AIRLINE OPERATIONS

BUSINESS ANALYTICS TOOLS 2 - Final Project

DePaul University - Kellstadt Graduate School of Business

Aryan Wagh

Chanukya Bolli

Sowmya Ram Erni

MAIN BUSINESS IDEA OR MOTIVATION

Flight delays stand as a formidable hurdle, significantly impacting the punctuality of airlines. They wield substantial influence, often driving down the on-time performance of carriers. Yet, within this challenge lies an opportunity for airlines to revolutionize their operations.

By meticulously pinpointing patterns inherent in delays, airlines can tailor strategic approaches to mitigate disruptions. The key lies in predictive prowess—a capability to anticipate the likelihood of delays before they unfold. This foresight empowers proactive interventions, enabling airlines to delve into the intricate web of delay causes.

From scrutinizing maintenance issues to dissecting the complexities of weather dynamics and unpunctual arrivals of aircraft, a comprehensive analysis unveils avenues for targeted solutions. Through these measures, airlines can elevate their operational efficiency, fostering a smoother journey for passengers and augmenting overall satisfaction.

INTRODUCTION OF TOPIC & DATA

Our research initiative focuses on 'Airline On-Time Performance and Flight Delay Causes,' leveraging the comprehensive dataset provided by the Bureau of Transportation Statistics (BTS). This extensive dataset is a repository of information encompassing airline punctuality metrics and the underlying causes of flight delays, spanning from January 2004 to December 2019.

Within this dataset, a wealth of intricate details are encapsulated, offering a comprehensive panorama of the aviation landscape. It catalogs crucial elements such as chronological markers including the year and month, specifics pertaining to individual airlines and airports, comprehensive flight counts, a description of delayed flights, and a breakdown of the diverse factors attributing to these

instances of delay. These factors are systematically classified into five principal categories: Air Carrier, Extreme Weather, National Aviation System (NAS), Late-Arriving Aircraft, and Security, providing a structured framework to discern the multifaceted roots of flight disruptions.

The depth and breadth of this dataset render it an invaluable asset for discerning trends and patterns inherent in flight delays within the U.S. aviation framework. Its analytical utility extends to benefit airlines, airports, policymakers, and travelers, offering nuanced insights to comprehend and strategically address the contributing factors behind flight delays. By leveraging this dataset, stakeholders can gain a profound understanding of the dynamics shaping delays, fostering informed decision-making and facilitating initiatives aimed at enhancing the efficiency and reliability of the U.S. aviation system.

RESEARCH QUESTION

How do the delay factors interact with each other? Are there synergies or dependencies that, when addressed together, could lead to more substantial improvements in on-time performance?

Procedure to solve the Research Question

We Explored delay factor interactions using descriptive analytics (mean, distribution, correlation, ANOVA) to identify patterns. Applied predictive analytics, including clustering and regression models, to understand group-specific influences on on-time performance. Utilized machine learning techniques for nuanced insights into nonlinear relationships and potential synergies among delay factors. Nac_ct and carrier_ct were the most significant variables in predicting the delay in flights, i.e more than 30 minutes.

DATA & EMPIRICAL METHODOLOGY

The dataset `flight_delay_2004_2019.csv` was taken from https://bigblue.depaul.edu/jlee141/econdata/BTS/ website.

Data Source: https://bigblue.depaul.edu/jlee141/econdata/BTS/flight_delay_2004_2019.csv

About Dataset
The data used in this analysis consists of airline on-time performance and causes of flight delays. The dataset spans from January 2004 to December 2019 and includes various variables such as the year, month, airline information, airport details, counts of delayed and canceled flights, and reasons for delays categorized into different factors like weather, National Aviation System (NAS), security, late-arriving aircraft, and carrier-related issues.

The primary variables of interest for this analysis include `nas_ct`, `_weather_ct`, `security_ct`, `late_aircraft_ct`, and `carrier_ct`. These variables represent counts of delays attributed to the National Aviation System, weather conditions, security issues, late-arriving aircraft, and the airline carrier, respectively.

The estimating equation(s) for this analysis will be based on a multivariate regression model, aiming to understand the relationship between the counts of delays in the specified categories and the potential influencing factors.

An estimated regression equation given by Y = b0 + b1x1 + b2X^2 + e.

Method

The methodology of using a multivariate regression model allows for a comprehensive analysis of the impact of various factors on different types of delays. This approach provides a quantitative framework to understand relationships, assess significance, and make predictions based on historical data. It offers a more nuanced understanding compared to univariate analyses and helps in identifying critical areas for improvement in airline operations and management.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
22	VAR22	Char	1	$1.	$1.
16	_arr_delay	Num	8	BEST12.	BEST32.
17	_carrier_delay	Num	8	BEST12.	BEST32.
2	_month	Num	8	BEST12.	BEST32.
10	_weather_ct	Num	8	BEST12.	BEST32.
5	airport	Char	5	$5.	$5.
6	airport_name	Char	58	$58.	$58.
14	arr_cancelled	Num	8	BEST12.	BEST32.
8	arr_del15	Num	8	BEST12.	BEST32.
15	arr_diverted	Num	8	BEST12.	BEST32.
7	arr_flights	Num	8	BEST12.	BEST32.
3	carrier	Char	4	$4.	$4.
9	carrier_ct	Num	8	BEST12.	BEST32.
4	carrier_name	Char	22	$22.	$22.
13	late_aircraft_ct	Num	8	BEST12.	BEST32.
21	late_aircraft_delay	Num	8	BEST12.	BEST32.
11	nas_ct	Num	8	BEST12.	BEST32.
19	nas_delay	Num	8	BEST12.	BEST32.
12	security_ct	Num	8	BEST12.	BEST32.
20	security_delay	Num	8	BEST12.	BEST32.
18	weather_delay	Num	8	BEST12.	BEST32.
1	year	Num	8	BEST12.	BEST32.

RESULTS

Descriptive statistics

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
year	265047	2011.34	4.7312526	2004.00	2019.00
_month	265047	6.5074043	3.4490732	1.0000000	12.0000000
arr_flights	264681	396.3093309	1054.48	1.0000000	21977.00
arr_del15	264625	78.0685045	207.9696510	0	6377.00
carrier_ct	264681	21.9451040	48.1662535	0	1792.07
_weather_ct	264681	2.7583871	10.4491942	0	717.9400000
nas_ct	264681	25.9388784	89.0688297	-0.0100000	4091.27
security_ct	264681	0.1772067	0.8294422	0	80.5600000
late_aircraft_ct	264681	27.2324723	79.5165030	0	1885.47
arr_cancelled	264681	6.7707051	28.4278117	0	1969.00
arr_diverted	264681	0.9181468	4.0941401	0	256.0000000
_arr_delay	264681	4481.77	13036.94	0	433687.00
_carrier_delay	264681	1322.86	3493.39	0	196944.00
weather_delay	264681	228.8352885	881.9452197	0	57707.00
nas_delay	264681	1196.08	4893.63	-1.0000000	238440.00
security_delay	264681	7.0259444	37.1587004	0	3194.00
late_aircraft_delay	264681	1726.97	5167.39	0	148181.00

The table summarizes flight delay data between 2004 and 2019. It shows on average, there were about 400 arrival flights including approximately 78 delayed flights per airport per month. The most common causes of delays were:

Late-arriving aircraft (average flights delayed 1727 and delayed by 27.5 minutes)

National Aviation System (average delay of 1196 minutes)

Carrier-related issues (average delay of 1323 minutes)

Weather-related delays were less frequent but could still be significant (average

of 229 minutes).

Cancellations and diversions were relatively rare events (average of 6 and 1 per airport per month, respectively).

The SUMMARY Procedure

Variable	Mean	Median	Minimum	Maximum
_arr_delay	4481.77	1330.00	0	433687.00
_carrier_delay	1322.86	476.0000000	0	196944.00

The UNIVARIATE Procedure
Variable: arr_del15

Moments			
N	264625	Sum Weights	264625
Mean	78.0685045	Sum Observations	20658878
Std Deviation	207.969651	Variance	43251.3757
Skewness	7.94454658	Kurtosis	95.3810816
Uncorrected SS	1.30582E10	Corrected SS	1.14454E10
Coeff Variation	266.393794	Std Error Mean	0.40428211

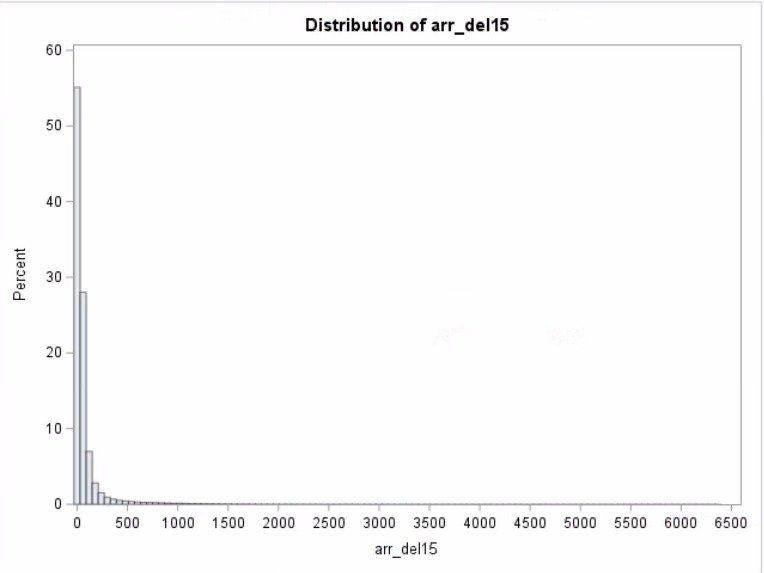
Quantiles (Definition 5)	
Level	Quantile
100% Max	6377
99%	1051
95%	303
90%	148
75% Q3	60
50% Median	25
25% Q1	10
10%	4
5%	2
1%	0
0% Min	0

Basic Statistical Measures			
Location		Variability	
Mean	78.06850	Std Deviation	207.96965
Median	25.00000	Variance	43251
Mode	5.00000	Range	6377
		Interquartile Range	50.00000

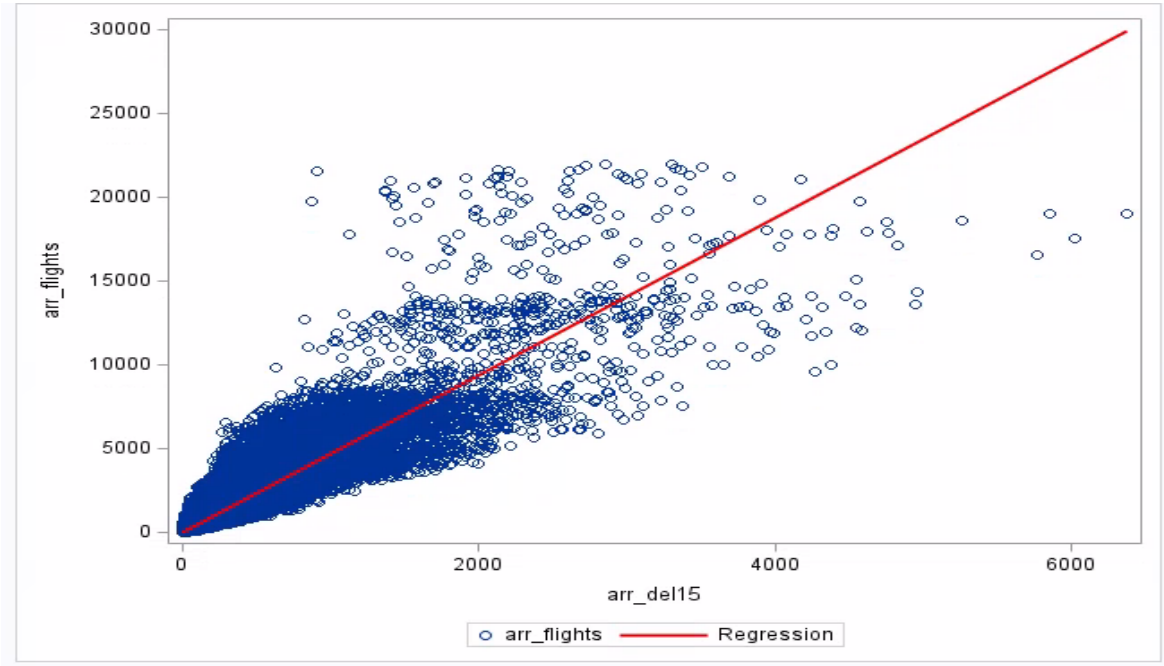
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	265031	5268	182771
0	265030	5778	1321
0	264953	5862	204162
0	264909	6029	6829
0	264882	6377	149221

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	193.104	Pr > |t|	<.0001
Sign	M	129534	Pr >= |M|	<.0001

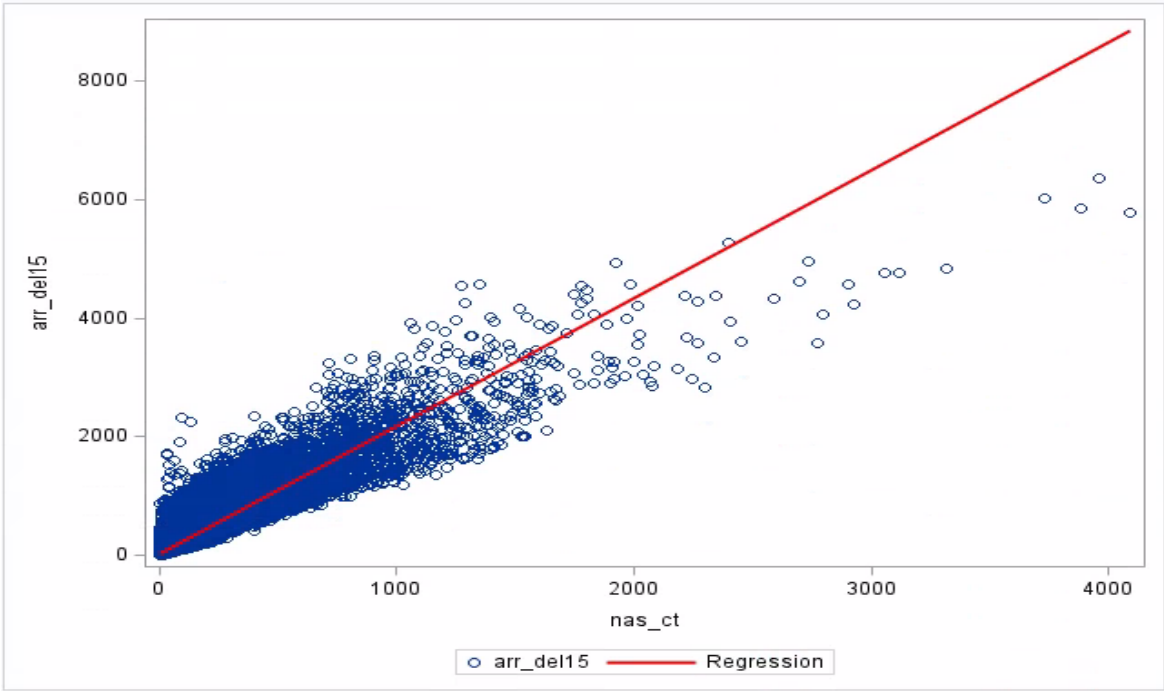
Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	422	0.16	100.00

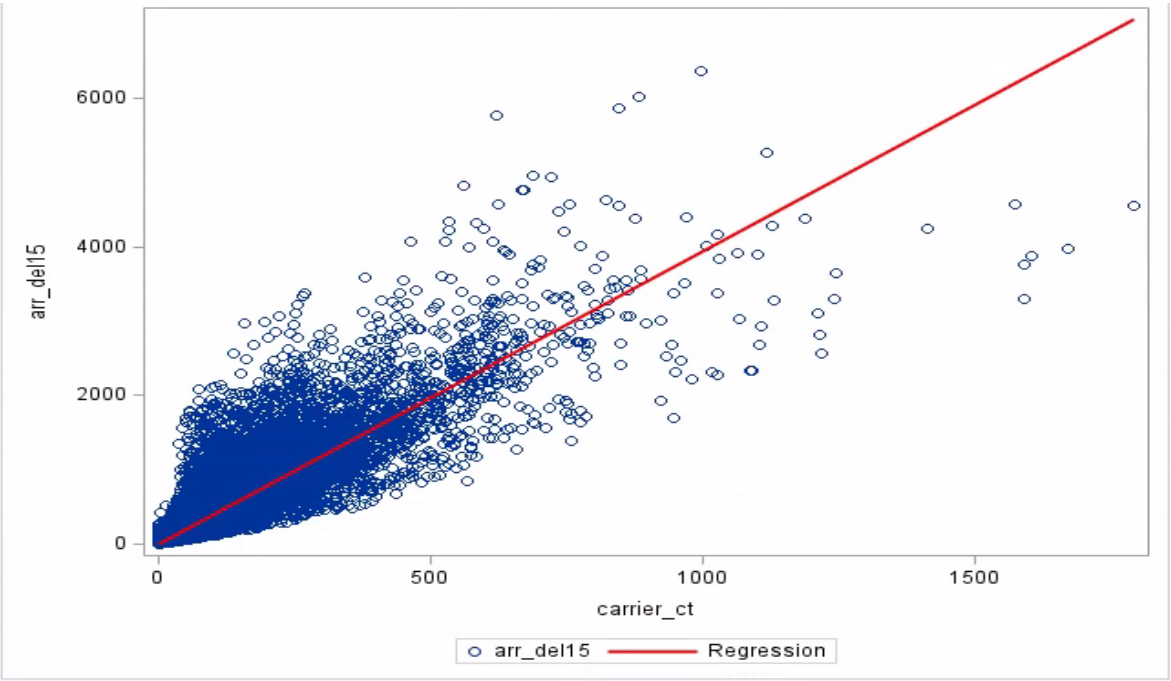


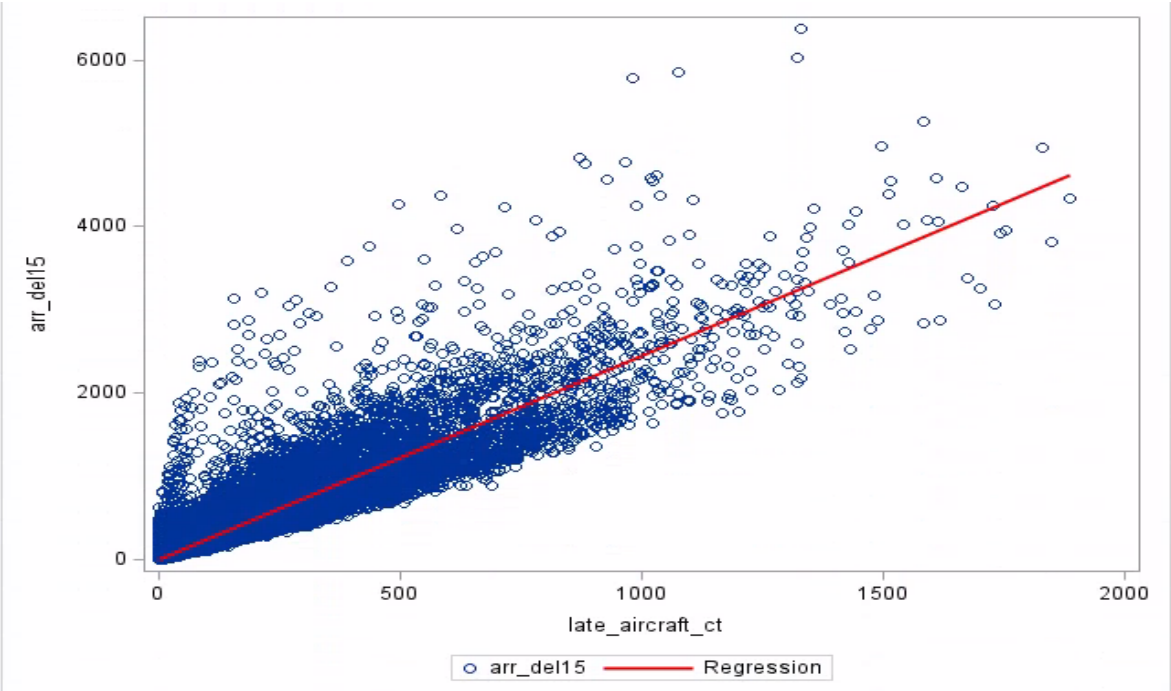
Highlight Skewed Skewed Distribution: The skewness and kurtosis values indicate a significant positive skewness and a heavy-tailed distribution,suggesting the presence of outliers and potentially extreme delays, which is an essential focus area for this study.

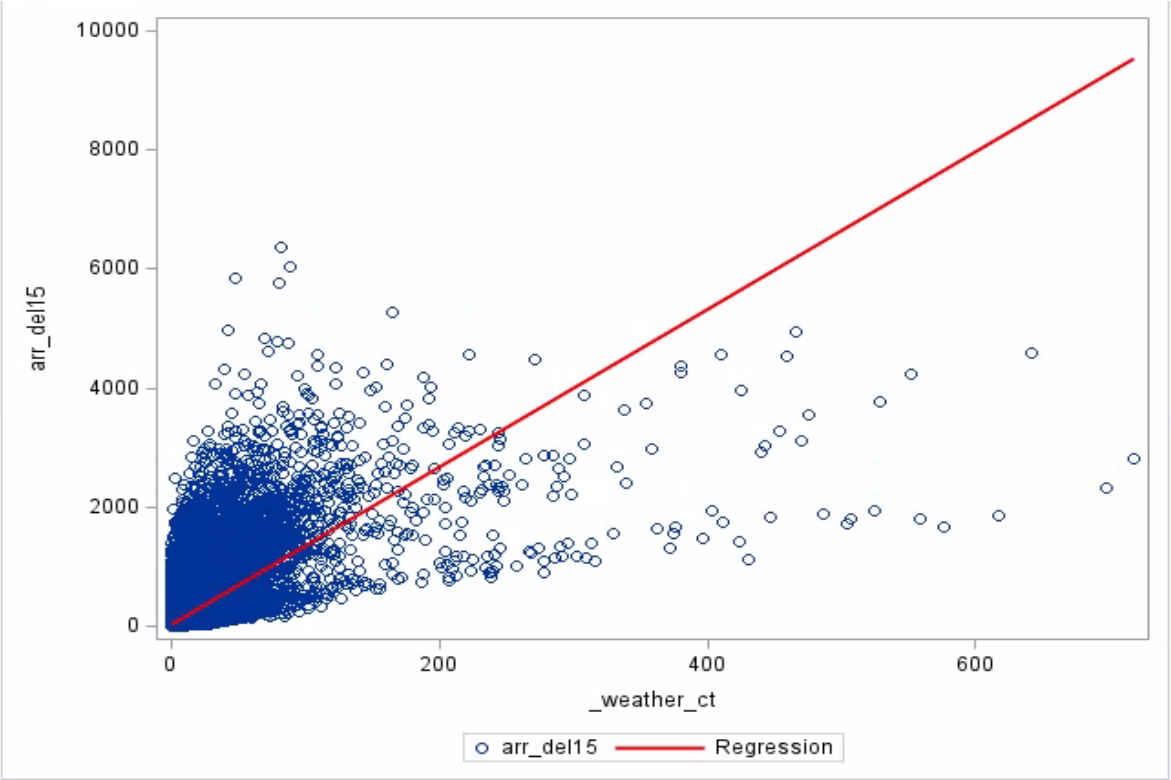


The plot shows a positive correlation between arr_flights and arr_del15, which means that as the number of flights arriving at an airport increases, the number of flights delayed by more than 15 minutes also increases. This is likely because more flights means more congestion at the airport, which can lead to delays.




arr_del15
6000
4000
2000
0
0
500
1000
1500
carrier_ct
arr_del15
Regression


arr_del15
6000
4000
2000
0
0
500
1000
1500
2000
late_aircraft_ct
arr_del15
Regression

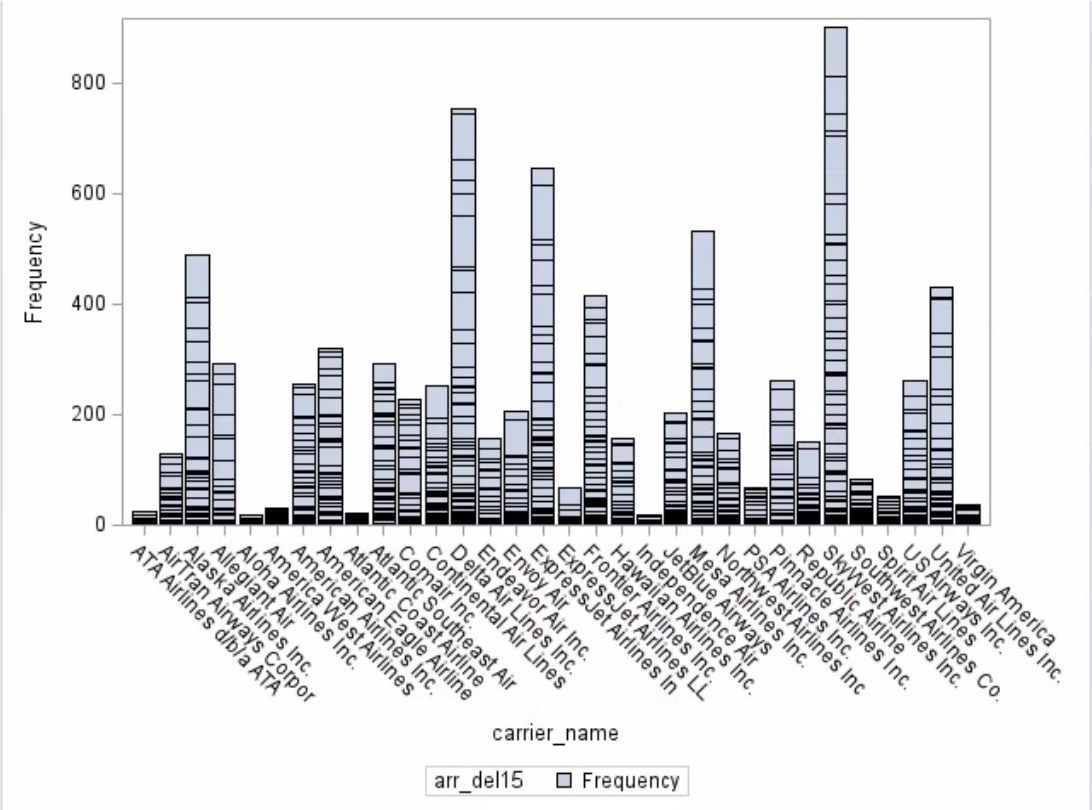


When exploring the relationship between the arrival delay indicator (arr_del15) and various contributing factors to flight delays such as late aircraft, carrier issues, and the national aviation system, observable patterns emerged. Plots depicting the relationship between arr_del15 and these factors demonstrated a noticeable linear correlation.

However, upon examining the relationship between arr_del15 and weather-related factors, the plot presented a distinct contrast. Unlike the other delay-contributing variables, the plot between arr_del15 and weather did not exhibit a robust or pronounced relationship. The lack of a strong correlation between these variables indicates that changes or variations in weather conditions might not have a direct or linear impact on the likelihood of arrival delays as compared to factors like late aircraft, carrier issues, or challenges within the national aviation system.

We will further explore and explain this in Regression Analysis.

Which Carrier has the highest delay time?



Skywest Airlines has the highest number of delay time period.

CORRELATION PROCEDURE

The CORR Procedure

8 Variables:	arr_del15 nas_ct _weather_ct security_ct late_aircraft_ct carrier_ct _month delayed

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
arr_del15	264625	78.06850	207.96965	20658878	0	6377
nas_ct	264681	25.93888	89.06883	6865528	-0.01000	4091
_weather_ct	264681	2.75839	10.44919	730093	0	717.94000
security_ct	264681	0.17721	0.82944	46903	0	80.56000
late_aircraft_ct	264681	27.23247	79.51650	7207918	0	1885
carrier_ct	264681	21.94510	48.16625	5808452	0	1792
_month	265047	6.50740	3.44907	1724768	1.00000	12.00000
delayed	265047	0.43780	0.49612	116038	0	1.00000

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations								
	arr_del15	nas_ct	_weather_ct	security_ct	late_aircraft_ct	carrier_ct	_month	delayed
arr_del15	1.00000 264625	0.92493 <.0001 264625	0.66434 <.0001 264625	0.49022 <.0001 264625	0.93256 <.0001 264625	0.91486 <.0001 264625	-0.00480 0.0136 264625	0.35475 <.0001 264625
nas_ct	0.92493 <.0001 264625	1.00000 264681	0.58070 <.0001 264681	0.38756 <.0001 264681	0.76117 <.0001 264681	0.75481 <.0001 264681	-0.00657 0.0007 264681	0.28456 <.0001 264681
_weather_ct	0.66434 <.0001 264625	0.58070 <.0001 264681	1.00000 264681	0.33412 <.0001 264681	0.53019 <.0001 264681	0.69641 <.0001 264681	-0.01720 <.0001 264681	0.23441 <.0001 264681
security_ct	0.49022 <.0001 264625	0.38756 <.0001 264681	0.33412 <.0001 264681	1.00000 264681	0.48294 <.0001 264681	0.51282 <.0001 264681	0.00676 0.0005 264681	0.20218 <.0001 264681
late_aircraft_ct	0.93256 <.0001 264625	0.76117 <.0001 264681	0.53019 <.0001 264681	0.48294 <.0001 264681	1.00000 264681	0.84444 <.0001 264681	-0.00245 0.2072 264681	0.33476 <.0001 264681
carrier_ct	0.91486 <.0001 264625	0.75481 <.0001 264681	0.69641 <.0001 264681	0.51282 <.0001 264681	0.84444 <.0001 264681	1.00000 264681	-0.00072 0.7093 264681	0.39857 <.0001 264681
_month	-0.00480 0.0136 264625	-0.00657 0.0007 264681	-0.01720 <.0001 264681	0.00676 0.0005 264681	-0.00245 0.2072 264681	-0.00072 0.7093 264681	1.00000 265047	-0.01651 <.0001 265047
delayed	0.35475 <.0001 264625	0.28456 <.0001 264681	0.23441 <.0001 264681	0.20218 <.0001 264681	0.33476 <.0001 264681	0.39857 <.0001 264681	-0.01651 <.0001 265047	1.00000 265047

interpretation:

Anova Procedure

As we went further with our research, we found that the variable Delayed (delayed by 30 min) is more suitable over the variable arr_del15 (delayed by 15 min), and gives us a better fitting model and result.

The ANOVA Procedure

Dependent Variable: delayed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17299	47254.87202	2.73165	37.73	<.0001
Error	247381	17911.25790	0.07240		
Corrected Total	264680	65166.12992			

R-Square	Coeff Var	Root MSE	delayed Mean
0.725145	61.37654	0.269079	0.438407

Source	DF	Anova SS	Mean Square	F Value	Pr > F
carrier_ct	17299	47254.87202	2.73165	37.73	<.0001

The ANOVA Procedure

Dependent Variable: delayed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20030	41403.86102	2.06709	21.28	<.0001
Error	244650	23762.26891	0.09713		
Corrected Total	264680	65166.12992			

R-Square	Coeff Var	Root MSE	delayed Mean
0.635359	71.08760	0.311653	0.438407

Source	DF	Anova SS	Mean Square	F Value	Pr > F
nas_ct	20030	41403.86102	2.06709	21.28	<.0001

The ANOVA Procedure

Dependent Variable: delayed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1070	6640.46724	6.20604	27.95	<.0001
Error	263610	58525.66268	0.22202		
Corrected Total	264680	65166.12992			

R-Square	Coeff Var	Root MSE	delayed Mean
0.101901	107.4768	0.471186	0.438407

Source	DF	Anova SS	Mean Square	F Value	Pr > F
security_ct	1070	6640.467240	6.206044	27.95	<.0001

The ANOVA Procedure

Dependent Variable: delayed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21228	42604.95162	2.00702	21.66	<.0001
Error	243452	22561.17830	0.09267		
Corrected Total	264680	65166.12992			

R-Square	Coeff Var	Root MSE	delayed Mean
0.653790	69.43793	0.304421	0.438407

Source	DF	Anova SS	Mean Square	F Value	Pr > F
late_aircraft_ct	21228	42604.95162	2.00702	21.66	<.0001

The ANOVA Procedure

Dependent Variable: delayed

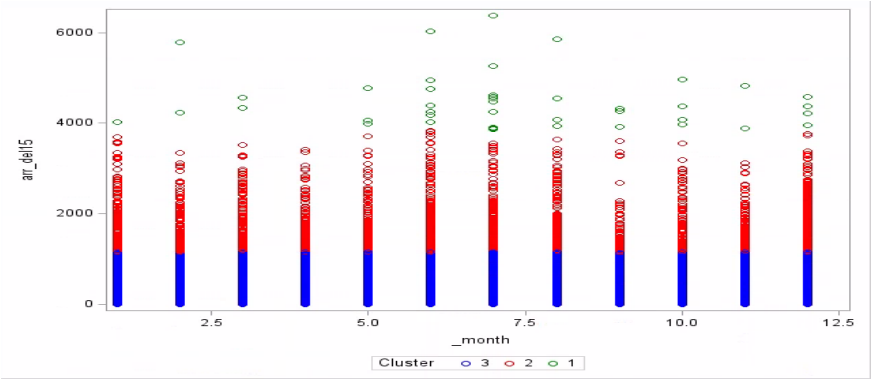
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5346	20329.22368	3.80270	21.99	<.0001
Error	259334	44836.90624	0.17289		
Corrected Total	264680	65166.12992			

R-Square	Coeff Var	Root MSE	delayed Mean
0.311960	94.84417	0.415803	0.438407

Source	DF	Anova SS	Mean Square	F Value	Pr > F
_weather_ct	5346	20329.22368	3.80270	21.99	<.0001

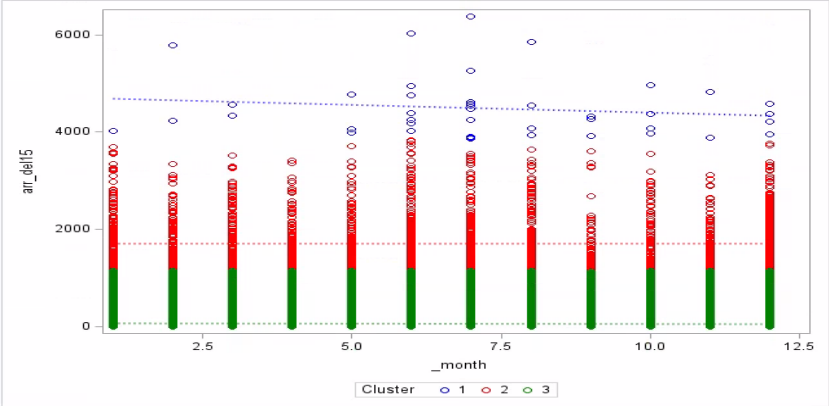
Non–hierarchical clustering

Simple Graph using K-means



The graph shows the number of clusters of flights by month obtained using K-means clustering. The number of clusters is highest in the summer months (June, July, and August), which suggests that there is more variability in flight delays during these months.

Graph with variable relationship



The line graph shows that the number of clusters of flights by month obtained using K-means clustering is generally increasing over time, with a peak in the summer months. This suggests that there is more variability in flight delays during the summer months, which may be due to factors such as increased travel demand and severe weather events.

Simple Regression:

The REG Procedure
Model: MODEL1
Dependent Variable: delayed

Number of Observations Read	265047
Number of Observations Used	264681
Number of Observations with Missing Values	366

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5276.69789	5276.69789	23320.2	<.0001
Error	264679	59889	0.22627		
Corrected Total	264680	65166			

Root MSE	0.47568	R-Square	0.0810
Dependent Mean	0.43841	Adj R-Sq	0.0810
Coeff Var	108.50204		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.39729	0.00096301	412.55	<.0001
nas_ct	1	0.00159	0.00001038	152.71	<.0001

- The estimated value of delayed is 0.39729 when nas_ct is zero.
- The estimated change in delayed for a one-unit increase in nas_ct is 0.00159.

The regression model explains a statistically significant amount of the variance in the delayed variable. The positive coefficient for nas_ct suggests that an increase in nas_ct is associated with an increase in the probability of a flight being delayed by more than 30 minutes. The F-statistic indicates that the overall model is significant.

After Splitting the Data

The FREQ Procedure

Selection Indicator				
Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	79514	30.00	79514	30.00
1	185533	70.00	265047	100.00

%let indep_var = carrier_ct _weather_ct nas_ct security_ct late_aircraft_ct _month

The REG Procedure
Model: MODEL1
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read	265047
Number of Observations Used	185264
Number of Observations with Missing Values	79783

Number in Model	Adjusted R-Square	R-Square	Variables in Model
6	0.1640	0.1641	carrier_ct _weather_ct nas_ct security_ct late_aircraft_ct _month
5	0.1640	0.1640	carrier_ct _weather_ct nas_ct late_aircraft_ct _month
5	0.1640	0.1640	carrier_ct _weather_ct nas_ct security_ct _month
4	0.1640	0.1640	carrier_ct _weather_ct nas_ct _month
4	0.1637	0.1637	carrier_ct _weather_ct late_aircraft_ct _month
5	0.1637	0.1637	carrier_ct _weather_ct security_ct late_aircraft_ct _month
5	0.1637	0.1637	carrier_ct _weather_ct nas_ct security_ct late_aircraft_ct
4	0.1637	0.1637	carrier_ct _weather_ct nas_ct late_aircraft_ct
4	0.1637	0.1637	carrier_ct _weather_ct nas_ct security_ct
3	0.1636	0.1637	carrier_ct _weather_ct nas_ct
4	0.1635	0.1636	carrier_ct _weather_ct security_ct _month
3	0.1635	0.1635	carrier_ct _weather_ct _month
3	0.1634	0.1634	carrier_ct _weather_ct late_aircraft_ct
4	0.1634	0.1634	carrier_ct _weather_ct security_ct late_aircraft_ct
3	0.1632	0.1632	carrier_ct _weather_ct security_ct
2	0.1632	0.1632	carrier_ct _weather_ct
4	0.1611	0.1611	carrier_ct nas_ct late_aircraft_ct _month
5	0.1611	0.1612	carrier_ct nas_ct security_ct late_aircraft_ct _month
3	0.1611	0.1611	carrier_ct nas_ct _month
4	0.1611	0.1611	carrier_ct nas_ct security_ct _month
4	0.1608	0.1609	carrier_ct nas_ct security_ct late_aircraft_ct
3	0.1608	0.1609	carrier_ct nas_ct late_aircraft_ct
2	0.1608	0.1608	carrier_ct nas_ct
4	0.1603	0.1604	carrier_ct security_ct late_aircraft_ct _month
2	0.1603	0.1603	carrier_ct _month
3	0.1603	0.1603	carrier_ct security_ct _month
2	0.1601	0.1601	carrier_ct late_aircraft_ct
3	0.1600	0.1601	carrier_ct security_ct late_aircraft_ct
1	0.1600	0.1600	carrier_ct
2	0.1600	0.1600	carrier_ct security_ct
5	0.1194	0.1194	_weather_ct nas_ct security_ct late_aircraft_ct _month
4	0.1191	0.1192	_weather_ct nas_ct security_ct late_aircraft_ct
4	0.1188	0.1188	_weather_ct security_ct late_aircraft_ct _month
3	0.1185	0.1185	_weather_ct security_ct late_aircraft_ct
4	0.1179	0.1179	_weather_ct nas_ct late_aircraft_ct _month
3	0.1176	0.1176	_weather_ct nas_ct late_aircraft_ct
3	0.1172	0.1172	_weather_ct late_aircraft_ct _month
2	0.1170	0.1170	_weather_ct late_aircraft_ct
4	0.1167	0.1168	nas_ct security_ct late_aircraft_ct _month
3	0.1165	0.1165	nas_ct security_ct late_aircraft_ct
3	0.1148	0.1148	nas_ct late_aircraft_ct _month
3	0.1148	0.1148	security_ct late_aircraft_ct _month
2	0.1145	0.1145	nas_ct late_aircraft_ct
2	0.1145	0.1145	security_ct late_aircraft_ct
2	0.1127	0.1127	late_aircraft_ct _month
1	0.1124	0.1124	late_aircraft_ct
4	0.0965	0.0966	_weather_ct nas_ct security_ct _month
3	0.0963	0.0963	_weather_ct nas_ct security_ct
3	0.0913	0.0913	nas_ct security_ct _month
2	0.0910	0.0911	nas_ct security_ct
3	0.0889	0.0889	_weather_ct nas_ct _month

2	0.0887	0.0887	_weather_ct nas_ct
2	0.0815	0.0816	nas_ct _month
1	0.0813	0.0813	nas_ct
3	0.0733	0.0733	_weather_ct security_ct _month
2	0.0731	0.0731	_weather_ct security_ct
2	0.0565	0.0566	_weather_ct _month
1	0.0564	0.0564	_weather_ct
2	0.0408	0.0408	security_ct _month
1	0.0405	0.0405	security_ct
1	0.0003	0.0003	_month

The output from PROC REG using the adjusted R-squared selection method provides information about different models based on the number of variables included and their corresponding adjusted R-squared values.

The table allows you to compare models with different combinations of variables based on their adjusted R-squared values.

Higher adjusted R-squared values generally indicate better-fitting models, but it's essential to consider the number of variables to avoid overfitting, hence we consider ‘delayed’ instead of ‘arr_del15’.

Stepwise Selection:

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	265047
Number of Observations Used	185264
Number of Observations with Missing Values	79783

Stepwise Selection: Step 1

Variable carrier_ct Entered: R-Square = 0.1600 and C(p) = 890.4430

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7299.42814	7299.42814	35293.5	<.0001
Error	185262	38316	0.20682		
Corrected Total	185263	45615			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.34768	0.00116	18517	89531.7	<.0001
carrier_ct	0.00414	0.00002201	7299.42814	35293.5	<.0001

Stepwise Selection: Step 2

Variable _weather_ct Entered: R-Square = 0.1632 and C(p) = 186.8148

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7444.67035	3722.33517	18066.2	<.0001
Error	185261	38171	0.20604		
Corrected Total	185263	45615			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.34572	0.00116	18235	88503.5	<.0001
carrier_ct	0.00470	0.00003056	4872.60070	23649.0	<.0001
_weather_ct	-0.00379	0.00014276	145.24221	704.93	<.0001

Stepwise Selection: Step 3

Variable nas_ct Entered: R-Square = 0.1637 and C(p) = 87.3434

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7465.55659	2488.51886	12084.5	<.0001
Error	185260	38150	0.20593		
Corrected Total	185263	45615			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.34490	0.00116	18062	87711.9	<.0001
carrier_ct	0.00493	0.00003828	3418.03113	16598.3	<.0001
_weather_ct	-0.00360	0.00014395	128.91292	626.01	<.0001
nas_ct	-0.00018574	0.00001844	20.88624	101.43	<.0001

Stepwise Selection: Step 4

Variable _month Entered: R-Square = 0.1640 and C(p) = 13.7377

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7481.11881	1870.27970	9085.93	<.0001
Error	185259	38134	0.20584		
Corrected Total	185263	45615			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.36217	0.00230	5096.61238	24759.6	<.0001
carrier_ct	0.00494	0.00003827	3424.98567	16638.8	<.0001
_weather_ct	-0.00363	0.00014396	130.88479	635.85	<.0001
nas_ct	-0.00018676	0.00001844	21.11611	102.58	<.0001
_month	-0.00266	0.00030565	15.56221	75.60	<.0001

Stepwise Selection: Step 5

Variable late_aircraft_ct Entered: R-Square = 0.1640 and C(p) = 6.3501

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7483.05110	1496.61022	7270.95	<.0001
Error	185258	38132	0.20583		
Corrected Total	185263	45615			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.36196	0.00230	5086.26824	24710.5	<.0001
carrier_ct	0.00503	0.00004977	2105.83283	10230.7	<.0001
_weather_ct	-0.00372	0.00014661	132.17133	642.12	<.0001
nas_ct	-0.00016454	0.00001981	14.19302	68.95	<.0001
late_aircraft_ct	-0.00008269	0.00002699	1.93229	9.39	0.0022
_month	-0.00266	0.00030564	15.59396	75.76	<.0001

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	carrier_ct		1	0.1600	0.1600	890.443	35293.5	<.0001
2	_weather_ct		2	0.0032	0.1632	186.815	704.93	<.0001
3	nas_ct		3	0.0005	0.1637	87.3434	101.43	<.0001
4	_month		4	0.0003	0.1640	13.7377	75.60	<.0001
5	late_aircraft_ct		5	0.0000	0.1640	6.3501	9.39	0.0022

The final model includes five variables: carrier_ct, _weather_ct, nas_ct, _month, and late_aircraft_ct. These variables together explain about 16.4% of the changes in the outcome. All the chosen variables are important, and there are no other variables that significantly improve the model beyond these.

Model 3:

In this regression model (MODEL3), we are trying to predict the dependent variable (y) using the independent variable carrier_ct. The analysis of variance (ANOVA) table shows that the model is statistically significant, with a very low p-value (< 0.0001), indicating that the relationship between the independent and dependent variables is not due to random chance.

The model's R-squared value is 0.1600, which means that about 16.0% of the variability in the dependent variable (y) is explained by the independent variable carrier_ct. The adjusted R-squared accounts for the number of predictors in the model and is also 0.1600.

Overall, the model suggests a strong and significant relationship between carrier_ct and the dependent variable y, with the chosen variables providing meaningful predictive power.

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	265047
Number of Observations Used	185264
Number of Observations with Missing Values	79783

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7299.42814	7299.42814	35293.5	<.0001
Error	185262	38316	0.20682		
Corrected Total	185263	45615			

Root MSE	0.45478	R-Square	0.1600
Dependent Mean	0.43851	Adj R-Sq	0.1600
Coeff Var	103.70947		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.34768	0.00116	299.22	<.0001
carrier_ct	1	0.00414	0.00002201	187.87	<.0001

The MEANS Procedure

Variable	N	Mean
rmse1	79417	0.4330057
rmse2	79417	0.4330237
rmse3	79417	0.4343601
mse1	79417	0.2065553
mse2	79417	0.2065626
mse3	79417	0.2077462
mae1	79417	0.4330057
mae2	79417	0.4330237
mae3	79417	0.4343601
mpe1	34798	0.5184672
mpe2	34798	0.5185056
mpe3	34798	0.5192500

For all metrics (RMSE, MSE, MAE, MPE), lower values indicate better model performance. Therefore, Model 1 appears to be the best-performing model, as it has the lowest values for each metric.

- RMSE: 0.4330057
- MSE: 0.2065553
- MAE: 0.4330057
- MPE: 0.5184672

Random Forest:

RF1

	estname	prob	true_total	truepos	falsneg	detection_rate	false_total	falspos	trueneg	false_pos_rate
1	RANDFOREST	0.1	34796	33795	1001	0.9712	44511	7065	37446	0.1587
2	RANDFOREST	0.2	34796	33328	1468	0.9578	44511	5204	39307	0.1169
3	RANDFOREST	0.3	34796	32822	1974	0.9433	44511	3947	40564	0.0887
4	RANDFOREST	0.4	34796	32332	2464	0.9292	44511	3040	41471	0.0683
5	RANDFOREST	0.5	34796	31825	2971	0.9146	44511	2334	42177	0.0524
6	RANDFOREST	0.6	34796	31251	3545	0.8981	44511	1760	42751	0.0395
7	RANDFOREST	0.7	34796	30602	4194	0.8795	44511	1251	43260	0.0281
8	RANDFOREST	0.8	34796	29823	4973	0.8571	44511	837	43674	0.0188
9	RANDFOREST	0.9	34796	28774	6022	0.8269	44511	456	44055	0.0102

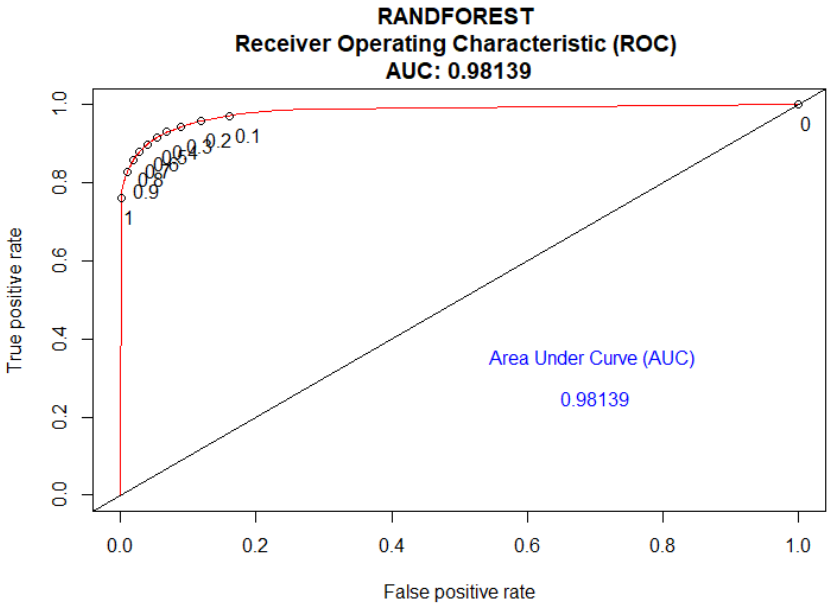
for prob=0.3:

At a probability threshold of 0.3, the model identified 32,822 instances correctly as positive out of a total of 34,796 true positives in the dataset. There were 1,974 false negatives, indicating instances that the model missed. The detection rate at this threshold is 94.33%, showing a high rate of correctly identifying positive instances. Out of 44,511 instances predicted as positive, 3,947 were false positives.

The false positive rate at this threshold is 8.87%, indicating the proportion of false positives among all instances predicted as positive.

AUC(RF1)

- The random forest model (rf1) demonstrates excellent predictive performance, as evidenced by the high AUC value of 0.98795 at a probability threshold of 0.3
- The model's high discriminative power at this threshold suggests that these variables contribute significantly to predicting flight delays which are more than 30 minutes.



- The Mtry value was taken by this output.The second value has the lowest errors.

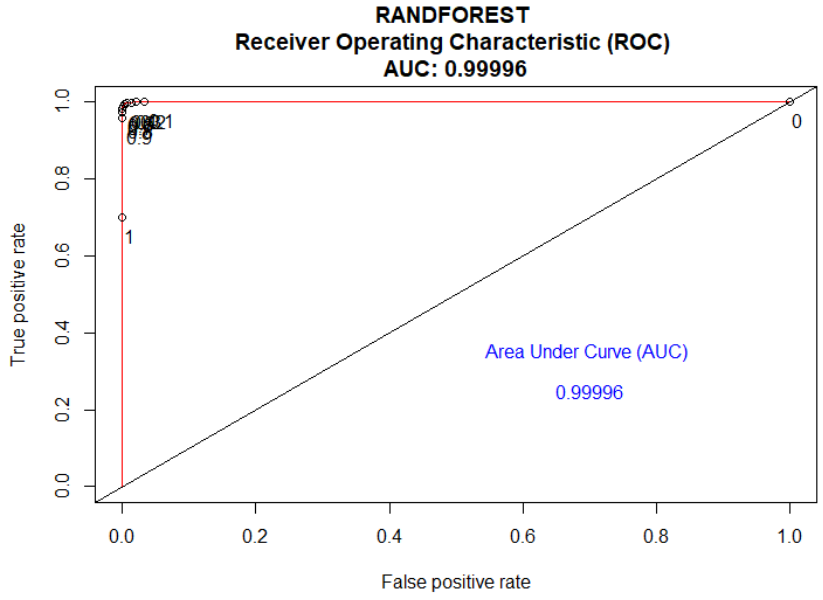
```
> cbind(1:12,oob.values)
      oob.values
 [1,]  1 0.006113815
 [2,]  2 0.003610011
 [3,]  3 0.004009325
 [4,]  4 0.004478788
 [5,]  5 0.005164096
 [6,]  6 0.005061570
 [7,]  7 0.005169492
 [8,]  8 0.005072362
 [9,]  9 0.005045381
[10,] 10 0.005093947
[11,] 11 0.005115531
[12,] 12 0.005153304
```

RF2:

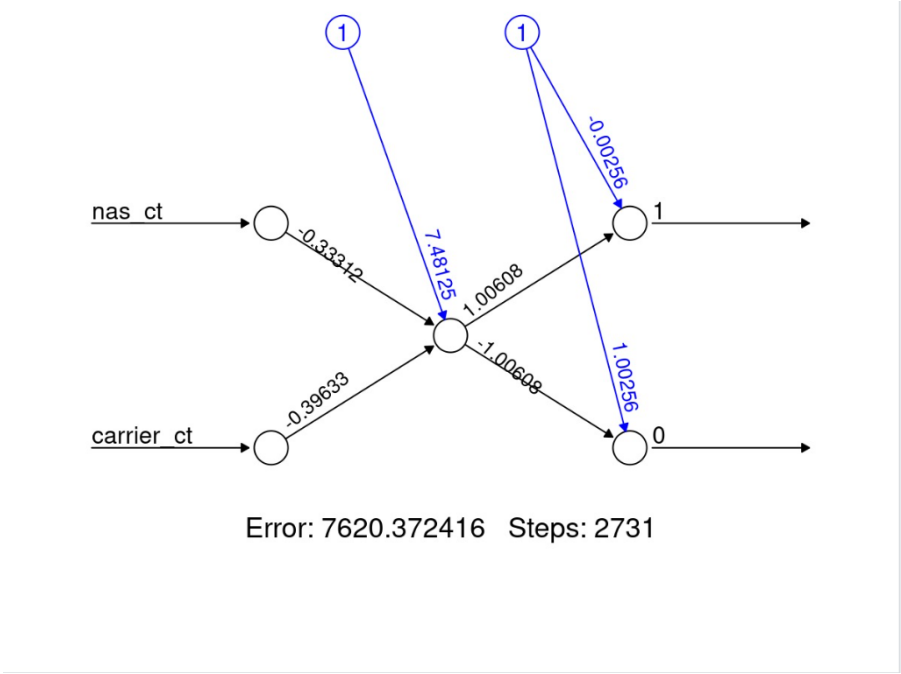
```
> conf_table(rfhat2,testy, "RANDFOREST")
     estname prob true_total truepos falsneg detection_rate false_total falspos trueneg false_pos_rate
1 RANDFOREST  0.1      34796   34795       1              1       44511    1427   43084         0.0321
2 RANDFOREST  0.2      34796   34791       5         0.9999       44511     928   43583         0.0208
3 RANDFOREST  0.3      34796   34780      16         0.9995       44511     578   43933          0.013
4 RANDFOREST  0.4      34796   34735      61         0.9982       44511     302   44209         0.0068
5 RANDFOREST  0.5      34796   34652     144         0.9959       44511     123   44388         0.0028
6 RANDFOREST  0.6      34796   34460     336         0.9903       44511      35   44476           8e-04
7 RANDFOREST  0.7      34796   34208     588         0.9831       44511      12   44499           3e-04
8 RANDFOREST  0.8      34796   33851     945         0.9728       44511       3   44508           1e-04
9 RANDFOREST  0.9      34796   33331    1465         0.9579       44511       0   44511               0
```

At a probability threshold of 0.3, the model achieved a high detection rate of 99.95%, indicating its effectiveness in identifying true positives. Similarly, the false positive rate at this threshold is relatively low at 1.3%, suggesting good specificity.

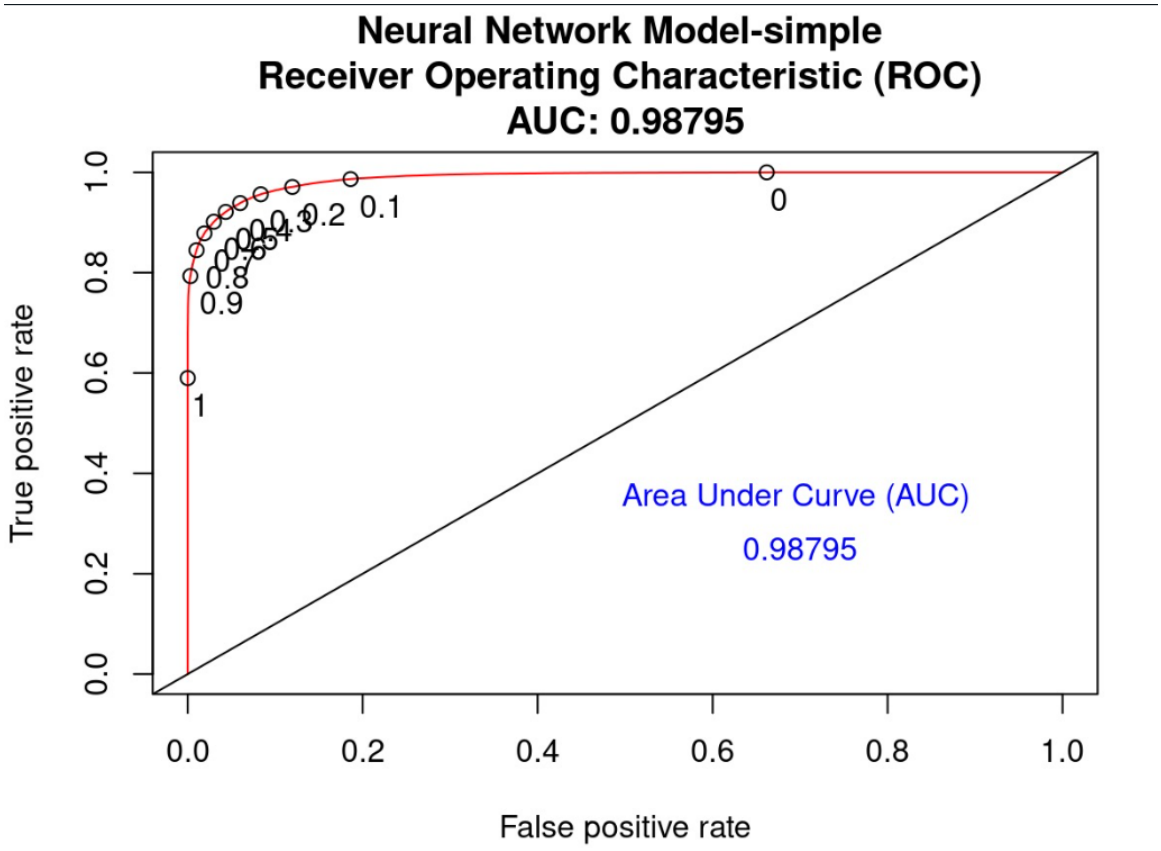
- The AUC (Area Under the Curve) plot value of 0.99996 indicates an exceptionally high discriminatory power of the model.
- But a higher AUC could also be a case of overfitting.



Neural Network:(nnet1)

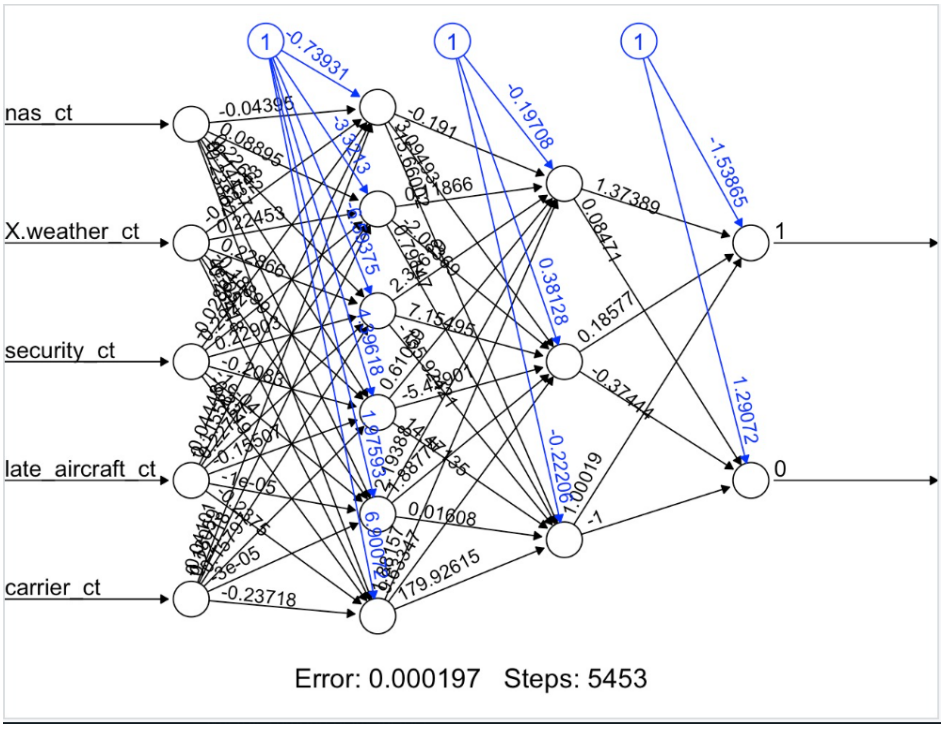


At a probability threshold of 0.3, the NeuralNet model was applied to 34,796 instances, aiming to predict whether flights were delayed by 30 minutes or more (delayed = 1) or not delayed (delayed = 0). The model correctly identified 33,269 instances where flights were indeed delayed (true positives) and missed 1,527 instances of delayed flights (false negatives). The detection rate, representing the proportion of actual delayed flights correctly identified by the model, was 95.61%, indicating a high sensitivity to detecting delayed flights. However, the model also misclassified 3,714 instances as delayed when they were not (false positives), resulting in a false positive rate of 8.34%. Despite these misclassifications, the model demonstrated an overall high accuracy and precision in predicting flights delayed by 30 minutes or more at this probability threshold.



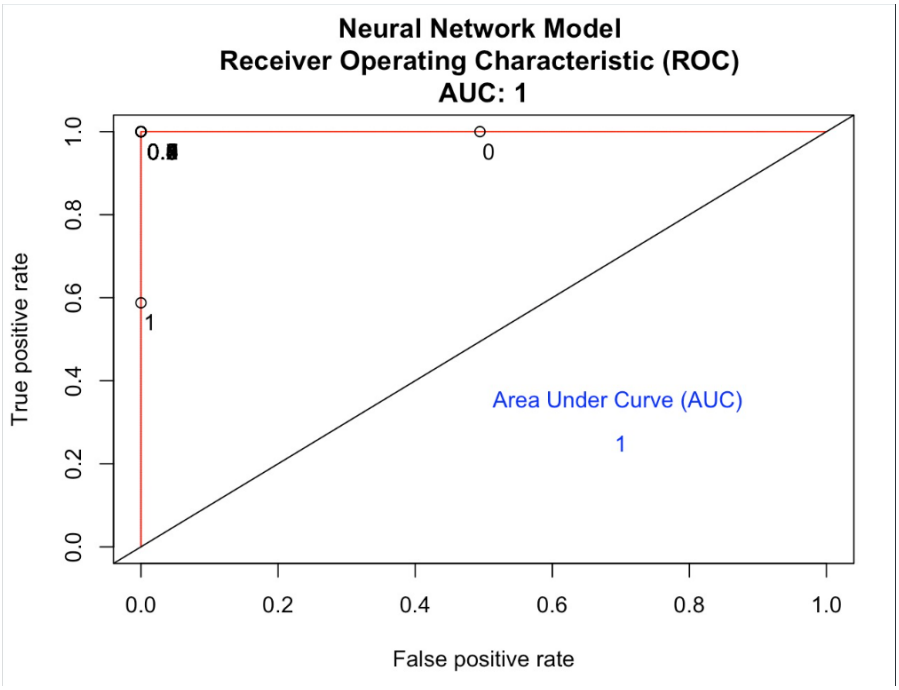
The high AUC value of 0.98795 indicates that the Random Forest model performed exceptionally well in distinguishing between delayed and non-delayed flights across various probability thresholds. It suggests that, on average, the model is very effective in assigning higher probabilities to delayed flights compared to non-delayed ones.

nnet2:



- The model correctly identified all instances of delayed flights (true positives) and non-delayed flights (true negatives). The false positive rate, which indicates the proportion of non-delayed flights incorrectly classified as delayed, is consistently zero for all probability thresholds.

	estname	prob	true_total	truepos	falsneg	detection_rate	false_total	falspos	trueneg	false_pos_rate
1	NeuralNet	0.1	34796	34796	0	1	44511	0	44511	0
2	NeuralNet	0.2	34796	34796	0	1	44511	0	44511	0
3	NeuralNet	0.3	34796	34796	0	1	44511	0	44511	0
4	NeuralNet	0.4	34796	34796	0	1	44511	0	44511	0
5	NeuralNet	0.5	34796	34796	0	1	44511	0	44511	0
6	NeuralNet	0.6	34796	34796	0	1	44511	0	44511	0
7	NeuralNet	0.7	34796	34796	0	1	44511	0	44511	0
8	NeuralNet	0.8	34796	34796	0	1	44511	0	44511	0
9	NeuralNet	0.9	34796	34796	0	1	44511	0	44511	0



This high level of accuracy in correctly classifying both delayed and non-delayed flights suggests a strong performance of the NeuralNet model with the extended set of predictor variables.

However, in a dataset of this size, such a result is highly unusual and suggests potential overfitting.

SUMMARY OF THE PROJECT

The analysis of flight delay data from 2004 to 2019 reveals key insights. On average, airports experience approximately 78 delayed flights per month out of 400 arrival flights. The most common delay causes include late-arriving aircraft, National Aviation System issues, and carrier-related problems. Weather-related delays are less frequent but can still be significant.

Exploring relationships, a positive correlation is observed between the number of flights and the number of flights delayed by over 15 minutes. Notably, weather-related factors show a weaker correlation with delays compared to other contributors like late aircraft and carrier issues.

Regression analysis, ANOVA, and clustering techniques further elucidate patterns. Skywest Airlines stands out for the highest delay periods. The choice of the delayed variable (30 min) over arr_del15 enhances model fitting.

Stepwise regression reveals a model with carrier_ct, _weather_ct, nas_ct, _month, and late_aircraft_ct as significant predictors, explaining about 16.4% of delay variability.

In machine learning, Random Forest and Neural Network models showcase high performance, with Random Forest exhibiting exceptional discrimination ability (AUC 0.98795).

In conclusion, the comprehensive analysis integrates statistical, machine learning, and predictive techniques to understand and predict flight delays, offering valuable insights for operational improvements in the airline industry.

BIBLIOGRAPHY

- Bureau of Transportation Statistics (BTS)

  https://bigblue.depaul.edu/jlee141/econdata/BTS/

  https://bigblue.depaul.edu/jlee141/econdata/BTS/flight_delay_2004_2019.csv

APPENDIX: SAS or R Command

SAS

```
filename webdat url "https://bigblue.depaul.edu/jlee141/econdata/BTS/flight_delay_2004_2019.csv" ;
PROC IMPORT OUT= airline DATAFILE= webdat DBMS=CSV REPLACE;
RUN;

proc contents data= airline ;run;

proc means data = airline; run;

proc  summary print data= airline mean median min max ;
var _arr_delay; var _carrier_delay;
run;

data air1; set airline;
if arr_del15 > 30 then delayed = 1 ; else delayed = 0 ;
run;

proc univariate data= airline;
var arr_del15;
run;

/* Plots */
proc sgplot data= airline;
scatter x=arr_del15 y=  arr_flights;
reg x=arr_del15 y=  arr_flights/ lineattrs=(color=red);
run;

proc sgplot data= airline;
scatter x=_weather_ct y=arr_del15 ;
reg x=_weather_ct y=  arr_del15/ lineattrs=(color=red);
run;


proc sgplot data= airline;
scatter x=carrier_ct y= arr_del15;
reg x=carrier_ct y=  arr_del15/ lineattrs=(color=red);
run;

proc sgplot data= airline;
scatter x=late_aircraft_ct y= arr_del15;
reg x=late_aircraft_ct y=  arr_del15/ lineattrs=(color=red);
run;

/* Correlation */

proc corr data= air1;
var arr_del15; var nas_ct; var _weather_ct; var security_ct;
var late_aircraft_ct; var carrier_ct ; var _month; var delayed;
run;

/* Anova Analysis */

proc anova data= air1;
class carrier_ct;
model delayed= carrier_ct ;
run;

proc anova data= air1;
class  nas_ct;
model delayed= nas_ct  ;
run;
```

```
proc anova data= air1;
 class  late_aircraft_ct;
 model delayed= late_aircraft_ct  ;
 run;

proc anova data= air1;
 class  security_ct;
 model delayed= security_ct ;
 run;

proc anova data= air1;
 class  _weather_ct;
 model delayed= _weather_ct  ;
 run;

 /* Clsutering */

proc fastclus data=air1 out=kmeandat maxclusters=4;
 var _month arr_del15 ;
 run;
 /* Simple Graph using K-means */
proc sgplot data=kmeandat ;
 scatter y= arr_del15 x= _month / group=cluster ;
 styleattrs datalinepatterns=(dot dash longdash);
 run;


proc sgplot data=kmeandat ;
 scatter y= arr_del15 x= _month  / group=cluster ;
 styleattrs datacontrastcolors=(blue red green)
 datalinepatterns=(dot dash longdash);
 reg  y= arr_del15 x= _month  / group=cluster;
 run;

 /*Regression*/

proc reg data = air1 ;
 model delayed = nas_ct;
 run;

 /* Spliting data */

proc surveyselect data= air1  method=srs seed=1234 outall
      samprate=0.7 out= air_split;
 run;

proc freq data=air_split; tables selected ;
  run;

 data regdata ; set air_split ;
    y = delayed ;
    if selected = 0 then y = . ;
 run ;


 %let indep_var =  carrier_ct _weather_ct nas_ct security_ct late_aircraft_ct _month ;
```

```
proc reg data=regdata;
    model  y = &indep_var / selection = adjrsq   ; output out=r1(where=(y=.)) p=yhat1;
    model  y = &indep_var / selection = stepwise ; output out=r2(where=(y=.)) p=yhat2;
    model  y = carrier_ct  ; output out=r3(where=(y=.)) p=yhat3;
 run ;

data allr ; merge r1 r2 r3  ;
 yorg = delayed;
 e1 = yorg - yhat1;
 e2 = yorg - yhat2;
 e3 = yorg - yhat3;

 rmse1 = ((e1)**2)**.5;
 rmse2 = ((e2)**2)**0.5 ;
 rmse3 = ((e3)**2)**0.5 ;

 mse1 = (e1)**2 ;
 mse2 = (e2)**2 ;
 mse3 = (e3)**2 ;

 mae1 = abs(e1) ;
 mae2 = abs(e2) ;
 mae3 = abs(e3) ;

 mpe1 = abs((e1)/yorg) ;
 mpe2 = abs((e2)/yorg) ;
 mpe3 = abs((e3)/yorg) ;
 run ;
proc means data=allr n mean ; var rmse: mse: mae: mpe: ;  run ;
```

Random Forest (R)

```
source("http://bigblue.depaul.edu/jlee141/econdata/R/func_lib.R")

airline <- read.csv("https://bigblue.depaul.edu/jlee141/econdata/BTS/flight_delay_2004_2019.csv")

str(airline)


airline$year = as.factor(airline$year)
airline$X.month = as.factor(airline$X.month)
airline$carrier = as.factor(airline$carrier)
airline$carrier_name = as.factor(airline$carrier_name)
airline$airport = as.factor(airline$airport)
airline$airport_name = as.factor(airline$airport_name)

indata <- airline
indata$delayed <- ifelse(indata$arr_del15 > 30, 1, 0)
#creating binary variable from delay 15 min to delay with 30 min
indata <- subset(indata, select = -X)
# X was a variable with N/A values.
indata<- na.omit(indata)
# To omit the N/A values

indata$delayed = as.factor(indata$delayed)
str(indata)


set.seed(1234)


train_idx = sample(c(TRUE, FALSE), nrow(indata), replace = TRUE, prob = c(0.7,0.3))
train <- indata[train_idx,]

test <- indata[!train_idx,]

testy <- test$delayed

# to check if the data is splitting
dim(test)
dim(train)
```

```
library('randomForest')
library('ROCR')

rf1 <- randomForest(formula=delayed ~ nas_ct + carrier_ct , data = train, mtry=2, ntree=200)
summary(rf1)

rfhat1 <- predict(rf1,newdata = test, type = "prob")
rfhat1 <- rfhat1[,2]
conf_table(rfhat1,testy, "RANDFOREST")
auc_plot(rfhat1,testy, "RANDFOREST")

oob.values <- vector(length = 12)
for (i in 1:12) {
  temp.model <- randomForest(formula = delayed ~ nas_ct + X.weather_ct +
                               security_ct + late_aircraft_ct + carrier_ct ,
                             data = train, mtry=i,ntree=800)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
cbind(1:12,oob.values)

mrf_tree <- randomForest(formula=delayed ~ nas_ct + X.weather_ct +
                           security_ct + late_aircraft_ct + carrier_ct ,
                         data = train, mtry=i, ntree=800)
Trees <- rep(1:nrow(rf_tree$err.rate))
Error.rate <- rf_tree$err.rate[,"OOB"]
plot(Trees, Error.rate, col="blue")

rf2 <- randomForest(formula=delayed ~ nas_ct + X.weather_ct + security_ct +
                      late_aircraft_ct + carrier_ct  , data = train, mtry=2, ntree=700)
summary(rf2)

rfhat2 <- predict(rf2,newdata = test, type = "prob")
rfhat2 <- rfhat2[,2]
conf_table(rfhat2,testy, "RANDFOREST")
auc_plot(rfhat2,testy, "RANDFOREST")
```

```r
source("http://bigblue.depaul.edu/jlee141/econdata/R/func_lib.R")

airline <- read.csv("https://bigblue.depaul.edu/jlee141/econdata/BTS/flight_delay_2004_2019.csv")

str(airline)


airline<- subset(airline, select = -X)

airline$year = as.factor(airline$year)
airline$X.month = as.factor(airline$X.month)
airline$carrier = as.factor(airline$carrier)
airline$carrier_name = as.factor(airline$carrier_name)
airline$airport = as.factor(airline$airport)
airline$airport_name = as.factor(airline$airport_name)

airline$year = as.numeric(as.factor(airline$year))
airline$X.month = as.numeric(as.factor(airline$X.month))
airline$carrier = as.numeric(as.factor(airline$carrier))
airline$carrier_name = as.numeric(as.factor(airline$carrier_name))
airline$airport = as.numeric(as.factor(airline$airport))
airline$airport_name = as.numeric(as.factor(airline$airport_name))

indata <- airline

indata$delayed <- ifelse(indata$arr_del15 > 30, 1, 0)

indata<- na.omit(indata)

indata$delayed = as.factor(indata$delayed)

str(indata)


set.seed(1234)

train_idx = sample(c(TRUE, FALSE), nrow(indata), replace = TRUE, prob = c(0.7,0.3))

train <- indata[train_idx,]

test <- indata[!train_idx,]
```