NYC Flights 13

Liam Frank

Last compiled on April 25, 2024

Contents

1 Introduction **3**

1.1 Background . . . 3

2 The Data **4**

2.1 Source . . . 4

2.2 Variables . . . 4

2.3 Observation . . . 5

3 Exploratory Data Analysis **6**

4 Methodology **16**

4.1 Types of Models . . . 16

4.2 Data Tranformations . . . 16

4.3 Model 1 Logistic Regression . . . 16

4.4 Model 2 K-Means Clustering . . . 18

5 Results **22**

5.1 Model 1 Logistic Regression . . . 22

5.2 Model 2 K-Means Clustering . . . 22

6 Discussion **23**

6.1 Final model interpolation . . . 23

6.2 Use of Model . . . 23

7 Furture Work **24**

8 References **25**

1 Introduction

Year after year more and more people choose to travel via air. Commercial aviation accounts for over 5% of annual GDP in The United States resulting in the operation of over 26,000 flights both foreign and domestic, carrying 2.6 million passengers daily (Airlines for America). According to the FAA, commercial aviation currently generates over 10,000,000 American jobs (FAA). The dataset chosen for this study is a part of the nycflights13 library in R. This dataset highlights all commercial flights departing from the three major airports in the vicinity of New York City, John F. Kennedy (JFK), LaGuardia (LGA), and Newark (EWR), in the calendar year of 2013. The airports highlighted in this study are amongst the busiest in The United States, JFK ranks 6th at 26.9 million passengers annually, EWR ranks 13th at 21.6 million passengers annually, and LGA ranks 19th at 14.4 million passengers annually (Baran). The library nycflights13 contains 5 separate tables that are linked together in a relational database schema. For this project, most all exploratory data analysis and model construction will be conducted using data found in the flights and weather tables, while other data tables will be referenced to help generate questions and draw potential conclusions from the exploratory data analysis process. This study aims to explore distributions, relationships, and employ classification and clustering techniques in regards to flight delay times.

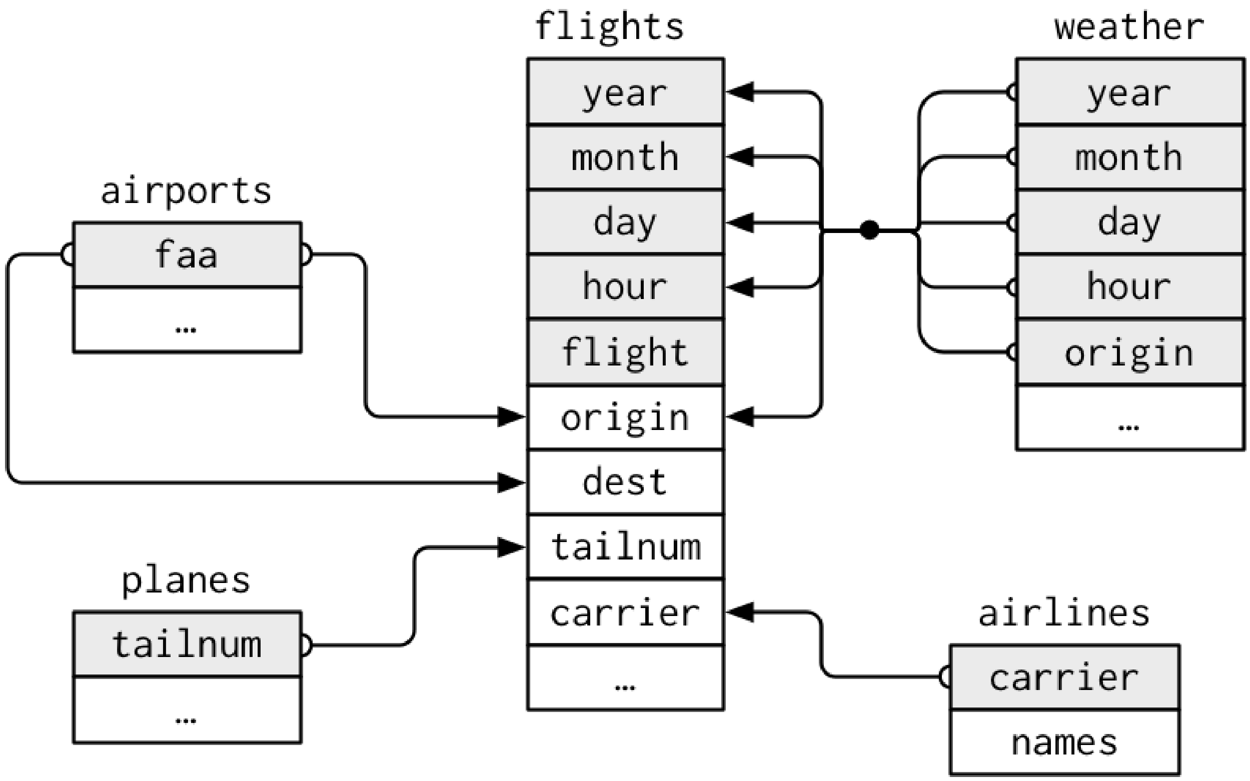
1.1 Background

With the rise in data-driven decision making, big data analytics has taken on a variety of different use cases in the commercial aviation industry, one of the most popular use cases being flight delay prediction. More than 20% of commercial flights experience an arrival delay of over 15 minutes, leading to both logistic and economic challenges for airlines and passengers alike. Previous studies have employed popular machine learning techniques such as regression, classification, and clustering with the goal of accurately predicting delay times. “Airline delay prediction by machine learning algorithms” used decision tree, cluster classification, and random forest to examine flight delays between US and Iranian airways. The study concluded that the most significant variables contributing to flight delay times are visibility, wind, and departure time. “Machine learning approach for flight departure delay prediction and analysis” utilized support vector machines to investigate patterns of delays at the three major New York airports. The study concluded that the most influential contributing factors to arrival delay include pushback delay, traffic volume, and weather. “Machine learning techniques for analysis of Egyptian flight delay” employed a variety of different decision trees to classify flight delays in Egyptian Airlines flight data. The accuracy of each model was compared with the highest accuracy percentage for a decision tree built being 83%.

2 The Data

2.1 Source

The data used comes from the R library nycflights13. This library is built upon a database schema as seen below. The flights table includes every flight from the calender year of 2013 that departed one of New York Cities three major airports, John F. Kennedy (JFK), LaGuardia (LGA), and Newark (EWR). The flights table contains flights of over 4,000 commercial aircraft flying to 105 unique destinations both foreign and domestic. The weather table includes Automated Weather Observation System (AWOS) data by the hour at each airport.



2.2 Variables

The inner_join() function was use to concatenate the flights and weather table.The data frame constructed contains 336,776 total observations and 29 variables. Each observation represents a flight departing from one of the three airports mentioned above.

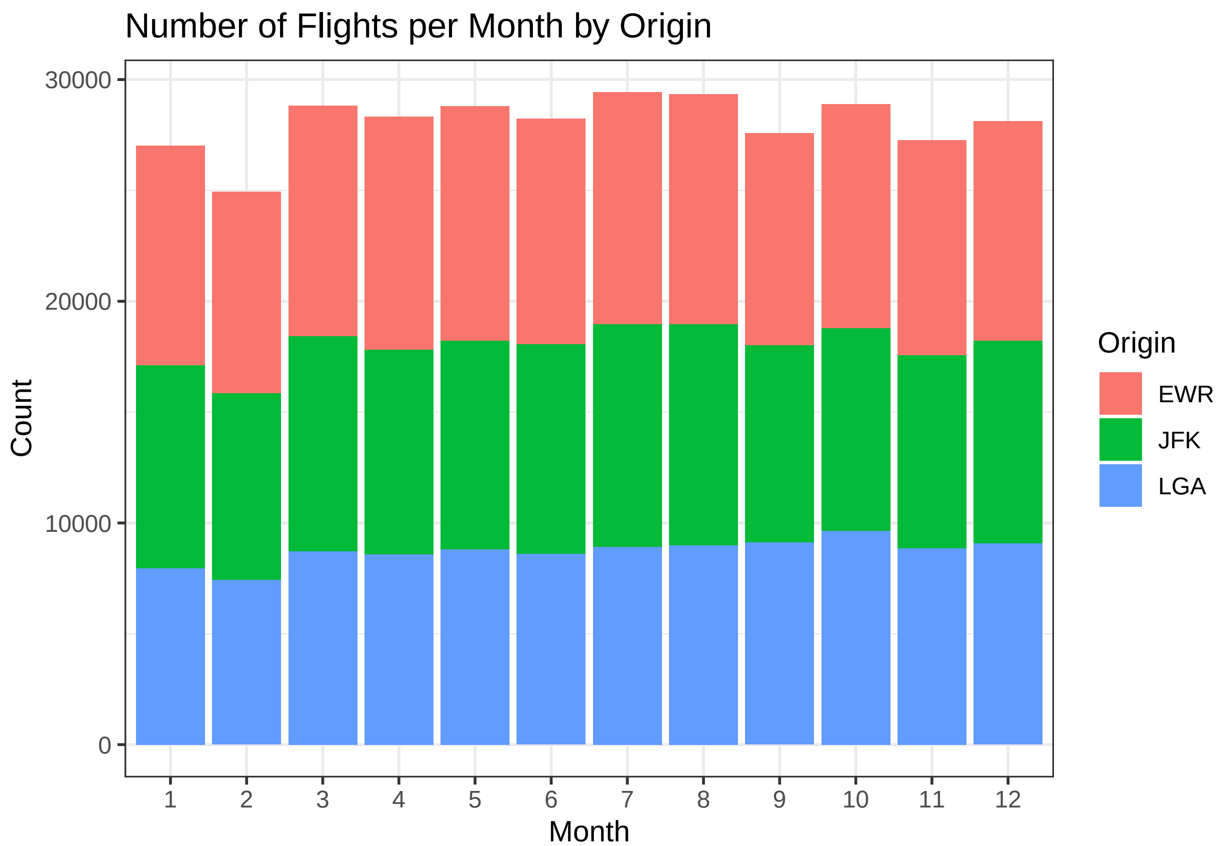
1) Year: Integer, being nycflights13 all observations are recorded as 2013
2) Month: Integer, 1 signifies January and so on to 12 for December. This will be converted to a factor as month should be treated as a categorical variable for this analysis
3) Day: Integer, day of the month that the recorded flight departed
4) Dep_time: Integer, departure time of flight recorded in 24-hour standard time
5) Sched_dep_time: Integer, scheduled departure time of flight recorded in 24-hour standard time
6) Dep_delay: Double precision, difference between scheduled departure time and actual departure time, positive value signifies a departure delay while a negative value signifies the flight left early
7) Arr_time: Integer, arrival time of flight recorded in 24-hour standard time
8) Sched_arr_time: Integer, scheduled arrival time of flight recorded in 24-hour standard time

9) Arr_delay: Double precision, difference between scheduled arrival time and actual arrival time, positive value signifies an arrival delay while a negative value signifies the flight arrived early
10) Carrier: Character, two letter abbreviation for the airline conducting the flight
11) Flight: Integer, three or four digit code that signifies the flight number
12) Tailnum: Character, the tail number of the aircraft that conducted the flight
13) Origin: Character, ICAO code for departure airport
14) Dest: Character, ICAO code for arrival airport
15) Air_time: Double precision, time in minutes between departure and arrival
16) Distance: Double precision, distance between departure and arrival airport in statute miles
17) Hour: Double precision, hour of scheduled departure time in 24-hour standard time
18) Minute: Double precision, minute of scheduled departure time in 24-hour standard time
19) Time_hour: Date-time, date and hour of scheduled departure
20) Made: Integer, amount of time a flight makes up over estimated time in the air. The formula for deriving "Made" was departure delay - arrival delay. A negative value indicates the flight lost time in the air, while a positive value indicates the flight made up time
21) Temp: Numeric, ambient air temperature in degrees Fahrenheit at time of departure
22) Dewp: Numeric, dew point at time of departure
23) Humid: Numeric, humidity at time of departure
24) Wind_dir: Integer, wind direction as a heading fix at time of departure
25) Wind_speed: Numeric, wind speed in statute miles per hour at time of departure
26) Wind_gust: Numeric, wind gusts in statute mile per hour at time of departure
27) Precip: Numeric, precipitation rate per hour in inches at time of departure
28) Pressure: Numeric, ambient air pressure in millibars at time of departure
29) Visib: Integer, visibility in statute miles at time of departure

2.3 Observation

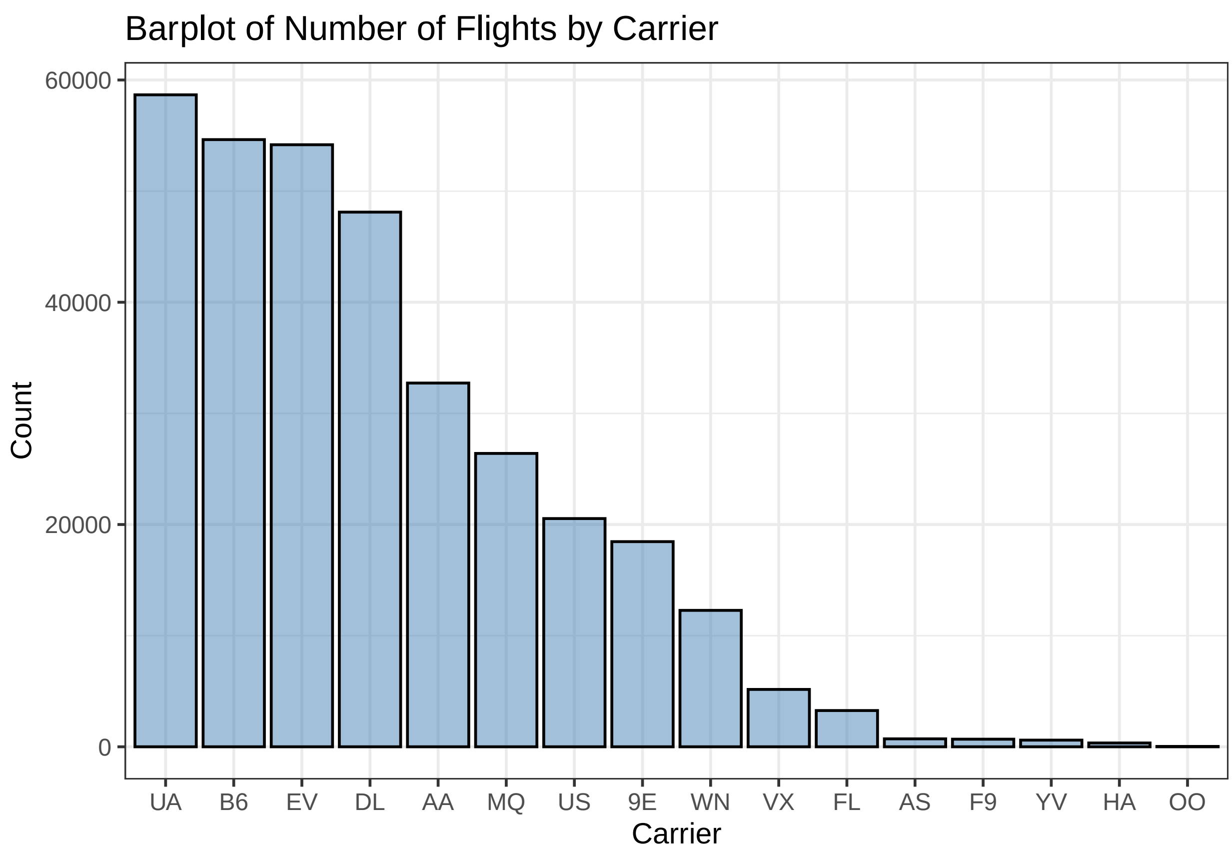
```
##                  year              month                day
##                "2013"               " 1"               " 1"
##              dep_time     sched_dep_time          dep_delay
##                " 517"             " 515"             "   2"
##              arr_time     sched_arr_time          arr_delay
##                " 830"             " 819"             "  11"
##               carrier             flight            tailnum
##                  "UA"             "1545"           "N14228"
##                origin               dest           air_time
##                 "EWR"              "IAH"              "227"
##              distance               hour             minute
##                "1400"               " 5"               "15"
##             time_hour               made               temp
## "2013-01-01 05:00:00"             "  -9"           " 39.02"
##                  dewp              humid           wind_dir
##               "28.04"           " 64.43"              "260"
##            wind_speed          wind_gust             precip
##            "12.65858"                 NA             "0.00"
##              pressure              visib
##              "1011.9"            "10.00"
```

3 Exploratory Data Analysis



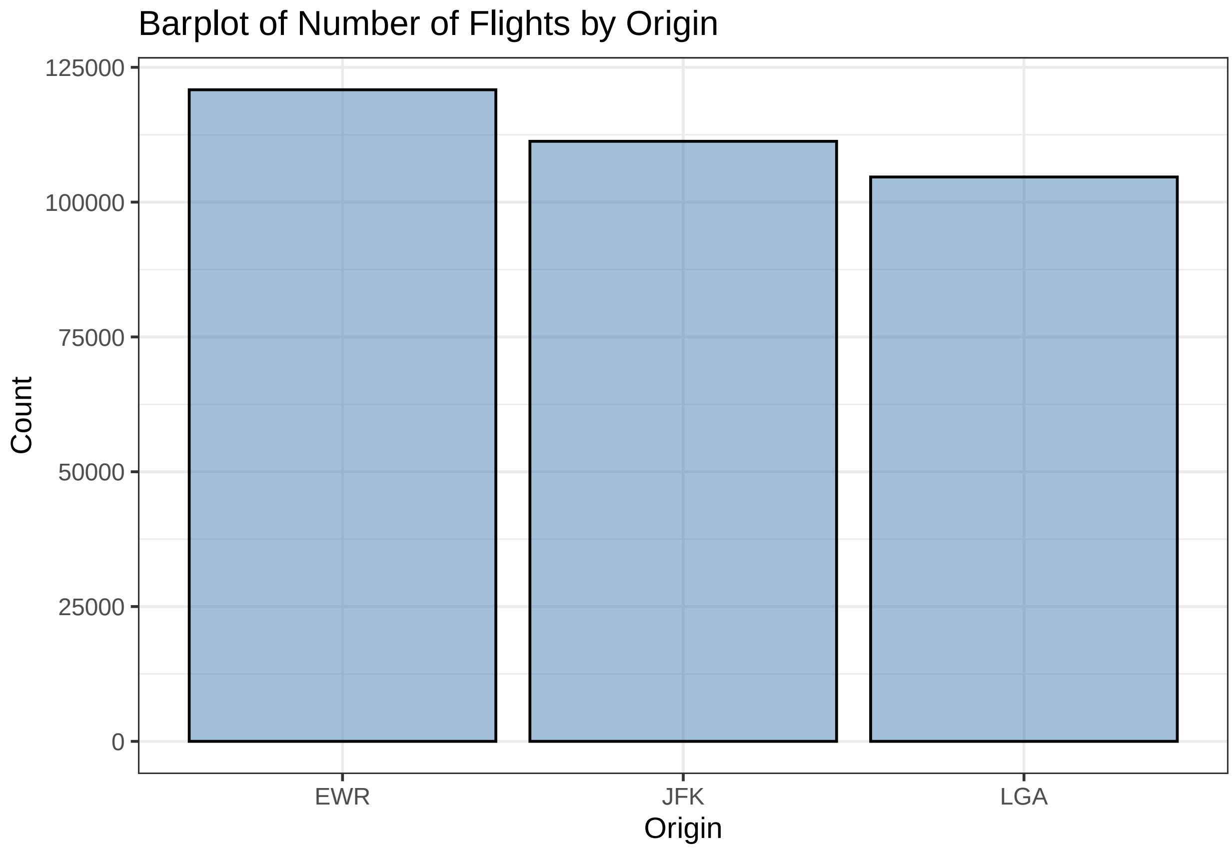

```
##       as.factor(month)
## origin     1     2     3     4     5     6     7     8     9    10    11    12
##    EWR  9893  9107 10420 10531 10592 10175 10475 10359  9550 10104  9707  9922
##    JFK  9161  8421  9697  9218  9397  9472 10023  9983  8908  9143  8710  9146
##    LGA  7950  7423  8717  8581  8807  8596  8927  8985  9116  9642  8851  9067
```

Through the bar plot it is evident that their is a increase in number of flight operations in both the summer months and winter months. The trend of increase and decrease in total operations over the months is standard across all three airports.



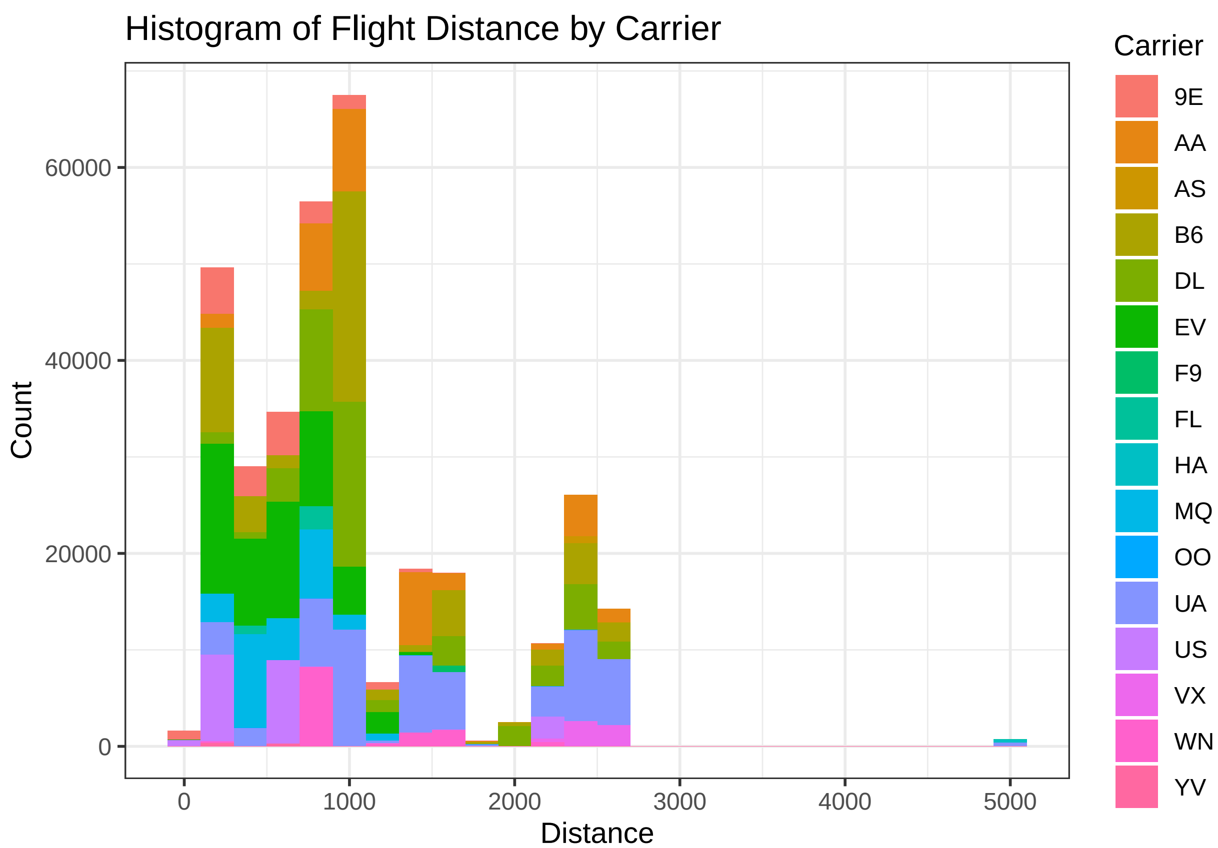

```
## carrier
##    9E    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO    UA    US
## 18460 32729   714 54635 48110 54173   685  3260   342 26397    32 58665 20536
##    VX    WN    YV
##  5162 12275   601
```

The bar plot above showcases the total number of fights conducted in 2013 by carrier. The top 5 airlines operating the most flight out of the three major airports in the New York City vicinity are United Airlines (UA), JetBlue (B6), ExpressJet (EV), Delta (DL), and American (AA). All five of these airlines use either JFK, or EWR as a hub.

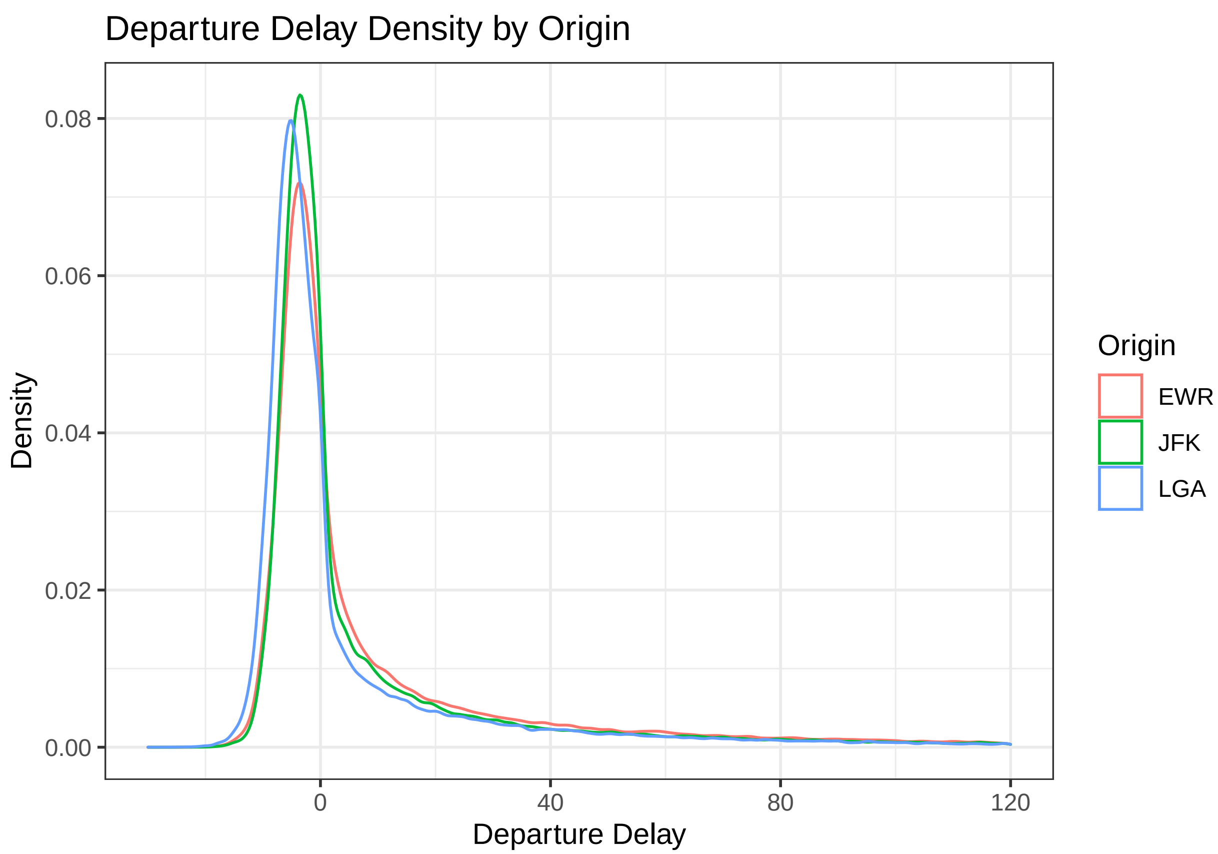



```
## origin
##    EWR    JFK    LGA
## 120835 111279 104662
```

Newark operated the most flights in 2013, with Kennedy and LaGuardia following closely.

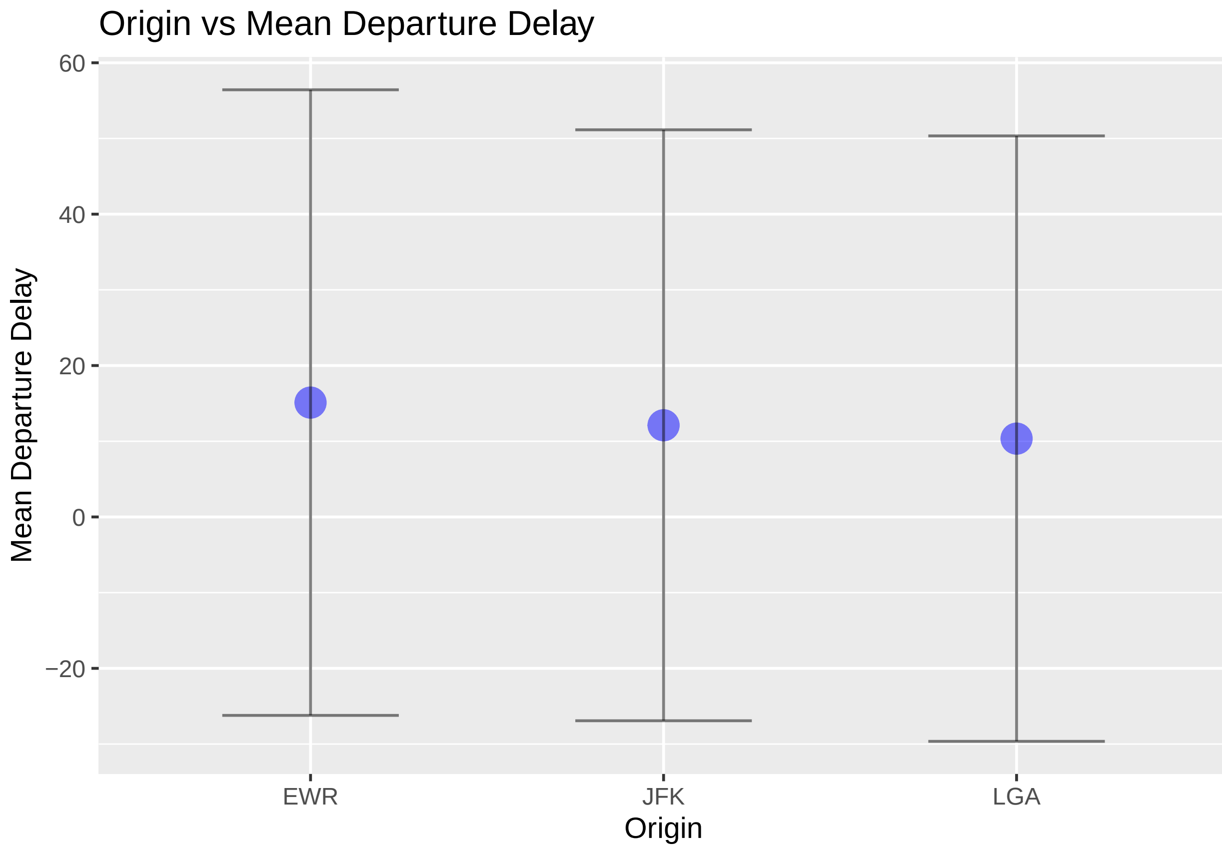


The histogram above reveals that the largest concentration of flight destinations fall around 1,000 miles of New York City. The outlier seen at 5,000 miles is a direct flight operated by Hawaiian Airlines from Kennedy to Honolulu.

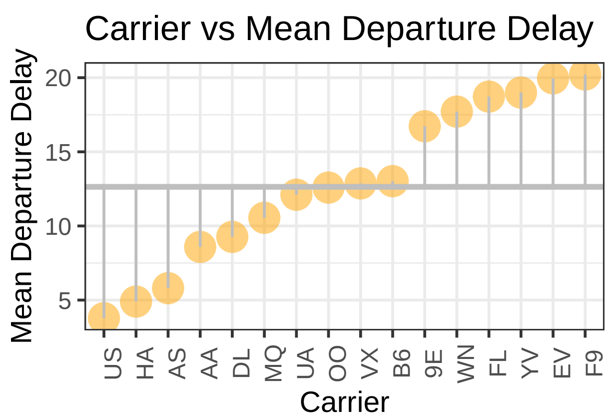


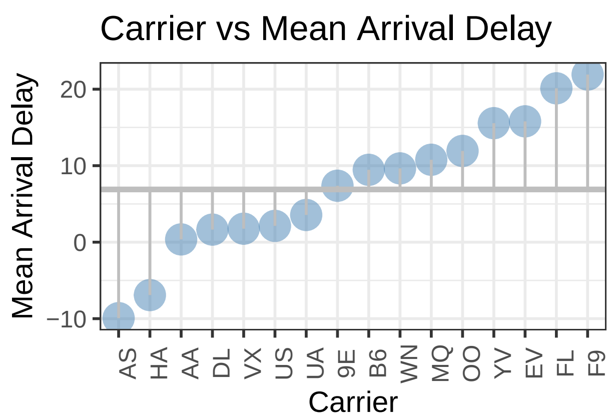

```
## flights$origin: EWR
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -25.00   -4.00   -1.00   15.11   15.00 1126.00    3239
## --------------------------------------------------------------
## flights$origin: JFK
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -43.00   -5.00   -1.00   12.11   10.00 1301.00    1863
## --------------------------------------------------------------
## flights$origin: LGA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -33.00   -6.00   -3.00   10.35    7.00  911.00    3153
```

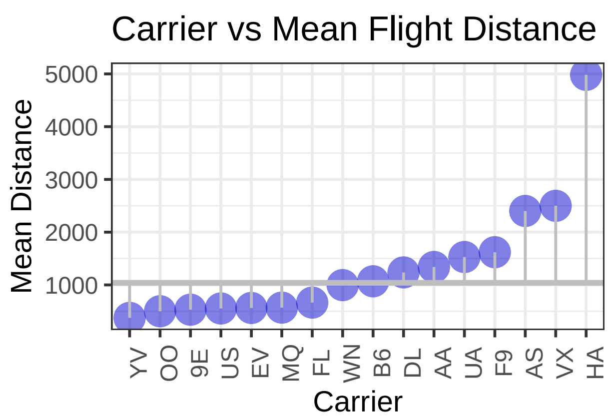
The departure delay density by origin is fairly similar but it does appear the LGA has the highest density of flights leaving early and EWR experiencing a greater density of delays. Examining summary statistics we can see this is true. Newark (EWR) had a mean departure delay time of 15.11 minutes, Kennedy (JFK) had a mean departure delay time of 12.11 minutes, and LaGuardia (LGA) had a mean departure delay time of 10.35 minutes.

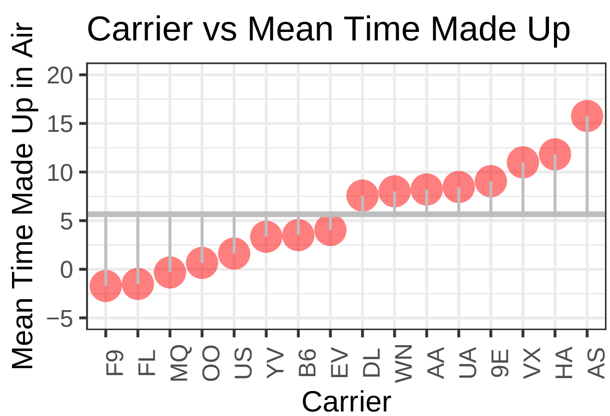


As mentioned earlier, their was a difference in mean departure delay times by origin. But, the error bar plot above confirms that the difference is not statistically significant.

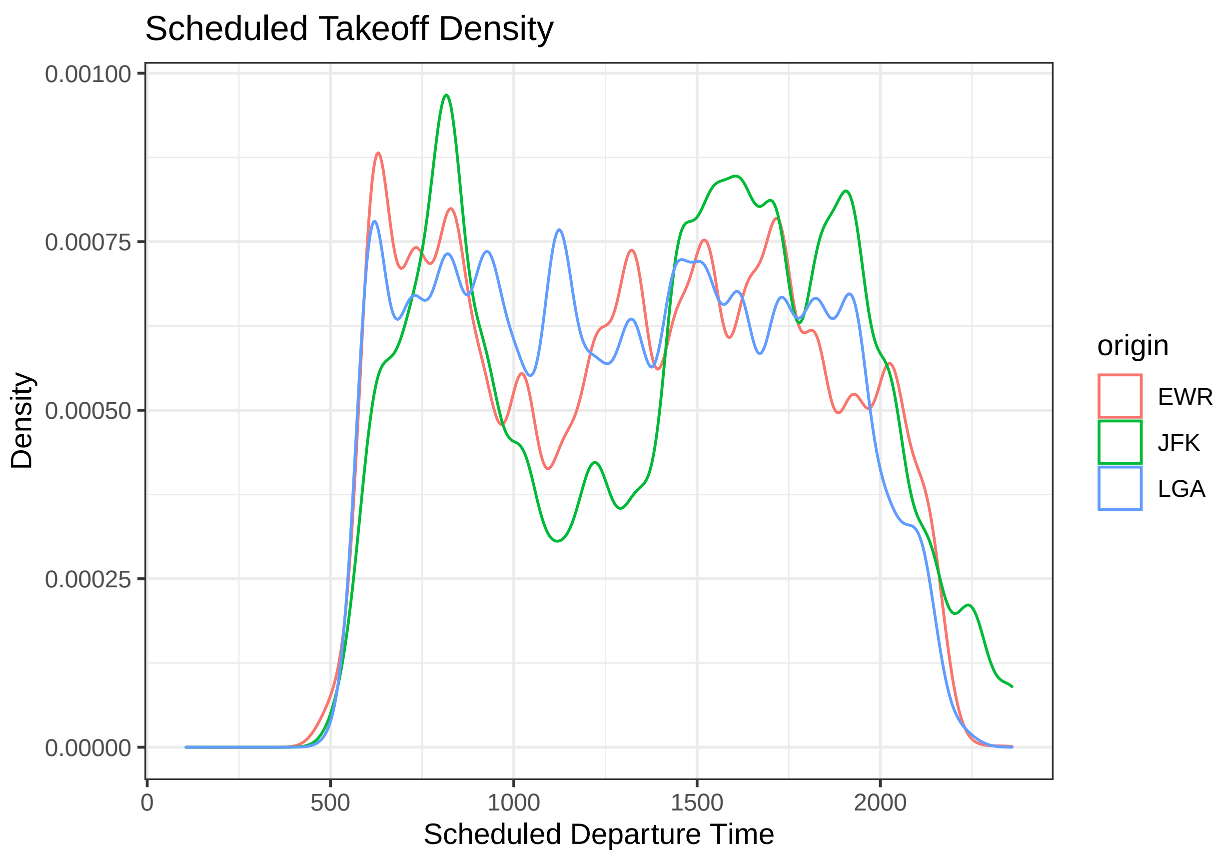




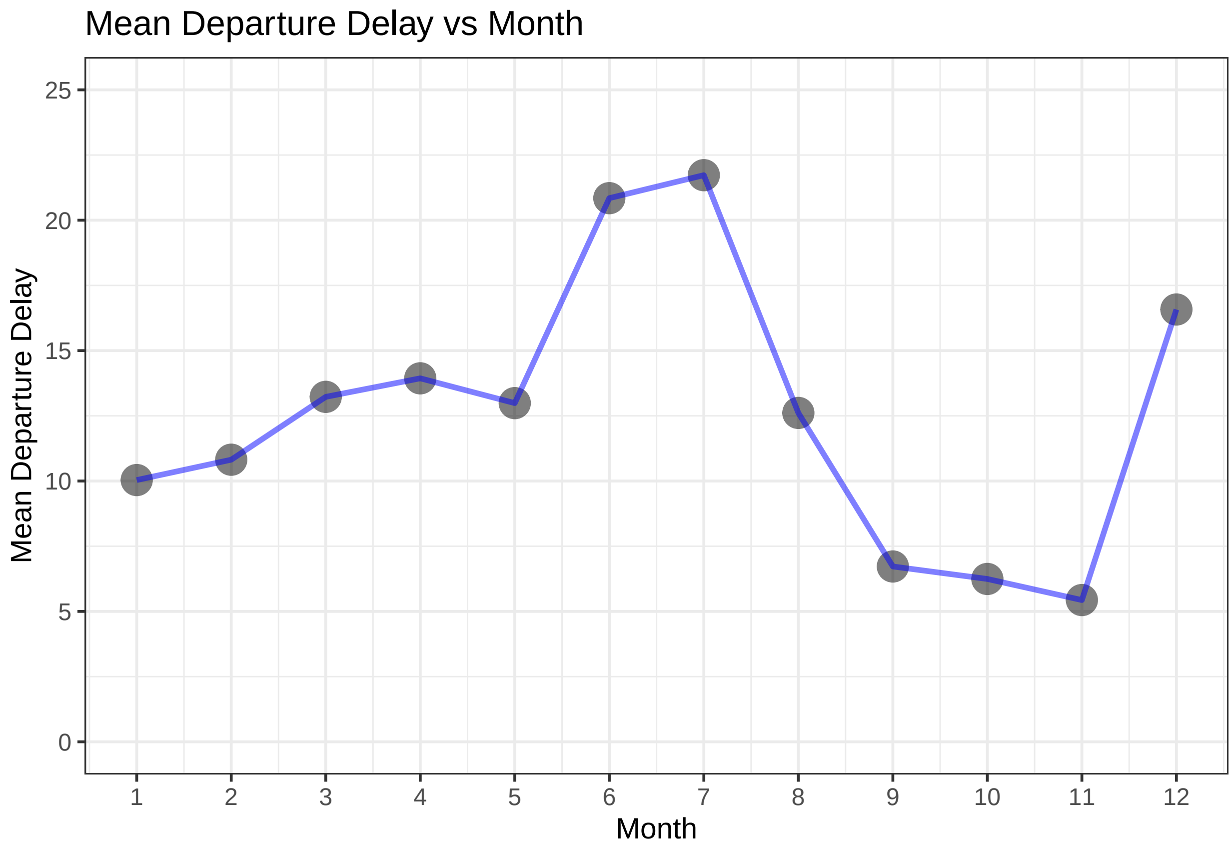




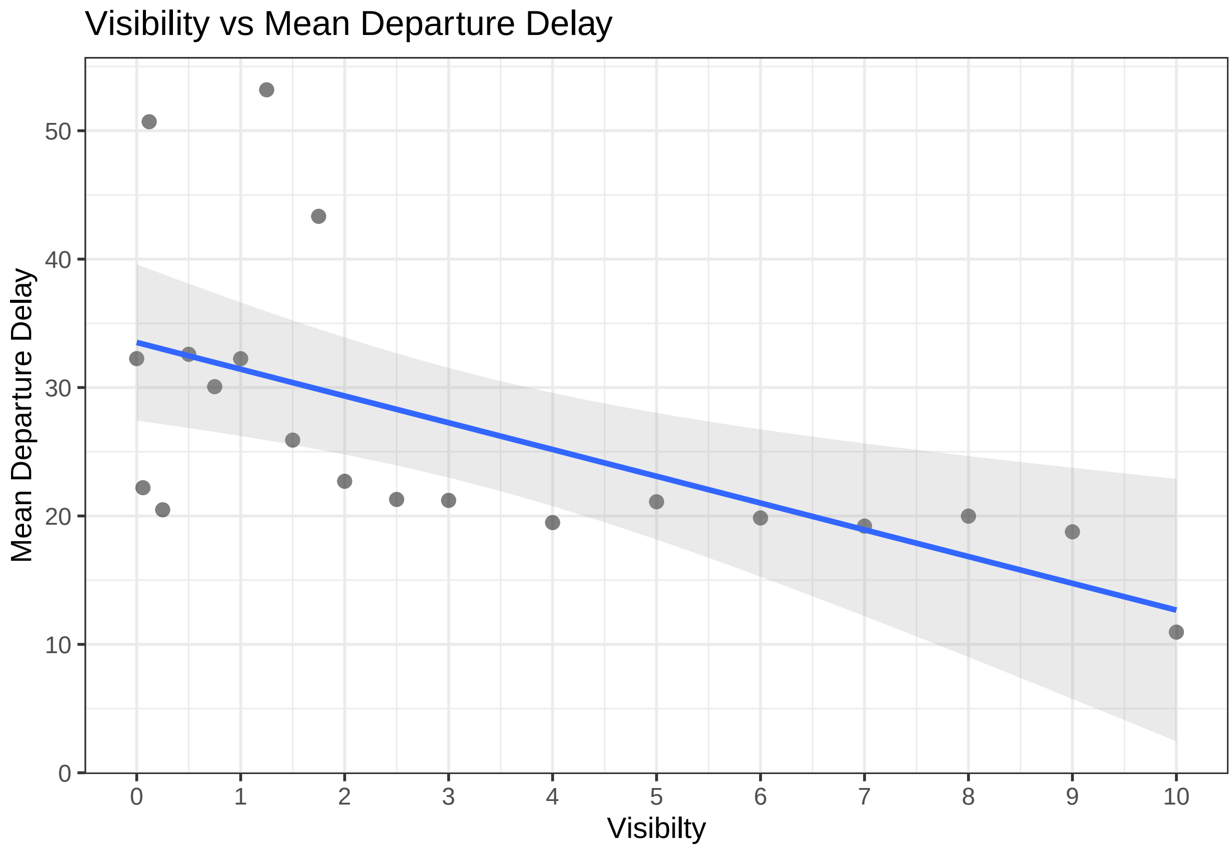
Examining each of the lollipop graphs above their are a few initial key takeaways. First, fairly straight forward and expected, a low average departure delay for a given airline typically leads to a low average arrival delay. Second, a trend is established between flight distance and time made up in the air. The airlines with higher mean flight distances also experience a greater makeup of time in the air. This trend does have an outlier though, Frontier (F9). Despite having one of the longest average flight distances, on average they lose time in the air. Frontier also experiences the highest average departure and arrival delay. Contrary to Frontier, Alaska (AS) has the earliest mean arrival times, and the greatest time made up in the air on average. Like Frontier, Alaska is also in the top 4 airlines for longest mean flight distance. Similar to Alaska (AS) is Hawaiian (HA). Hawaiian Airlines has one of the lowest average departure and arrival delay times, with the longest mean flight distance, and second in mean time made up in the air. The relationships shown in the lollipop graphs tend to signify a strong correlation between distance and time made up in the air. As well as time made up in the air and arrival delay.



The density plot above highlights the density of departing flights from the three airports. JFK operates the most international flights of the three airports. This is seen through the density plot with the greatest density of flights leaving either morning or late evening. EWR operates a more equal balance of domestic and international flights, so the mid-day dip in departure density is less pronounced. LGA operates mostly domestic flights, in turn there density plot for departures is much flatter signifying a steadier flow of traffic throughout the day.



The line plot for mean departure delay by month is rather telling. It follow the same trend seen above pertaining to number of flights per month.



Mean departure delay for each factor of visibility was plot. The overall relationship is rather linear in nature indicating that as visibility worsens, mean flight departure delay increases my a measurable amount.

4 Methodology

4.1 Types of Models

Both Logistic Regression and K-Means Clustering were techniques employed in this study. Logistic regression is a statistical method commonly used for binary classification tasks, where the outcome variable is categorical with two levels (e.g., yes/no, 0/1). It models the relationship between one or more independent variables and the probability of the outcome occurring. The logistic regression model applies the logistic function, also known as the sigmoid function to the linear combination of the independent variables, transforming the output into a probability between 0 and 1. By setting a threshold or optimal cutoff, logistic regression classifies observations into one of the two categories.

On the other hand, k-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into distinct clusters based on similarity. It aims to group observations into k clusters, where each observation belongs to the cluster with the nearest centroid. The algorithm iteratively assigns observations to the nearest centroid and updates the centroids based on the mean of the assigned observations. K-means clustering works by minimizing the within-cluster sum of squares, seeking to minimize the variance within clusters and maximize the variance between clusters.

4.2 Data Tranformations

To test the classification accuracy of the logistic model, the data was split into a test and train set using the standard 70/30 split. A subset of the original data set was comprised of only continuous variables was created for k-means clustering as it is a distance based algorithm not applicable to categorical variables.

4.3 Model 1 Logistic Regression

A logistic regression model was created to classify observations based on a derived variable of significant arrival delay. Significant arrival delay, denoted as sig_arr_delay is a binary factor with a 1 corresponding to a flight arriving at its destination greater than 7 minutes late, and a 0 corresponds to a flight being less than 7 minutes late. The mean arrival delay for all observations is 6.88 minutes, that is why the value of 7 was decided on for the split of the data into a binary factor suitable for logistic regression.

```
##
## Call:
## glm(formula = sig_arr_delay ~ air_time + made + dep_time + distance +
##     wind_speed + visib, family = "binomial", data = Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3508  -0.7088  -0.4184   0.5909   4.2693
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -1.598e+00  3.279e-02  -48.734  < 2e-16 ***
## air_time    -2.158e-03  5.107e-04   -4.226 2.38e-05 ***
## made        -8.733e-02  4.919e-04 -177.527  < 2e-16 ***
## dep_time     1.452e-03  1.211e-05  119.870  < 2e-16 ***
## distance     3.277e-04  6.591e-05    4.972 6.63e-07 ***
## wind_speed   1.331e-02  9.899e-04   13.444  < 2e-16 ***
## visib       -1.163e-01  2.733e-03  -42.540  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282961  on 228016  degrees of freedom
## Residual deviance: 204546  on 228010  degrees of freedom
##   (56 observations deleted due to missingness)
## AIC: 204560
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 0.2771216
```

```
## [1] 0
```

As previously mentioned the response variable of significant arrival delay was used. All explanatory variables used are significant with the absolute value of the z-values being greater than 2 and the p-values being less than 0.05. The explanatory variables used were air time, time made up in the air, departure time, distance, wind speed, and visibility. The model had a McFadden's Pseudo R squared of .277 with p-value less than 0.05.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##   threshold
## 1 0.3578069
```

The probability optimal cutoff found for binary classification is 0.3578069.

```
##    labels sig
## 6       1   1
## 14      0   0
## 15      1   1
## 25      0   0
## 30      0   0
## 37      0   0
```

```
##
## labels     0     1
##      0 55185  9599
##      1 12134 20806
```

```
##
##     FALSE      TRUE
## 0.2223416 0.7774333
```

```
## [1] 0.8184465
```

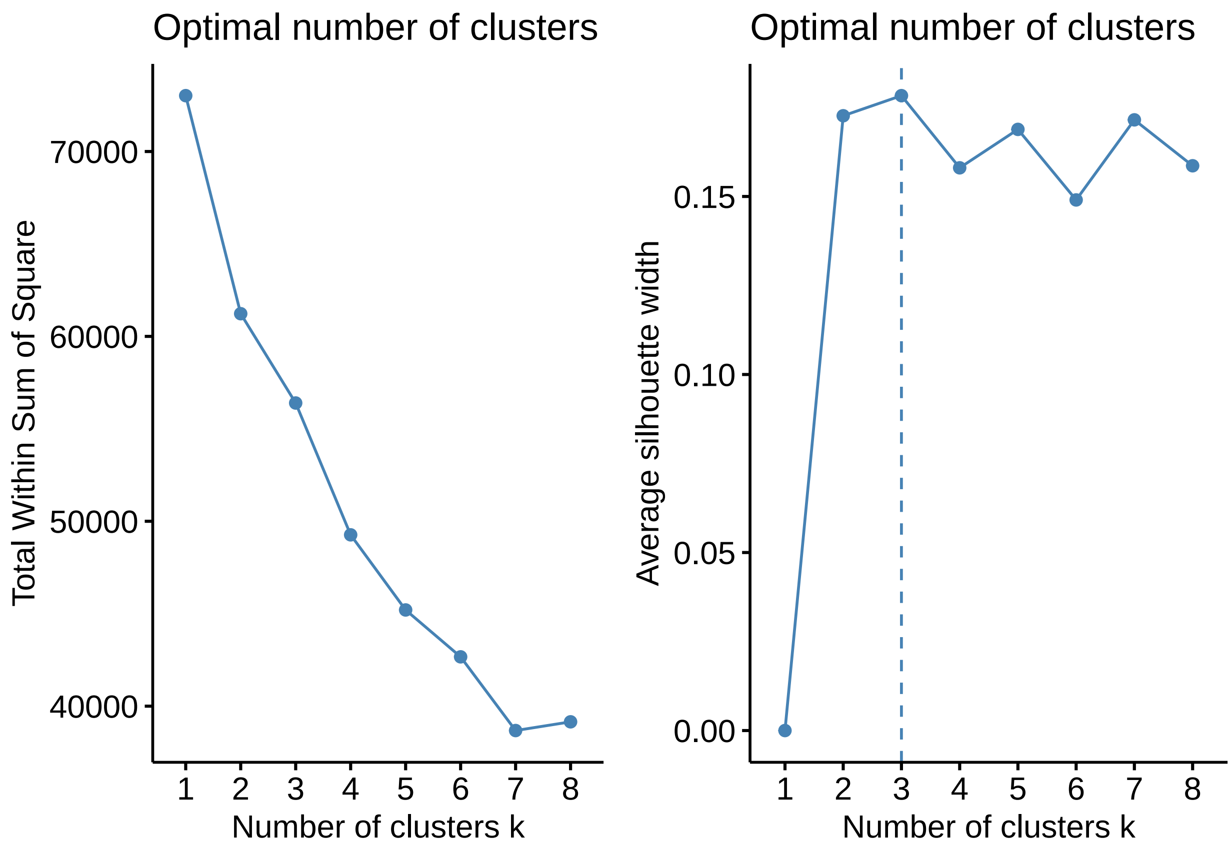
```
## [1] 0.6860056
```

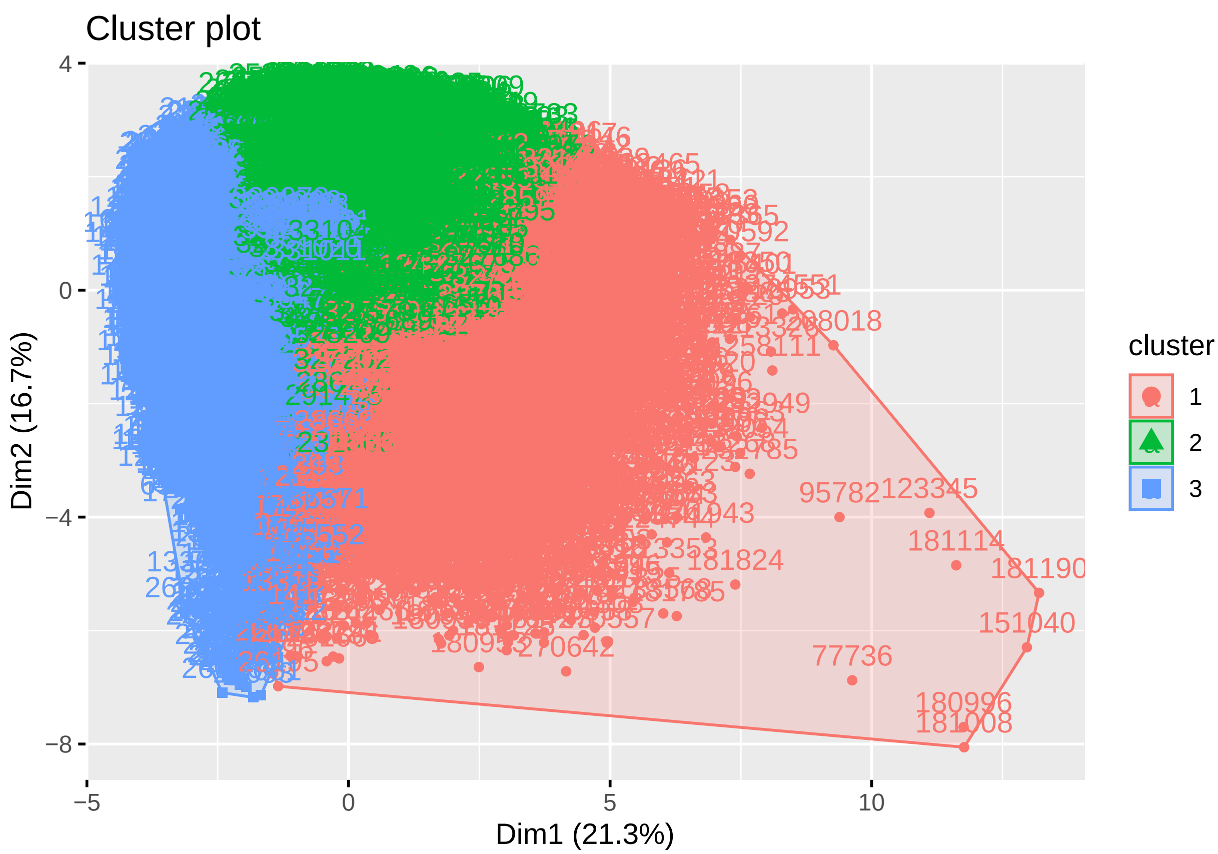
```
## [1] 0.22276
```

Examining the confusion matrix and calculating measures of accuracy, the model sensitivity or true positive rate was 82% and a specificity or true negative rate was 69%. The total mis-classification error rate was 22%. Resulting in a prediction accuracy of just shy of 80%.

4.4 Model 2 K-Means Clustering

K-means clustering was utilized in an attempt to group similar observations for pattern recognition of common characteristics. A subset of the nycflights data was created containing only numerical variables. This data was then scaled to insure no single feature has a greater influence in distance calculations. In order to combat computational limitations a wss elbow plot and silhouette plot to find the optimal number of clusters were created with a subset 25,000 observations. The k-means algorithm then employed all 325,000 observations.






```
## [1] 311682.5 254361.0 211602.6

## [1] 777646.1

## [1] 226057.9

## [1] 17445 34949 24815

##   cluster  dep_time sched_dep_time  dep_delay arr_time arr_delay air_time
## 1       1 1728.1583      1664.4999 49.5615363 1770.018 52.992147 148.6251
## 2       2 1583.0332      1577.2936  4.8835732 1784.105 -3.618387 149.9497
## 3       3  945.0491       955.3652  0.4918396 1174.504 -4.369293 148.4178
##    distance      made     temp wind_dir wind_speed wind_gust    visib cluster
## 1  997.7457 -3.430610 47.50902 252.3158   20.79241  30.02147 9.149586       1
## 2 1047.6204  8.501960 62.13778 246.6134   14.35711  22.35741 9.922029       2
## 3 1010.7296  4.861132 45.78576 253.8420   17.28911  25.71672 9.624746       3
```

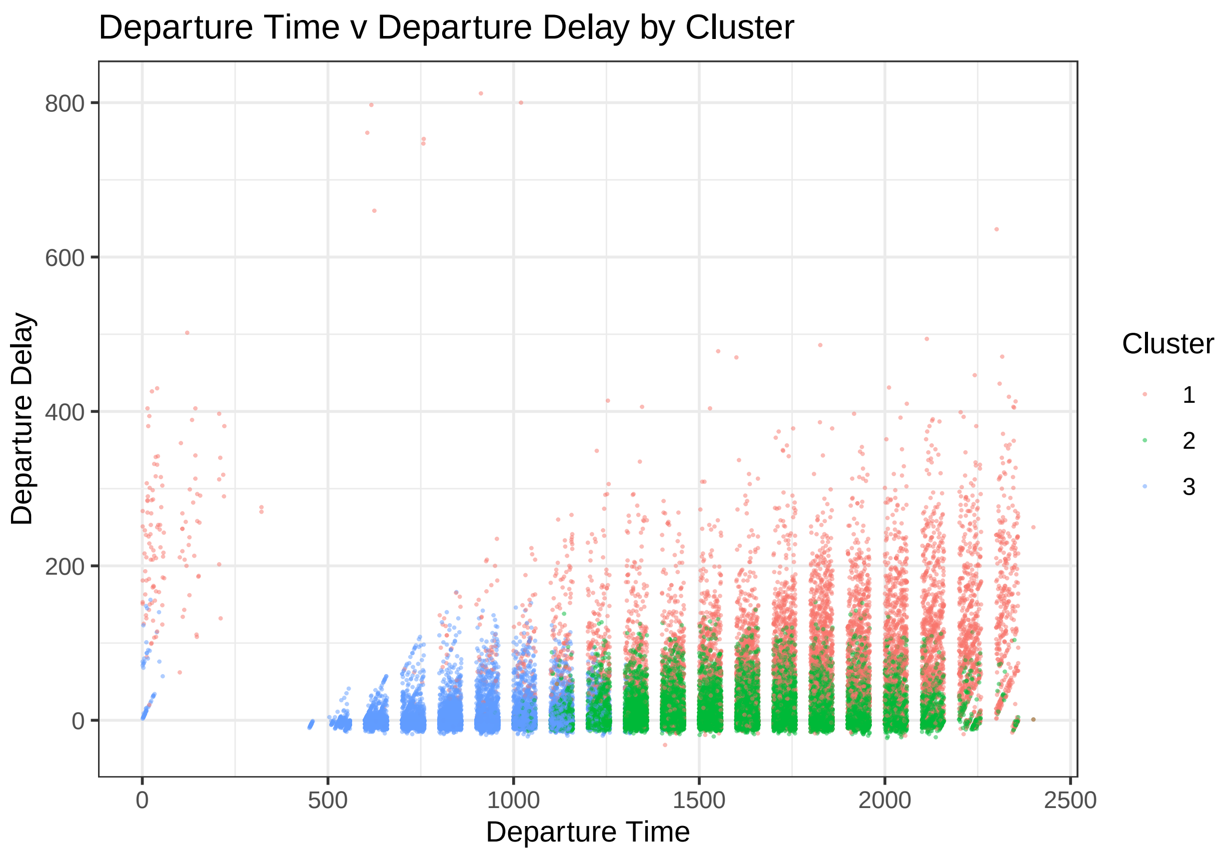
Cluster 1 is the smallest by observation numbers, but contains the largest within cluster sum of squares. The spread out nature of cluster 1 can also be noticed in the plot above. 22.5% of variability in the data set is accounted for by the variability between clusters, this suggests a lackluster level of distinction between clusters. Examining the cluster means the only noticeable differences seem to arise in departure time, scheduled departure time, departure delay, arrival time, arrival delay, and time made up in the air.

```
##   dep_time sched_dep_time dep_delay arr_time arr_delay air_time distance made
```

```
## 1     2400           2359         1      515        30       230      1617  -29
## 2     2400           2359         1      324       -14       186      1576   15
## 3     2400           2359         1      338        -1       196      1576    2
## 4     2400           1950       250      107       217       101       733   33
## 5     2359           2359         0      440        -5       203      1617    5
## 6     2359           2255        64      123        87        34       187  -23
##    temp wind_dir wind_speed wind_gust visib cluster
## 1 35.96      300   20.71404  27.61872    10       1
## 2 42.98      330   13.80936  21.86482    10       2
## 3 48.02      330   19.56326  25.31716    10       1
## 4 91.94      270   10.35702  19.56326    10       1
## 5 30.02      300   18.41248  24.16638    10       1
## 6 33.98      260   17.26170  24.16638    10       1

##   dep_time sched_dep_time dep_delay arr_time arr_delay air_time distance made
## 1        1           2100       181      124       179      127      725    2
## 2        1           2245        76      121        87       56      273  -11
## 3        1           2128       153      247       172      234     1626  -19
## 4        1           2250        71      120        75       54      264   -4
## 5        1           1930       271      106       245       36      200   26
## 6        1           2359         2      336        -5      189     1576    7
##    temp wind_dir wind_speed wind_gust visib cluster
## 1 32.00      260   21.86482  35.67418    10       1
## 2 33.98      320   16.11092  25.31716    10       3
## 3 33.98       80   18.41248  25.31716     3       1
## 4 39.92      300   17.26170  21.86482    10       3
## 5 59.00      120   12.65858  20.71404     9       1
## 6 51.08      300   29.92028  35.67418    10       3

##   dep_time sched_dep_time dep_delay arr_time arr_delay air_time distance made
## 1      912           1940       812     1228       821      174     1010   -9
## 2     1020           2100       800     1336       784      335     2475   16
## 3      617           1700       797      858       783      313     2248   14
## 4      606           1725       761      923       783      222     1417  -22
## 5      758           1925       753     1049       744      149      950    9
## 6      757           1930       747     1013       744       85      541    3
##    temp wind_dir wind_speed wind_gust visib cluster
## 1 62.06      170   17.26170  25.31716  2.50       1
## 2 33.98       80   18.41248  25.31716  3.00       1
## 3 57.02      180   25.31716  33.37262  0.12       1
## 4 57.02      180   25.31716  33.37262  0.12       1
## 5 62.06      170   17.26170  25.31716  2.50       1
## 6 33.98      360   14.96014  24.16638 10.00       1

##   dep_time sched_dep_time dep_delay arr_time arr_delay air_time distance made
## 1     1408           1440       -32     1549       -10       52      229  -22
## 2     2006           2029       -23     2134       -42       69      419   19
## 3     2137           2159       -22     2232       -44       38      269   22
## 4     2044           2106       -22     2143       -30       40      200    8
## 5     1038           1059       -21     1218       -36       74      479   15
## 6     2008           2029       -21     2225         9       67      419  -30
##    temp wind_dir wind_speed wind_gust visib cluster
## 1 51.80      330   25.31716  36.82496    10       1
```

```
## 2 57.92       280   17.26170  28.76950    10        2
## 3 30.92       300   12.65858  20.71404    10        2
## 4 33.08       320   11.50780  24.16638    10        2
## 5 44.06       250   18.41248  23.01560    10        3
## 6 69.98        20   12.65858  21.86482    10        2
```

Arranging some observations we can see that flights with late departure times seem to be assigned to cluster 1 for the most part. Flights with early departure times are assigned to a mixed bag of either cluster 1 or 3, suggesting the model doesn't know where to cluster these observations. Flights with the greatest departure delay are all assigned to cluster 1, and the flights with the smallest departure delay (left early) are assigned to a mixed number of clusters.



The plot above further explores the significant variables that contributed towards cluster assignments. Cluster 1 for the most part is assigned to observations with higher values for departure delay. Cluster 2 and cluster 3 appear to split the observations with less significant delays based of departure time.

5 Results

5.1 Model 1 Logistic Regression

The logistic regression model built correctly predicted the binary outcome of significant arrival delay 77.7% of the time. Significant arrival delay was a derived binary factor variable from arrival delay, if a flight arrival time was more than 7 minutes late the value of true (1) was assigned, less than 7 minutes late the value of false was assigned (0). The explanatory variables used to build the model were air time, time made up in the air, departure time, flight distance, wind speed, and visibility. These 6 independent variables were all significant predictors with p values less than 0.05 and the absolute value of their z values being greater than 2 signifying statistical significance from zero. The deviance residuals are centered close to zero and the residual deviance was less than the null deviance indicating the explanatory variables contribute to model fit. Examining the coefficient estimates, The intercept or the log odds of a flight being delayed is -1.598. For a one unit increase in air time, the log odds for the flight being delayed decrease 0.00215. For a one unit increase in time made up in the air, the log odds for the flight being delayed decrease 0.08733. For a one unit increase in departure time, the log odds of the flight being delayed increases 0.001452. For a one unit increase in distance, the log odds of the flight being delayed increases 0.000328. For a one unit increase in wind speed, the log odds of the flight being delayed increases 0.01331. For a one unit increase in visibility, the log odds of the flight being delayed decrease 0.1163.

5.2 Model 2 K-Means Clustering

The results from the K-Means Clustering model built appear to be less conclusive. This can be attributed to the lack of variance retained from the original numeric data set in dimensionality reduction. For K-Means, Principal Component Analysis is used to reduce the input data to two dimensions, making it suitable for the distance based K-Means algorithm. Dimension 1 accounted for 21.3% of variance in the original data set and dimension 2 accounted for 16.7% of variance in the original data set, summing to 38%. Three distinct clusters were formed although these clusters did see vast overlap among outer points and points further from the cluster centroids. Examining cluster means, cluster 1 averaged the latest departure time in the day for departing flights, it also featured the highest departure and arrival delay among clusters. While being the most delayed, cluster 1 did also average a value of -3.4 for time made up in the air, signifying that flights on average spent 3.4 more minutes in the air than scheduled. Cluster 2 featured a mean value for departure time, departure delay, and arrival delay between that of cluster 1 and cluster 3. Cluster 2 had the highest mean value for time made up in the air of 8.5 minutes, denoting that on average flights in cluster 2 spent 8.5 minutes less in the air than predicted. Cluster 3 feature the lowest mean departure time, departure delay, and arrival delay while the mean time made up in the air fell between that of cluster 1 and 2.

Multiple conclusions can be drawn from the cluster means. First, the later the departure time in the day, the higher the average departure and arrival delay among flights. Second, flights leaving earliest in the day arrive earliest in comparison to their scheduled arrival times despite not making up the most time in the air among clusters. Third, flights departing in the afternoon to late afternoon make up the most time in the air on average.

6 Discussion

6.1 Final model interpolation

As mentioned above, both models created give valuable insights in the daily operations of commercial aircraft. The logistic regression model allows prediction of probability for a binary outcome, then classifying that observation based on its predicted probability. K-Mean Clustering groups similar observations allowing for trends and patterns to be recognized. Conclusions were drawn from logistic regression using the log odds for coefficient estimates and test set to test model accuracy. Conclusions were drawn from K-Mean using cluster means and cluster assignments for various observations.

6.2 Use of Model

Future use cases for the logistic regression model built to classify significant arrival delay include but are not limited to, flight scheduling, arrival predictions, traffic management, and operation logistics planning. The model could be integrated into airport capacity planning systems to optimize resource allocation and mitigate congestion during peak hours. Furthermore, airlines could utilize the model to enhance customer service by proactively managing delays and informing passengers about potential disruptions. Additionally, government agencies responsible for transportation infrastructure could leverage the insights generated by the model to improve overall system efficiency and reliability. Overall, the versatility of the logistic regression model extends beyond solely arrival delay classification.

Future use cases for the K-Means Clustering model built include travel patterns, demand forecasting, and route optimization. For instance, airlines can identify high-demand routes or peak travel times, allowing for more strategic scheduling and resource allocation. Moreover, k-means clustering enables airlines to segment their customer base more effectively, tailoring services and marketing efforts to different traveler preferences. Overall, leveraging logistic regression and k-means clustering in commercial aviation facilitates data-driven decision-making and offers valuable insights for enhancing operational efficiency and customer satisfaction.

7 Furture Work

To build more accurate and tailored models deployable to production level standards, more time should be spent methodically imputing missing values and detecting potential outliers specific to the agency or corporation utilizing the the various models built. By creating subsets of the data set tailored to specific airports or airlines, organizations can fine-tune models to fit their unique data characteristics and operational requirements accurately. This customization enables airports or airlines to extract insights directly applicable to their operations, optimizing resource allocation, enhancing customer service, and ultimately improving overall efficiency. By prioritizing data pre-processing techniques and customization efforts, organizations can develop models that not only meet but exceed production standards, driving meaningful impact within the aviation industry.

8 References

"Air Traffic by the Numbers." Air Traffic By The Numbers | Federal Aviation Administration, www.faa.gov/air_traffic/by_the_numbers. Accessed 22 Mar. 2024.

"Airlines for America." Airlines For America, www.airlines.org/impact/. Accessed 22 Mar. 2024. Baran, Michelle. "These Are the 20 Busiest Airports in the United States." AFAR Media, AFAR Media, 17 Mar. 2024, www.afar.com/magazine/busiest-airports-in-the-us.

Khaksar, H., and A. Sheikholeslami. "Airline Delay Prediction by Machine Learning Algorithms." Scientia Iranica, Sharif University of Technology, 1 Oct. 2019, scientiairanica.sharif.edu/article_20020.html.

Tang, Yuemin. "Airline Flight Delay Prediction Using Machine Learning Models." Airline Flight Delay Prediction Using Machine Learning Models, dl.acm.org/doi/fullHtml/10.1145/3497701.3497725#bib3. Accessed 22 Mar. 2024.

```r
knitr::opts_chunk$set(echo = TRUE)
library(nycflights13) #used for data
library(ggplot2) #used for visualizations
library(cluster) #used for k-means
library(factoextra) #used for k-means visualizations
library(ISLR) #used for logistic glm
library(pROC) #used for calculating optimal cutoff
library(tidyverse) #used for data wrangling
library(tidyr) #used for data wrangling
library(dplyr) #used for data wrangling
library(gridExtra) #used for plot arrangments
data <- read.csv("NYCF.csv")
set.seed(123)
data <- read.csv("NYCF.csv")
transposed_data <- t(data)
transposed_data[,1]
ggplot(data=flights, aes(x=as.factor(month),fill=origin))+
  geom_bar()+
  theme_bw()+
  labs(title="Number of Flights per Month by Origin",x="Month",y="Count",fill="Origin")
xtabs(~origin+as.factor(month), flights)
ggplot(data=flights, aes(x=fct_infreq(carrier)))+
  geom_bar(fill="steelblue",color="black",alpha=0.5)+
  theme_bw()+
  labs(title="Barplot of Number of Flights by Carrier",x="Carrier",y="Count")
xtabs(~carrier, flights)
ggplot(data=flights, aes(x=fct_infreq(origin)))+
  geom_bar(fill="steelblue",color="black",alpha=0.5)+
  theme_bw()+
  labs(title="Barplot of Number of Flights by Origin",x="Origin",y="Count")
xtabs(~origin, flights)
ggplot(data=flights, aes(x=distance, fill=carrier))+
  geom_histogram(binwidth = 200)+
  theme_bw()+
  labs(title="Histogram of Flight Distance by Carrier",x="Distance",y="Count",fill="Carrier")
flights %>%
  ggplot(aes(x=dep_delay, color=origin))+
  geom_density(alpha=0.3)+
  labs(title="Departure Delay Density by Origin",y="Density",x="Departure Delay",color="Origin")+
  theme_bw()+ xlim(-30,120)
by(flights$dep_delay, flights$origin, summary)
flights %>%
  mutate(origin = as.factor(origin))%>%
  group_by(origin)%>%
  drop_na(dep_delay)%>%
  summarise(mean_d=mean(dep_delay),sd_d=sd(dep_delay))%>%
  ggplot(aes(origin,mean_d))+
  geom_point(size=5,color="blue", alpha=0.5)+
  geom_errorbar(aes(x=origin,
                    ymin=mean_d - sd_d,
                    ymax=mean_d + sd_d,
                    width=0.5),
                alpha=0.5, color="black" )+
```

```
  labs(title="Origin vs Mean Departure Delay",x="Origin",y="Mean Departure Delay")

flights2 <- flights %>% drop_na(dep_delay)

p1 <- flights %>%
  group_by(carrier)%>%
  drop_na(dep_delay)%>%
  summarise(mean_d=mean(dep_delay)) %>%
  mutate(carrier=fct_reorder(carrier,mean_d))%>%
  ggplot(aes(carrier,mean_d))+
  geom_point(size=5,color="orange", alpha=0.5)+
  geom_segment(aes(x=carrier,
                   y=mean(flights2$dep_delay),
                   xend=carrier,
                   yend=mean_d),
               color="grey")+
  geom_hline(yintercept=mean(flights2$dep_delay),
             color="grey",
             size=1)+
  theme_bw()+
  theme(axis.text.x=element_text(angle=90))+
  labs(y="Mean Departure Delay",x="Carrier",title="Carrier vs Mean Departure Delay")

flights4 <- flights %>% drop_na(arr_delay)

p2<-flights %>%
  group_by(carrier)%>%
  drop_na(arr_delay)%>%
  summarise(mean_d=mean(arr_delay)) %>%
  mutate(carrier=fct_reorder(carrier,mean_d))%>%
  ggplot(aes(carrier,mean_d))+
  geom_point(size=5,color="steelblue", alpha=0.5)+
  geom_segment(aes(x=carrier,
                   y=mean(flights4$arr_delay),
                   xend=carrier,
                   yend=mean_d),
               color="grey")+
  geom_hline(yintercept=mean(flights4$arr_delay),
             color="grey",
             size=1)+
  theme_bw()+
  theme(axis.text.x=element_text(angle=90))+
  labs(y="Mean Arrival Delay",x="Carrier",title="Carrier vs Mean Arrival Delay")

flights3 <- flights %>% drop_na(distance)

p3<-flights %>%
  group_by(carrier)%>%
  drop_na(distance)%>%
  summarise(mean_d=mean(distance)) %>%
  mutate(carrier=fct_reorder(carrier,mean_d))%>%
  ggplot(aes(carrier,mean_d))+
  geom_point(size=5,color="blue3", alpha=0.5)+
```

```
  geom_segment(aes(x=carrier,
                   y=mean(flights3$distance),
                   xend=carrier,
                   yend=mean_d),
               color="grey")+
  geom_hline(yintercept=mean(flights3$distance),
             color="grey",
             size=1)+
  theme_bw()+
  theme(axis.text.x=element_text(angle=90))+
  labs(y="Mean Distance",x="Carrier",title="Carrier vs Mean Flight Distance")

flights$made <- flights$dep_delay - flights$arr_delay
flights5 <- flights %>% drop_na(made)

p4<-flights %>%
  group_by(carrier)%>%
  drop_na(made)%>%
  summarise(mean_d=mean(made)) %>%
  mutate(carrier=fct_reorder(carrier,mean_d))%>%
  ggplot(aes(carrier,mean_d))+
  geom_point(size=5,color="red", alpha=0.5)+
  geom_segment(aes(x=carrier,
                   y=mean(flights5$made),
                   xend=carrier,
                   yend=mean_d),
               color="grey")+
  geom_hline(yintercept=mean(flights5$made),
             color="grey",
             size=1)+
  theme_bw()+
  ylim(-5,20)+
  theme(axis.text.x=element_text(angle=90))+
  labs(x="Carrier",y="Mean Time Made Up in Air",title="Carrier vs Mean Time Made Up")
grid.arrange(p1, p2, p3, p4, nrow = 2)
par(mfrow=c(1,1))
flights %>%
  ggplot(aes(x=sched_dep_time,color=origin))+
  geom_density(alpha=0.3)+
  labs(title="Scheduled Takeoff Density",x="Scheduled Departure Time",y="Density")+
  theme_bw()
flights %>%
  group_by(month)%>%
  drop_na(dep_delay)%>%
  summarise(mean_d=mean(dep_delay)) %>%
  ggplot(aes(month,mean_d))+
  geom_point(size=5,alpha=0.5)+
  geom_line(size=1, alpha=0.5, color="blue")+
  theme_bw()+
  labs(title="Mean Departure Delay vs Month",x="Month",y="Mean Departure Delay")+
  ylim(0,25)+
  scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12))
flightsw<- flights %>% inner_join(weather)
```

```
flightsw %>%
  group_by(visib) %>%
  drop_na(wind_speed) %>%
  drop_na(dep_delay) %>%
  summarise(mean_w=mean(wind_speed),mean_d=mean(dep_delay)) %>%
  ggplot(aes(visib,mean_d))+
  geom_point(size=2,alpha=0.5,color="black")+
  geom_smooth(size=1,alpha=0.2,method=lm)+
  scale_x_continuous(breaks=c(0,1,2,3,4,5,6,7,8,9,10))+
  theme_bw()+
  labs(title="Visibility vs Mean Departure Delay",x="Visibilty",y="Mean Departure Delay")
set.seed(123)
data <- read.csv("NYCF.csv")
data$sig_arr_delay <- ifelse(data$arr_delay > 7, "1", "0")
data <- data[!is.na(data$sig_arr_delay), ]
data$sig_arr_delay <- as.factor(data$sig_arr_delay)
Split <- sample(nrow(data), 0.70*nrow(data), replace=FALSE)
Train <- data[Split,]
Test <-  data[-Split,]
m1 <- glm(sig_arr_delay ~ air_time+made+dep_time+distance+wind_speed+visib, data=Train, family ='binomia
summary(m1)
ll.null <- m1$null.deviance/-2
ll.proposed <- m1$deviance/-2

## McFadden's Pseudo R^2 = [ LL(Null) - LL(Proposed) ] / LL(Null)
(ll.null - ll.proposed) / ll.null

## The p-value for the R^2
1 - pchisq(2*(ll.proposed - ll.null), df=(length(m1$coefficients)-1))
predictions<-predict(m1, Test, type = 'response')

roc_curve <- roc(Test$sig_arr_delay, predictions)

optimal_cutoff <- coords(roc_curve, "best", ret = "threshold")

print(optimal_cutoff)
labels <- ifelse(predictions> 0.3578069, '1', '0')
labels <- rep(labels, length.out = length(Test$sig_arr_delay))

conf_matrix <- table(labels, Test$sig_arr_delay)

df <- data.frame(labels = labels, sig = Test$sig_arr_delay)
head(df)

conf_matrix
table(labels == Test$sig_arr_delay)/length(Test$sig_arr_delay)
TP <- 55097   # True positives
TN <- 20858   # True negatives
FP <- 9547   # False positives
FN <- 12222 # False negatives

# Calculate sensitivity (true positive rate)
sensitivity <- TP / (TP + FN)
```

```
# Calculate specificity (true negative rate)
specificity <- TN / (TN + FP)

# Calculate misclassification error rate
misclassification_error <- (FP + FN) / sum(conf_matrix)

sensitivity
specificity
misclassification_error
par(mfrow=c(2,2))
numeric_data1 <- data  %>% slice(1:25000) %>%
  select_if(is.numeric)
numeric_data1 <- numeric_data1[,c("dep_time","sched_dep_time","dep_delay","arr_time","arr_delay","air_t
numeric_data1 <- na.omit(numeric_data1)
numeric_data <- numeric_data1[, !names(numeric_data1) %in% "precip"]
numeric_data <- numeric_data[, !names(numeric_data) %in% "year"]
numeric_data <- scale(numeric_data)
p1<-fviz_nbclust(numeric_data, kmeans, method = "wss", k.max = 8)
p2<-fviz_nbclust(numeric_data, kmeans, method = "silhouette", k.max = 8)
grid.arrange(p1, p2, nrow = 1)
set.seed(123)
par(mfrow=c(1,1))
numeric_data1 <- data %>% select_if(is.numeric)
numeric_data1 <- numeric_data1[,c("dep_time","sched_dep_time","dep_delay","arr_time","arr_delay","air_t
numeric_data1 <- na.omit(numeric_data1)
numeric_data <- numeric_data1[, !names(numeric_data1) %in% "precip"]
numeric_data <- numeric_data[, !names(numeric_data) %in% "year"]
numeric_data <- scale(numeric_data)

k3 <- kmeans(numeric_data, centers = 3, nstart = 1, iter.max = 10)
fviz_cluster(k3, data= numeric_data)
k3$withinss
k3$tot.withinss
k3$betweenss
k3$size
numeric_data1$cluster <- k3$cluster
aggregate(numeric_data1, by=list(cluster=numeric_data1$cluster), mean)
f <- arrange(numeric_data1, desc(dep_time))
a <- arrange(numeric_data1, dep_time)
head(f)
head(a)

f2 <- arrange(numeric_data1, desc(dep_delay))
a2 <- arrange(numeric_data1, dep_delay)
head(f2)
head(a2)
ggplot(numeric_data1, aes(x=dep_time,y=dep_delay, color=as.factor(cluster)))+geom_point(alpha=0.5, size=
  labs(title="Departure Time v Departure Delay by Cluster",y="Departure Delay",x="Departure Time",color=
```