The datasets that are going to be used are:

Dutch airports: <https://opendata.cbs.nl/statline/#/CBS/en/dataset/37478eng/table>

https://opendata.cbs.nl/statline/#/CBS/en/dataset/37478eng/table?dl=9CC7C

Ireland airports: <https://ws.cso.ie/public/api.jsonrpc?data=%7B%22jsonrpc%22:%222.0%22,%22method%22:%22PxStat.Data.Cube_API.ReadDataset%22,%22params%22:%7B%22class%22:%22query%22,%22id%22:%5B%5D,%22dimension%22:%7B%7D,%22extension%22:%7B%22pivot%22:null,%22codes%22:false,%22language%22:%7B%22code%22:%22en%22%7D,%22format%22:%7B%22type%22:%22JSON-stat%22,%22version%22:%222.0%22%7D,%22matrix%22:%22TAM07%22%7D,%22version%22:%222.0%22%7D%7D>

Process draft (descriptive for understanding a predictive for prediction.

Both databases are loaded into jupyter.

Netherlands Database:

Requires cleaning for the information. Information is divided per quarters, years and months. Quarter information is dropped as the analysis will be done per month. Cleaning of column titles is done and information divided by month.

Understanding the Total numbers throughout the year:

Barplots are created for the total of dutch airports per year, where it is visible that the outliers identified later correspond to the 2020 pandemic. There barplots are developed as an initial insight into the behaviour of the data throughout the years.

Insights: From these initial barplots, it is evident that the general number of flights and passengers dropped heavily during the 2020 pandemic and during the next couple of years, but the cargo transported didn´t have such a heavy impact during this period.

Description of all the features is obtained, in order to validate the values of the boxplots to be developed, and validate that the outliers are indeed extreme values.

These bpxlots are developed and outliers identified. As described earlier, most of these outliers can be tracked to months corresponding to the 2020 pandemic and the subsequent recovery of world travel.

Most outliers are dropped

Heat map developed to see relationship between features. From heatmap it is possible to see that there is a relationship mainly between the total numbers and the scheduled either flights, passengers and cargo. This could mean that the total numbers of these 3 are almost exclusively dependent on scheduled features, with unscheduled being not closely related to the total number.

It is important to define scheduled and non-scheduled services…

Ireland database:

Database is loaded and similar division is identified: Flights, passengers and Freight (cargo).

Information comes per month and year, but in one cell, therefore the column is divided and the month is translated into a categorical value.

Countries come divided differently to the Netherlands database. Therefore, the analysis will be made using the total countries and total airports in the country mirroring the previous dataframe.

The Irish dataframe was pivoted to represent the information in the same format as the Netherlands one, the names of the columns were modified to match the Netherlands Dataframe te, for this it is assumed that:

Cargo is equivalent to Freight in the Irish database.

Unloaded cargo is equivalent to arrival cargo and loaded cargo to departure cargo.

Commercial flights are equivalent to total flights

Same initial analysis and outlier elimination is carried out as in the Netherlands database.

Similar to the Netherlands dataframe, with the help of a heatmap, it is evident that the total number of flights, passengers and cargo are dependent on the scheduled features, with almost no relation to the non-scheduled ones. There is also a relation between the flights and passengers as expected.

Creating a scatter plot, we confirm that the passenger numbers are directly related to the number of flights, but not to the cargo transported.

Feedback CA1

Statistics

The code for describing and plotting the statistical features is good. Regarding binomial and normal distibution, they need to be improved. The graph that you plotted for a normal distribution does not match its properties. There is no proper difference between the two distributions highlighted.

Data Preparation

-10% late submission Good work! A lot of your report detailed what you did but unfortunately was quite light in explaining why you did it and how it helped you eg. You Engaged in Basic EDA. How did it help specifically? You didnt check for outliers, malformed data, Why? You used a lot of Bar Plots where there other visualizations that would have been appropriate. Very little design discussion as requested also You did not make reference to Tufts Principles of design in Part 4.

Machine Learning

3644 words used including title page, TOCs and text in the report. The report is based on an abstract, introduction, EDA and other sections. In ML section, the student compared three project management frameworks and CRISP-DM methodology is chosen. ML approaches are discussed for modelling. Three ML models are compared in the Table on page 25. GridSearchCV is employed for hyperparameter tunning in the case of Lasso regression. Conclusions and references are provided at the end of the report. Jupyter notebook is provided for the practical implementation of ML models. GitHub is used during the development of this project and 10 commits performed. A more depth in the interpretation, similarities and differences of ML modelling results are required in this report. Overall the rationales are justified nicely.