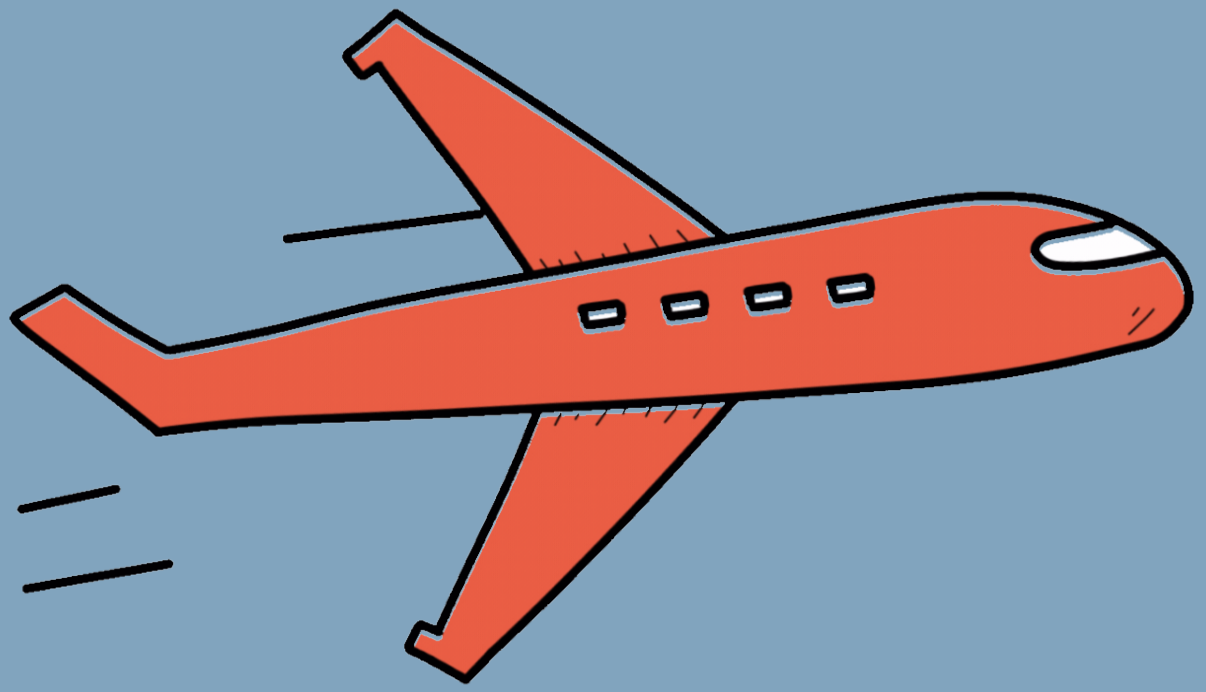
Predicting Flight Delays

Caleb Ward and
Jessica Moloney

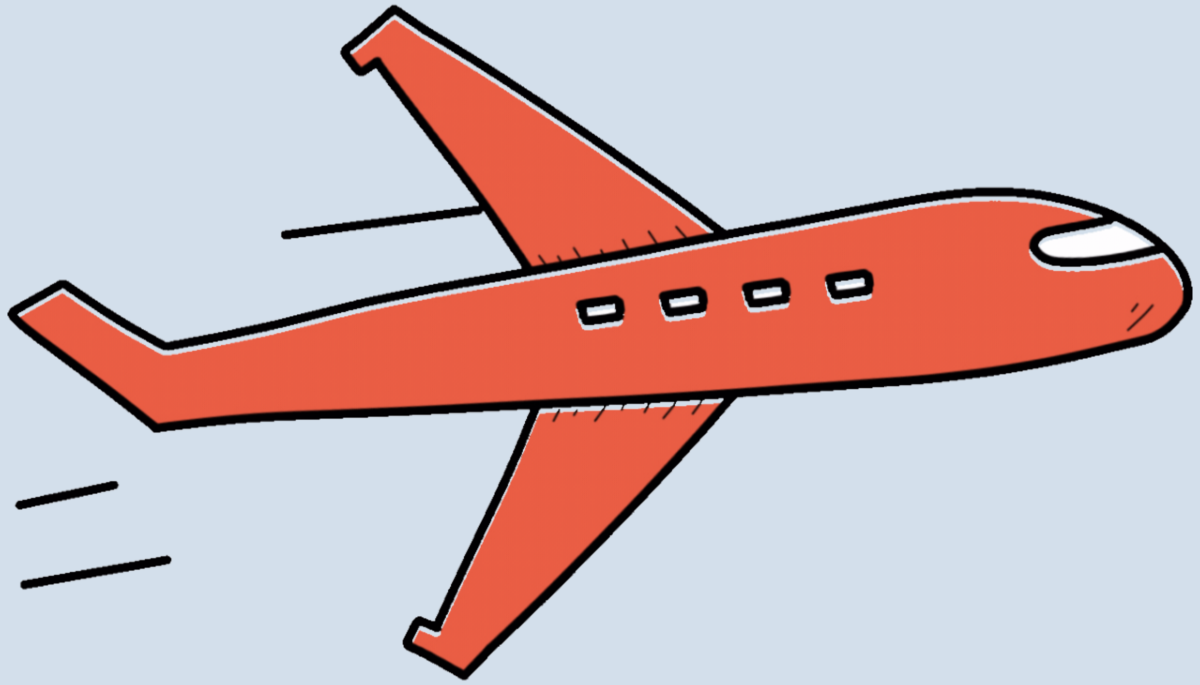
Problem

Can we predict how delayed flights will be one week in advance?



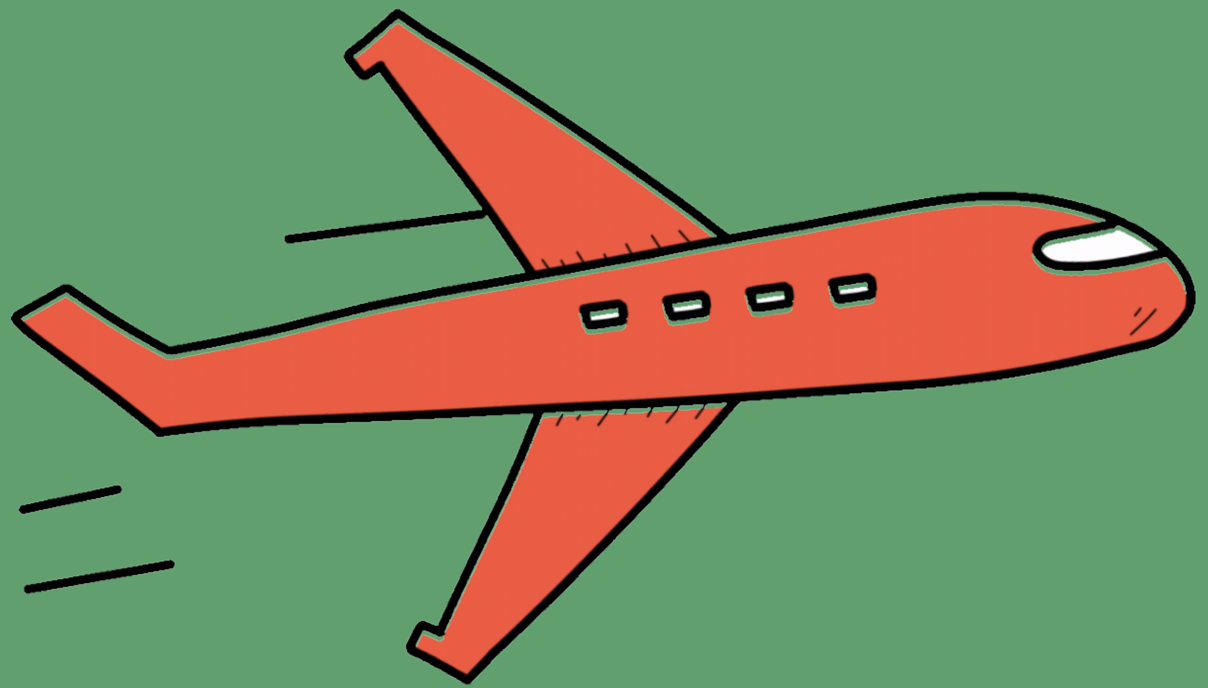
Data

- 15 million records of US flights in 2018 and 2019
- Information about carrier, flight number, tail number, planned departure / arrival time, origin / destination, actual delays



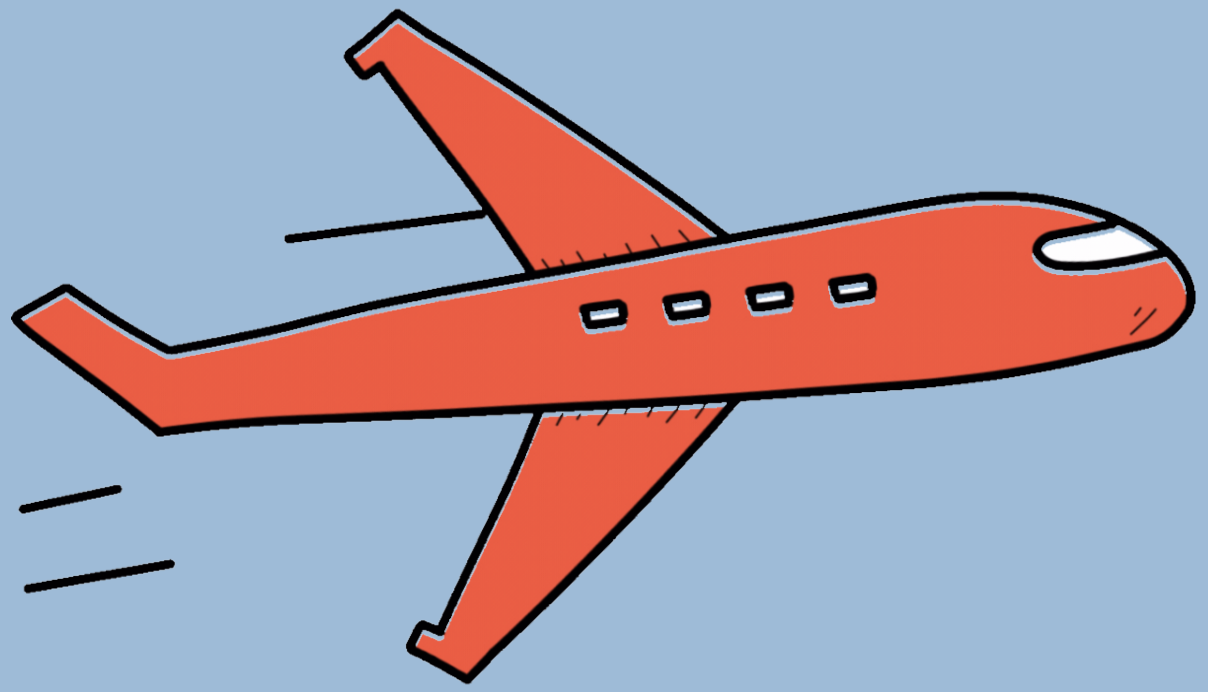
Additional Sources

- NOAA - NWS Climate Prediction Center
- Government Source
- Historical forecasts, not historical weather records
- Exhaustive coverage
- Easily accessible and free

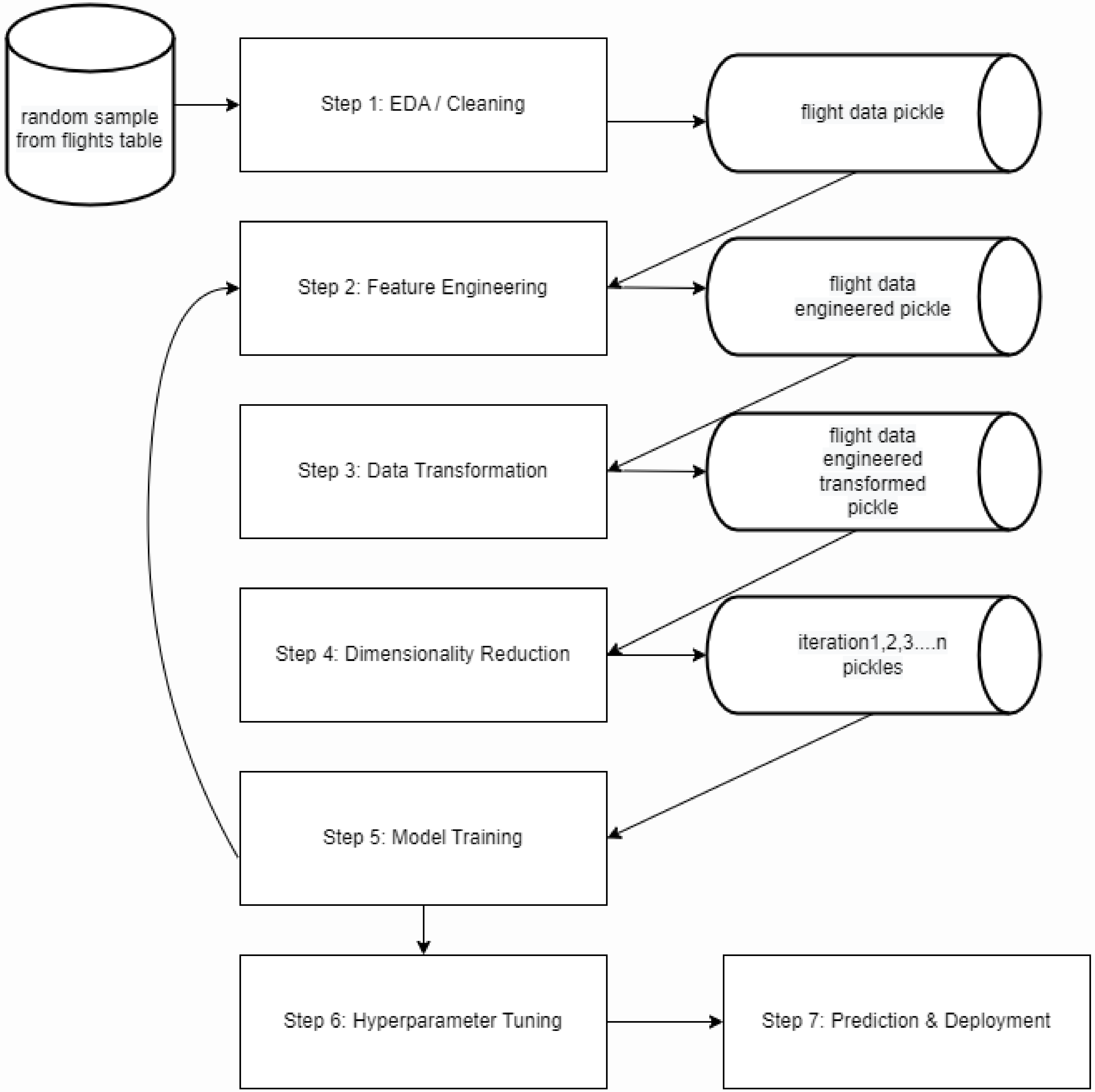


Goal

- Use machine learning to find patterns in the data and predict future delays

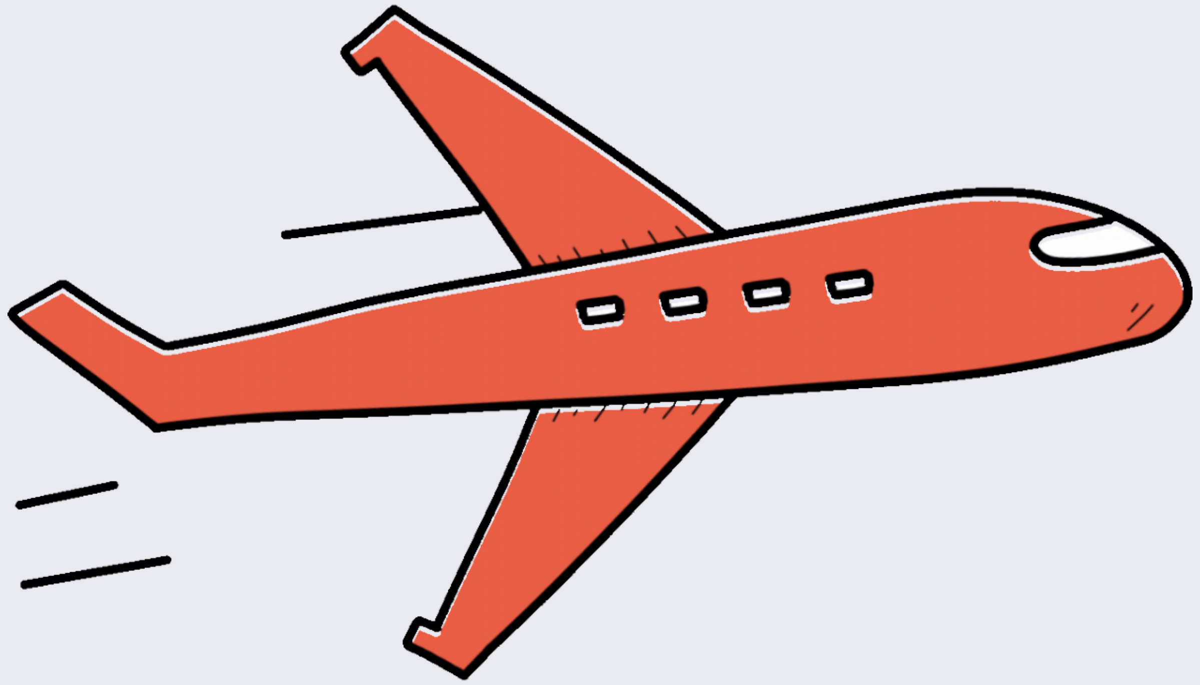


Approach

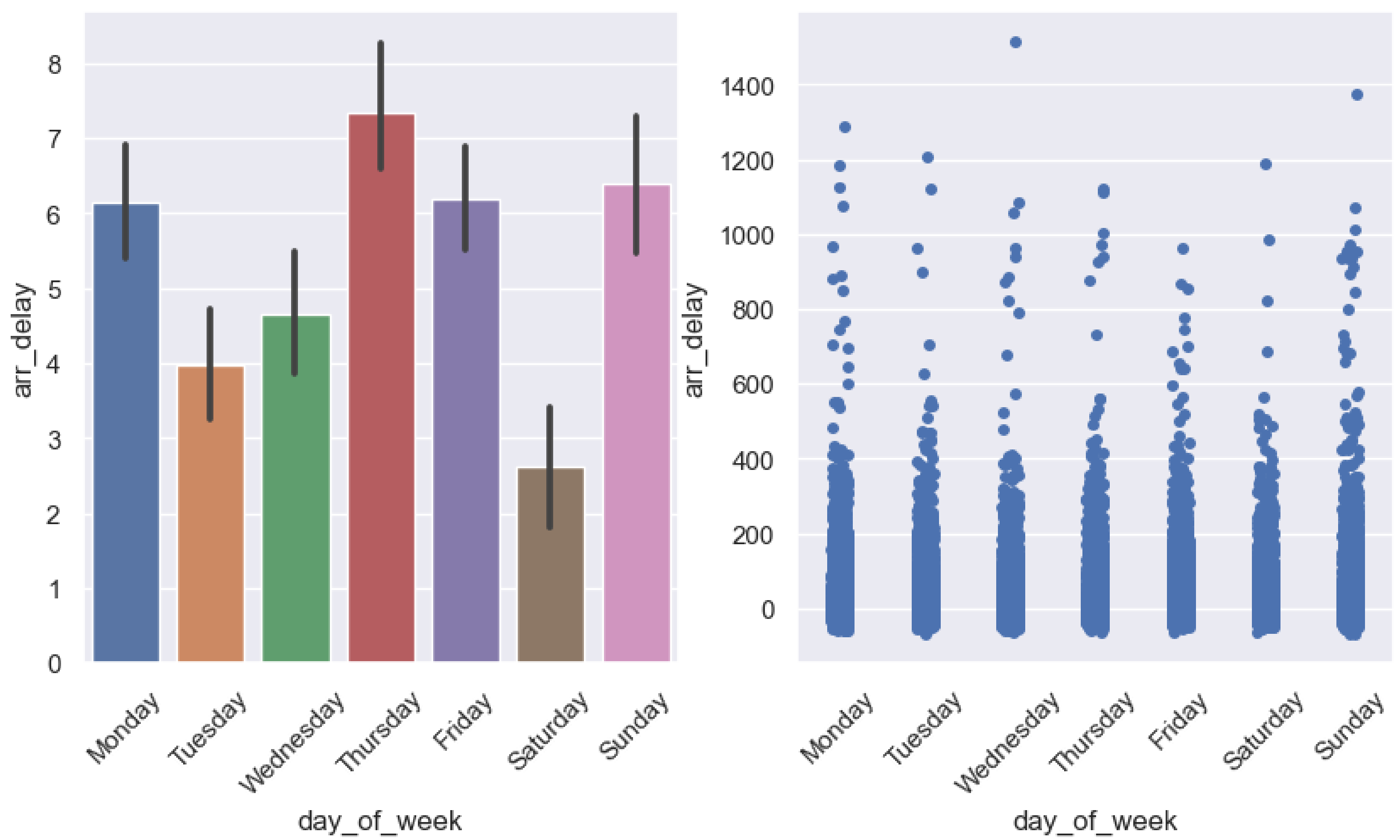


random sample from flights table
Step 1: EDA / Cleaning
flight data pickle
Step 2: Feature Engineering
flight data engineered pickle
Step 3: Data Transformation
flight data engineered transformed pickle
Step 4: Dimensionality Reduction
iteration1,2,3....n pickles
Step 5: Model Training
Step 6: Hyperparameter Tuning
Step 7: Prediction & Deployment

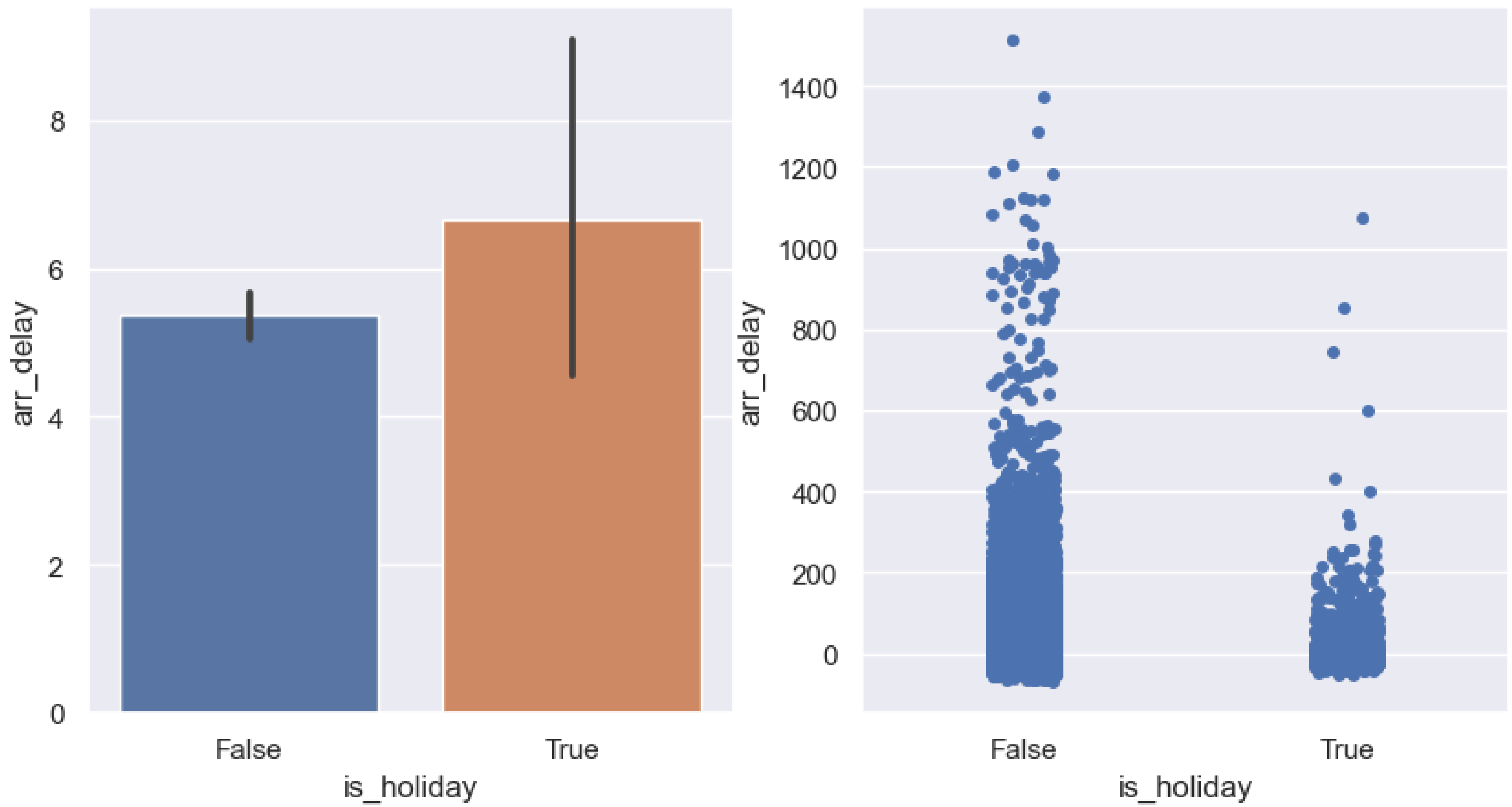
Features based on Flight Date



Arrival Delay by Day of Week

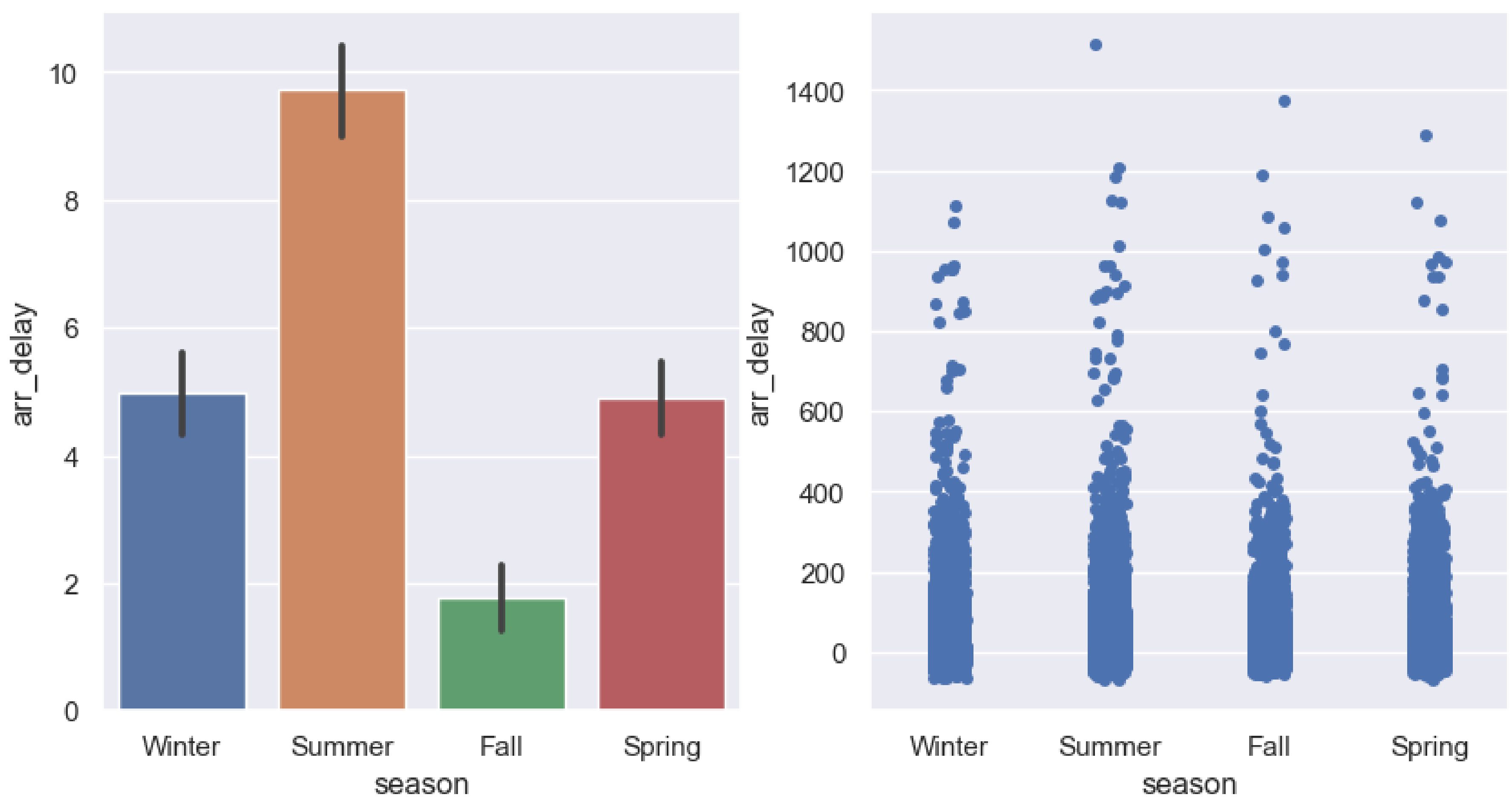


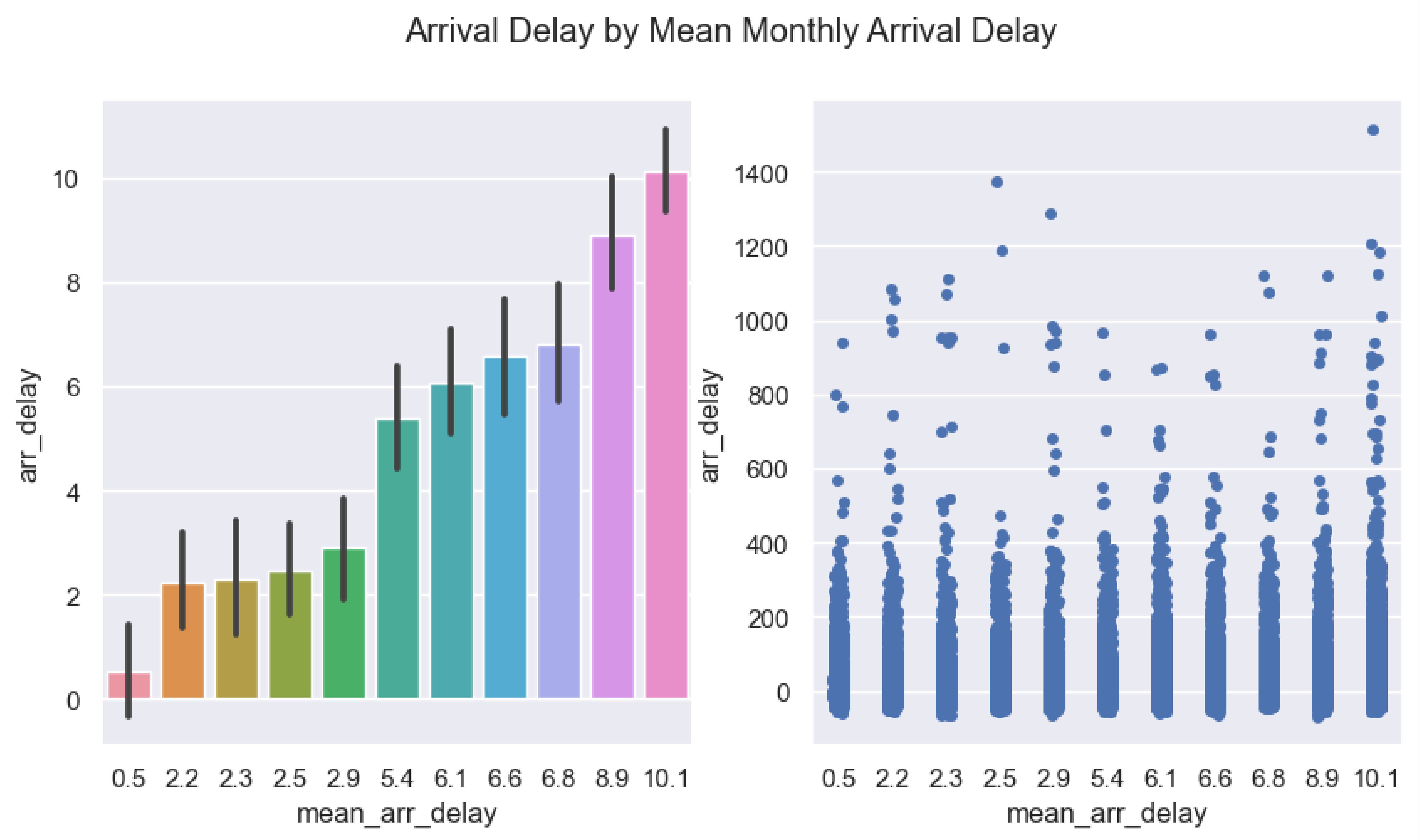
Arrival Delay by Normal Day vs. Holiday



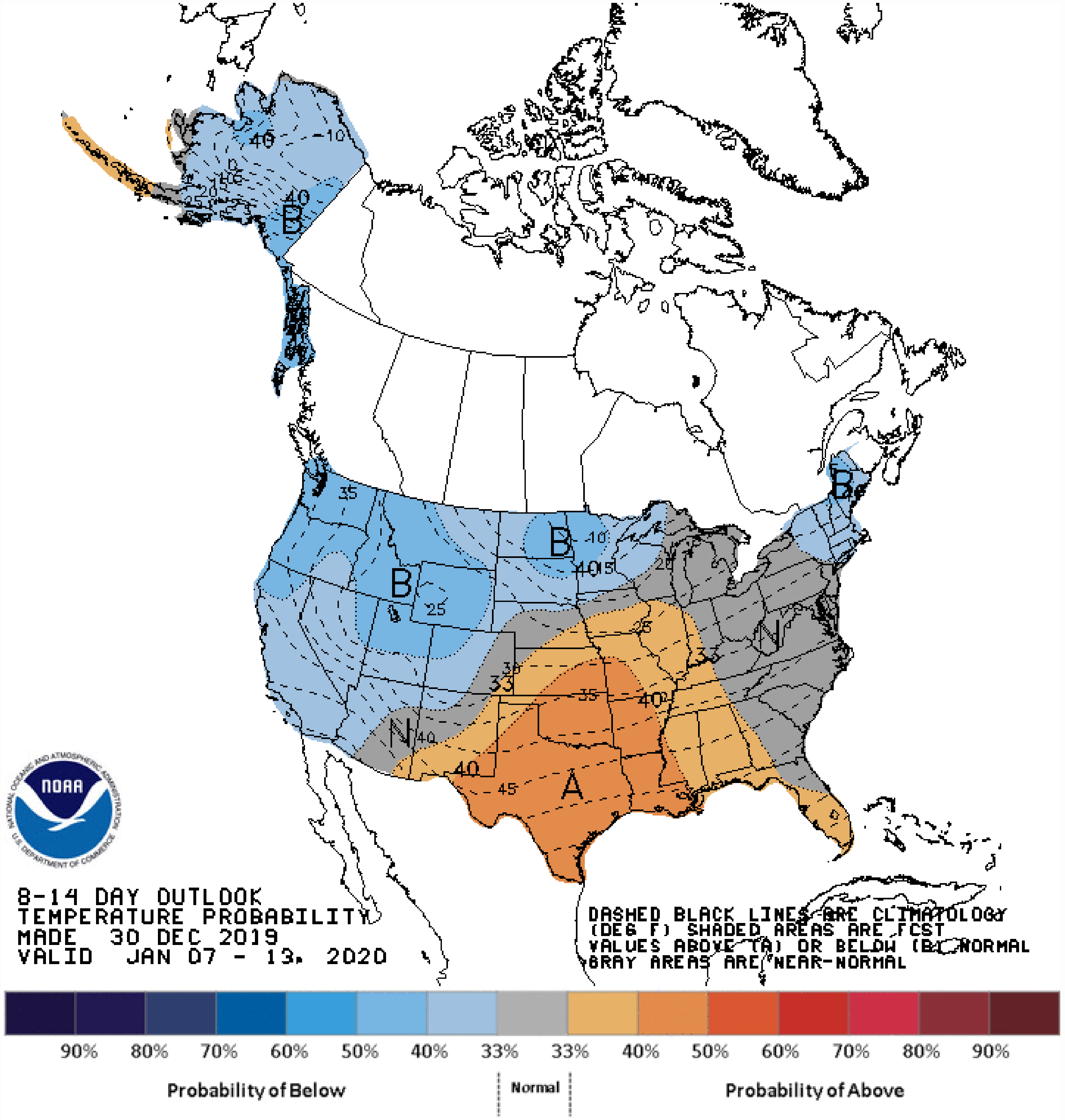
Arrival Delay by Weekday vs. Weekend

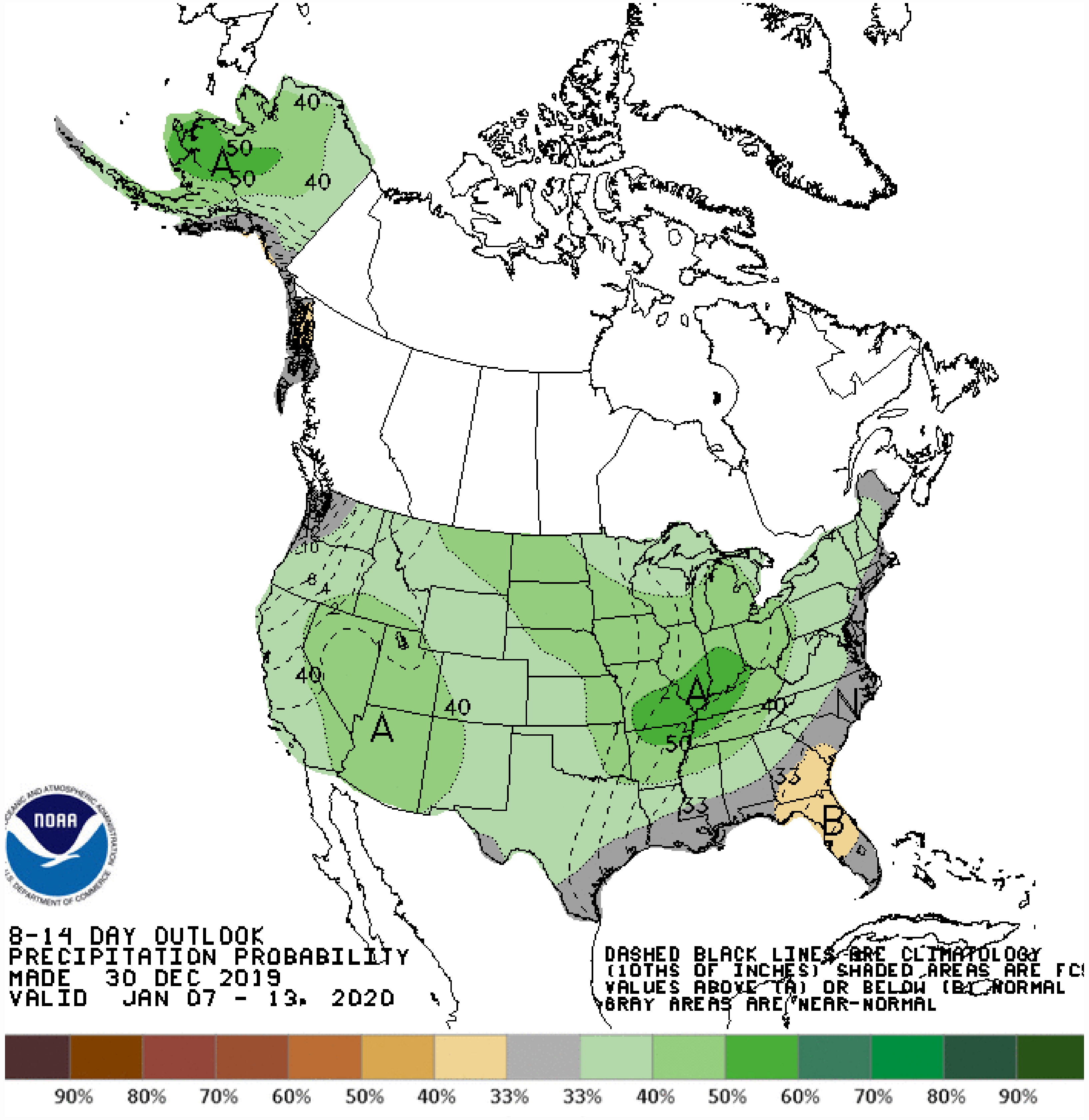
Arrival Delay by Season



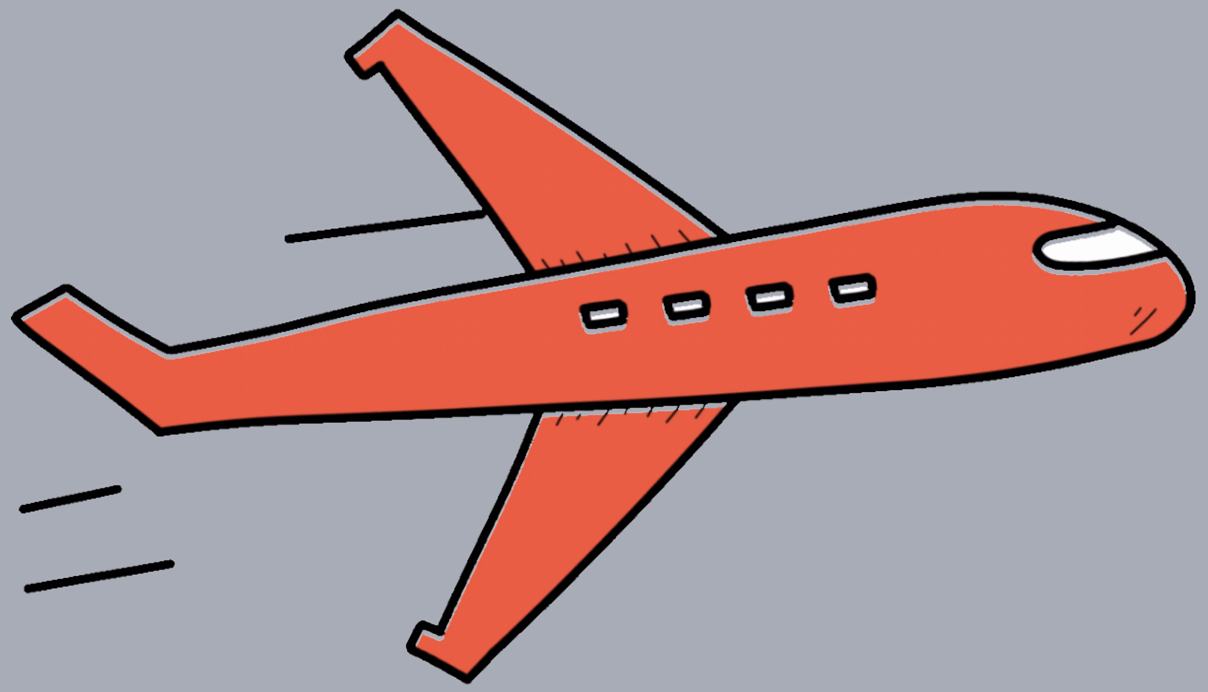


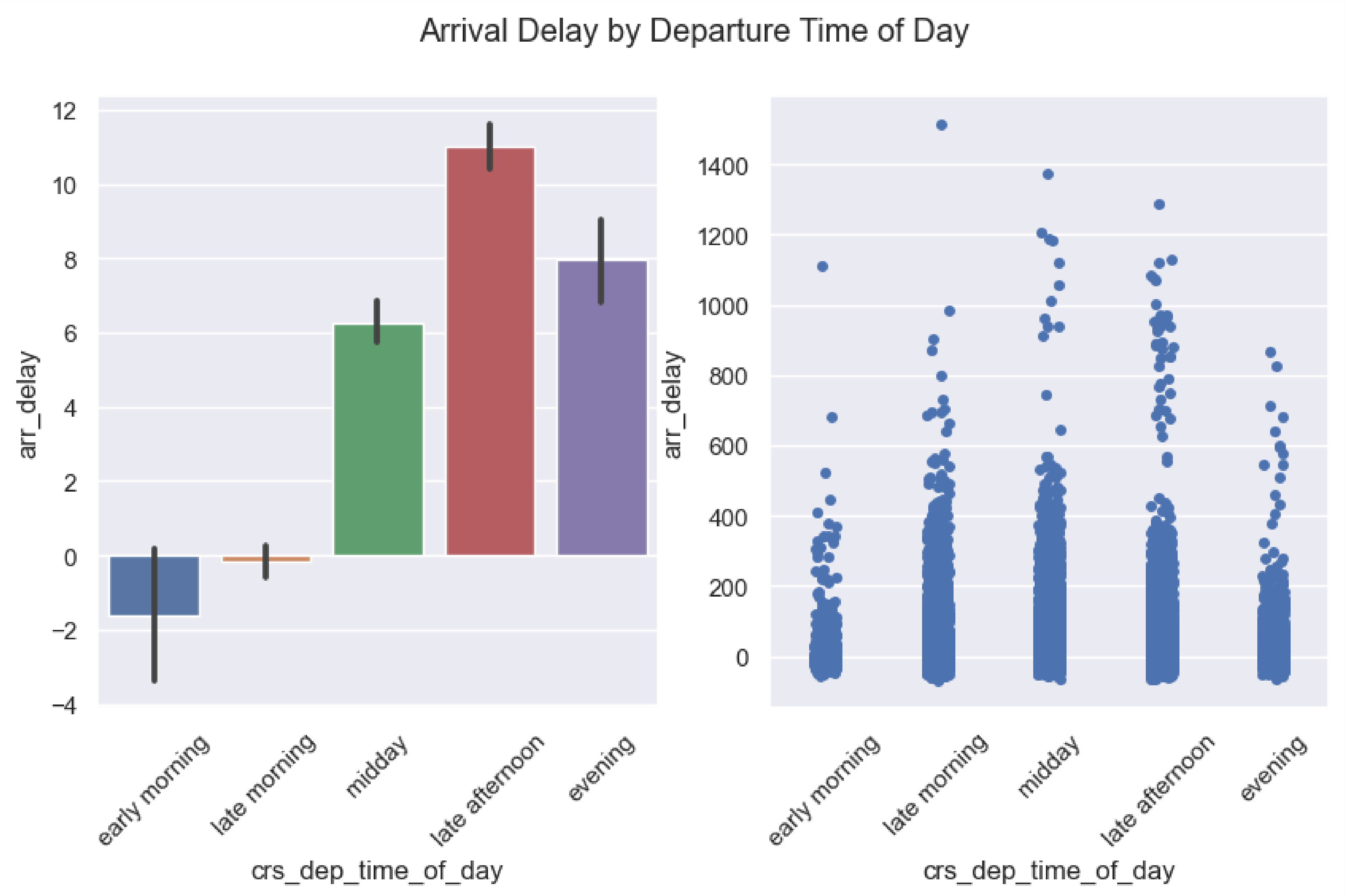
Arrival Delay by Mean Monthly Arrival Delay
arr_delay
0
2
4
6
8
10
0.5 2.2 2.3 2.5 2.9 5.4 6.1 6.6 6.8 8.9 10.1
mean_arr_delay
arr_delay
0
200
400
600
800
1000
1200
1400
0.5 2.2 2.3 2.5 2.9 5.4 6.1 6.6 6.8 8.9 10.1
mean_arr_delay

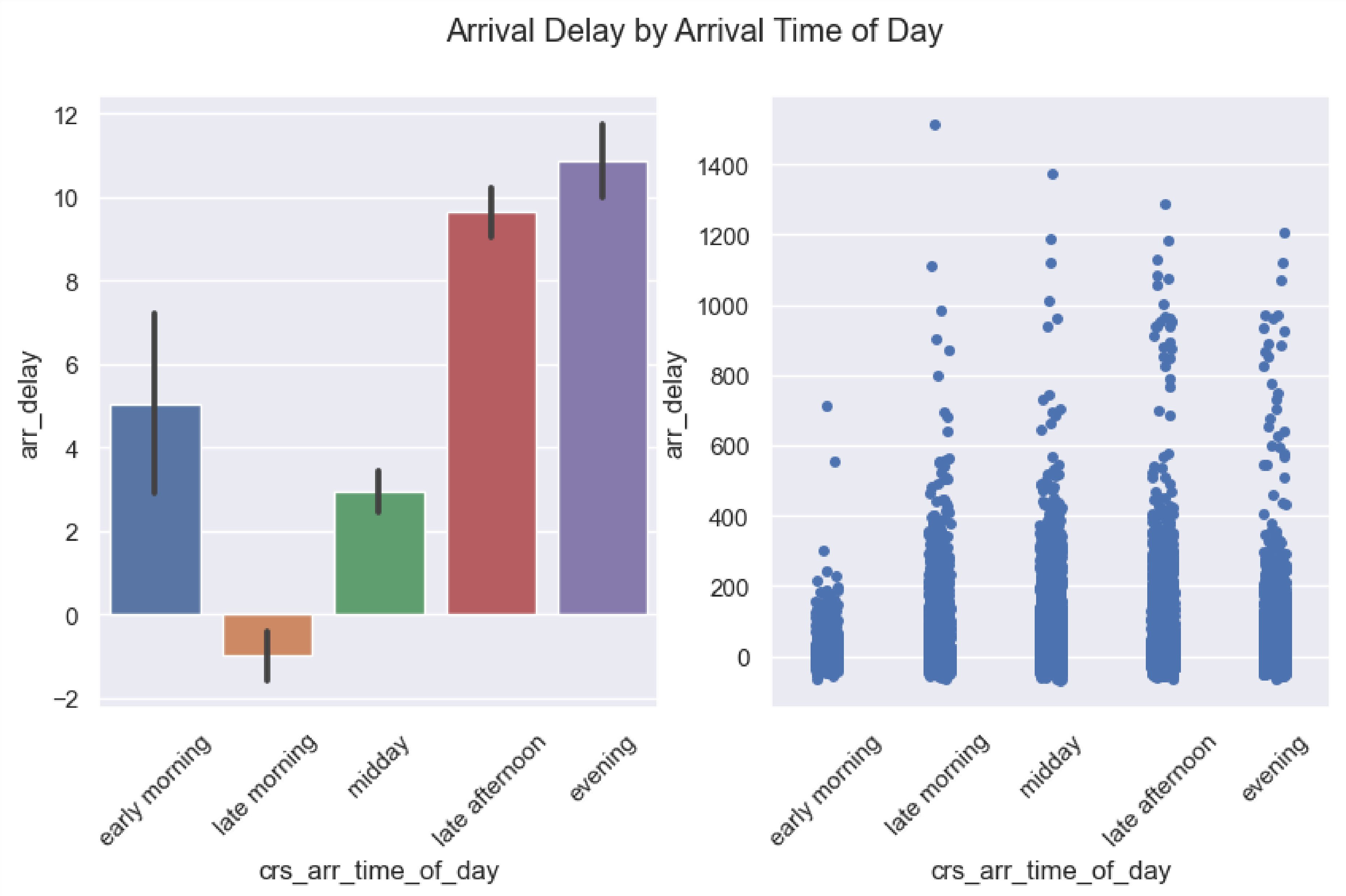

8-14 DAY OUTLOOK
TEMPERATURE PROBABILITY
MADE 30 DEC 2019
VALID JAN 07 - 13, 2020
DASHED BLACK LINES ARE CLIMATOLOGY
(DEG F) SHADED AREAS ARE FCST
VALUES ABOVE (A) OR BELOW (B) NORMAL
GRAY AREAS ARE NEAR-NORMAL
90%
80%
70%
60%
50%
40%
33%
33%
40%
50%
60%
70%
80%
90%
Probability of Below
Normal
Probability of Above


8-14 DAY OUTLOOK
PRECIPITATION PROBABILITY
MADE 30 DEC 2019
VALID JAN 07 - 13, 2020
DASHED BLACK LINES ARE CLIMATOLOGY
(10THS OF INCHES) SHADED AREAS ARE FC
VALUES ABOVE (A) OR BELOW (B) NORMAL
GRAY AREAS ARE NEAR-NORMAL
90%
80%
70%
60%
50%
40%
33%
33%
40%
50%
60%
70%
80%
90%

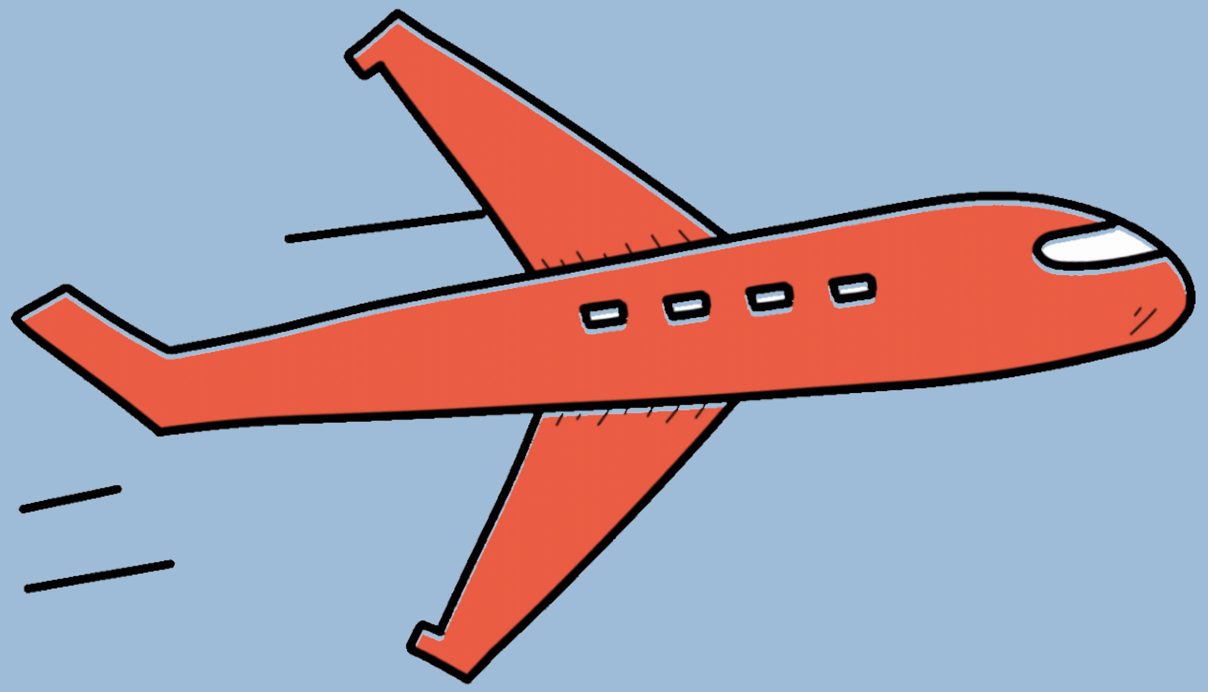
Features based on Departure/Arrival Time

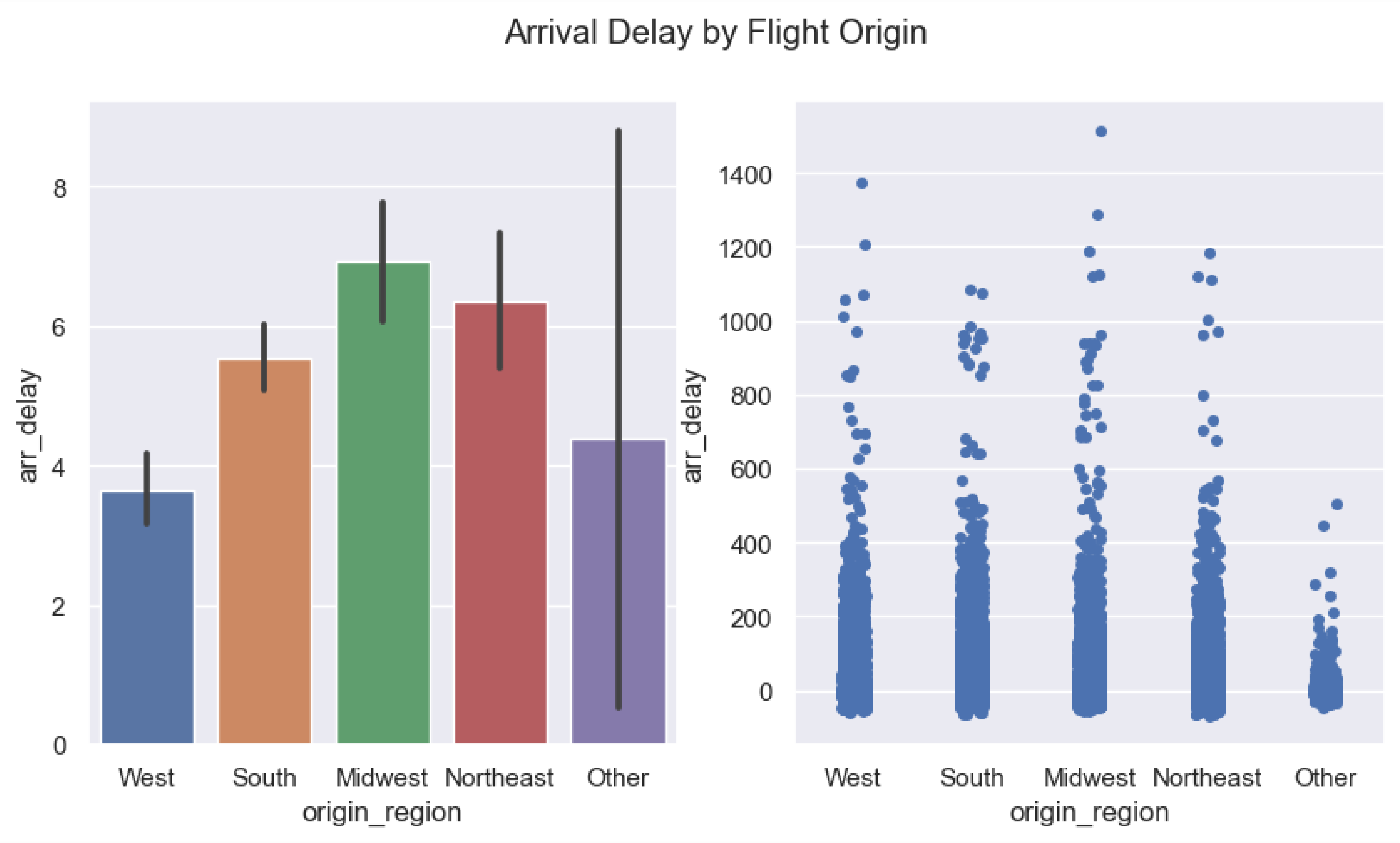


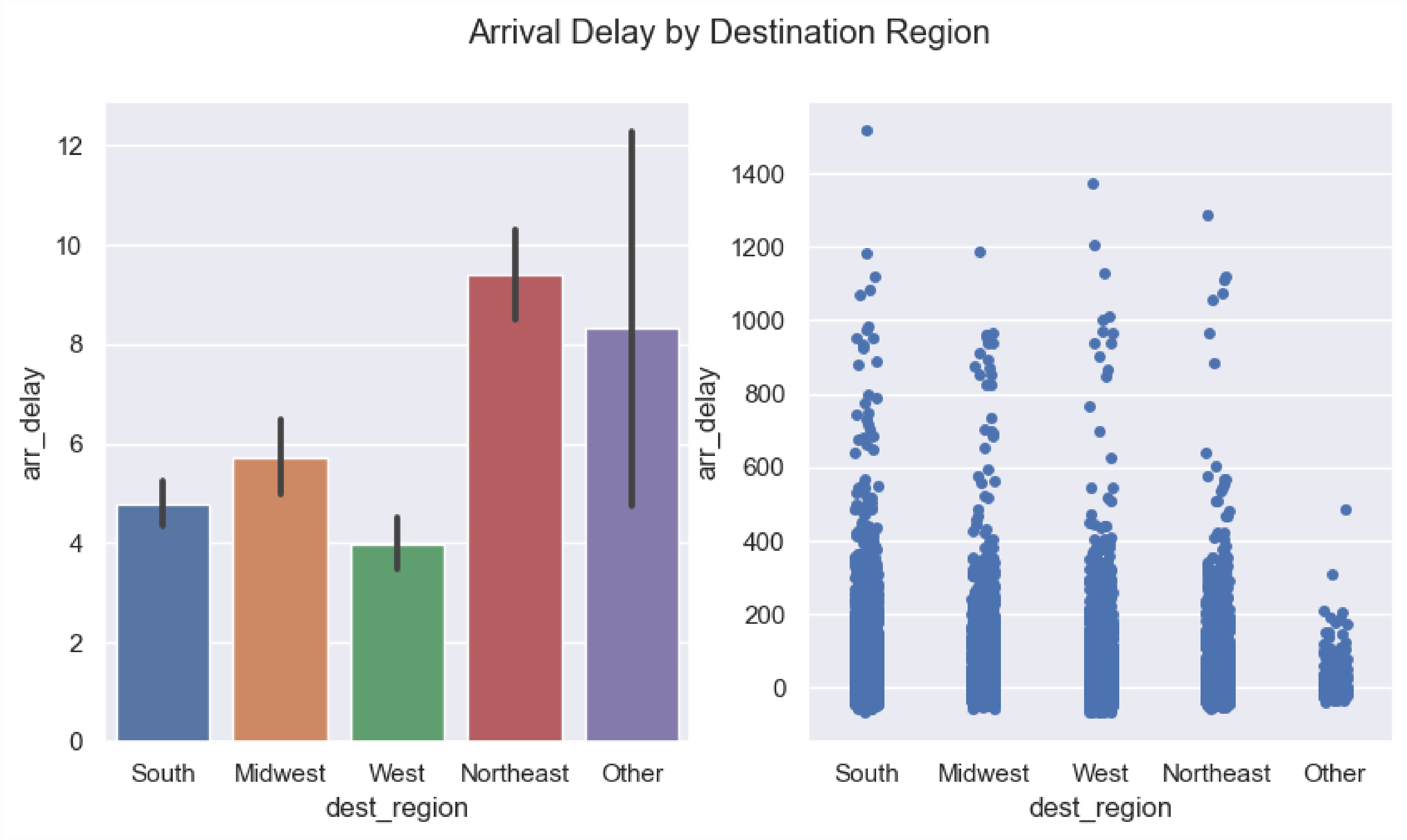

Arrival Delay by Departure Time of Day
arr_delay
crs_dep_time_of_day
early morning
late morning
midday
late afternoon
evening
12
10
8
6
4
2
0
−2
−4
1400
1200
1000
800
600
400
200
0


Arrival Delay by Arrival Time of Day
arr_delay
crs_arr_time_of_day
early morning
late morning
midday
late afternoon
evening
12
10
8
6
4
2
0
−2
1400
1200
1000
800
600
400
200
0

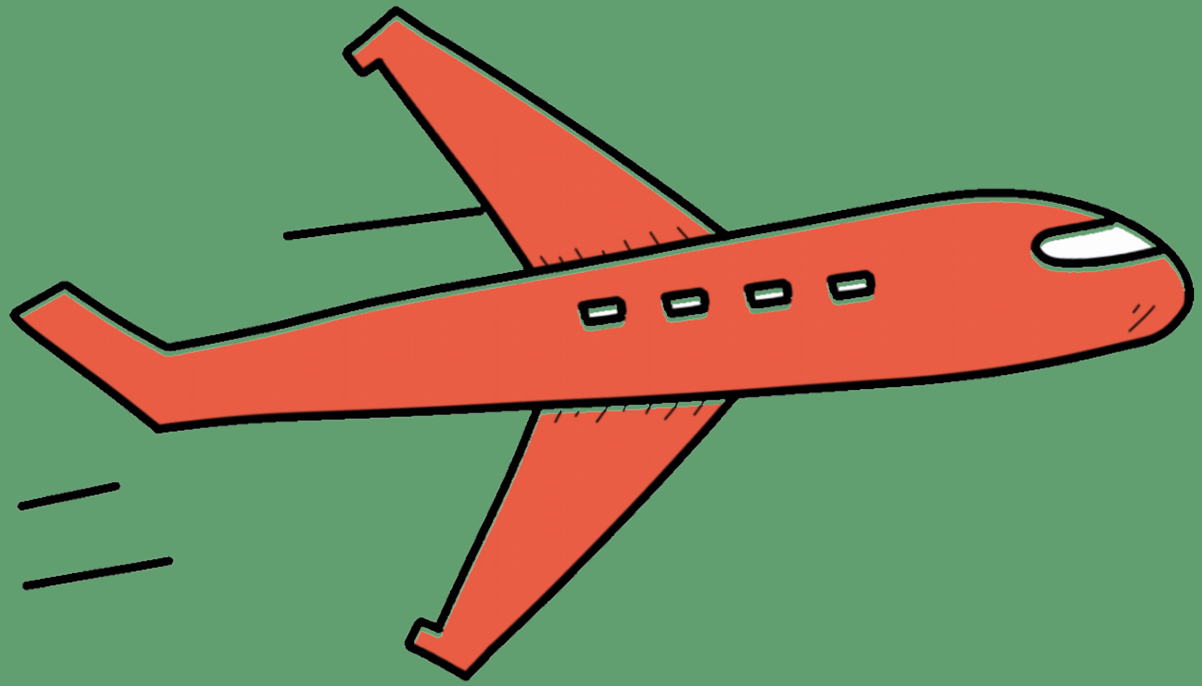
Features based on Origin/Destination City

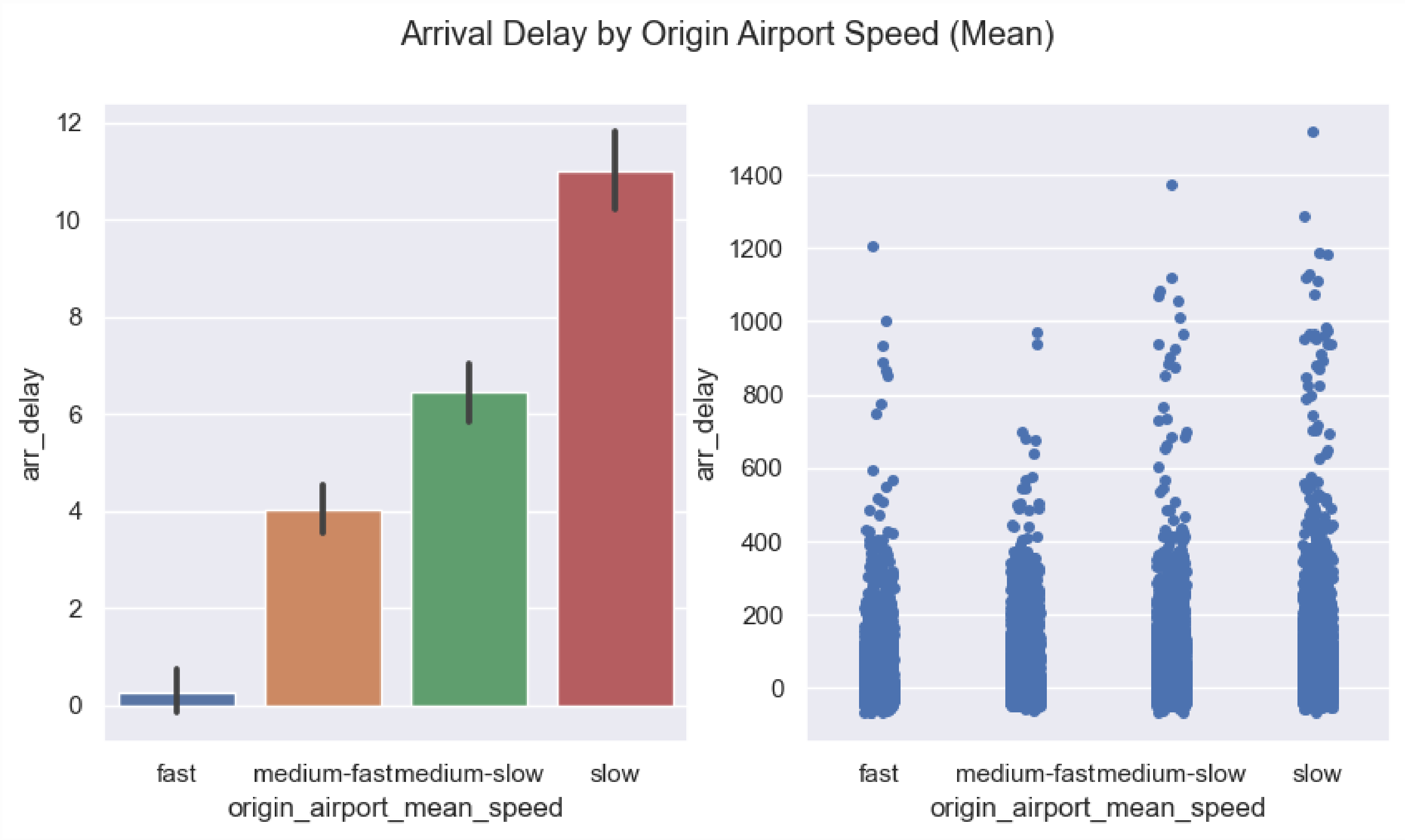


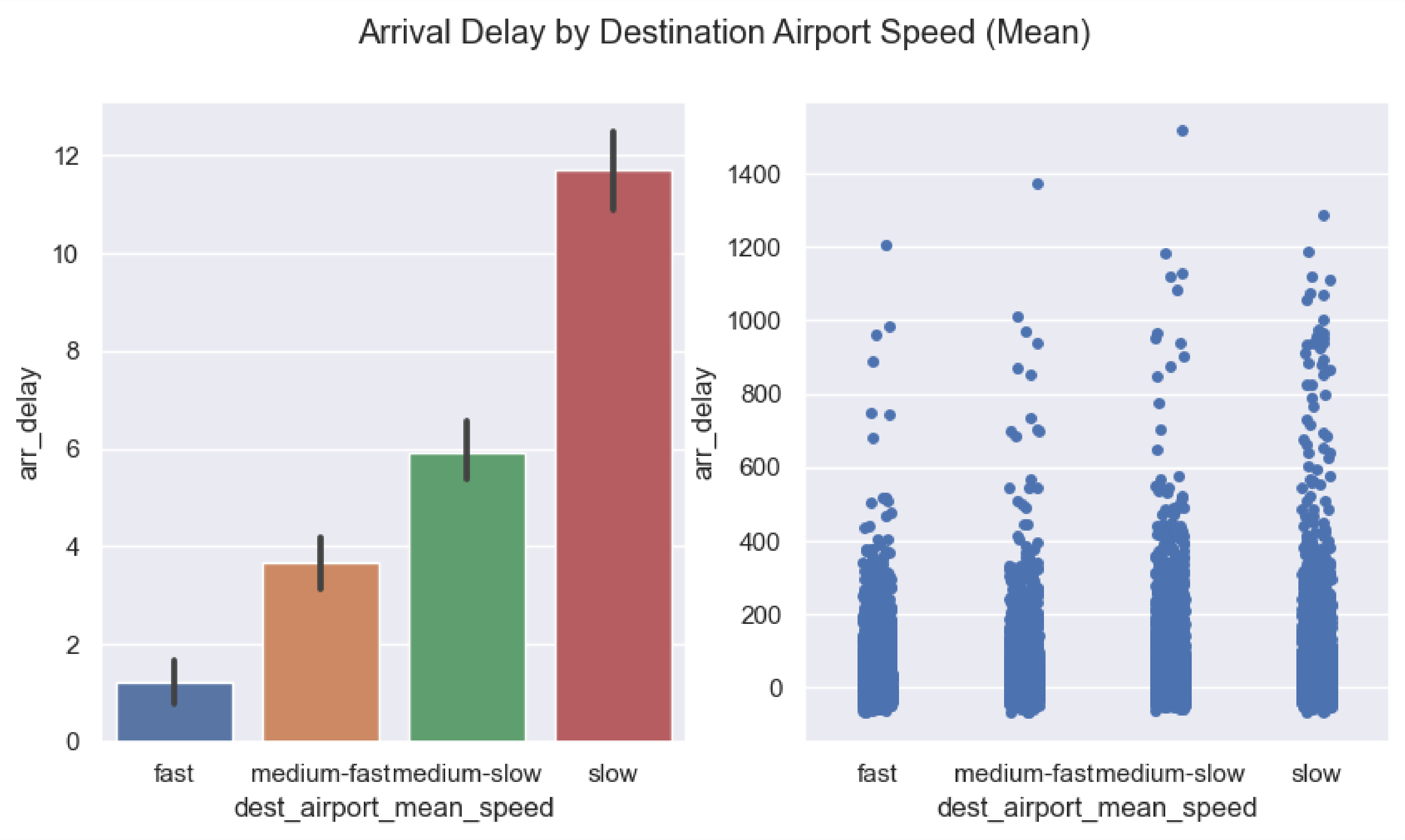

Arrival Delay by Flight Origin
arr_delay
origin_region
West
South
Midwest
Northeast
Other
0
2
4
6
8
0
200
400
600
800
1000
1200
1400


Arrival Delay by Destination Region
arr_delay
dest_region
South
Midwest
West
Northeast
Other
0
2
4
6
8
10
12
0
200
400
600
800
1000
1200
1400

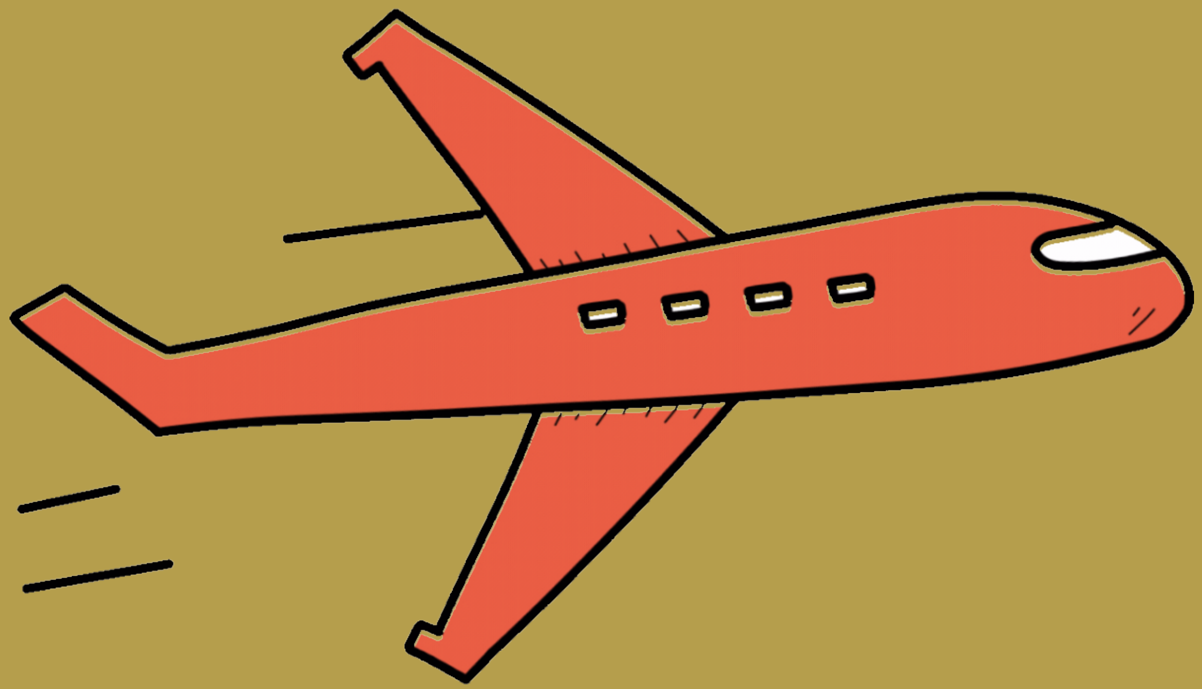
Features based on Airport Id

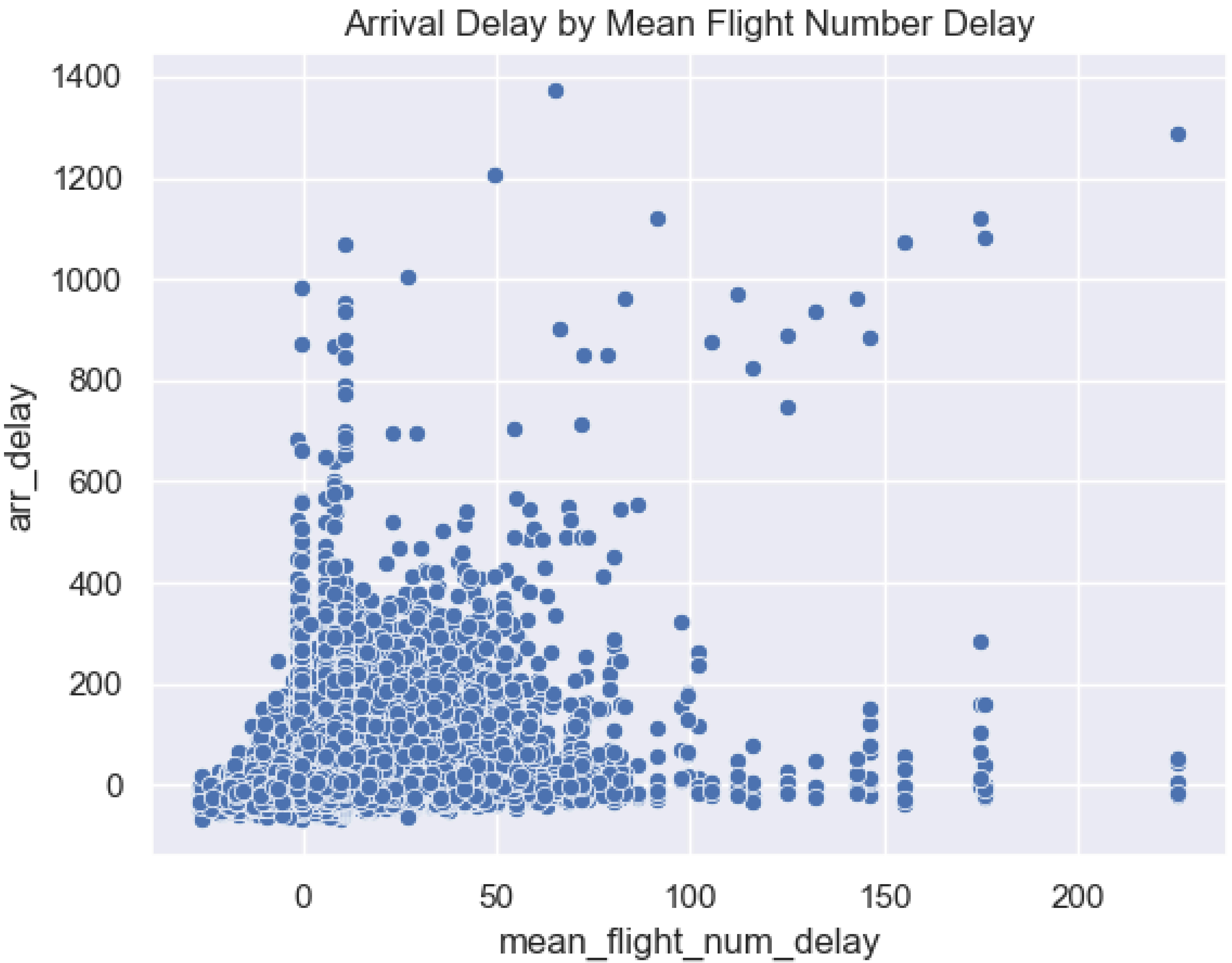



Arrival Delay by Origin Airport Speed (Mean)
arr_delay
origin_airport_mean_speed
fast
medium-fast
medium-slow
slow
0
2
4
6
8
10
12
0
200
400
600
800
1000
1200
1400


Arrival Delay by Destination Airport Speed (Mean)
arr_delay
dest_airport_mean_speed
fast
medium-fast
medium-slow
slow
0
2
4
6
8
10
12
0
200
400
600
800
1000
1200
1400

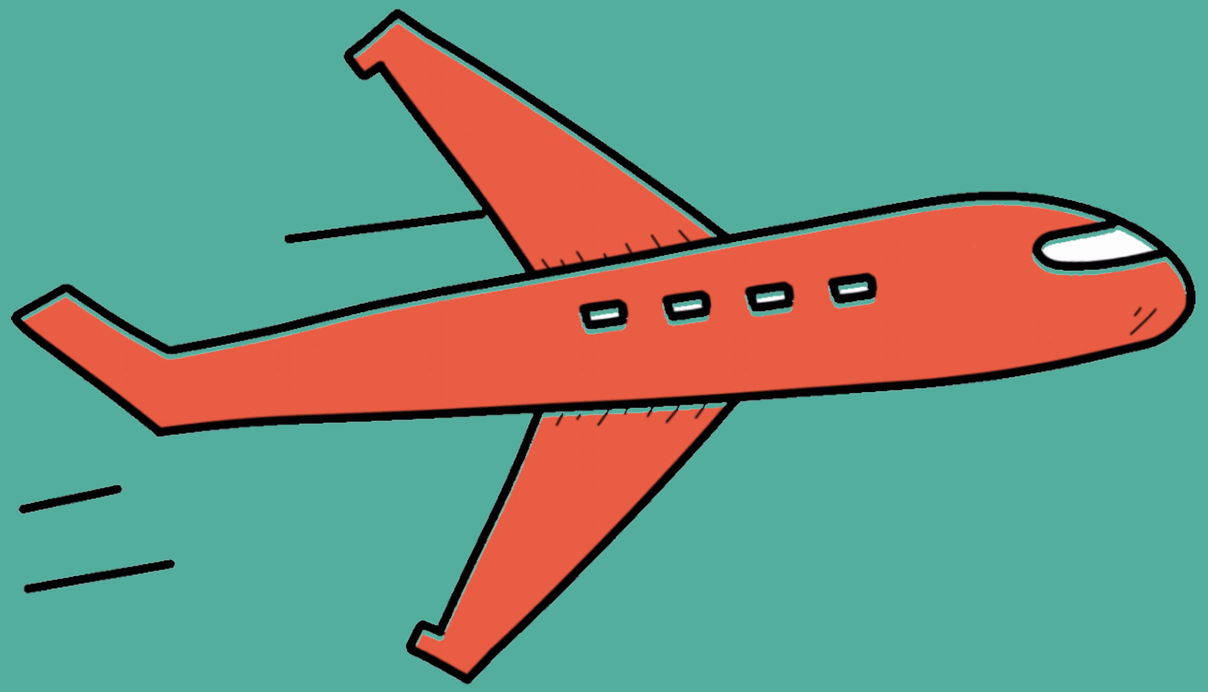
Features based on Flight Routes

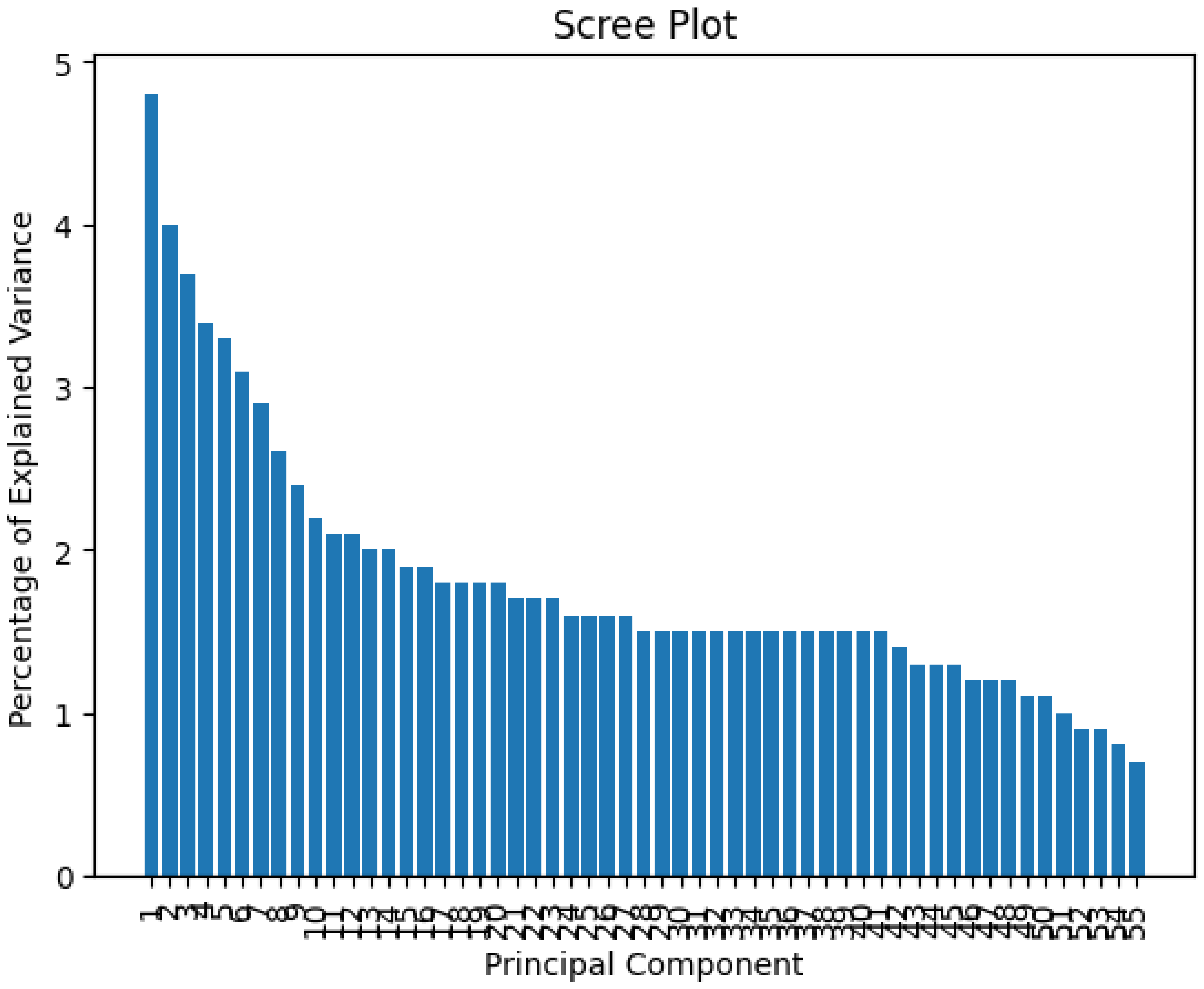




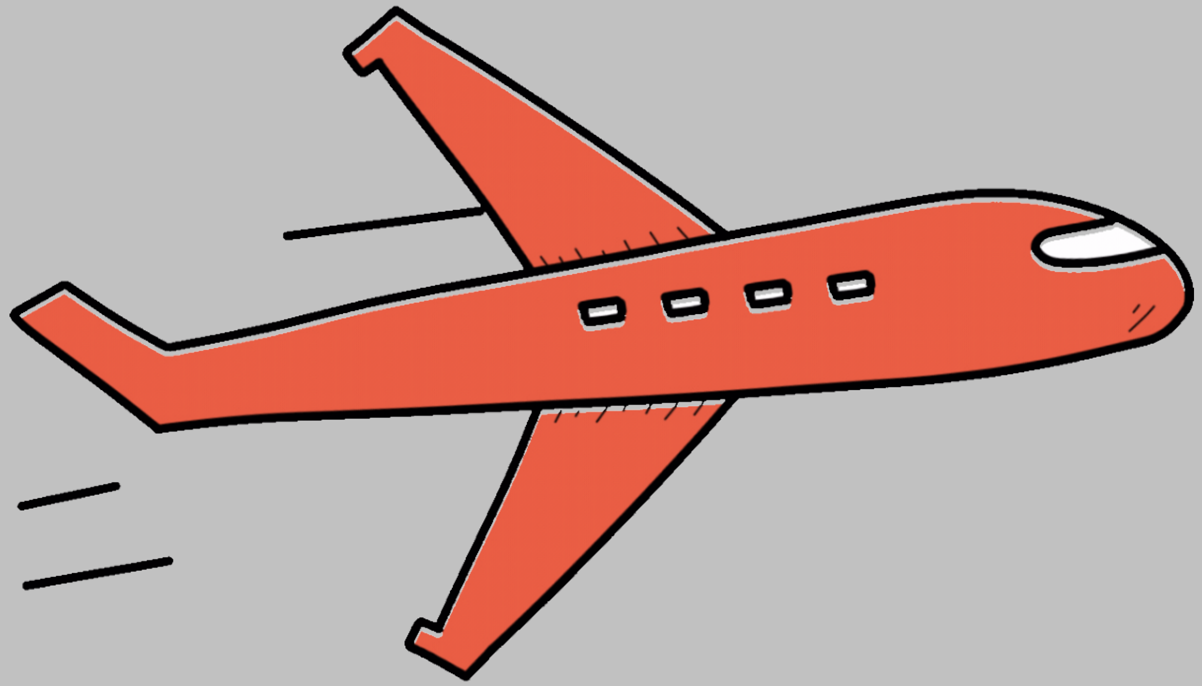
Correlation Coefficient = 0.27

Dimensionality Reduction




Scree Plot
Percentage of Explained Variance
Principal Component
0
1
2
3
4
5

Training Models



	MAE_test	MAE_train	RMSE_test	RMSE_train	R2_test	R2_train	ADJR2_test	ADJR2_train
linear	15.221647	15.392797	36.407460	34.151856	0.492709	0.485384	0.492709	0.485384
ridge	15.221711	15.392862	36.407493	34.151856	0.492708	0.485384	0.492708	0.485384
polynomial	15.197644	15.355645	36.486101	34.140478	0.490516	0.485727	0.490516	0.485727
lasso	15.370936	15.545292	36.493779	34.166493	0.490301	0.484943	0.490301	0.484943
sgd	15.309494	15.750729	36.497936	34.245050	0.490185	0.482572	0.490185	0.482572
r_forest	18.072724	18.122804	39.210367	35.901627	0.411593	0.431301	0.411593	0.431301
g_boost	15.522921	15.364055	39.487800	33.617242	0.403237	0.501370	0.403237	0.501370
voting_r	15.691872	15.270631	39.558562	33.401120	0.401096	0.507760	0.401096	0.507760
xgb	15.983414	15.201328	43.680043	33.057213	0.269799	0.517845	0.269799	0.517845
d_tree	16.588934	15.123663	45.439398	32.918182	0.209792	0.521892	0.209792	0.521892

What's Next

Check Tailnumber against the FAA's Registry

Examine how different types of airplanes impact delays

Integrating the CPC's Hurricane Season Outlook

Examine the impact of extreme weather conditions

Fine Tuning our Hyper-Parameters

Lessons Learned

Don't automatically drop outliers - only if they clearly don't make sense

Multiple dimensions presenting the same information in different ways weakens model performance

Split into train / test set before feature engineering

Thank you!