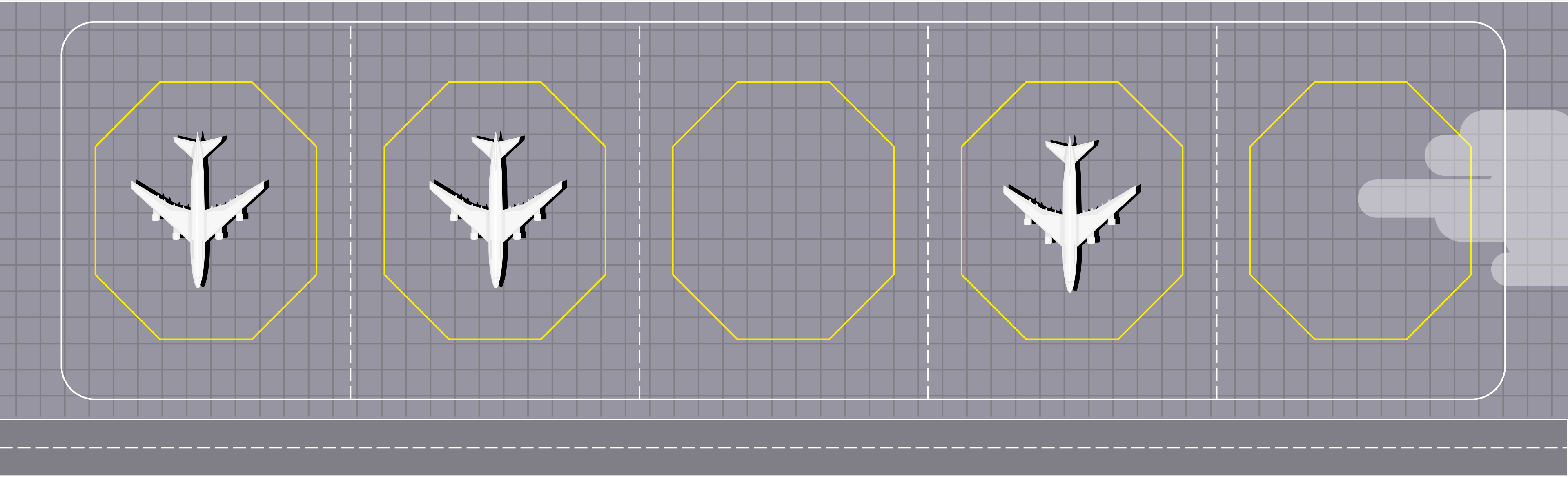
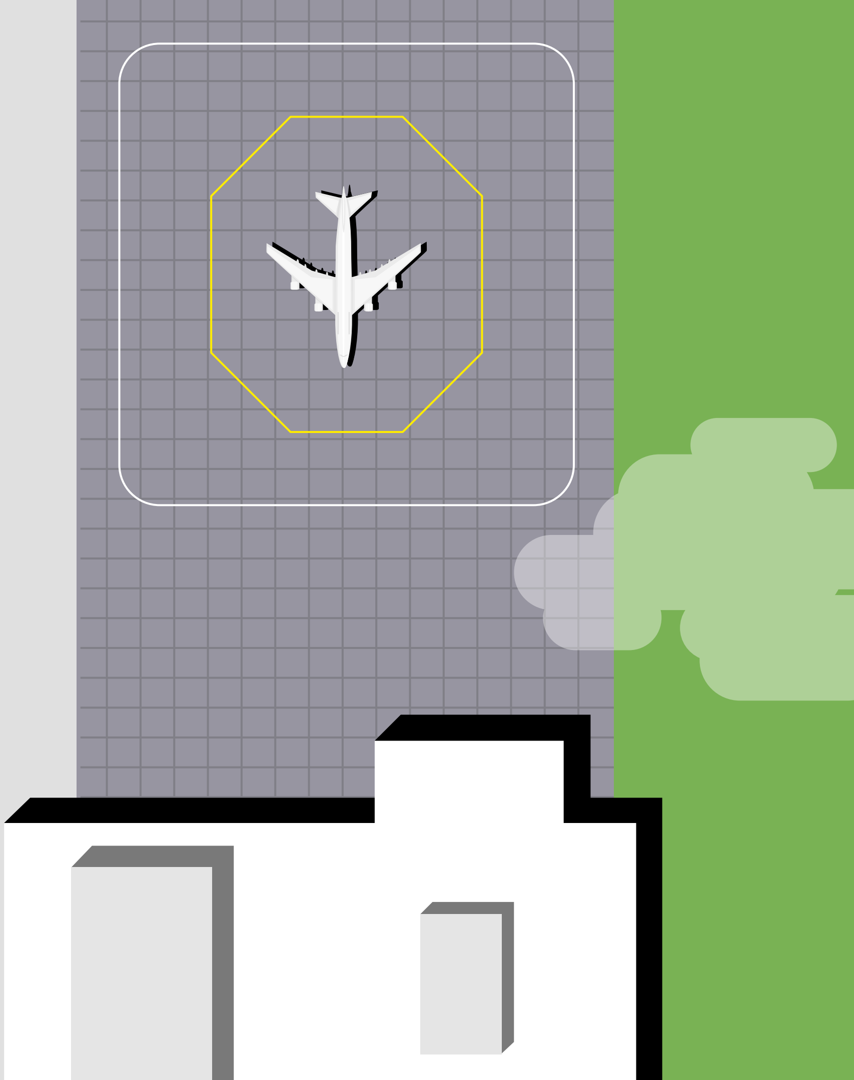
Flight Delay

Members : Aryan, Hon Joo, Adrian
Team 10



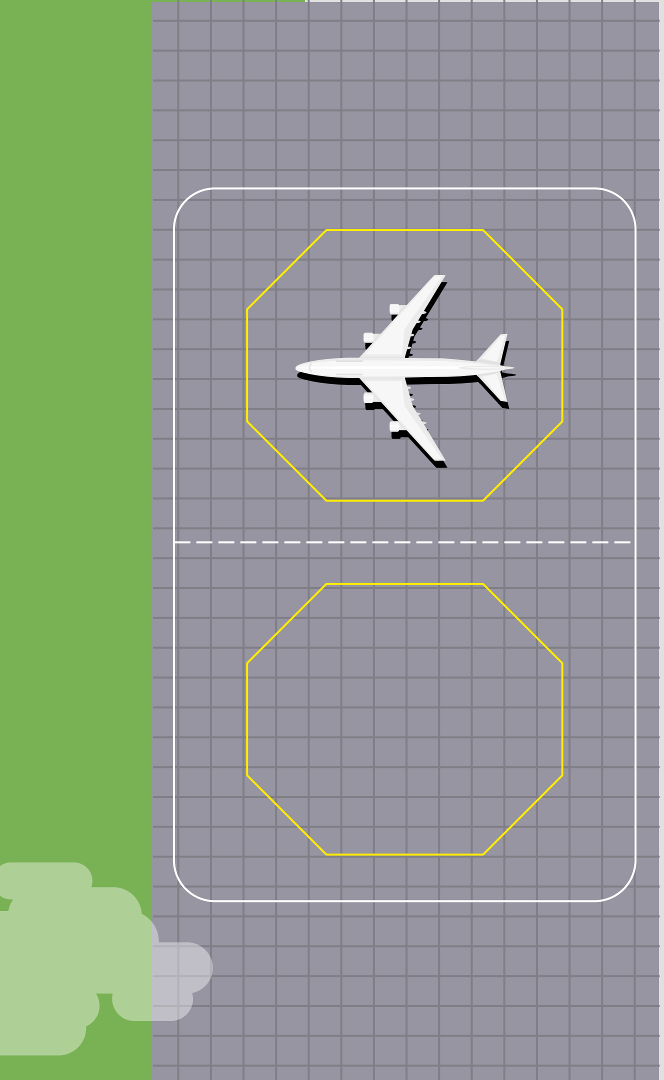
Problem Formulation

What's the problem?



Our problem definition

Everyone hates delays, but why are delays caused in the first place? Can they be prevented and predicted?



Data Preparation

What dataset would be suitable? What are the characteristics of the dataset? Are there missing values? Are there categorical variables?

Dataset

- Dataset of flight data in 2018
- Contains delay data and type of delay (NAS, etc.).
- Other data like destination, elapsed time are provided

```
FL_DATE
OP_CARRIER
OP_CARRIER_FL_NUM
ORIGIN
DEST
CRS_DEP_TIME
DEP_TIME
DEP_DELAY
TAXI_OUT
WHEELS_OFF
WHEELS_ON
TAXI_IN
CRS_ARR_TIME
ARR_TIME
ARR_DELAY
CANCELLED
CANCELLATION_CODE
DIVERTED
CRS_ELAPSED_TIME
ACTUAL_ELAPSED_TIME
AIR_TIME
DISTANCE
CARRIER_DELAY
WEATHER_DELAY
NAS_DELAY
SECURITY_DELAY
LATE_AIRCRAFT_DELAY
```

Why a 2018 dataset?

- 2018 is the earliest year in which data is still reliable
- Flights in part of 2019 and 2020 onwards affected by COVID-19

On time and early flights

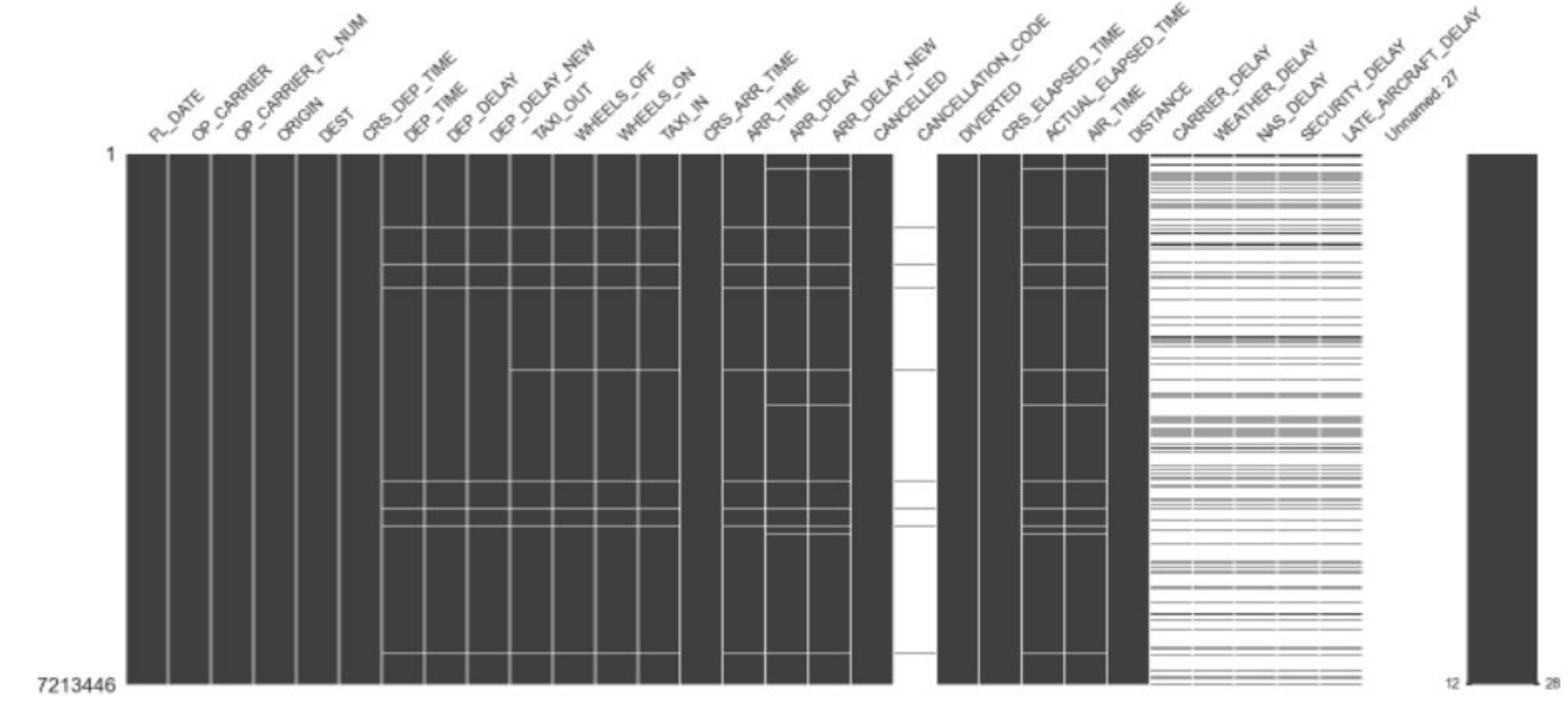
- Represented by zero and negative values in ARR_DELAY and DEP_DELAY
- Treat both cases as the same
- Set to zero for both cases
- Creation of DEP_DELAY_NEW and ARR_DELAY_NEW

	ARR_DELAY	ARR_DELAY_NEW	DEP_DELAY	DEP_DELAY_NEW
0	-23.0	0.0	-5.0	0.0
1	-24.0	0.0	-8.0	0.0
2	-13.0	0.0	-5.0	0.0
3	-2.0	0.0	6.0	6.0
4	14.0	14.0	20.0	20.0
5	-11.0	0.0	3.0	3.0
6	-16.0	0.0	-3.0	0.0
7	-19.0	0.0	-6.0	0.0
8	-2.0	0.0	13.0	13.0
9	-17.0	0.0	-2.0	0.0
10	-16.0	0.0	-5.0	0.0
11	129.0	129.0	121.0	121.0
12	-26.0	0.0	-3.0	0.0
13	-3.0	0.0	11.0	11.0
14	73.0	73.0	76.0	76.0
15	55.0	55.0	54.0	54.0
16	25.0	25.0	72.0	72.0
17	29.0	29.0	47.0	47.0
18	-18.0	0.0	-6.0	0.0
19	-21.0	0.0	9.0	9.0

Missing values

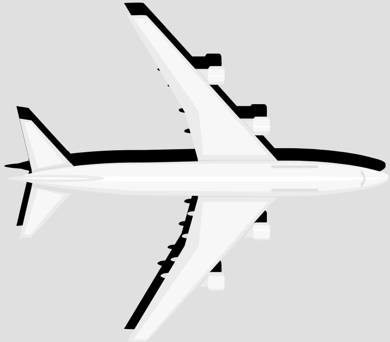
- Some columns contain missing values
- To clean the dataset, we remove these entries
- Visually represented using Missingno library

Missingno representation



Before and after removing missing delay values

Column	Non-Null Count
FL_DATE	7213446 non-null
OP_CARRIER	7213446 non-null
OP_CARRIER_FL_NUM	7213446 non-null
ORIGIN	7213446 non-null
DEST	7213446 non-null
CRS_DEP_TIME	7213446 non-null
DEP_TIME	7101129 non-null
DEP_DELAY	7096212 non-null
DEP_DELAY_NEW	7096212 non-null
TAXI_OUT	7097616 non-null
WHEELS_OFF	7097617 non-null
WHEELS_ON	7094200 non-null
TAXI_IN	7094200 non-null
CRS_ARR_TIME	7213446 non-null
ARR_TIME	7094201 non-null
ARR_DELAY	7076406 non-null
ARR_DELAY_NEW	7076406 non-null
CANCELLED	7213446 non-null
CANCELLATION_CODE	116584 non-null
DIVERTED	7213446 non-null
CRS_ELAPSED_TIME	7213436 non-null
ACTUAL_ELAPSED_TIME	7079004 non-null
AIR_TIME	7079004 non-null
DISTANCE	7213446 non-null
CARRIER_DELAY	1352710 non-null
WEATHER_DELAY	1352710 non-null
NAS_DELAY	1352710 non-null
SECURITY_DELAY	1352710 non-null
LATE_AIRCRAFT_DELAY	1352710 non-null



Column	Non-Null Count
FL_DATE	7071818 non-null
OP_CARRIER	7071818 non-null
OP_CARRIER_FL_NUM	7071818 non-null
ORIGIN	7071818 non-null
DEST	7071818 non-null
CRS_DEP_TIME	7071818 non-null
DEP_TIME	7071818 non-null
DEP_DELAY	7071818 non-null
DEP_DELAY_NEW	7071818 non-null
TAXI_OUT	7071818 non-null
WHEELS_OFF	7071818 non-null
WHEELS_ON	7071818 non-null
TAXI_IN	7071818 non-null
CRS_ARR_TIME	7071818 non-null
ARR_TIME	7071818 non-null
ARR_DELAY	7071818 non-null
ARR_DELAY_NEW	7071818 non-null
CANCELLED	7071818 non-null
CANCELLATION_CODE	0 non-null
DIVERTED	7071818 non-null
CRS_ELAPSED_TIME	7071818 non-null
ACTUAL_ELAPSED_TIME	7071817 non-null
AIR_TIME	7071817 non-null
DISTANCE	7071818 non-null
CARRIER_DELAY	1352375 non-null
WEATHER_DELAY	1352375 non-null
NAS_DELAY	1352375 non-null
SECURITY_DELAY	1352375 non-null
LATE_AIRCRAFT_DELAY	1352375 non-null

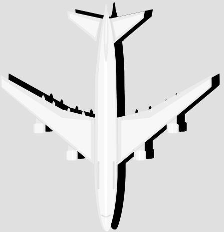
Observation: Flights with missing delay data are actually cancelled flights!

Encoding categorical data

- Columns like DEST and OP_CARRIER are categorical
- Encode them for easier machine learning
- Assign unique index to each carrier and airport

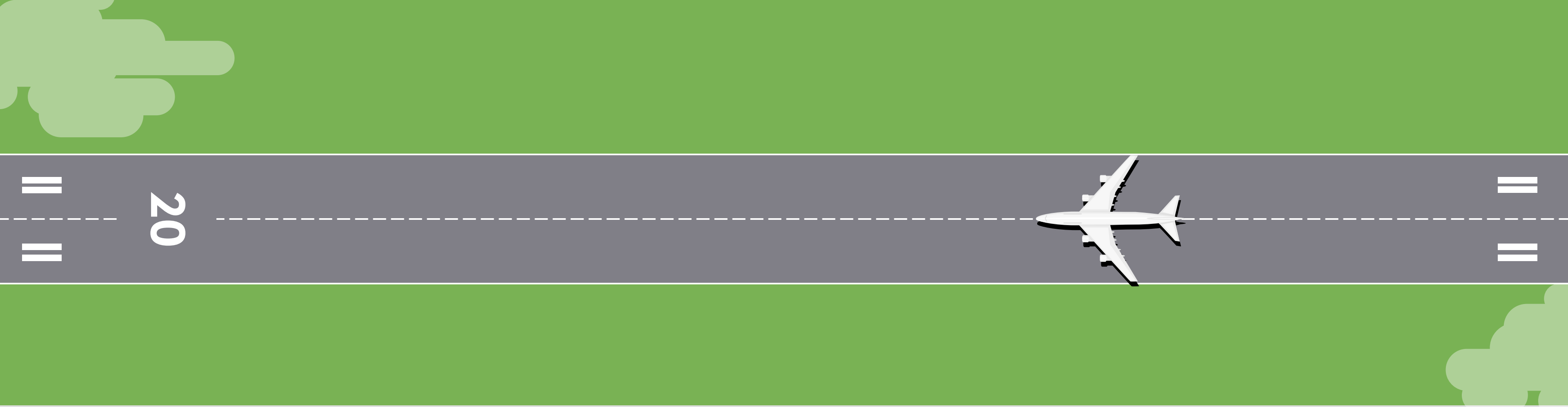
Before and after encoding

	FL_DATE	OP_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	DEST	CRS_DEP_TIME	DEP_TIME	DEP_DELAY
0	2018-01-01	UA	2429	EWR	DEN	1517	1512.0	-5.0
1	2018-01-01	UA	2427	LAS	SFO	1115	1107.0	-8.0
2	2018-01-01	UA	2426	SNA	DEN	1335	1330.0	-5.0



	FL_DATE	OP_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	DEST	CRS_DEP_TIME	DEP_TIME	DEP_DELAY
0	2018-01-01	0	2429	7	0	1517	1512.0	-5.0
1	2018-01-01	0	2427	5	1	1115	1107.0	-8.0
2	2018-01-01	0	2426	52	0	1335	1330.0	-5.0

Note: We can maintain a dictionary mapping indexes to each airport/carrier for future reference

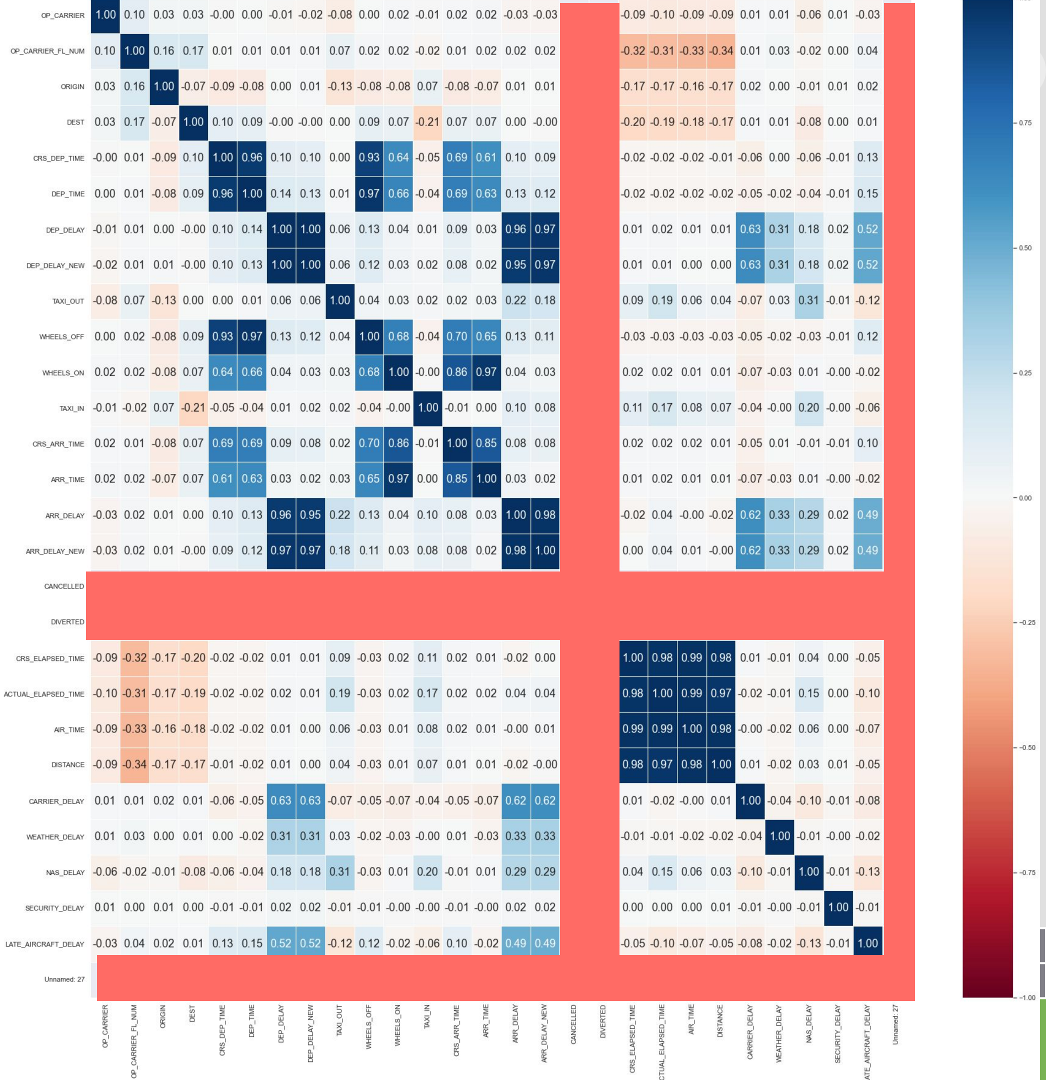


Exploratory Data Analysis

What is the mean departure delay experienced? How is the departure delay distributed? Are there outliers?

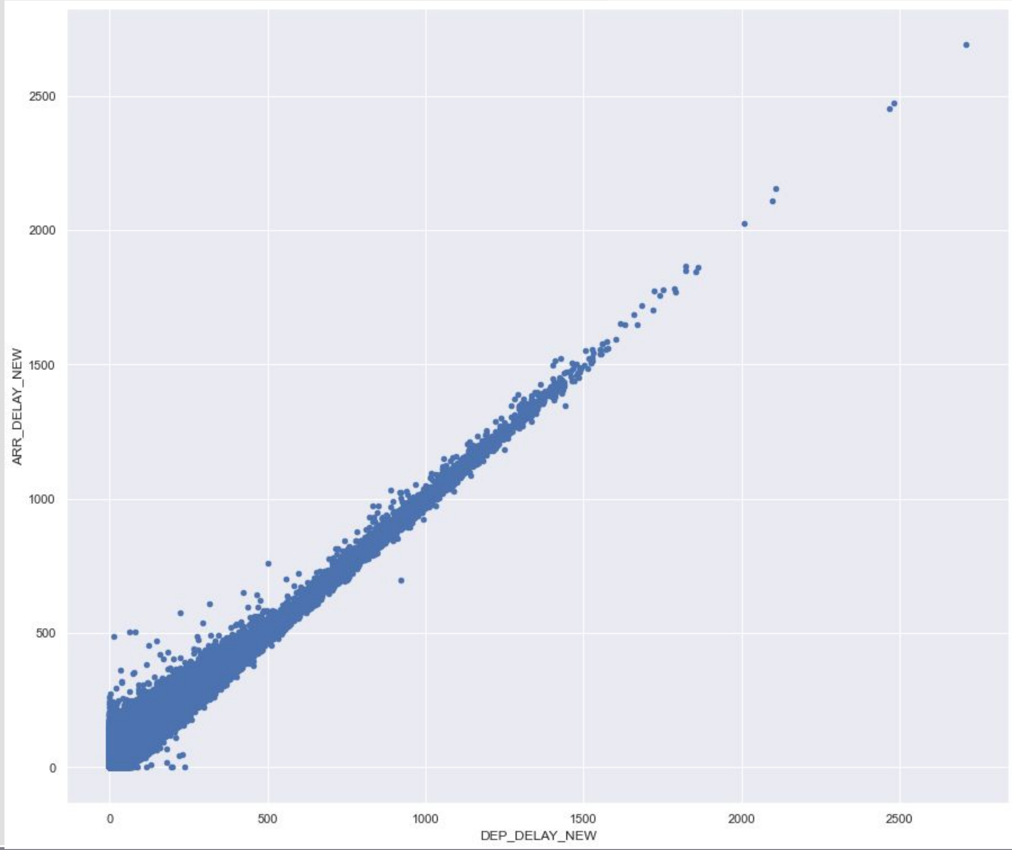
Choosing predictors with high correlation

- After using heatmap, we observed that some correlation are too low or NULL which we do not need



Arrival and Departure delay

Based on the observation, we can see that DEP_DELAY_NEW and ARR_DELAY_NEW are highly correlated.



Outliers

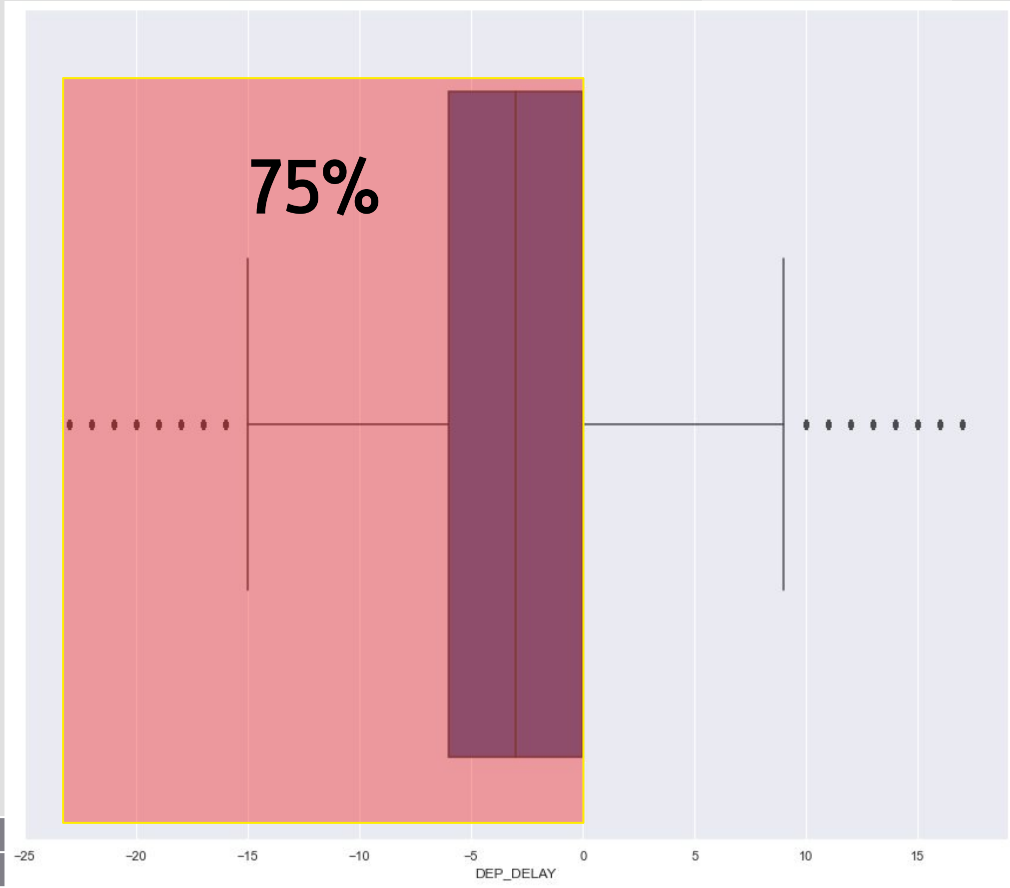
	ARR_DELAY_NEW	DEP_DELAY_NEW
count	7.071818e+06	7.071818e+06
mean	1.342195e+01	1.313459e+01
std	4.348008e+01	4.352941e+01
min	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00
75%	8.000000e+00	7.000000e+00
max	2.692000e+03	2.710000e+03

Applying formula to remove outliers

	ARR_DELAY_NEW	DEP_DELAY_NEW
count	3.670944e+06	3.670944e+06
mean	1.144973e+00	1.398853e+00
std	3.228388e+00	3.481568e+00
min	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00
75%	0.000000e+00	0.000000e+00
max	2.000000e+01	1.700000e+01

Departure Delay

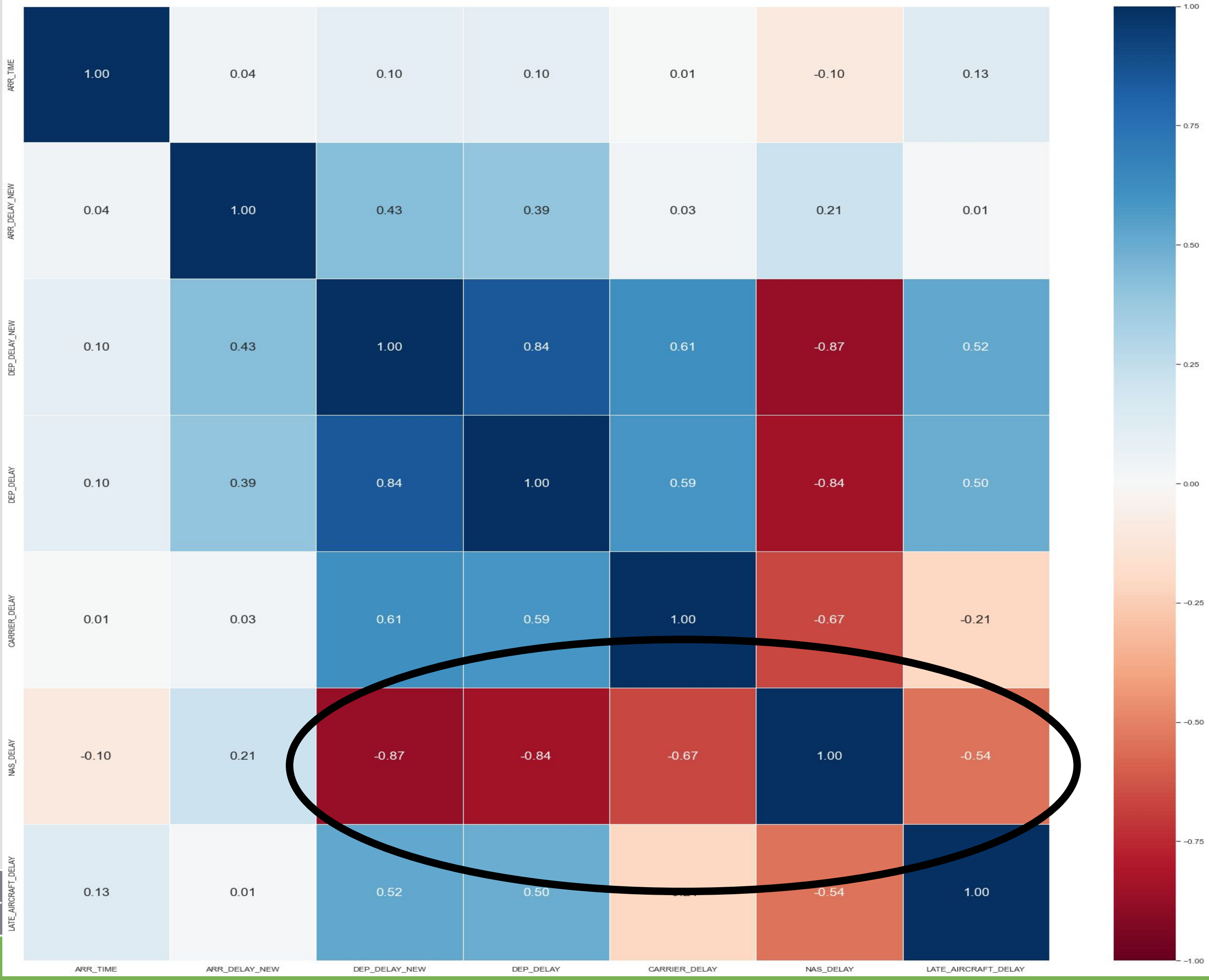
- Delays are mostly negative
- Convert all those negative delays to zeros
- Change to DEP_DELAY_NEW



Heatmap (After removing outliers)

NAS delay is now highly correlated !!!

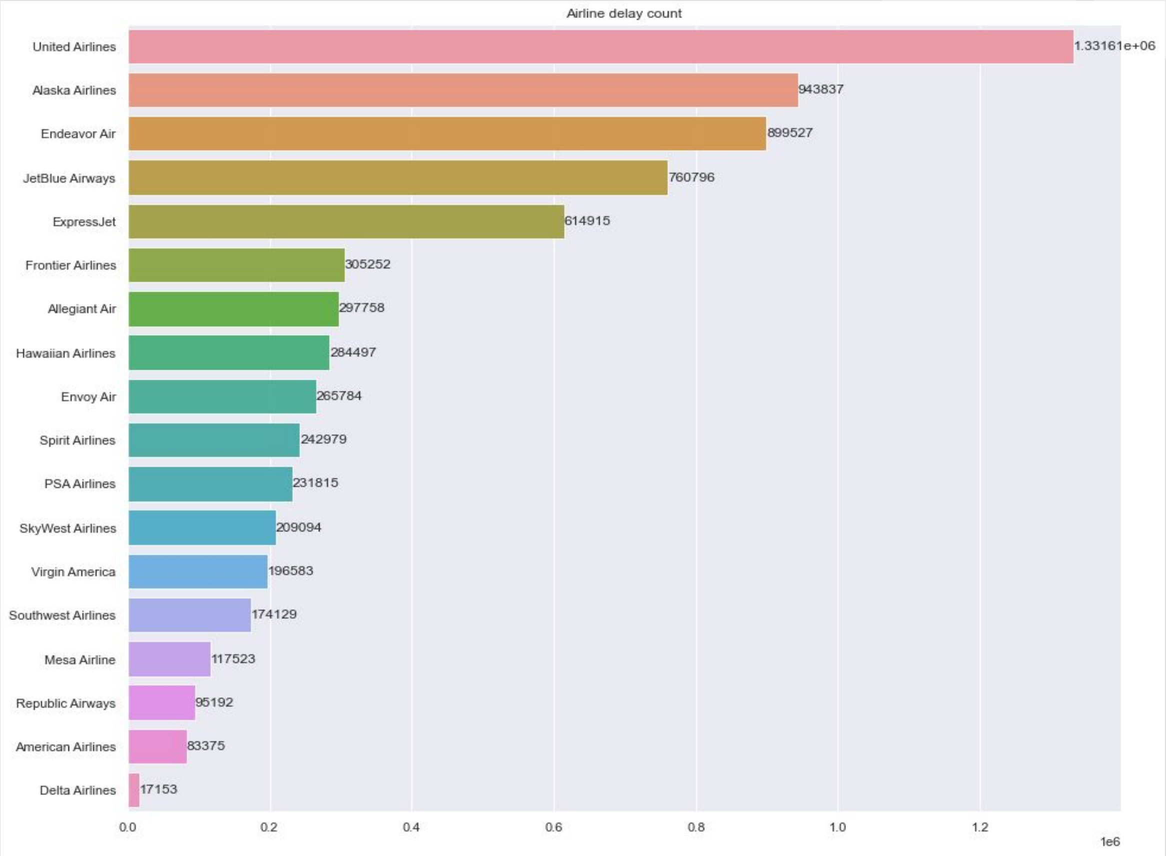
Corr : **0.29-> -0.87**



Airline

TOP 5 Airlines with highest Delay Count

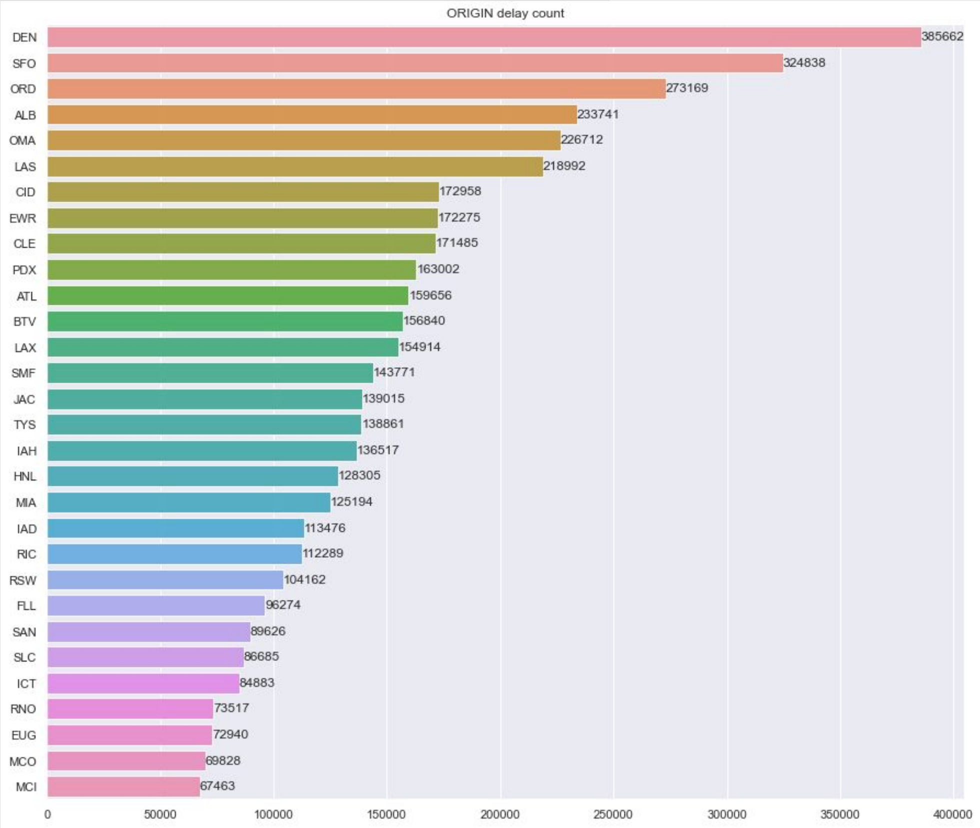
1) United Airlines
2) Alaska Airlines
3) Endeavor Air
4) JetBlue Airways
5) ExpressJet



Origin (Cities)

Top 5 States with highest delay count

1) Denver
2) San Francisco
3) O'Hare Airport (Chicago)
4) Albany International Airport (New York)
5) Omaha (Eppley Airfield)



Formulating the Prediction Problem

Response Variable: DEP_DELAY_NEW

Predictors: (NAS_DELAY, ORIG, DEST, LATE_AIRCRAFT_DELAY, CARRIER_DELAY, OP_CARRIER)

Loss Function: Mean Squared Error

Metric: Explained Variance, Mean Squared Error

Dataset: 80-20 split of both datasets: Missing Value Removed and Missing Value Imputed

Models for Regression

- Linear Regression: capture linear trends in the data, if any.
- Support Vector Machine: capture complex non-linear trends in the data, if any.
- Dense Neural Network: capture complex non-linear trends and linear trends present in data, if any, by making use of linear and non-linear activations like Rectified Linear Units.
- GridSearch: tune important hyperparameters of SVM.
- XGBoost: apply Gradient Boosting techniques

RESULTS

Which models had the lowest loss?
Which generalized the best?

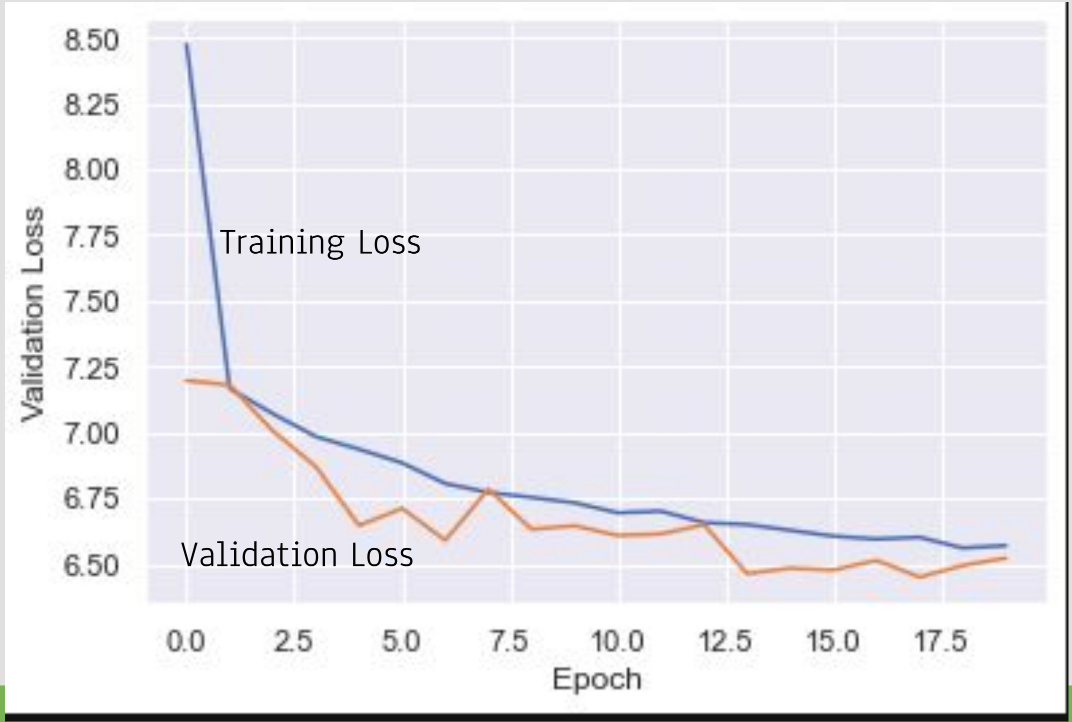
Linear Regression Performance

	Missing Value Imputed Training	Missing Value Removed Training	Missing Value Imputed Validation	Missing Value Removed Validation
Explained Variance	0.044	0.815	0.040	0.814
Mean Squared Error	11.59	7.29	11.61	7.3

SVM Regression Performance

	Missing Value Imputed Training	Missing Value Removed Training	Missing Value Imputed Validation	Missing Value Removed Validation
Explained Variance	-0.17	0.78	-0.18	0.78
Mean Squared Error	14.3	8.44	14.28	8.5

Neural Network Performance

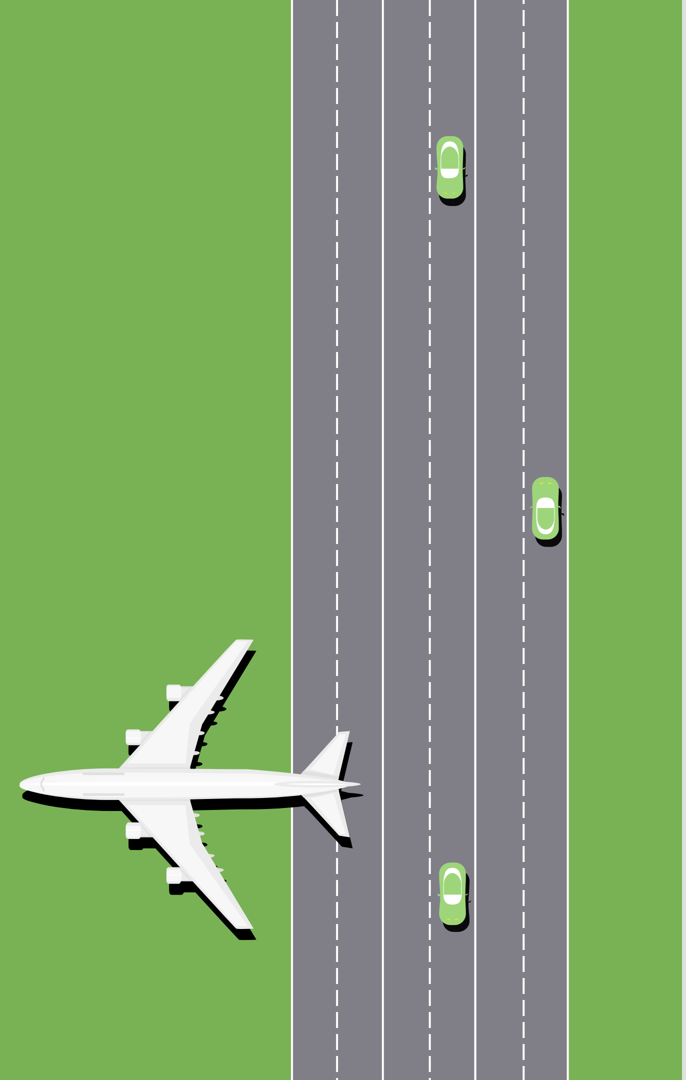


XGBoost Regression Performance

	Missing Value Removed Training	Missing Value Removed Validation
Mean Squared Error	4.79	5.77

SLIGHT OVERFITTING!

	Train Loss	Validation Loss
XGBoost	4.79	5.77
Deep Neural Network	6.3061	6.3209
Support Vector Machine	8.445	8.536
Linear Regression	7.294507	7.317763



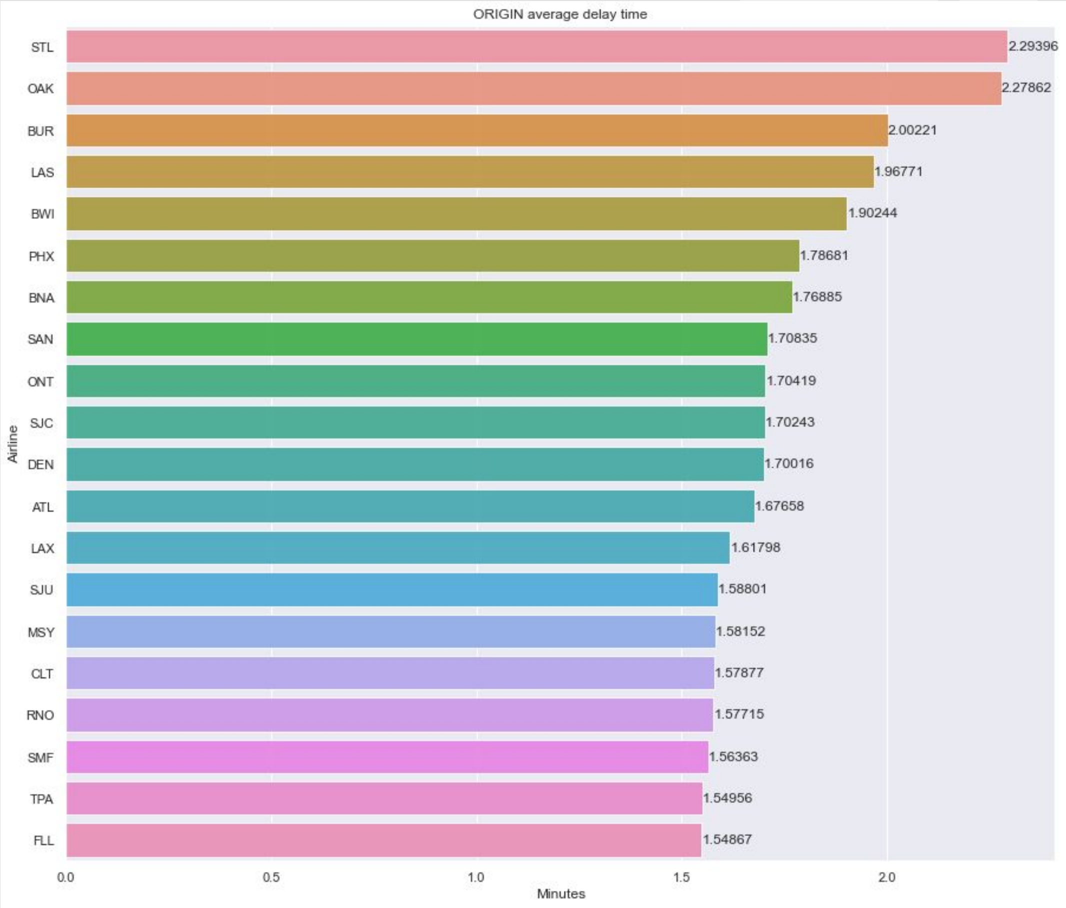
Recommendations and Insights

What really causes the delay? Can a passenger choose to do something to avoid the delay?

Origin (Cities)

TOP 5 cities with highest average delay time

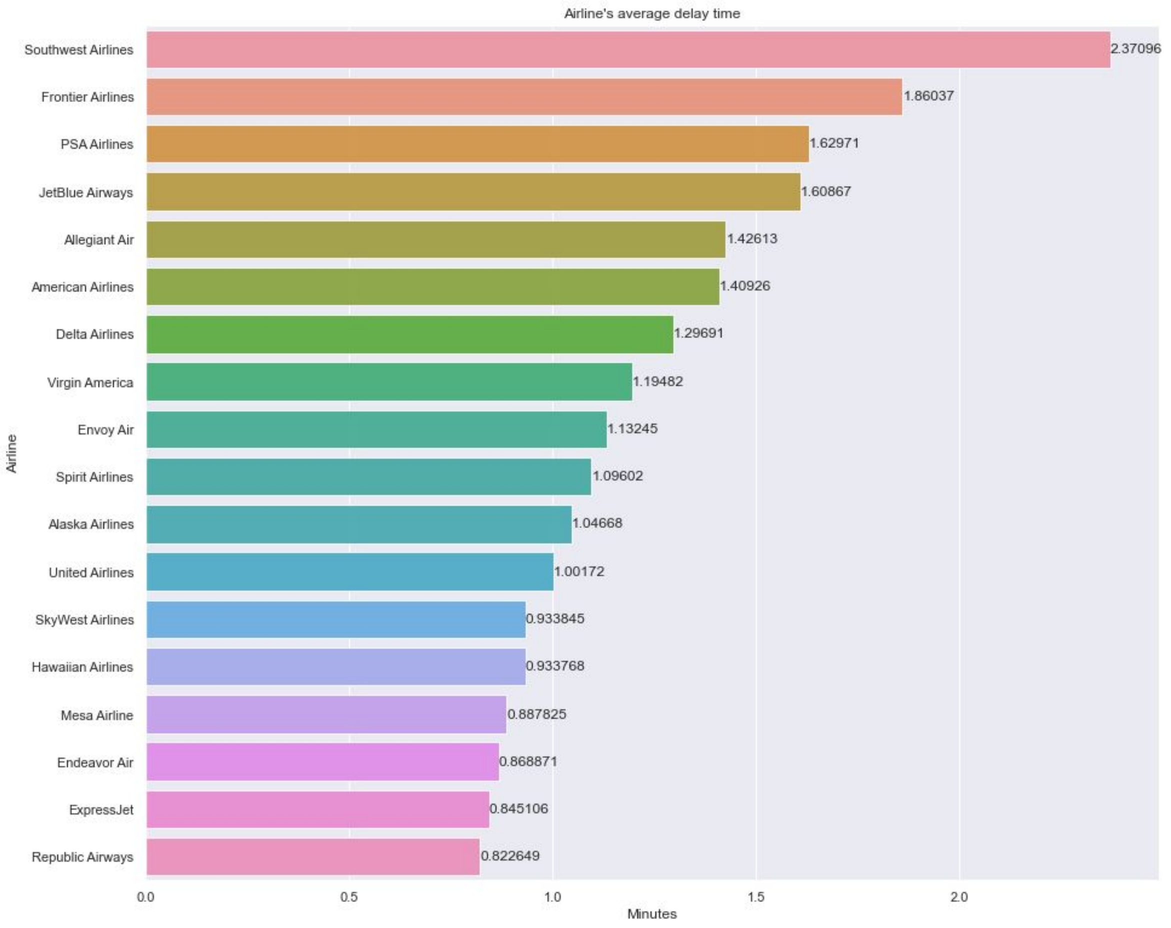
1) STL (St. Louis)
2) OAK (Oakland)
3) BUR (Burbank)
4) LAS (Las vegas)
5) BWI (Baltimore/D.C)

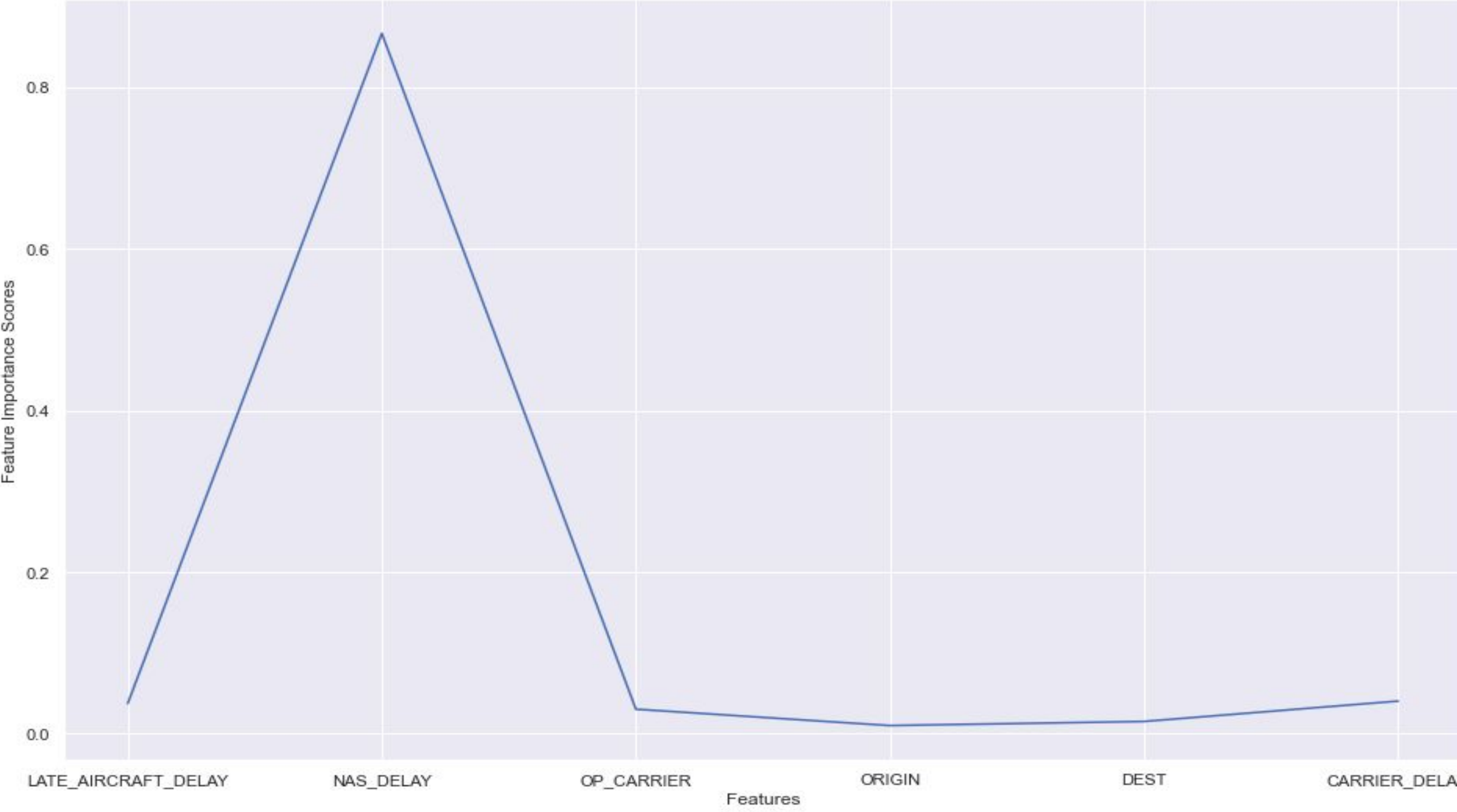


Airline

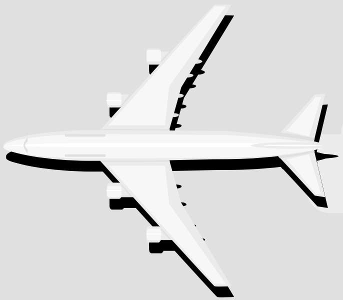
TOP 5 Airlines with highest average delay time

1) Southwest Airlines
2) Frontier Airlines
3) PSA Airlines
4) JetBlue Airways
5) Allegiant Air





0.8
0.6
0.4
0.2
0.0
Feature Importance Scores
LATE_AIRCRAFT_DELAY
NAS_DELAY
OP_CARRIER
ORIGIN
DEST
CARRIER_DELA
Features



THANK YOU AND HAVE A SAFE FLIGHT!