FLIGHT DELAYS

Itua Etiobhio
Yusuf Olodo

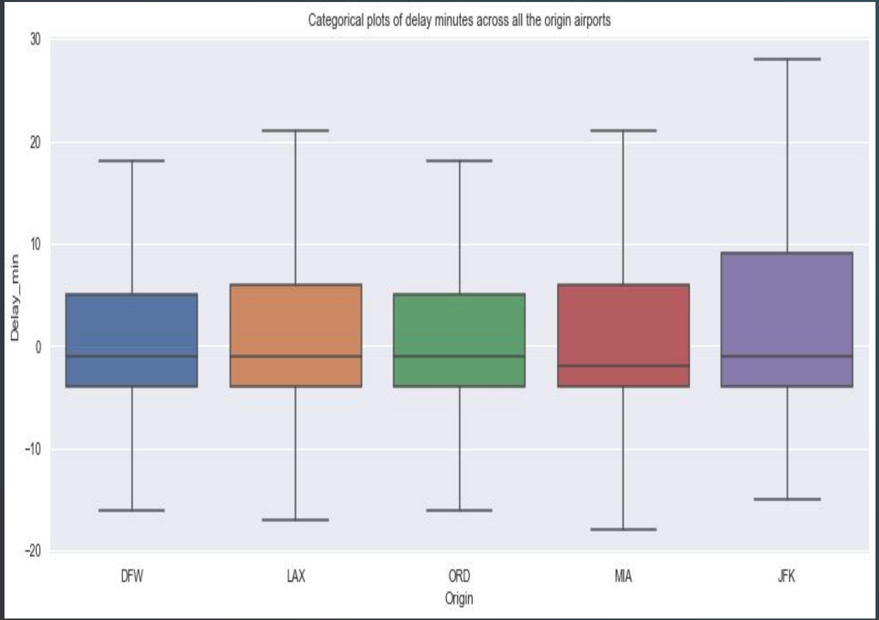
BUSINESS PROBLEM

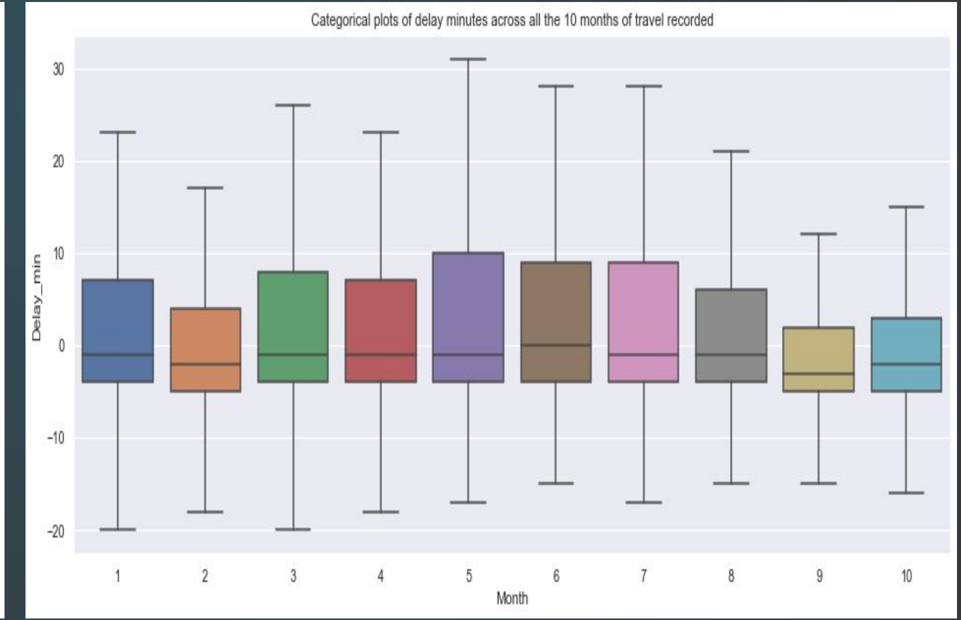
- A US airline has approached our data science company and are interested in knowing what factors affect the arrival delays of their flights.
- There are 6 questions the airline was interested in finding out that affects their business
- Our mission is to develop a robust and high performing model that can predict the arrival delays based on certain predictor variables

DATA ANALYSIS

1.Which airports have the highest and lowest average delay minutes amongst LAX,DFW,MIA, JFK, ORD

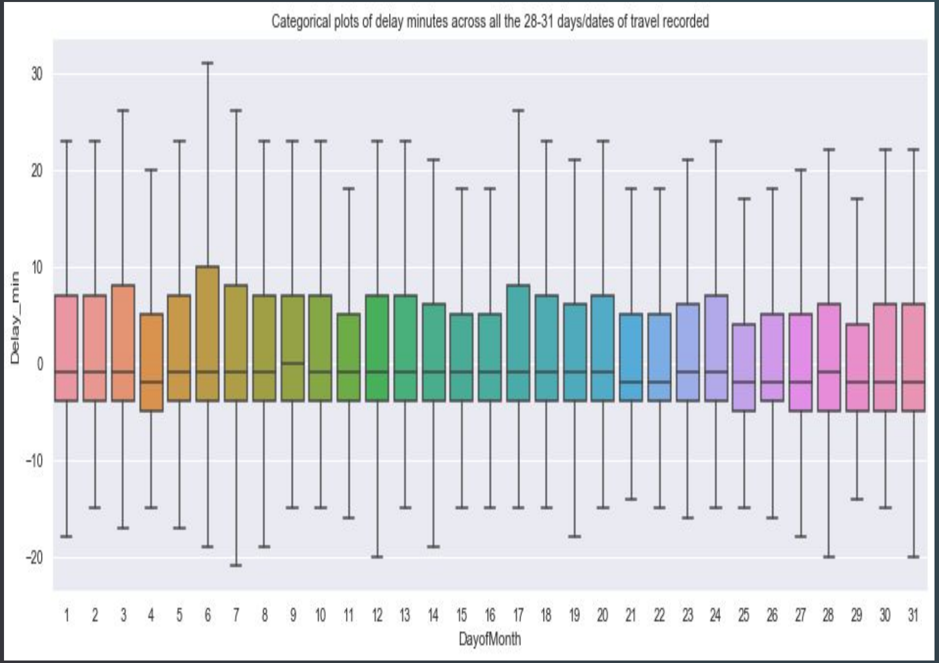
2.Which months have the highest and lowest average delay minutes amongst the months of travel?

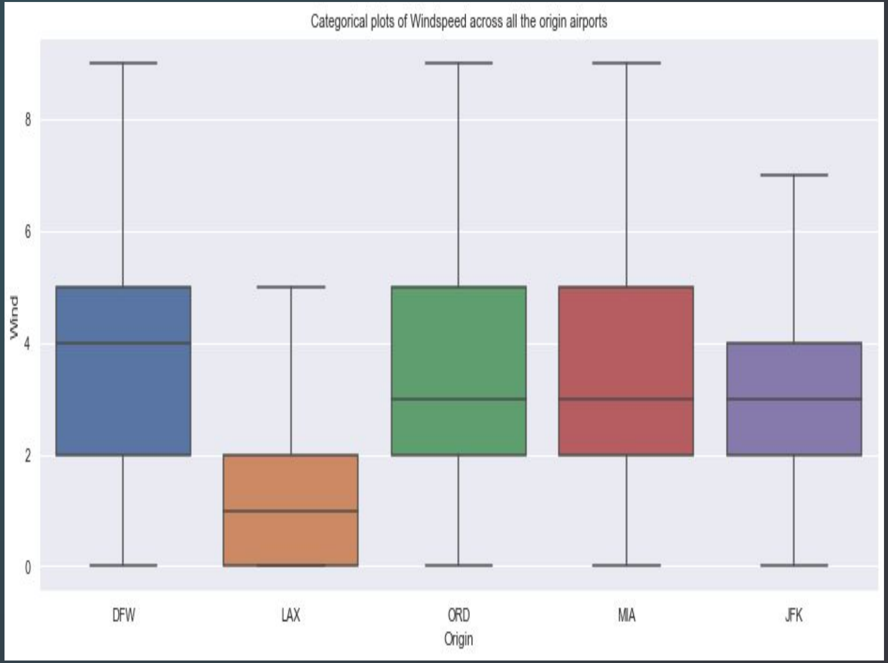




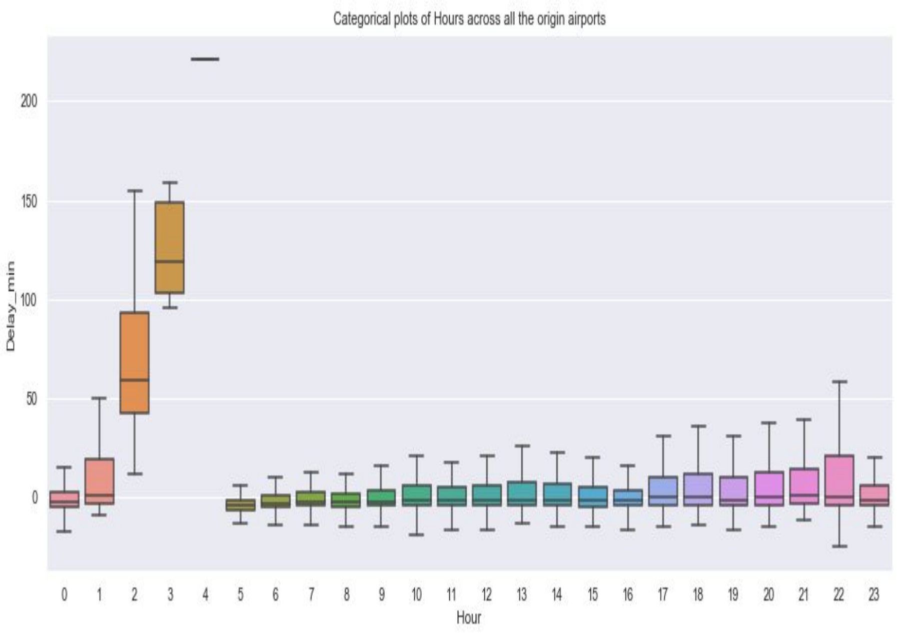
3.Which days have the highest and lowest average delay minutes?

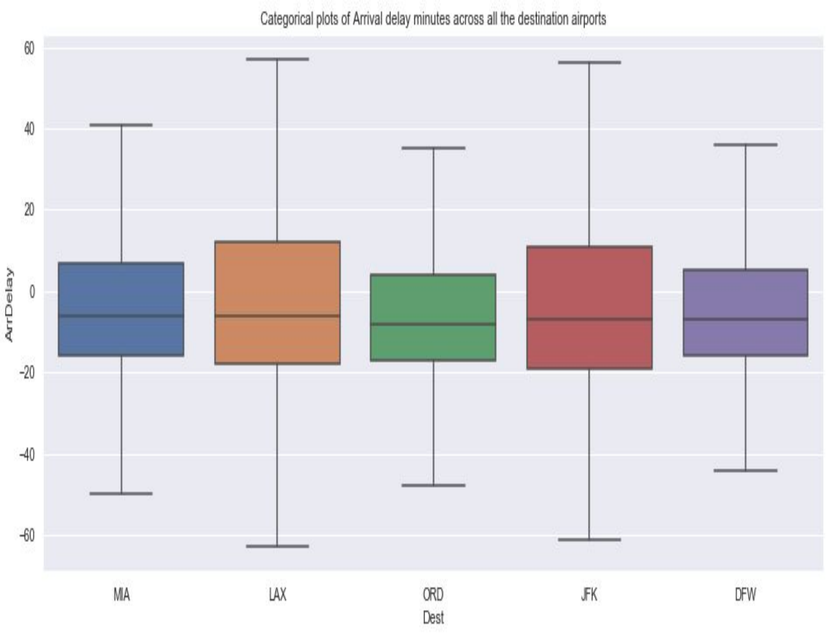
4.Which airports have the highest and lowest average wind speed (in terms of weather affecting service disruption)?





5.Which hours of the day have the highest and lowest average delay minutes across all the international airports ?
6.Which international airports have the highest and lowest average Arrival delay minutes?





RECOMMENDATIONS

We can confirm between the hours of 02:00 -03:00 are the worst times across all five airports

We can confirm that Miami International Airport has the highest average delay minutes in arrival followed by Los Angeles International Airport and the lowest average arrival delay minutes is O'Hare International Airport Chicago

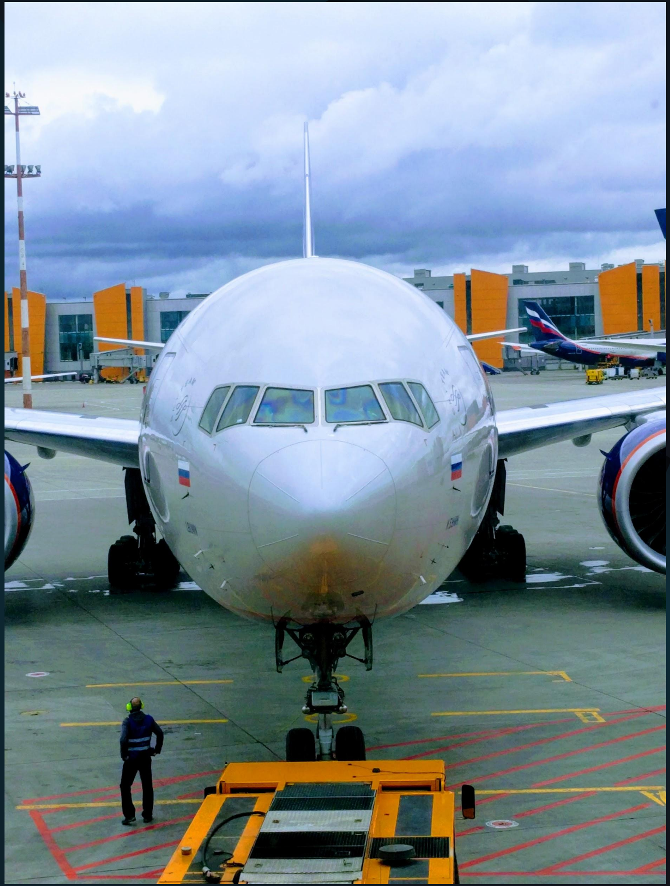
We can also confirm that Dallas Forth Worth Texas International airport has the worst conditions for wind speed and Los Angeles International airport has the best conditions

We can confirm that the highest average in terms of delay minutes occurs in the month of June lowest average occurs in the month of September

Building Regression Model

The main question we wanted to answer with building a regression model is:

1. which model can best explain the proportion of variance in the data



Building Regression Model

	Baseline Model	Lasso	Ridge	Polynomial
Training R-squared:	0.8248	0.7740	0.8248	0.8273
Testing R-squared	0.8059	0.7562	0.8058	0.8098
Training Mean Squared Error	361.73	466.60	361.73	18.88
Testing Mean Squared Error	379.73	476.76	379.72	19.38

Validation

Cross Validation for Polynomial model

Cross Validation Mean r2: 0.8237

Cross Validation Mean MSE: 357.29

Cross Validation 10 Fold Score: [0.75004849 0.83874685 0.83635621 0.81775607 0.84195497 0.85156864

0.82067991 0.84161514 0.82467689 0.81414576]

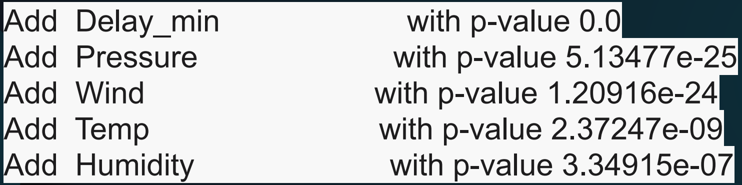
Cross Validation 10 Fold mean squared error [365.61985076 346.75390226 368.03679407 358.38808653
371.53955419 350.21998836 364.45530232 348.29094393 345.81107194 353.80965609]

THANK YOU!

Q&A

Appendix

RFE
Optimum number of features: 5
Score with 5 features: 0.804161
['Temp', 'Humidity', 'Wind', 'Pressure', 'Delay_min']



```
Add  Delay_min                      with p-value 0.0
Add  Pressure                       with p-value 5.13477e-25
Add  Wind                           with p-value 1.20916e-24
Add  Temp                           with p-value 2.37247e-09
Add  Humidity                       with p-value 3.34915e-07
```