Data 1030 Midterm Presentation

Repository Link

Chai Harsha

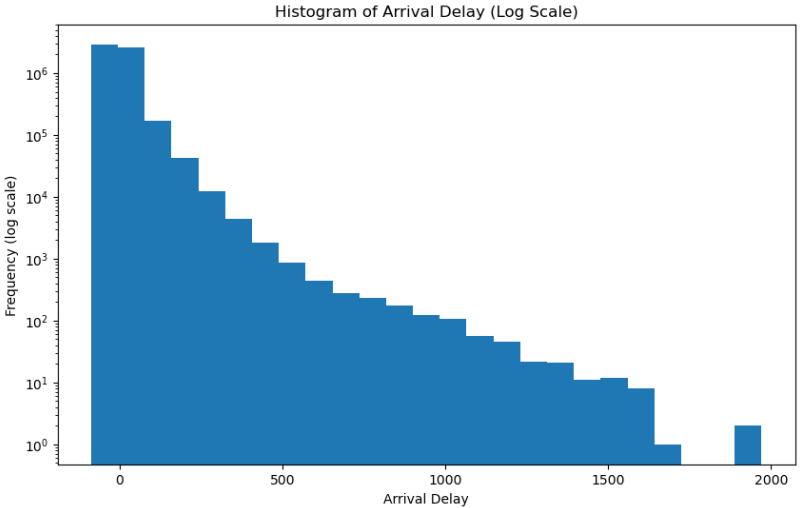
Brown University

October 20, 2025

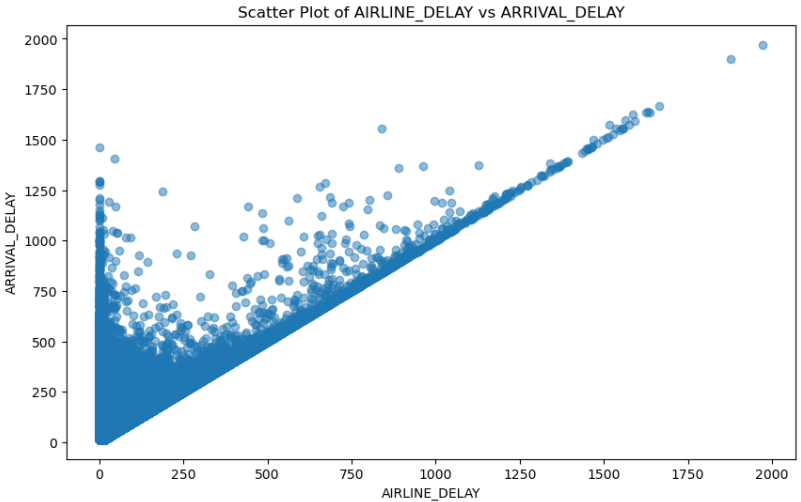
Flight Data

- Thousands of flights + thousands of airports = vast potential for delays
- Airport design, airline inefficiencies, etc
- Can we predict delays?
- USDOT Bureau of Transportation Statistics collects on-time performance of major carriers
- Kaggle dataset - 2015 flight delays
- This is a very large dataset (>5 million rows) and also has many missing values
- Continuous data and some categorical data
- There are other factors not in this database - weather being the obvious one

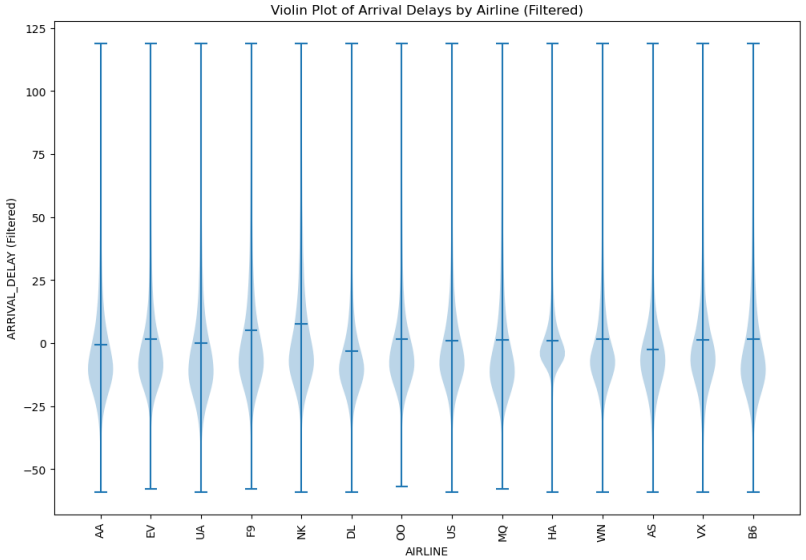
EDA


Histogram of Arrival Delay (Log Scale)
Frequency (log scale)
10^0
10^1
10^2
10^3
10^4
10^5
10^6
0
500
1000
1500
2000
Arrival Delay

EDA


Scatter Plot of AIRLINE_DELAY vs ARRIVAL_DELAY
ARRIVAL_DELAY
AIRLINE_DELAY
0
250
500
750
1000
1250
1500
1750
2000

EDA


Violin Plot of Arrival Delays by Airline (Filtered)
ARRIVAL_DELAY (Filtered)
125
100
75
50
25
0
−25
−50
AA
EV
UA
F9
NK
DL
OO
US
MQ
HA
WN
AS
VX
B6
AIRLINE

Preprocessing

- 105071 rows (1.81% of dataset) missing the target feature - dropped
- Too many airports - selected top 20 + "Other"
- Standard scaler on continuous data
- Minmax scaler on time-of-day data
- One-Hot encoded categorical data
- Dropped columns like cancellation reason and air system delay for too many missing data ($> 80\%$)