Purpose/Goal

Flight delay is the serious and extensive problem in United States. Increasing flight delays place a significant strain on the U.S. air travel system and cost airlines, passengers and society many billions of dollars each year. According to Federal Aviation Administration (FAA), the total cost of air transportation delay in 2007 is approximately **$31.2 billion**. This includes $16.7 billion direct cost to passengers. Flight delay also has indirect effect on U.S economy. Our study/Analysis will help the FAA to predict the airline arrival delay so that they can make the precautionary action to avoid/reduce the cost incurred due to flight delays. Below are the objectives of our analysis.

- Which airline and airport has the highest delay time?
- Predicting the arrival delay and identifying the factors which influence the delay.
- Which day of the month in a year has the highest delay time?

About data

This analysis is based on the sample data collected from the population comes originally from The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) for the year **2008**. We have sampled the data based on carrier and originating airport with the equal proportion of each category. The sample has **10,997 rows** and **21 columns** which contains the data for five popular airlines (Alaska Airlines Inc., JetBlue Airways, Delta Air Lines Inc., Hawaiian Airlines Inc., Southwest Airlines Co.) and five origin Airports (Chicago O'Hare, Dallas/Fort Worth, Denver, Hartsfield–Jackson Atlanta, Los Angeles). Description of variables can be found in *appendix section A.1.*

Initial analysis of the data

- We observed that variables ArrivalTime, and ArrivalDelay have about 0.28% of missing data and ActualElapsedTime, AirTime, ArrivalDelay, and TaxiIn have about 0.37% of missing data. Hence, we imputed the median to the missing fields across the columns.
- Columns such as Departure Time, Scheduled Departure Time, Arrival Time, Scheduled Arrival time were in ‘hhmm’ format. Hence these columns are normalized to minutes.

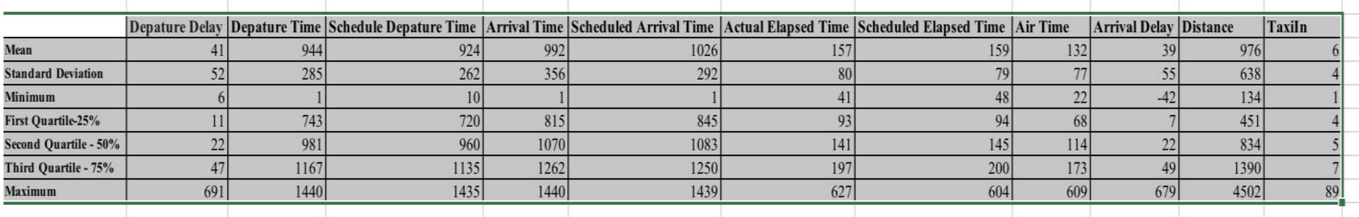
Methods and Tools

Below are the methods and tools used for the analysis;

- Descriptive Statistics
- Normal distribution, Histograms, ANOVA, Multiple Linear Regression, Factor Analysis, Clustering
- Correlation
- Python, JMP, and Tableau

Analysis

Before we initiate our analysis let’s do some exploratory data analysis which will help us to better understand the data. Rather than exploring on all the variables let’s assume the significant factors which may influence our objectives. Below are the descriptive statistics for the variables.



	Depature Delay	Depature Time	Schedule Depature Time	Arrival Time	Scheduled Arrival Time	Actual Elapsed Time	Scheduled Elapsed Time	Air Time	Arrival Delay	Distance	TaxiIn
Mean	41	944	924	992	1026	157	159	132	39	976	6
Standard Deviation	52	285	262	356	292	80	79	77	55	638	4
Minimum	6	1	10	1	1	41	48	22	-42	134	1
First Quartile-25%	11	743	720	815	845	93	94	68	7	451	4
Second Quartile - 50%	22	981	960	1070	1083	141	145	114	22	834	5
Third Quartile - 75%	47	1167	1135	1262	1250	197	200	173	49	1390	7
Maximum	691	1440	1435	1440	1439	627	604	609	679	4502	89

Let’s get started with findings from this statistic:

Departure Delay

Count	Mean	Median	Standard Deviation
10997	41	22	52

Breaking down the above fundamental statistics,

Count: Number of records collected for the sample is 10,997.

Mean: The average departure delay time is 41 minutes. We could say if the Departure delay is more than 41 minutes, then the flight is delayed. But practically passengers will lose the patient if they need to wait for so long.

Median: The median is the center point in the dataset. Comparing the mean of 41 minutes and median of 22 minutes lead to the suspicion that the data may have outliers. Hence, we could assume mean is influenced by outliers and delay time of 22 minutes derived from median is acceptable.

Standard Deviation: This tells how spread the data are from the mean. The Std.Dev of 52 shows that about 68% of departure delay fall within one standard deviation.

75% of the data (Departure delay) falls within 47 minutes. However, the maximum value of 691 minutes tells us the data is **right skewed**. This can be clearly seen in the histogram plot shown in *appendix section A.2.*

Other variable follows similar break down with the difference in the value. Hence going forward, we will highlight the major findings for other rating sites.

Departure Time

Most of the flights start departing from 720th minutes of each day. The frequency of flight departures is less during the start of the day. Hence, we could say the data is **left skewed**.

Scheduled Departure Time

Distribution of Scheduled Departure Time is almost similar to Departure Time.

AirTime

On an average the duration of travel is 132 minutes and 75% of flight duration is within 173 minutes which substantiate that the distribution is **right skewed**. Since the air time depends of Distance between origin and destination we could say, Air Time and Distance should be highly corelated. However, airtime may be influenced by other factors like weather.

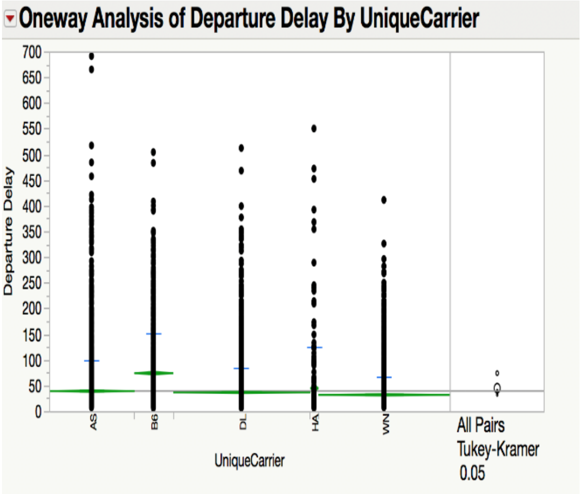
Arrival Delay

Less than 10% of the time the flight arrived at the destination earlier than the scheduled arrival time. However, the arrival delay follows the same distribution as departure delay. Hence, it is evident that the flight departing late will arrive late. The average arrival delay is 39 minutes which is close to average departure delay. For instance, the 10 minutes delayed departure will cause 10 minutes delay in arrival. Departure delay is the most influencing factor for predicting Arrival delay.

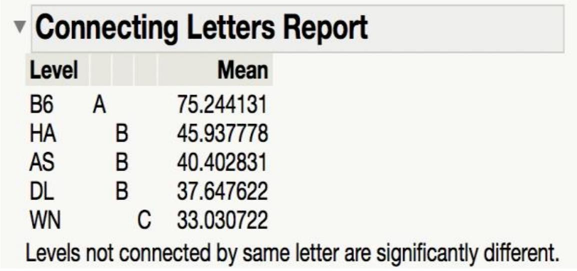
The distribution of these variables can be seen in *appendix A.2.*

Which airline and airport has the highest delay time?

To determine which airline has the highest departure delay and arrival delay, ANOVA was performed on the data set. Initially we performed ANOVA on departure delay based on airline. From the analysis, we found that the **unequal variance assumption failed** due to less p-value (**0.0001)** of levene's test. Hence, we considered the p-value (0.0001) of Welch's Test to continue our analysis.



From the analysis of variance and connecting letter report, we have observed that ***Jet blue (B6)*** has the highest average departure delay time of **75 minutes** and **Southwest Airlines Co. (WN**) has the lowest average departure delay time of **33 minutes** compared to other airlines. There is no significant difference exist between Alaska (AS) and Delta(DL) airline.



Connecting Letters Report

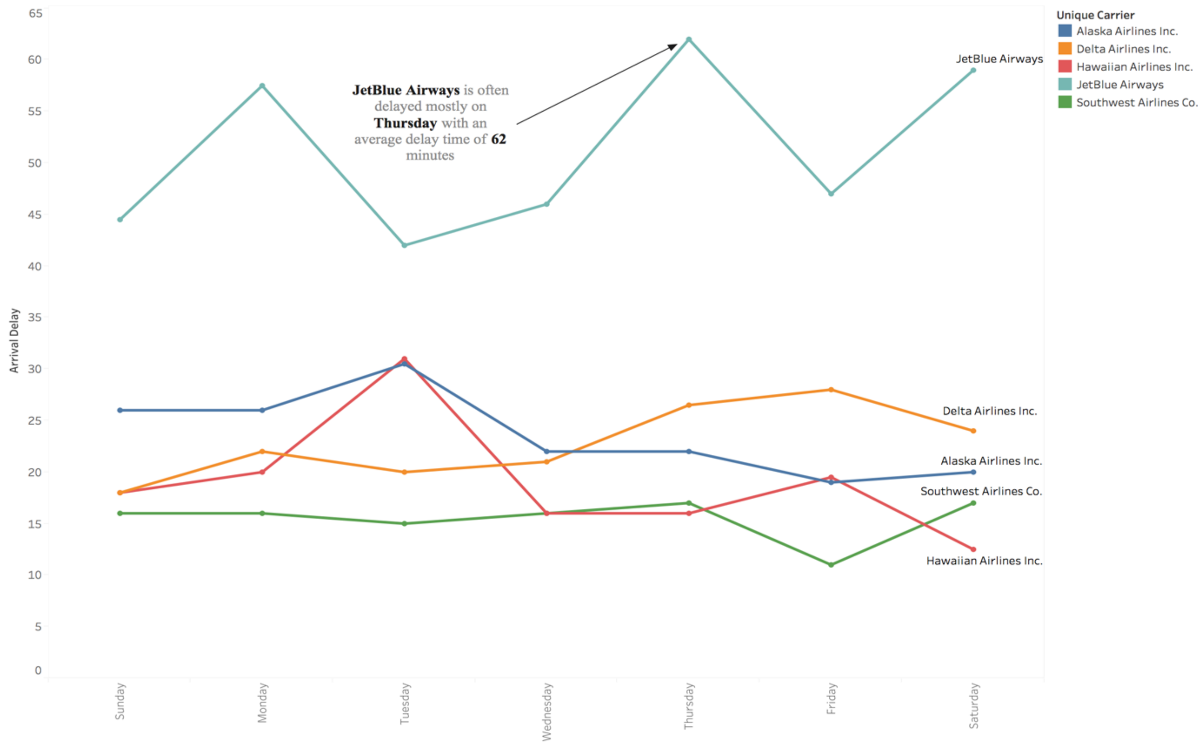
Level				Mean
B6	A			75.244131
HA		B		45.937778
AS		B		40.402831
DL		B		37.647622
WN			C	33.030722

Levels not connected by same letter are significantly different.

Similarly, ***Chicago O'Hare international airport (ORD)*** has the highest departure delay time of **72 minutes** compared to all other airports. *See the appendix A.3* for ANOVA output. We considered the Welch's test p-value for our analysis since the **unequal variance assumption failed.**

Since, the ***Jet blue (B6)*** and ***Chicago O'Hare international airport (ORD)*** has the highest departure delay, we could say this airline and airport will have the highest **arrival delay**. This can be evidently seen from the below report from Tableau. Both Jet Blue (B6) and Chicago O'Hare international airport (ORD) are **often delayed on Thursday** with average arrival delay time of **62 minutes.**

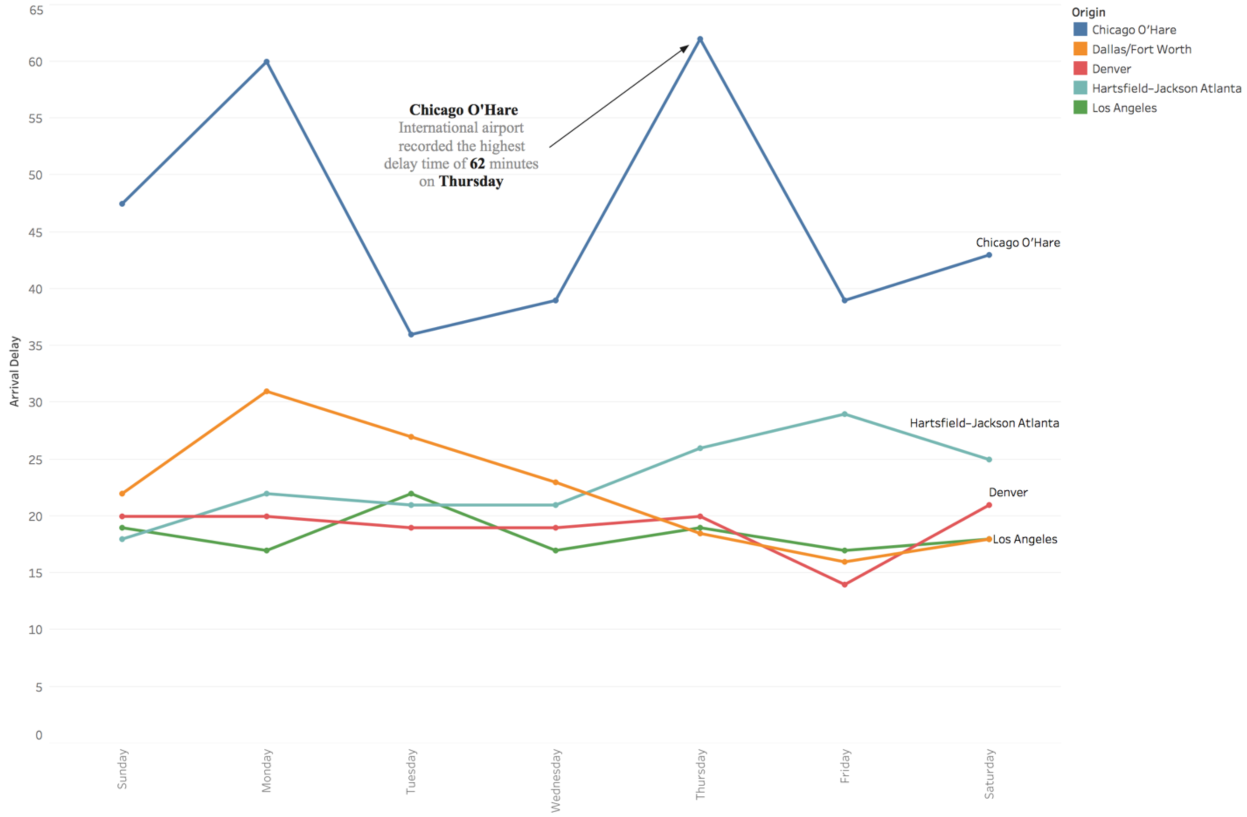
Which airline is often delayed?



JetBlue Airways is often delayed mostly on **Thursday** with an average delay time of **62** minutes

The trend of median of Arr Delay for Day Of Week. Color shows details about Unique Carrier. The marks are labeled by Unique Carrier.

Most delay occured in which International Airport?



Chicago O'Hare International airport recorded the highest delay time of **62** minutes on **Thursday**

Predicting the arrival delay and identifying the factors which influence the delay.

Before building the model to predict the arrival delay, lets identify the influencing factors which are correlated with the variable Arrival Delay.

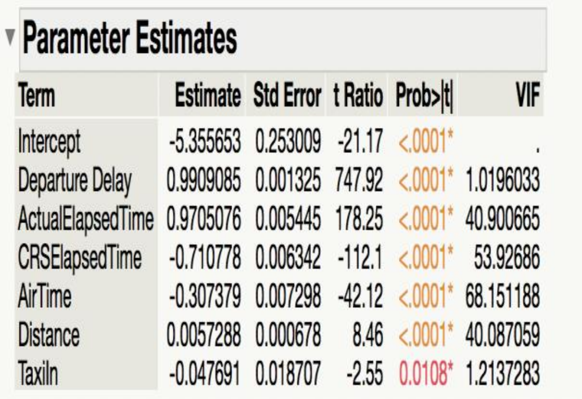
See appendix A.4 for pair plot and correlation matrix. From the two graph, it is evident that Departure Delay is highly corelated with arrival delay. It has the correlation value of **0.95** (i.e.) 95% of variability in arrival delay can be explained Departure delay. Hence, it is one of the most influencing factor to predict the arrival delay. We could also see the independent variables such as ActualElapsedTime, AirTime, Distance, and CRSElapsedTime are highly corelated. This cause the suspicion of multicollinearity. We used the VIF value to check the multicollinearity between these variables.

We used multiple linear regression to predict the model. In the initial analysis, we observed that, except Departure Delay, ActualElapsedTime, CRSElapsedTime, AirTime, Distance, and TaxiIn, all the other variables are not significant to the model since the p-value is greater than 0.05. Hence, we removed those variables from the model and identified the below model with below model with RSquare value of **0.98 (98%).**

<u>Model 1:</u>

Arrival Delay = -5.56 + 0.99 * Departure Delay + 0.97 * ActualElapsedTime – 0.71 * CRSElapsedTime – 0.31 * AirTime + 0.005 * Distance – 0.05 * TaxiIn

The ANOVA **p- value (0.0001)** for this model is less than 0.05 and all the independent variables are significant. However, as stated above multicollinearity exist between the variables ActualElapsedTime, CRSElapsedTime, Airtime, and Distance, since the **VIF** value is greater than **5.0** (The threshold value we used in the model).

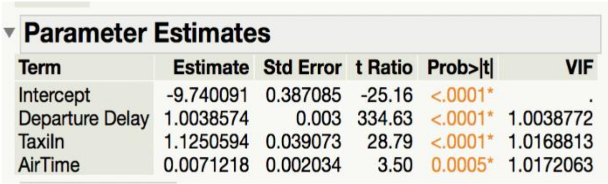


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-5.355653	0.253009	-21.17	<.0001*	.
Departure Delay	0.9909085	0.001325	747.92	<.0001*	1.0196033
ActualElapsedTime	0.9705076	0.005445	178.25	<.0001*	40.900665
CRSElapsedTime	-0.710778	0.006342	-112.1	<.0001*	53.92686
AirTime	-0.307379	0.007298	-42.12	<.0001*	68.151188
Distance	0.0057288	0.000678	8.46	<.0001*	40.087059
TaxiIn	-0.047691	0.018707	-2.55	0.0108*	1.2137283

Since, we have multicollinearity exist between the variables, the above model is not adequate. Hence, we used backward elimination process to remove these variables from the model.

After removing the variable ActualElapsedTime, CRSElapsedTime, and Distance, the VIF value of Airtime reduced to **1.02.**



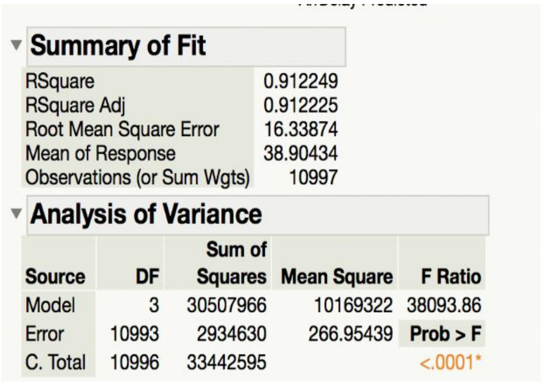
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-9.740091	0.387085	-25.16	<.0001*	.
Departure Delay	1.0038574	0.003	334.63	<.0001*	1.0038772
TaxiIn	1.1250594	0.039073	28.79	<.0001*	1.0168813
AirTime	0.0071218	0.002034	3.50	0.0005*	1.0172063

Model 2:

Arrival Delay = -9.74 + 1.00 * Departure Delay + 1.13 * TaxiIn + 0.01 * AirTime

The above model is adequate with RSquare value of **0.91.** (i.e. **91%** of the variability of Arrival delay can be explained by these Departure Delay, TaxiIn, and AirTime). Even though the RMSE value (16.33) is high compared to previous model (RMSE: approx. 7), this model is model stable since all the multiclonality variables are removed.



Summary of Fit

RSquare	0.912249
RSquare Adj	0.912225
Root Mean Square Error	16.33874
Mean of Response	38.90434
Observations (or Sum Wgts)	10997

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	30507966	10169322	38093.86
Error	10993	2934630	266.95439	Prob > F
C. Total	10996	33442595		<.0001*

Interpreting the coefficients

For every one-minute increase in Departure Delay, Arrival delay will increase by **one minute.**

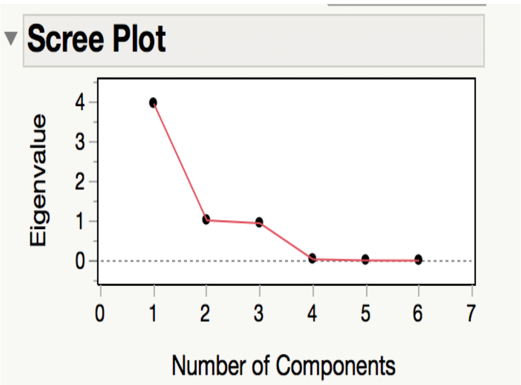
For every one-minute increase in TaxiIn, Arrival delay will increase by **1.13 minute.**

For one-minute increase in AirTime, Arrival delay will increase by **0.01 minute.**

PCA / Factor Analysis

Before concluding model 2 is the best model to predict the arrival delay, let's do PCA / factor analysis to reduce the multicollinearity between these variables.

Let's determine the number of components to be used to our analysis using scree plot.



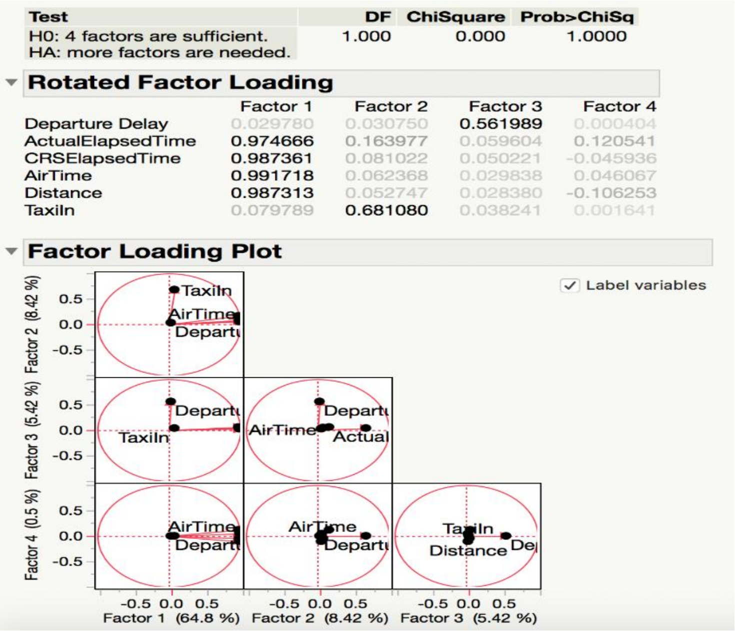
As per the scree plot, let's try with 2 factors. Since the ChiSq value (0.0001) is less than 0.05, we need more factors. Hence, we try 4 factors.

The greater ChiSq value (**1.000**) suggest that 4 factors are sufficient for the analysis.

As the factor values for the variables are greater than 0.50 except factor 4. factor 1, factor 2 and factor 3 are significant to run the MLR. Factor 1 explains 64.8 % variability.

We ran the MLR with these three factors identified. The overall model is significant as the ANOVA p-value is less than 0.05. This also has the high RSquare value of 0.90. However, the factor 1 is not significant to the model since, the p-value **(0.1896)** is high.

Hence, we removed this factor from the model and identified the below adequate model with RSquare value of **0.90** and all the factors are significant.



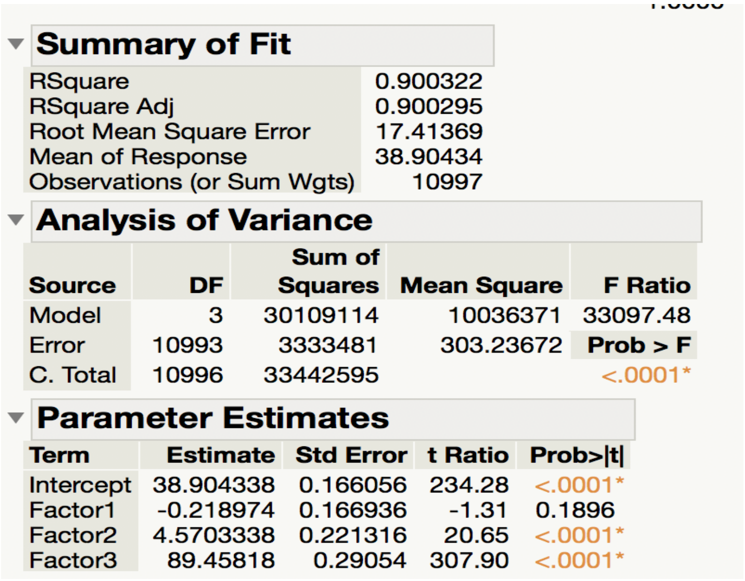
Test	DF	ChiSquare	Prob>ChiSq
H0: 4 factors are sufficient. HA: more factors are needed.	1.000	0.000	1.0000

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4
Departure Delay	0.029780	0.030750	0.561989	0.000404
ActualElapsedTime	0.974666	0.163977	0.059604	0.120541
CRSElapsedTime	0.987361	0.081022	0.050221	-0.045936
AirTime	0.991718	0.062368	0.029838	0.046067
Distance	0.987313	0.052747	0.028380	-0.106253
TaxiIn	0.079789	0.681080	0.038241	0.001641

Factor Loading Plot

☑ Label variables



Summary of Fit

RSquare	0.900322
RSquare Adj	0.900295
Root Mean Square Error	17.41369
Mean of Response	38.90434
Observations (or Sum Wgts)	10997

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	30109114	10036371	33097.48
Error	10993	3333481	303.23672	Prob > F
C. Total	10996	33442595		<.0001*

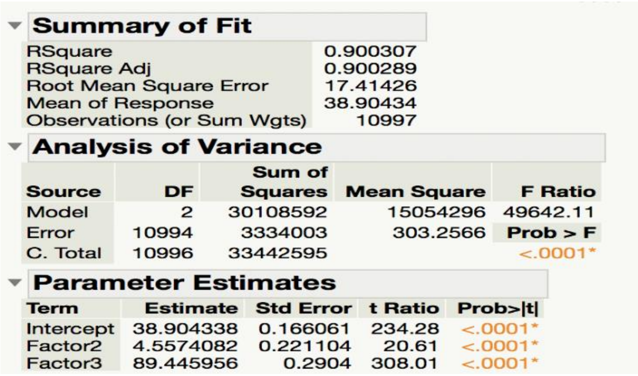
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38.904338	0.166056	234.28	<.0001*
Factor1	-0.218974	0.166936	-1.31	0.1896
Factor2	4.5703338	0.221316	20.65	<.0001*
Factor3	89.45818	0.29054	307.90	<.0001*

Model 3:

Arrival Delay = 38.90 + 4.55 * TaxiIn +89.45 * Departure Delay

Even though, this model is adequate, **Model 2 is the best model** to predict the arrival delay. It has the highest RSquare value **0.91** with less RMSE value of **16.34** compared to model 3.



Summary of Fit

RSquare	0.900307
RSquare Adj	0.900289
Root Mean Square Error	17.41426
Mean of Response	38.90434
Observations (or Sum Wgts)	10997

Analysis of Variance

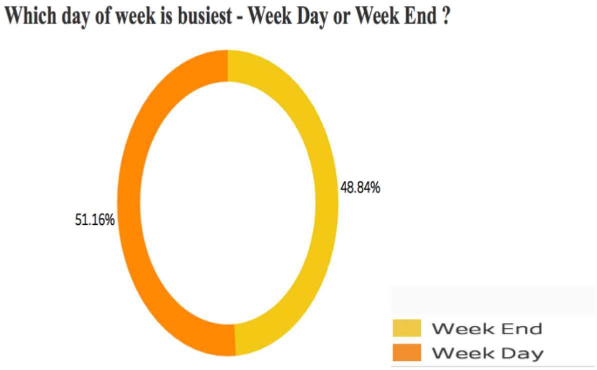
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	30108592	15054296	49642.11
Error	10994	3334003	303.2566	Prob > F
C. Total	10996	33442595		<.0001*

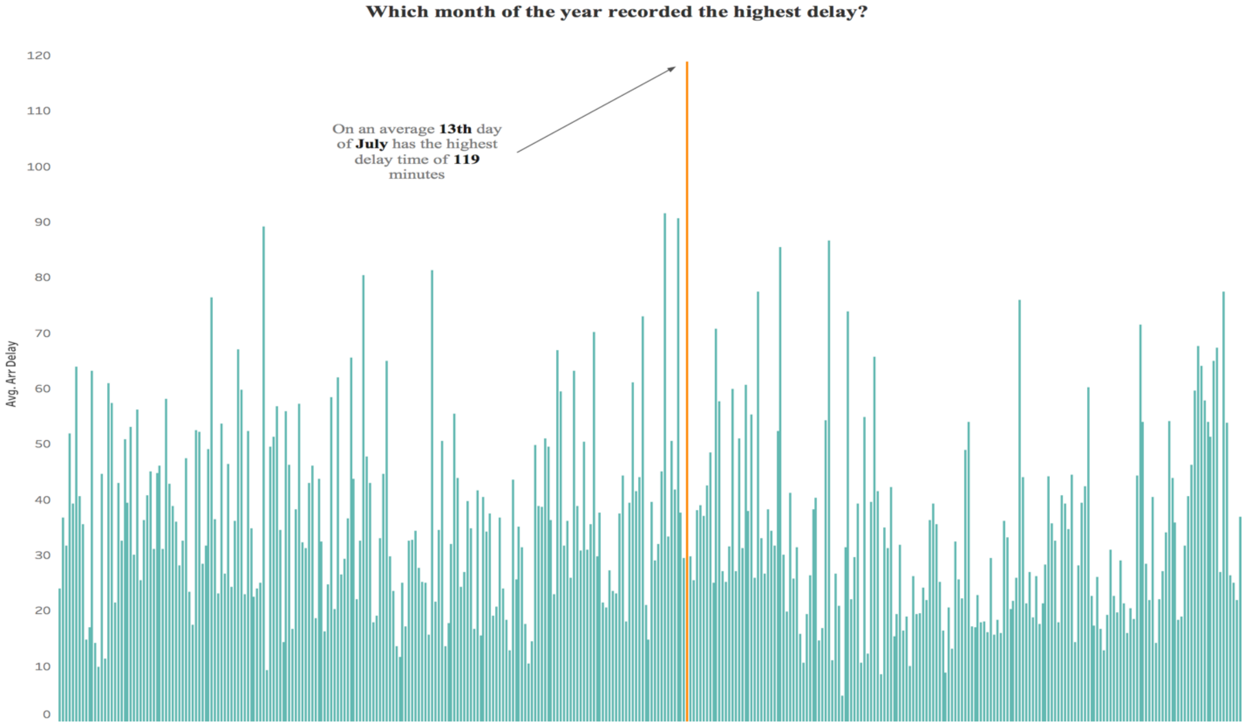
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38.904338	0.166061	234.28	<.0001*
Factor2	4.5574082	0.221104	20.61	<.0001*
Factor3	89.445956	0.2904	308.01	<.0001*

Which day of the month in a year has the highest delay time?

Exploring the data, we observed that arrival delay in weekday is more compared to weekend. As seen in the figure, weekday has **51.16 %** of flight delay compared to 48.84% in weekend. However, this difference is not significant. On an average, 13th day of July has the highest delay time of 119 minutes in the year 2008.



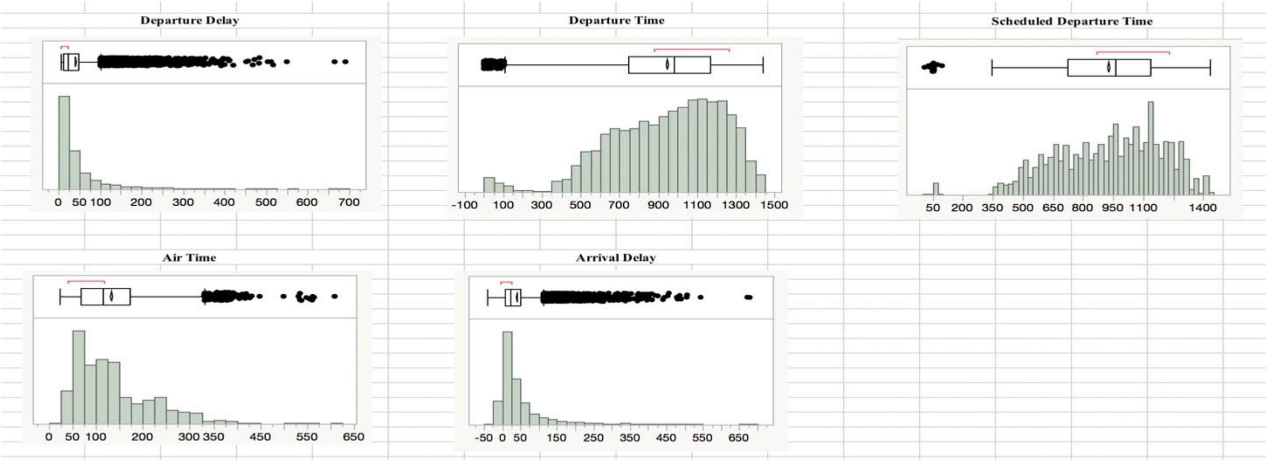


Appendix

A.1 Variable Description

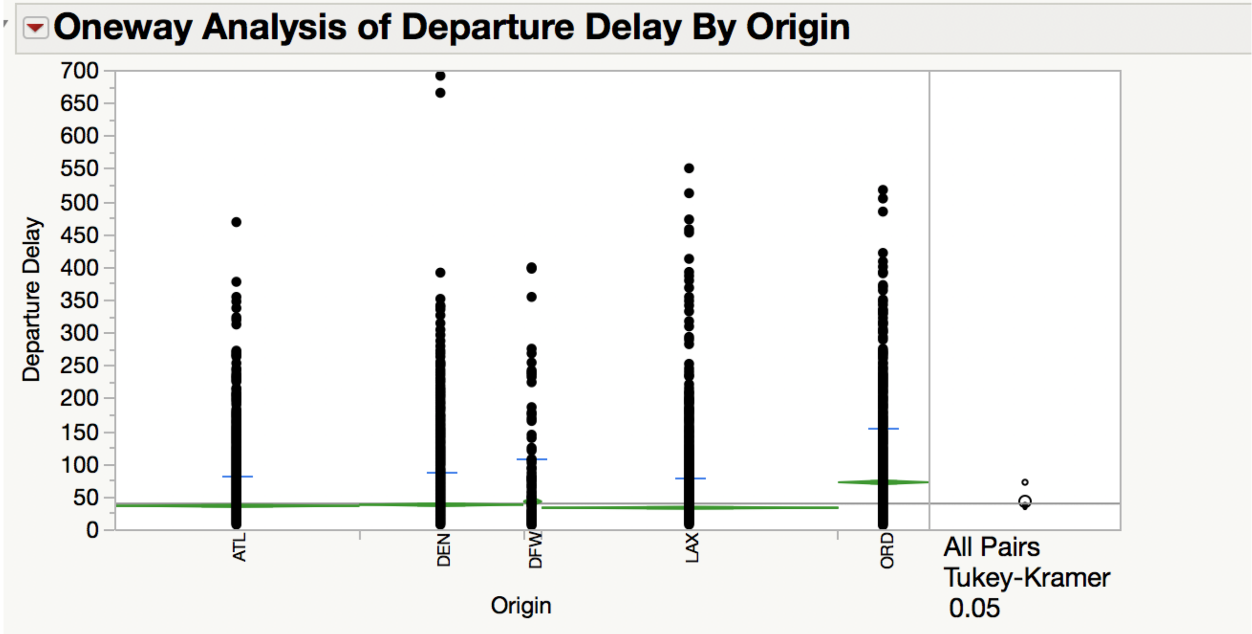
Variable descriptions		
Sl. No	**Name**	**Description**
1	Year	2008
2	Month	12-Jan
3	DayofMonth	31-Jan
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes

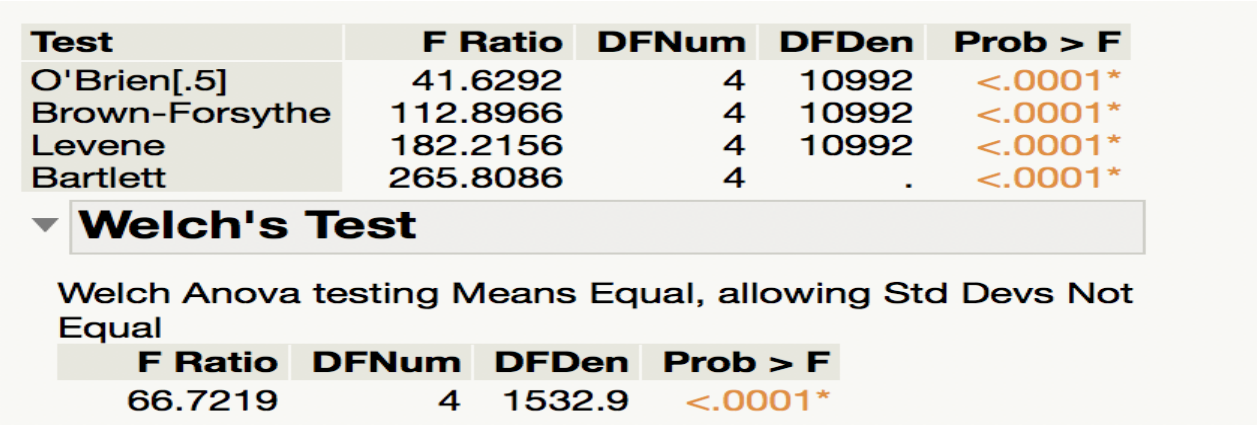
A.2 Histogram Plot



A.3 Analysis of Variance

One-way analysis of departure delay by Origin





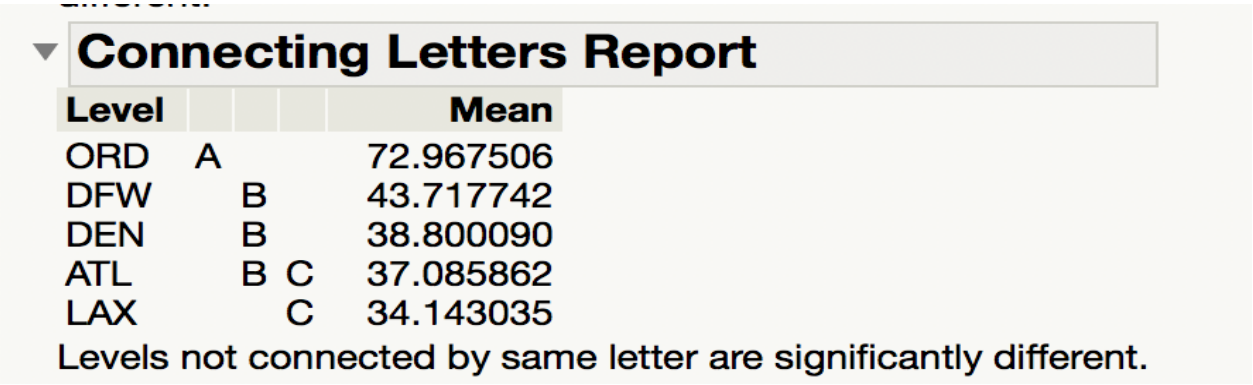
Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	41.6292	4	10992	<.0001*
Brown-Forsythe	112.8966	4	10992	<.0001*
Levene	182.2156	4	10992	<.0001*
Bartlett	265.8086	4	.	<.0001*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
66.7219	4	1532.9	<.0001*

different.



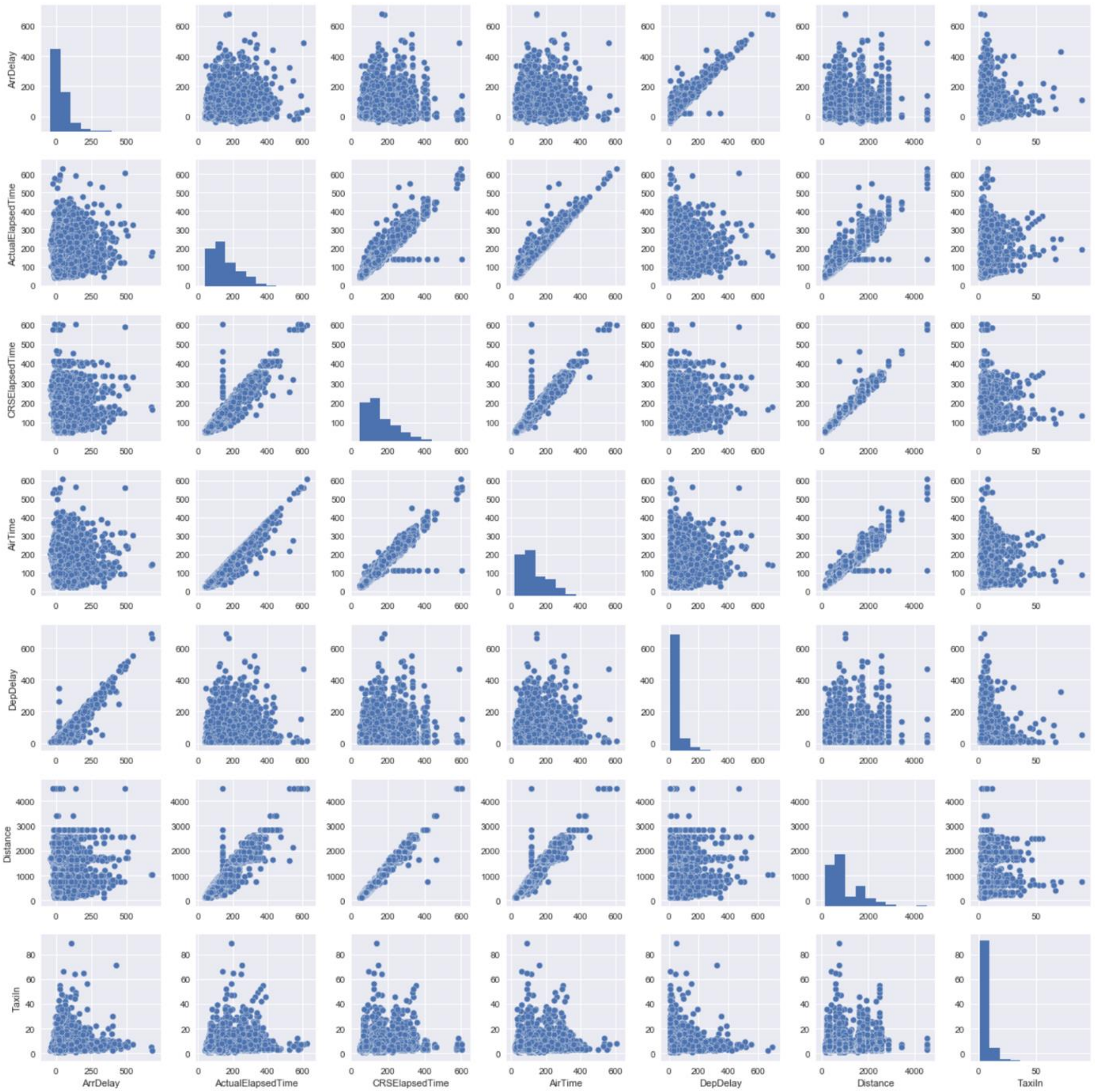
Connecting Letters Report

Level				Mean
ORD	A			72.967506
DFW		B		43.717742
DEN		B		38.800090
ATL		B	C	37.085862
LAX			C	34.143035

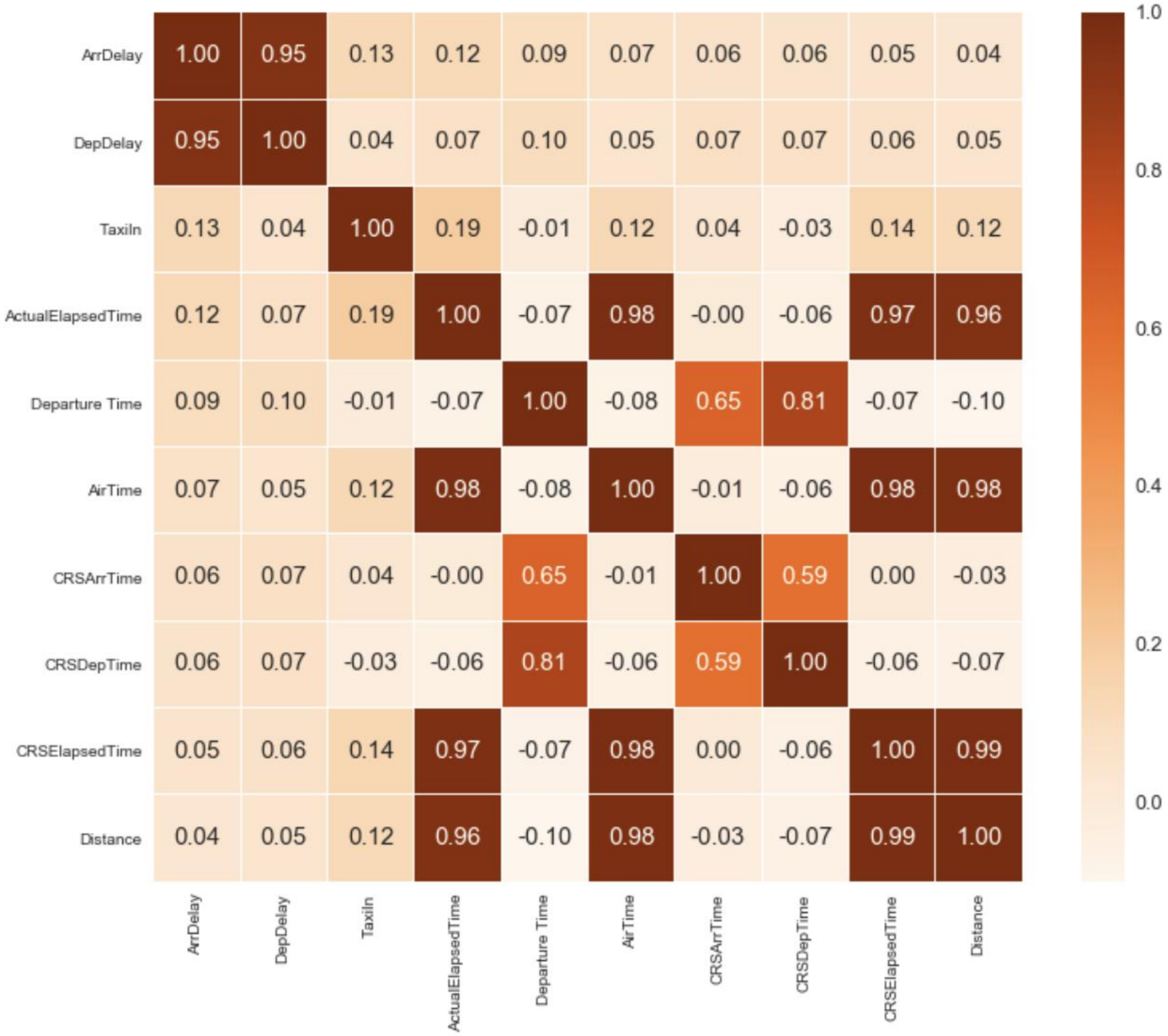
Levels not connected by same letter are significantly different.

A.4 Identifying the correlation between the variables

Pair Plot

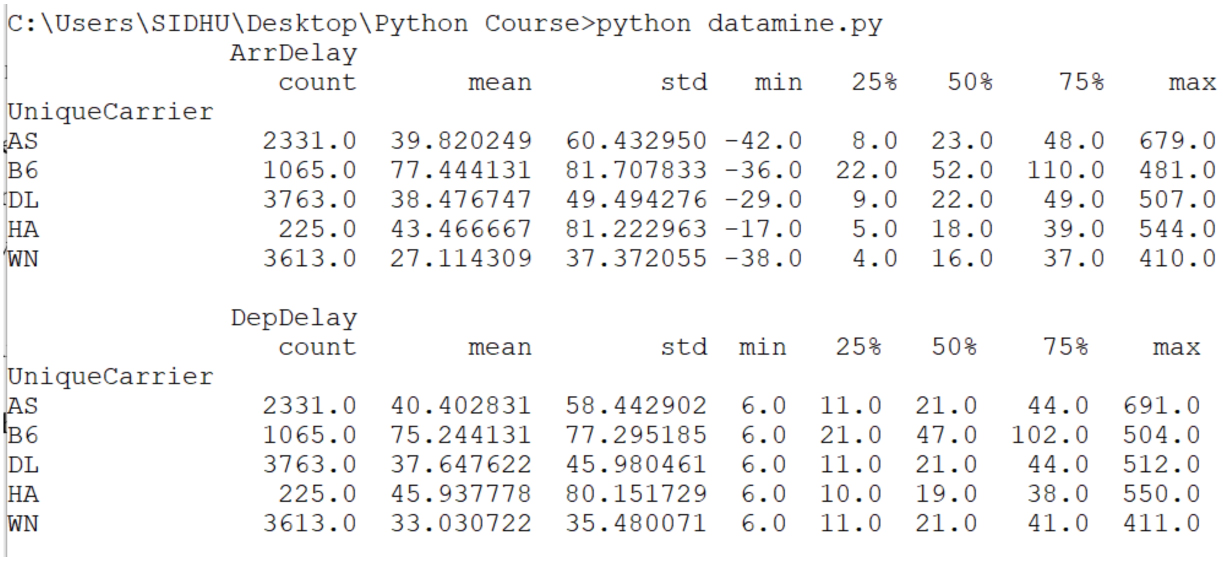


Correlation Matrix



	ArrDelay	DepDelay	TaxiIn	ActualElapsedTime	Departure Time	AirTime	CRSArrTime	CRSDepTime	CRSElapsedTime	Distance
ArrDelay	1.00	0.95	0.13	0.12	0.09	0.07	0.06	0.06	0.05	0.04
DepDelay	0.95	1.00	0.04	0.07	0.10	0.05	0.07	0.07	0.06	0.05
TaxiIn	0.13	0.04	1.00	0.19	-0.01	0.12	0.04	-0.03	0.14	0.12
ActualElapsedTime	0.12	0.07	0.19	1.00	-0.07	0.98	-0.00	-0.06	0.97	0.96
Departure Time	0.09	0.10	-0.01	-0.07	1.00	-0.08	0.65	0.81	-0.07	-0.10
AirTime	0.07	0.05	0.12	0.98	-0.08	1.00	-0.01	-0.06	0.98	0.98
CRSArrTime	0.06	0.07	0.04	-0.00	0.65	-0.01	1.00	0.59	0.00	-0.03
CRSDepTime	0.06	0.07	-0.03	-0.06	0.81	-0.06	0.59	1.00	-0.06	-0.07
CRSElapsedTime	0.05	0.06	0.14	0.97	-0.07	0.98	0.00	-0.06	1.00	0.99
Distance	0.04	0.05	0.12	0.96	-0.10	0.98	-0.03	-0.07	0.99	1.00

Descriptive Analysis Comparing Airlines based on Departure Delay and Arrival Delay:

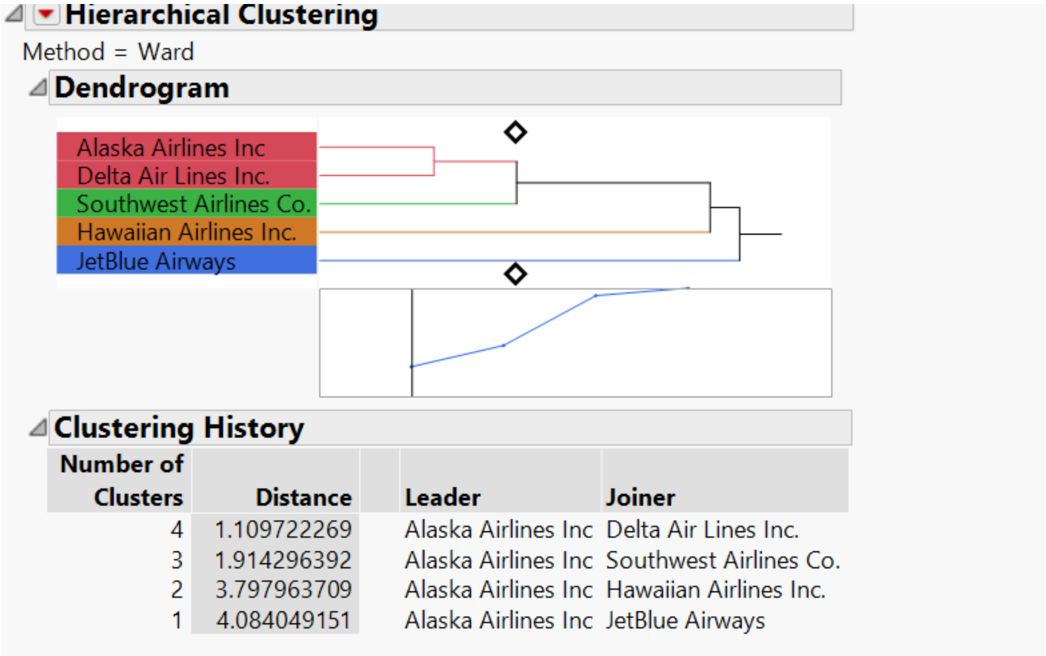


```
C:\Users\SIDHU\Desktop\Python Course>python datamine.py
              ArrDelay
                 count       mean        std   min   25%   50%    75%    max
UniqueCarrier
AS              2331.0  39.820249  60.432950 -42.0   8.0  23.0   48.0  679.0
B6              1065.0  77.444131  81.707833 -36.0  22.0  52.0  110.0  481.0
DL              3763.0  38.476747  49.494276 -29.0   9.0  22.0   49.0  507.0
HA               225.0  43.466667  81.222963 -17.0   5.0  18.0   39.0  544.0
WN              3613.0  27.114309  37.372055 -38.0   4.0  16.0   37.0  410.0

              DepDelay
                 count       mean        std  min   25%   50%    75%    max
UniqueCarrier
AS              2331.0  40.402831  58.442902  6.0  11.0  21.0   44.0  691.0
B6              1065.0  75.244131  77.295185  6.0  21.0  47.0  102.0  504.0
DL              3763.0  37.647622  45.980461  6.0  11.0  21.0   44.0  512.0
HA               225.0  45.937778  80.151729  6.0  10.0  19.0   38.0  550.0
WN              3613.0  33.030722  35.480071  6.0  11.0  21.0   41.0  411.0
```

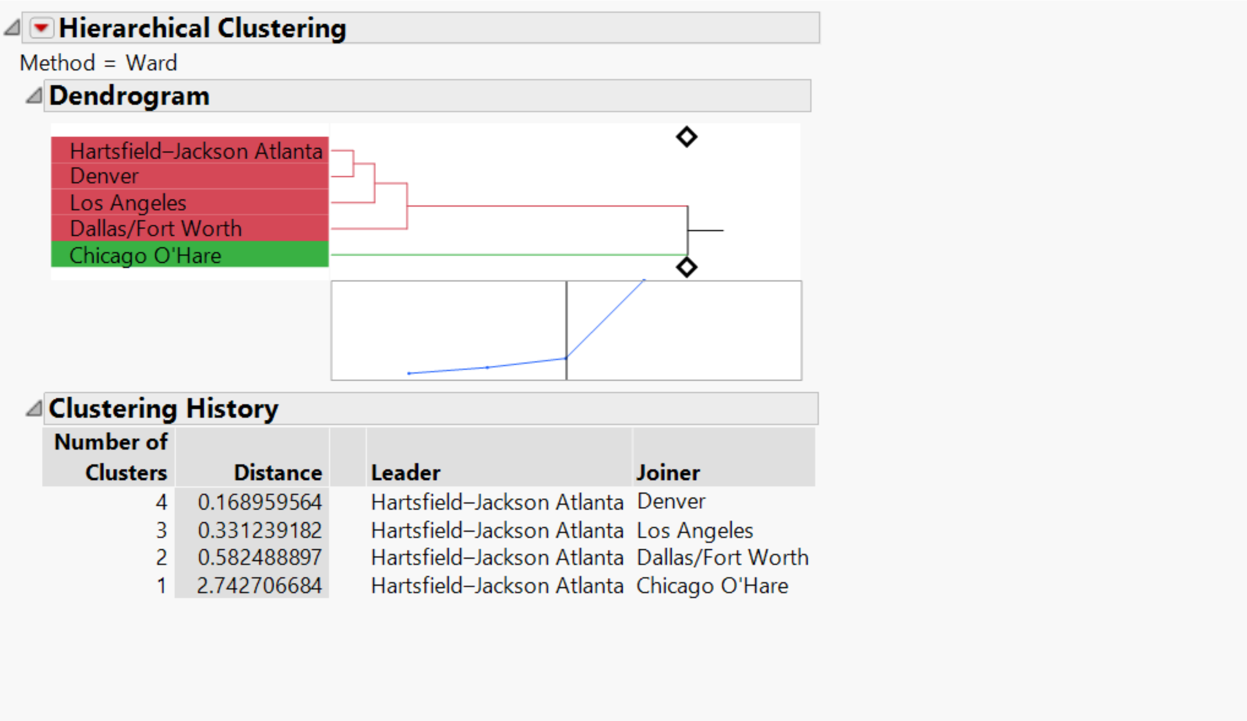
Clustering:

1.1 Comparison of Airlines



Number of Clusters	Distance	Leader	Joiner
4	1.109722269	Alaska Airlines Inc	Delta Air Lines Inc.
3	1.914296392	Alaska Airlines Inc	Southwest Airlines Co.
2	3.797963709	Alaska Airlines Inc	Hawaiian Airlines Inc.
1	4.084049151	Alaska Airlines Inc	JetBlue Airways

1.2 Comparison of Origin Airports based on Departure and Arrival Delay



Hierarchical Clustering

Method = Ward

Dendrogram

Clustering History

Number of Clusters	Distance	Leader	Joiner
4	0.168959564	Hartsfield–Jackson Atlanta	Denver
3	0.331239182	Hartsfield–Jackson Atlanta	Los Angeles
2	0.582488897	Hartsfield–Jackson Atlanta	Dallas/Fort Worth
1	2.742706684	Hartsfield–Jackson Atlanta	Chicago O'Hare