**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

Screenshot of GitHub Submission…………………………………………………………………..

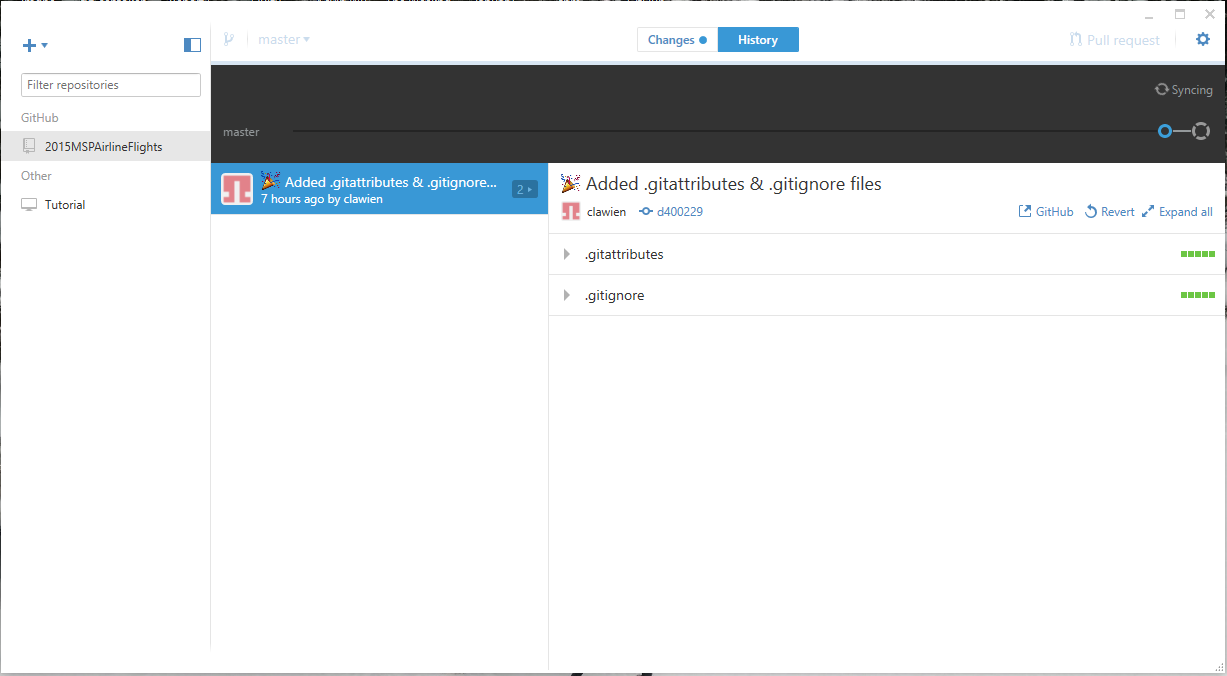
Corrected Contents from Project #2……………………………………………………………….

Data Understanding……………………………………………………………………………………....

Data Preparation…………………………………………………………………………………………..

Descriptive Analytics…...

Screenshot of GitHub Submission



\*Application was not syncing with GitHub

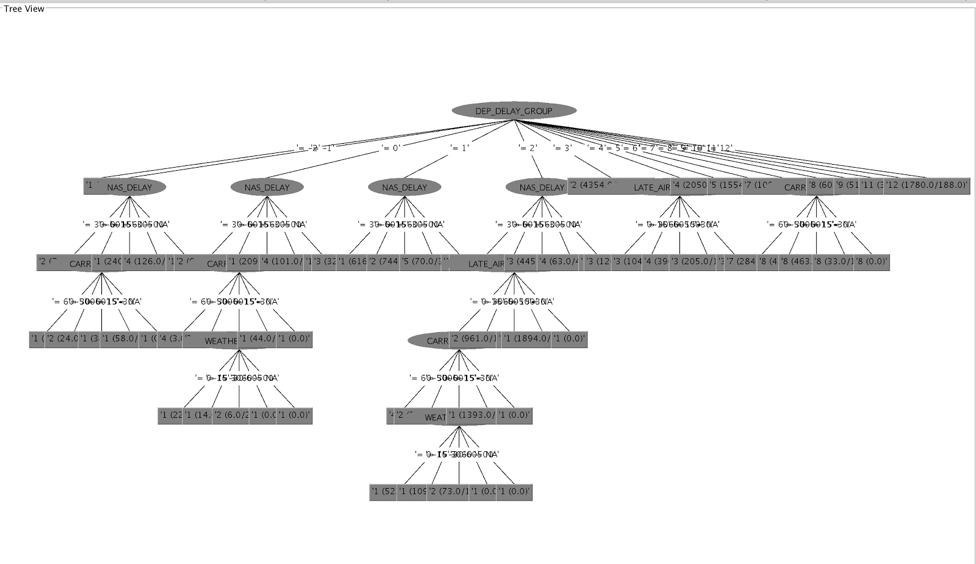
Project #2 Revisions

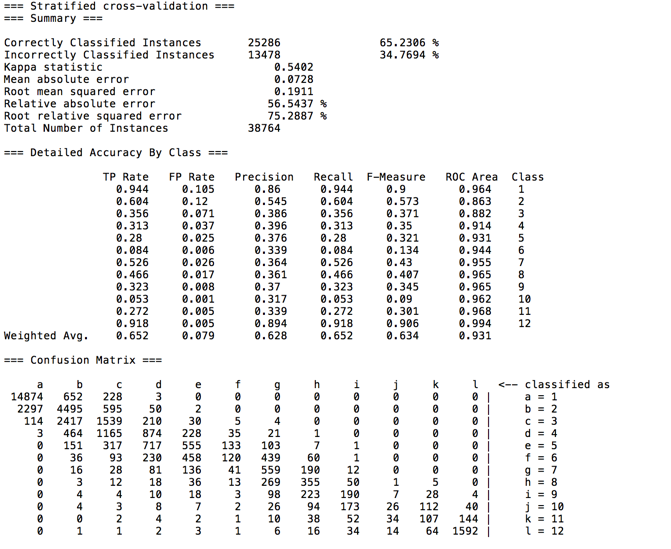
**Descriptive Analytics: 25/30 -** “Good work but could have used scatter plots or other charts to show not just individual performance but also patterns.”

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: 10/20 -** “Good start but more can be done here. Also there is no conclusion provided from the model.”

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.





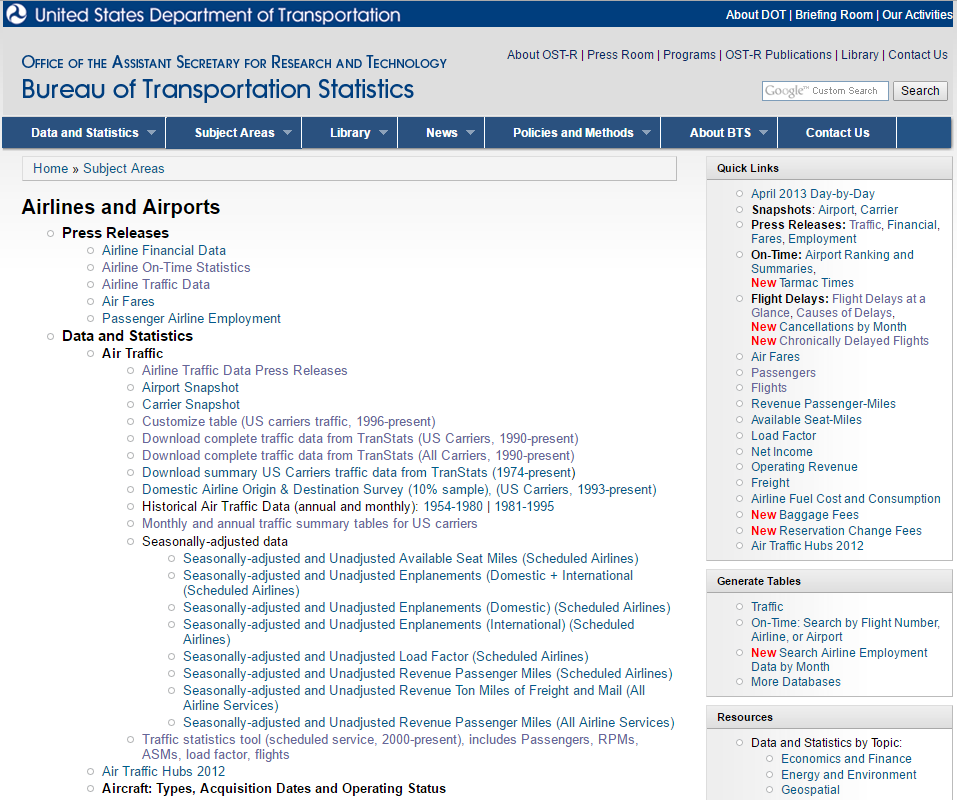
Conclusion:

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.



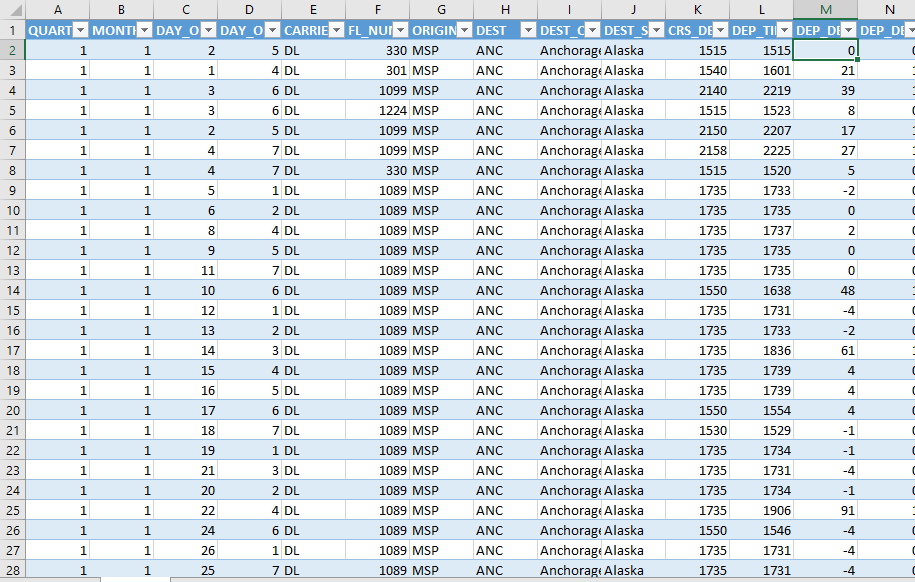
With the large amount of data included in this dataset, narrowing down the size was a top priority. For Project #2 it was decided to focus only on the month of January 2016. While working with this data it soon became clear that finding accurate results would be difficult given such a small timeframe 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…

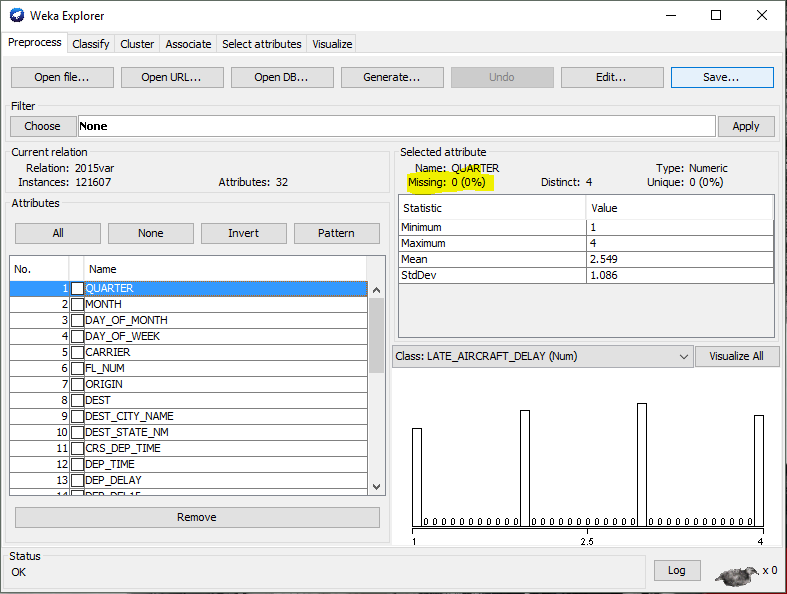


Data Preparation

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.

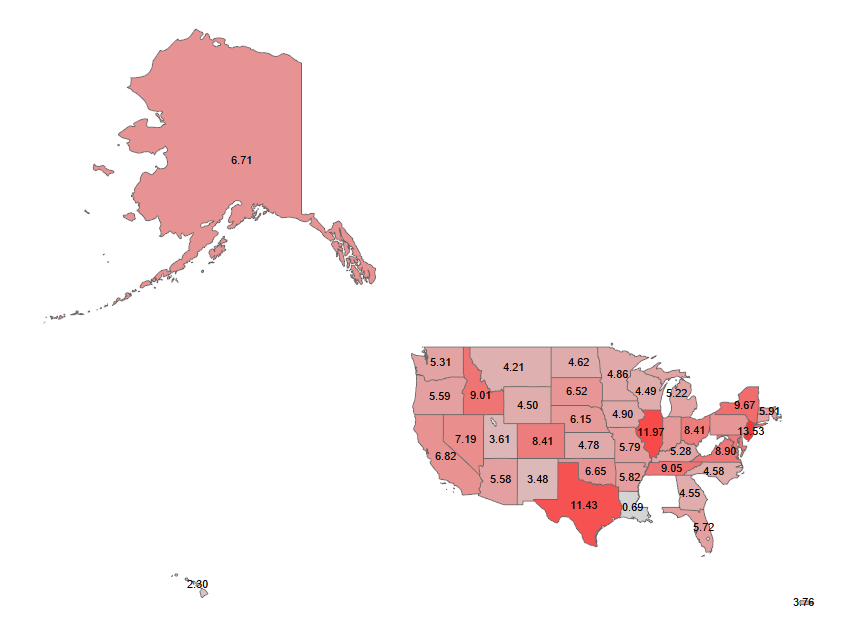
A concern all along was dealing with missing data. When first selecting which variables to use, only variables with an average of 0.0005% missing values were used to help limit the number of missing values in our final dataset. Weka includes a great feature that counts the number of missing entries in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



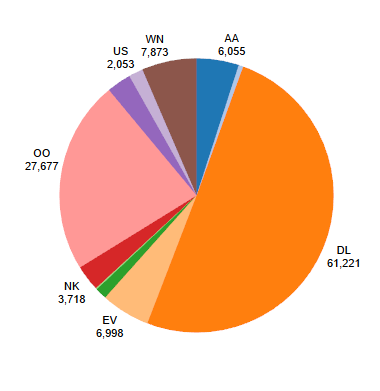
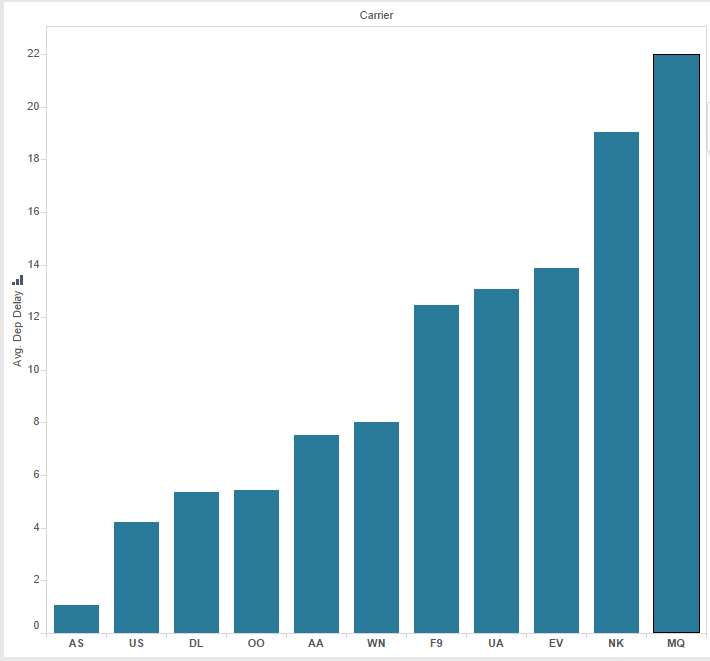


Descriptive Analytics

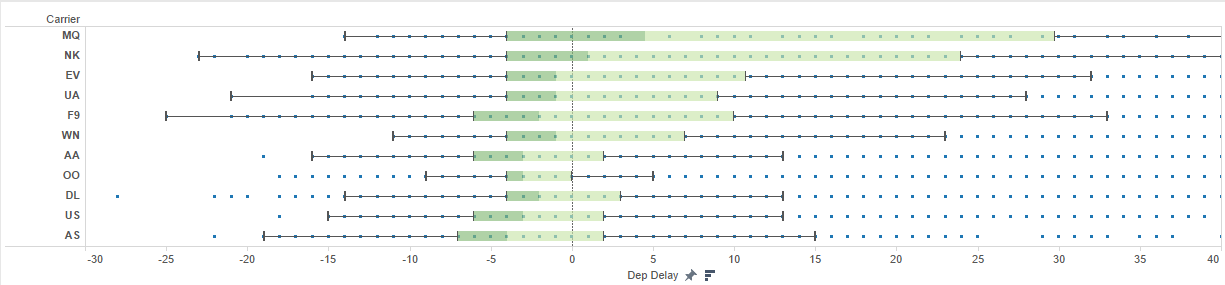
Tableau was the primary visualization tool used for this project. Saving a copy of the dataset in a .xlsx format allows Tableau to easily read and open the data set. Through this program the data can be manipulated in thousands of ways to gain early insights on trends.



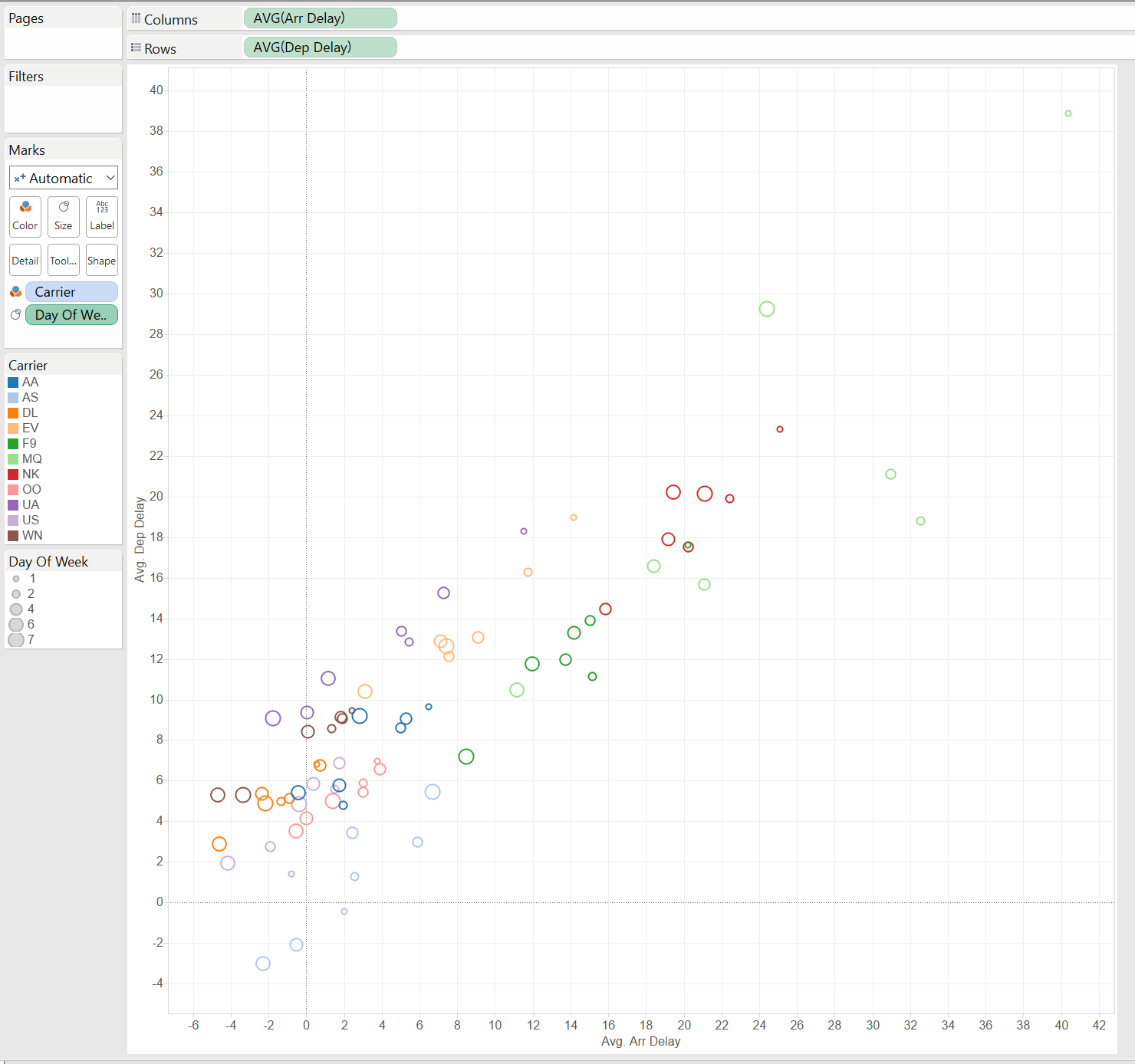
The chart above shows the average delay times of flights departing the Minneapolis International Airport (MSP) in the year of 2015 and arriving at airports in each state. Some states are missing from the chart due to no flights actually flying to those states. This chart shows you which arrival airports 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.



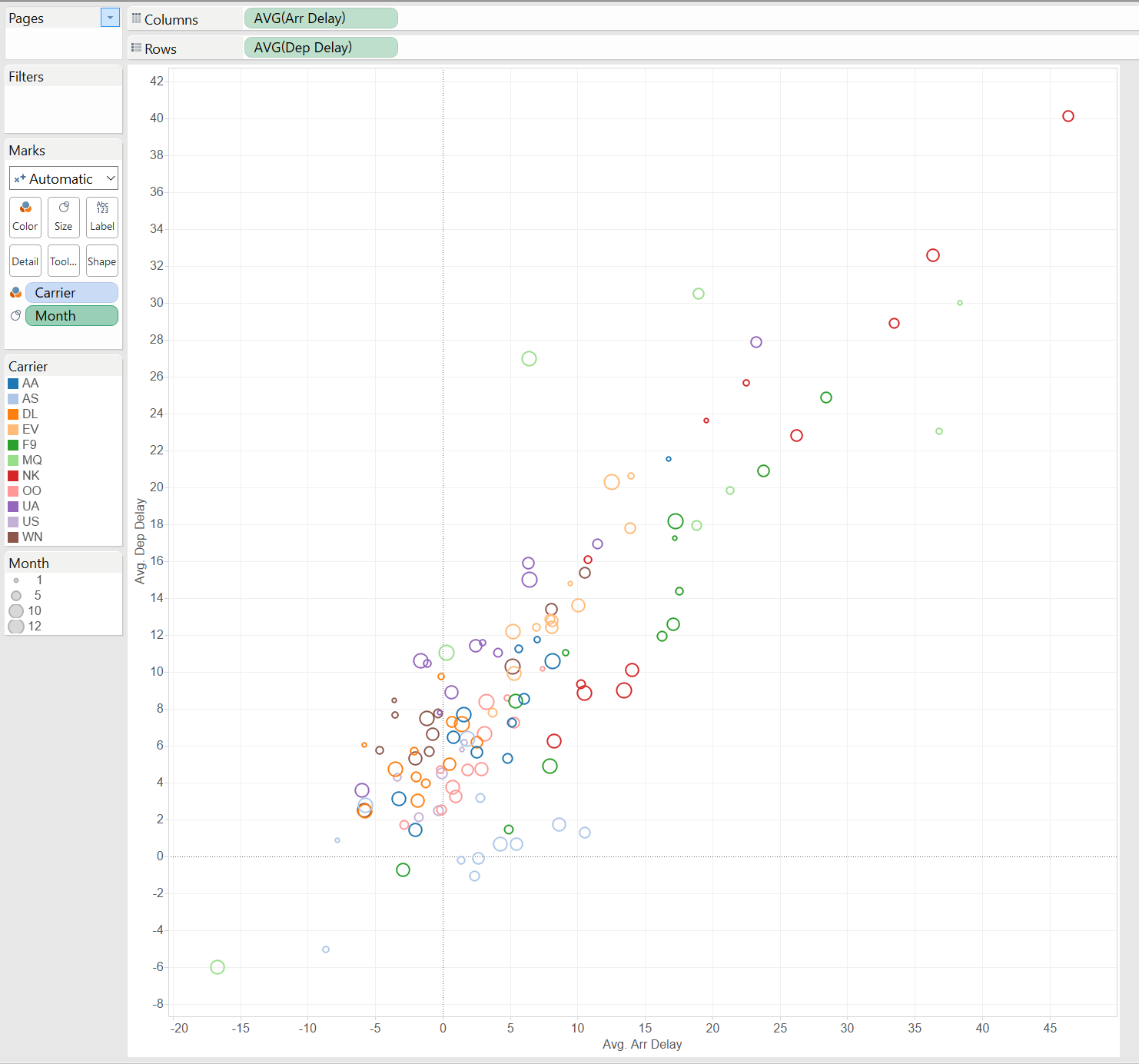
In the graph and pie chart above the data shows that departure delay times by airline are not related to the number of flights in the year by each of those airlines. The Minneapolis International Airport is one of Delta’s main hubs in the United States, accounting for nearly half the total flights departing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



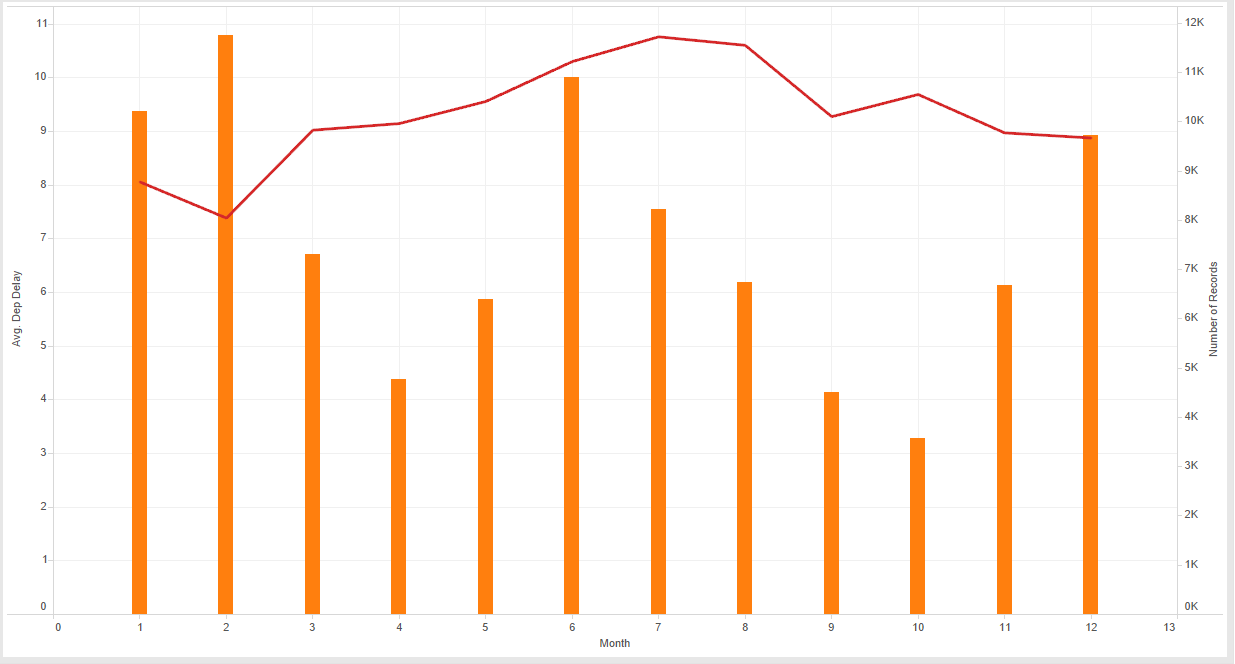
The box plot above shows the Upper Whisker, Upper Quartile, Median, Lower Quartile, and Lower Whisker of the delay times of the 11 airlines that fly out of the Minneapolis International Airport.



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.

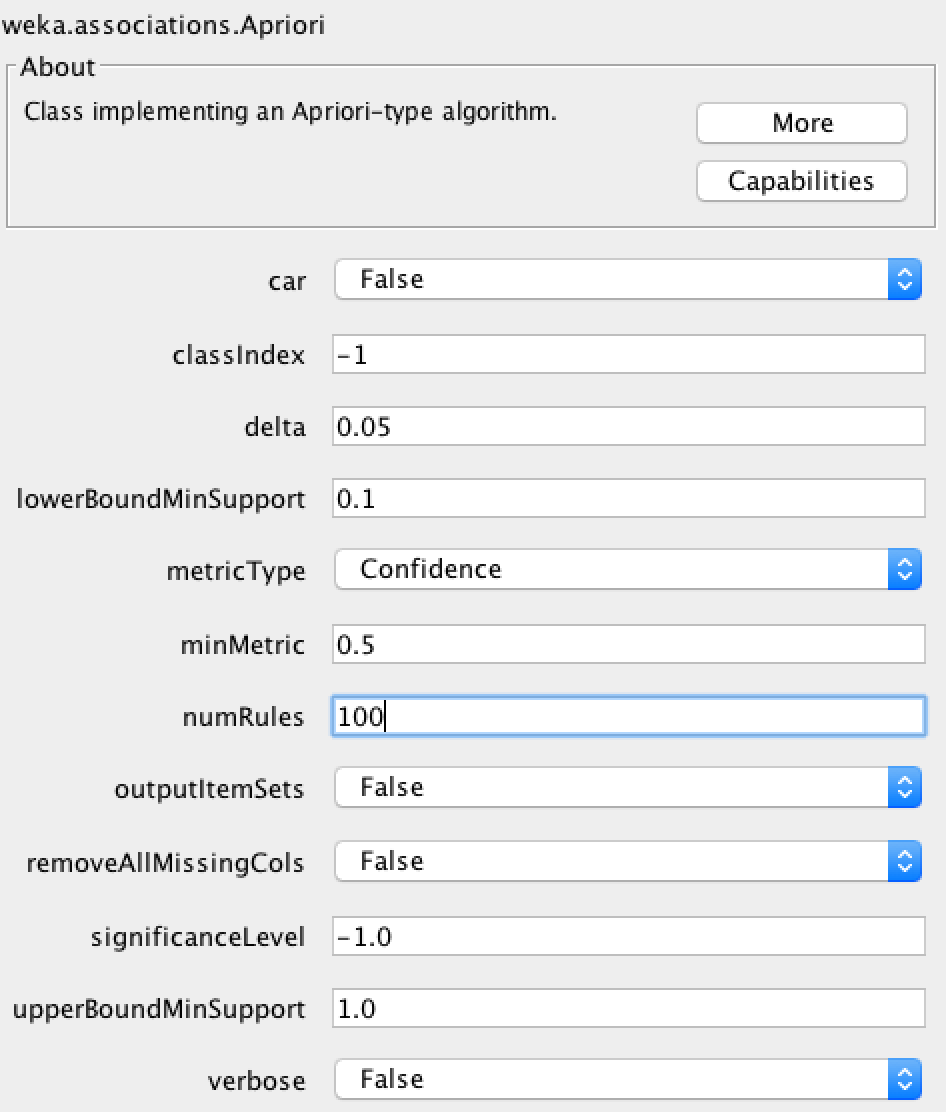


Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a [relational database](http://searchsqlserver.techtarget.com/definition/relational-database) or other information repository. We want to know which variables are related to the delay time. If 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.

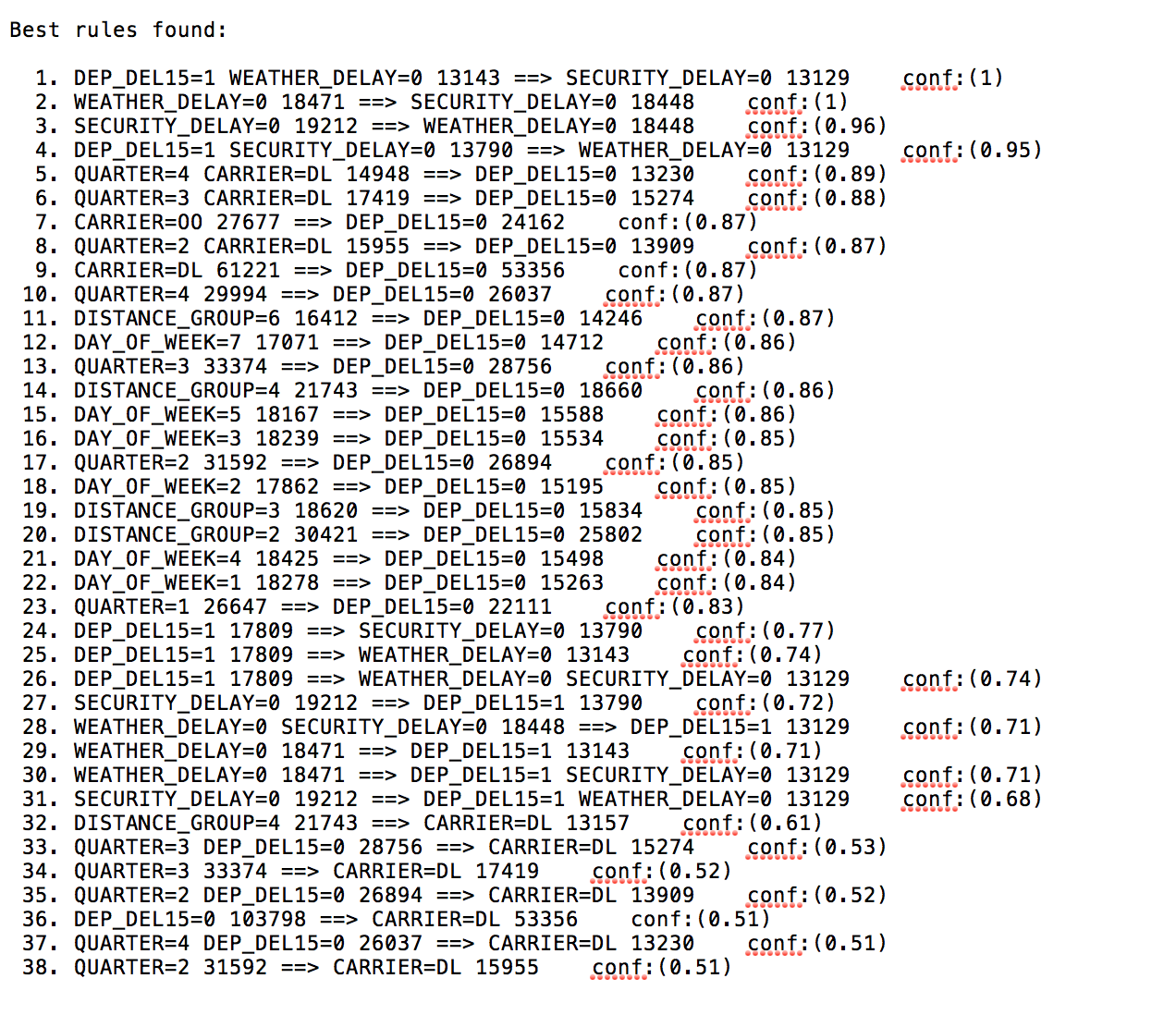
Expanded Modeling and Predictive Analytics

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:



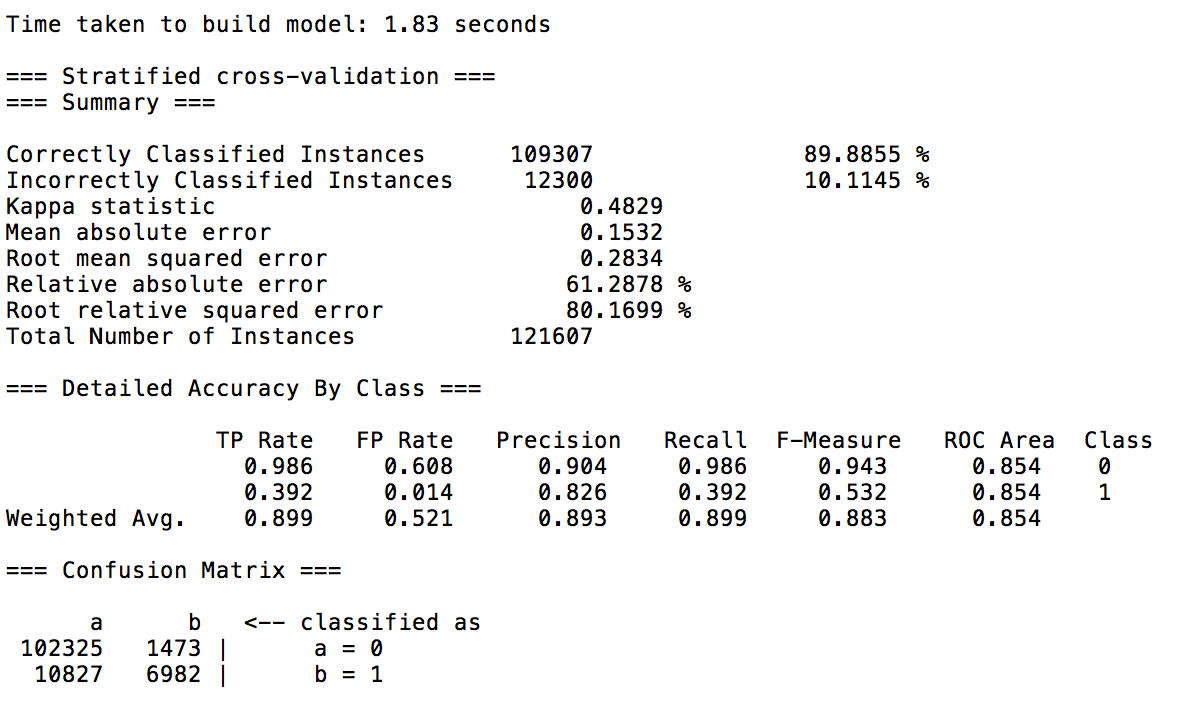
From the result, we can see that there are several rules show that when CARRIER=DL, it has high confidence that not delayed. It shows that QUARTER=4 and QUARTER=3 also have high confidence that not delayed. The most useful and highest confidence rule is the fifth one. It shows that when QUARTER=4 and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

Evaluation

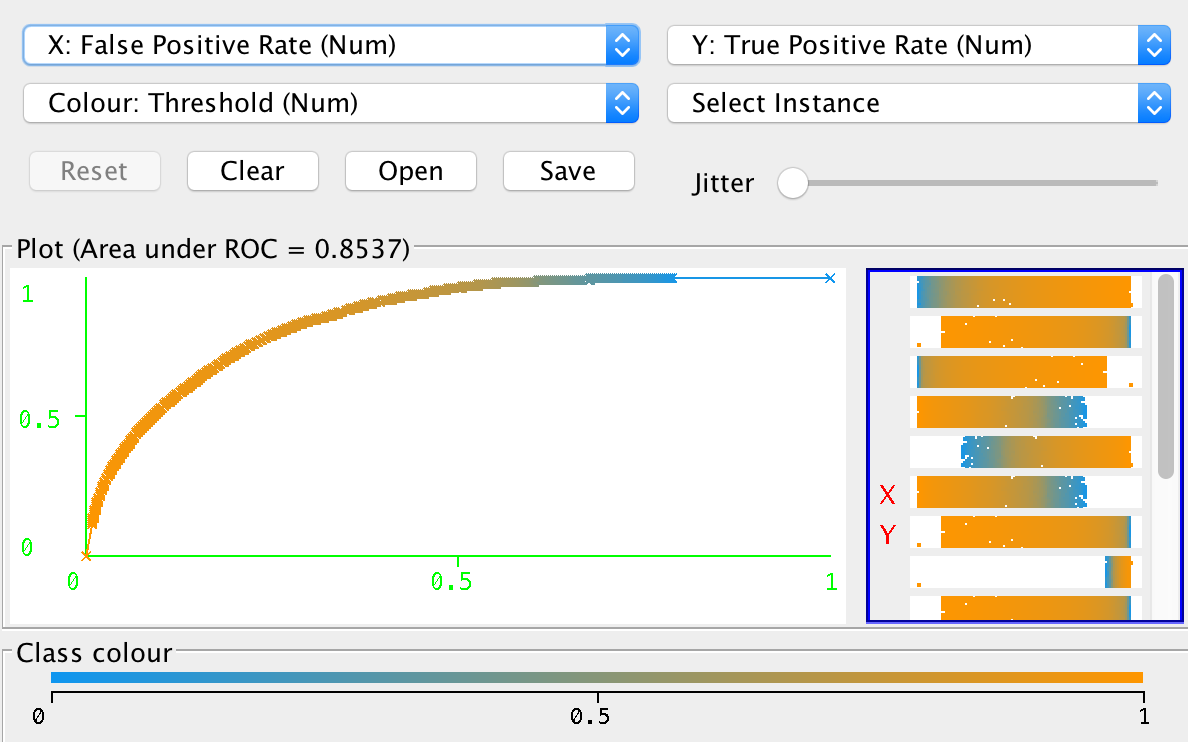
We used two models total. At first, we ran the J48 by using the dataset which concludes all the data from January in 2016. From the result, we got that the most informative attribute that causes the airline delay is NAS\_Delay. However, the goal of our project is to find the useful information to give 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 folders 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.



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.

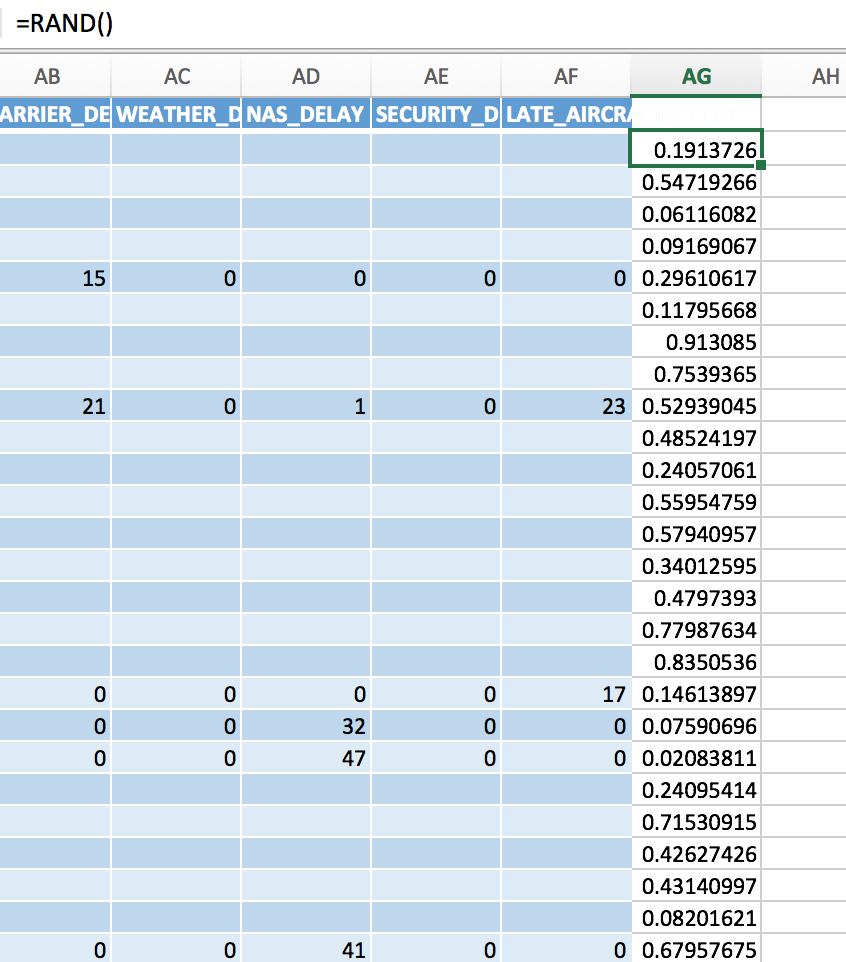
This model got around 90% accuracy. Overall, according to the result which we get from J48, our associate rule model is reliable and accurate. From the evaluation, we got that when analyze a business case, the first thing need to be decided is the goal of the whole project. Then, thinking about using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

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.



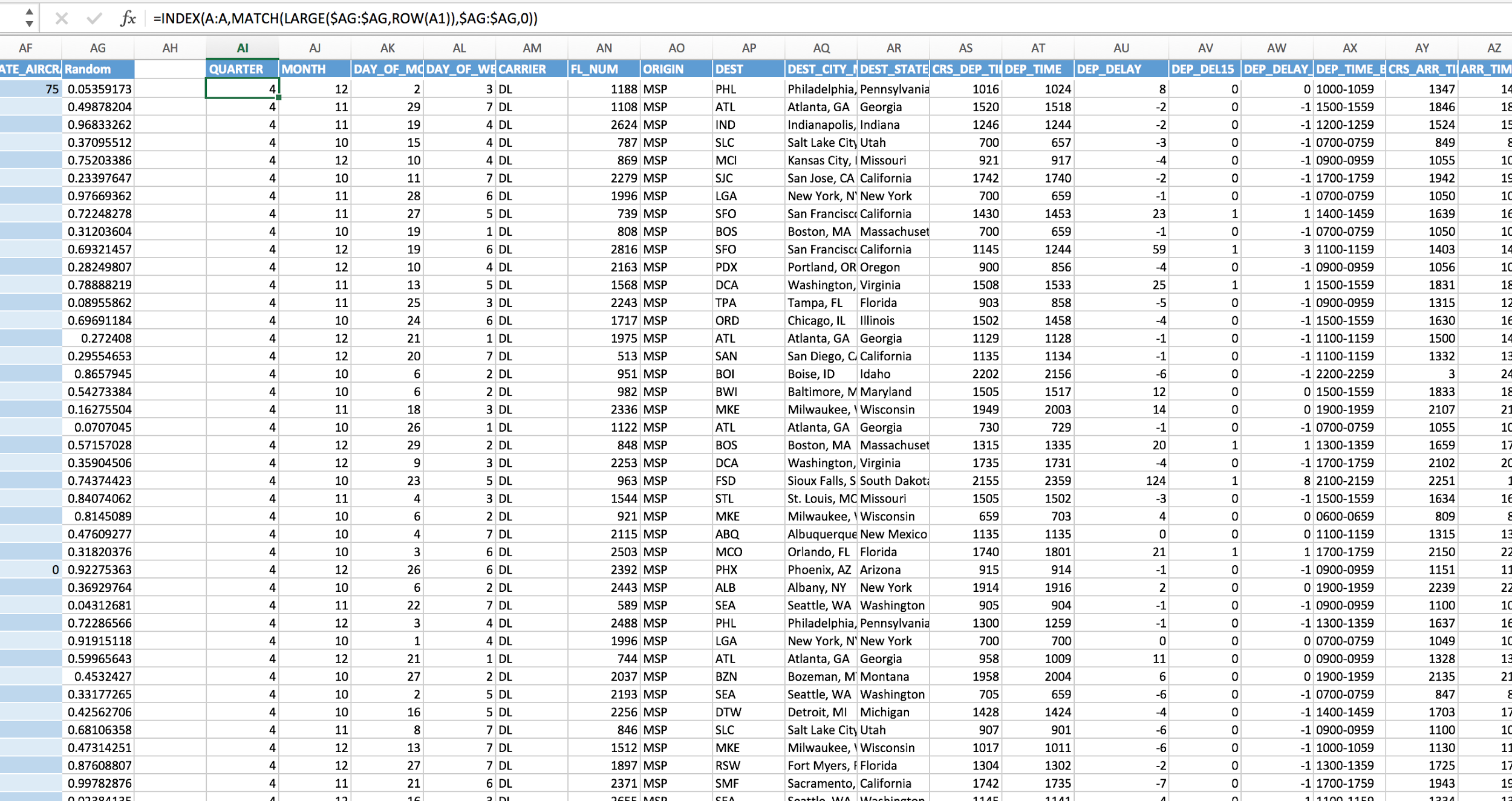
The second step is add the formula in anywhere in the spreadsheet.

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

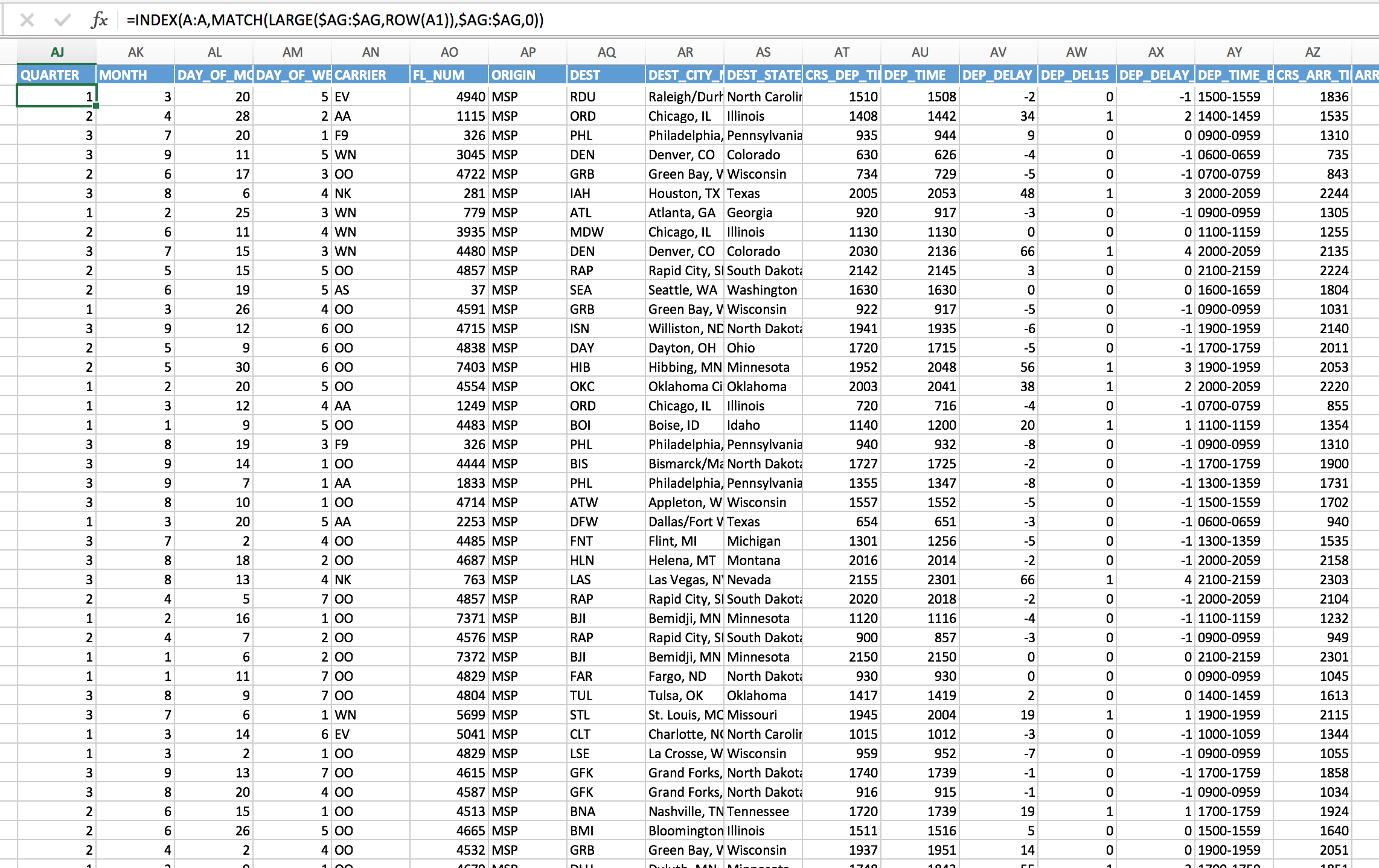


Then, drag the lower right corner of the cell to spread the whole sheet to 1000 rows. That are the random 1000 instances which we choose from the dataset.

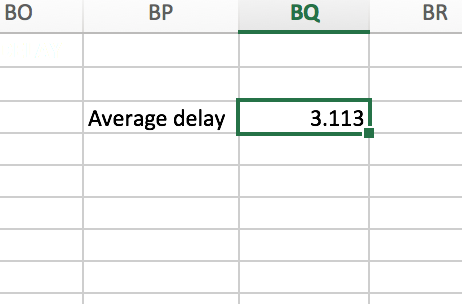
QUARTER=4 and CARRIER=DL



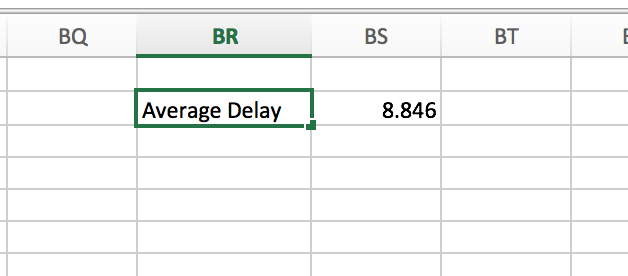
QUARTER does not =4 and CARRIER does not = DL



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

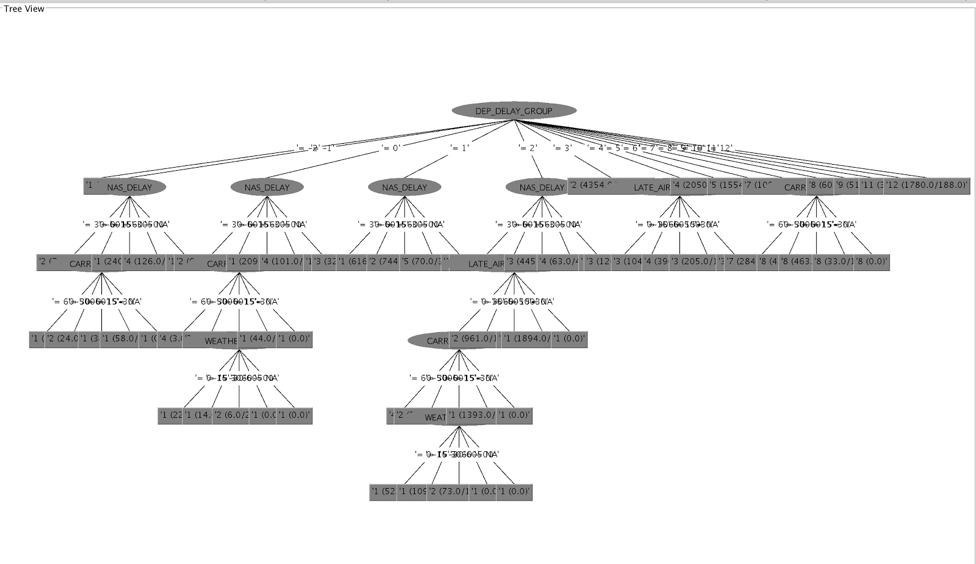


Conclusion

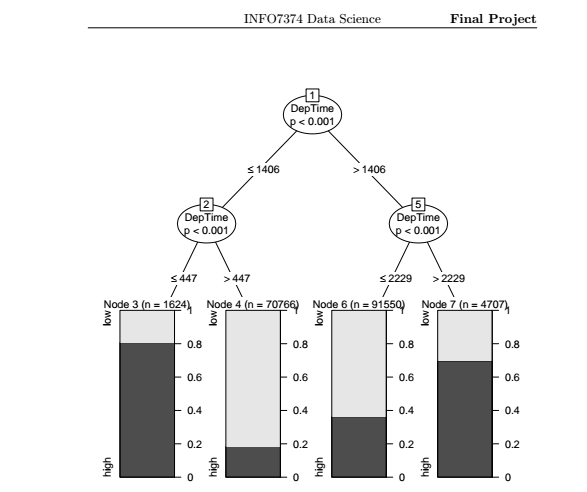
This simulation prove that the result which we get from the association rule is right. It means that if people choose flight in quarter 4 and the carrier is DL, they can save more time and have lower probability to meet the departure delay situation.

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.



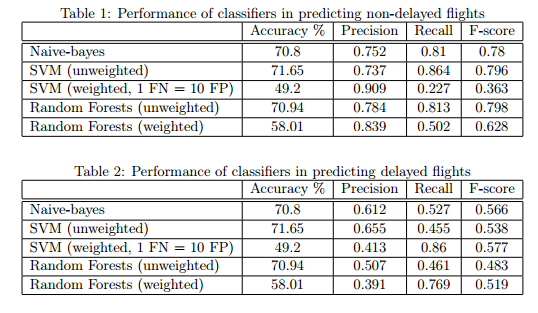
Here is our decision tree done in weka which differs from what was done in the other studies. We did this do discover what type of delay is most common while other studies went other routes to find the time of day with most delays.



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.

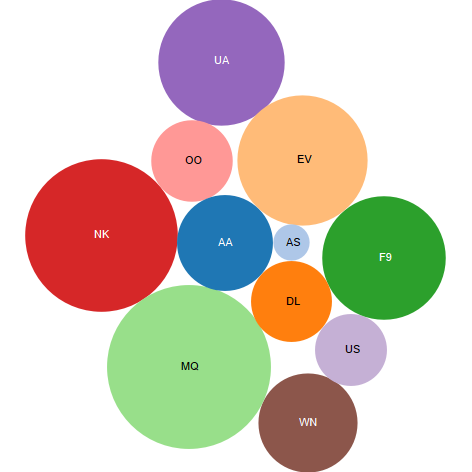


By using a Naive-bayes algorithm this study was able to determine that they were better at predicting non-delayed flights than delayed ones. Our model differs because we did not use a Naive-bayes algorithm to predict whether or not a customer will face a delay.

Deployment(Nik)

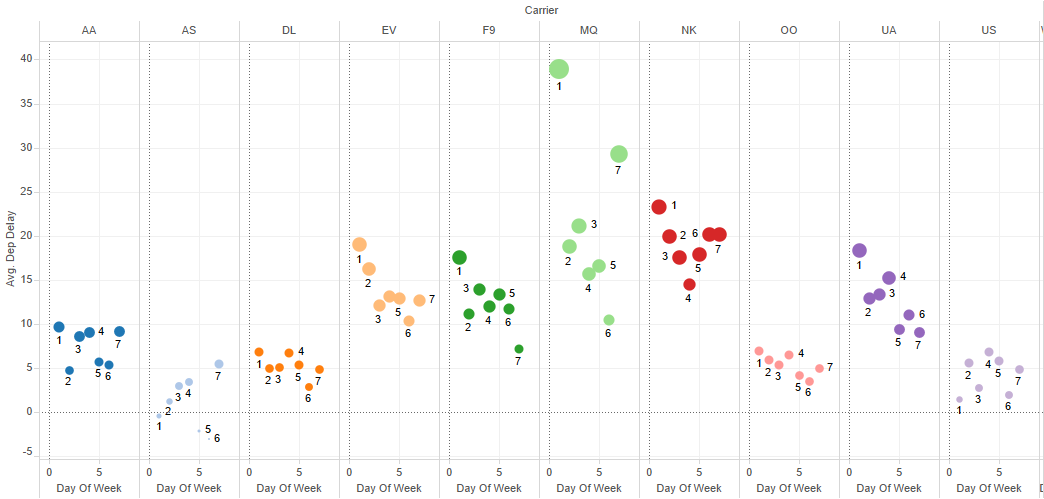
Discuss how the results will be deployed - 5

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 vizualization 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.

Identify risks and how you would mitigate them (at least 5) - 5

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.

Putting false information