong positive trend between the appointment months and the number of appointments. This needs to be addressed as the rapid growth may overwhelm the NHS.

Overall Recommendations from the complete analysis:

- Increase staff in:
  NHS Norfolk and Waveney ICB - 26A
  NHS Kent and Medway ICB - 91Q
  NHS North West London ICB - W2U3Z
  NHS Bedfordshire Luton and Milton Keynes ICB - M1J4Y
  NHS Greater Manchester ICB - 14L
- Increase staff in November and months
- Account for high workload frequency in Autumn months
- Hire more General Practitioners
- Encourage patients to do online or telephone appointments for general consultations
- Improve data handling, error control, and general administration to reduce unmapped and inconsistent appointments.
- Create more awareness of the dangers of drug use
- Try and improve recognition of Nurses in the NHS
- Attempt to create more jobs in the NHS that use Artificial intelligence.
- Increase overall staff over staff due to increasing trend in time-series.

- Do not allow same-day appointments or switch to only non in person Appointments during busy months.
- Re-distribute staff from the less busy Holiday period into The more busy ones such as Autumn and Spring.