FLIGHT PREDICTION PROJECT

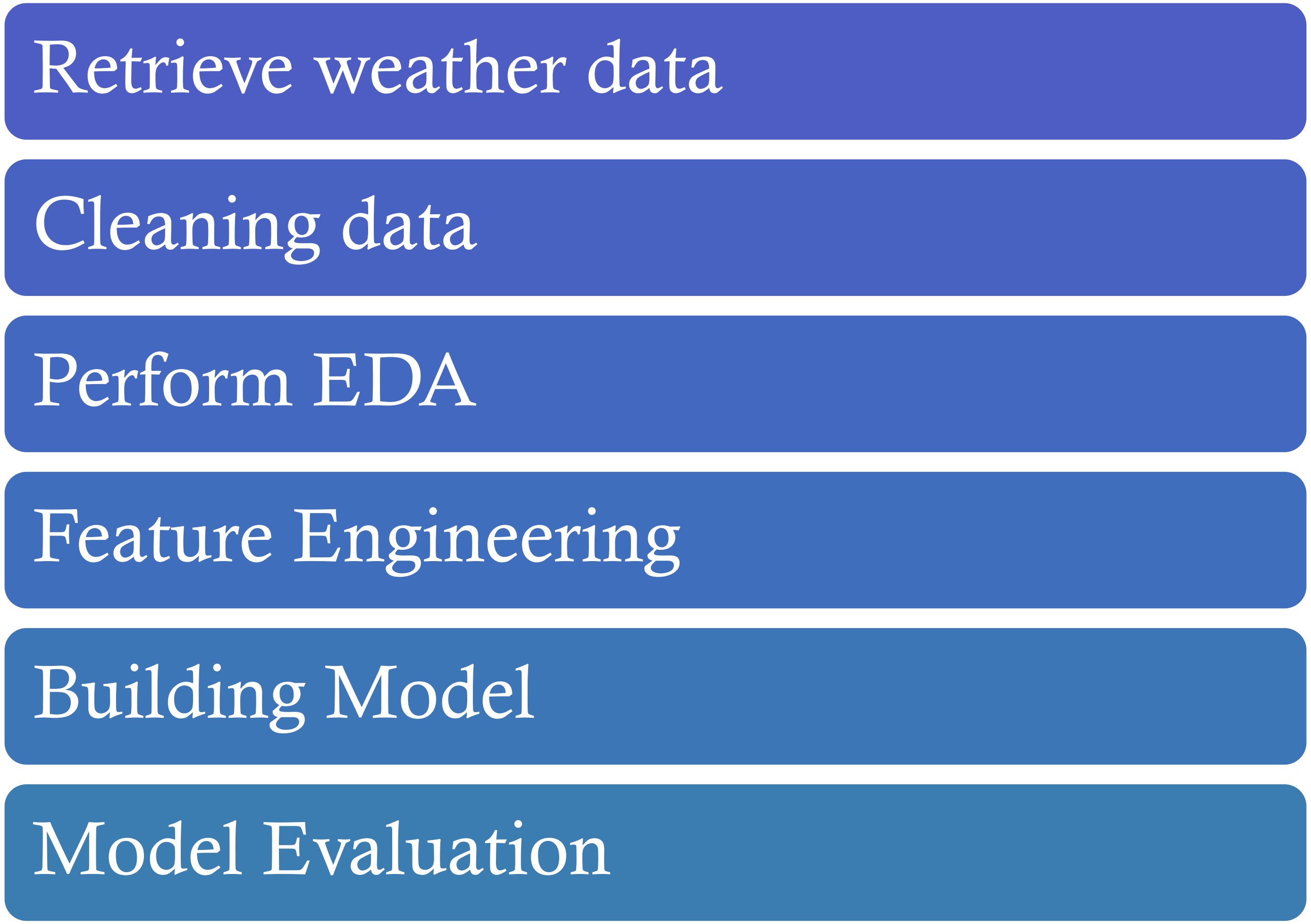
Ali Faisal Raza

Colin Tran

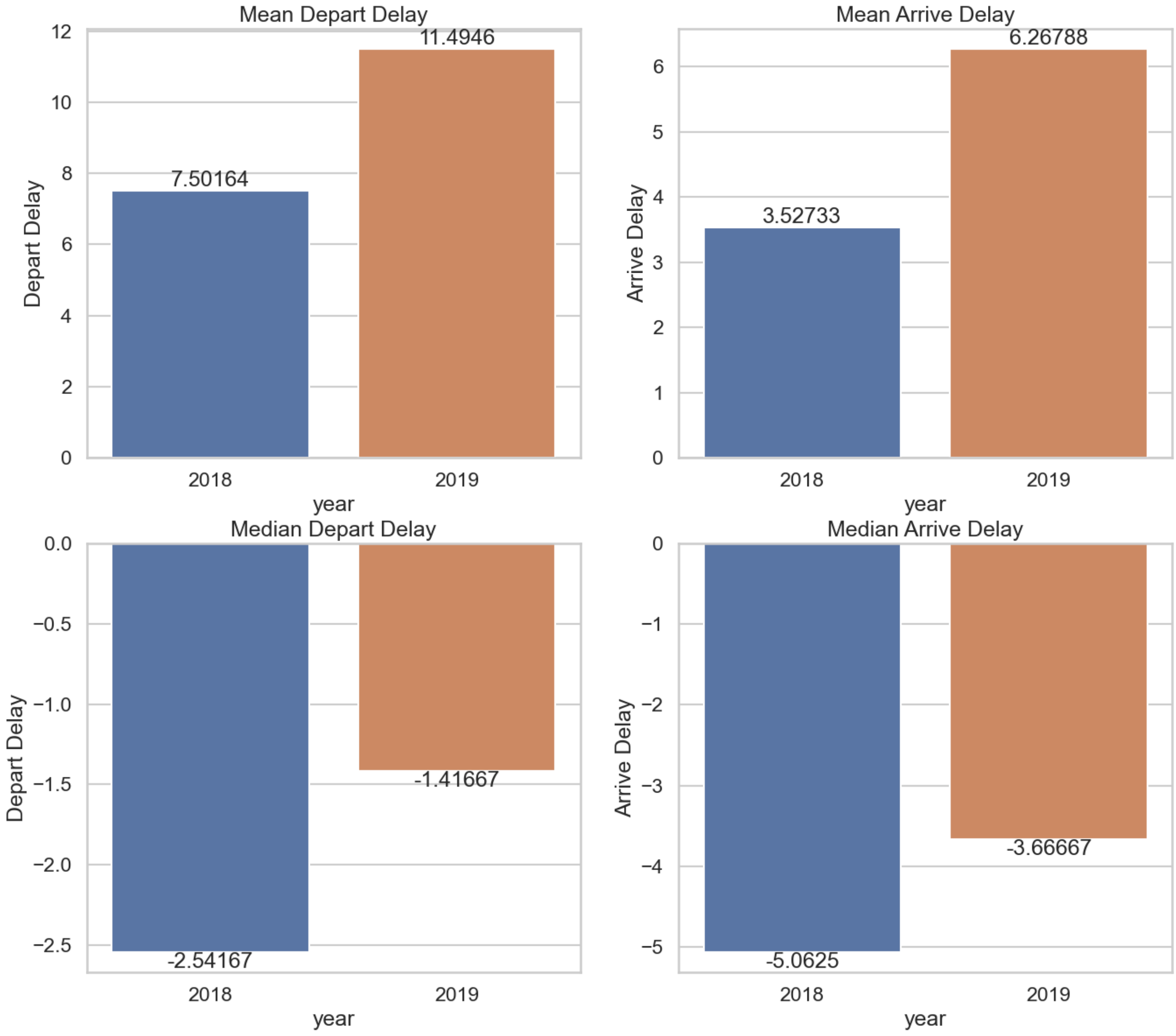
AGENDA

- Project Flow Structure
- Exploratory Data Analysis
- Results
- Biggest Challenge

PROJECT FLOW STRUCTURE



- Retrieve weather data
- Cleaning data
- Perform EDA
- Feature Engineering
- Building Model
- Model Evaluation

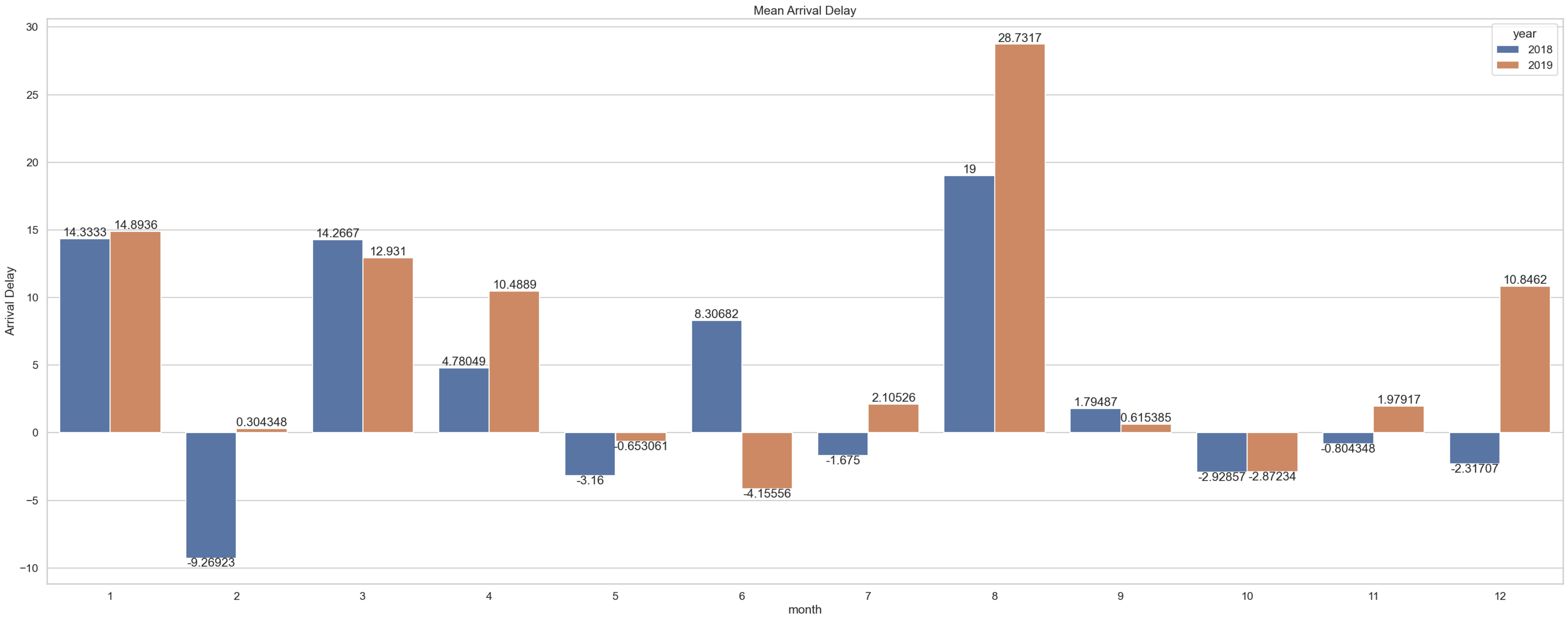


EDA

In 2019, the average duration of flight delay was roughly 177% greater than it was in 2018

EDA

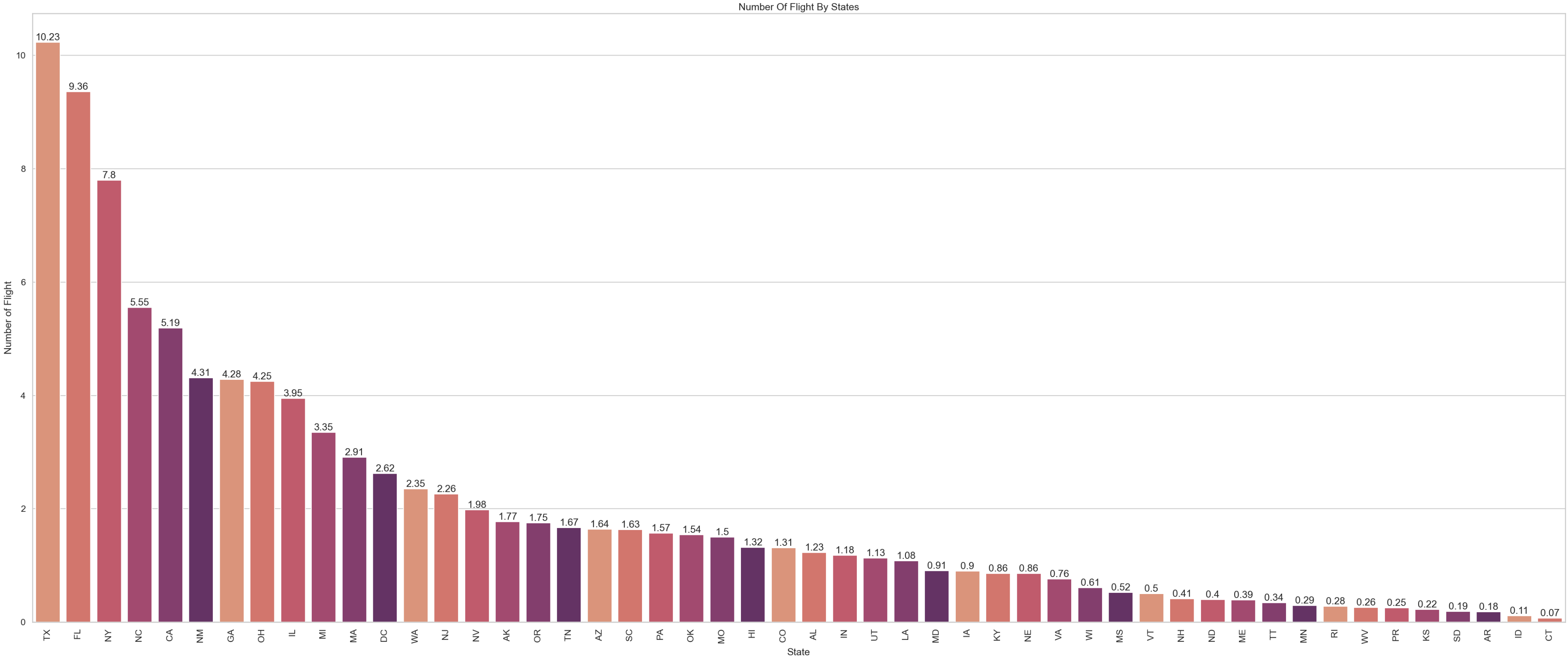
August has the highest average flight delay in both 2018 and 2019



There are 7 states that cover 50% of US air traffic

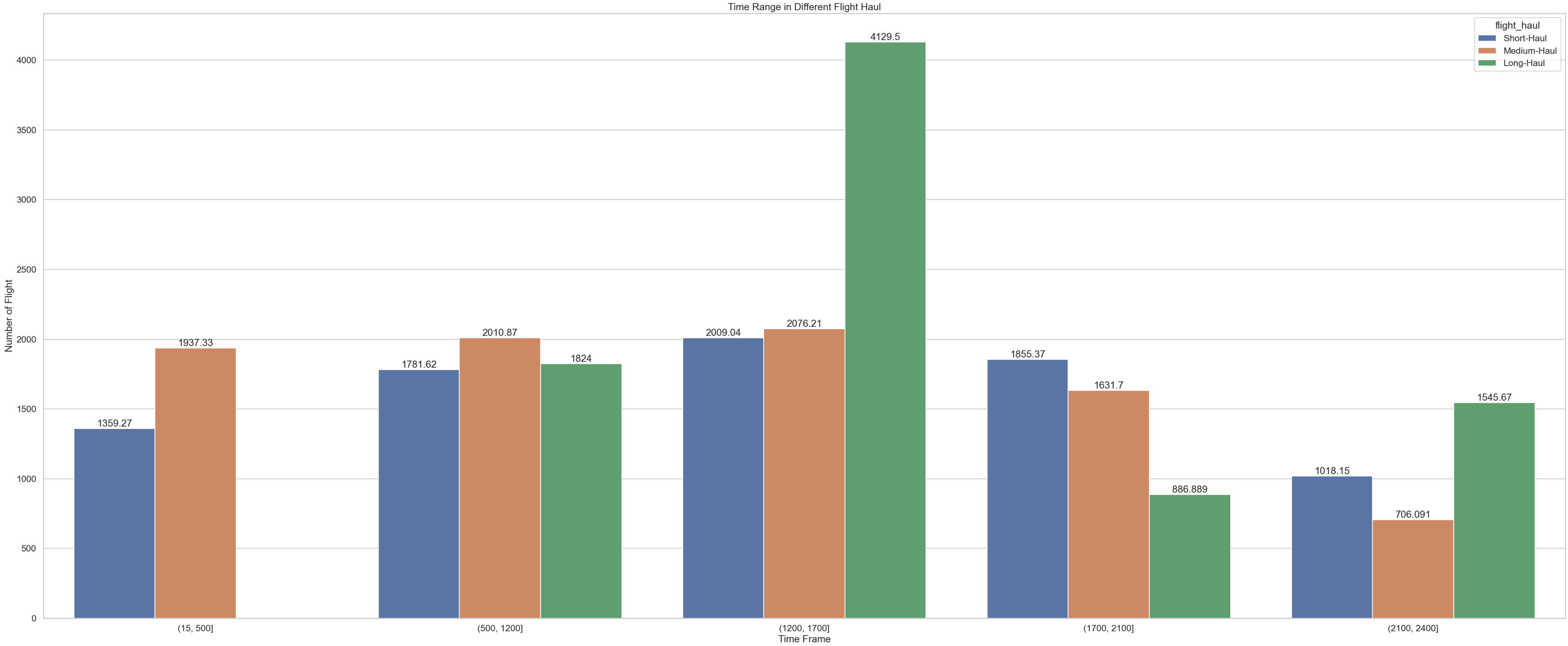
EDA

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

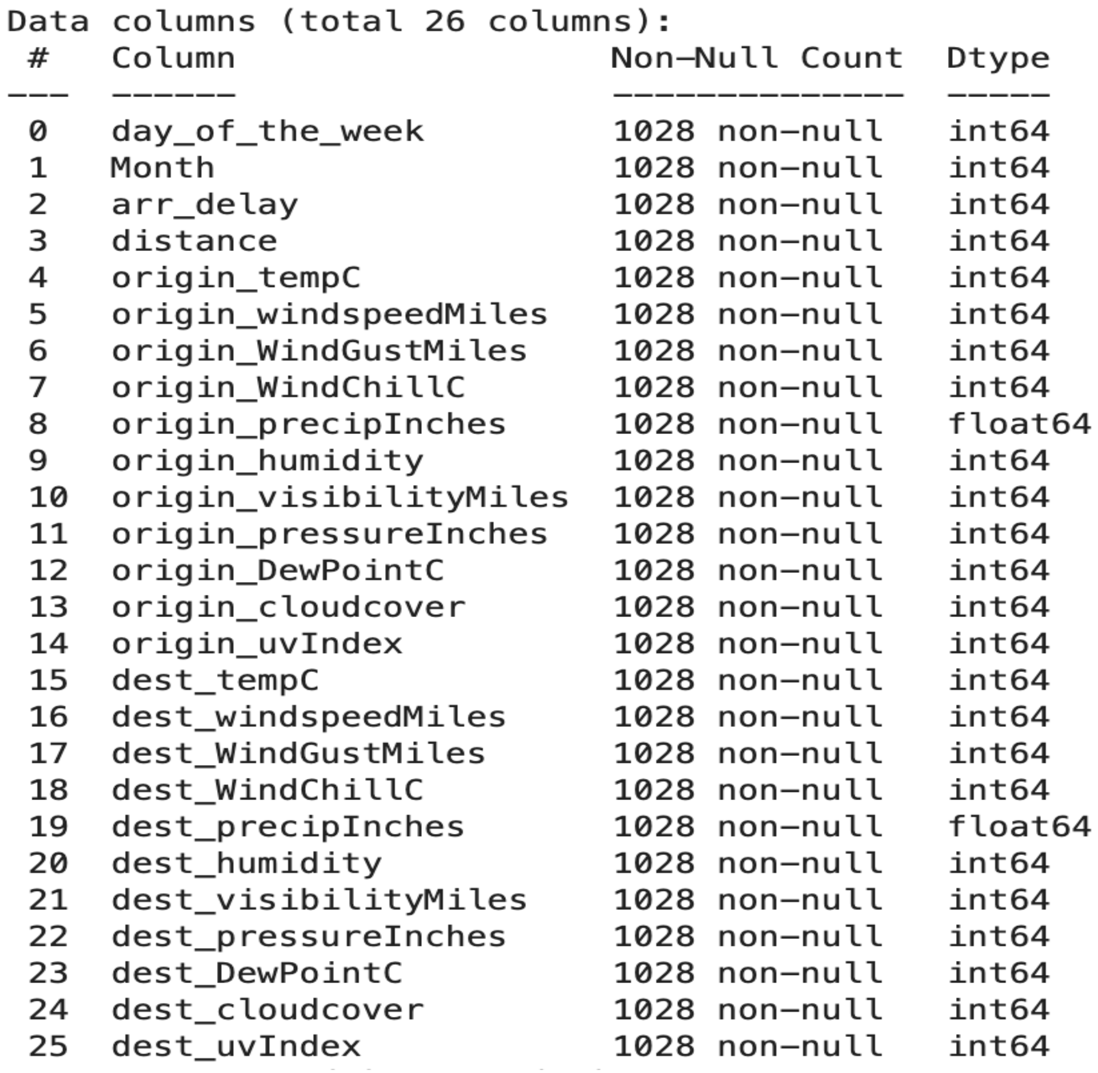


EDA

Most of the long-haul flight takes off from 12pm-17pm

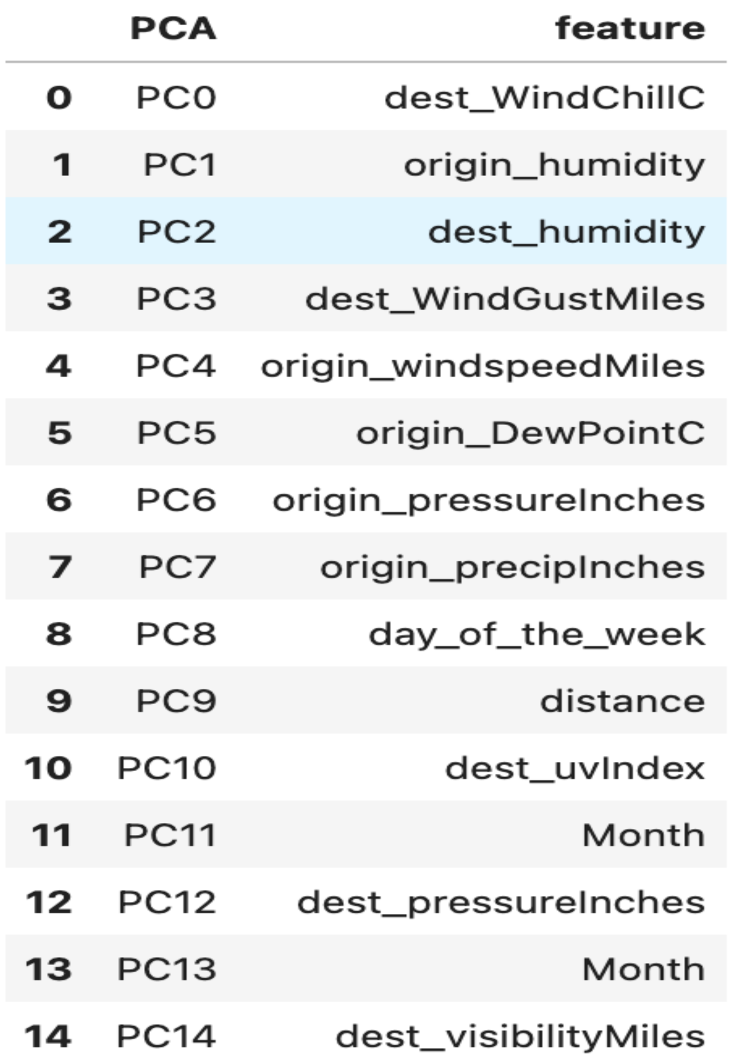


SELECTED FEATURES



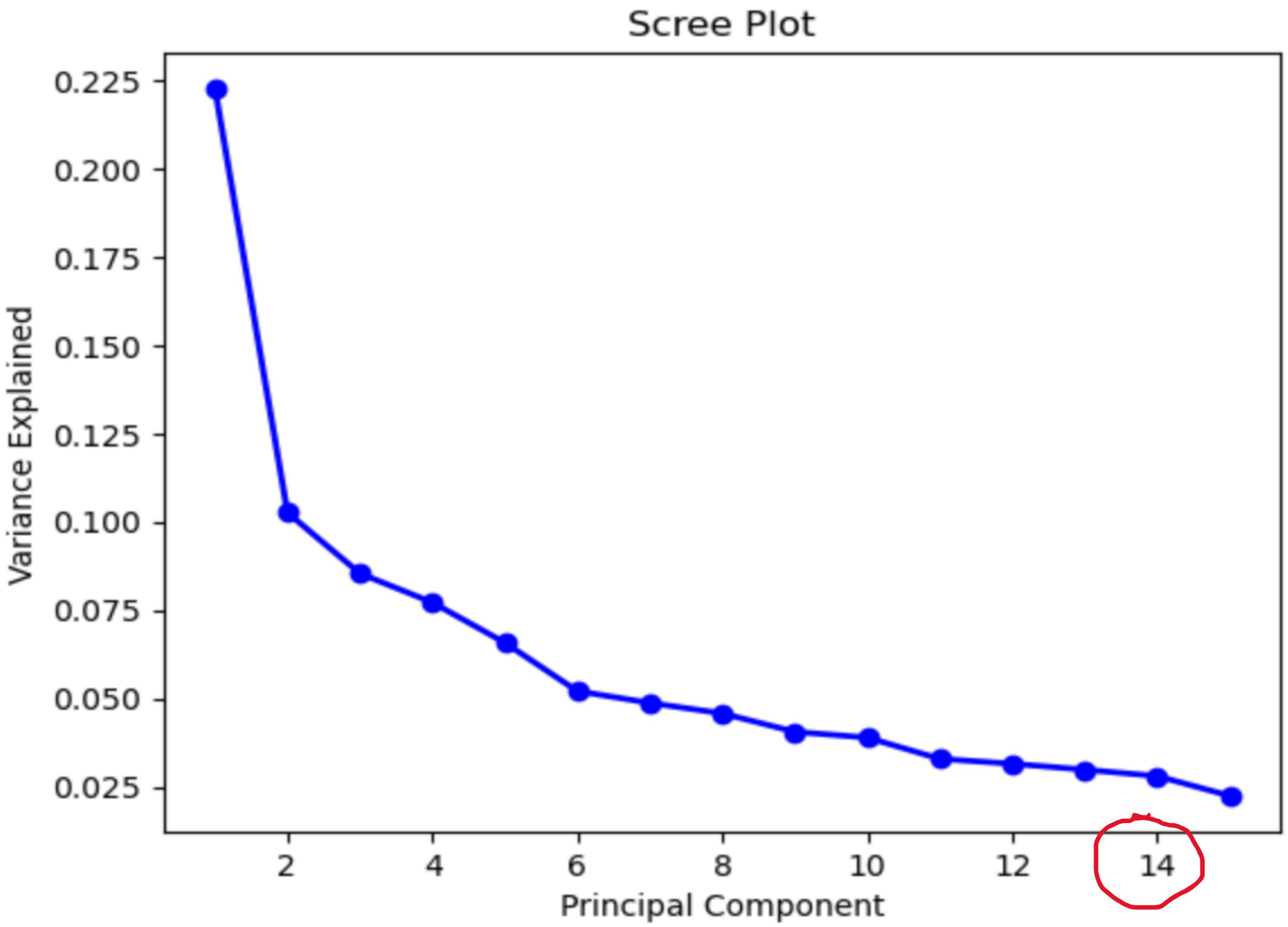
```
Data columns (total 26 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   day_of_the_week         1028 non-null   int64  
 1   Month                   1028 non-null   int64  
 2   arr_delay               1028 non-null   int64  
 3   distance                1028 non-null   int64  
 4   origin_tempC            1028 non-null   int64  
 5   origin_windspeedMiles   1028 non-null   int64  
 6   origin_WindGustMiles    1028 non-null   int64  
 7   origin_WindChillC       1028 non-null   int64  
 8   origin_precipInches     1028 non-null   float64
 9   origin_humidity         1028 non-null   int64  
 10  origin_visibilityMiles  1028 non-null   int64  
 11  origin_pressureInches   1028 non-null   int64  
 12  origin_DewPointC        1028 non-null   int64  
 13  origin_cloudcover       1028 non-null   int64  
 14  origin_uvIndex          1028 non-null   int64  
 15  dest_tempC              1028 non-null   int64  
 16  dest_windspeedMiles     1028 non-null   int64  
 17  dest_WindGustMiles      1028 non-null   int64  
 18  dest_WindChillC         1028 non-null   int64  
 19  dest_precipInches       1028 non-null   float64
 20  dest_humidity           1028 non-null   int64  
 21  dest_visibilityMiles    1028 non-null   int64  
 22  dest_pressureInches     1028 non-null   int64  
 23  dest_DewPointC          1028 non-null   int64  
 24  dest_cloudcover         1028 non-null   int64  
 25  dest_uvIndex            1028 non-null   int64  
```

SIGNIFICANT FEATURES



	PCA	feature
0	PC0	dest_WindChillC
1	PC1	origin_humidity
2	PC2	dest_humidity
3	PC3	dest_WindGustMiles
4	PC4	origin_windspeedMiles
5	PC5	origin_DewPointC
6	PC6	origin_pressureInches
7	PC7	origin_precipInches
8	PC8	day_of_the_week
9	PC9	distance
10	PC10	dest_uvIndex
11	PC11	Month
12	PC12	dest_pressureInches
13	PC13	Month
14	PC14	dest_visibilityMiles

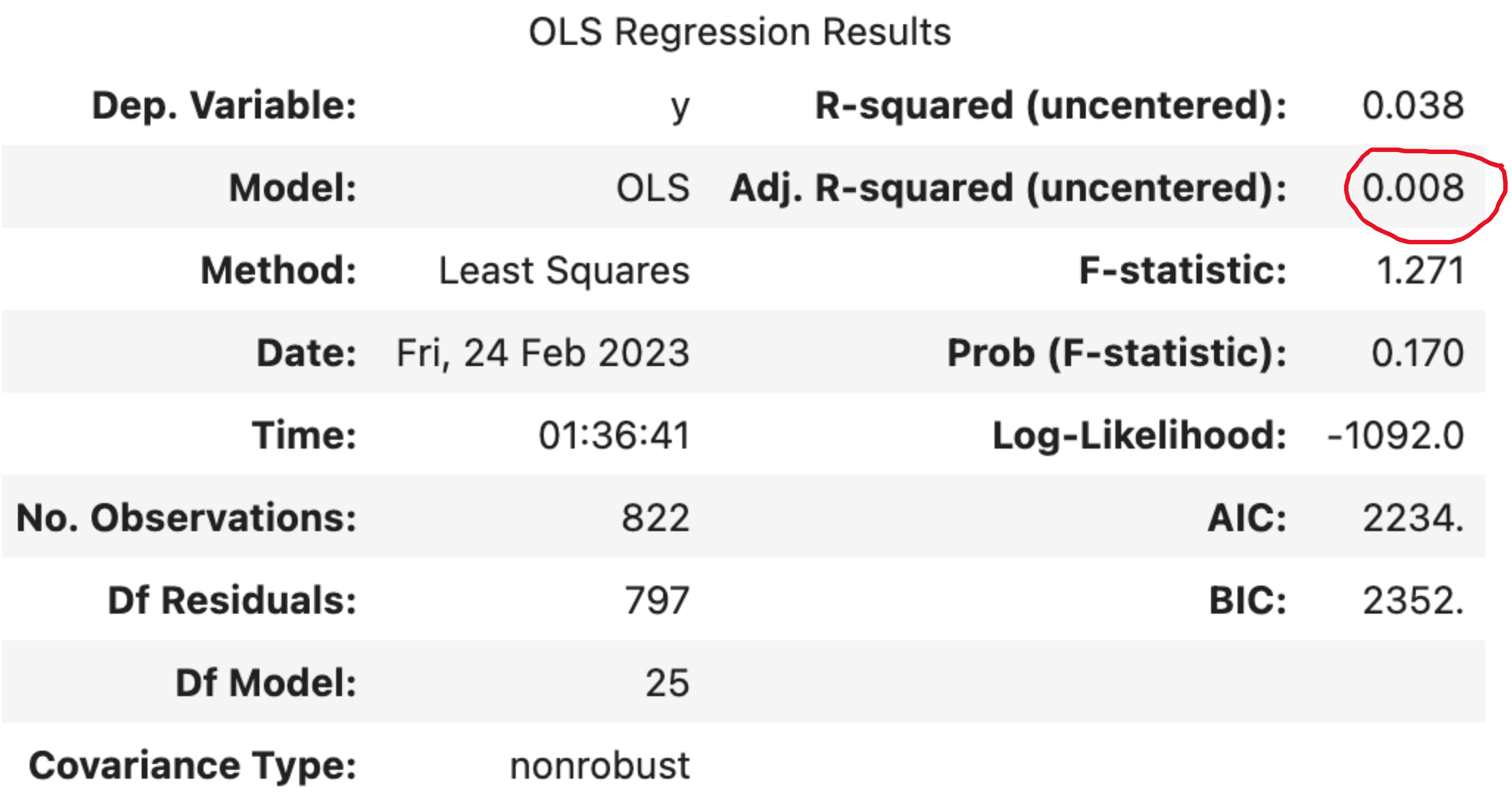
PCA ANALYSIS



```
cumulative variance components:  0.9239605276680356
```

No dominant PCA vectors.

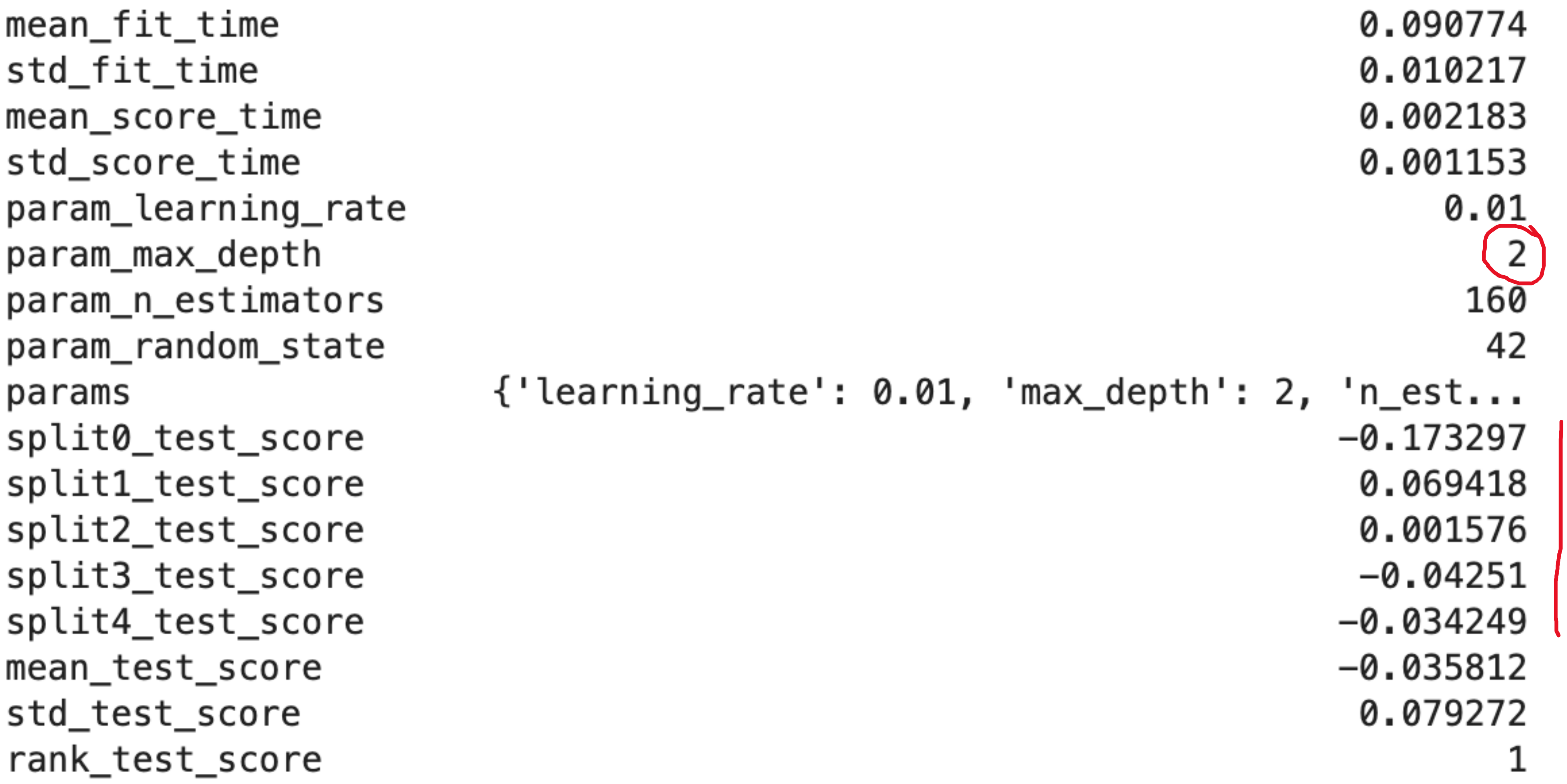
LINEAR REGRESSION FIT



OLS Regression Results

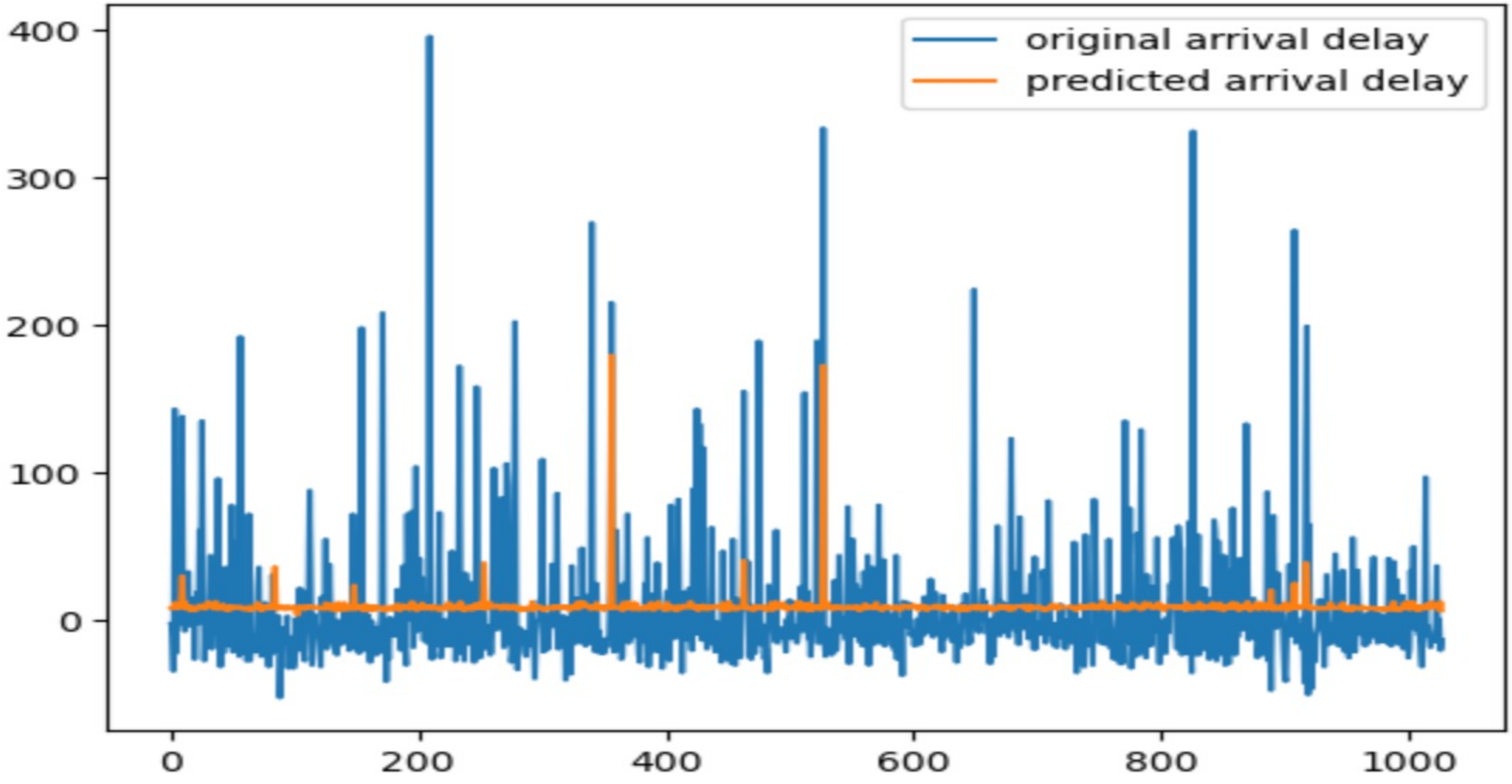
Dep. Variable:	y	**R-squared (uncentered):**	0.038
Model:	OLS	**Adj. R-squared (uncentered):**	0.008
Method:	Least Squares	**F-statistic:**	1.271
Date:	Fri, 24 Feb 2023	**Prob (F-statistic):**	0.170
Time:	01:36:41	**Log-Likelihood:**	-1092.0
No. Observations:	822	**AIC:**	2234.
Df Residuals:	797	**BIC:**	2352.
Df Model:	25		
Covariance Type:	nonrobust		

XGBOOST FIT (5 KFOLDS WITH GRID SEARCH TUNING)

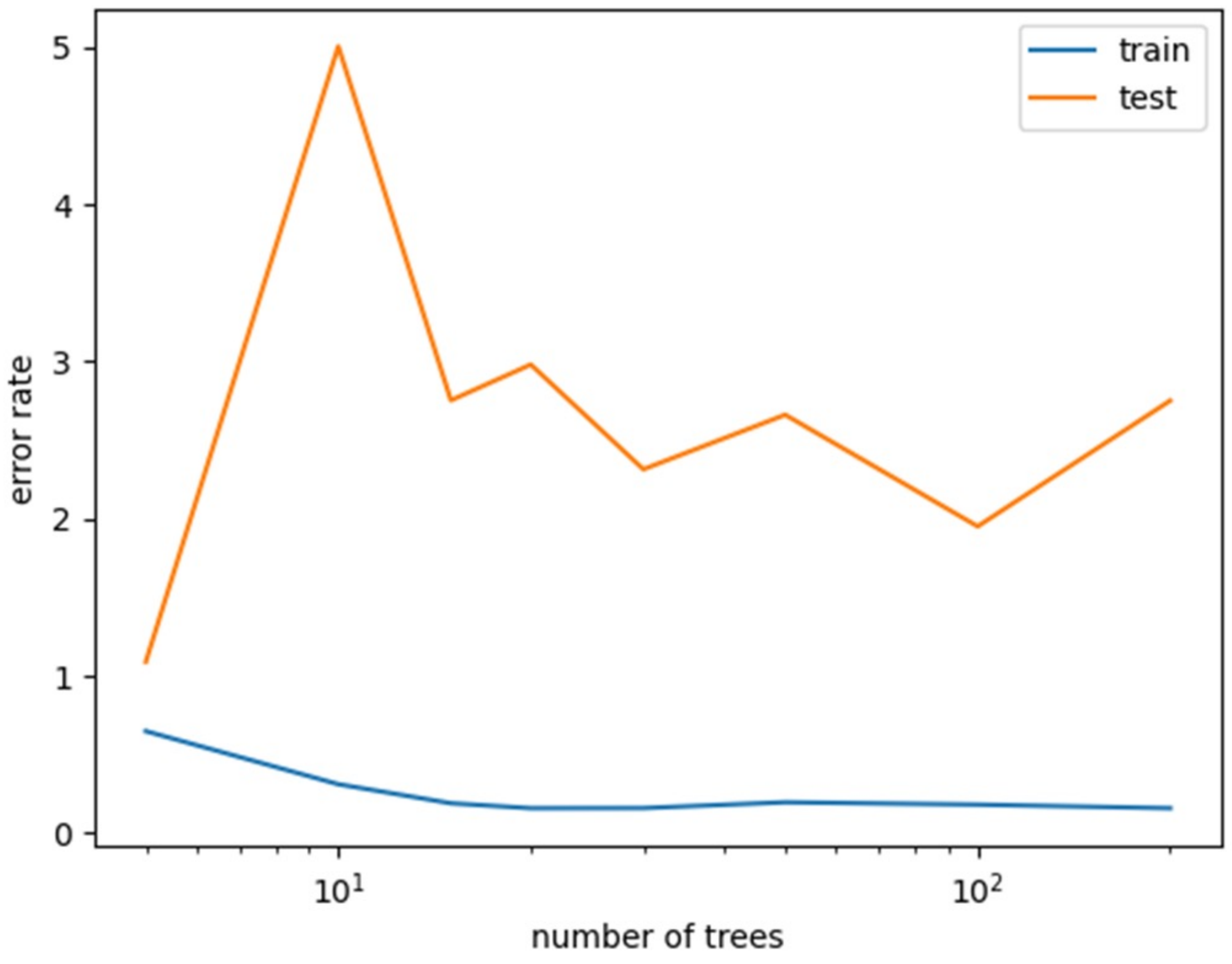


```
mean_fit_time                                                  0.090774
std_fit_time                                                   0.010217
mean_score_time                                                0.002183
std_score_time                                                 0.001153
param_learning_rate                                                0.01
param_max_depth                                                       2
param_n_estimators                                                  160
param_random_state                                                   42
params                 {'learning_rate': 0.01, 'max_depth': 2, 'n_est...
split0_test_score                                             -0.173297
split1_test_score                                              0.069418
split2_test_score                                              0.001576
split3_test_score                                              -0.04251
split4_test_score                                             -0.034249
mean_test_score                                               -0.035812
std_test_score                                                 0.079272
rank_test_score                                                       1
```

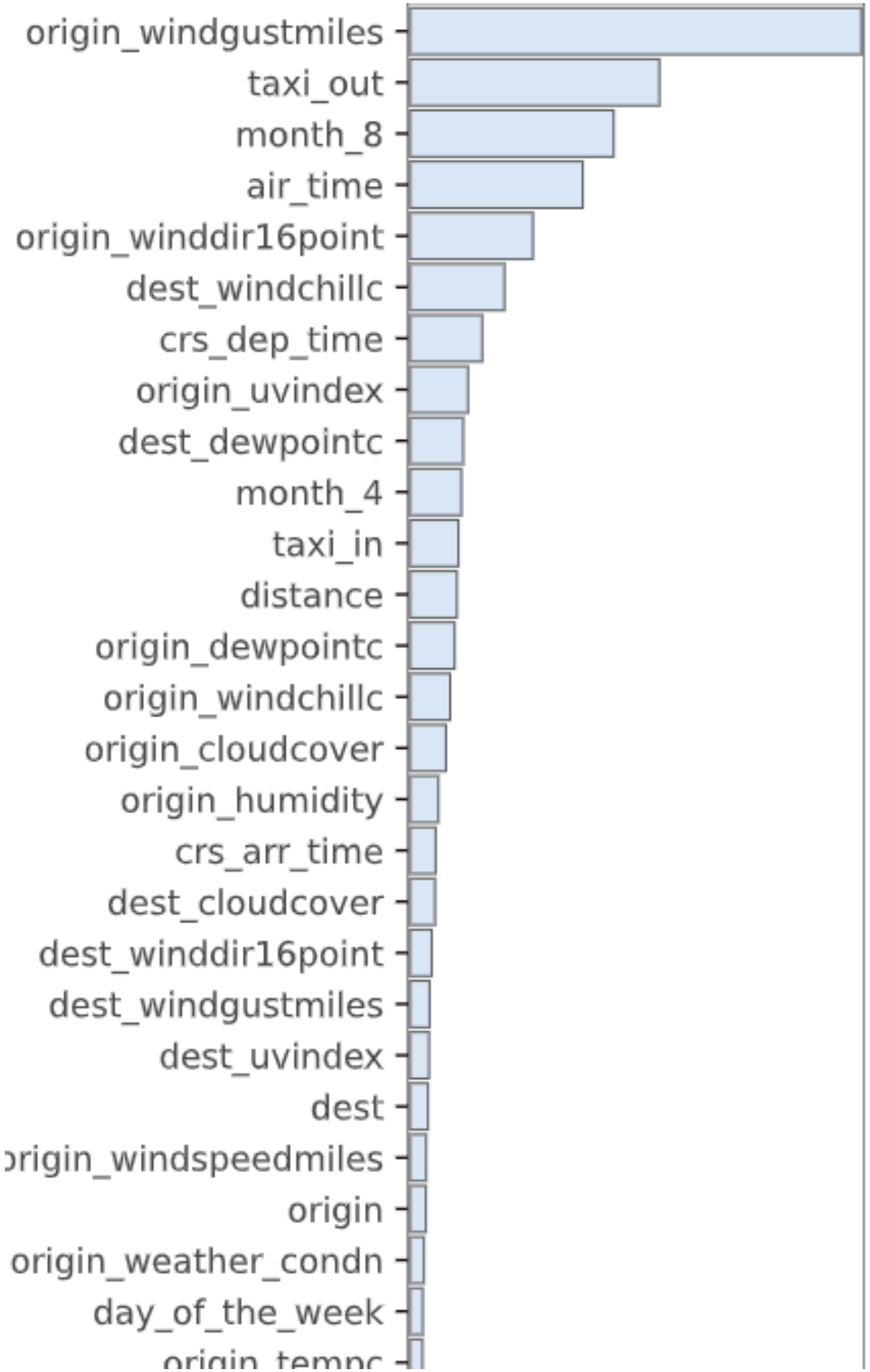
ORIGINAL DELAY VS PREDICTED DELAY (TRAINING DATA)



RANDOM FOREST



IMPORTANT FEATURE OF RANDOM FOREST



CHALLENGE

The random nature of data

The restriction on weather API leading to small sample dataset

THANK YOU FOR YOUR LISTENING