Raport for Flight Delay Predicion Project

Dominik Olejarz, Patryk Rybak, Maksymilian Wnuk

February 8, 2024

1 Problem Statement

Our goal is to predict flight delays and, in the future, expand the program to calculate the most probable delay values. Throughout our work, we will utilize a dataset containing flight information from 2017 to 2018, provided by the Bureau of Transportation Statistics in conjunction with the Weather API. The primary focus is to examine the correlation between weather conditions and flight delays.

Our approach centers around machine learning techniques, and we will employ statistical learning models using Python. The aim is to forecast flight delays and in future provide precise values of delays. We believe that our project can contribute to improving the travel experience by offering more accurate information about potential flight delays.

2 Data downloading

Getting data for our machine learning models requires two combined datasets, flights and weather:

2.1 Flights data

We are getting 2017-2018 flights data from `kaggle.com` website. It contains of columns:

- Airline
- Origin city
- Destnation city
- Arr time
- Cancelled
- Delay
- Distance

2.2 Weather data

Fetching data for weather requires more work than just downloading a csv file. We will be downloading it by making api requests to `https://www.visualcrossing.com/`. Thanks to Visual Crossing® company, we were able to make those requests in unlimited way. We contacted their support and then they gave us access to unlimited api key. We get columns:

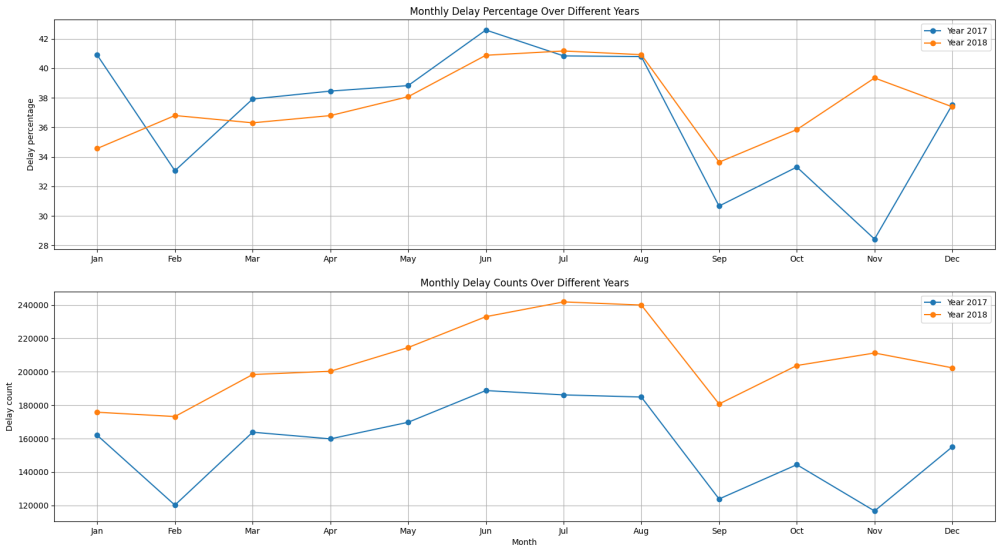
- Temperature
- Cloudcover
- Monphase
- Windspeed, windgust and winddir
- Snow and snowdepth
- Dew humidity
- Ice and freezing rain

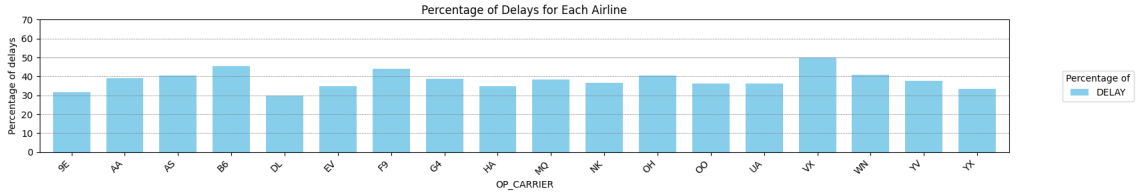
3 Merging data

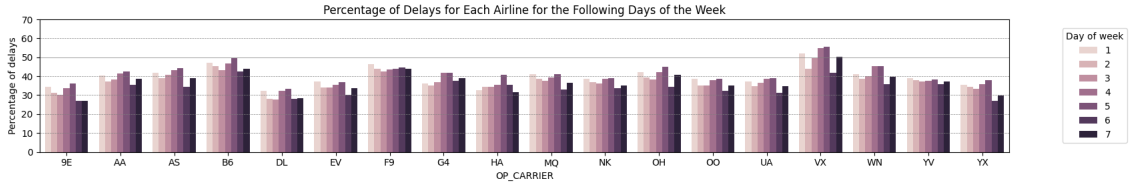
DOMINIK

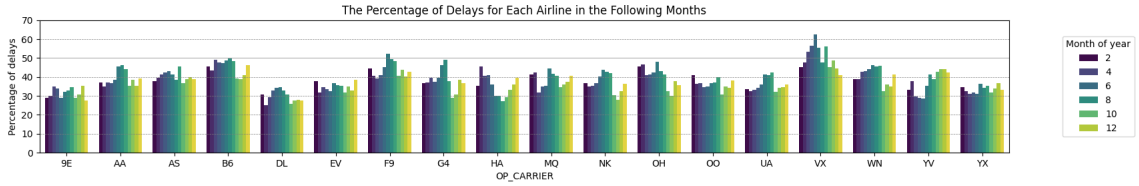
4 Plots

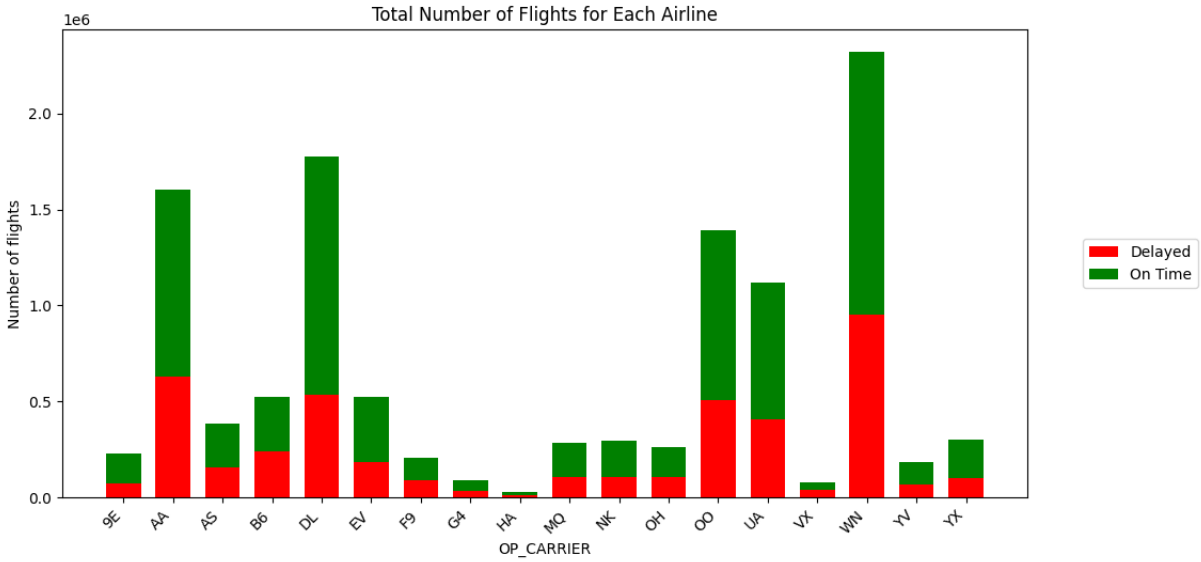
In this section we will look at plots describing data in our dataset.

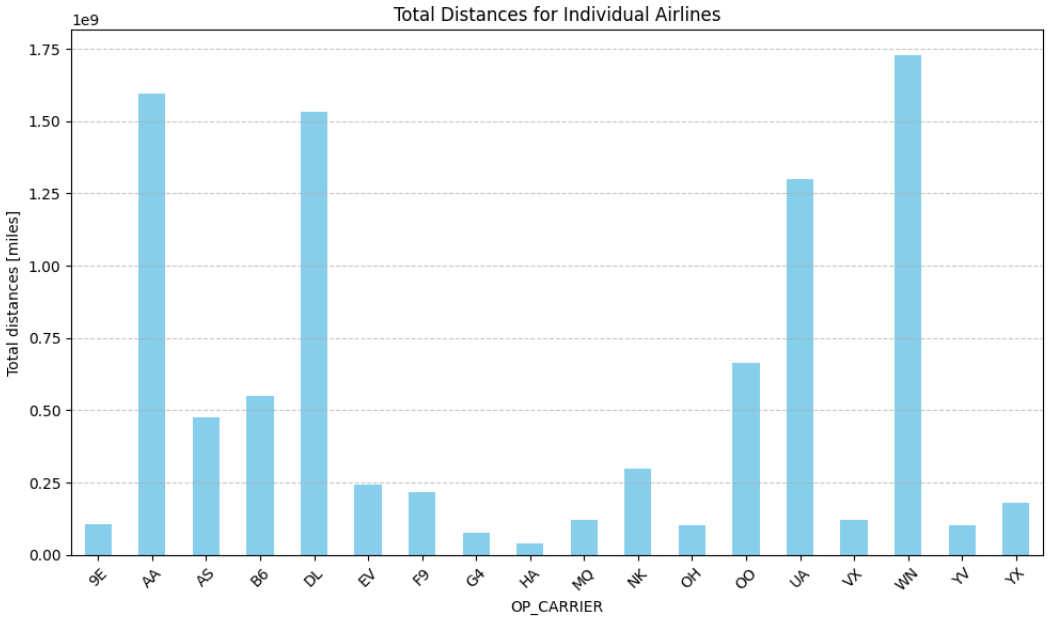



Percentage of Delays for Each Airline
Percentage of delays
OP_CARRIER
9E AA AS B6 DL EV F9 G4 HA MQ NK OH OO UA VX WN YV YX
Percentage of
DELAY


Percentage of Delays for Each Airline for the Following Days of the Week
Percentage of delays
OP_CARRIER
9E AA AS B6 DL EV F9 G4 HA MQ NK OH OO UA VX WN YV YX
Day of week
1
2
3
4
5
6
7


The Percentage of Delays for Each Airline in the Following Months
Percentage of delays
OP_CARRIER
9E AA AS B6 DL EV F9 G4 HA MQ NK OH OO UA VX WN YV YX
Month of year
2
4
6
8
10
12


Total Number of Flights for Each Airline
1e6
Number of flights
OP_CARRIER
9E AA AS B6 DL EV F9 G4 HA MQ NK OH OO UA VX WN YV YX
Delayed
On Time



5 Data preprocessing

Dataset contains of lots of NaN's. We removed all of them, by removing rows that contained them. Data loss was incredibly low, we lost about 10,000 of rows that contained at least one NaN, which is great. Additionally, we removed columns that contained A LOT of NaN's. Those columns were:

- Windgust -75.22% NaNs
- Snow - 8.13% NaNs
- Snowdepth - 8.13% NaNs
- ID,id - 55% and 44% respectively,no idea why so many nans, its just id's whose we don't even use in our model.

However, we are not sure if that was proper way of doing this, hence we ask ourselves: does snow affect flight delays? How about snowdepth and wingust? Later on we can always retrieve those columns and check correlation.

6 Machine learning part