Predicting Flight Delays

Jagvir & Noah

01

Project Flow

Steps Taken

1. Database exploration by SQL Query
2. Exploratory Data Analysis in Python
3. Building a pipeline

- automating data retrieval/cleaning in python

4. Modelling cycle
   - Feature engineer
   - Model
   - Evaluate
   - Optimize
   - Repeat

Database Exploration

Addressing large scale flights data

Understanding the scale:

Query for size of flights table:
4.267GB

Query metadata for estimate of rows:
15`207`047 rows (from pgclass)

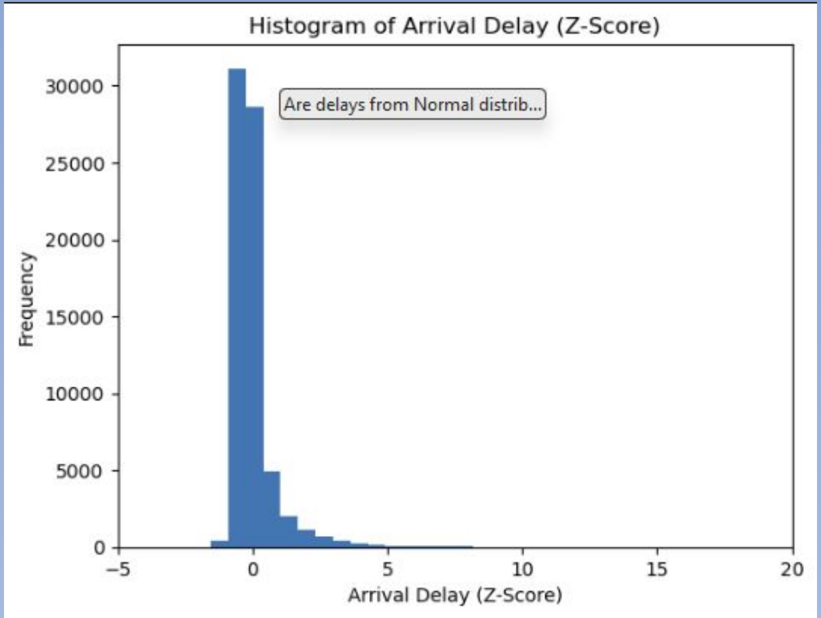
Accounting for scale:

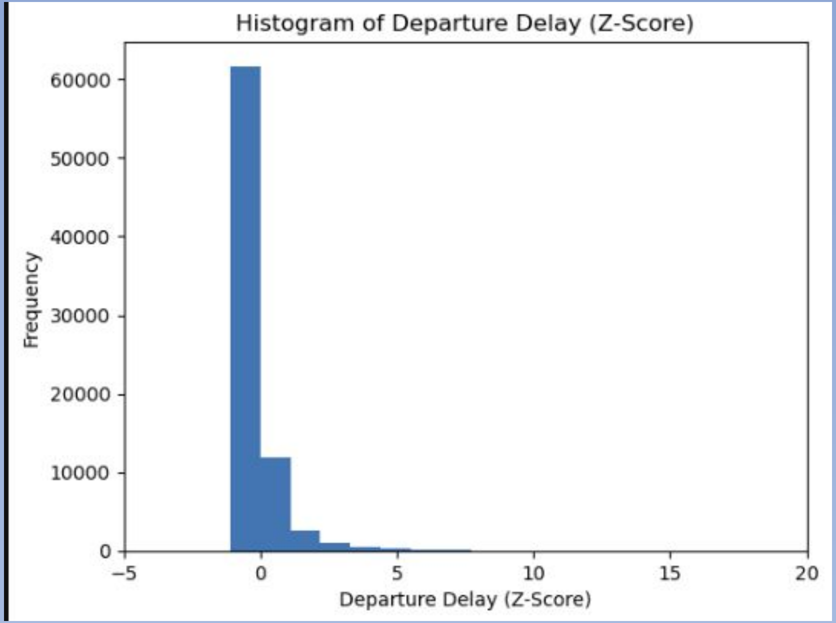
Random sampling of **100`000** flights records (to begin)

Coding with upscale in mind

EDA Relationships

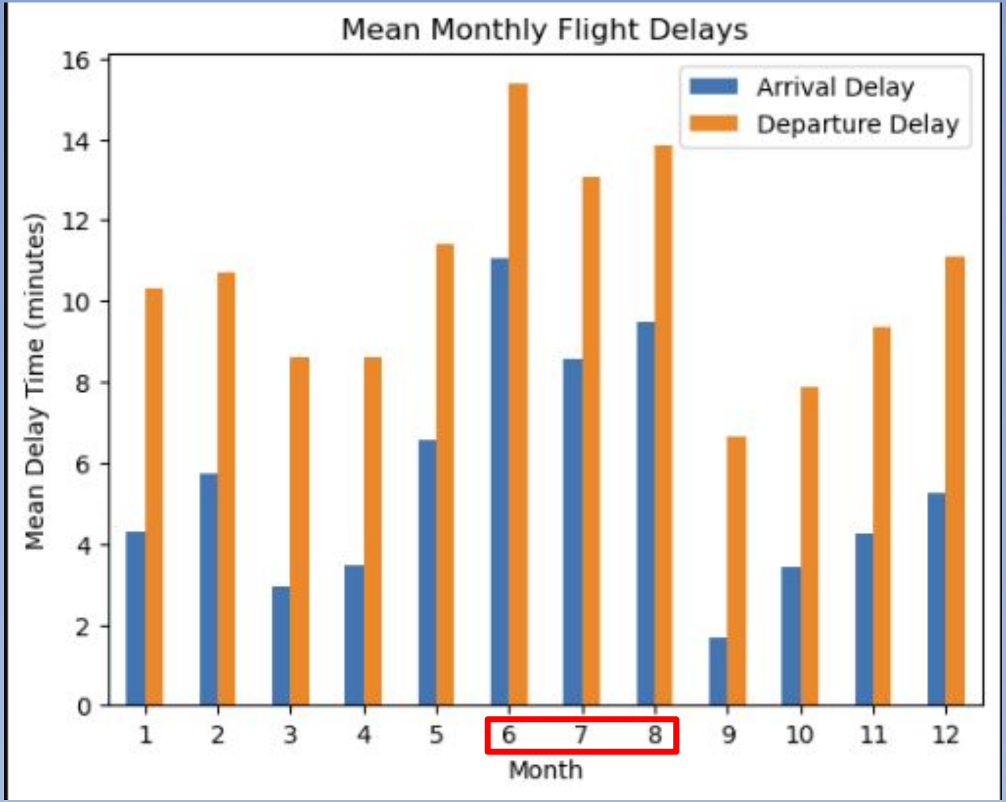
Are the delays from normal distribution and the mean of the delay is 0?



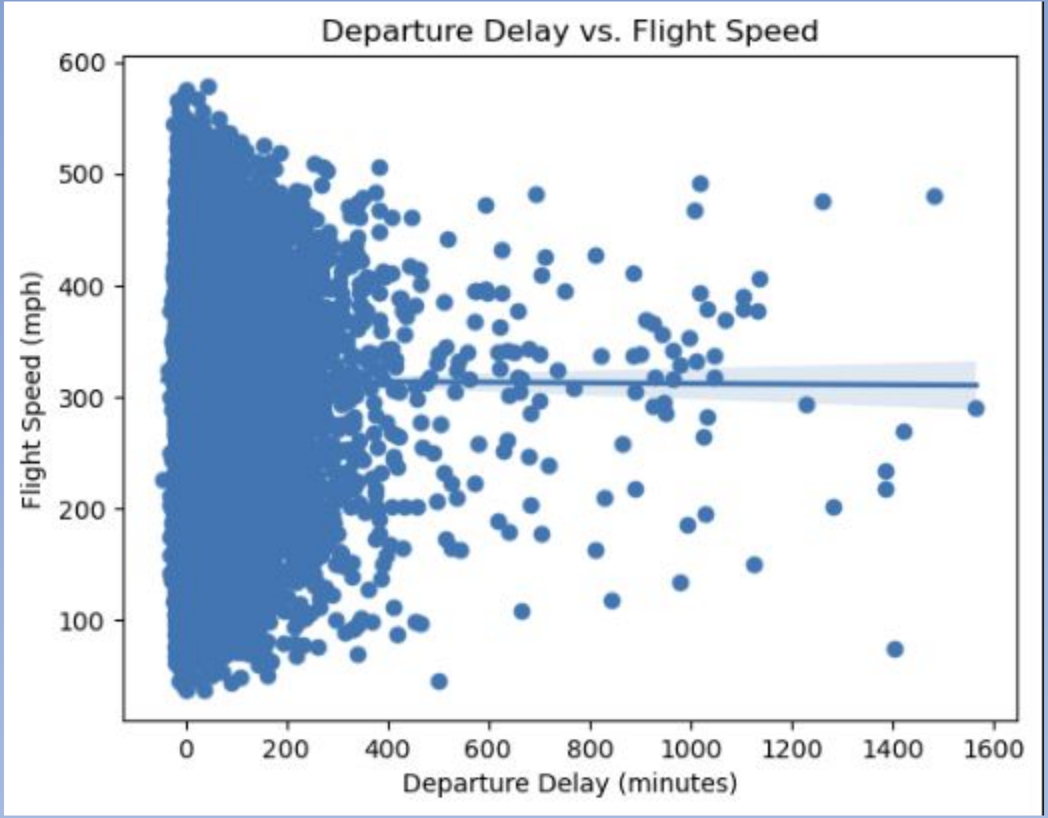


EDA Relationships

Is average/median monthly delay different during the year?



EDA Relationships



Test the hypothesis whether planes fly faster when there is the departure delay?

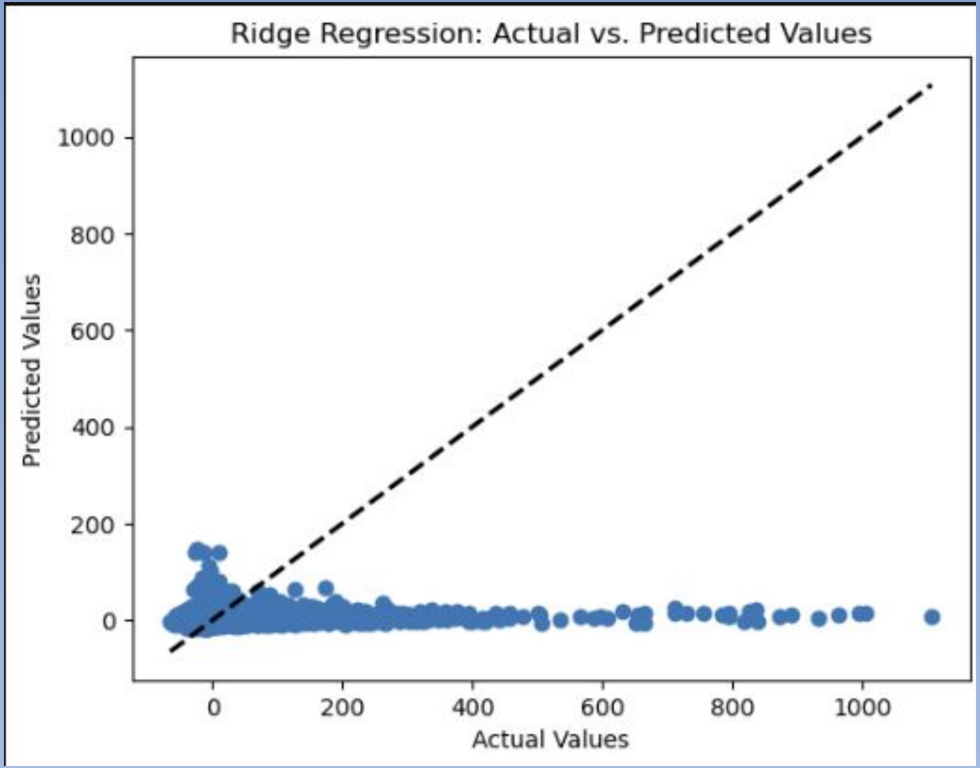
02

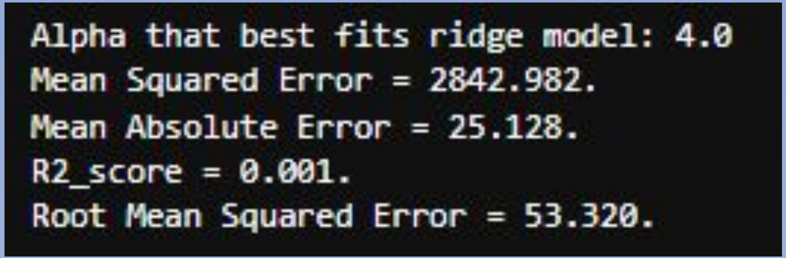
Results

Feature Importance

- Finding missing values and imputing with appropriate values
- Dropping redundant columns
- Parse ‘fl_date’ as datetime and extracting year, month, day
- Encoding categorical features
- Scaling numerical features (StandardScaler)

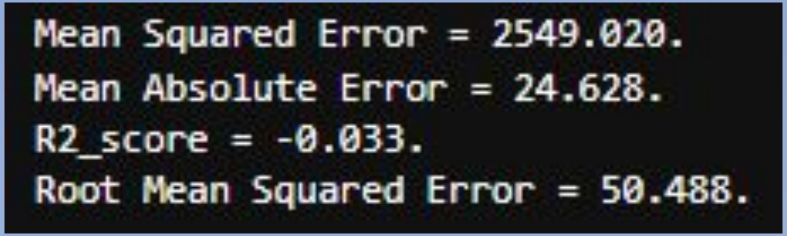
Ridge Regression Model



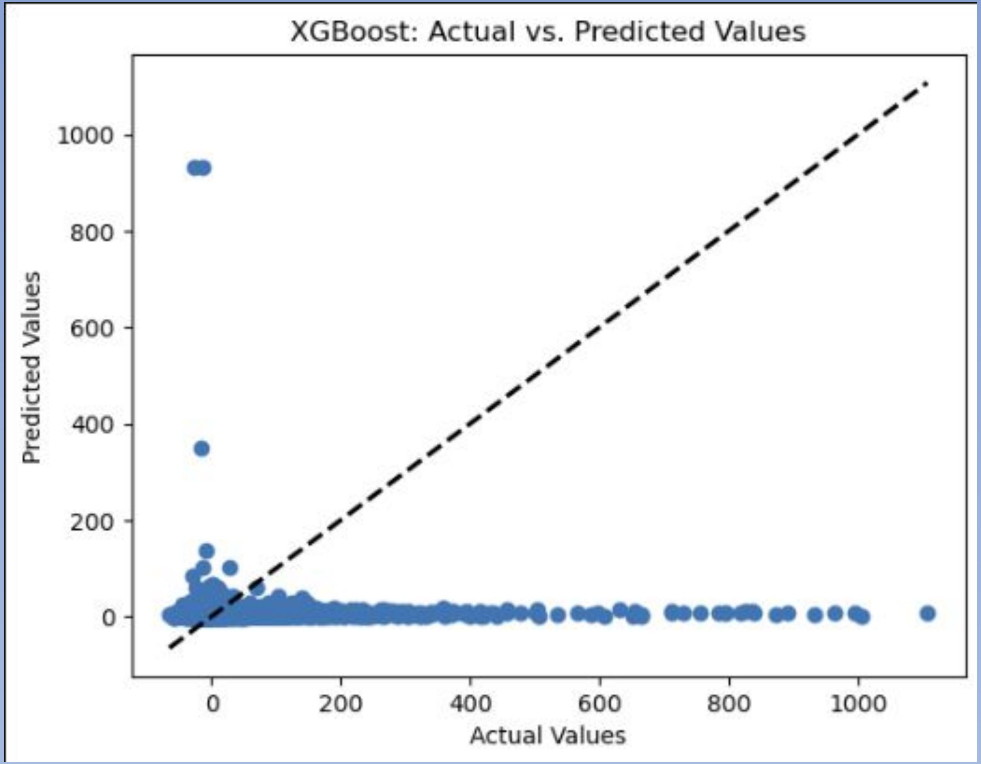


```
Alpha that best fits ridge model: 4.0
Mean Squared Error = 2842.982.
Mean Absolute Error = 25.128.
R2_score = 0.001.
Root Mean Squared Error = 53.320.
```

XGBoost Model



```
Mean Squared Error = 2549.020.
Mean Absolute Error = 24.628.
R2_score = -0.033.
Root Mean Squared Error = 50.488.
```



03

Challenges & Future

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**