Machine Learning Project EDA - Exploratory Data Analysis Report

Team :
Simon Duchesne 11297912
Piero Matos 11339060
Balogog Georges 11337767

March 15, 2024

1 Introduction

This report presents an exploratory analysis of the dataset containing flight information. The dataset consists of 32 columns/features, including flight date, airline information, departure/arrival details, delays, cancellation information, and more.

2 Data Overview

2.1 Columns/Features

The different columns/features in the dataset are as follows:

- FL_DATE (Flight Date: yyyy-mm-dd)
- AIRLINE_CODE (Reporting Airline: Unique Carrier Code)
- DOT_CODE (DOT ID Reporting Airline: Identification number assigned by US DOT to identify a unique airline)
- FL_NUMBER (Flight Number Reporting Airline: Flight Number)
- ORIGIN (Origin Airport)
- ORIGIN_CITY (Origin City Name)
- DEST (Destination Airport)
- DEST_CITY (Destination City Name)
- CRS_DEP_TIME (CRS Departure Time: Scheduled departure time in local time)

- DEP_TIME (Actual Departure Time: Actual departure time in local time)
- DEP_DELAY (Departure Delay: Difference in minutes between scheduled and actual departure time)
- TAXI_OUT (Taxi Out Time: Time taken for taxiing out in minutes)
- WHEELS_OFF (Wheels Off Time: Time when the aircraft leaves the ground)
- WHEELS_ON (Wheels On Time: Time when the aircraft touches the ground)
- TAXI_IN (Taxi In Time: Time taken for taxiing in after landing in minutes)
- CRS_ARR_TIME (CRS Arrival Time: Scheduled arrival time in local time)
- ARR_TIME (Actual Arrival Time: Actual arrival time in local time)
- ARR_DELAY (Arrival Delay: Difference in minutes between scheduled and actual arrival time)
- CANCELLED (Cancelled Flight Indicator: 1 if flight is cancelled, 0 otherwise)
- CANCELLATION_CODE (Cancellation Code: Reason for cancellation)
- DIVERTED (Diverted Flight Indicator: 1 if flight is diverted, 0 otherwise)
- CRS_ELAPSED_TIME (CRS Elapsed Time: Scheduled elapsed time of flight in minutes)
- ELAPSED_TIME (Actual Elapsed Time: Actual elapsed time of flight in minutes)
- AIR_TIME (Air Time: Flight time in minutes)
- DISTANCE (Distance: Distance between airports in miles)
- DELAY_DUE_CARRIER (Carrier Delay: Delay due to carrier in minutes)
- DELAY_DUE_WEATHER (Weather Delay: Delay due to weather in minutes)
- DELAY_DUE_NAS (NAS Delay: National Air System delay in minutes)
- DELAY_DUE_SECURITY (Security Delay: Delay due to security in minutes)
- DELAY_DUE_LATE_AIRCRAFT (Late Aircraft Delay: Delay due to late aircraft in minutes)

2.2 Shape of Data

The dataset contains 3,000,000 rows and 32 columns.

2.3 Unique Values per Column

Column	Unique Values
FL_DATE	1704
AIRLINE	18
AIRLINE_DOT	18
AIRLINE_CODE	18
DOT_CODE	18
FL_NUMBER	7111
ORIGIN	380
ORIGIN_CITY	373
DEST	380
DEST_CITY	373
CRS_DEP_TIME	1384
DEP_TIME	1440
DEP_DELAY	1513
TAXI_OUT	179
WHEELS_OFF	1440
WHEELS_ON	1440
TAXI_IN	202
CRS_ARR_TIME	1435
ARR_TIME	1440
ARR_DELAY	1527
CANCELLED	2
CANCELLATION_CODE	4
DIVERTED	2
CRS_ELAPSED_TIME	640
ELAPSED_TIME	696
AIR_TIME	666
DISTANCE	1727
DELAY_DUE_CARRIER	1291
DELAY_DUE_WEATHER	812
DELAY_DUE_NAS	671
DELAY_DUE_SECURITY	172
DELAY_DUE_LATE_AIRCRAFT	958

Table 1: Unique Values per Column

2.4 Nans Values Proportion

Column	Nans	% Nans
FL_DATE	0.0	0.000000
AIRLINE	0.0	0.000000
AIRLINE_DOT	0.0	0.000000
AIRLINE_CODE	0.0	0.000000
DOT_CODE	0.0	0.000000
FL_NUMBER	0.0	0.000000
ORIGIN	0.0	0.000000
ORIGIN_CITY	0.0	0.000000
DEST	0.0	0.000000
DEST_CITY	0.0	0.000000
CRS_DEP_TIME	0.0	0.000000
DEP_TIME	77615.0	2.587167
DEP_DELAY	77644.0	2.588133
TAXI_OUT	78806.0	2.626867
WHEELS_OFF	78806.0	2.626867
WHEELS_ON	79944.0	2.664800
TAXI_IN	79944.0	2.664800
CRS_ARR_TIME	0.0	0.000000
ARR_TIME	79942.0	2.664733
ARR_DELAY	86198.0	2.873267
CANCELLED	0.0	0.000000
CANCELLATION_CODE	2920860.0	97.362000
DIVERTED	0.0	0.000000
CRS_ELAPSED_TIME	14.0	0.000467
ELAPSED_TIME	86198.0	2.873267
AIR_TIME	86198.0	2.873267
DISTANCE	0.0	0.000000
DELAY_DUE_CARRIER	2466137.0	82.204567
DELAY_DUE_WEATHER	2466137.0	82.204567
DELAY_DUE_NAS	2466137.0	82.204567
DELAY_DUE_SECURITY	2466137.0	82.204567
DELAY_DUE_LATE_AIRCRAFT	2466137.0	82.204567

Table 2: NaN Values Proportion

2.5 Statistics of the Dataset

Column	Count	Mean	Std
DOT_CODE	3.000×10^6	1.998×10^4	3.773×10^2
FL_NUMBER	3.000×10^6	2.512×10^3	1.747×10^3
CRS_DEP_TIME	3.000×10^6	1.327×10^3	4.859×10^2
DEP_TIME	2.922×10^6	1.330×10^3	4.993×10^2

DEP_DELAY	2.922×10^6	1.012×10^1	4.925×10^1
TAXI_OUT	2.921×10^6	1.664×10^1	9.193×10^0
WHEELS_OFF	2.921×10^6	1.352×10^3	5.009×10^2
WHEELS_ON	2.920×10^6	1.462×10^3	5.272×10^2
TAXI_IN	2.920×10^6	7.679×10^0	6.270×10^0
CRS_ARR_TIME	3.000×10^6	1.491×10^3	5.115×10^2
ARR_TIME	2.920×10^6	1.467×10^3	5.318×10^2
ARR_DELAY	2.914×10^6	4.261×10^0	5.117×10^1
CANCELLED	3.000×10^6	2.638×10^{-2}	1.603×10^{-1}
DIVERTED	3.000×10^6	2.352×10^{-3}	4.844×10^{-2}
CRS_ELAPSED_TIME	3.000×10^6	1.423×10^2	7.156×10^1
ELAPSED_TIME	2.914×10^6	1.366×10^2	7.168×10^1
AIR_TIME	2.914×10^6	1.123×10^2	6.975×10^1
DISTANCE	3.000×10^6	8.094×10^2	5.879×10^2
DELAY_DUE_CARRIER	5.339×10^5	2.476×10^1	7.177×10^1
DELAY_DUE_WEATHER	5.339×10^5	3.985×10^0	3.241×10^1
DELAY_DUE_NAS	5.339×10^5	1.316×10^1	3.316×10^1
DELAY_DUE_SECURITY	5.339×10^5	1.459×10^{-1}	3.582×10^0
DELAY_DUE_LATE_AIRCRAFT	5.339×10^5	2.547×10^1	5.577×10^1

3 Cancellation Analysis

3.1 Number of Occurrences of Different Classes

- Class 0 (Not Cancelled): 2,920,860 occurrences (97.362%)
- Class 1 (Cancelled): 79,140 occurrences (2.638%)

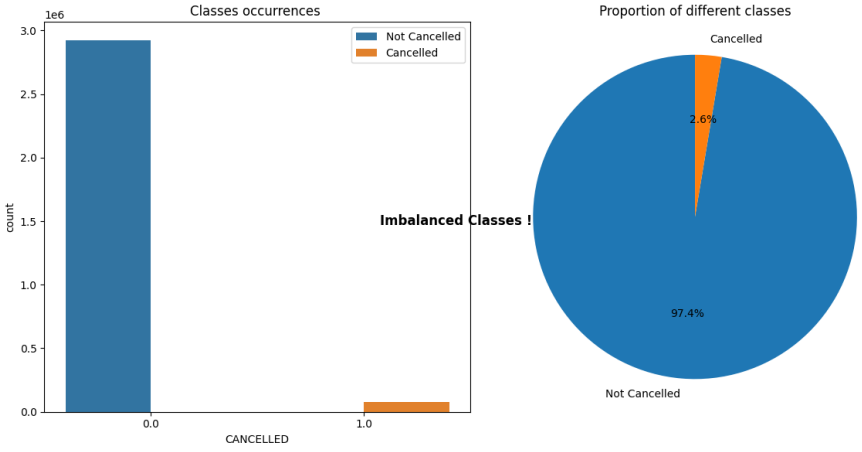

Figure 1: Classes proportion

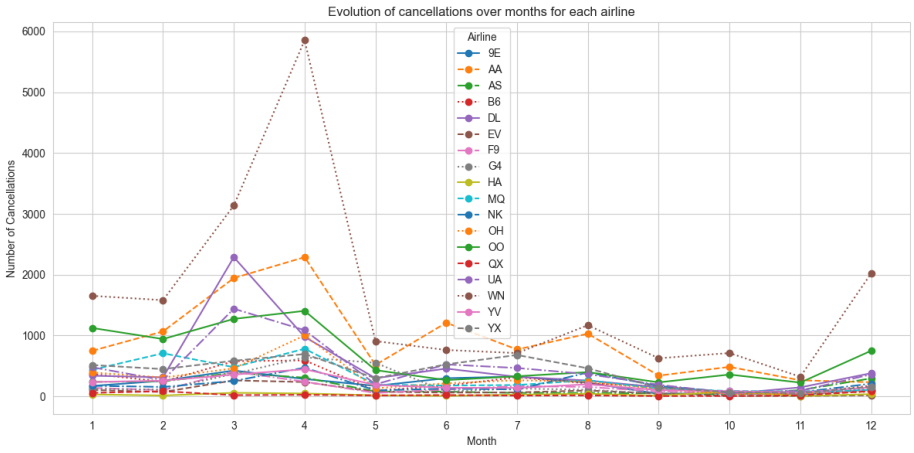

Figure 2: Evolution of cancellations over months for each airline

4 Conclusion

This exploratory analysis provides an overview of the dataset, including column details, data shape, unique values, missing values proportion, statistics, and cancellation analysis.

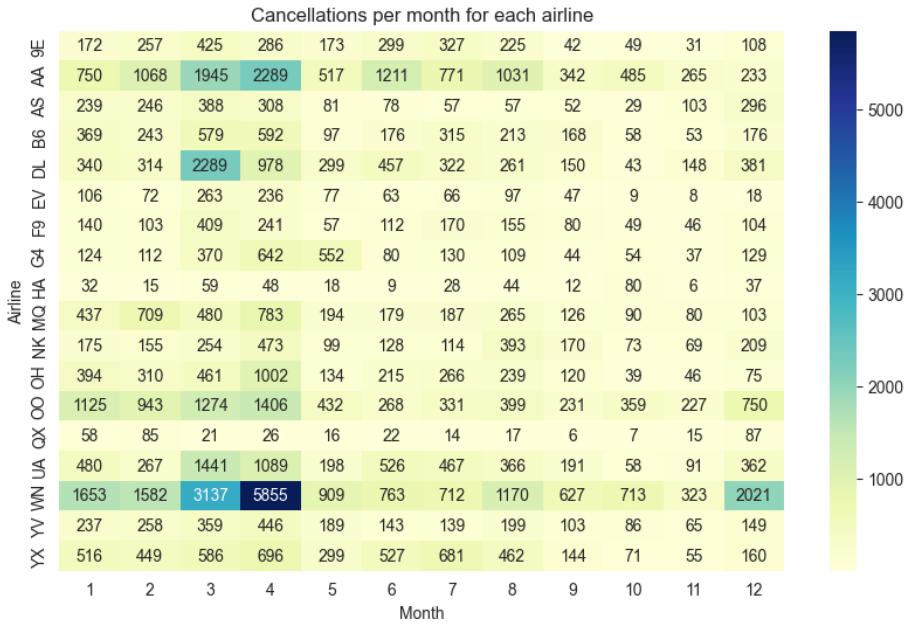

Figure 3: Cancellations per month for each airline

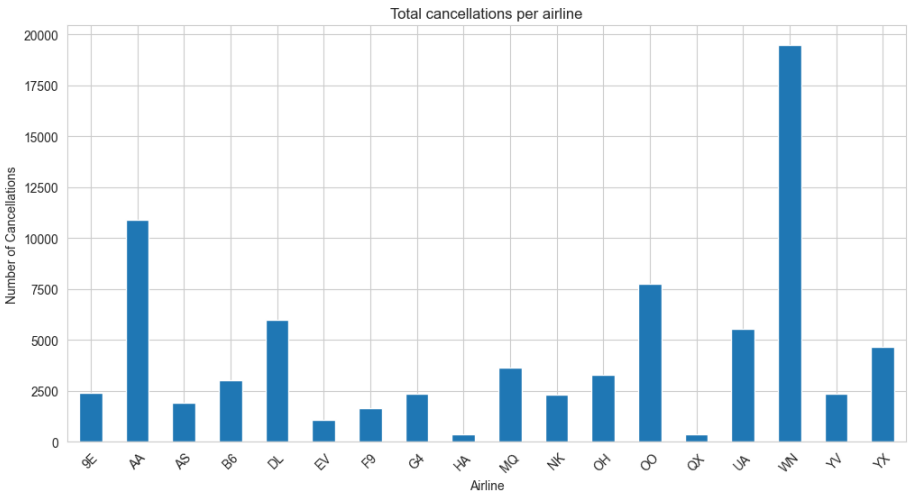

Figure 4: Total cancellations per airline

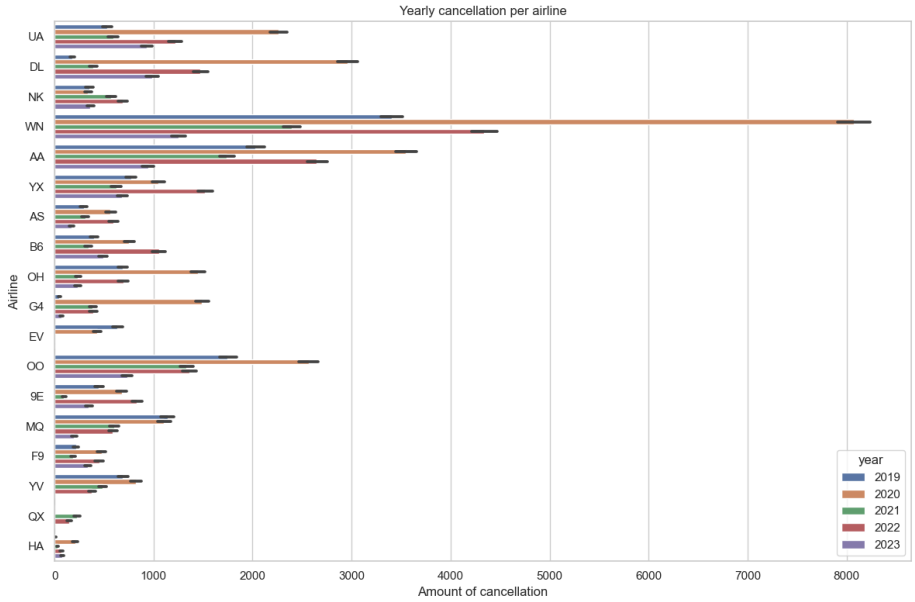

Figure 5: Yearly cancellation per airline

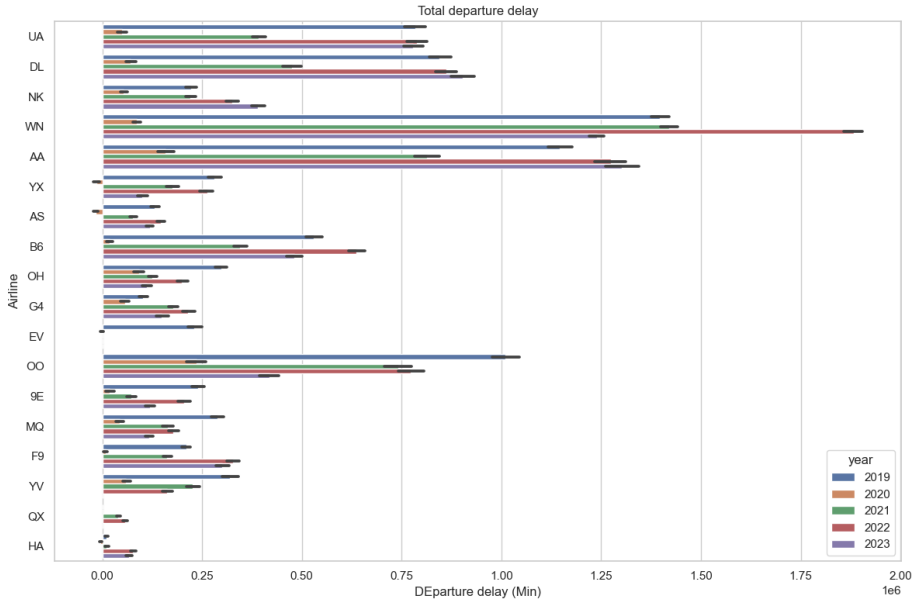

Figure 6: Total departure delay

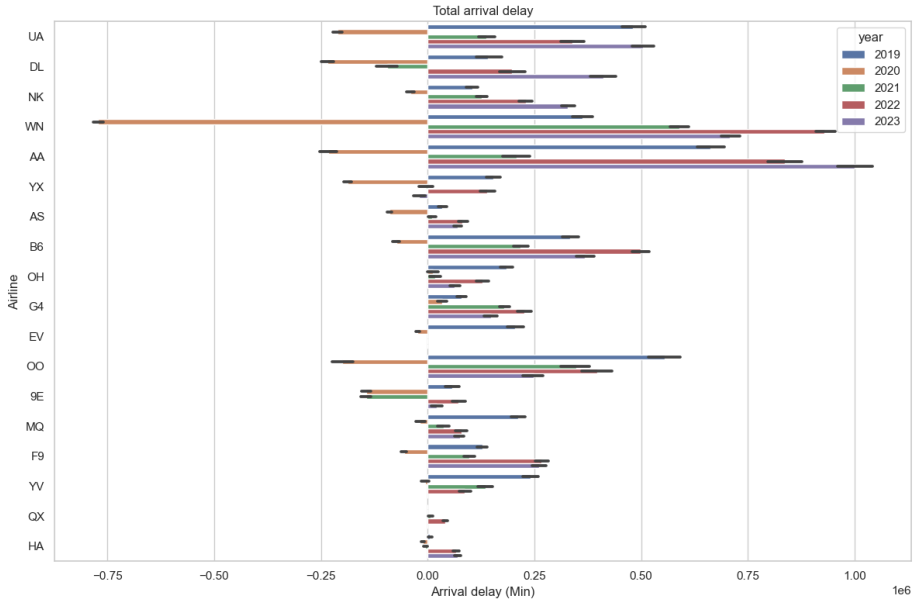

Figure 7: Total arrivall delay