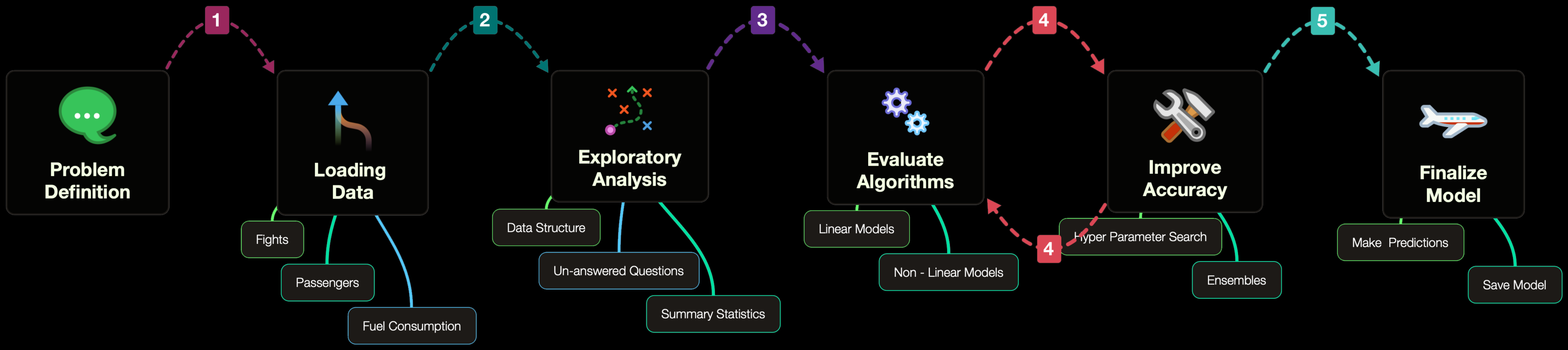
PREDICTING FLIGHT DELAYS

SUPERVISED MACHINE LEARNING

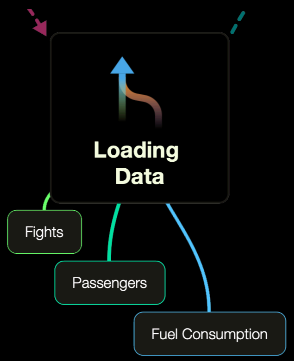
PROCESS



1
2
3
4
5
Problem Definition
Loading Data
Fights
Passengers
Fuel Consumption
Exploratory Analysis
Data Structure
Un-answered Questions
Summary Statistics
Evaluate Algorithms
Linear Models
Non - Linear Models
4
Improve Accuracy
Hyper Parameter Search
Ensembles
Finalize Model
Make Predictions
Save Model

WHAT IS THE PROBLEM?

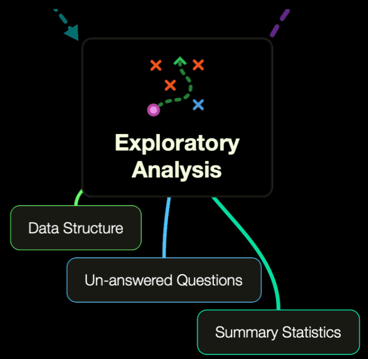
- Significant implications for airlines, affecting their profitability and customer satisfaction.
- Accurate estimation is crucial for airlines to make informed decisions and optimize their operations.
- Understanding the factors affecting flight delays is essential for developing accurate prediction models.



DATASETS

Four separate tables related to US the air travel industry.

- Flights — departure and arrival information 2018 and 2019.
- Fuel Consumption — different airlines from years 2015-2019 aggregated per month.
- Passengers — totals on different routes from years 2015-2019 aggregated per month.
- Flights test — test dataset for flights in January 2020

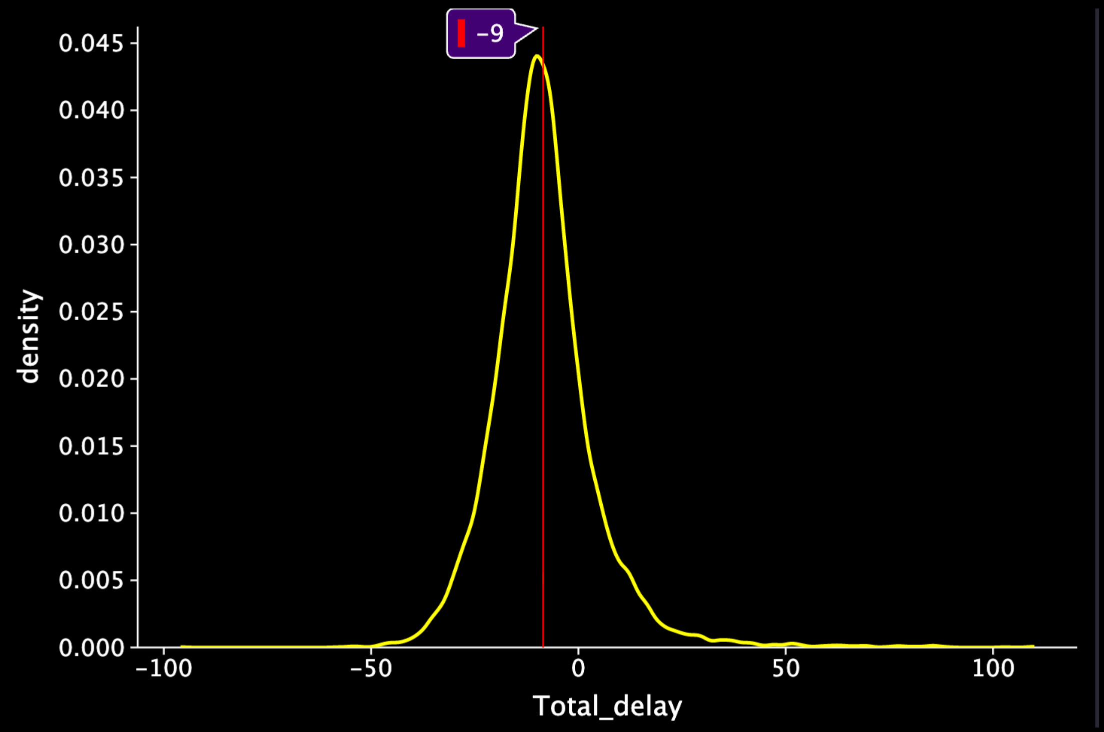


EDA

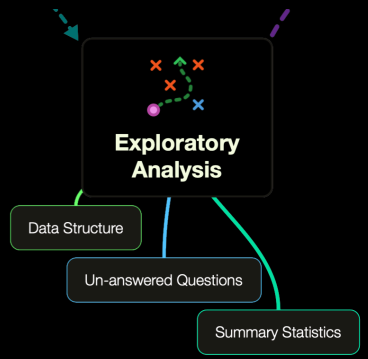
Is distribution of delays normal?

- By day of week, most flights leave on time, by on average 7 mins earlier.
- ATL was the busiest airport with 1602 flights with 179K passengers passing though in just one month.
- Flights that left late night were more likely to be delayed

Shapiro - Wilk Test

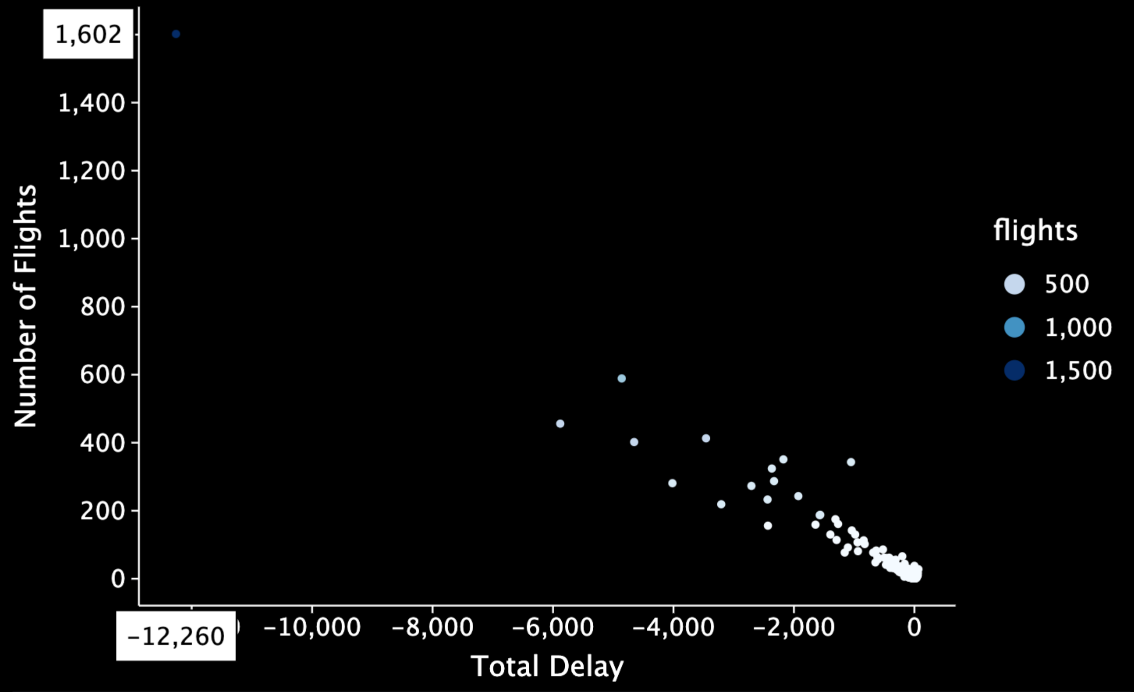


P > alpha = 0.05
Mean total delay is not 0 (-9 minutes)



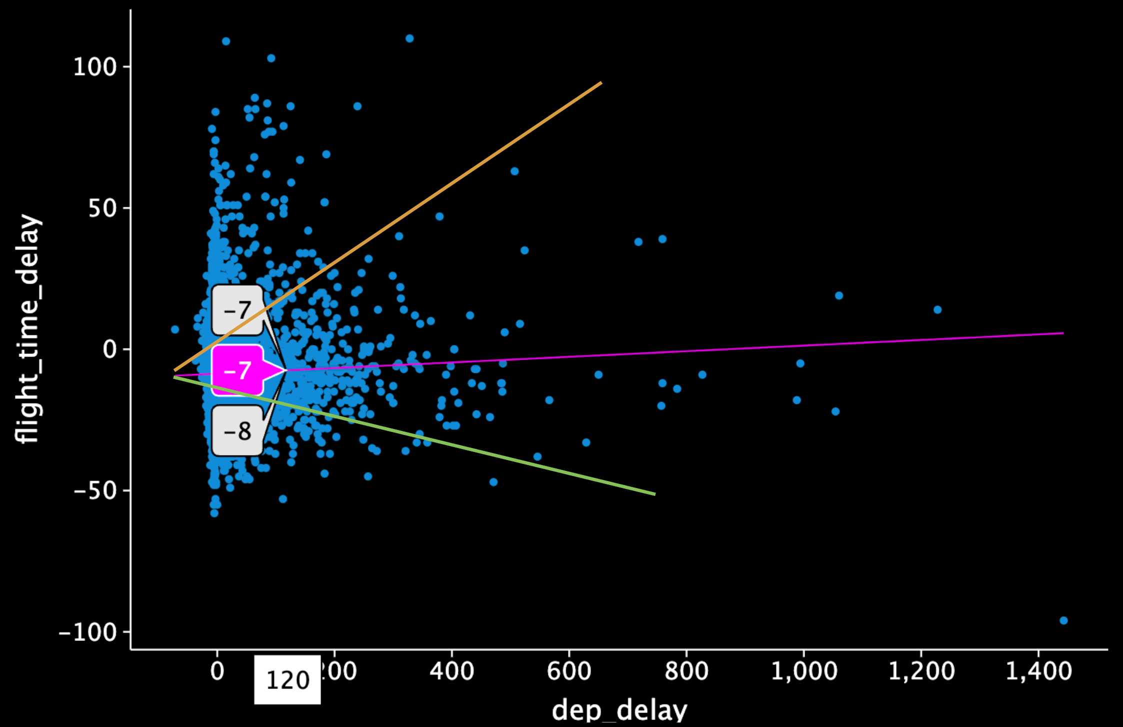
EDA

Delay Vs Number of Previous Flights

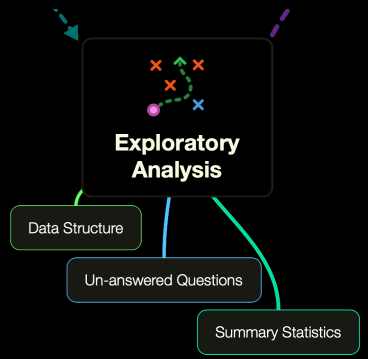


There is a strong association between total delay and Number of previous flights.

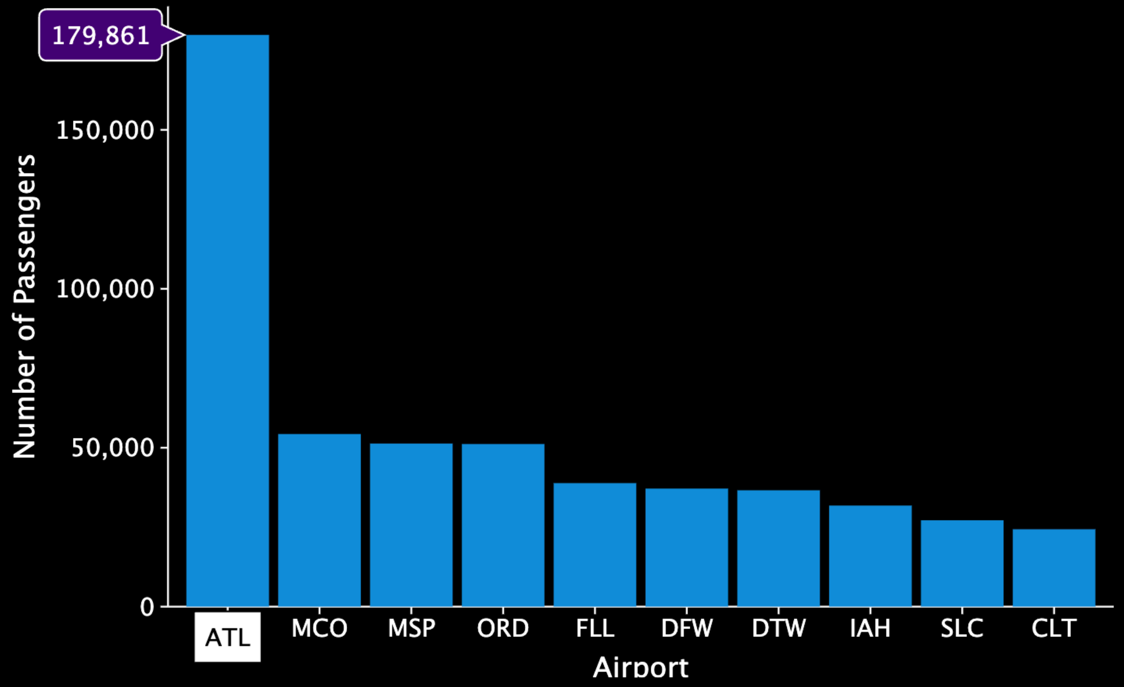
Will my pilot fly faster when departure was delayed?



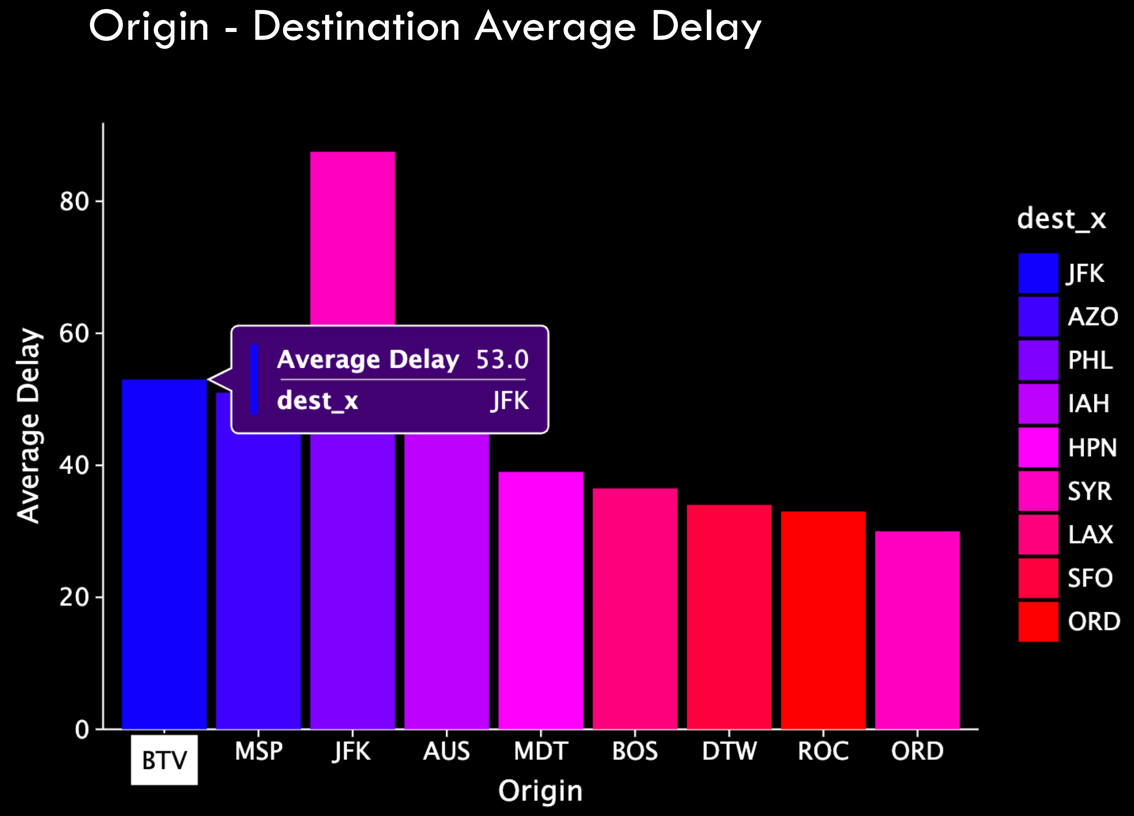
There is a positive linear relationship, but very Close to zero. At 2 hours late, they'll only fly faster By 7 minutes.



EDA



Atlanta - home to Delta Airways was the busiest airport 179,000 passengers and 1,602 flights in one month



Flying between BTV and JFK has highest average Delay of 53 minutes. Worse combination possible

FEATURE IMPORTANCE

- Origin city average departure delay (both origin and dest).
- Day of week.
- Day of week average departure delay (both origin and dest).
- op_unique_carrier average arrival delay.
- Airtime avg of distance group.
- Number of passenger average of distance group.
- Payload average of distance group.

MODEL SELECTION

- Grid Search for finding best hyper-parameters.
- Use Ridge and Lasso Model.
- Use three Ensemble Techniques : 1) Random Forest Regressor 2) Gradient Boosting 3) XGBoost
- In which Random Forest Regressor gives minimum mean squared error
- Use linear regression model and find summary in which we observed
  1) R2 of model is 0.1031937
  2) Saturday has largest positive coefficient.
  3)Monday has smallest negative coefficient.

CHALLENGES

- Weather related data was not readily available.
- Disjoint in table keys. The most important key (flight number) was missing in the flight dataset. When merged we lost a good chunk of data.

CONCLUSION

- An okay project - not the results we were expecting.
- Sample across multiple years, data enrichment.

CONCLUSION