Mid-Term Project

Exploring the Air Travel Industry

Isabelle de Robert

& Maggie Fiander

Brief Agenda

- Project Overview
- Exploratory Data Analysis
  - Review of Tasks
- Modeling and Evaluations
  - Review of Process
  - Model Evaluation Metrics
  - Model Selection
- Conclusions and Next Steps

The Data Structure

passengers

The observations of passenger totals - spanning 2015-2019

fuel

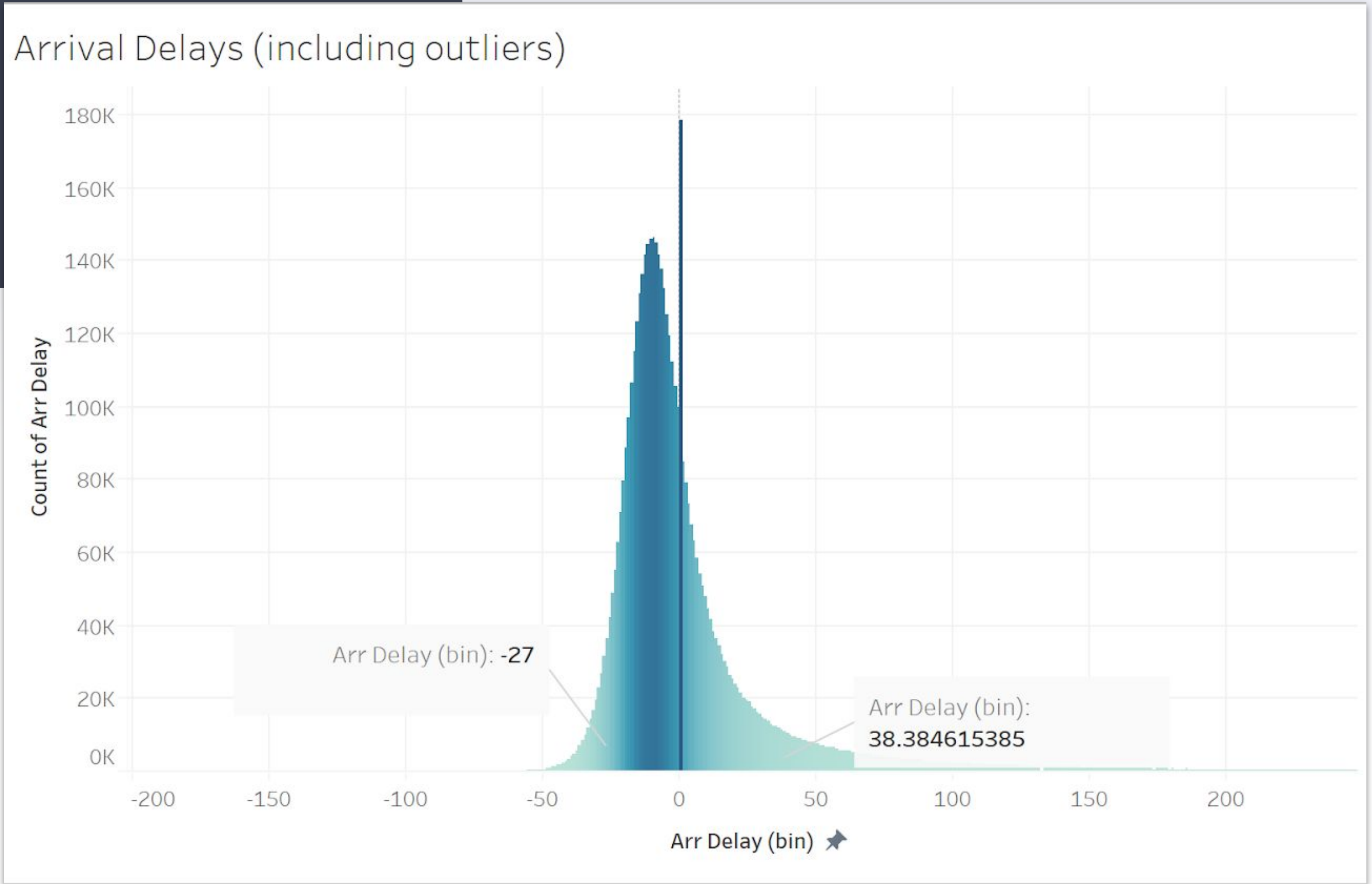
The observations on fuel consumption - spanning 2015 - 2019

flights (flights_test)

The observations on flight arrival and departure patterns - data spanning 2018 and 2019

1

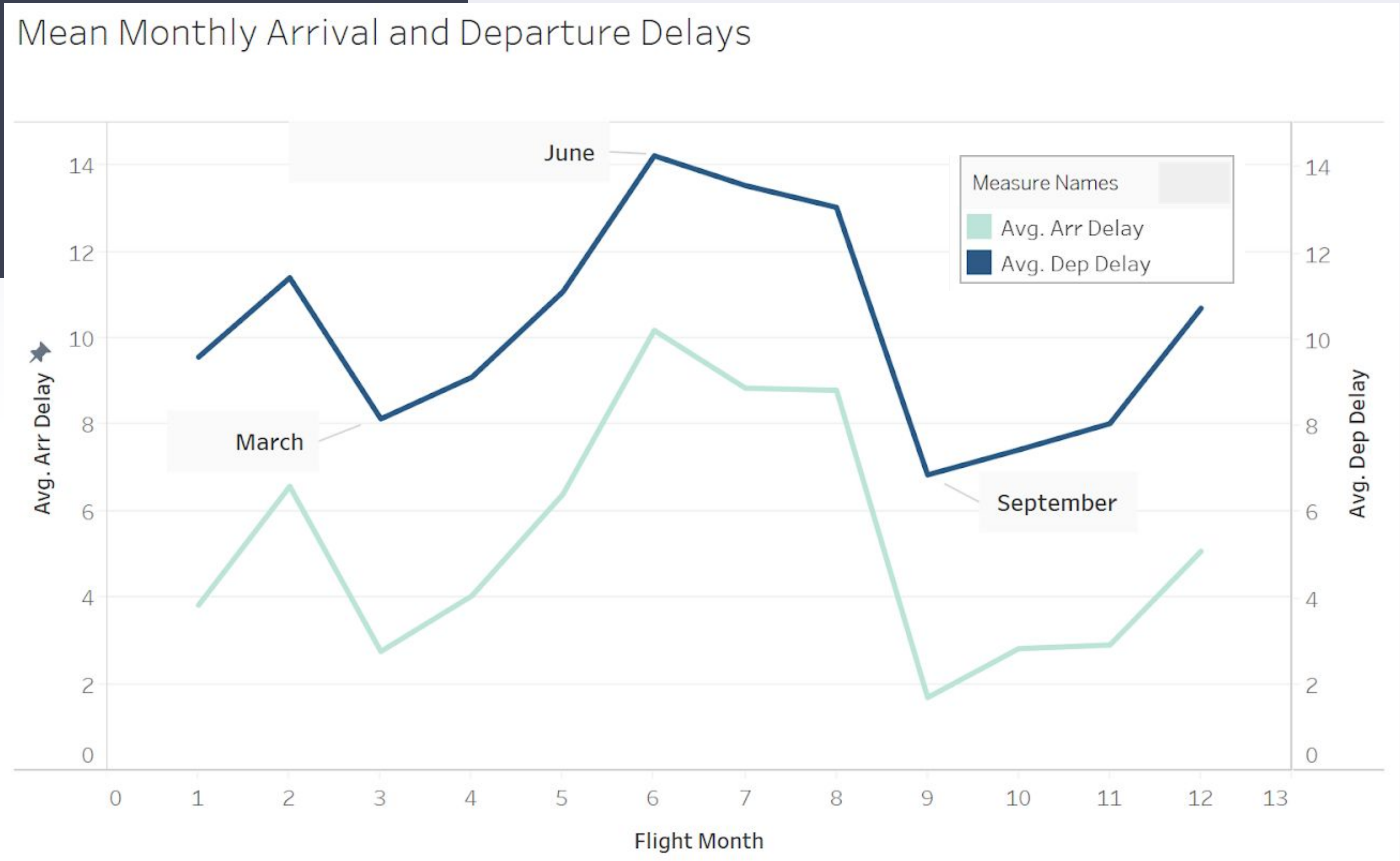
Exploratory Data Analysis



Arrival Delays (including outliers)

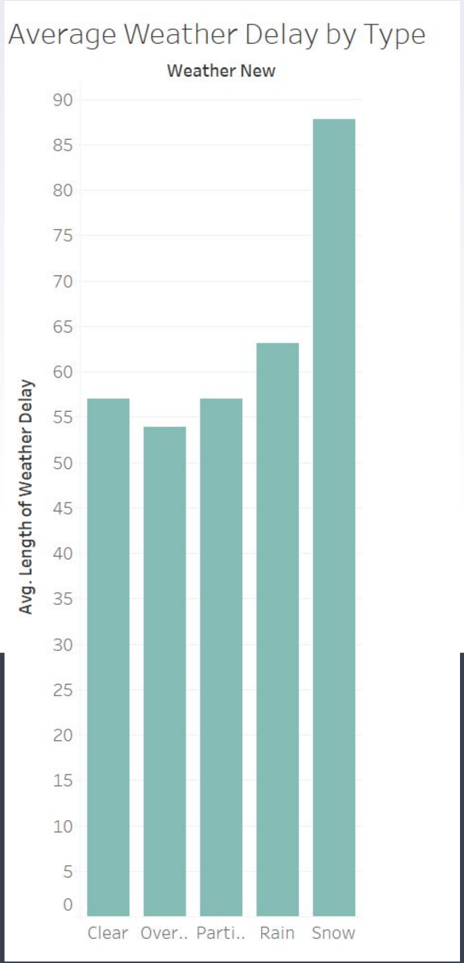
Arr Delay (bin): -27

Arr Delay (bin): 38.384615385

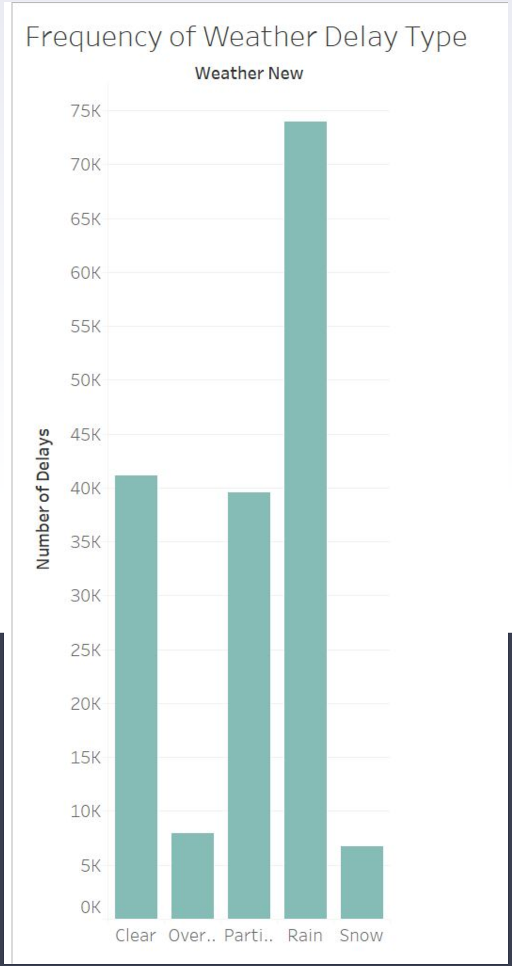


Mean Monthly Arrival and Departure Delays

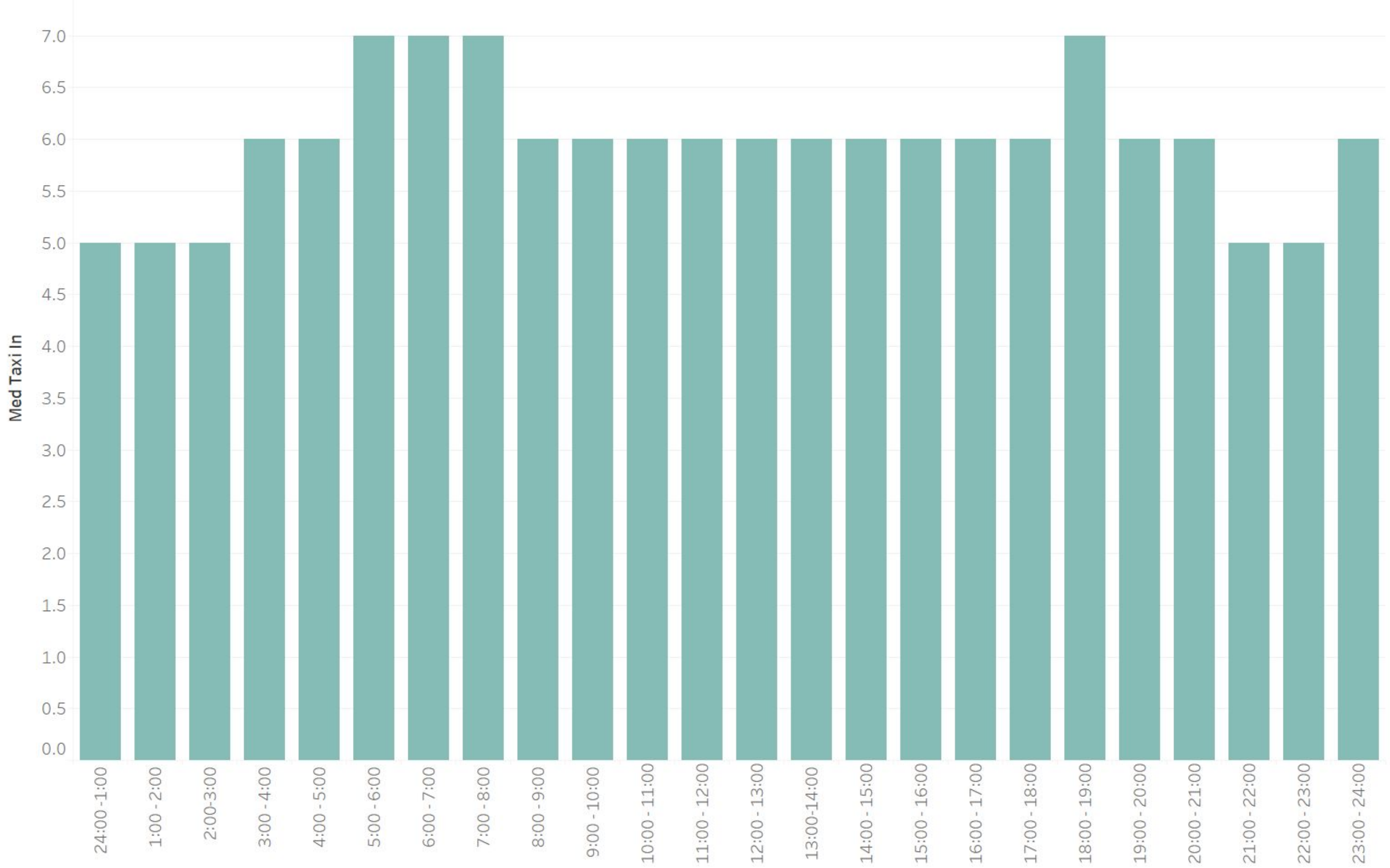
Delay Length Impacts due to Weather

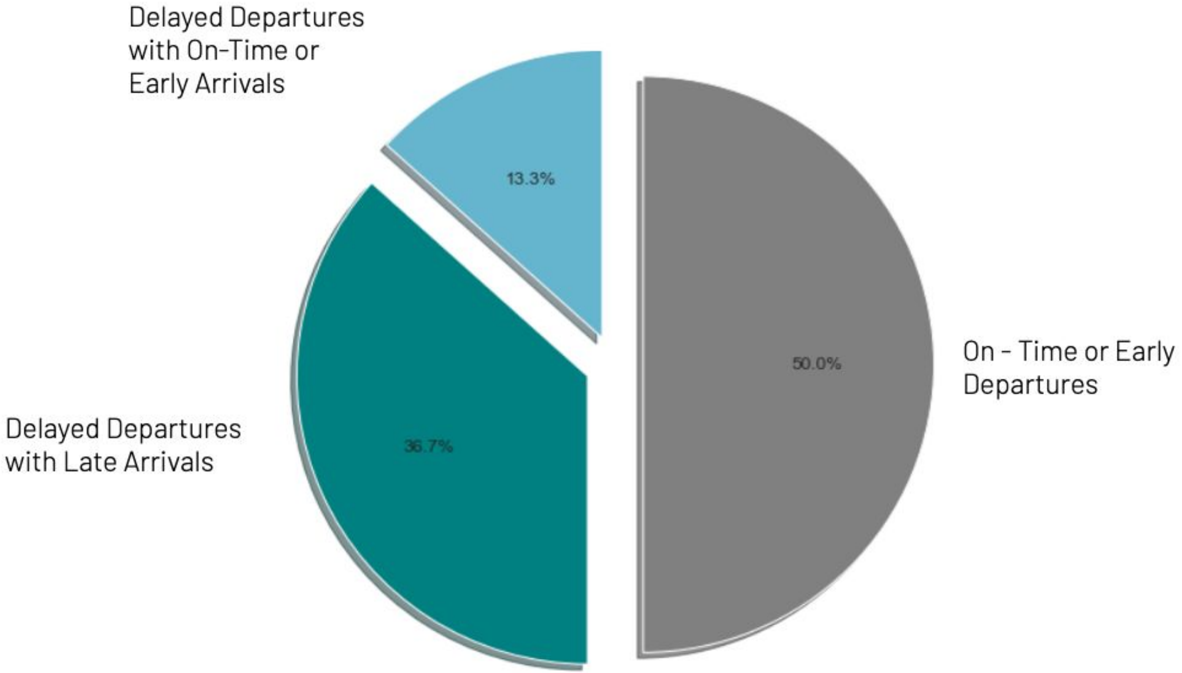


Delay Frequency – Weather

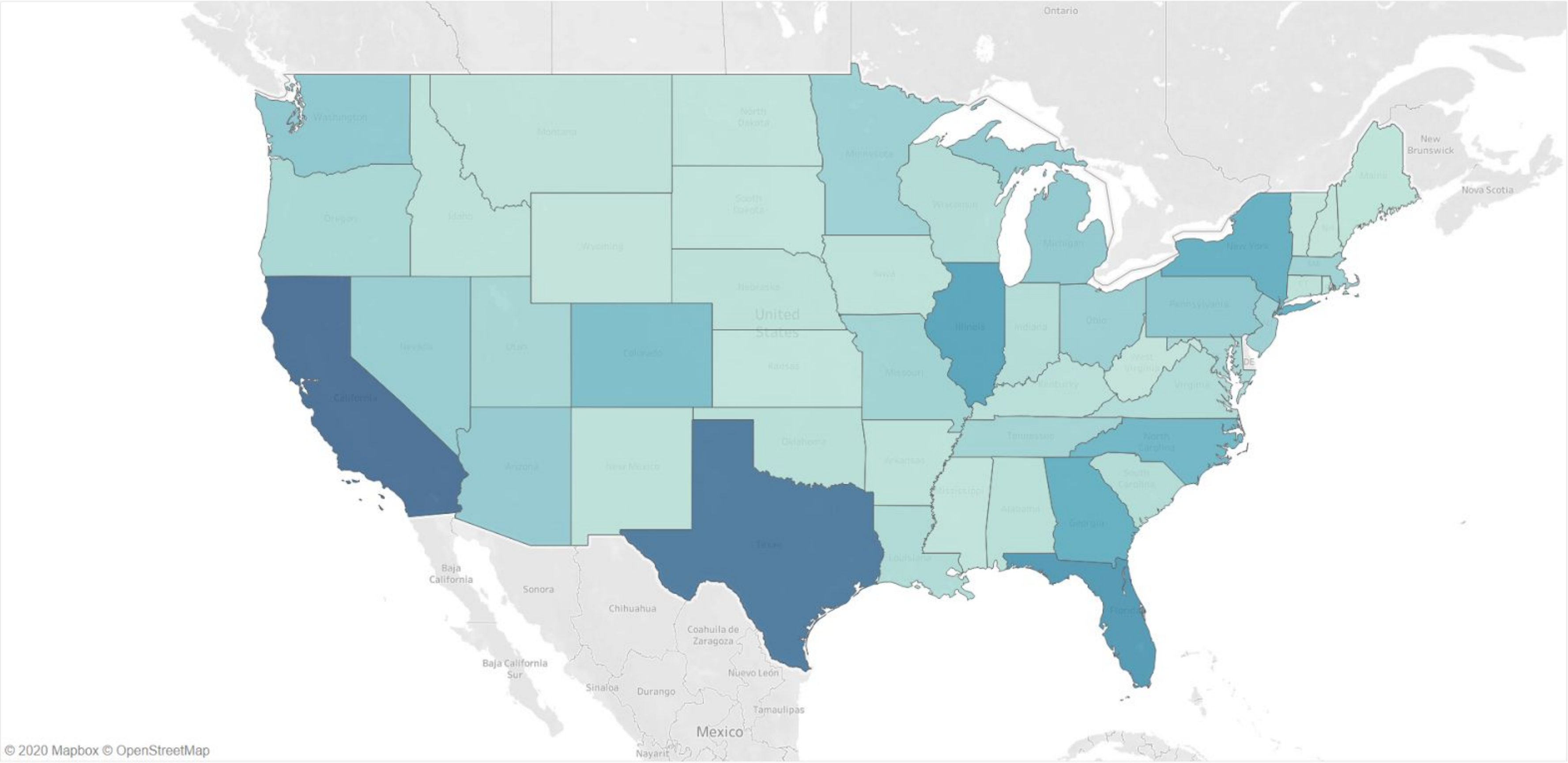


Median Taxi in Time by Scheduled Arrival Time

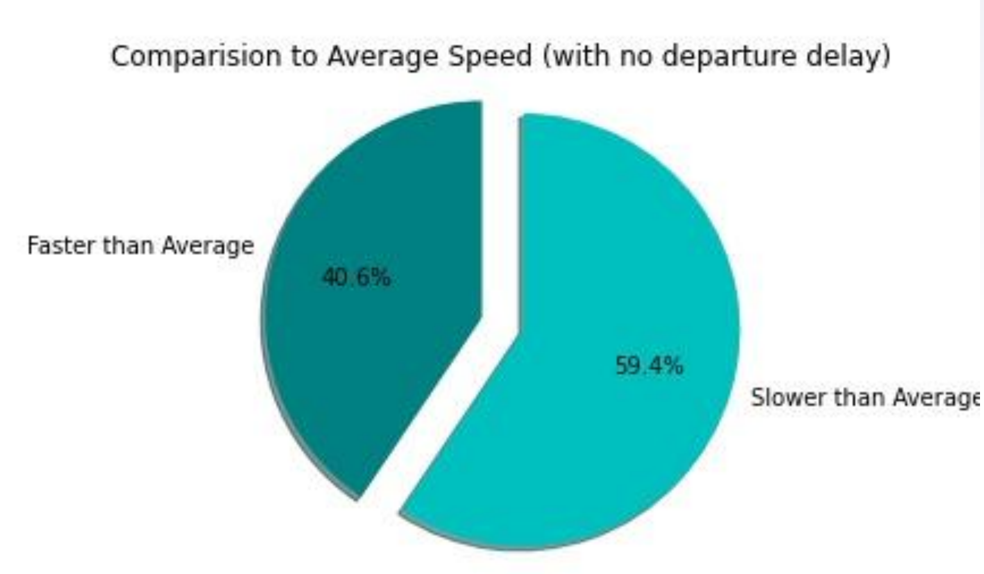


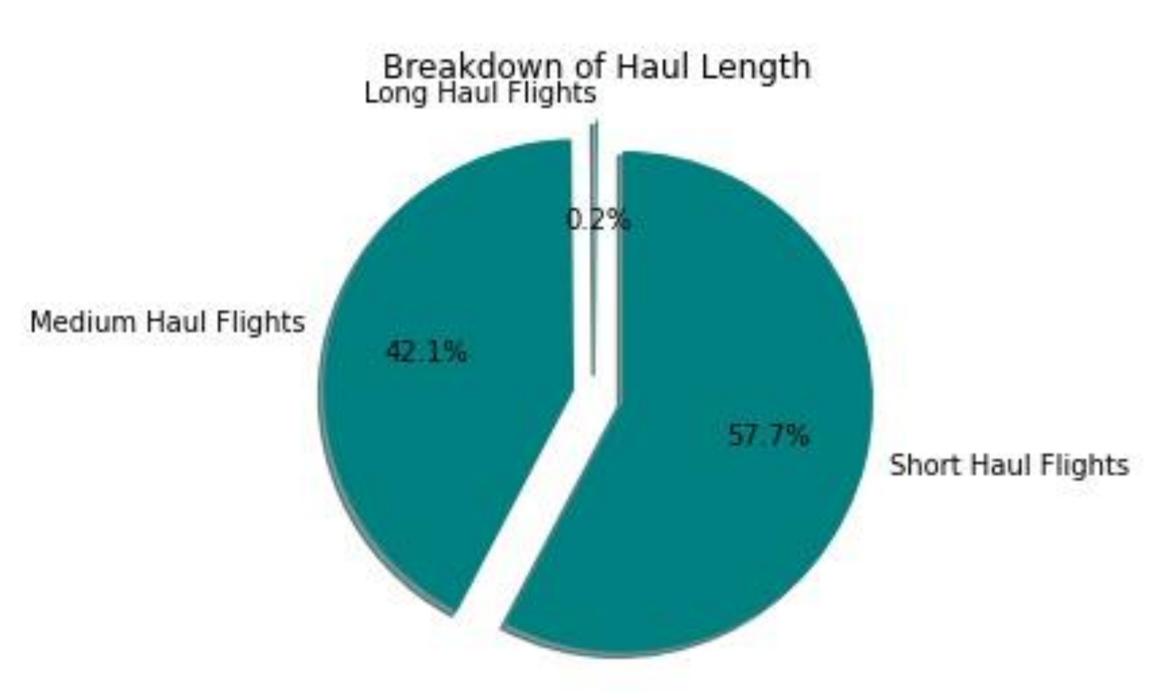

Delayed Departures with On-Time or Early Arrivals
13.3%
On - Time or Early Departures
50.0%
Delayed Departures with Late Arrivals
36.7%

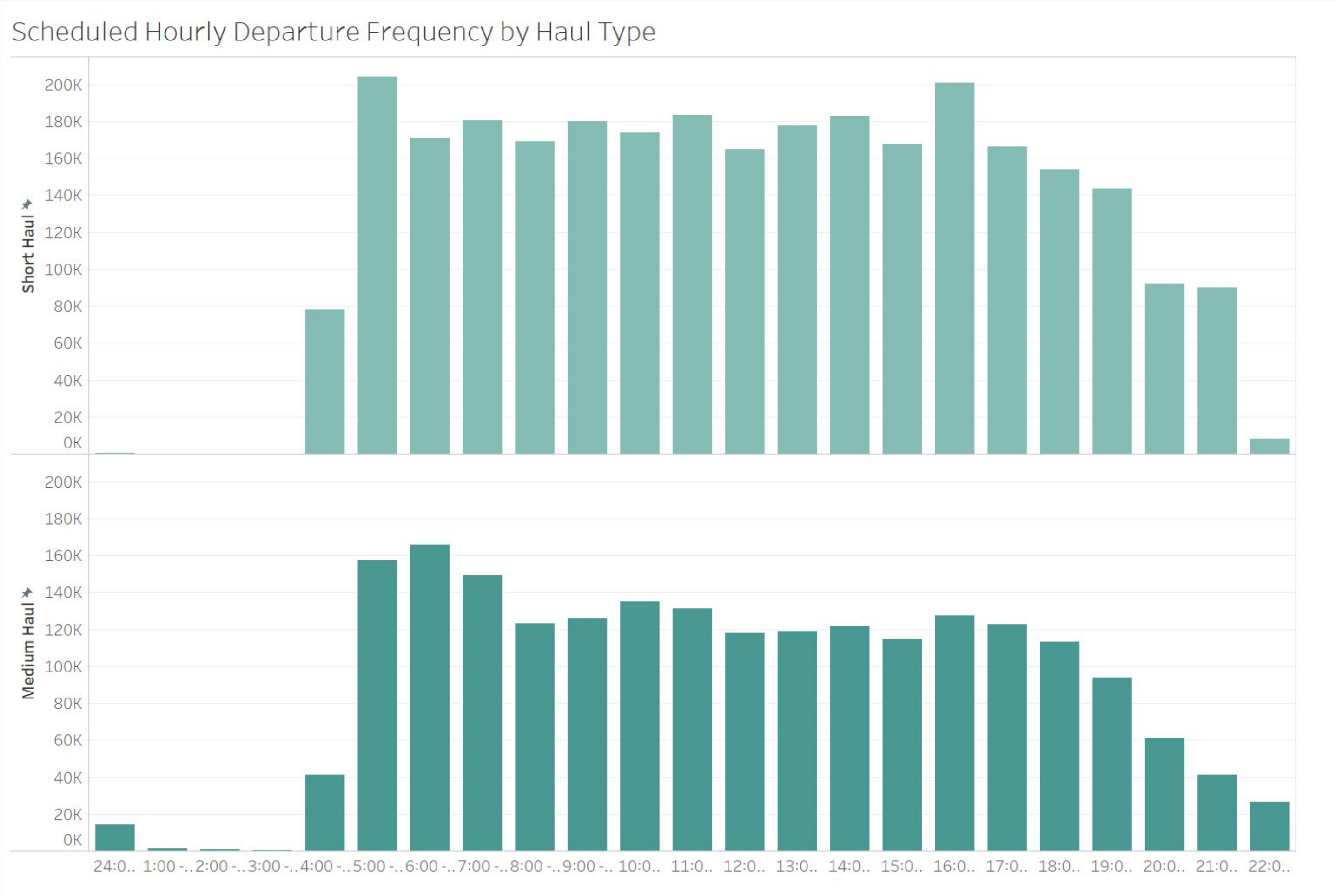
Percent of Air Traffic by State
(using departing flights)



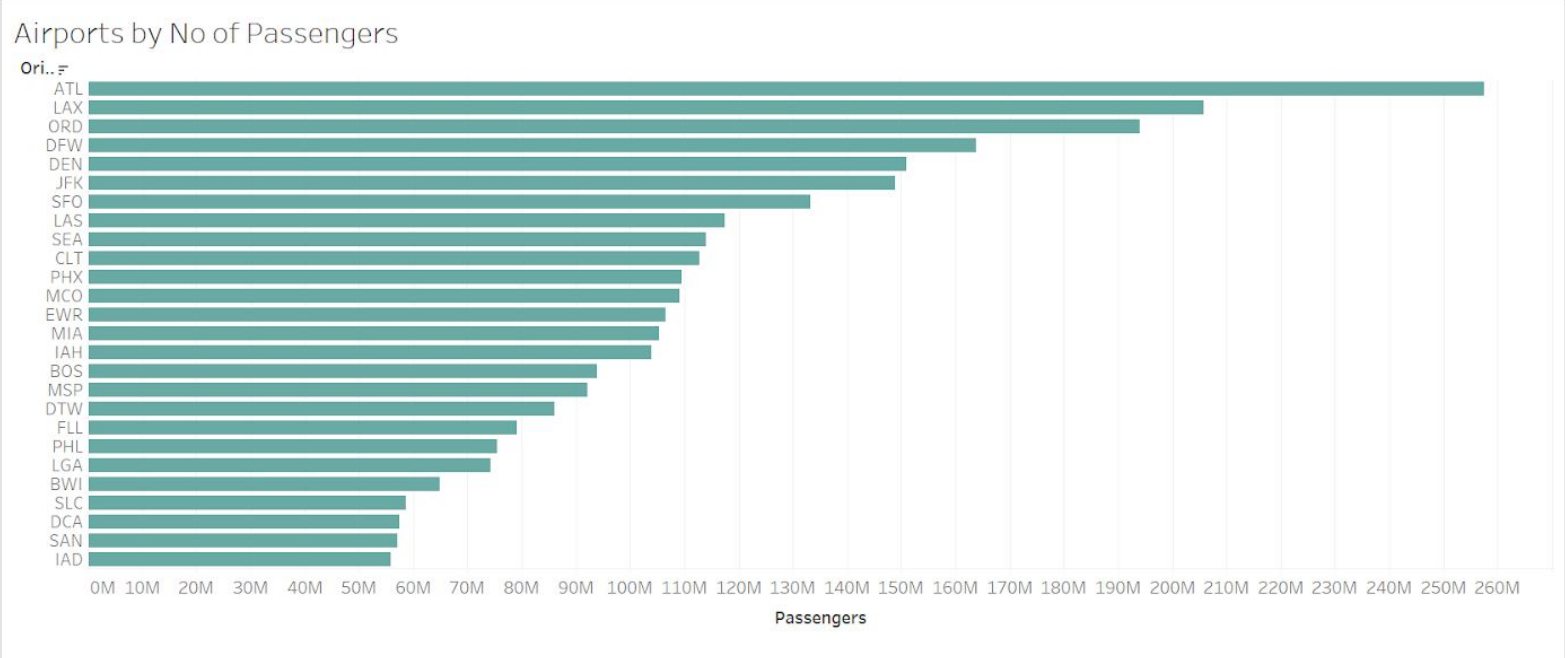
© 2020 Mapbox © OpenStreetMap


Comparision to Average Speed (with no departure delay)
Faster than Average
40.6%
59.4%
Slower than Average

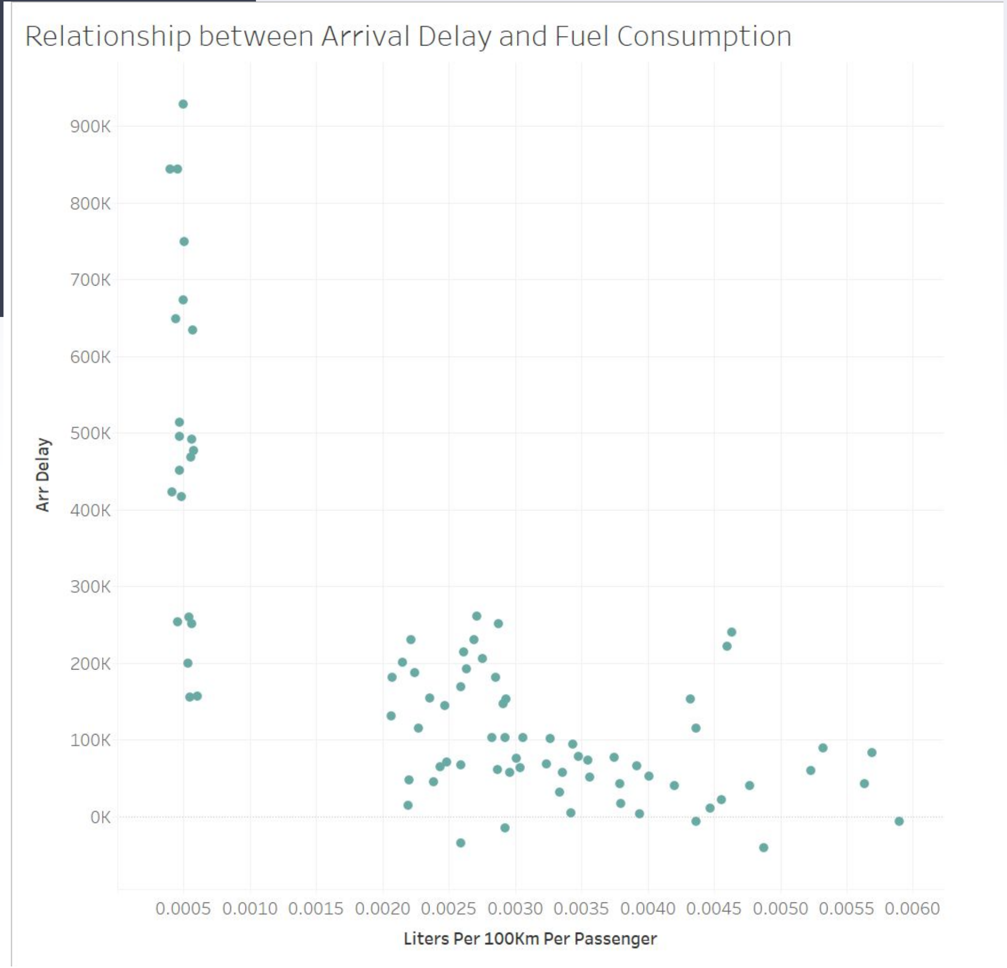

Breakdown of Haul Length
Long Haul Flights
0.2%
Medium Haul Flights
42.1%
57.7%
Short Haul Flights

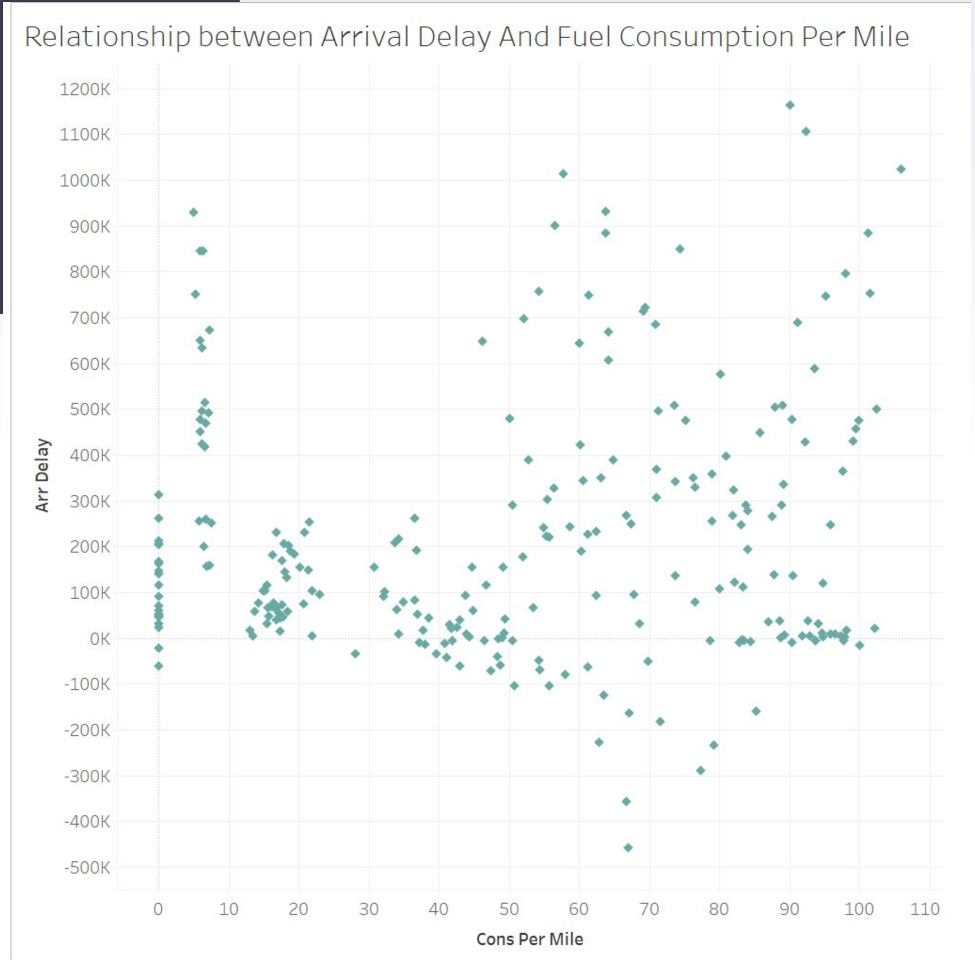


Scheduled Hourly Departure Frequency by Haul Type



Airports by No of Passengers
Ori..
ATL
LAX
ORD
DFW
DEN
JFK
SFO
LAS
SEA
CLT
PHX
MCO
EWR
MIA
IAH
BOS
MSP
DTW
FLL
PHL
LGA
BWI
SLC
DCA
SAN
IAD
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M 180M 190M 200M 210M 220M 230M 240M 250M 260M
Passengers


Relationship between Arrival Delay and Fuel Consumption
Arr Delay
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
0.0040
0.0045
0.0050
0.0055
0.0060
Liters Per 100Km Per Passenger

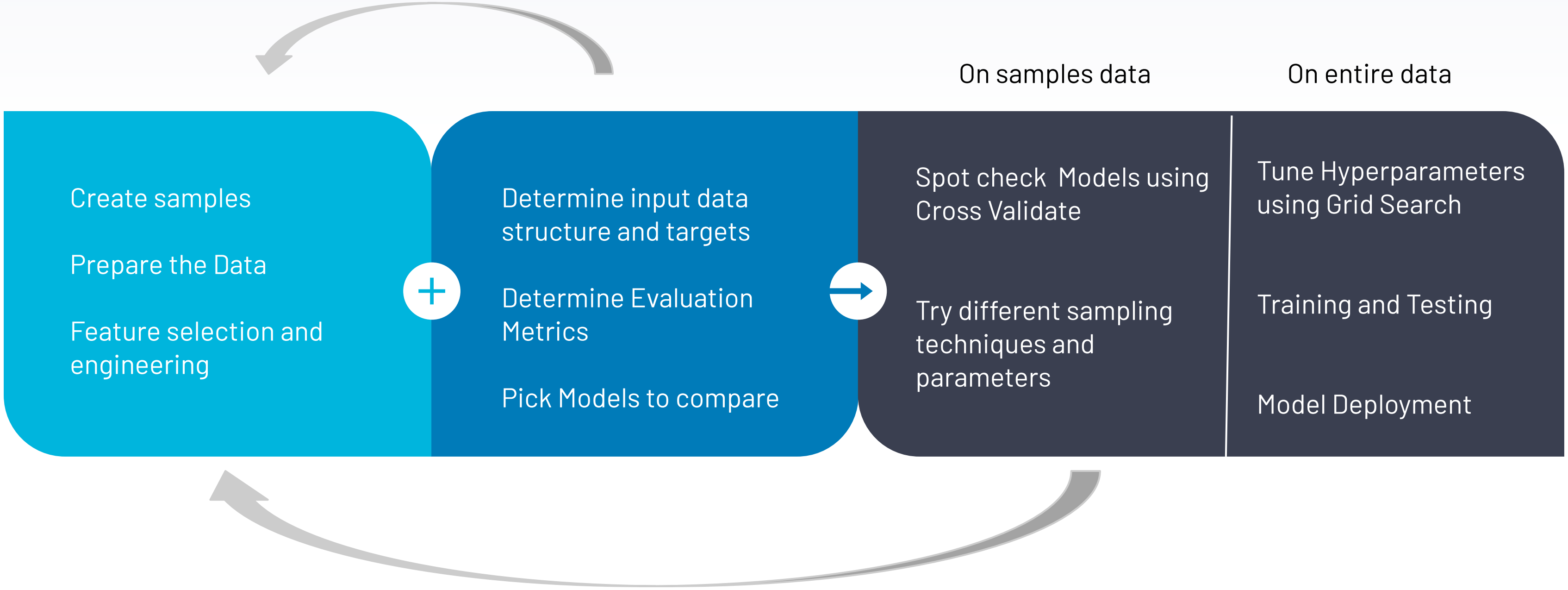


Relationship between Arrival Delay And Fuel Consumption Per Mile

2

Modeling and Evaluation

Process



On samples data

On entire data

Create samples

Prepare the Data

Feature selection and engineering

Determine input data structure and targets

Determine Evaluation Metrics

Pick Models to compare

Spot check Models using Cross Validate

Try different sampling techniques and parameters

Tune Hyperparameters using Grid Search

Training and Testing

Model Deployment

Feature Engineering

- Including an average measure delay from historical data - linked to flight month
- Dealing with Categorical Variables
- Time Related Data – Binning
- Dimensionality reduction (PCA)

Evaluation Metrics

- Regression : R^2 and MSE

- Multiclass Classification :

  F1 Score (weighted), Accuracy (balanced) and AUC

- Binary Classification :

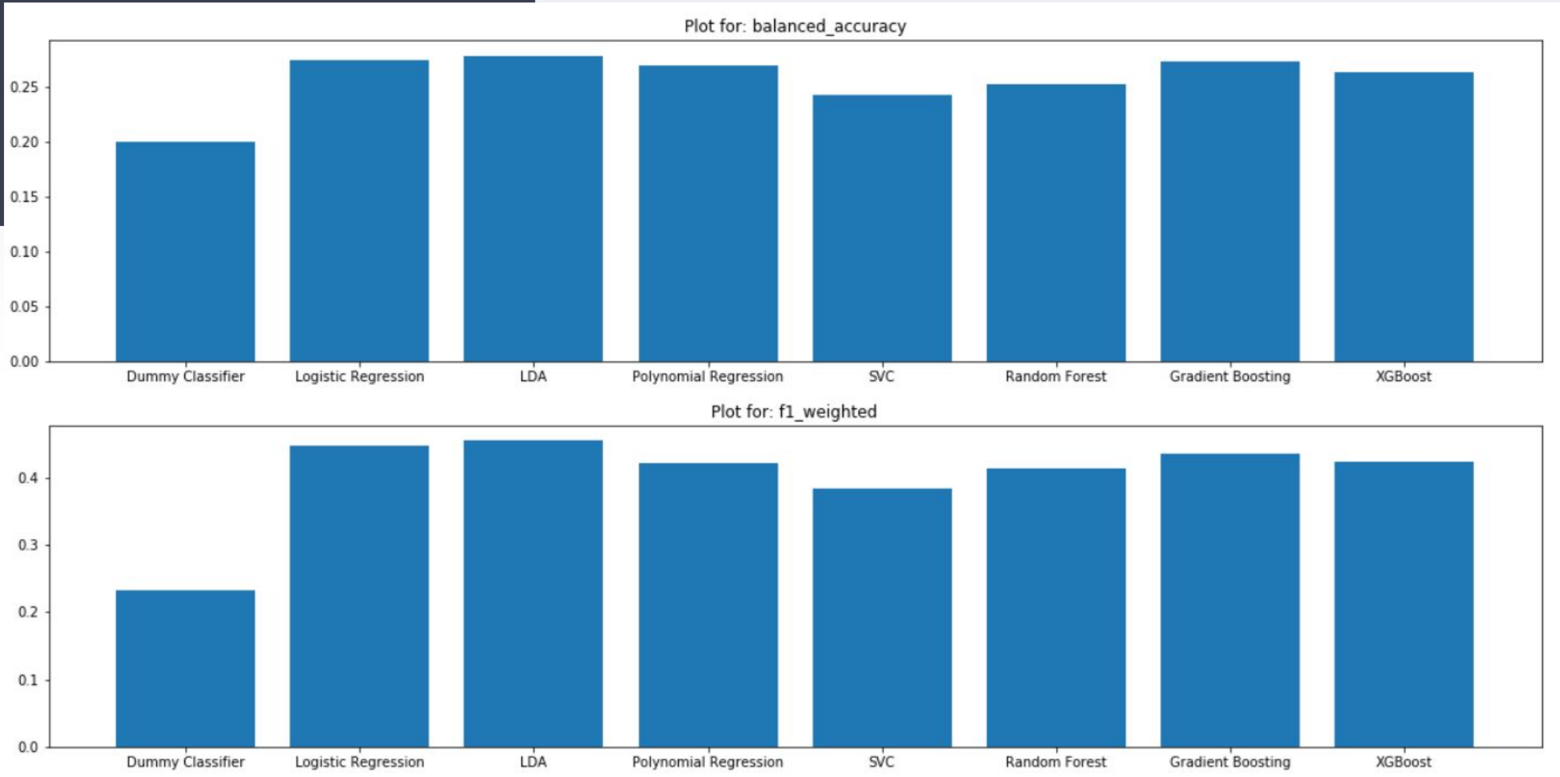
  Brier Score, Precision-Recall AUC and F1 Score(weighted)

Model Selection

Used scikit-learn's cross_validate function to compare algorithms

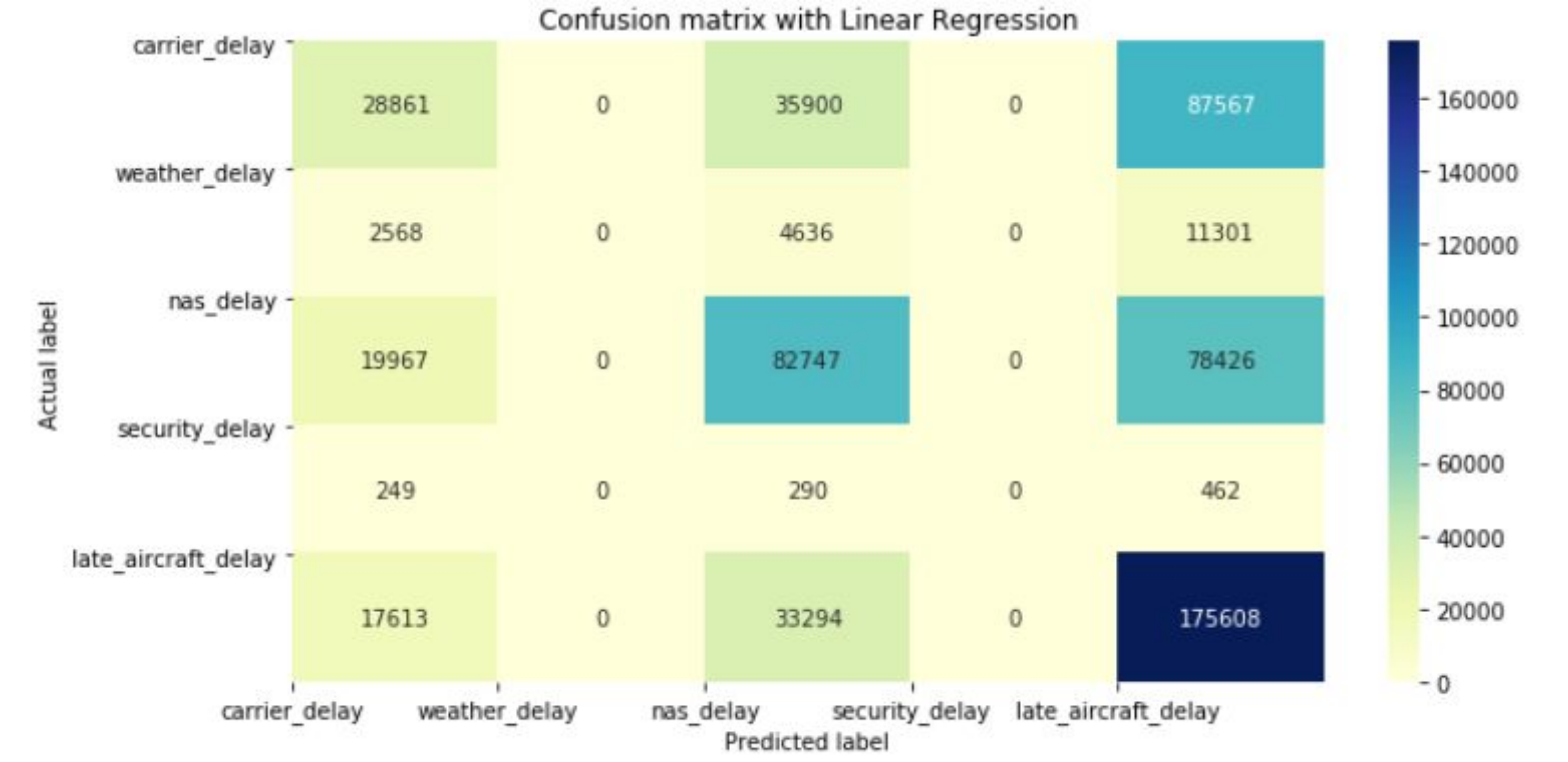
Tests performed with Linear, SVM and Ensemble Models

- Regression : XGBoost
- Multiclass Classification : LDA
- Binary Classification : Logistic Regression


Plot for: balanced_accuracy
0.00
0.05
0.10
0.15
0.20
0.25
Dummy Classifier
Logistic Regression
LDA
Polynomial Regression
SVC
Random Forest
Gradient Boosting
XGBoost
Plot for: f1_weighted
0.0
0.1
0.2
0.3
0.4
Dummy Classifier
Logistic Regression
LDA
Polynomial Regression
SVC
Random Forest
Gradient Boosting
XGBoost

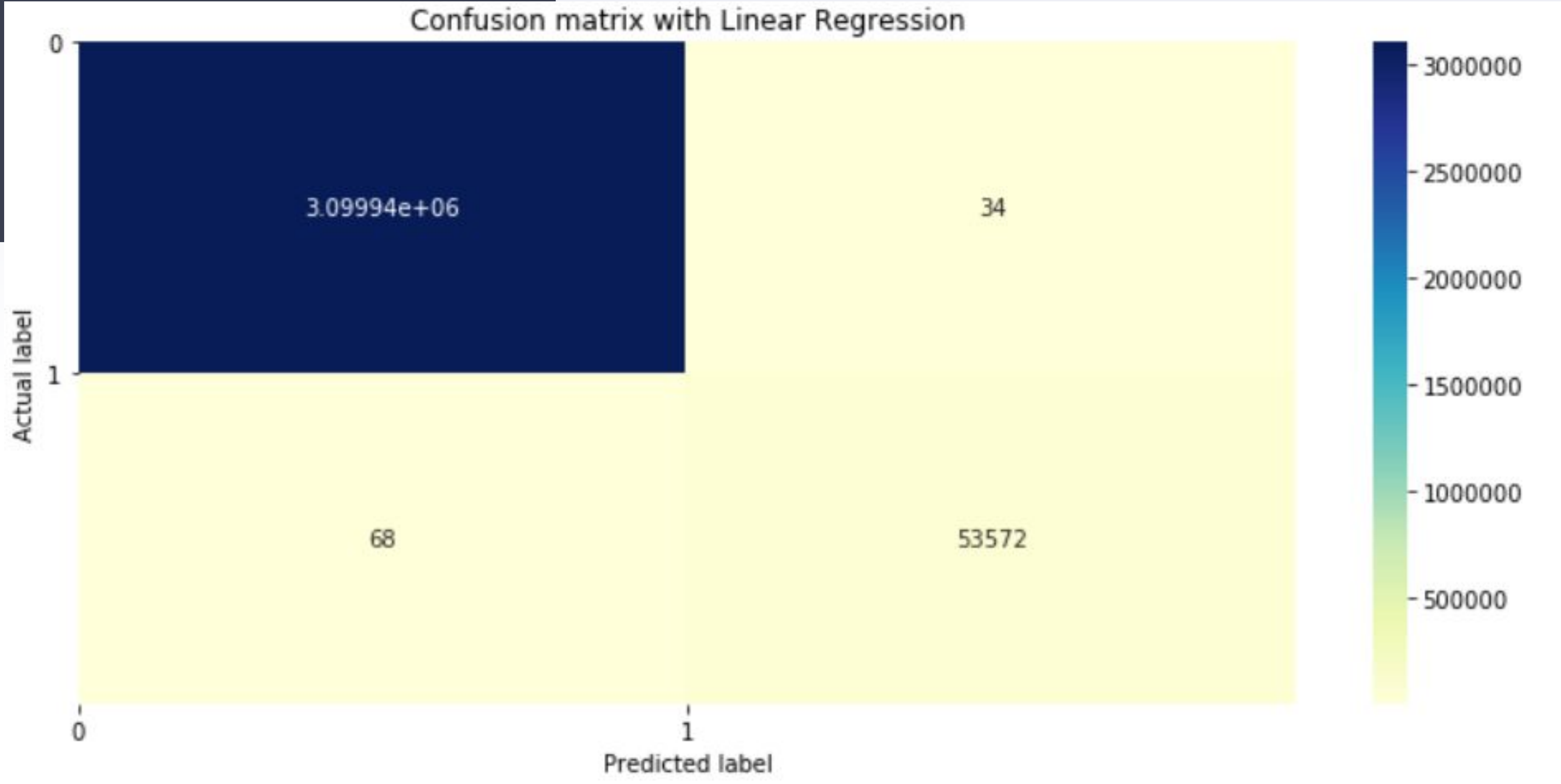
Model Performance

Some visualizations of the output we obtained in our modeling process



Confusion matrix with Linear Regression

Actual label \ Predicted label	carrier_delay	weather_delay	nas_delay	security_delay	late_aircraft_delay
carrier_delay	28861	0	35900	0	87567
weather_delay	2568	0	4636	0	11301
nas_delay	19967	0	82747	0	78426
security_delay	249	0	290	0	462
late_aircraft_delay	17613	0	33294	0	175608



Confusion matrix with Linear Regression

3

Conclusions and Next Steps

Challenges

- Remote Collaboration and Version Control
  - File Share (large .csv)
- Data Prep and Feature Engineering
  - Time Required for All Iterations
- Bug Resolution Progression
  - Propagated Errors

Lessons

This project showed the value of:

- Prototyping and Working with Samples
- Establishing a Plan
- Clear and Reusable Code
- Understanding the Data (targets and observations)

Next Steps

- Back to the sandbox on things we had particular challenges
- Prepare an "ideal" workflow for the changes in how we would approach the problem with what we know now
- Learning ways to integrate API data into model
- Learn more about scripts and packages

Any questions?

You can find our repo at

- github.com/Isabelle-Dr/MidTerm-Project

Mid-Term Project

Exploring the Air Travel Industry

Isabelle de Robert

& Maggie Fiander