Mid-Term Project

SHILPI & SHUBHANGI

Objective

PREDICTING DELAYS

WHY ???

- Challenging, lot of learning
- Can check accuracy
- Supervised ML Algorithms

Steps

- **Data Gathering**

  Used VS Code editor and SQL

- **Data Exploration**

  Explored different features & their relationship with different variables

- **Data Cleaning**

  Null values replacement, Outliers detection,

- **Feature Engineering**

  Binning, Encoding

- **Feature Selection**

  feature importance and heatmaps

Approach

RESEARCH

- Research what factors affect flight delay
- Deciding how to sample dataset for modeling

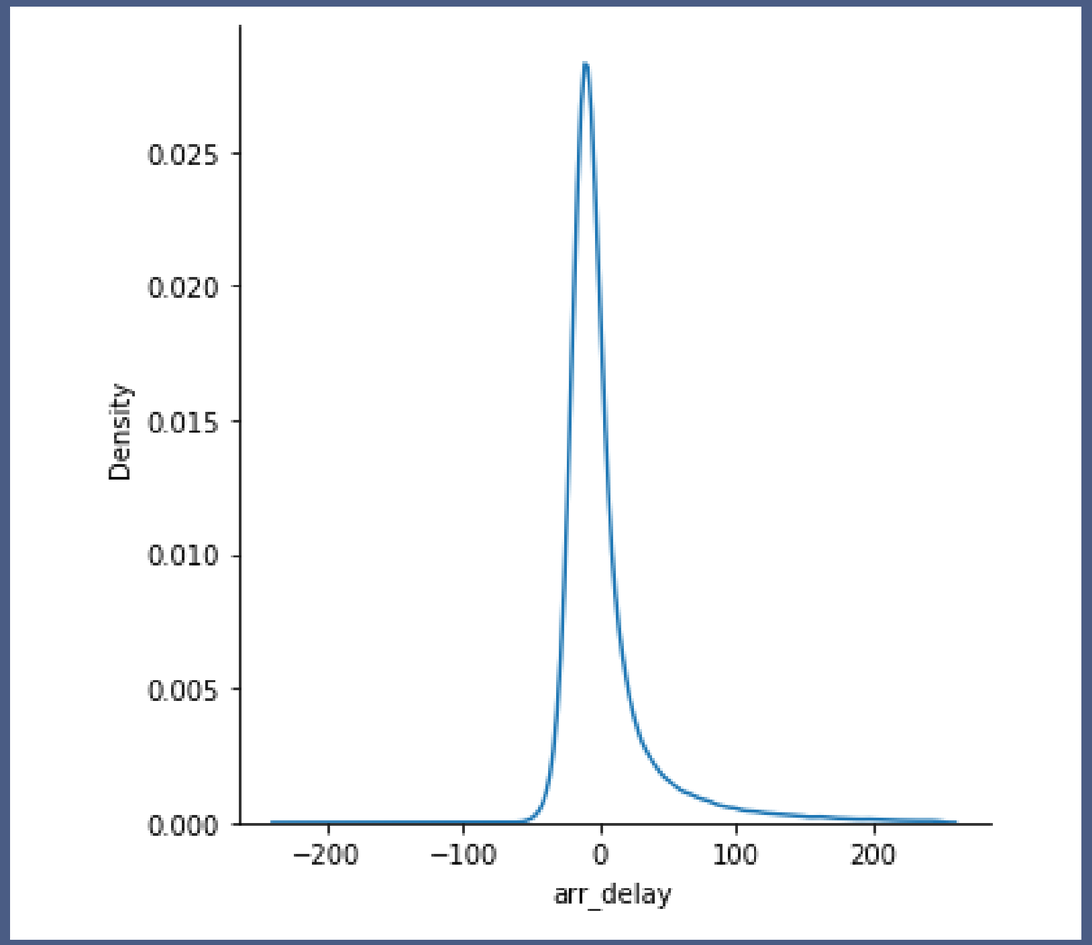
TOOLS & TECHNIQUES

- Jupyter Lab/Notebook
- Google Collab
- Pycaret
- Google Drive/Slack

CODE QUALITY

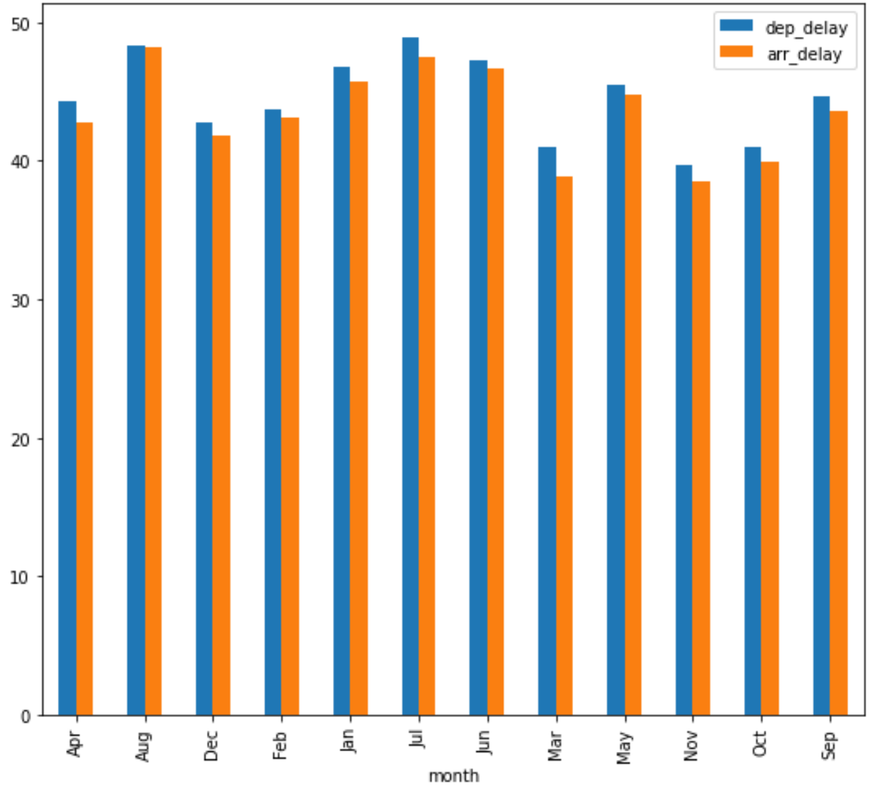
- Organize Codes
- Add comments & docstring
- Relevant names of variable
- Organized Dataset
- Storey-Telling

Exploratory Data Analysis

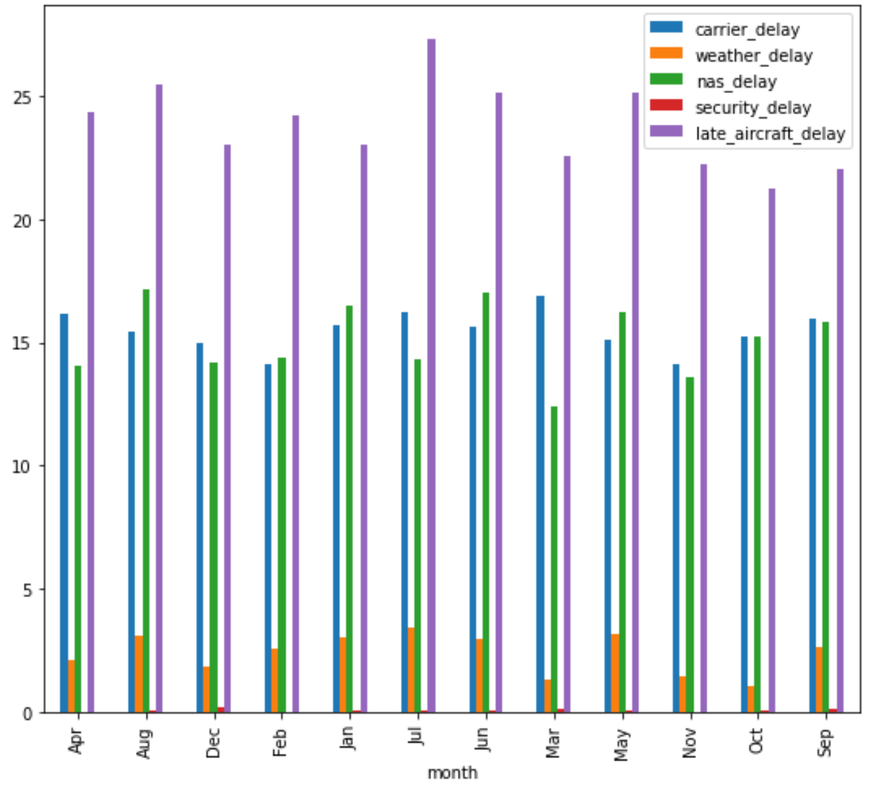


TEST HYPOTHESIS FOR NORMAL DISTRIBUTION

- Sampled Dataset Randomly of samle size 150K
- Plotted graph to check distribution
- Shepiro test on sample data < 5000

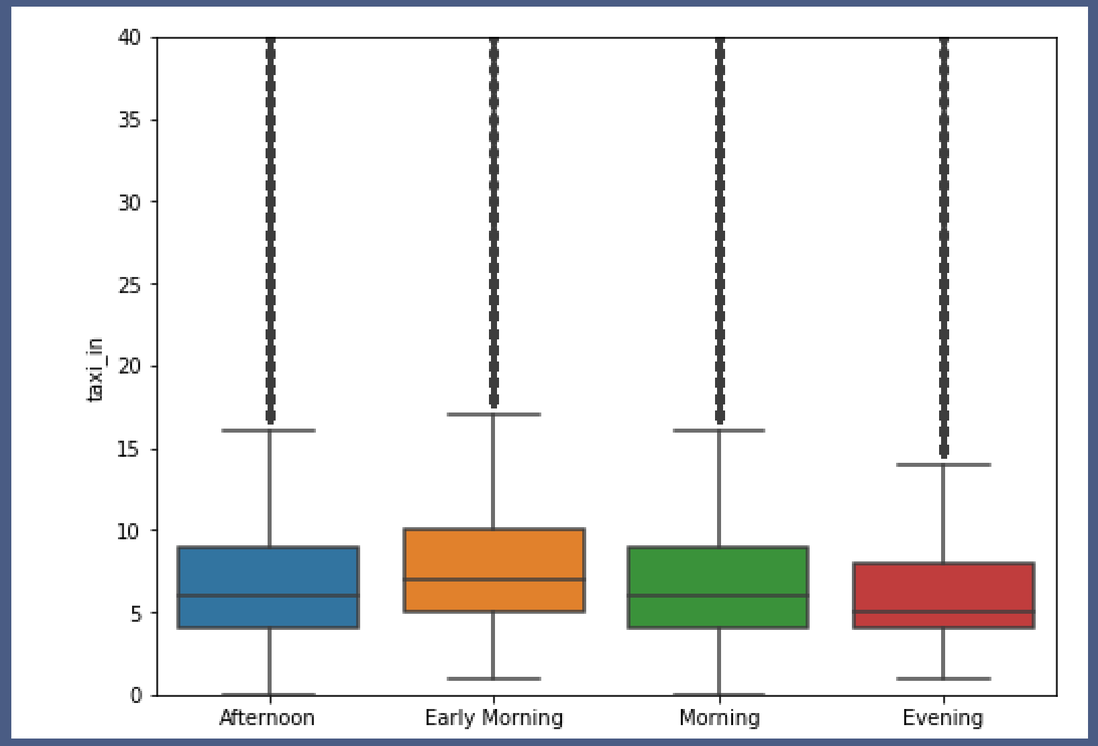


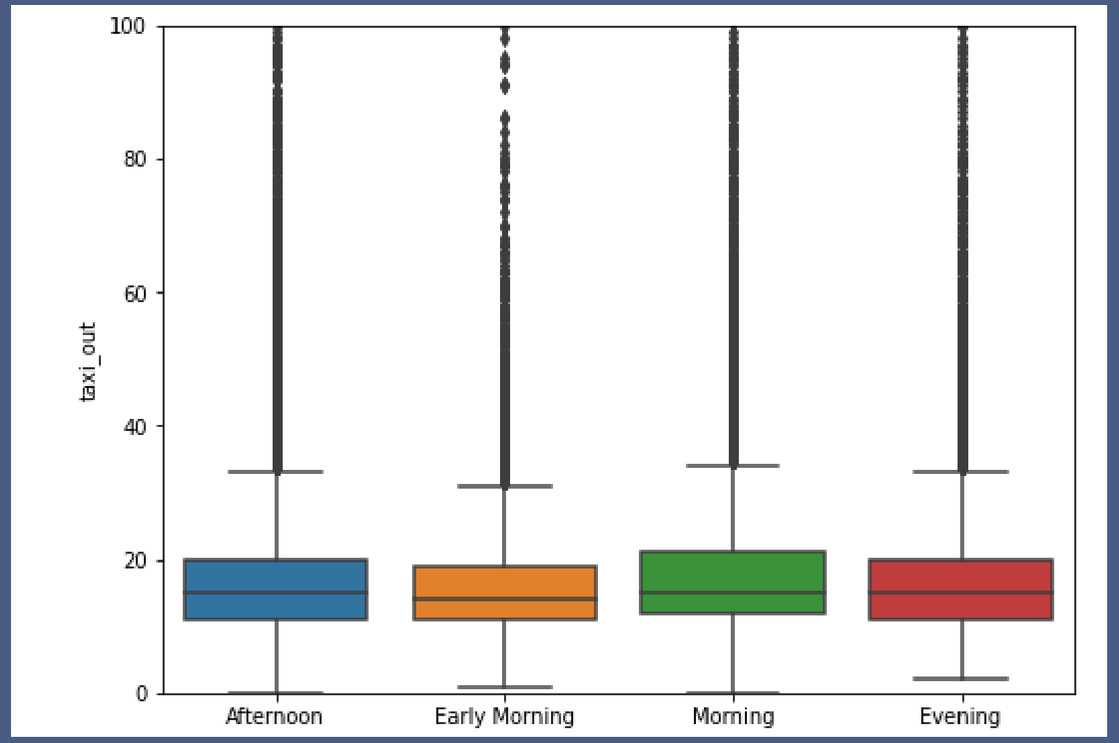
Monthly distribution of arrival and departure delays



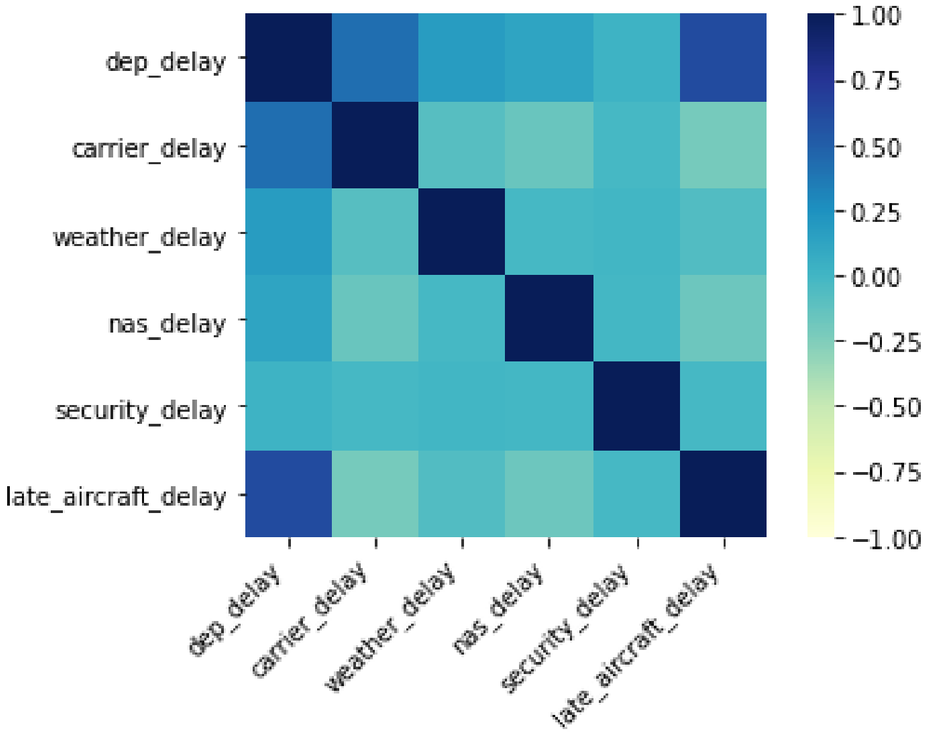
Monthly distribution of reasons of arrival delays

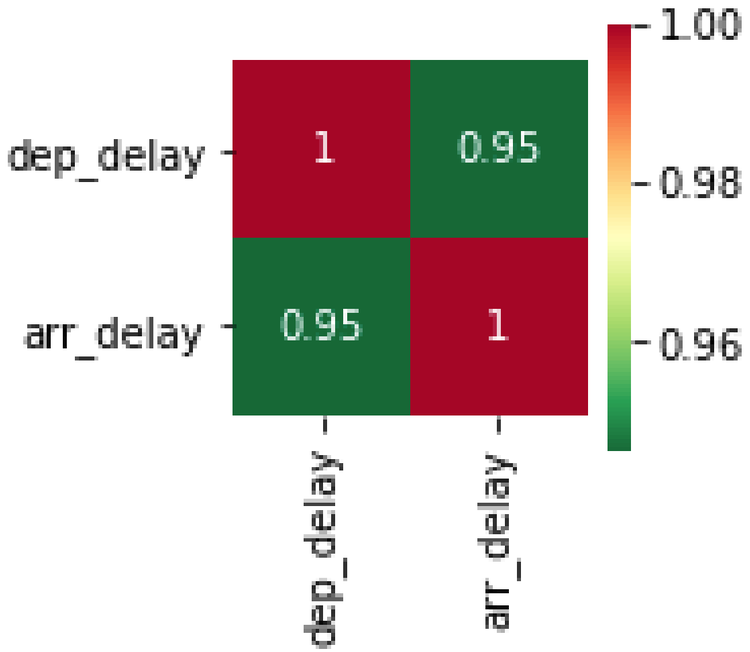
Taxi in and out during the day



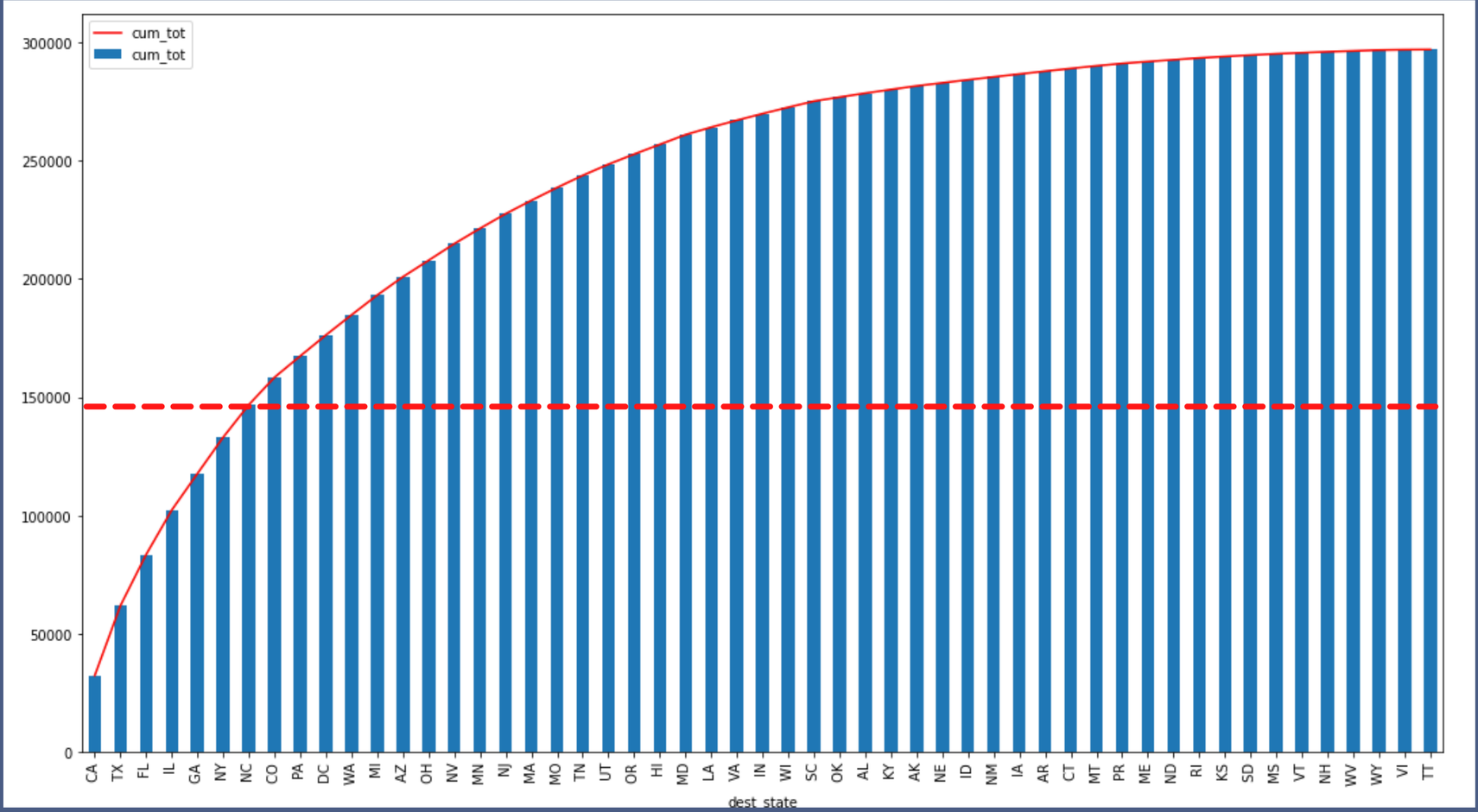


Relationship between arrival and departure



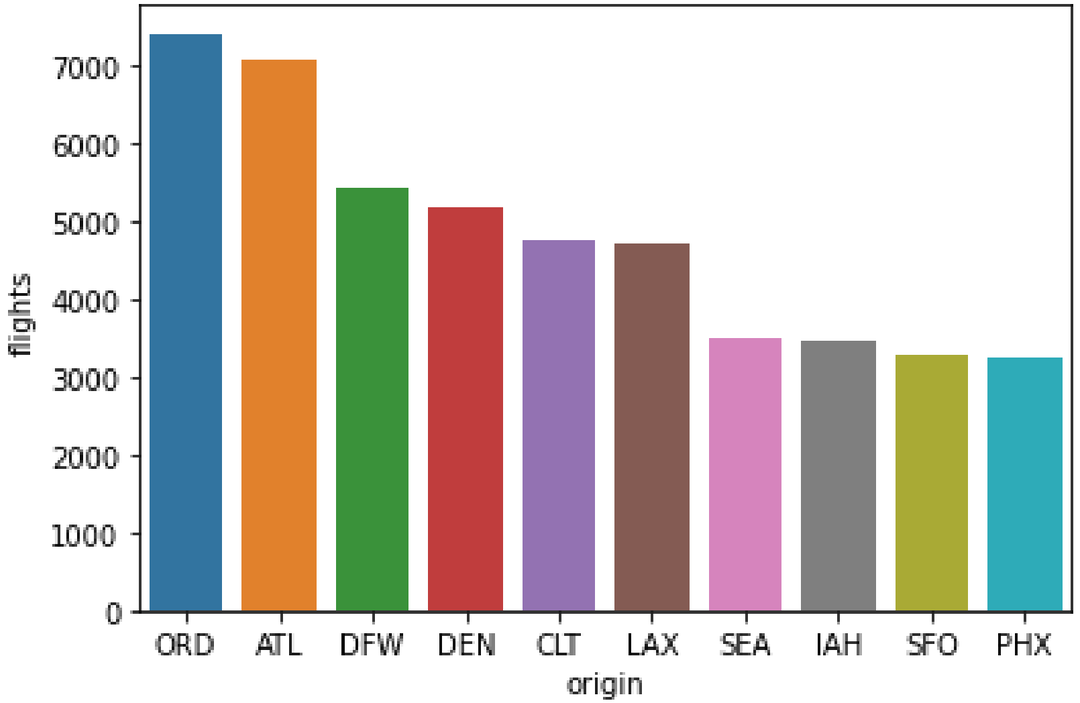


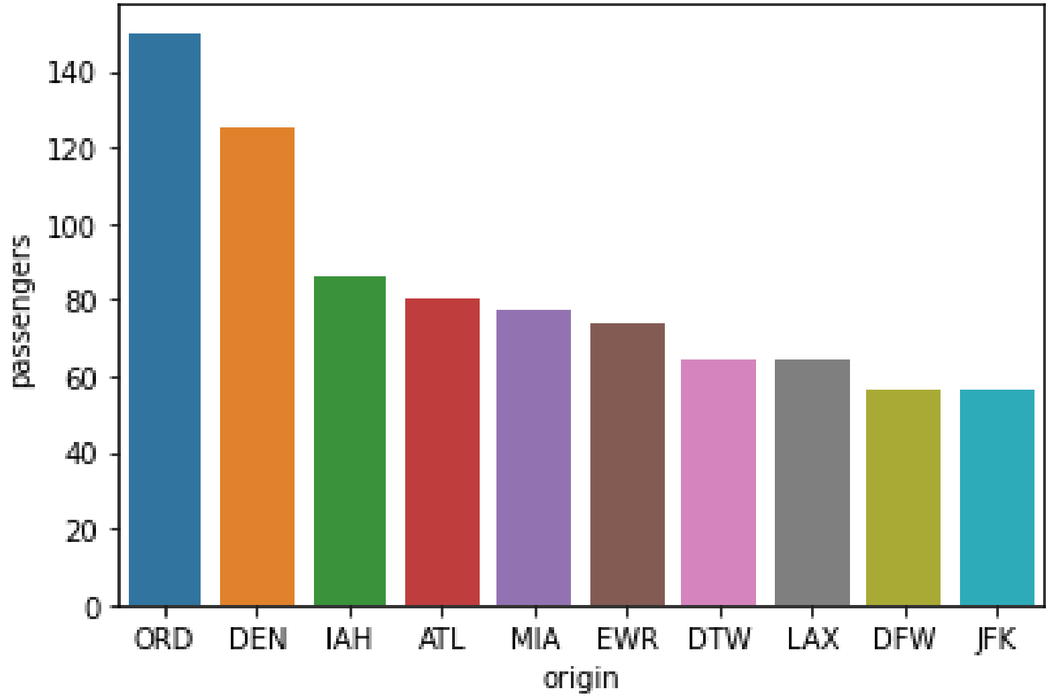
Air traffic on the basis of states

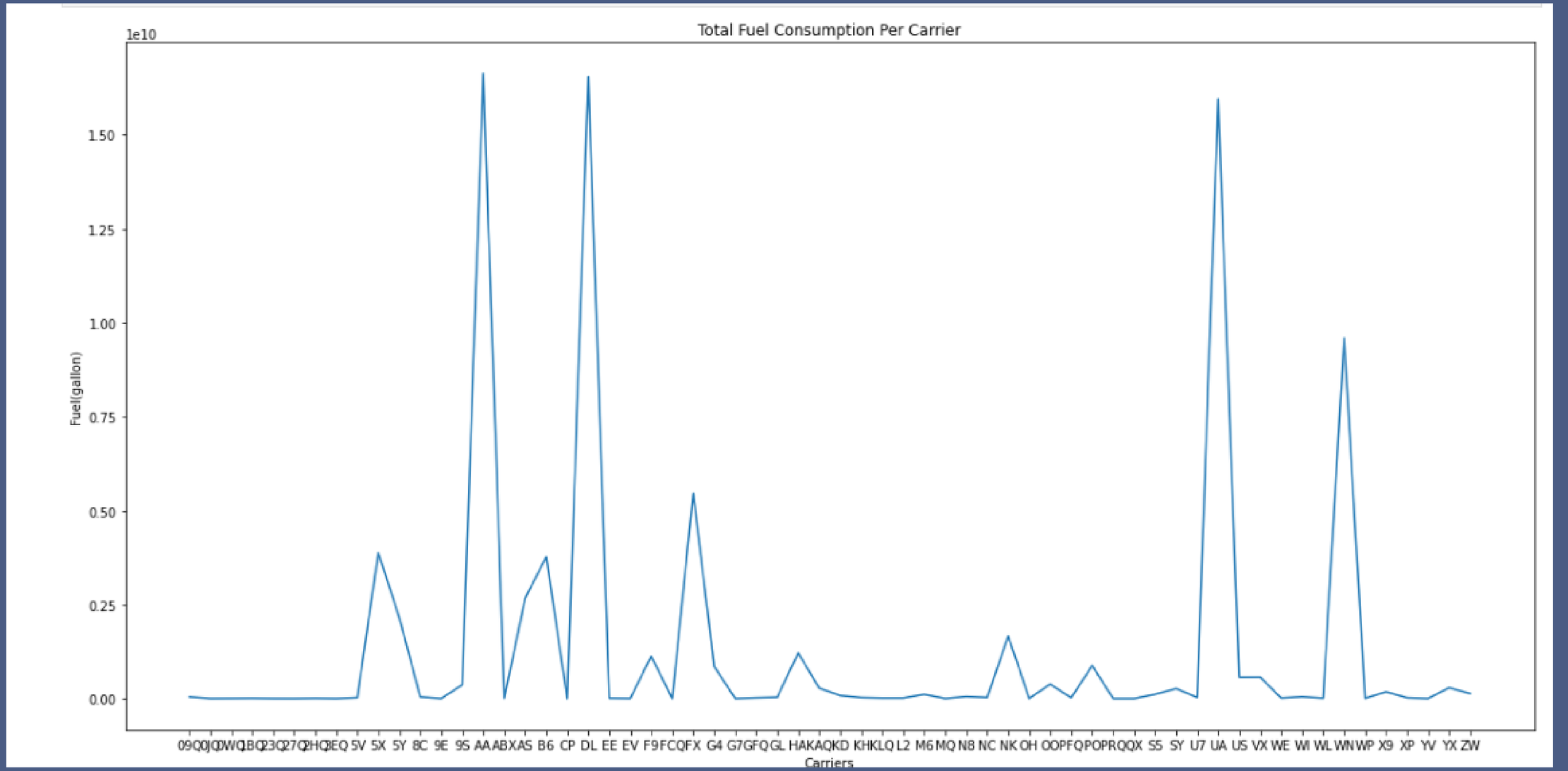


States covering 50% air trafic

- California
- Texas
- Florida
- Illinois
- Georgia
- New York
- North Carolina


7000
6000
5000
4000
3000
2000
1000
0
flights
ORD
ATL
DFW
DEN
CLT
LAX
SEA
IAH
SFO
PHX
origin


140
120
100
80
60
40
20
0
passengers
ORD
DEN
IAH
ATL
MIA
EWR
DTW
LAX
DFW
JFK
origin


Total Fuel Consumption Per Carrier
1e10
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Fuel(gallon)
09 Q0 JQ 0W Q1 BQ 23 Q2 7Q 2H Q3 EQ 5V 5X 5Y 8C 9E 9S AA AB XA S B6 CP DL EE EV F9 FC QF X G4 G7 GF QG L HA KA QK D KH KL Q L2 M6 MQ N8 NC NK OH OO PF QP OP RQ QX S5 SY U7 UA US VX WE WI WL WN WP X9 XP YV YX ZW
Carriers

ML Algorithms

Logistic Regression

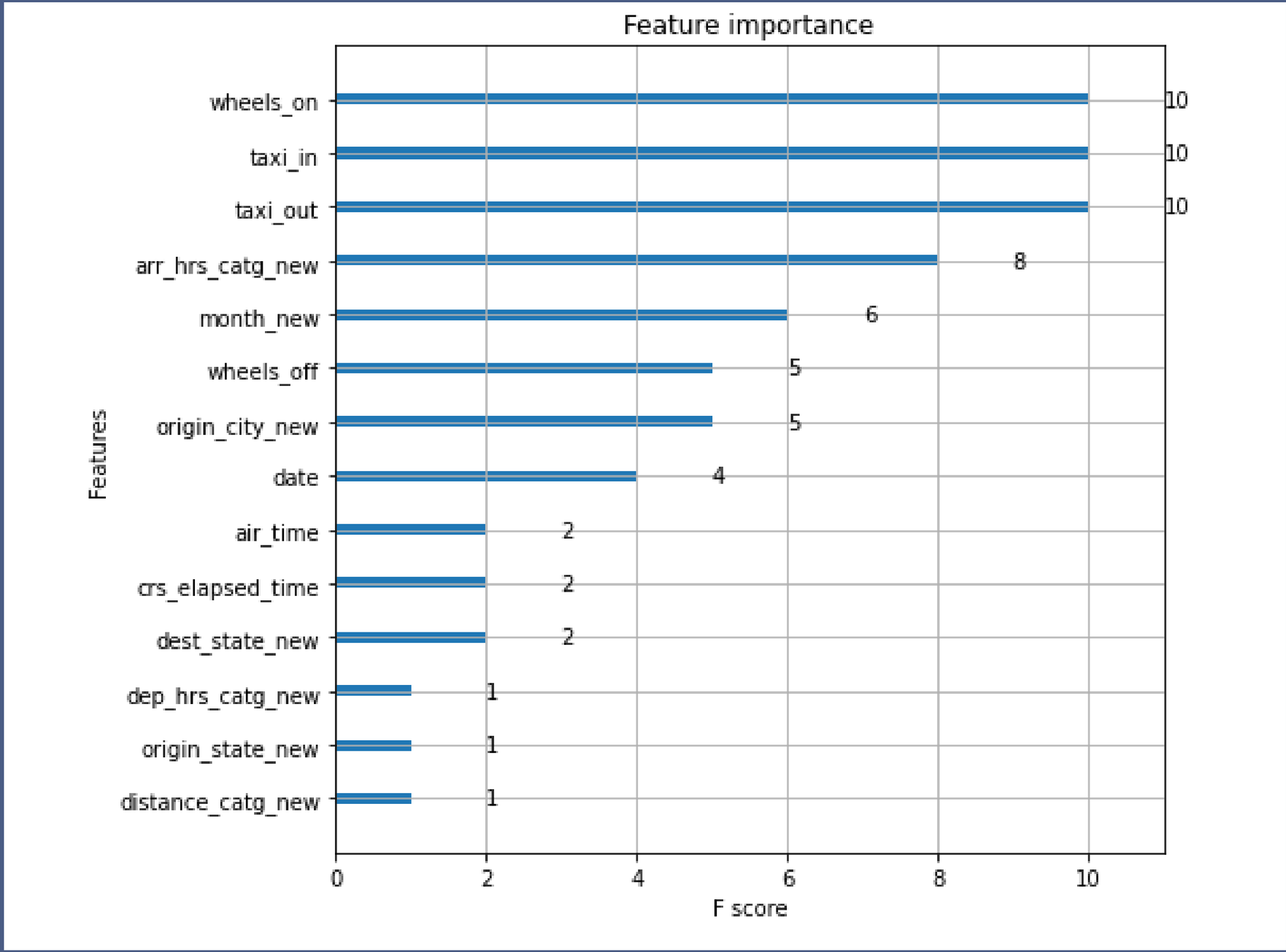
Random Forest

XG Boost

Model Evaluation

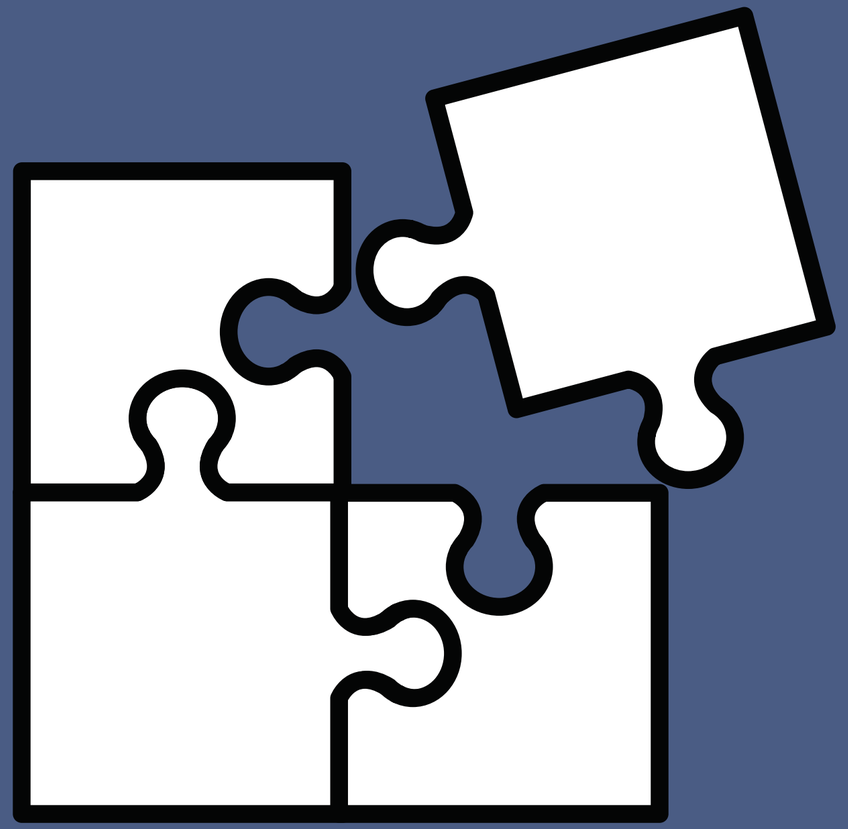
Random Forest: Accuracy 5%

XG Boost: MSE 33.81, MSE 29


Feature importance
wheels_on
taxi_in
taxi_out
arr_hrs_catg_new
month_new
wheels_off
origin_city_new
date
air_time
crs_elapsed_time
dest_state_new
dep_hrs_catg_new
origin_state_new
distance_catg_new
Features
10
10
10
8
6
5
5
4
2
2
2
1
1
1
0
2
4
6
8
10
F score

Challenges

- Technical issues
- Sampling dataset
- Many new concepts to learn
- Tools - Google Colab, Pycaret
- No definite answer
- Limited Timeline



THANKYOU