University of Minnesota Duluth

Airline Flight Delays

Project #3: Final Report

Nikolaus Schroeder
Daniel Brula
Jiawei Yu
Corbin Lawien

FMIS 3295: Business Analytics
Nik R. Hassan

4/26/2016

Executive Summary

From the beginning, our group has always believed the time of a customer is just as valuable as the money they spend to get the product or service they are purchasing. When it comes to air travel, customers as a whole, lose millions of hours each year due to canceled and delayed flights. *Berkely News* conducted a study and found that approximately $32.9 billion is lost each year due to these delays and cancellations. The purpose of this data mining project is to determine how certain factors are related to these delays and to provide suggestions on how to solve flight delay issues.

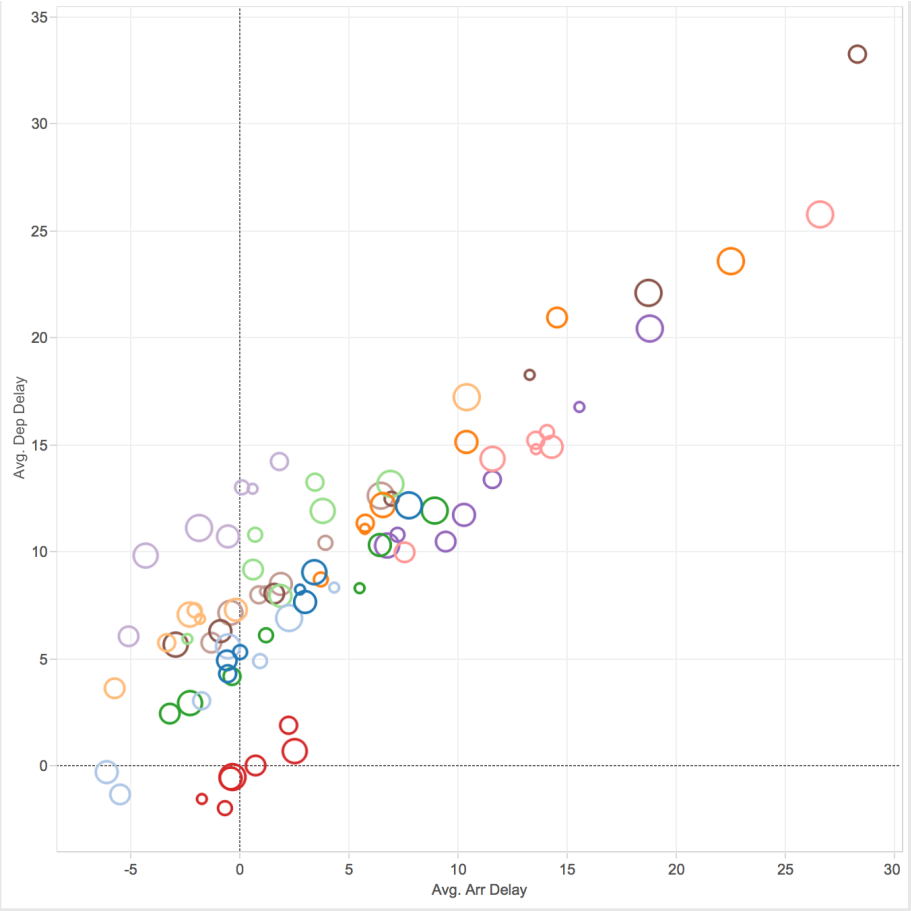
The United States Department of Transportation along with the Federal Aviation Administration monitor every flight that takes off and lands around the world. This data has been collected since the mid-1980's, but has increased dramatically over the years. One primary reason for this is the increase of security risks on airplanes since the September 11th terrorist attacks. The Department of Transportation stores all of this information on their website in several databases and is available to the general public. Through the use of these databases, our group can manipulate the data to find relationships between the variables.

Over the course of this project we have run into many obstacles, ultimately leading to the decision to scrap the dataset used in our previous report and start over with completely new data. We believe this decision has greatly improved the accuracy of our data and makes it much easier to determine actual relationships.

Table of Contents

Project #2 Revisions

Descriptive Analytics:

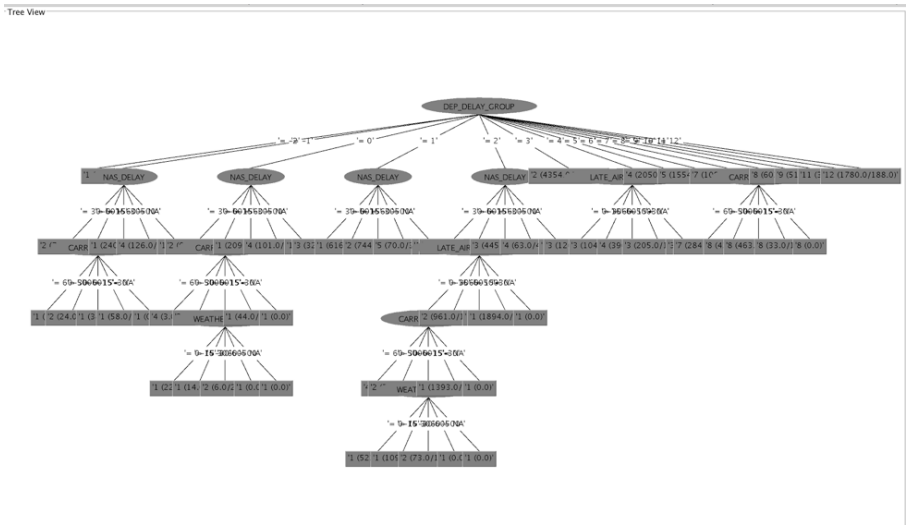


Scatter plot of Project #2 dataset

The charts we presented in Project #2 made it difficult to understand the overall goal of this project. For Project #3, we are using all new charts, including scatter plots and box and whisker plots. These charts give the reader an earlier understanding of trends and relationships between delay times and other variables in our data set. These new charts are located in the "Descriptive Analytics" portion of this report.

Modeling and Predictive Analytics:

At first, we decided to use J48 decision tree as the main model for our project because J48 decision tree can automatically build a tree which has many leave node to show which variable is the most informative to the class variable. Also, it can show the percentage and confusion matrix to present the accuracy of classification. According to the result of J48 decision tree, we can find a next step's direction of our project.




```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       25286               65.2306 %
Incorrectly Classified Instances     13478               34.7694 %
Kappa statistic                          0.5402
Mean absolute error                      0.0728
Root mean squared error                  0.1911
Relative absolute error                 56.5437 %
Root relative squared error             75.2887 %
Total Number of Instances            38764

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.944     0.105       0.86      0.944     0.9        0.964    1
                 0.604     0.12        0.545     0.604     0.573      0.863    2
                 0.356     0.071       0.386     0.356     0.371      0.882    3
                 0.313     0.037       0.396     0.313     0.35       0.914    4
                 0.28      0.025       0.376     0.28      0.321      0.931    5
                 0.084     0.006       0.339     0.084     0.134      0.944    6
                 0.526     0.026       0.364     0.526     0.43       0.955    7
                 0.466     0.017       0.361     0.466     0.407      0.965    8
                 0.323     0.008       0.37      0.323     0.345      0.965    9
                 0.053     0.001       0.317     0.053     0.09       0.962    10
                 0.272     0.005       0.339     0.272     0.301      0.968    11
                 0.918     0.005       0.894     0.918     0.906      0.994    12
Weighted Avg.    0.652     0.079       0.628     0.652     0.634      0.931

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k     l   <-- classified as
 14874   652   228     3     0     0     0     0     0     0     0     0 |     a = 1
  2297  4495   595    50     2     0     0     0     0     0     0     0 |     b = 2
   114  2417  1539   210    30     5     4     0     0     0     0     0 |     c = 3
     3   464  1165   874   228    35    21     1     0     0     0     0 |     d = 4
     0   151   317   717   555   133   103     7     1     0     0     0 |     e = 5
     0    36    93   230   458   120   439    60     1     0     0     0 |     f = 6
     0    16    28    81   136    41   559   190    12     0     0     0 |     g = 7
     0     3    12    18    36    13   269   355    50     1     5     0 |     h = 8
     0     4     4    10    18     3    98   223   190     7    28     4 |     i = 9
     0     4     3     8     7     2    26    94   173    26   112    40 |     j = 10
     0     0     2     4     2     1    10    38    52    34   107   144 |     k = 11
     0     1     1     2     3     1     6    16    34    14    64  1592 |     l = 12
```

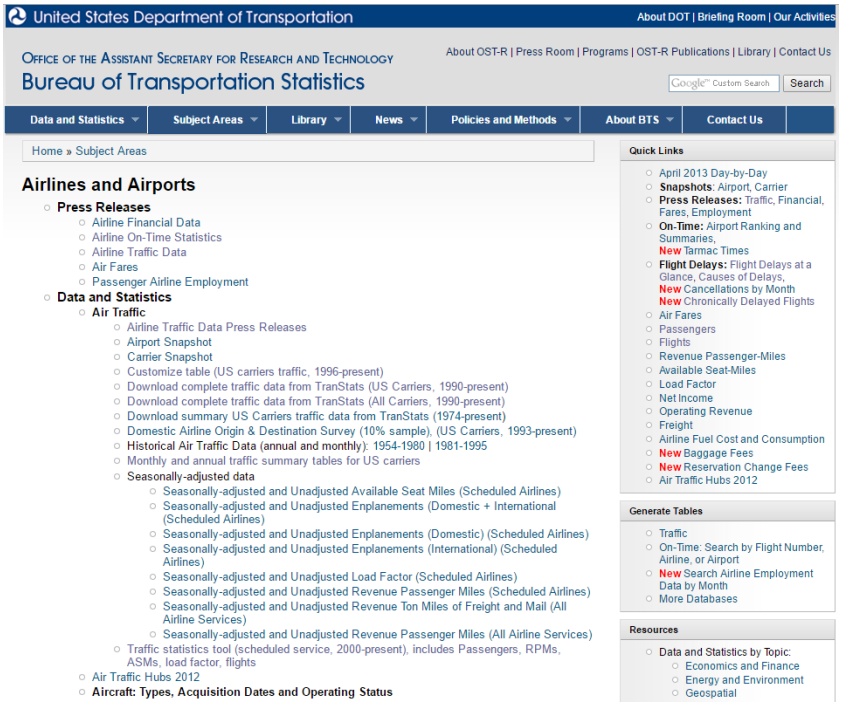
Conclusion:

From the result picture, we can find that the NAS_Delay is the most informative attribute, which means that main cause of the airline delay is NAS_Delay. If the airport want to reduce the delay time, they can focus on how to improve this part. However, from the correctly classified instances percentage, which is only 65%, it may not be a very useful model for our project.

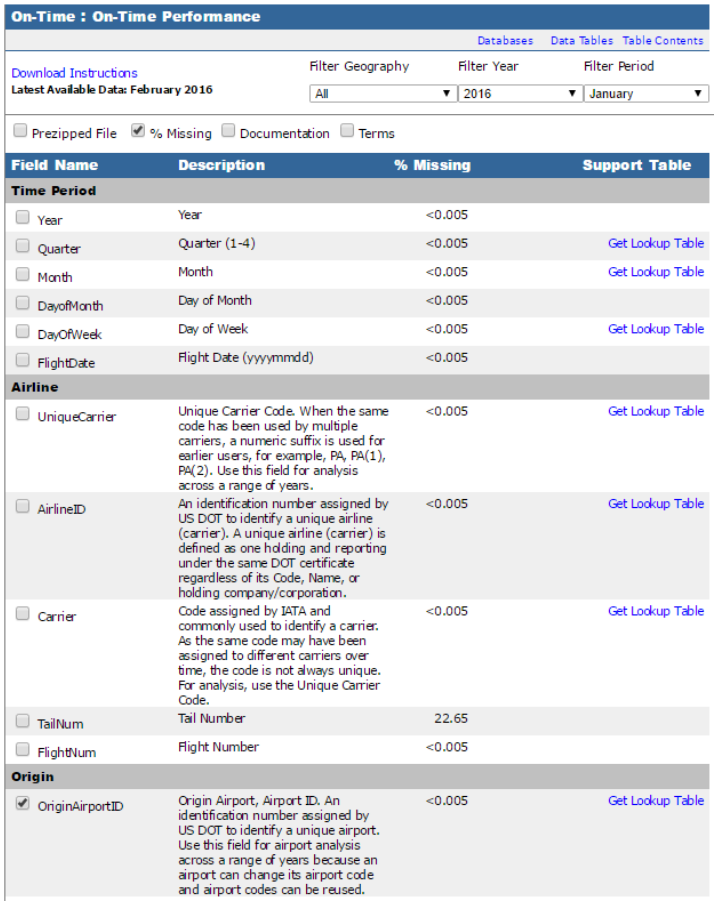
We did the J48 decision tree to present the most informative attribute for airline delay. The conclusion we got is that the NAS_delay has the highest information gain. So, if airline companies want to reduce the delay time, they can pay more attention on the NAS_delay part. Also, the accuracy of J48 decision tree is around 65%, which means that it may be a really good model for our project. So we built the new association rule in the expand model part.

Data Understanding

The data used throughout this project was collected from the United States Department of Transportation website (http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/subject_areas/airline_information/index.html). Located on this website are nearly a hundred different data sets, all related to the air transportation industry. The majority of these datasets go back as far as 1987, making it simple to measure trends over time.



With flight delay times being the target variable of this project, this limited the number of datasets which would be useful. This project focused on the data available in the "On Time Performance" dataset. This dataset, broken up by months, includes 109 different variables and over 500,000 entries per month, including, date, carrier, distance, airport locations, and etc.



On-Time : On-Time Performance

Databases | Data Tables | Table Contents

Download Instructions

Latest Available Data: February 2016

Filter Geography: All | Filter Year: 2016 | Filter Period: January

☐ Prezipped File ☑ % Missing ☐ Documentation ☐ Terms

Field Name	Description	% Missing	Support Table
Time Period			
☐ Year	Year	<0.005	
☐ Quarter	Quarter (1-4)	<0.005	Get Lookup Table
☐ Month	Month	<0.005	Get Lookup Table
☐ DayofMonth	Day of Month	<0.005	
☐ DayOfWeek	Day of Week	<0.005	Get Lookup Table
☐ FlightDate	Flight Date (yyyymmdd)	<0.005	
Airline			
☐ UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.	<0.005	Get Lookup Table
☐ AirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.	<0.005	Get Lookup Table
☐ Carrier	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.	<0.005	Get Lookup Table
☐ TailNum	Tail Number	22.65	
☐ FlightNum	Flight Number	<0.005	
Origin			
☑ OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.	<0.005	Get Lookup Table

With the large amount of data included in this dataset, narrowing down the size was a top priority. For Project #2 it was decided to focus only on the month of January 2016. While working with this data it soon became clear that finding accurate results would be difficult given such a small timeframe and the countless number of airports, of all sizes, included in the data. For the final project, new constraints were created to improve the accuracy of the project…

- Data will include the entire year of 2015.
- Only flights departing the Minneapolis International Airport (MSP).
- Domestic Flights only.
- Only variables related to departure delay included.

With 109 variables in the dataset, many of which had high percentages of missing data, reducing this to ones needed was the next step. On the Department of Transportation website is the ability to select which variables are needed before downloading the entire dataset. The final 27 variables used in this projects are…

Variable	Description
Quarter	Quarter of year (1-4)
Month	Month of year the flight occurred
Day_Of_Month	Day of month the flight occurred
Day_Of_Week	Day of the week the flight occurred
Carrier	Code assigned by the IATA to each specific carrier
Fl_Num	Flight number
Origin	Airport where the flight departs
Dest	Airport where the flight arrives
Dest_City_Name	City where the airport is located
Dest_State_Nm	State where the airport is located
Crs_Dep_Time	Scheduled departure time of flight
Dep_Time	Actual departure time of flight
Dep_Delay	Difference (in minutes) between Crs_Dep_Time and Dep_Time
Dep_Del15	Departure delay indicator (=1 if Dep_Delay is >= 15 minutes, otherwise =0)
Dep_Delay_Group	Departure delay in intervals of 15 minutes
Dep_Time_Blk	Time block of Crs_Dep_Time
Crs_Arr_Time	Scheduled arrival time of flight
Arr_Time	Actual arrival time of flight

Arr_Delay	Difference (in minutes) between Crs_Arr_Time and Arr_Time
Arr_Del15	Arrival delay indicator (=1 if Arr_Delay is >= 15 minutes, otherwise =0)
Arr_Delay_Group	Arrival delay in intervals of 15 minutes
Arr_Time_Blk	Time block of Crs_Arr_Time
Crs_Elapsed_Time	Scheduled time flight is in air
Actual_Elapsed_Time	Actual time flight is in air
Air_Time	Flight time
Distance	Miles between the two airports
Distance_Group	Distance at intervals of 250 miles

Once these variables are selected, the data can be downloaded directly from the Department of Transportation website in Excel file format with the click of a button. The downloaded data appears as the following…

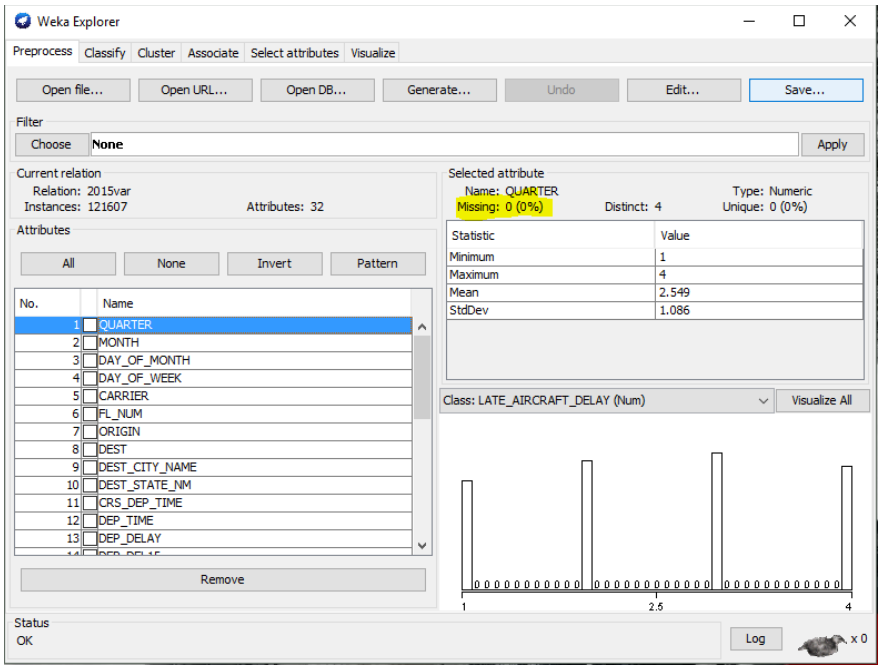
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	QUART	MONTH	DAY_O	DAY_O	CARRIE	FL_NUI	ORIGIN	DEST	DEST_C	DEST_S	CRS_DE	DEP_TI	DEP_DE	DEP_DE
2	1	1	2	5	DL	330	MSP	ANC	Anchorag	Alaska	1515	1515	0	
3	1	1	1	4	DL	301	MSP	ANC	Anchorag	Alaska	1540	1601	21	
4	1	1	3	6	DL	1099	MSP	ANC	Anchorag	Alaska	2140	2219	39	
5	1	1	3	6	DL	1224	MSP	ANC	Anchorag	Alaska	1515	1523	8	
6	1	1	2	5	DL	1099	MSP	ANC	Anchorag	Alaska	2150	2207	17	
7	1	1	4	7	DL	1099	MSP	ANC	Anchorag	Alaska	2158	2225	27	
8	1	1	4	7	DL	330	MSP	ANC	Anchorag	Alaska	1515	1520	5	
9	1	1	5	1	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1733	-2	
10	1	1	6	2	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1735	0	
11	1	1	8	4	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1737	2	
12	1	1	9	5	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1735	0	
13	1	1	11	7	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1735	0	
14	1	1	10	6	DL	1089	MSP	ANC	Anchorag	Alaska	1550	1638	48	
15	1	1	12	1	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1731	-4	
16	1	1	13	2	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1733	-2	
17	1	1	14	3	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1836	61	
18	1	1	15	4	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1739	4	
19	1	1	16	5	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1739	4	
20	1	1	17	6	DL	1089	MSP	ANC	Anchorag	Alaska	1550	1554	4	
21	1	1	18	7	DL	1089	MSP	ANC	Anchorag	Alaska	1530	1529	-1	
22	1	1	19	1	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1734	-1	
23	1	1	21	3	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1731	-4	
24	1	1	20	2	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1734	-1	
25	1	1	22	4	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1906	91	
26	1	1	24	6	DL	1089	MSP	ANC	Anchorag	Alaska	1550	1546	-4	
27	1	1	26	1	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1731	-4	
28	1	1	25	7	DL	1089	MSP	ANC	Anchorag	Alaska	1735	1731	-4	

Data Preparation

As shown in the previous section, data preparation starts on the Department of Transportation website. Since this data is collected by the United States government, it's clear that it's of the highest quality and very accurate. Once the variables are selected, data is downloaded in Excel file format by month. This creates 12 different Excel files, one for each month, that each need to be cleaned and combined into one file.

Opening each file in Excel, data is sorted by *Origin*, and every entry without "MSP" as the *Origin* variable is removed. This leaves, roughly, 10,000 entries per file. The next step is to combine the data into one file. Using Excel's table functions and having all common variables makes this a very easy process. After combining the tables, a total of 121,607 entries are ready to be analyzed. The dataset is then saved in two file formats, .xlsx for use in Tableau, and .csv for use in R and Weka.

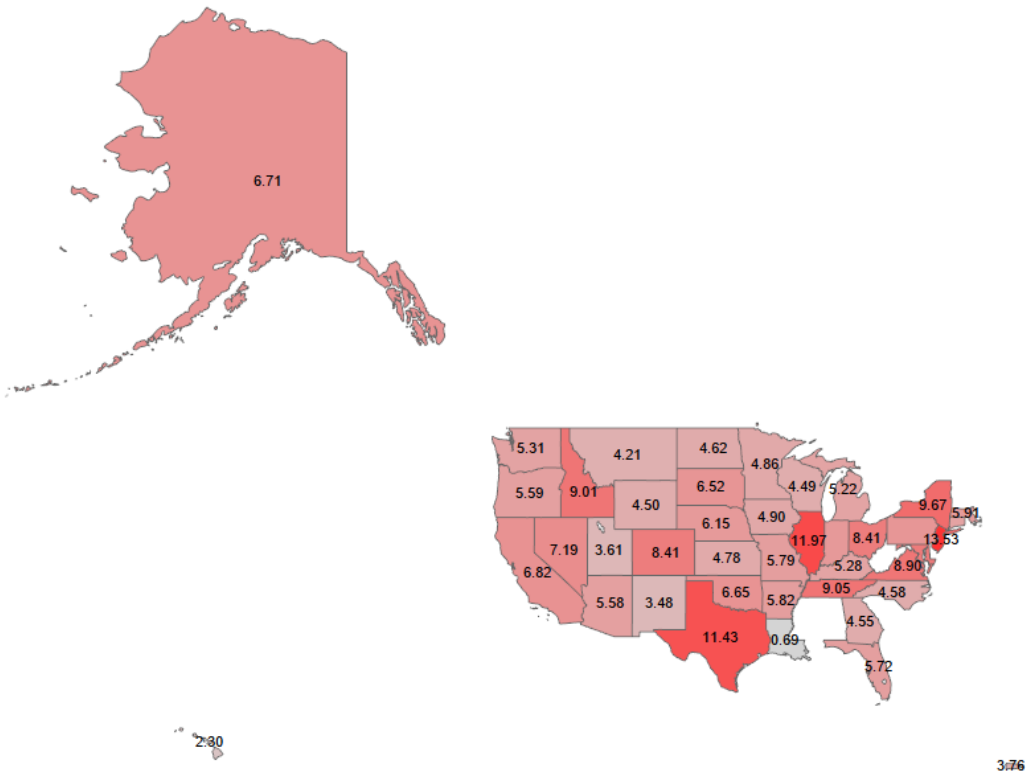
A concern all along was dealing with missing data. When first selecting which variables to use, only variables with an average of 0.0005% missing values were used to help limit the number of missing values in our final dataset. Weka includes a great feature that counts the number of missing entries in each variable. After opening the final, consolidated dataset in the program, it was discovered that there were ZERO missing values in each variable.



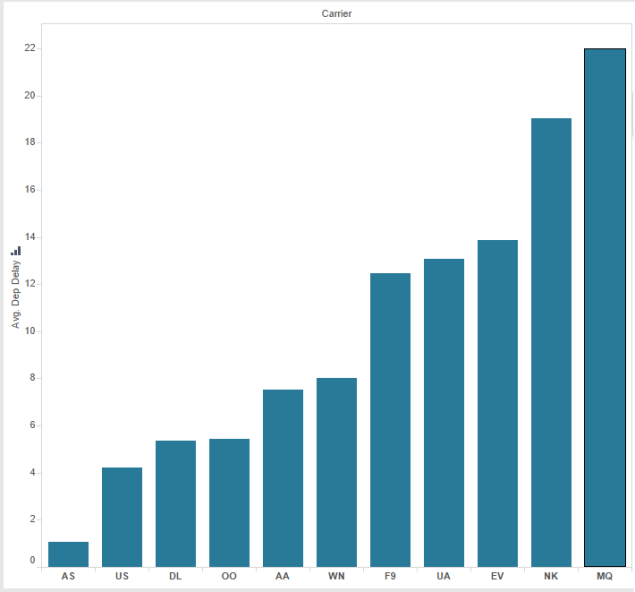
Carrier	Name
9E	Pinnacle
AA	American
AQ	Aloha
AS	Alaska
B6	JetBlue
CO	Continental
DL	Delta
EV	Atlantic
F9	Frontier
FL	AirTran
HA	Hawaiian
HP	America West
MQ	American Eagle
NW	Northwest
OH	Comair
OO	Skywest
UA	United
US	US Airways
WN	Southwest
XE	Expressjet
YV	Mesa

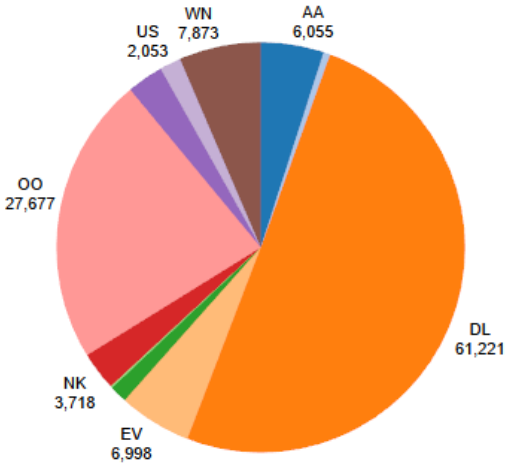
Descriptive Analytics

Tableau was the primary visualization tool used for this project. Saving a copy of the dataset in a .xlsx format allows Tableau to easily read and open the data set. Through this program the data can be manipulated in thousands of ways to gain early insights on trends.

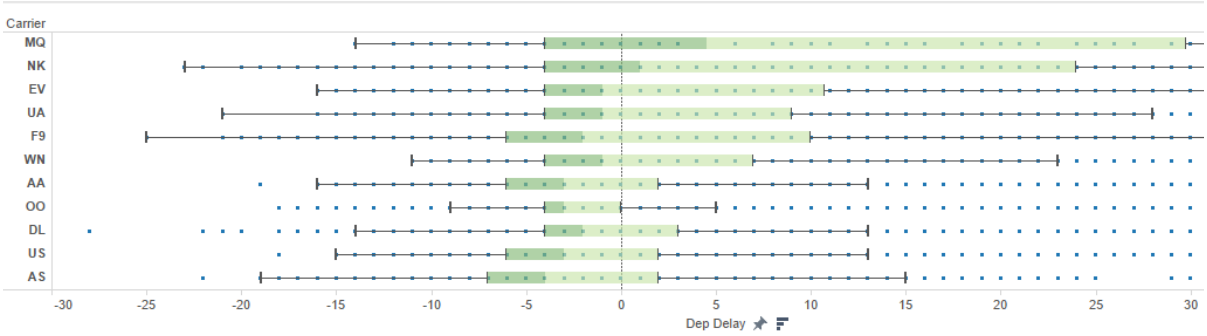


The chart above shows the average delay times of flights departing the Minneapolis International Airport (MSP) in the year of 2015 and arriving at airports in each state. Some states are missing from the chart due to no flights actually flying to those states. This chart shows you which arrival airports in states will likely have longer departure delays. At the same time, this chart also shows that the distance between airports may not be related to departure delays, as flights departing to airports in Minnesota had a higher average delay times than flights departing to airports in states like Utah, Georgia, and Montana.

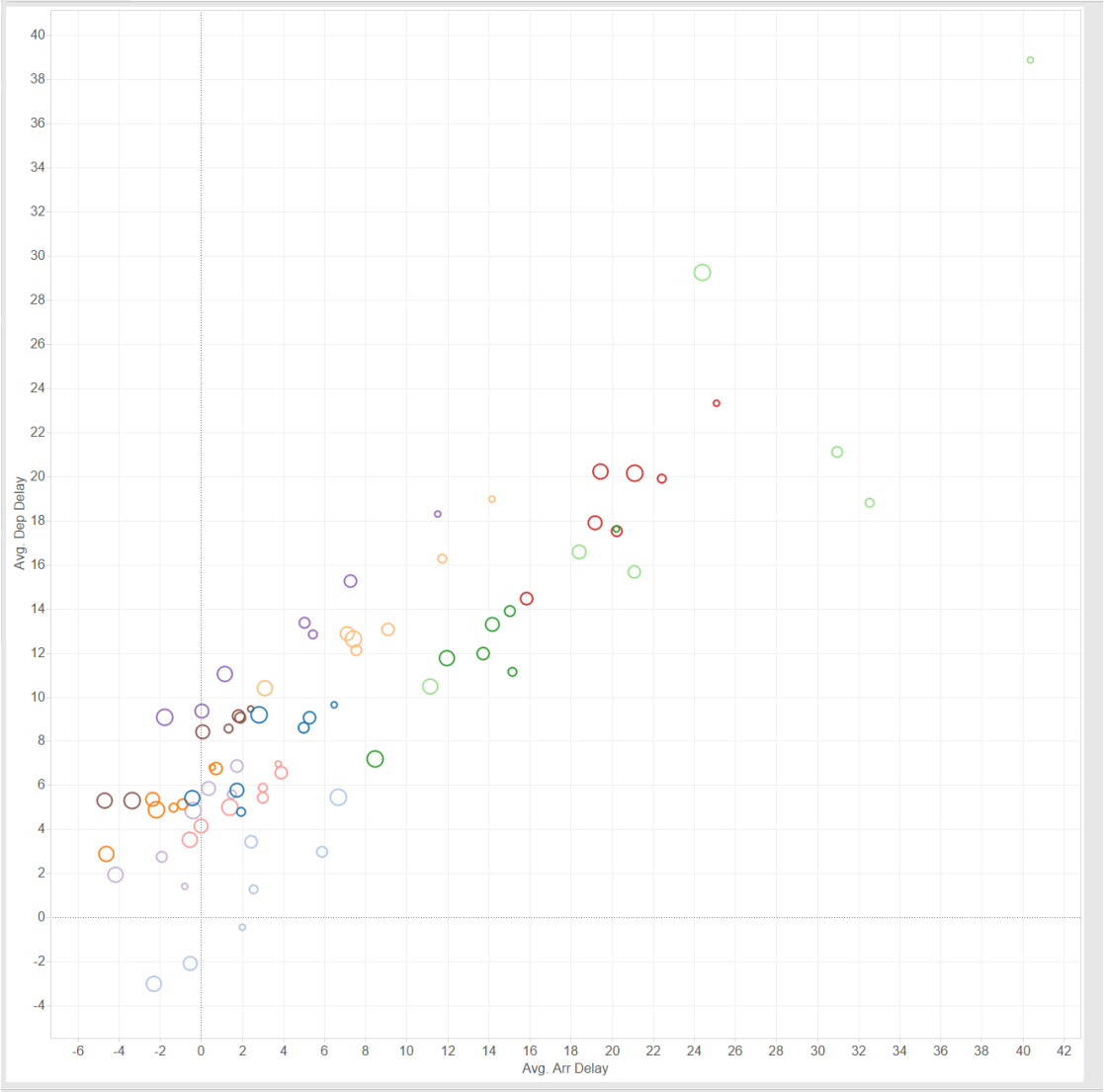




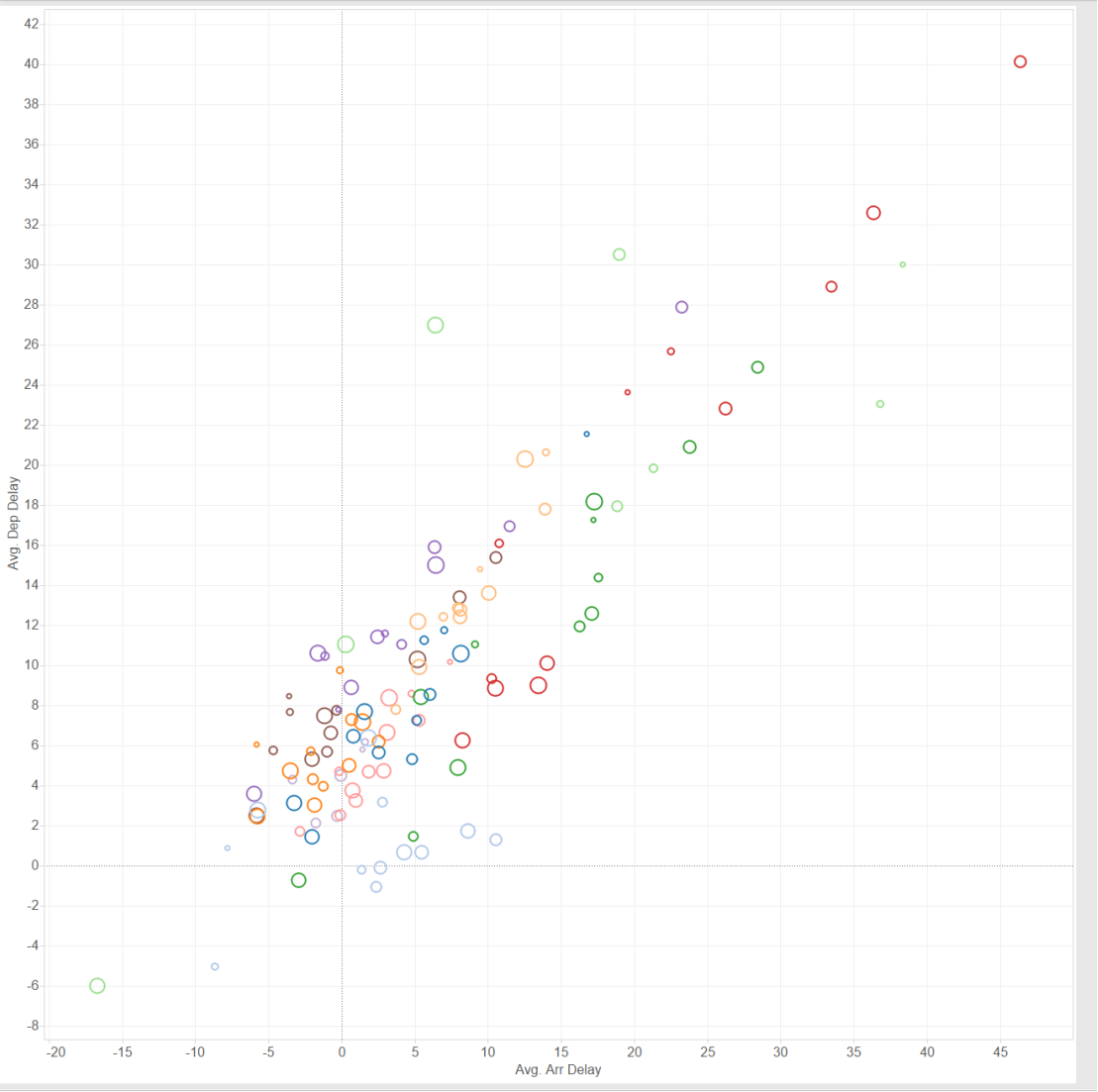
In the graph and pie chart above the data shows that departure delay times by airline are not related to the number of flights in the year by each of those airlines. The Minneapolis International Airport is one of Delta's main hubs in the United States, accounting for nearly half the total flights departing from the airport, but also has the third lowest average delay time.



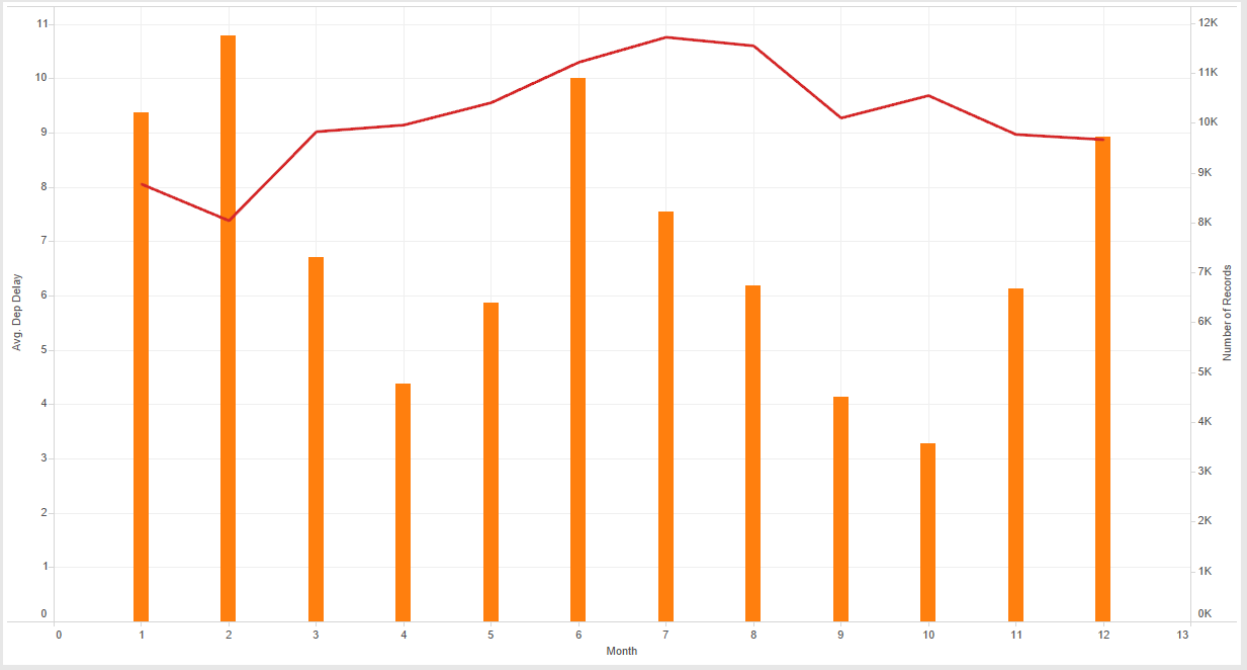
The box plot above shows the Upper Whisker, Upper Quartile, Median, Lower Quartile, and Lower Whisker of the delay times of the 11 airlines that fly out of the Minneapolis International Airport.



The first scatter plot that we were able to create involved plotting the average arrival delay and the average departure delay for the whole year based on day of the week along with color coding the different carriers. After looking at the scatter plot, it looks like most airlines have longer delays towards the beginning of the week compared to the end of the week. Another thing we notice is that the airline American Eagle (MQ) is on the higher end of both the average arrival and departure delay throughout the whole week. Delta airlines (DL) along with SouthWest (WN) are usually on the lower end of the plot when it comes to both average departure delay and average arrival delay.



For our second scatter plot, we used the same measures but changed day of the week to month. When looking at this scatter plot, we get a little bit of a different scenario compared to the day of the week scatter plot. We do find one similarity between the two scatter plots right away, Delta and SouthWest are still near the lower end of the plot showing low numbers in both arrival and departure delays. Another thing we notice is that the earlier months in the year seem to have higher delays compared to the middle of the year.



This final chart shows us the total number of departing flights and the average departure delay by month at the Minneapolis International Airport. This shows a slight relationship between the two variables with higher numbers of flights in the summer months along with higher average delay time. At the same time however, January and February also had high flight delay times, but the lowest amount of total flights.

With many hours in Tableau, many relationships were found. Flying with Alaska Airlines (AS), US Airways (US), and Delta Airlines (DL) lowers risks of having severely delayed flights. Distance between the Minneapolis International Airport and the arrival airport was not related to the severity of flight delays, but the size of arrival airports does. Flight delay times spike in the months the winter and summer months, but decline in the spring and summer months.

Modeling and Descriptive Analytics

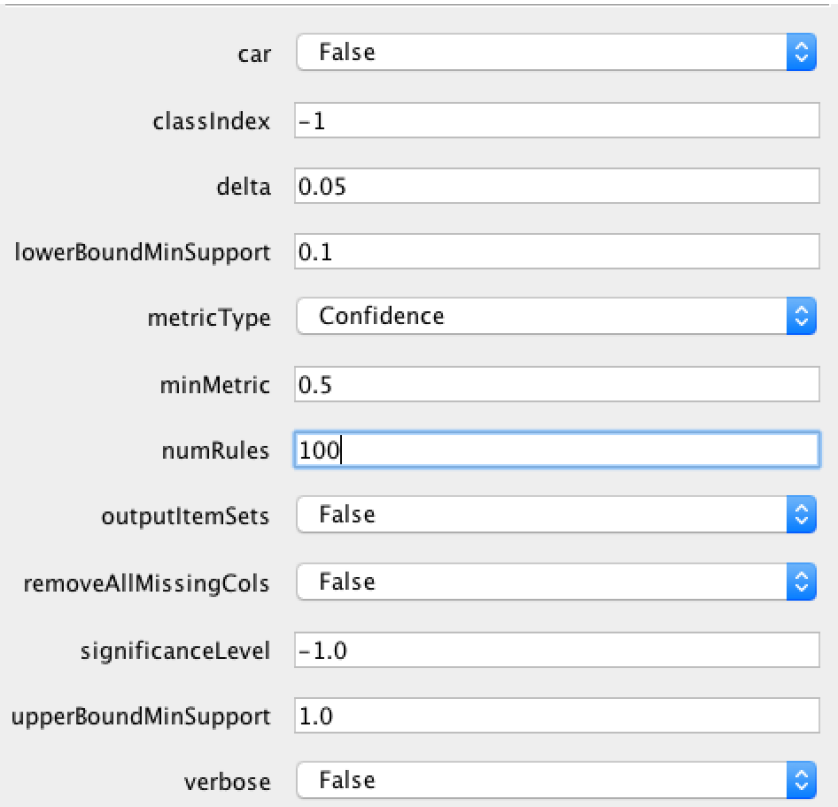
After we did the J48 decision tree model, we got that the NAS_delay is the most informative factor to the airline delay. But our goal is helping customers to decide which airline and which day they choose can save their time. Also, it's so difficult to achieve this goal among all origins. Therefore, we decided to choose all the airlines which depart from MSP airport. This time, we got 121607 instances.

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. We want to know which variables are related to the delay time. If we can get a high confidence association rule, we can give customers advice based on the rule. Therefore, this model can solve the problem that how do customers choose to reduce their waiting time.

Expanded Modeling and Predictive Analytics

Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. After considering air flight performance, we realized that there is something special and important that can address with so that it is quite beneficial to improve the quality of air flights' on-time performance and save customers' time.

Hence, based on this dataset, we built the associate rule, and use the Apriori method. We set the minimum confidence as 0.5 and the numRules as 100:



After running this method, we got many useful rules. Here is the result:

```
Best rules found:

 1. DEP_DEL15=1 WEATHER_DELAY=0 13143 ==> SECURITY_DELAY=0 13129    conf:(1)
 2. WEATHER_DELAY=0 18471 ==> SECURITY_DELAY=0 18448    conf:(1)
 3. SECURITY_DELAY=0 19212 ==> WEATHER_DELAY=0 18448    conf:(0.96)
 4. DEP_DEL15=1 SECURITY_DELAY=0 13790 ==> WEATHER_DELAY=0 13129    conf:(0.95)
 5. QUARTER=4 CARRIER=DL 14948 ==> DEP_DEL15=0 13230    conf:(0.89)
 6. QUARTER=3 CARRIER=DL 17419 ==> DEP_DEL15=0 15274    conf:(0.88)
 7. CARRIER=OO 27677 ==> DEP_DEL15=0 24162    conf:(0.87)
 8. QUARTER=2 CARRIER=DL 15955 ==> DEP_DEL15=0 13909    conf:(0.87)
 9. CARRIER=DL 61221 ==> DEP_DEL15=0 53356    conf:(0.87)
10. QUARTER=4 29994 ==> DEP_DEL15=0 26037    conf:(0.87)
11. DISTANCE_GROUP=6 16412 ==> DEP_DEL15=0 14246    conf:(0.87)
12. DAY_OF_WEEK=7 17071 ==> DEP_DEL15=0 14712    conf:(0.86)
13. QUARTER=3 33374 ==> DEP_DEL15=0 28756    conf:(0.86)
14. DISTANCE_GROUP=4 21743 ==> DEP_DEL15=0 18660    conf:(0.86)
15. DAY_OF_WEEK=5 18167 ==> DEP_DEL15=0 15588    conf:(0.86)
16. DAY_OF_WEEK=3 18239 ==> DEP_DEL15=0 15534    conf:(0.85)
17. QUARTER=2 31592 ==> DEP_DEL15=0 26894    conf:(0.85)
18. DAY_OF_WEEK=2 17862 ==> DEP_DEL15=0 15195    conf:(0.85)
19. DISTANCE_GROUP=3 18620 ==> DEP_DEL15=0 15834    conf:(0.85)
20. DISTANCE_GROUP=2 30421 ==> DEP_DEL15=0 25802    conf:(0.85)
21. DAY_OF_WEEK=4 18425 ==> DEP_DEL15=0 15498    conf:(0.84)
22. DAY_OF_WEEK=1 18278 ==> DEP_DEL15=0 15263    conf:(0.84)
23. QUARTER=1 26647 ==> DEP_DEL15=0 22111    conf:(0.83)
24. DEP_DEL15=1 17809 ==> SECURITY_DELAY=0 13790    conf:(0.77)
25. DEP_DEL15=1 17809 ==> WEATHER_DELAY=0 13143    conf:(0.74)
26. DEP_DEL15=1 17809 ==> WEATHER_DELAY=0 SECURITY_DELAY=0 13129    conf:(0.74)
27. SECURITY_DELAY=0 19212 ==> DEP_DEL15=1 13790    conf:(0.72)
28. WEATHER_DELAY=0 SECURITY_DELAY=0 18448 ==> DEP_DEL15=1 13129    conf:(0.71)
29. WEATHER_DELAY=0 18471 ==> DEP_DEL15=1 13143    conf:(0.71)
30. WEATHER_DELAY=0 18471 ==> DEP_DEL15=1 SECURITY_DELAY=0 13129    conf:(0.71)
31. SECURITY_DELAY=0 19212 ==> DEP_DEL15=1 WEATHER_DELAY=0 13129    conf:(0.68)
32. DISTANCE_GROUP=4 21743 ==> CARRIER=DL 13157    conf:(0.61)
33. QUARTER=3 DEP_DEL15=0 28756 ==> CARRIER=DL 15274    conf:(0.53)
34. QUARTER=3 33374 ==> CARRIER=DL 17419    conf:(0.52)
35. QUARTER=2 DEP_DEL15=0 26894 ==> CARRIER=DL 13909    conf:(0.52)
36. DEP_DEL15=0 103798 ==> CARRIER=DL 53356    conf:(0.51)
37. QUARTER=4 DEP_DEL15=0 26037 ==> CARRIER=DL 13230    conf:(0.51)
38. QUARTER=2 31592 ==> CARRIER=DL 15955    conf:(0.51)
```

From the result, we can see that there are several rules show that when CARRIER=DL, it has high confidence that not delayed. It shows that QUARTER=4 and QUARTER=3 also have high confidence that not delayed. The most useful and highest confidence rule is the fifth one. It shows that when QUARTER=4 and CARRIER=DL has 89% confidence that not delayed.

Evaluation

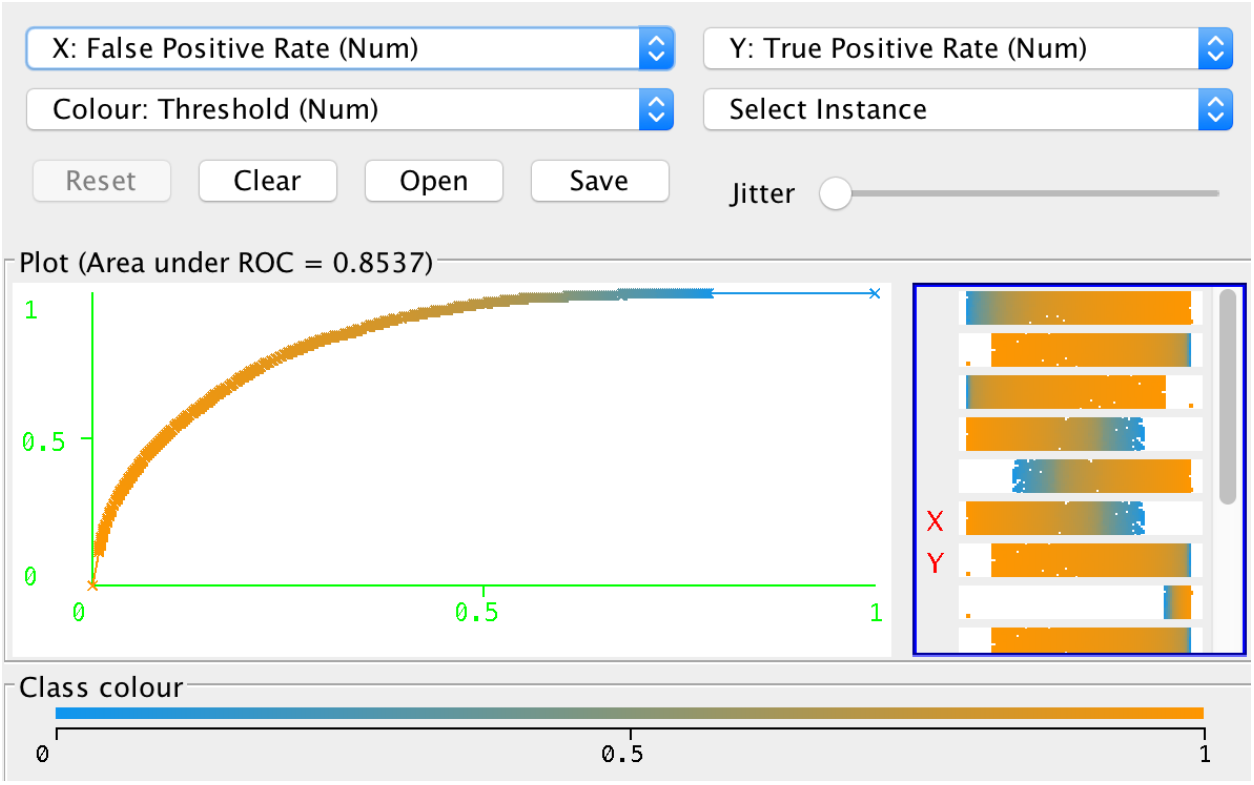
We used two models total. At first, we ran the J48 by using the dataset which concludes all the data from January in 2016. From the result, we got that the most informative attribute that causes the airline delay is NAS_Delay. However, the goal of our project is to find the useful information to give some advice to customer for saving their time, such as which quarter or which carrier customers choose can decrease the probability to meet the airline delay situation. The J48 decision tree model cannot provide specific answer to our goal. Also, the accuracy of the J48 decision tree is around 65% which means this model cannot be the main model of our project.

We focused on the airlines which depart from the MSP in 2015. We deleted some attributes which have more than 99% missing, and some other attributes which contain the same meaning. For example, from the attribute DEST, we can know the dest_state and dest_city, so we kept the DEST and deleted the other two attributes. Then, we ran the association rule.

After we got the result from association rule, we used the same attributes to do the J48 decision tree. At first time, we used 10 folds as cross-validation, and we got 89.8855% accuracy. Also, the model just made some minor errors, such as the mean absolute error is 0.1532.

```
Time taken to build model: 1.83 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      109307               89.8855 %
Incorrectly Classified Instances     12300               10.1145 %
Kappa statistic                          0.4829
Mean absolute error                      0.1532
Root mean squared error                  0.2834
Relative absolute error                 61.2878 %
Root relative squared error             80.1699 %
Total Number of Instances           121607

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0.608      0.904     0.986     0.943      0.854    0
                 0.392     0.014      0.826     0.392     0.532      0.854    1
Weighted Avg.    0.899     0.521      0.893     0.899     0.883      0.854

=== Confusion Matrix ===

      a      b   <-- classified as
 102325   1473 |      a = 0
  10827   6982 |      b = 1
```

Then, we plot the ROC curve of this model.



From this ROC curve, we can see that the Y axis means True Positive Rate and the X axis means the False Positive Rate. The true positive means that the model correctly classified the airlines which are not delayed. The trend of the curve shows that the True positive rate is higher than the false positive rate. It means that our model does the good job to classify the airlines which are not delayed. According to the curve, it shows that the area under ROC = 0.8573.

Then we used 30% as test set and trained the remaining data to prevent the overfitting. The percentage of accuracy around 92%, which means our model did a good job.

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      112472               92.4881 %
Incorrectly Classified Instances      9135                7.5119 %
Kappa statistic                          0.6396
Mean absolute error                      0.1253
Root mean squared error                  0.2503
Relative absolute error                 50.1031 %
Root relative squared error             70.7841 %
Total Number of Instances           121607
```

This model got around 90% accuracy. Overall, according to the result which we get from J48, our associate rule model is reliable and accurate. From the evaluation, we got that when analyze a business case, the first thing need to be decided is the goal of the whole

project. Then, thinking about using what kind of methods can approach to the final goal. And J48 decision tree model can help to figure out which attribute is related to the aim.

Prescriptive Analytics

We did the simulation to test the result which we get. Our result shows that when QUARTER=4 and CARRIER=DL, it will have less delay time.

First, we inserted a table into our original dataset, and divided it into two datasets. One of them being QUARTER=4 and CARRIER=DL, this dataset contains 14949 instances. Another one is QUARTER does not equal to 4 and CARRIER does not equal to DL, and this dataset contains 45341 instances.

Then, we used a formula which can random choose 1000 instances from the dataset. This formula contains two steps. The first step is add a RAND() cell at the end of the first row, and drag that down.

=RAND()

AB	AC	AD	AE	AF	AG	AH
ARRIER_DE	WEATHER_D	NAS_DELAY	SECURITY_D	LATE_AIRCR		
					0.1913726	
					0.54719266	
					0.06116082	
					0.09169067	
15	0	0	0	0	0.29610617	
					0.11795668	
					0.913085	
					0.7539365	
21	0	1	0	23	0.52939045	
					0.48524197	
					0.24057061	
					0.55954759	
					0.57940957	
					0.34012595	
					0.4797393	
					0.77987634	
					0.8350536	
0	0	0	0	17	0.14613897	
0	0	32	0	0	0.07590696	
0	0	47	0	0	0.02083811	
					0.24095414	
					0.71530915	
					0.42627426	
					0.43140997	
					0.08201621	
0	0	41	0	0	0.67957675	

The second step is add the formula in anywhere in the spreadsheet.

Formula: INDEX(A:A,MATCH(LARGE($AG:$AG,ROW(A1)),$AG:$AG,0))

fx =INDEX(A:A,MATCH(LARGE($AG:$AG,ROW(A1)),$AG:$AG,0))

Then, drag the lower right corner of the cell to spread the whole sheet to 1000 rows. That are the random 1000 instances which we choose from the dataset.

QUARTER=4 and CARRIER=DL

=INDEX(A:A,MATCH(LARGE($AG:$AG,ROW(A1)),$AG:$AG,0))

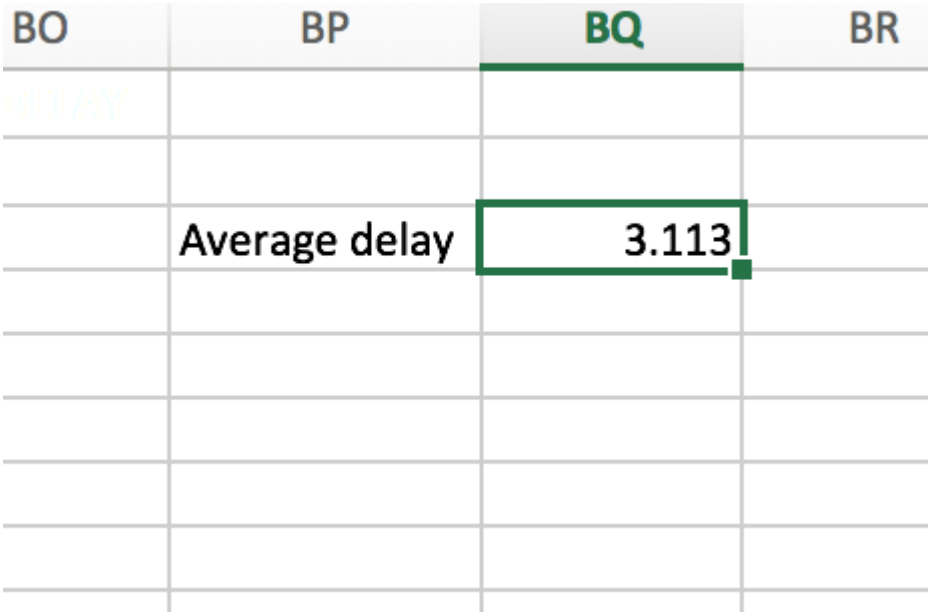
AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
ATE_AIRCR	Random		QUARTER	MONTH	DAY_OF_MC	DAY_OF_WE	CARRIER	FL_NUM	ORIGIN	DEST	DEST_CITY_I	DEST_STATE	CRS_DEP_TI	DEP_TIME	DEP_DELAY	DEP_DEL15	DEP_DELAY_	DEP_TIME_B	CRS_ARR_TI	ARR_TIM
75	0.05359173		4	12	2	3	DL	1188	MSP	PHL	Philadelphia,	Pennsylvania	1016	1024	8	0	0	1000-1059	1347	14
	0.49878204		4	11	29	7	DL	1108	MSP	ATL	Atlanta, GA	Georgia	1520	1518	-2	0	-1	1500-1559	1846	18
	0.96833262		4	11	19	4	DL	2624	MSP	IND	Indianapolis,	Indiana	1246	1244	-2	0	-1	1200-1259	1524	15
	0.37095512		4	10	15	4	DL	787	MSP	SLC	Salt Lake Cit	Utah	700	657	-3	0	-1	0700-0759	849	8
	0.75203386		4	12	10	4	DL	869	MSP	MCI	Kansas City, I	Missouri	921	917	-4	0	-1	0900-0959	1055	10
	0.23397647		4	10	11	7	DL	2279	MSP	SJC	San Jose, CA	California	1742	1740	-2	0	-1	1700-1759	1942	19
	0.97669362		4	11	28	6	DL	1996	MSP	LGA	New York, N	New York	700	659	-1	0	-1	0700-0759	1050	10
	0.72248278		4	11	27	5	DL	739	MSP	SFO	San Francisc	California	1430	1453	23	1	1	1400-1459	1639	16
	0.31203604		4	10	19	1	DL	808	MSP	BOS	Boston, MA	Massachuset	700	659	-1	0	-1	0700-0759	1050	10
	0.69321457		4	12	19	6	DL	2816	MSP	SFO	San Francisc	California	1145	1244	59	1	3	1100-1159	1403	14
	0.28249807		4	12	10	4	DL	2163	MSP	PDX	Portland, OR	Oregon	900	856	-4	0	-1	0900-0959	1056	10
	0.78888219		4	11	13	5	DL	1568	MSP	DCA	Washington,	Virginia	1508	1533	25	1	1	1500-1559	1831	18
	0.08955862		4	11	25	3	DL	2243	MSP	TPA	Tampa, FL	Florida	903	858	-5	0	-1	0900-0959	1315	12
	0.69691184		4	10	24	6	DL	1717	MSP	ORD	Chicago, IL	Illinois	1502	1458	-4	0	-1	1500-1559	1630	16
	0.272408		4	12	21	1	DL	1975	MSP	ATL	Atlanta, GA	Georgia	1129	1128	-1	0	-1	1100-1159	1500	14
	0.29554653		4	12	20	7	DL	513	MSP	SAN	San Diego, C	California	1135	1134	-1	0	-1	1100-1159	1332	13
	0.8657945		4	10	6	2	DL	951	MSP	BOI	Boise, ID	Idaho	2202	2156	-6	0	-1	2200-2259	3	24
	0.54273384		4	10	6	2	DL	982	MSP	BWI	Baltimore, M	Maryland	1505	1517	12	0	0	1500-1559	1833	18
	0.16275504		4	11	18	3	DL	2336	MSP	MKE	Milwaukee, \	Wisconsin	1949	2003	14	0	0	1900-1959	2107	21
	0.0707045		4	10	26	1	DL	1122	MSP	ATL	Atlanta, GA	Georgia	730	729	-1	0	-1	0700-0759	1055	10
	0.57157028		4	12	29	2	DL	848	MSP	BOS	Boston, MA	Massachuset	1315	1335	20	1	1	1300-1359	1659	17
	0.35904506		4	12	9	3	DL	2253	MSP	DCA	Washington,	Virginia	1735	1731	-4	0	-1	1700-1759	2102	20
	0.74374423		4	10	23	5	DL	963	MSP	FSD	Sioux Falls, S	South Dakot	2155	2359	124	1	8	2100-2159	2251	1
	0.84074062		4	11	4	3	DL	1544	MSP	STL	St. Louis, MC	Missouri	1505	1502	-3	0	-1	1500-1559	1634	16
	0.8145089		4	10	6	2	DL	921	MSP	MKE	Milwaukee, \	Wisconsin	659	703	4	0	0	0600-0659	809	8
	0.47609277		4	10	4	7	DL	2115	MSP	ABQ	Albuquerque	New Mexico	1135	1135	0	0	0	1100-1159	1315	13
	0.31820376		4	10	3	6	DL	2503	MSP	MCO	Orlando, FL	Florida	1740	1801	21	1	1	1700-1759	2150	22
0	0.92275363		4	12	26	6	DL	2392	MSP	PHX	Phoenix, AZ	Arizona	915	914	-1	0	-1	0900-0959	1151	11
	0.36929764		4	10	6	2	DL	2443	MSP	ALB	Albany, NY	New York	1914	1916	2	0	0	1900-1959	2239	22
	0.04312681		4	11	22	7	DL	589	MSP	SEA	Seattle, WA	Washington	905	904	-1	0	-1	0900-0959	1100	10
	0.72286566		4	12	3	4	DL	2488	MSP	PHL	Philadelphia,	Pennsylvania	1300	1259	-1	0	-1	1300-1359	1637	16
	0.91915118		4	10	1	4	DL	1996	MSP	LGA	New York, N	New York	700	700	0	0	0	0700-0759	1049	10
	0.59965643		4	12	21	1	DL	744	MSP	ATL	Atlanta, GA	Georgia	958	1009	11	0	0	0900-0959	1328	13
	0.4532427		4	10	27	2	DL	2037	MSP	BZN	Bozeman, M	Montana	1958	2004	6	0	0	1900-1959	2135	21
	0.33177265		4	10	2	5	DL	2193	MSP	SEA	Seattle, WA	Washington	705	659	-6	0	-1	0700-0759	847	8
	0.42562706		4	10	16	5	DL	2256	MSP	DTW	Detroit, MI	Michigan	1428	1424	-4	0	-1	1400-1459	1703	17
	0.68106358		4	11	8	7	DL	846	MSP	SLC	Salt Lake Cit	Utah	907	901	-6	0	-1	0900-0959	1100	10
	0.47314251		4	12	13	7	DL	1512	MSP	MKE	Milwaukee, \	Wisconsin	1017	1011	-6	0	-1	1000-1059	1130	11
	0.87608807		4	12	27	7	DL	1897	MSP	RSW	Fort Myers, I	Florida	1304	1302	-2	0	-1	1300-1359	1725	17
	0.99782876		4	11	21	6	DL	2371	MSP	SMF	Sacramento,	California	1742	1735	-7	0	-1	1700-1759	1943	19

QUARTER does not =4 and CARRIER does not = DL

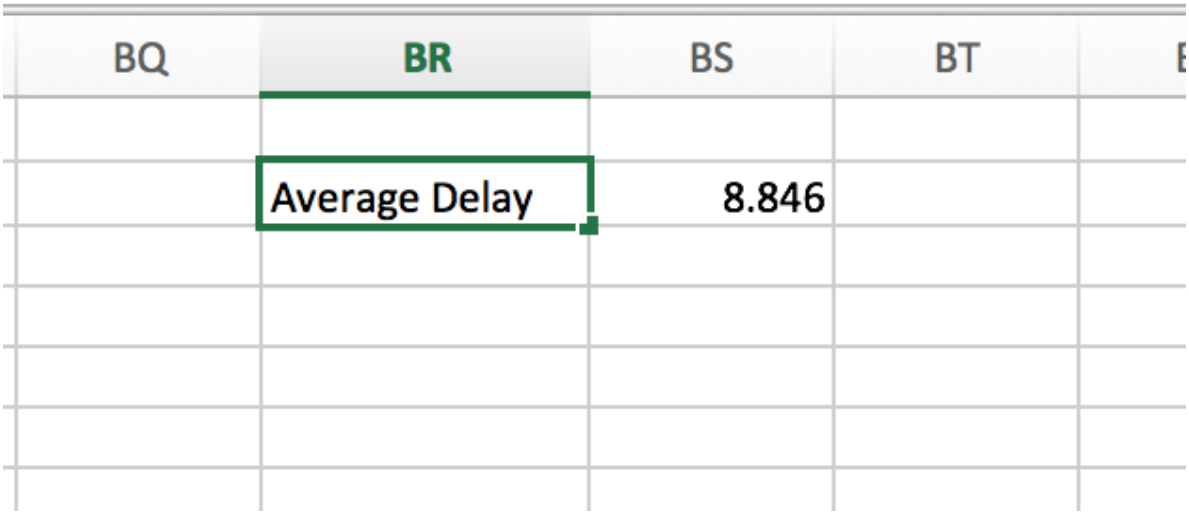
=INDEX(A:A,MATCH(LARGE($AG:$AG,ROW(A1)),$AG:$AG,0))

AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
QUARTER	MONTH	DAY_OF_MC	DAY_OF_WE	CARRIER	FL_NUM	ORIGIN	DEST	DEST_CITY_I	DEST_STATE	CRS_DEP_TI	DEP_TIME	DEP_DELAY	DEP_DEL15	DEP_DELAY_	DEP_TIME_B	CRS_ARR_TI
1	3	20	5	EV	4940	MSP	RDU	Raleigh/Durh	North Carolir	1510	1508	-2	0	-1	1500-1559	1836
2	4	28	2	AA	1115	MSP	ORD	Chicago, IL	Illinois	1408	1442	34	1	2	1400-1459	1535
3	7	20	1	F9	326	MSP	PHL	Philadelphia,	Pennsylvania	935	944	9	0	0	0900-0959	1310
3	9	11	5	WN	3045	MSP	DEN	Denver, CO	Colorado	630	626	-4	0	-1	0600-0659	735
2	6	17	3	OO	4722	MSP	GRB	Green Bay, V	Wisconsin	734	729	-5	0	-1	0700-0759	843
3	8	6	4	NK	281	MSP	IAH	Houston, TX	Texas	2005	2053	48	1	3	2000-2059	2244
1	2	25	3	WN	779	MSP	ATL	Atlanta, GA	Georgia	920	917	-3	0	-1	0900-0959	1305
2	6	11	4	WN	3935	MSP	MDW	Chicago, IL	Illinois	1130	1130	0	0	0	1100-1159	1255
3	7	15	3	WN	4480	MSP	DEN	Denver, CO	Colorado	2030	2136	66	1	4	2000-2059	2135
2	5	15	5	OO	4857	MSP	RAP	Rapid City, S	South Dakot	2142	2145	3	0	0	2100-2159	2224
2	6	19	5	AS	37	MSP	SEA	Seattle, WA	Washington	1630	1630	0	0	0	1600-1659	1804
1	3	26	4	OO	4591	MSP	GRB	Green Bay, V	Wisconsin	922	917	-5	0	-1	0900-0959	1031
3	9	12	6	OO	4715	MSP	ISN	Williston, ND	North Dakot	1941	1935	-6	0	-1	1900-1959	2140
2	5	9	6	OO	4838	MSP	DAY	Dayton, OH	Ohio	1720	1715	-5	0	-1	1700-1759	2011
2	5	30	6	OO	7403	MSP	HIB	Hibbing, MN	Minnesota	1952	2048	56	1	3	1900-1959	2053
1	2	20	5	OO	4554	MSP	OKC	Oklahoma Ci	Oklahoma	2003	2041	38	1	2	2000-2059	2220
1	3	12	4	AA	1249	MSP	ORD	Chicago, IL	Illinois	720	716	-4	0	-1	0700-0759	855
1	1	9	5	OO	4483	MSP	BOI	Boise, ID	Idaho	1140	1200	20	1	1	1100-1159	1354
3	8	19	3	F9	326	MSP	PHL	Philadelphia,	Pennsylvania	940	932	-8	0	-1	0900-0959	1310
3	9	14	1	OO	4444	MSP	BIS	Bismarck/Ma	North Dakot	1727	1725	-2	0	-1	1700-1759	1900
3	9	7	1	AA	1833	MSP	PHL	Philadelphia,	Pennsylvania	1355	1347	-8	0	-1	1300-1359	1731
3	8	10	1	OO	4714	MSP	ATW	Appleton, W	Wisconsin	1557	1552	-5	0	-1	1500-1559	1702
1	3	20	5	AA	2253	MSP	DFW	Dallas/Fort V	Texas	654	651	-3	0	-1	0600-0659	940
3	7	2	4	OO	4485	MSP	FNT	Flint, MI	Michigan	1301	1256	-5	0	-1	1300-1359	1535
3	8	18	2	OO	4687	MSP	HLN	Helena, MT	Montana	2016	2014	-2	0	-1	2000-2059	2158
3	8	13	4	NK	763	MSP	LAS	Las Vegas, N	Nevada	2155	2301	66	1	4	2100-2159	2303
2	4	5	7	OO	4857	MSP	RAP	Rapid City, S	South Dakot	2020	2018	-2	0	-1	2000-2059	2104
1	2	16	1	OO	7371	MSP	BJI	Bemidji, MN	Minnesota	1120	1116	-4	0	-1	1100-1159	1232
2	4	7	2	OO	4576	MSP	RAP	Rapid City, S	South Dakot	900	857	-3	0	-1	0900-0959	949
1	1	6	2	OO	7372	MSP	BJI	Bemidji, MN	Minnesota	2150	2150	0	0	0	2100-2159	2301
1	1	11	7	OO	4829	MSP	FAR	Fargo, ND	North Dakot	930	930	0	0	0	0900-0959	1045
3	8	9	7	OO	4804	MSP	TUL	Tulsa, OK	Oklahoma	1417	1419	2	0	0	1400-1459	1613
3	7	6	1	WN	5699	MSP	STL	St. Louis, MC	Missouri	1945	2004	19	1	1	1900-1959	2115
1	3	14	6	EV	5041	MSP	CLT	Charlotte, N	North Carolir	1015	1012	-3	0	-1	1000-1059	1344
1	3	2	1	OO	4829	MSP	LSE	La Crosse, W	Wisconsin	959	952	-7	0	-1	0900-0959	1055
3	9	13	7	OO	4615	MSP	GFK	Grand Forks,	North Dakot	1740	1739	-1	0	-1	1700-1759	1858
3	8	20	4	OO	4587	MSP	GFK	Grand Forks,	North Dakot	916	915	-1	0	-1	0900-0959	1034
2	6	15	1	OO	4513	MSP	BNA	Nashville, TN	Tennessee	1720	1739	19	1	1	1700-1759	1924
2	6	26	5	OO	4665	MSP	BMI	Bloomington	Illinois	1511	1516	5	0	0	1500-1559	1640
2	4	2	4	OO	4532	MSP	GRB	Green Bay, V	Wisconsin	1937	1951	14	0	0	1900-1959	2051

After that, we used the AVERAGE() function to calculate the average dep_delay time. The calculation shows that when the QUARTER=4 and CARRIER=DL the average dep_delay time =3.113



Another simulation shows that the average dep_delay time= 8.846



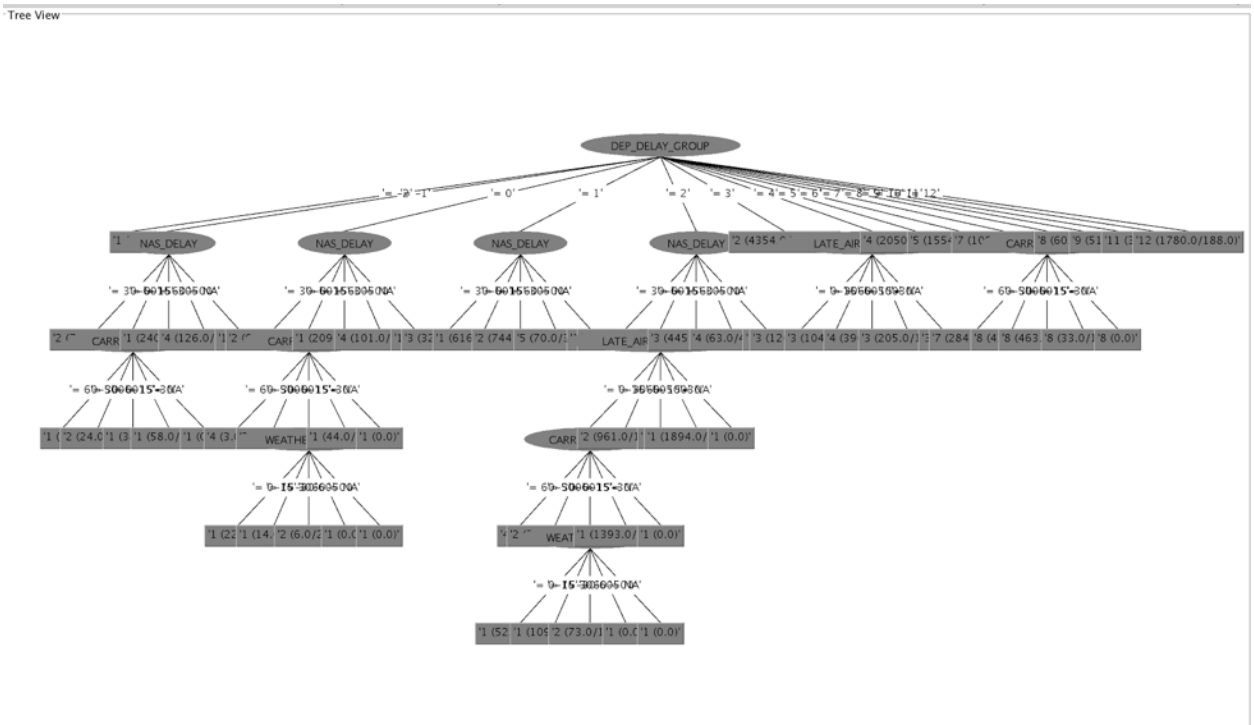
Because we used the RAND() function, every time we clicked the Calculate button, the excel will automatically choose another 1000 random instances. We clicked that button several times, and all of the results show that if the instances follow our rule, the average delay time is less than those instances which do not follow our rule.

Conclusion

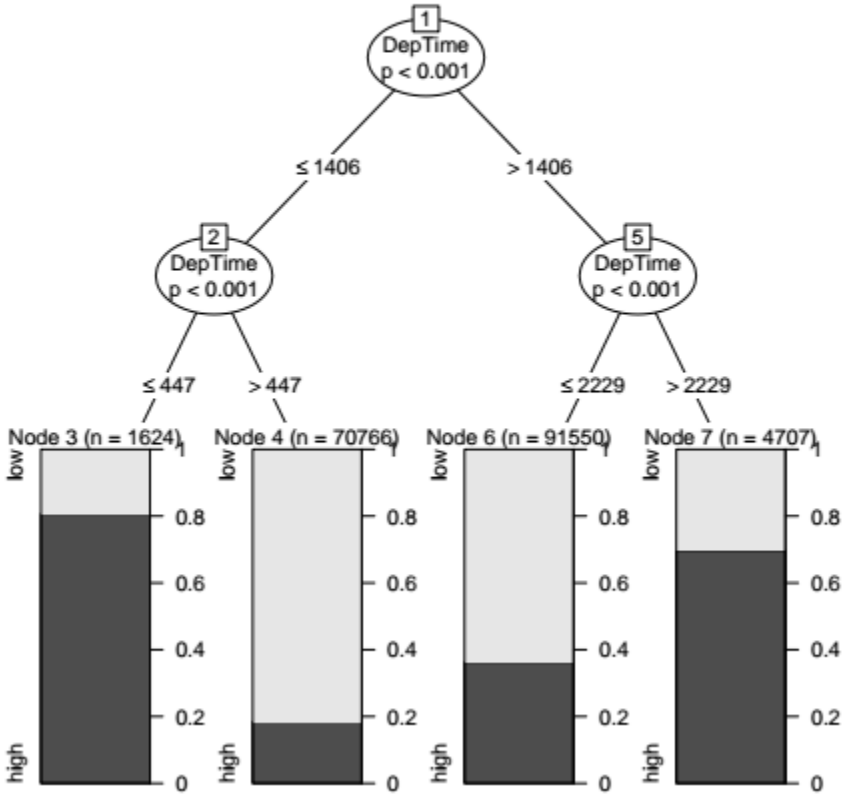
This simulation proves that the result which we get from the association rule is right. It means that if people choose flight in quarter 4 and the carrier is DL, they can save more time and have lower probability to meet the departure delay situation.

Comparison with Documented Results

One of the documented results that we are comparing with is an overall Data Analysis of U.S. Airlines On-time Performance authored by Yanxiang Zhu, Nilesh Padwal, and Mingxuan Li. Their document was finished in June of 2014. Their data only contained 17 attributes with about 168,000 instances while our dataset had more attributes with 53 but fewer instances, 38,764. One other major difference is that they mainly did all of their work in R and used Rpart and Ctree for their decision making trees, while we also used R for grouping and organizing our data, we decided to use Weka and the tool J48 for our decision making trees. They also used cluster analysis through PAM and KMeans. They were essentially looking for the variables that have a significant impact when it comes to delays while we wanted to focus on first finding the most common delay that is happening and finding out the correlations to it such as what day and what time they occur. We proceeded to find the variables that have an impact on delay. Our variables of interest are day of week, distance, travel company and time of day. There are others that we are looking into as well.



Here is our decision tree done in weka which differs from what was done in the other studies. We did this do discover what type of delay is most common while other studies went other routes to find the time of day with most delays.



Here is an image of the CTree used in the other study that shows the time of day in relation of the delays. It shows that at 10:30PM- 5:00 AM delays are more common than compared to the day.

Another documented result that we are going to be comparing with is a report by Raj Bandyopadhyay and Rafael Guerrero called Predicting Airline Delays. This report uses similar tools as ours such as Weka. However, they go about their analysis a little bit differently. One of the major differences with our report and theirs is that they chose only one airline (American

Airlines) from one airport (O'Hare in Chicago). They also chose to use a Naive-Bayes classifier instead of a J48 decision tree. The J48 decision tree showed us what the most common form of delay is.

To further our project we will model our project and compare it after this. Instead of the Chicago airport we will be using data from the Minneapolis/St. Paul airport because it is more relevant to our class.

Table 1: Performance of classifiers in predicting non-delayed flights

	Accuracy %	Precision	Recall	F-score
Naive-bayes	70.8	0.752	0.81	0.78
SVM (unweighted)	71.65	0.737	0.864	0.796
SVM (weighted, 1 FN = 10 FP)	49.2	0.909	0.227	0.363
Random Forests (unweighted)	70.94	0.784	0.813	0.798
Random Forests (weighted)	58.01	0.839	0.502	0.628

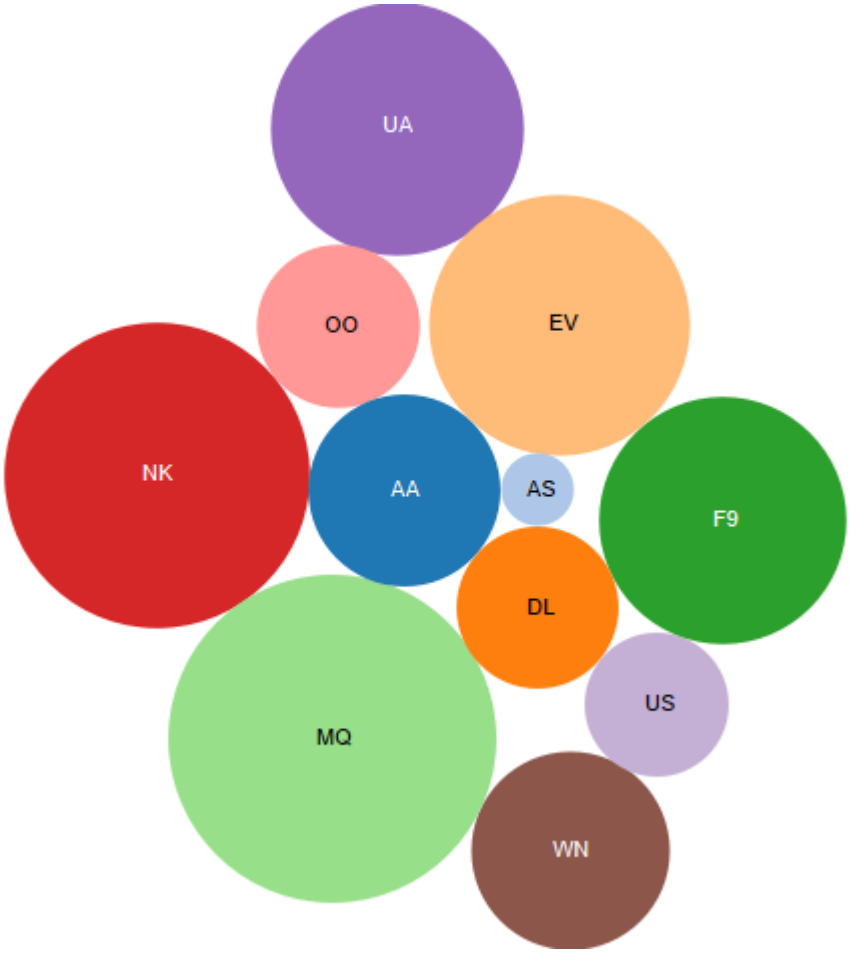
Table 2: Performance of classifiers in predicting delayed flights

	Accuracy %	Precision	Recall	F-score
Naive-bayes	70.8	0.612	0.527	0.566
SVM (unweighted)	71.65	0.655	0.455	0.538
SVM (weighted, 1 FN = 10 FP)	49.2	0.413	0.86	0.577
Random Forests (unweighted)	70.94	0.507	0.461	0.483
Random Forests (weighted)	58.01	0.391	0.769	0.519

By using a Naive-bayes algorithm this study was able to determine that they were better at predicting non-delayed flights than delayed ones. Our model differs because we did not use a Naive-bayes algorithm to predict whether or not a customer will face a delay.

Deployment

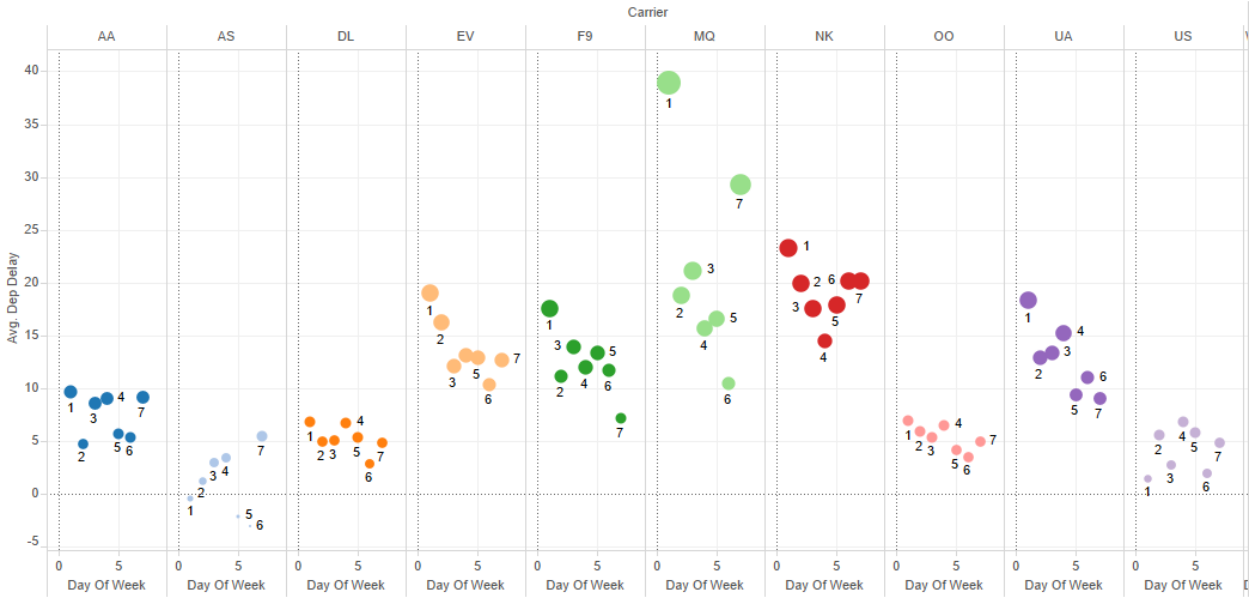
Throughout this project we have found a lot of useful information. Our goal to deploy this is to create a website that is linked to the Department of Transportation that will use visualization to show people when the best time to book a flight is. Customers will be able to see for themselves which airline carriers are the most reliable when it comes to departure and arrival delays. We will include visualizations like the following which shows the average departure delay per carrier.



Our website will be modeled after the Leapfrog website that helps to improve the safety and quality of healthcare. By informing customers what the likelihood of a delay is the

quality of their travel will improve greatly. The website we deploy will have links to the DoT airline data so customers can look through it themselves, we will also have the charts and data be dynamic so it is changing in time with the airlines. This way airline companies can improve their image if they begin as a lagger in the industry.

To further improve travel quality we will show customers the day of week with the highest probability of delay. The website will contain a graph the resembles this which shows the average delay for each day of the week at each airline. With this customers will be able to plan trips based on which airline they would like to travel with if they can choose between going on different days.



Issues in our deployment would be to link our models to the DoT data or to involve our findings in their website directly. An ethical problem that comes up with this is the integrity of the data. With the DoT website there were many missing values that could alter the data. If we are to link to this website some airline companies will be at a disadvantage

because they do not have the data to properly represent them. Our team will be sure to give every airline a fair shot at being represented in a good light.

This information does have some risks that may cause problems with the airline companies. Since our conclusion is the best time of year to take a flight is quarter four at Delta this may cause an increase in demand at this time and place. Without proper warning or preparation they may reach this unexpected demand without having a way to reach it. We would resolve this issue by informing the airlines of our findings and tell them we plan on publishing them so they will be aware of the possible influx of new customers at this time.

Another issue is getting sued by airlines by publishing stats that paint their business in a bad light and causes them to lose business. To avoid this our team will make sure to use the clean data that is always accurate. We will also work with legal teams to ensure that we are not stepping out of our means and publishing information that should not be published. A mistake we could make is modeling the data incorrectly or interpreting it incorrectly. To avoid this we will work with each other to double check our work and guarantee that our publications are accurate.