FLIGHT DELAYS

Prediction

Rajwant Kaur, Philipp Werdenbach, Oliver D. Mücke

Table of Contents

Introduction

Why are flights delayed?

- Aircraft delay from previous flight
- Extreme weather conditions
- Air traffic control restrictions
- Waiting for crew or staff strike
- Mechanical issues
- Bird strike etc…

Objective

Goal:

To predict the estimated duration of **flight delay in minutes** for each flight mainly flying from or to the Tunis-Carthage International Airport
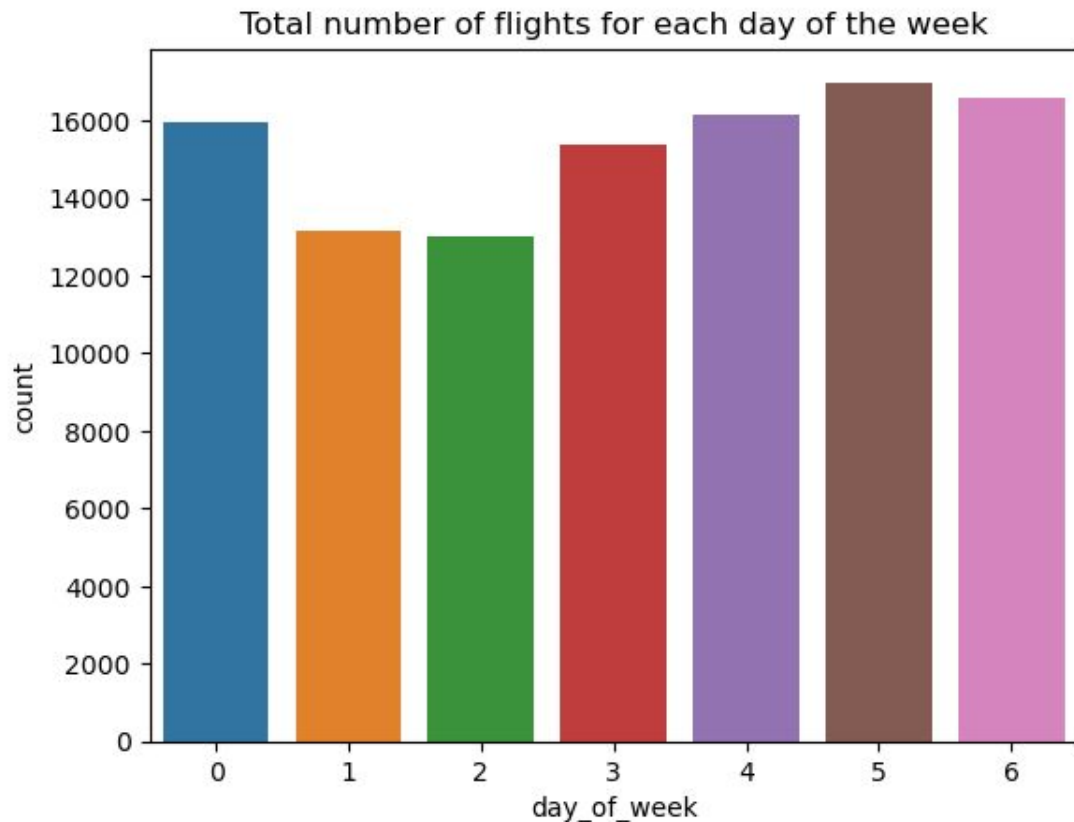
Why do we need predictions?

- To decrease of capital costs caused by reallocation of flight crews and aircraft
- To minimize the negative impact on passenger demand

Data Analysis

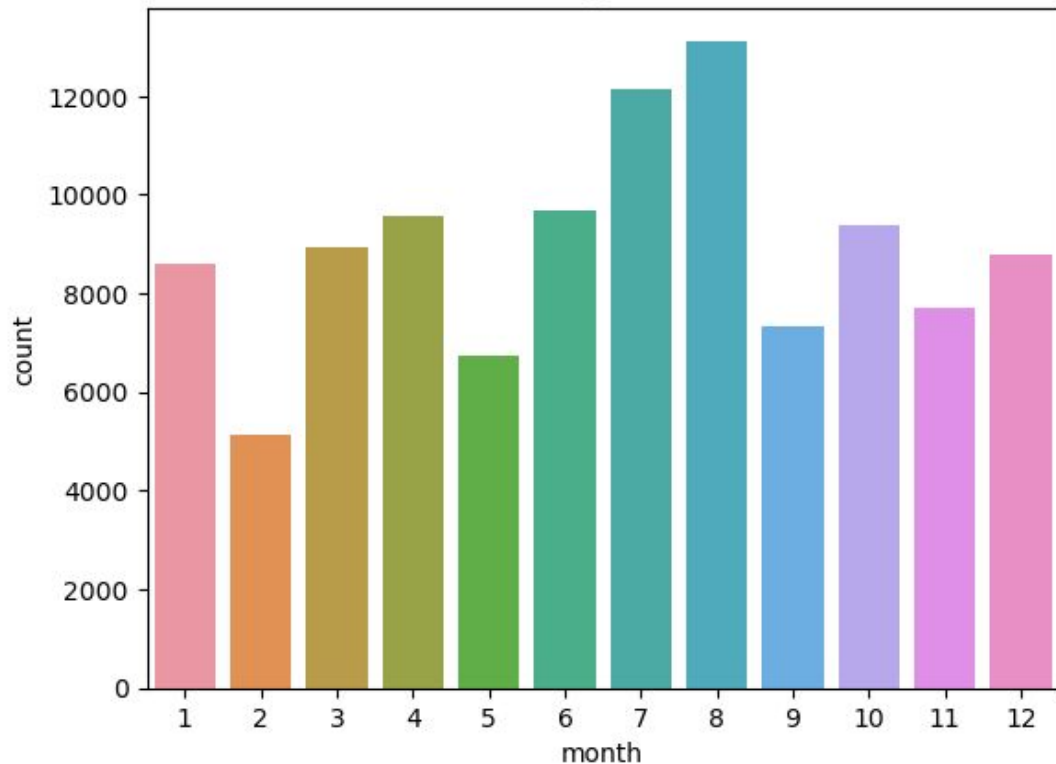- years 2016 to 2018
- ca. 100,000 samples
- 9 attributes

Data Analysis

Total number of flights for each day of the week
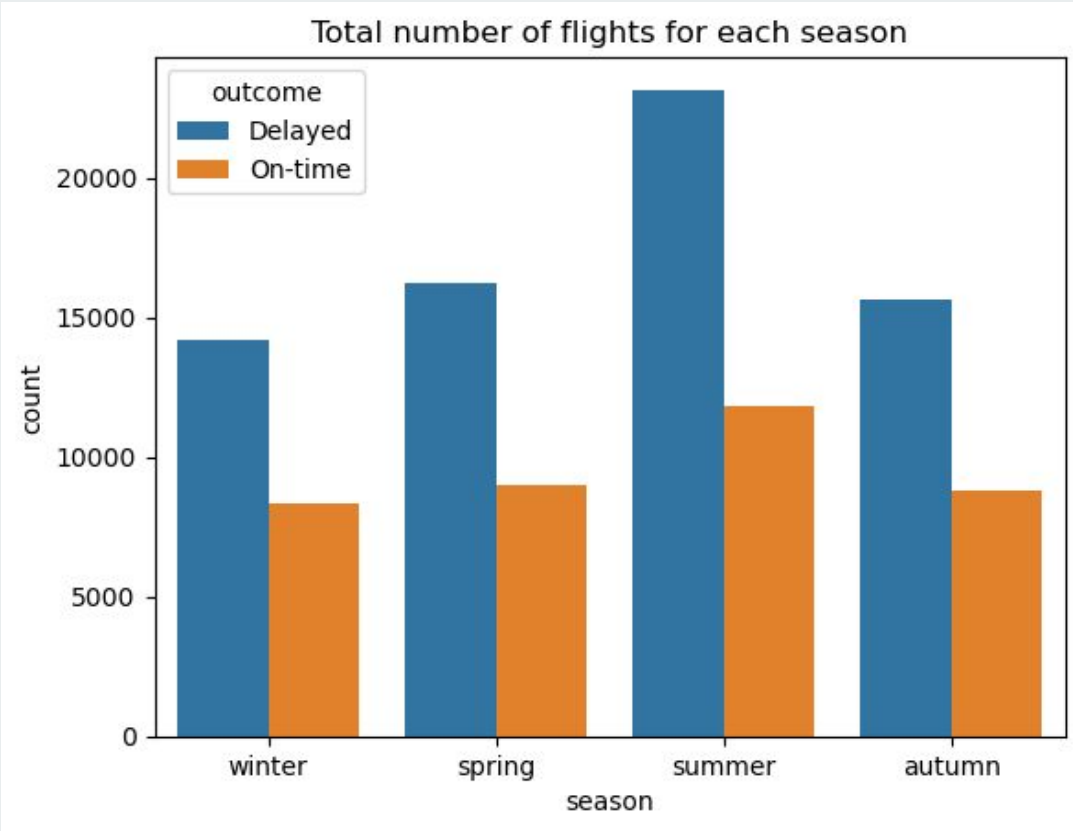
slightly more flights on weekend

Data Analysis

Total number of flights for each month

More flights in the summer months July and August

Data Analysis

Total number of flights for each season
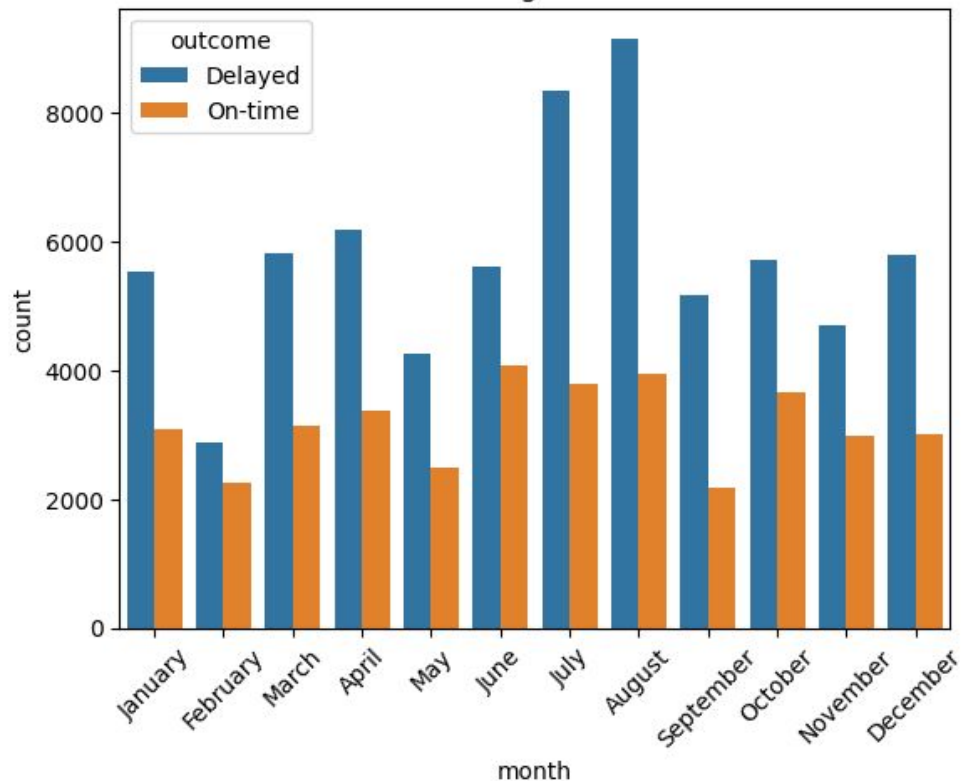
Absolute number of delayed flights is larger in the summer season
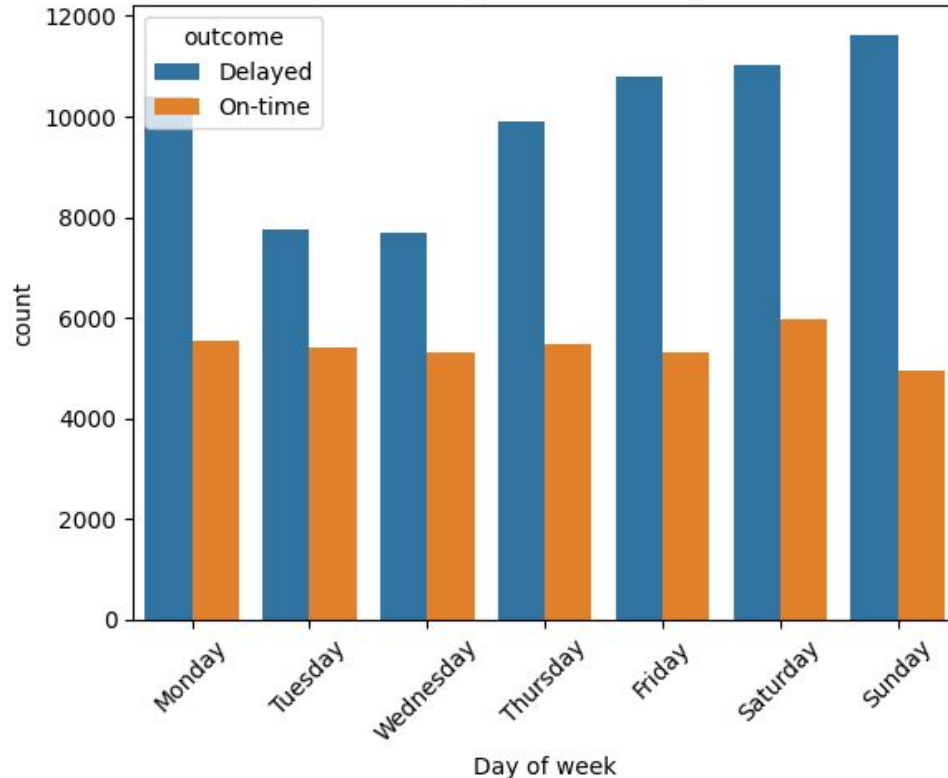
Data Analysis

Total number of flights for each month

Absolute number of delayed flights is larger for months July and August
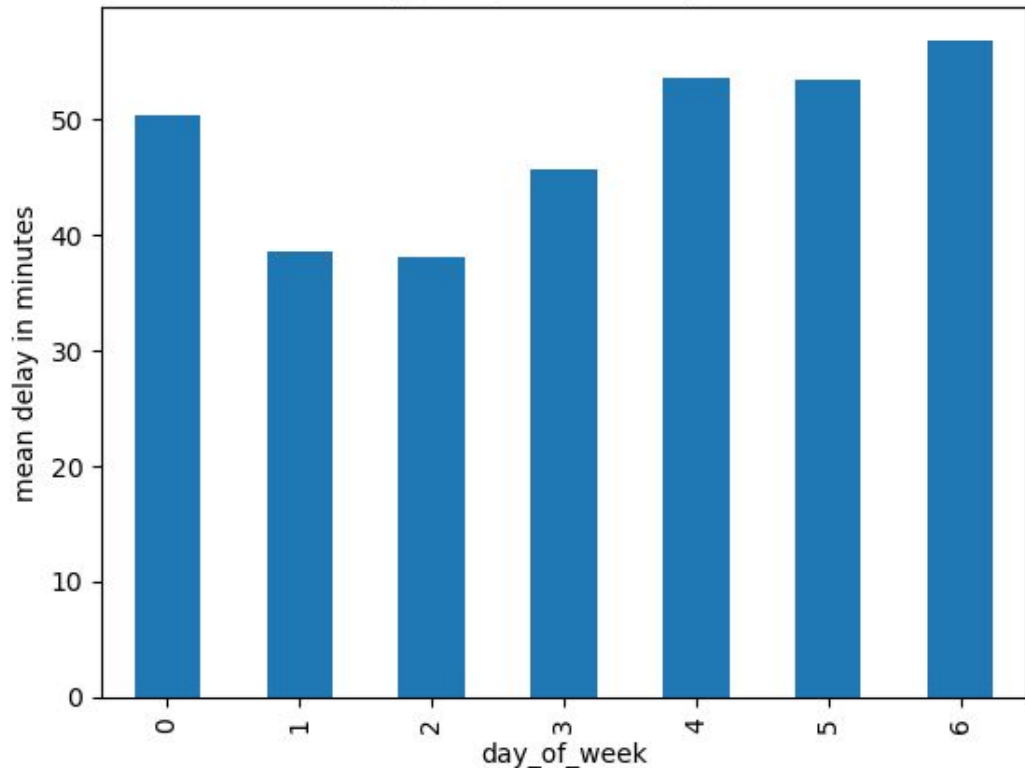
Data Analysis

Total number of flights for each day of the week

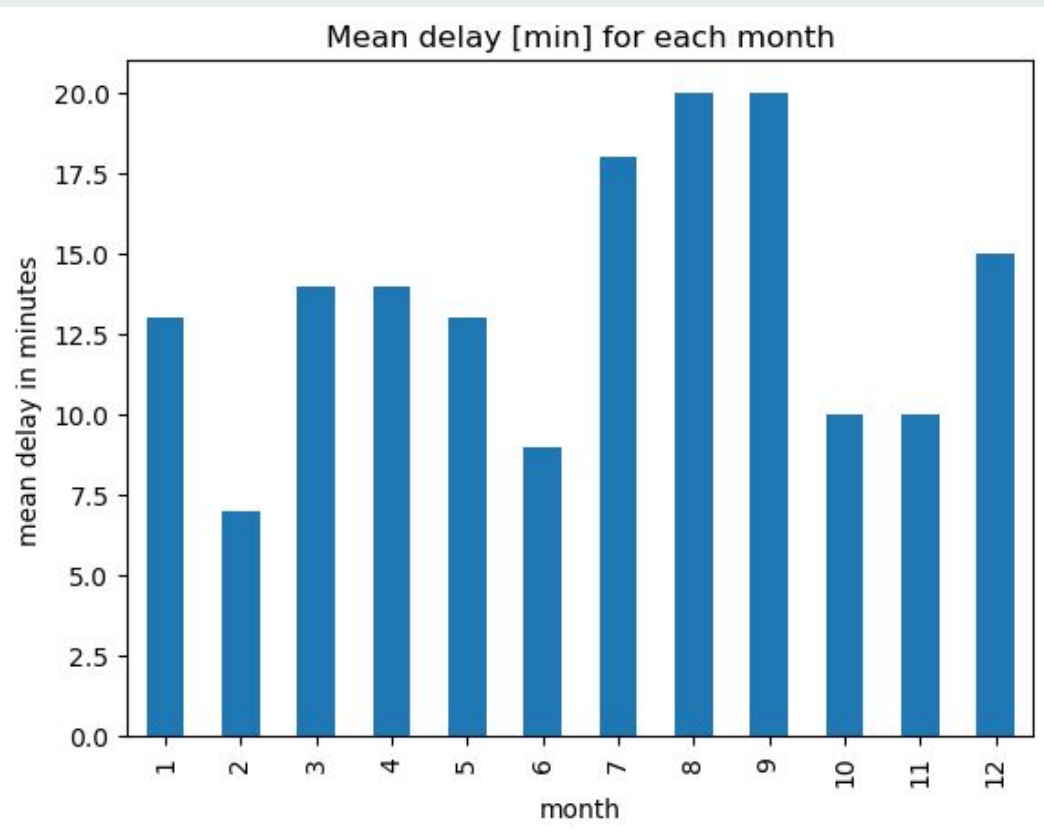Absolute number of delayed flights is smallest for Tuesdays and Wednesdays, increases afterwards

EDA

Mean delay [min] for each day of the week

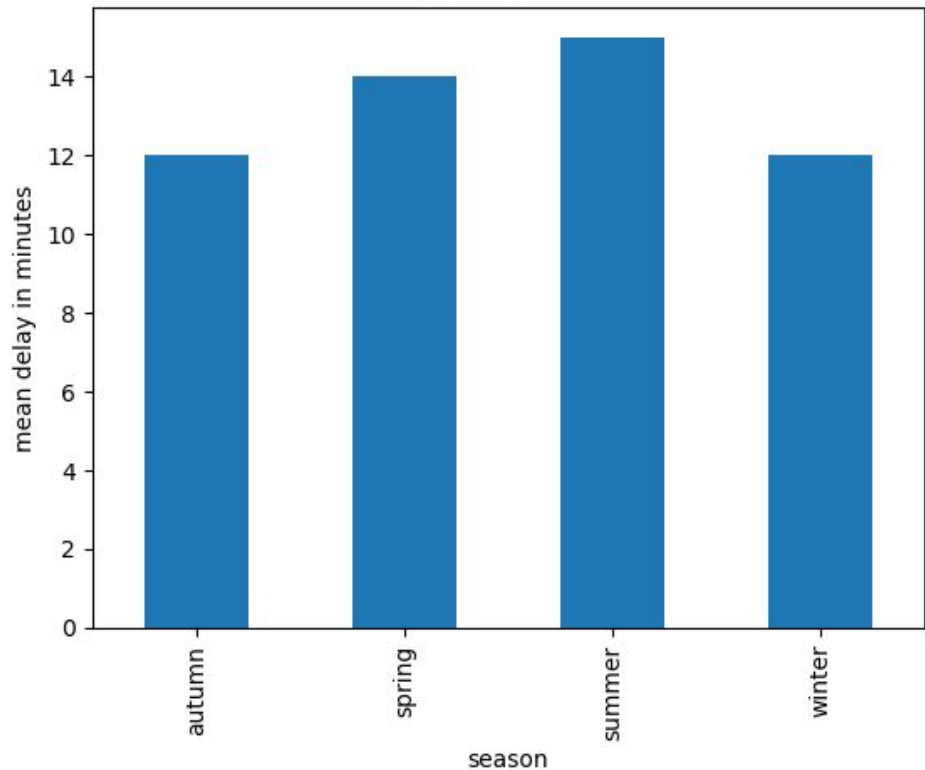Mean delay is largest on Sundays and smallest on Tuesdays/Wednesdays

EDA

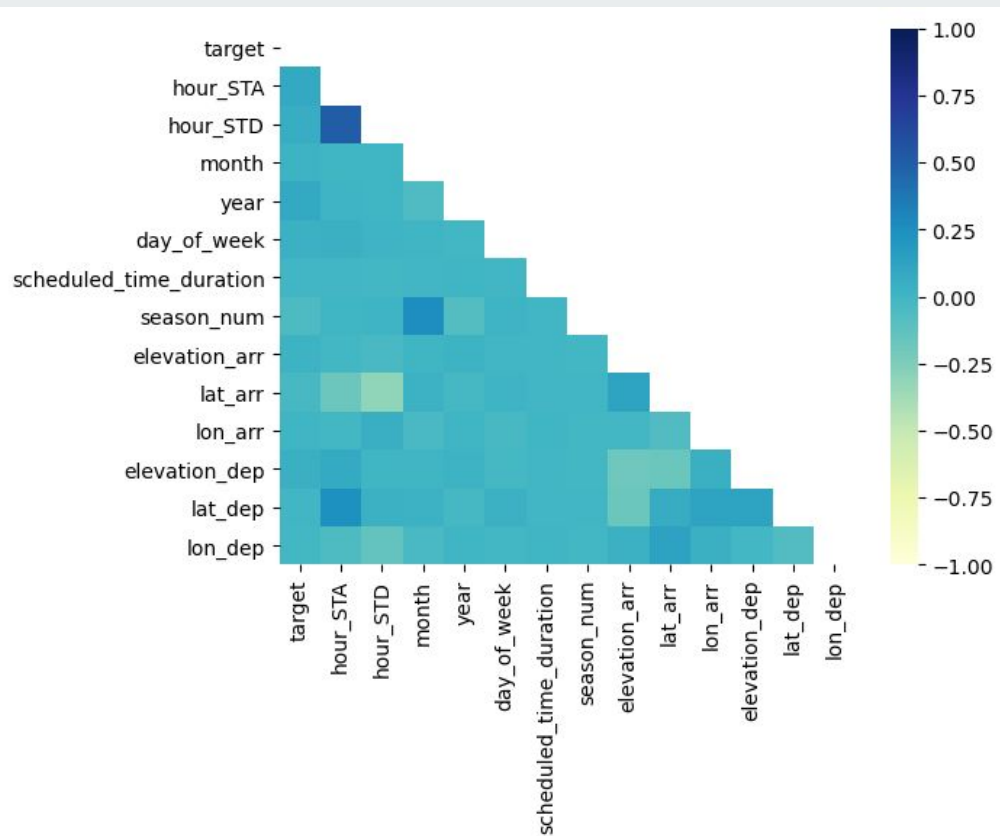Mean delay is largest in July-September

EDA

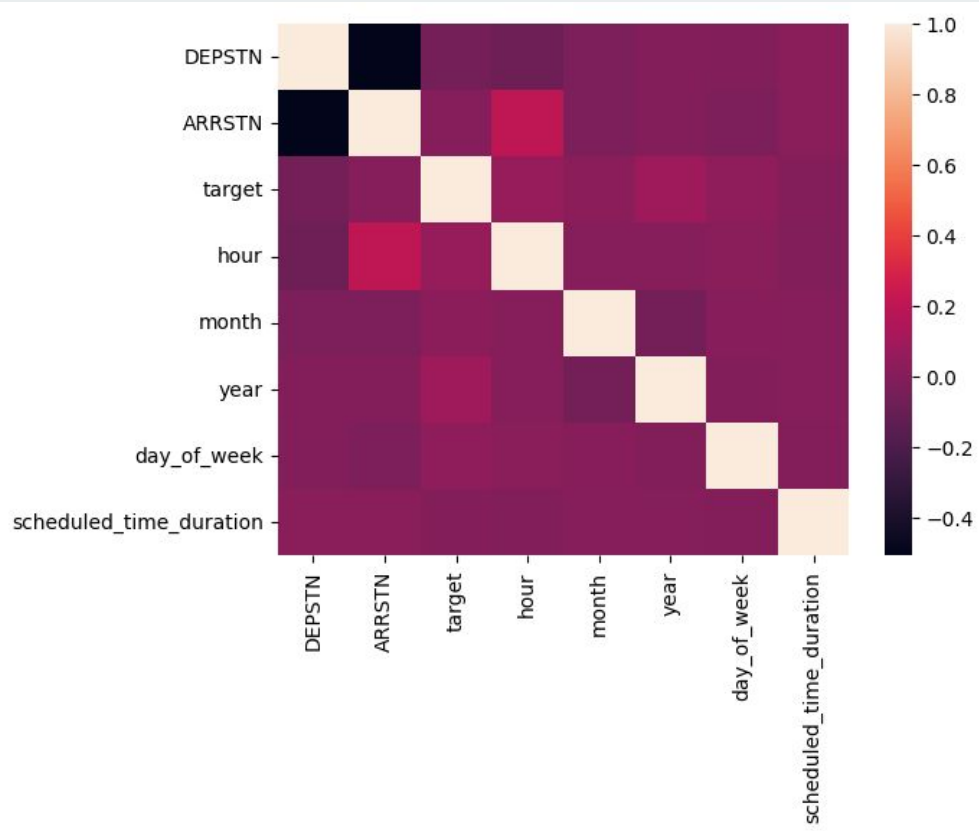

Mean delay [min] for each season
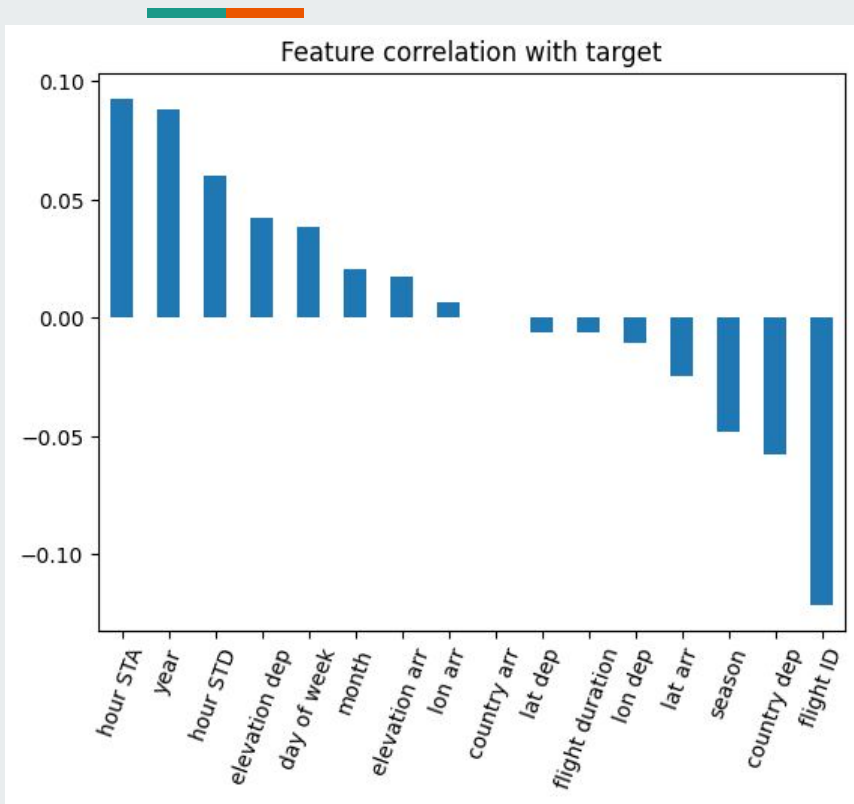
Mean delay is largest in summer

Data Analysis

Attributes are mainly uncorrelated and the correlation with target is very small (<0.1)

EDA

Data Analysis

Feature correlation with target

Attributes are mainly uncorrelated and the correlation with true label is very small (<0.15)

Predictions

How good are Models?

- Lower the values of Error ⬇
- Closer the predicted values are to the test 〰
- Better is the model 👍

Our Models:
- Regression analysis performed
- Different regressors/models are chosen
- Error analysis

Predictions

Linear regression (base model)   RMSE (test): 112.60 min

Linear regression (optimized)   RMSE (test): 111.311 min

Decision Tree   RMSE (test): 110.675 min

XGBoost   RMSE (test): 105.045 min

Predictions - Average

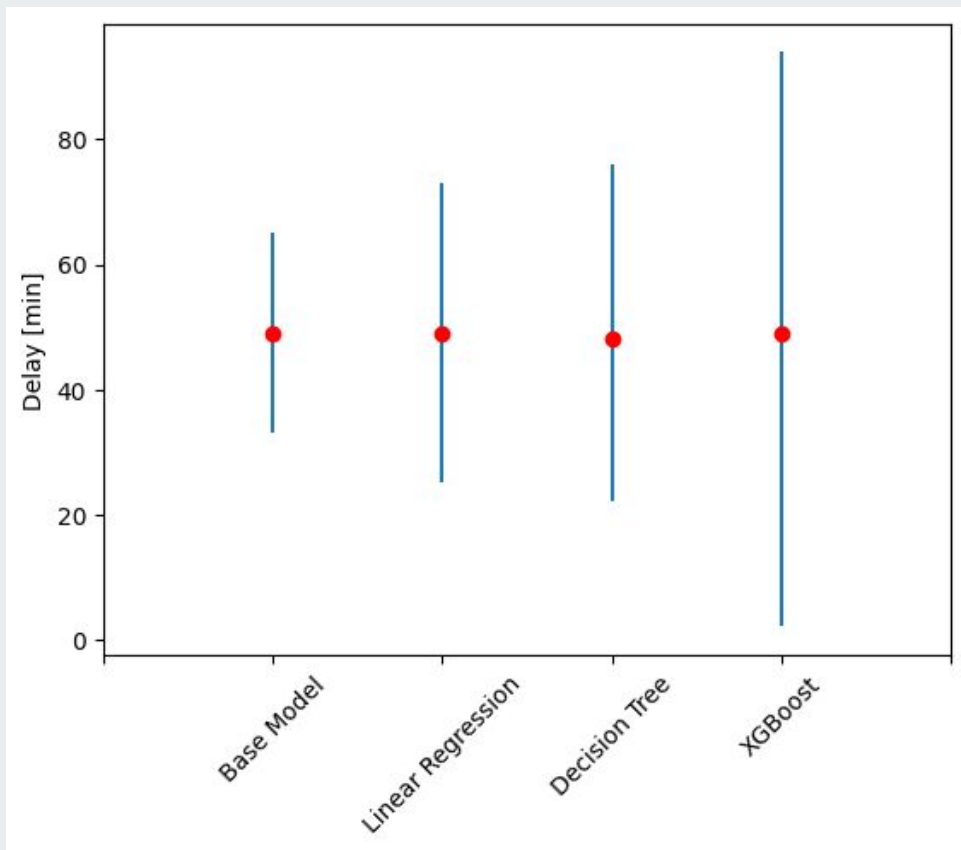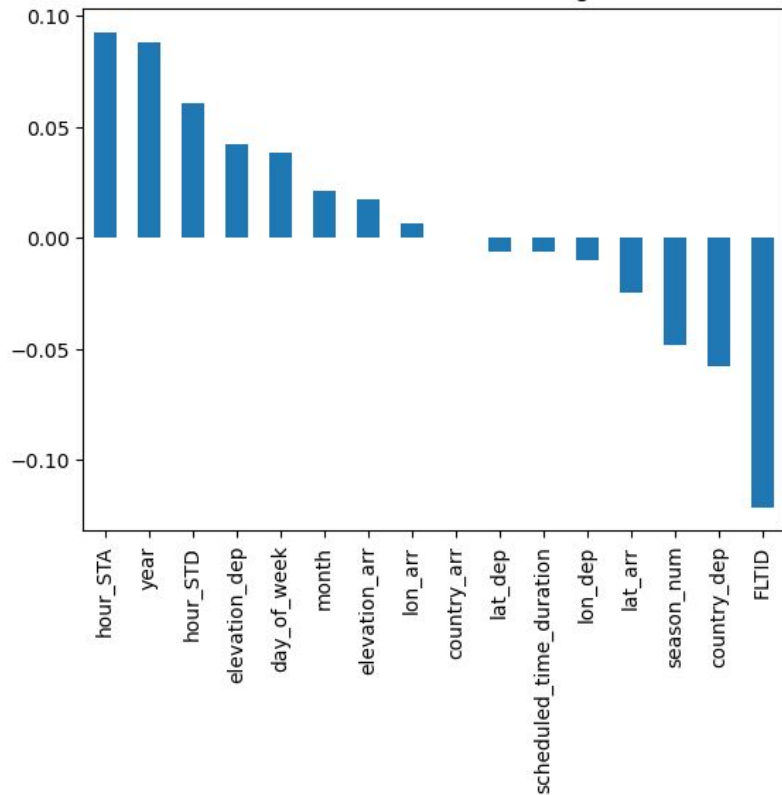|	
Actual delay	48 min
Linear Regression (base model)	49 min (+/- 16 min)
Linear Regression (optimized)	49 min (+/- 24 min)
Decision Tree	49 min (+/- 27 min)
XGBoost	48 min (+/- 46 min)

Predictions - Average
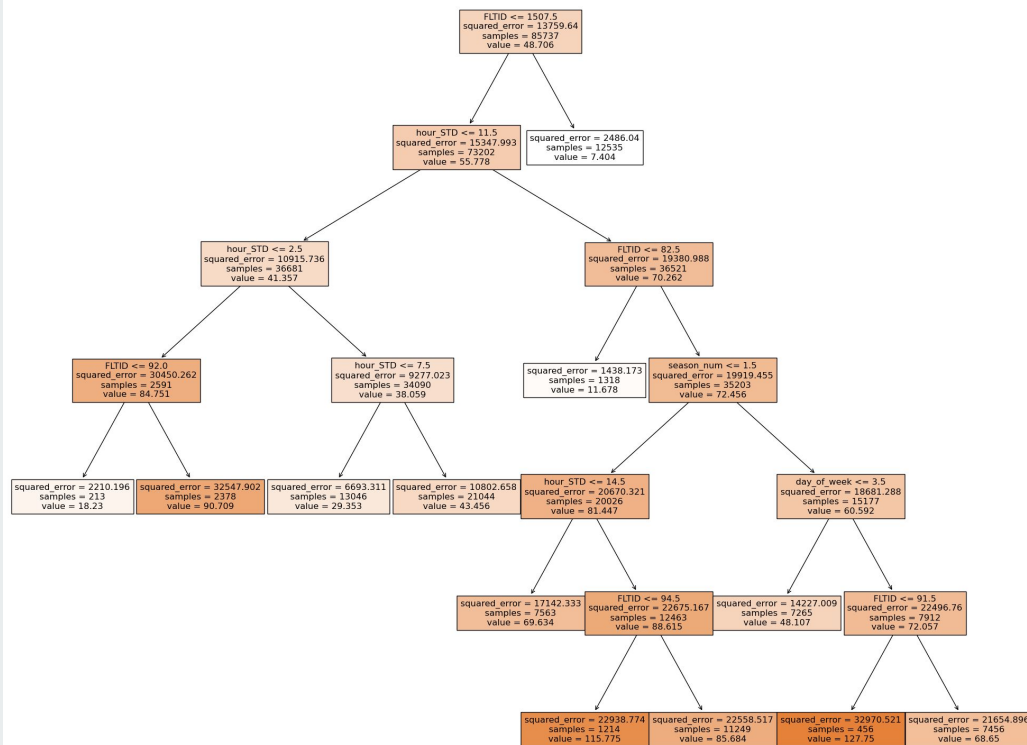
Decision tree for regression

Feature correlation with target

Overall the features show only weak correlation with the target. Despite, the features having most influence are:

- FLTID (-0.121)

- hour_STA (+0.093)

- year (+0.088)

- hour_STD (+0.060)

- country_dep (-0.058)

- season (-0.048)

- elevation_dep (+0.042)

- day_of_week (0.039)

Decision tree for regression

with Grid search:

R-squared (train): 0.053
R-squared (test): 0.052
RMSE (train): 114.125 min
RMSE (test): 110.675 min

XGBoost regressor

Metrics on training data
RMSE: 97.88 min
R2: 0.315

Metrics on test data
RMSE: 105.045 min
R2: 0.083

Milestone 3: Error Analysis

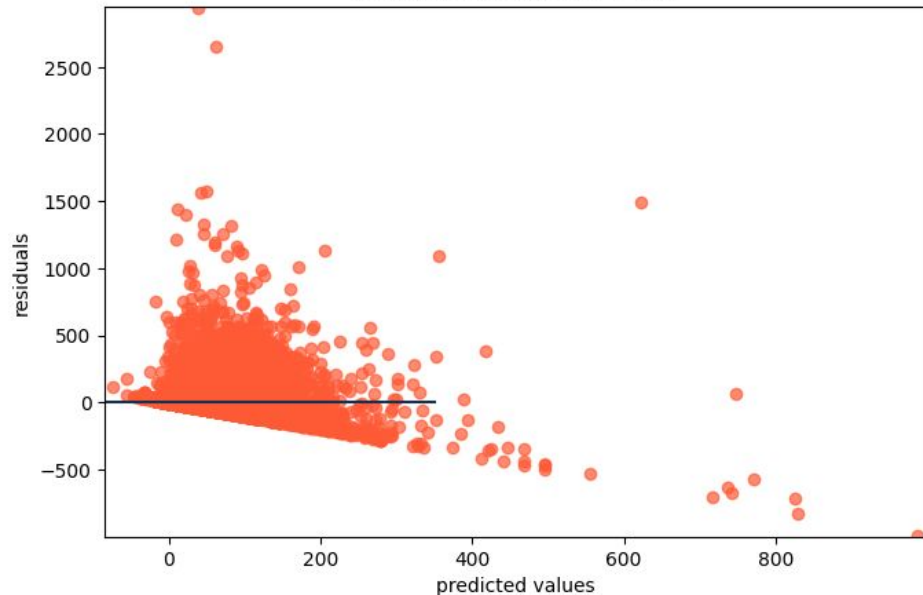**Model: XGBRegressor**

RMSE(test): 100.2 minutes
RMSE(train): 98.7 minutes

Error Analysis
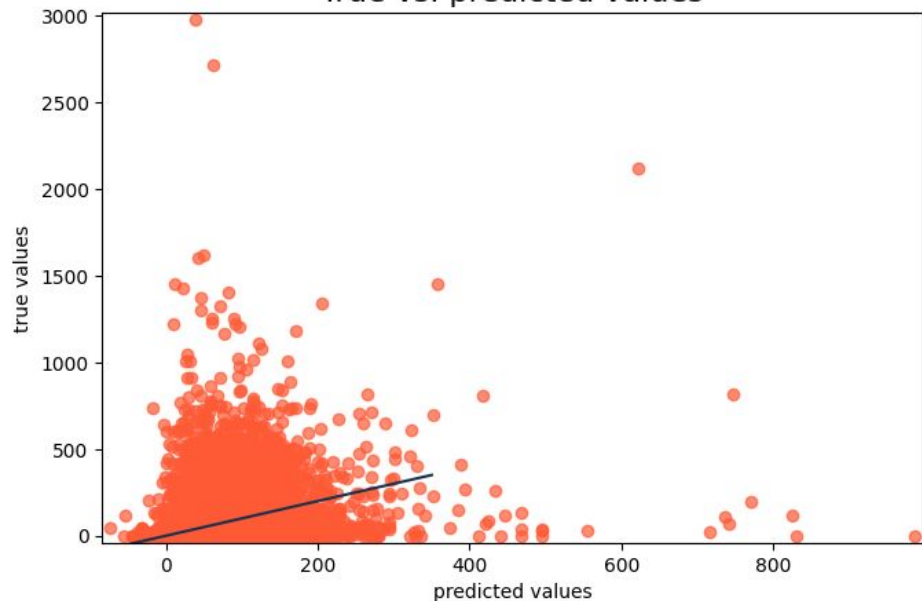
Conclusions

- Predictions contain large errors

- Attributes are no good indicator for delay

- More meaningful attributes might be gathered