Deliverable 1

Gabriele Di Palma

Francesco DI Stefano

Francesco Tramontano

3.1) We decided to select this dataset https://www.kaggle.com/datasets/parulpandey/us-international-air-traffic-data?select=International_Report_Departures.csv which originally consisted of two tables. However, we will only be utilizing the departures table as it is sufficient for our analysis.

The description of the dataset found in Kaggle states that: *The table contains data about all flights between US and non-US gateways, regardless of their origin or destination. Each observation in the table provides details about a specific airline and a pair of airports, one in the US and the other outside. The table includes three main columns that record the number of flights - Scheduled, Charter, and Total.*

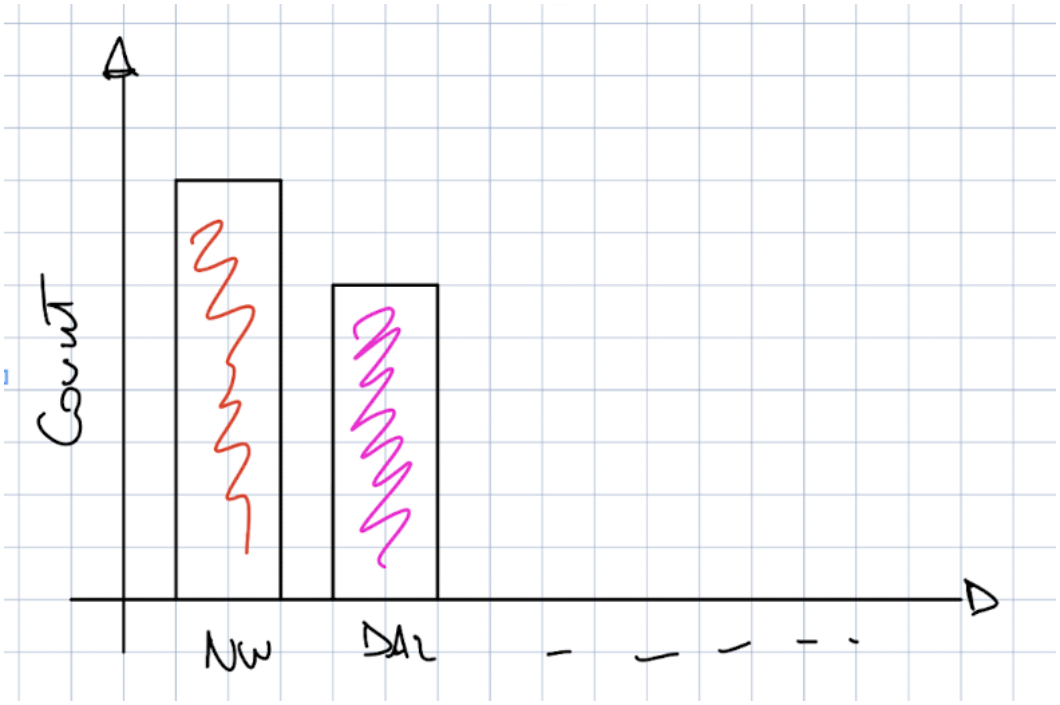
We thoroughly assessed the data for missing values, inconsistencies, and outliers, and eliminated all the rows containing missing values as they were minimal compared to the entire dataset. Although there exist some outliers, we retained them since they hold significance for our analysis. Furthermore, there are no instances of duplicates in the dataset.

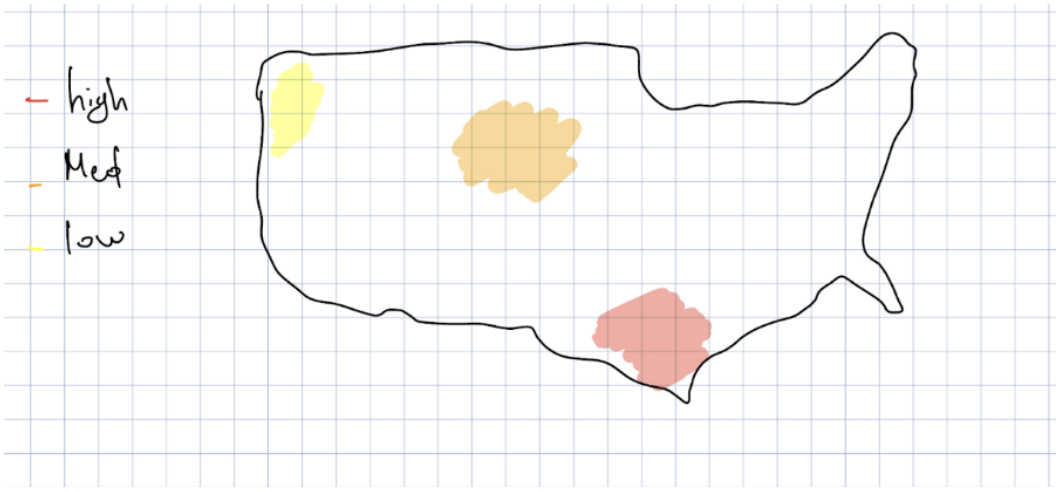
3.2) We propose conducting some relevant analyses including:

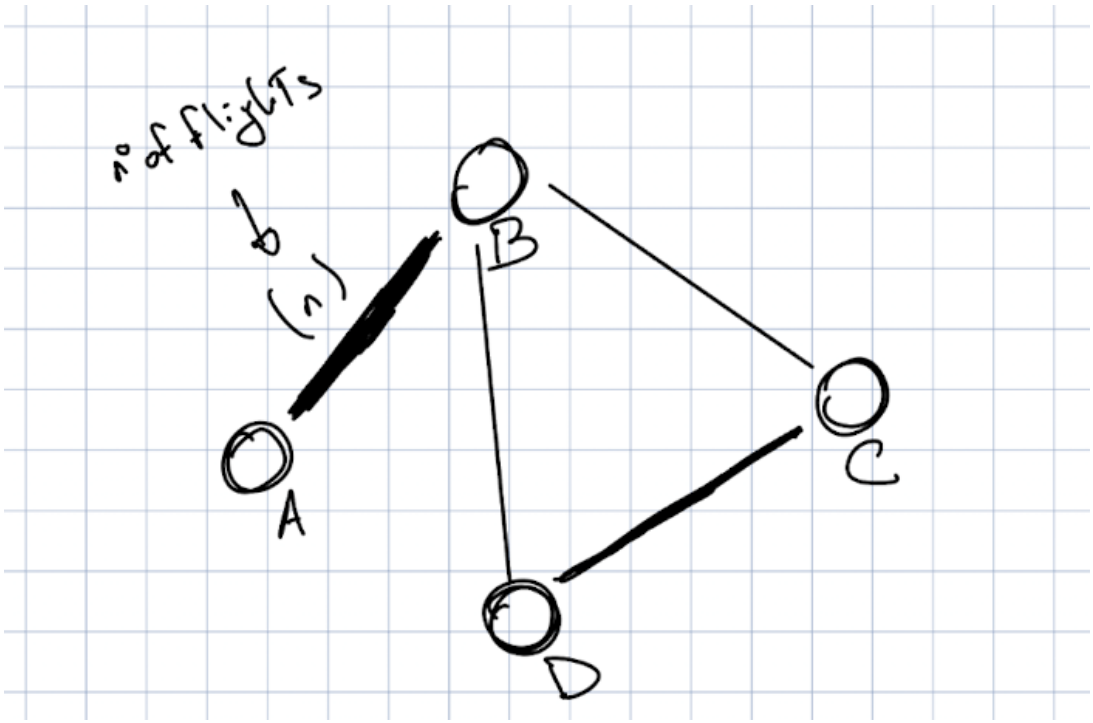
- Network analysis that examines the airports and their connections
- Descriptive statistics on airports (e.g. most popular destinations, number of flights per airport)
- Distribution of flights over some time periods.

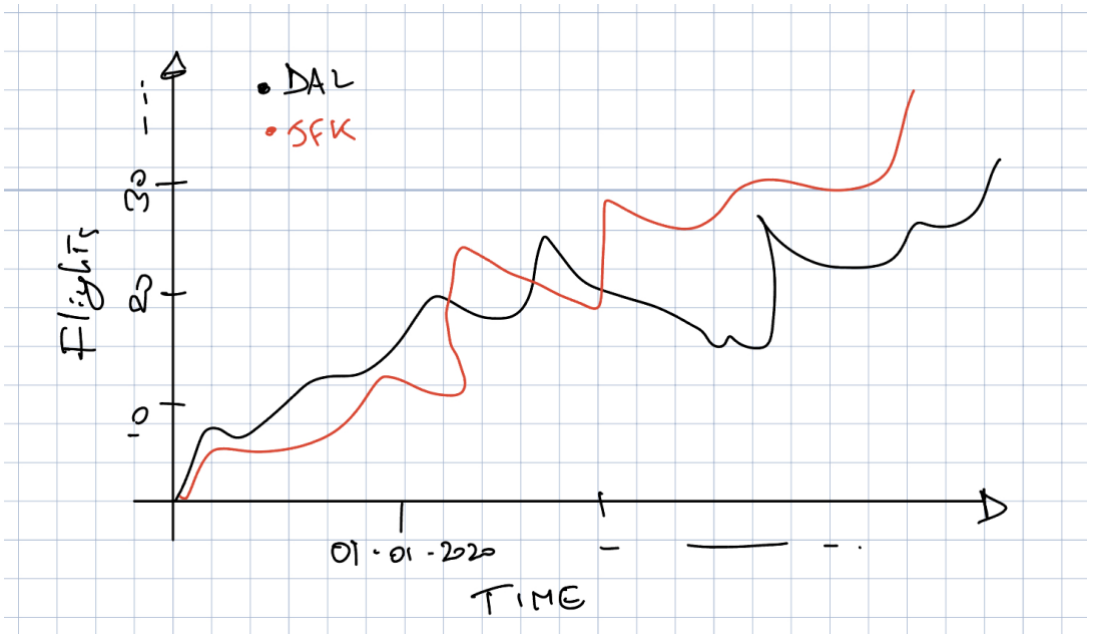
3.3) To accomplish the proposed analyses, we suggest utilizing the following visualizations. These visualizations will be implemented using various Python libraries.

1. Static barplot that showcases the top ten airports based on the total number of flights.
2. Dynamic geographical heatmap that highlights the distribution of airports.
3. Dynamic graph analysis that focuses on airports and their connections. In this visualization, airports will be used as nodes, and flights will be used as edges. Since the dataset is not respective of origin and destination, the edges will be undirected and weighted based on the total number of flights.
4. A static time series analysis that examines flights per airport over time.


Count
NW
SAL


high
Med
low


no of flights
b
(n)
A
B
C
D


DAL
SFK
Flights
30
20
10
01-01-2020
TIME