Deliverable 1

Gabriele Di Palma

Francesco DI Stefano

Francesco Tramontano

3.1) We decided to select this datset https://www.kaggle.com/datasets/parulpandey/us-international-air-traffic-data?select=International_Report_Departures.csv which originally was composed of two tables but we will only focus on the departures' one because it is enough for our purposes.

The table contains data on all flights between US gateways and non-US gateways, irrespective of origin and destination.
Each observation provides information on a specific airline for a pair of airports, one in the US and the other outside. Three main columns record the number of flights: Scheduled, Charter, and Total.

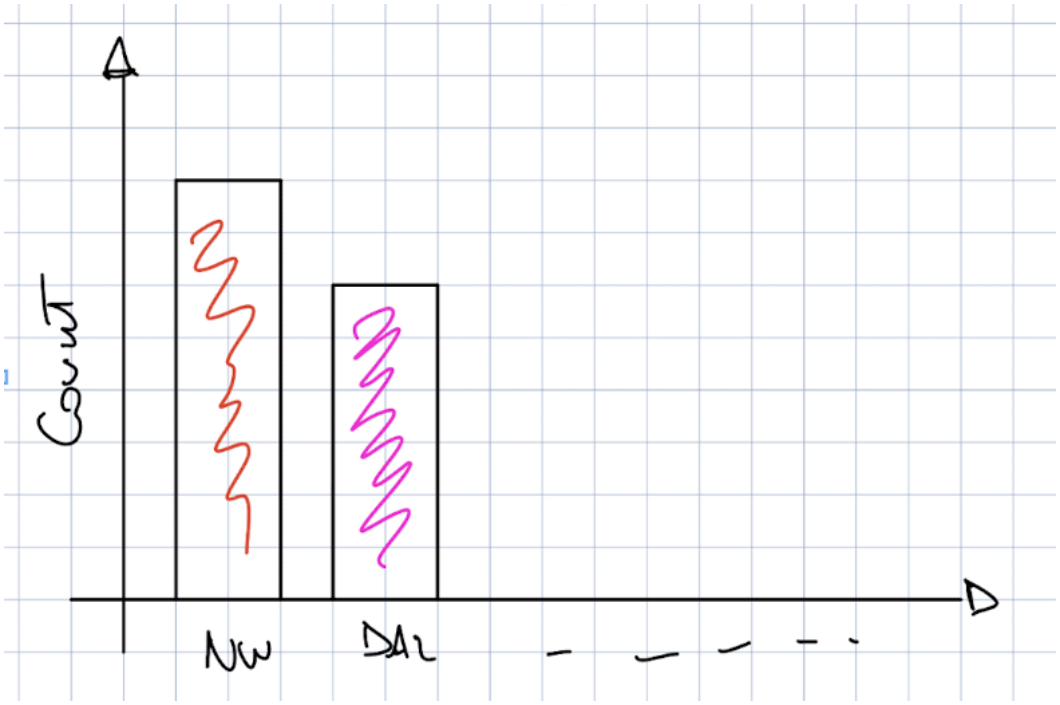
We have analysed missing values, inconsistencies and outliers, we deleted all the rows containing missing values since they were few compared to the entire data frame. There are some outliers but we kept them because they are meaningful for the analysis. There are no duplicates.

3.2) We believe that some relevant analysis might be:

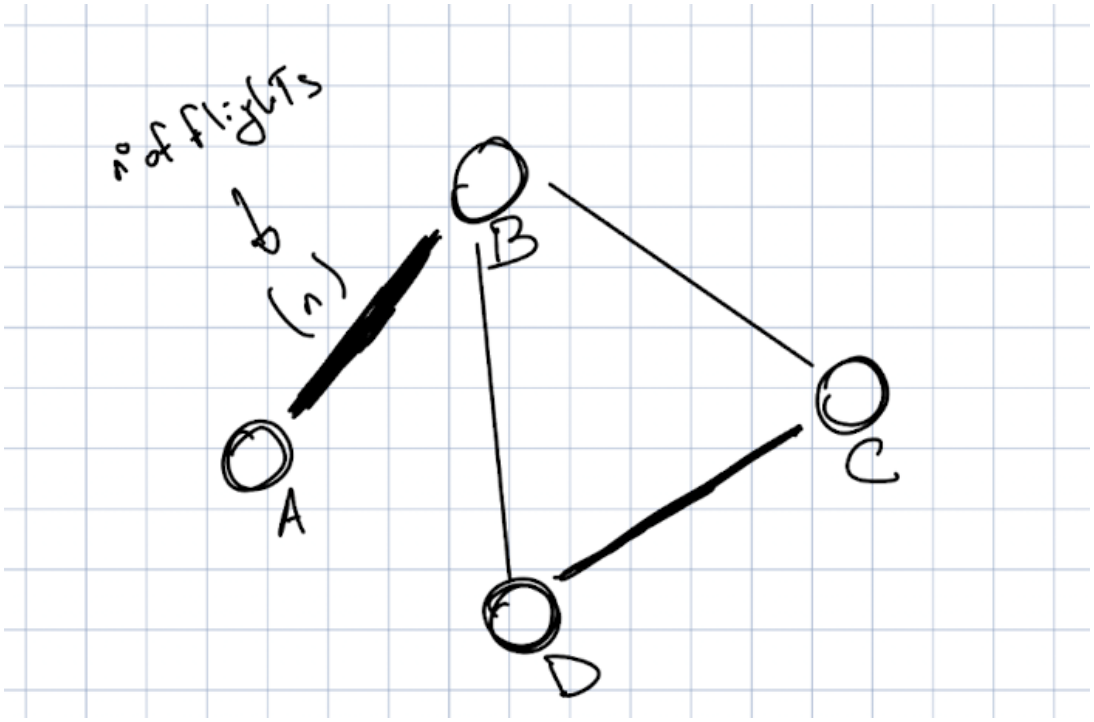
- Network analysis on the airports and connection between them.
- Descriptive statistics on airports (e.g. most popular destinations, number of flights per airports)
- Distribution of flights over some time periods.

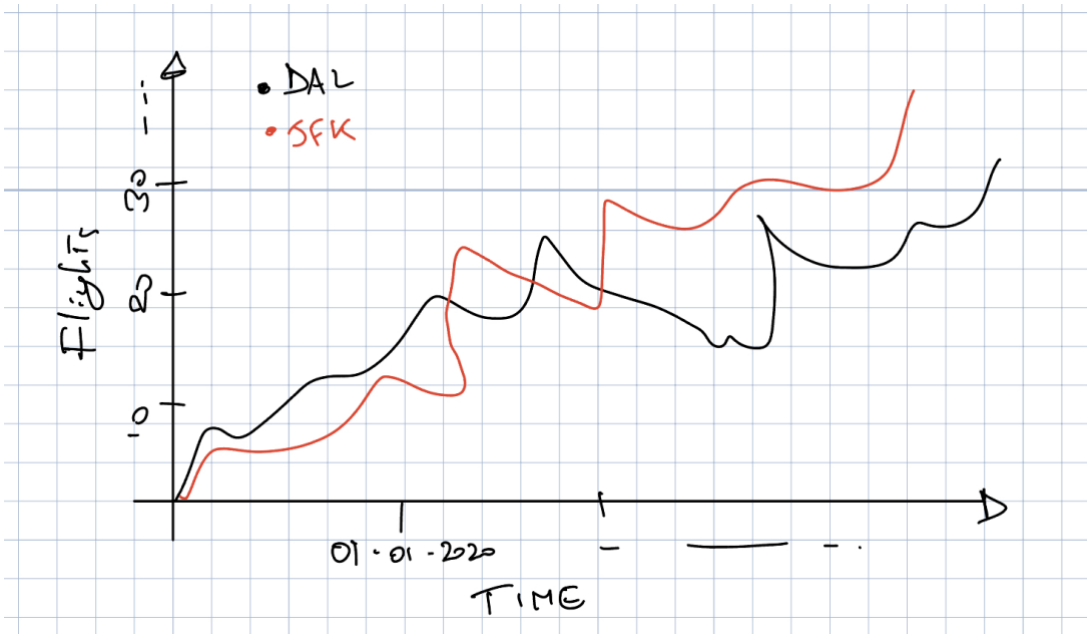
3.3) To accomplish the analysis specified before we thought that the following visualization might be helpful. We will use different Python libraries to implement them.

1. Static barplot of top ten airports based on the total number of flights
2. Geographical heatmap of the airports distribution (dynamic)
3. Dynamic graph analysis of airports and their connection, we will use airports as nodes and flights as edges (weighted on total flights, undirected since the df is not respective of origin and destination)
4. A time series analysis on flights per airport over time (static)


Count
D
NW
DA2


high
Med
low


no of flights
(1)
A
B
C
D


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