Data Mining with Supervised Machine Learning

Analysis on Domestic Flight Delays

and

Predicting Arrival Delays

USA 2015

By:

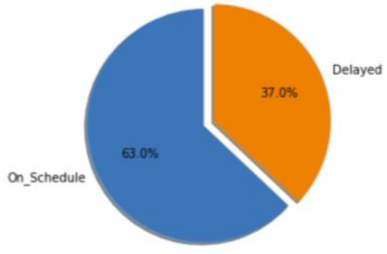
Swati Kohli

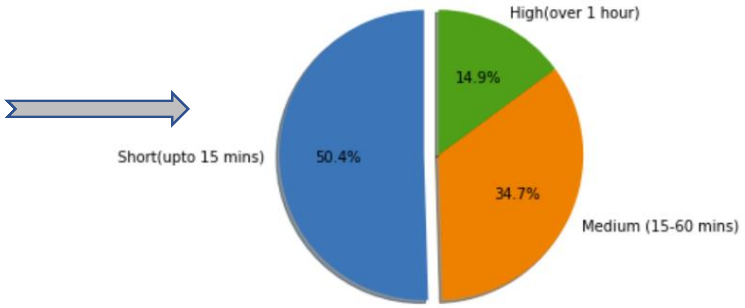
FLIGHT DELAY ANALYSIS

TABLE OF CONTENTS

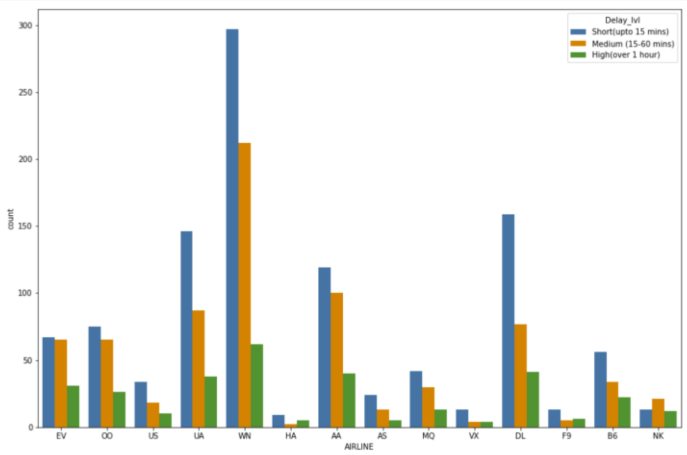
INTRODUCTION

Taking a flight for travellers has become an experience underlying uncertainty and anxiety pertaining to delays. As per statistical data of flights in 2015, more than one third of the total flights were delayed. A large majority of delay upto 15 minutes is visible in the bar graph below. However, this low value could be a consequence of most flights taking off on schedule. One third of the delays are between 15-60 minutes. Although 50% of the flights had short delays, regardless of the airlines, out of the total delays, 15% were delayed by more than an hour!





Observation: Pie Chart shows 37% flights delayed.



These delays affect both, the traveller and the airlines incurring cost and time, both of which are valuable commodities of today's fast life. Also, the reputation of airlines is at stake because delay is amongst the top key performance indicator. The delay could be caused by various reasons, anticipated or unforeseen, such as weather conditions, etc. However, some of the reasons can be dealt by taking informed decision based on a statistical approach. Therefore, analysis on flight delays is effective in understanding of flight performance. This study serves as a tool for largely three types of audience- Passengers, airline industry stakeholders and data analysts.

OBJECTIVE

This paper evaluates through a data-driven approach, the effects of various attributes of a flight lifecycle (air journey from point A to B) towards the performance (on-time or delay). These attributes are related to airlines, airports, timing/schedule, distance, flying process(taxi, wheel out, air time etc) and reasons of delay. Do some of these variables have a significant effect on delay? If so, which ones? Can variables be identified which can predict delays? Since the study is data-backed, this summary is useful for both travellers and airlines to improve their decision process vis a vis flight time, expenses, marketing strategy etc. Multiple regression method is used to investigate with a particular level of accuracy, which parameters are related significantly and to what extent they are influential contributors to the valuation of delay. In other words, the purpose is to present an equation (as best possible) to predict the arrival delay through relevant quantitative and qualitative input variables of flight delay.

Different statistical visualisations are created to observe the trends about the delays for coherent understanding.

DATASET & ATTRIBUTES

The analysis is presented from flight delay data of the United States in 2015 comprising various factors responsible and associated with flight delays for different airlines. There are 31 such attributes in the dataset explained below. The dataset contains records for 5821 Flights across 14 airlines.

- **YEAR, MONTH, DAY, DAY_OF_WEEK**: dates of the flight
- **AIRLINE**: An identification number assigned to identify a unique airline
- **FLIGHT_NUMBER, TAIL_NUMBER**: Flight and Aircraft identifier
- **ORIGIN_AIRPORT** and **DESTINATION_AIRPORT**: code attributed by IATA to identify the airports
- **SCHEDULED_DEPARTURE** , **SCHEDULED_ARRIVAL and SHEDULED_TIME** : scheduled times of take-off and landing and the planned time for trip.
- **DEPARTURE_TIME** and **ARRIVAL_TIME**: real times at which take-off and landing took place
- **DEPARTURE_DELAY** and **ARRIVAL_DELAY**: difference (in minutes) between planned and real times
- **DISTANCE**: distance (in miles)
- **WHEELS_OFF** and **WHEELS_ON**: When Aircraft's wheels leave the ground and touch the ground
- **TAXI_OUT** and **TAXI_IN**: The time duration elapsed between departure from the origin airport gate and wheels off and later between wheels-on and gate arrival at the destination airport
- **ELAPSED_TIME:** AIR_TIME+TAXI_IN+TAXI_OUT
- **AIR_TIME:** The time duration between wheels_off and wheels_on time
- **DIVERTED and CANCELLED:** Aircraft landed on airport that out of schedule and Flight Cancelled (1 = cancelled)
- **CANCELLATION_REASON,AIR_SYSTEM_DELAY,SECURITY_DELAY,AIRLINE_DELAY, LATE_AIRCRAFT_DELAY AND WHEATHER_DELAY:** Flight cancellation and delay reasons

DATA CLEANING

There are some missing values pertaining to delay, times, delay reasons and cancellation reasons. The missing values w.r.t delay (departure and arrival) do not match because Departure Delay and Arrival delay are calculated based on recorded actual time & scheduled time respectively, i.e.

Departure Delay = Departure time - Scheduled Departure

Since, the data is not recorded for Departure time (91 records) & Arrival time (94), there are missing values for Departure Delay and Arrival delay.

The remaining 14 records for arrival delay (108-94 = 14), have missing values which are not calculated even though there are records for Arrival time and Arrival Departure. Exploring if diverted and canceled feature are reasons for these missing values, 93 flights were canceled, so there is no departure delay, and 15 were diverted so there is no arrival record. However, there are still 2 flights that were canceled but have a departure delay. This could be because they took off but landed at the origin airport due to some technical difficulty. However, for this study all the missing data (delays and weather delay reason) has been removed as they would affect the analysis except cancellation reason which does not have a direct influence on delay in this study.

Further, arrival delay outliers are removed to explain (in the best way possible) arrival delays with some significant predictors through an equation.

ANALYSIS FOR DELAYS – INTUITIVE & THROUGH REGRESSION

While there could be many ways to investigate why there are delays like analyzing particular routes, or airlines etc, we look at some intuitive relations with visualizations of how some of the metrics may or may not impact delays by doing exploratory and regression analysis. These are related to origin airport, day of the week, time of the day, distance of the trip, airlines, departure time and delay reasons.

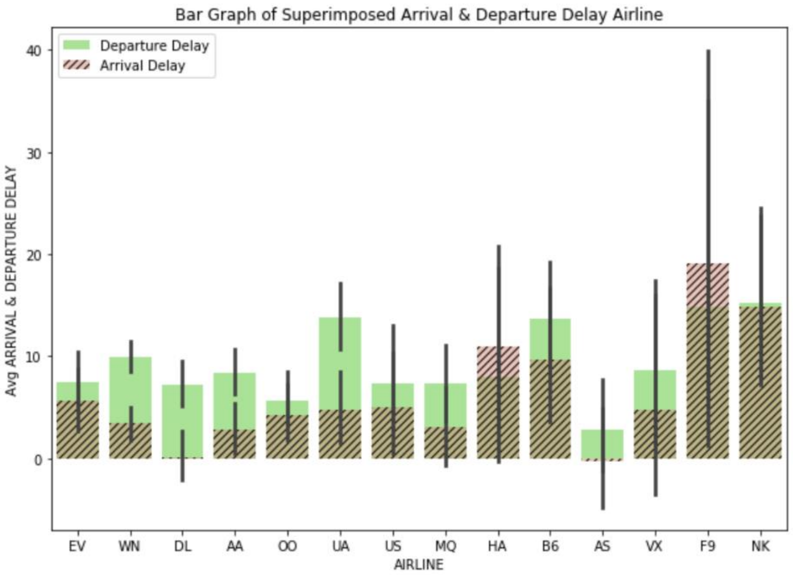
PART 1 - INTUITIVE RELATIONS - EXPLORATORY ANALYSIS

AVERAGE DEPARTURE AND ARRIVAL DELAYS

Understanding central distribution of data (mean, median, quartiles) gives a good start to analyse the dataset. Most airlines have minimal departure delays. NK (Spirit Air Lines) and F9 (Frontier Airlines) show more variability/some extreme values. Arrival delay variability seems similar with few outliers. Again, Frontier Airlines has the most extreme arrival delay.

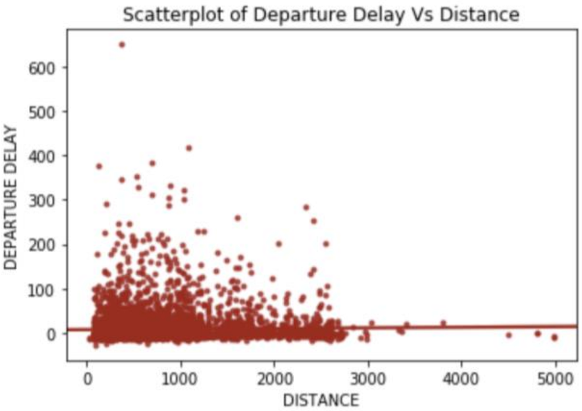
Median values reflect that flights actually depart and arrive early, approximately at 2 & 5 minutes respectively. However, it is inconclusive as the median data is robust and only dependent on the number of records. Further, it is observed that on an average departure delays are around 9 minutes. However, arrival delays are less than half of it, on an average almost 4 minutes. So there is little longer departure delay compared to arrival delays probably because airplane fly faster to make up for the delay. Exploring this result to diagnose if this is the case for majority airlines, the graph below shows visualization of the average departure and arrival delay vs Airlines.

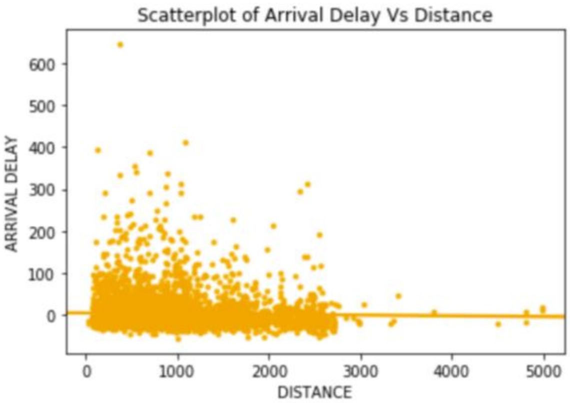
The superimposed bar graph shows a similar trend of average departure delays predominantly (10 out of 14) under 10 minutes. Also, the delays at arrival are generally lower than at departure. Intuitively, it seems quite logical that aircraft increase their speed to minimize delays and arrival delays are lesser as a consequence of greater distance. In other words, longer flights (bigger distance) makeup better for the departure delays. However, no conclusion can be drawn yet from mean or median values without diagnosis.



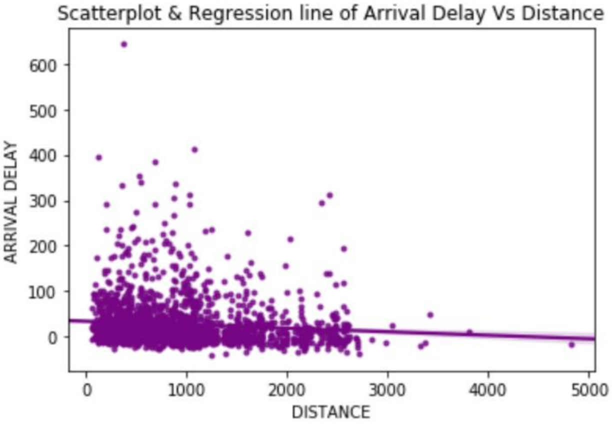
IMPACT OF DISTANCE ON DELAYS

To check the hypothesis above, the behaviour of departure and arrival delay is analyzed as a function of distance. The Correlation matrix shows a very low correlation (under 0.03) of distance with both departure and arrival delays.





Scatter plot shows no particular pattern or trend of distance with departure or arrival delay. If we plot a regression line, it is almost parallel to X axis which implies predictor Distance has no implication on Departure or Arrival Delay.

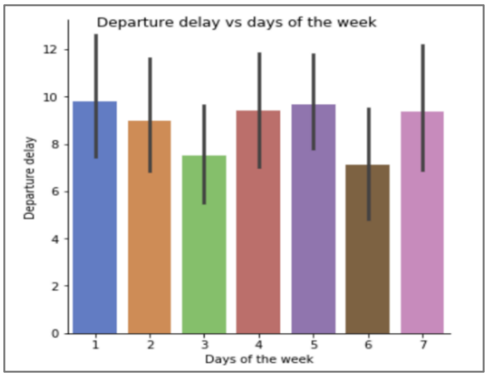


Diving deeper, if only positive departure delays are considered (since the dataset indicates early departure or arrival as delays with a negative value) , the results between distance and arrival delays are still inconclusive. Therefore, visualizations and statistical analysis through correlation, regression and p-value shows inconsequential relation between delays and distance.

IMPACT OF ORIGIN AIRPORT ON DEPARTURE DELAYS

Another aspect could be the origin airport influencing the delays. The study revealed FAR (Hector International Airport at North Dakota) and 12898 airport as the top two departure delay airports. The reason could be any, for example, bad weather of the state or the most frequented airline has a delay tendency etc. However, on further exploration it was found that both these airports had only one record for the delay which was due to air system and aircraft delay which indicates an exceptional case. Further study could be done in conjunction with routes or airlines to understand if other airports have any influence on departure delays.

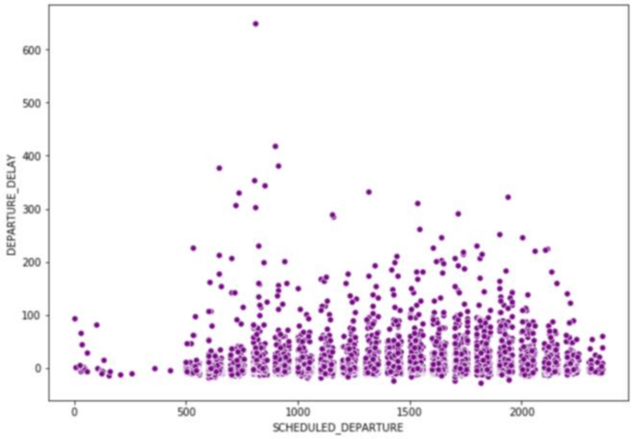
IMPACT OF DAY OF THE WEEK ON DEPARTURE DELAYS

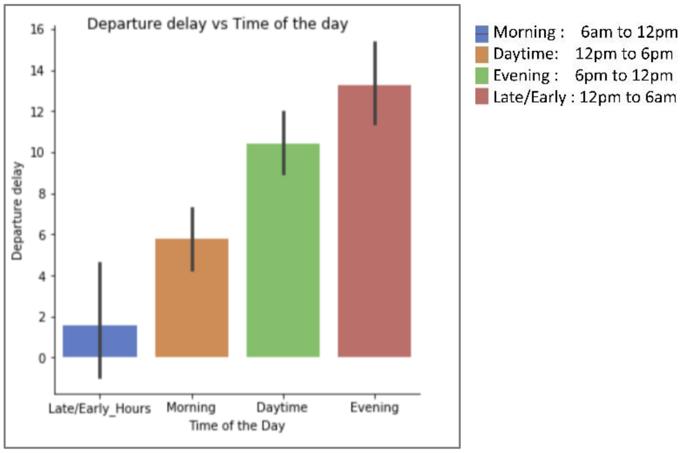


Numerical summary shows Wednesdays and Saturdays have a lower average departure delay, while the rest of the week is quite consistent. Boxplot shows no particular trend because of outliers, therefore, departure delays does not have any relation with the days of the week. It is also seen in the correlation matrix where the value is extremely low (-0.004) between the two. Days of the week are assigned numbers currently. Therefore, correlation matrix might give misleading results. Categorical variables can be formed to check if the results still seem insignificant or not.

IMPACT OF TIME OF THE DAY ON DEPARTURE DELAYS

Temporal variability could reveal some insight into departure delays. On analysis if any trend is observed, further study can be done to explore this aspect.





The scatter and bar-graph shows an interesting insight that average departure delay tends to increase with time of the day! Graph shows minimal delays from 12pm to 6am. In other words, flights usually leave on time in early morning or very late at night and the delay keeps growing upto 15minutes progressively till evening. Travellers can consider this insight to plan their flights based around this time. This insight suggests that time of the day (departure time) could be an important variable of prediction in the modelling of arrival delays.

PART 2 REGRESSION ANALYSIS - MODELING ARRIVAL DELAY

Further analysis is done with the insights from exploratory analysis how arrival delay can be best explained by certain predictors. For this task, missing data from weather delays is removed (refer Cleaning Data).

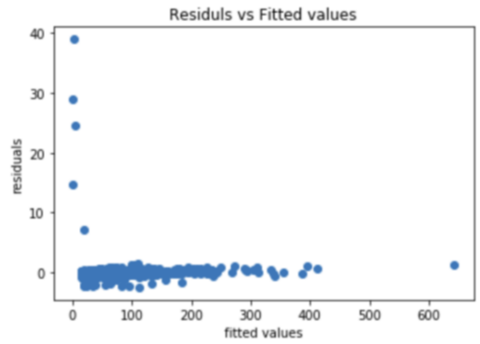
STEP 1

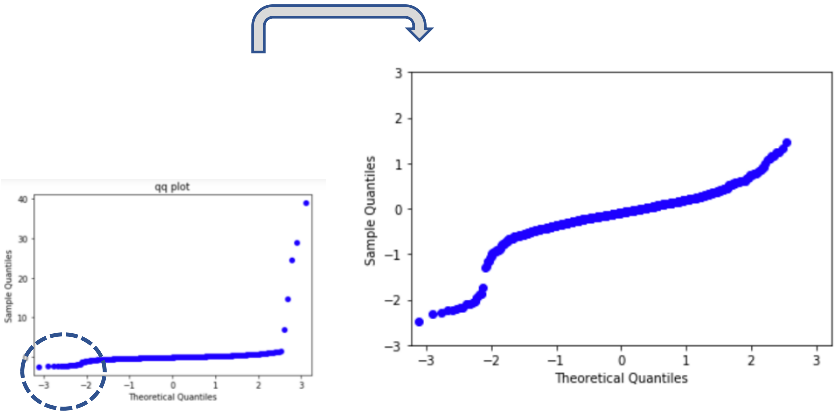
Firstly, The regression results using stats model show LATE_AIRCRAFT_DELAY, AIRLINE_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DEPARTURE_DELAY and only one airline (AS) have a significance in predicting arrival delay checked on 5% significance of p-value.

INTERPRETATION

ARRIVAL_DELAY = 0.55 + 0.98 **LATE_AIRCRAFT_DELAY** + 0.98 **LATE_AIRLINE_DELAY** + 0.98 **AIR_SYSTEM_DELAY** + 0.98 **WEATHER_DELAY** + 0.01 **DEPARTURE_DELAY** + 1.8 **AS**

- If aircraft is delayed by 10 minutes, keeping all other variables constant, on an average, the flight arrives late by 9.8 minutes.
- Departure delay of 10 minutes would result in arrival delay by 0.1 minute or 6 seconds, keeping all other variables constant.
- Coefficients reveal that the four types of delays have similar contribution to the arrival delay. Closer look at the data set shows most of these delays have value 0, meaning that in most cases they don't play a role in estimating arrival delay.
  The remaining numerical varibles are almost all nonsignificant except for departure delay.
- Focusing on the sign of the coefficients, it shows that the earlier the flight is, the less arrival delay there is, which aligs with reality, as early flights are usually on time, while flights later in the day suffers from the delays stacked up from before. One interesting observation though, is that distance has a positive effect on arrival delay.

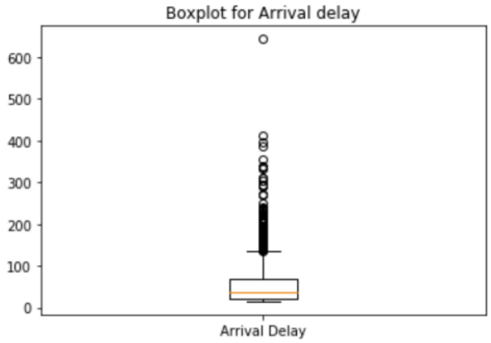




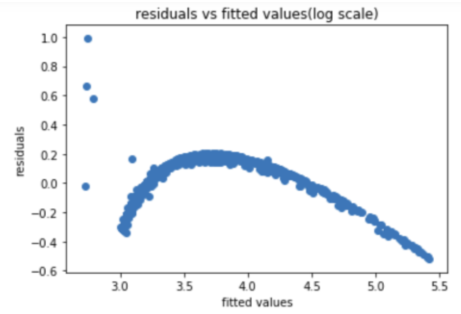
Firstly, there are quite a few outliers in the data set which are affecting the visualization. Removing outliers will reveal how good the model fit is. Checking assumptions, the residual vs fitted values plot shows slight curvature that error terms does not follow homoskedasticity assumption of linearity. The $r2$ is very high showing goodness-of-fit but that's not enough to say the model is suitable. Further, QQ plot shows the error terms deviate from normality assumption.

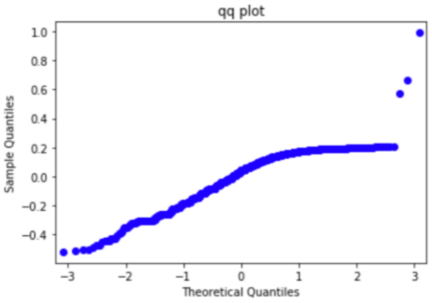
STEP 2

Outliers are observed and model is re-run for diagnosis by removing them from arrival delay and refitting the regression line by applying transformation as log of arrival delay and removing airline parameter.

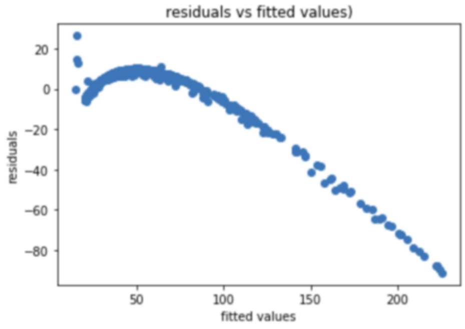


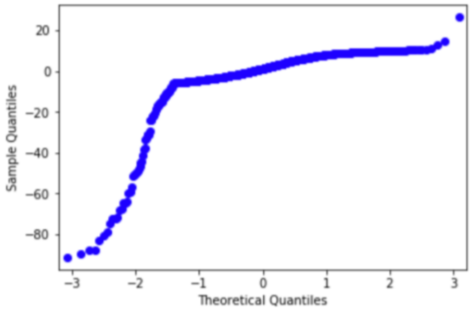
There is improvement in QQ plot which shows normality of error terms. However, it is still not enough to conclude normality assumption of linearity. In fact, we can see there is a strong nonlinear relationship between the inputs and output, as the residual plot shows a curvature pattern.





Converting log scale to normal scale, the graphs shows following results.





The R-square value reduced from 0.99 to 0.917 in the new model. This is because of reduction in the number of predictor variables and is not necessarily bad for the analytics considering overfitting.

Also, skew is reduced to -0.367 from 16.618.

Log(ARRIVAL_DELAY) = 2.707 + 0.019 **LATE_AIRCRAFT_DELAY** + 0.019 **LATE_AIRLINE_DELAY** + 0.0203 **AIR_SYSTEM_DELAY** + 0.019 **WEATHER_DELAY** + 0.0009 **DEPARTURE_DELAY**

CONCLUSION

As expected and is logical, departure delay and reasons of delay have significant influence on the arrival delays. However, these are still not conclusive and more study is required to model arrival delays better.

- To improve the model further, stepwise regression should be done to add or remove variables from the model. Adding interaction terms to the model could be explored as well.
- Data visualization shows arrival delay has some pattern with the time of the day. This can be included in model to check if the model performance improves further.
- Even though graph shows no relation between day of the week and departure delays, day of week variable is an ordinal variable and the correlation might be misleading. Hence it should be converted to categorical variable and its dummy variables can be added in the model to study the effect of new variable in the model.
- Indicator variables Cancelled and Diverted could be used to check their impact on arrival delay.
- Further, as described above, the graph shows nonlinearity between inputs and output. One might consider transforming some of the input variables (to polynomial terms) to solve this issue. For instance, adding a squared term might result in a better residual graph. As for the variables with too many zeros, there are several ad-hoc approaches one can take. For example, convert the variable into categorical variables, or perform some special transformations, or even use different models.