Github Repository link: https://github.com/thoppdev/LSE_DA_NHS_WTL

Final Assignment

The business problem: The NHS would like to identify the reason behind the missed appointments throughout the country and therefore reducing the enormous financial implications it generates. The immediate questions the NHS poses are of : a). Adequate staffing in the networks and b). actual utilisation of the resources.

The data analysis process. To better understand the problem, data analysis on the existing data sets are imperative to draw out any patterns, inferences, and finally any suggestion in addressing the issues. There are Four different datasets that are being analysed. The respective datasets are first loaded onto the Python platform, Pandas library for analysis. The components involved in this analysis cycle are Data Ingestion, data wrangling, visualisations of the data, Twitter API access and analysis.

Firstly, the necessary python libraries are imported. Pandas and NumPy libraries for data analysis, Seaborn, Matplotlib for visualisations, Twitter API access files. The respective files are then loaded on to pandas and creating the appropriates DataFrames for further analysis. This explains the step of Data Ingestion.

Secondly, the operations including describe, shape, dtypes, head, info etc are used in the process of Data Wrangling. The usage of groupby and aggregate function along with the statistical methods of, count, mean, sum, agg, value counts, sort_values are extensievely used in the data wrangling process. The main Three datasets: 'actual_duration.csv', 'national_categories.xlsx' and 'appointments_regional' are sense-checked for any errors, missing values. Then the 'describe ()' is used on the integer columns to get the statistical data.

As the datasets were not in a normalised status, in spite of having a common code 'icb_ons_code' and data integrity constraint(duplicates for primary key) , I did not apply the dataframe method 'merge' on the datasets.

Investigations:

The number of locations, service settings, context types, national categories, and appointment statuses are calculated. This achieved by creating dataframes and performing operation like groupby- based on the necessary fields and aggregate to find the 'sum' and 'counts'.

The date range has been found the service setting reported the most appointments are calculated.

There were 106 total number of locations. The topmost location was NHS North West London ICB - W2U3Z.
The top five locations are:

NHS Northwest London ICB - W2U3Z
NHS Kent and Medway ICB - 91Q
NHS Devon ICB - 15N
NHS Hampshire and Isle Of Wight ICB - D9Y0V
NHS North East London ICB - A3A8R

Then the total number of Service settings were calculated.

The top five locations are:

service settings: General Practice	359274
Primary Care Network	183790
Other	138789
Extended Access Provision	108122
Unmapped	27419

The Context Types were : Care related encounter, Inconsistent mapping and unmapped.
Within the National Categories, there were 18 overall categories. The main ones being Inconsistent mapping, General Consultation Routine, General consultation acute, planned clinics, clinical triage.
The total appointment statuses were, attended (232137), Unknown (201324) and DNA (163360).
The appointment modes were, Face-to-face, telephonic, home visit, Unknown, video /online.

The date ranges were calculated. The appropriate date conversion methods were used in capturing this data. Further up, the number of appointments for the six month period (from 01/01/2022 to 01/06/2022), and ‘sub_icb_location_name ‘ = ‘NHS North West London ICB – W2U3Z’ was calculated.
Of all the data, NHS north East London ICB- A3A8R, which is ‘General :Pracitce’ turns out with maximum appointments of 16527.
Next, total monthly appointments were calculated. The date range of the provided data sets and which service settings reported for the period of 01/01/2022 to 01/06/2022 were calculated as above.
I have calculated the month which had the highest number of appointments- September 2021, 24561648
The breakdown of total number of records have been calculated.

Various line plots have been plotted to find out in identifying service setting, context type and national categories. General practice had the highest number of appointments compared to others. In the Context type, Care related encounter had the maximum appointments.
I have filtered the data based on four seasons into separate dataframes and have used the lineplot logic to plot the graph but unfortunately it did not work for me.

From the point of analysis, I have undertook, it gives me the pattern of umapped, unknown, inconsistent mapping on records in the system.