Predictive Analysis of Flight Delays

Student Names: Rutvik Bhopatrao

Background and Introduction:

We often wonder, “Did we make the right decision while choosing this airline for our journey?”. We are tired of the airline’s delays and we must compromise this for lesser flight cost. Can this be avoided if we are able to find a pattern in the delays. We will try to answer this question in our case study. We will perform a predictive analysis model for the data and deduce patterns which will help us in further implementing it for finding which airlines, which destinations or which times are the best to avoid delays in our journeys.

The aim of the case study is to use data mining techniques to predict delays of flights depending on various factors like:

* Weather conditions
* Carrier Delays
* NAS Delays
* Security delays
* Late aircraft delays

Source: <https://data.world/data-society/airlines-delay/workspace/file?filename=airlinedelaycauses%2FDelayedFlights.csv>

Solution Design:

The solution design involves the following steps:

1)Data Cleaning and Pre-processing: Explore the Data set, identify missing, redundant and irrelevant observations and modify them

2)Dimension Reduction: Explore correlations among the variables and find variables of interest which impact the predictor variable (here in our delays in aircarfts delays)

3)Data Partitioning: Partition the data into test, validation and training data to build a model and support it for analysis.

4) Implementation of Data Mining Algorithms on Training Dataset: Apply Various Data Mining algorithms to build a mathematical model based on training data

5) Performance evaluation:Test the performance of various data mining algorithms based on validation data set and choose the algorithm which gives value of predicted variable close to the value of predictor variable in validation data set.

Data Collection:

The dataset was obtained from data.world website and it consists of 1936758 records of 30 variables of flight journeys for the whole year 2008 in USA.

The dataset contains the following variables:

|  |  |  |
| --- | --- | --- |
|  | **Variable Name** | **Description** |
| 1 | X | Serial No. |
| 2 | Year | 2008 |
| 3 | Month | 1-12 |
| 4 | DayofMonth | 1-31 |
| 5 | DayOfWeek | 1 (Monday) -7 (Sunday) |
| 6 | DepTime | actual departure time (local, hhmm) |
| 7 | CRSDepTime | scheduled departure time (local, hhmm) |
| 8 | ArrTime | actual arrival time (local, hhmm) |
| 9 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 10 | UniqueCarrier | unique carrier code |
| 11 | FlightNum | flight number |
| 12 | TailNum | plane tail number |
| 13 | ActualElapsedTime | in minutes |
| 14 | CRSElapsedTime | in minutes |
| 15 | AirTime | in minutes |
| 16 | ArrDelay | arrival delay, in minutes |
| 17 | DepDelay | departure delay, in minutes |
| 18 | Origin | origin IATA airport code |
| 19 | Dest | destination IATA airport code |
| 20 | Distance | in miles |
| 21 | TaxiIn | taxi in time, in minutes |
| 22 | TaxiOut | taxi out time in minutes |
| 23 | Cancelled | was the flight cancelled? |
| 24 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 25 | Diverted | 1 = yes, 0 = no |
| 26 | CarrierDelay | in minutes |
| 27 | WeatherDelay | in minutes |
| 28 | NASDelay | in minutes |
| 29 | SecurityDelay | in minutes |
| 30 | LateAircraftDelay | in minutes |

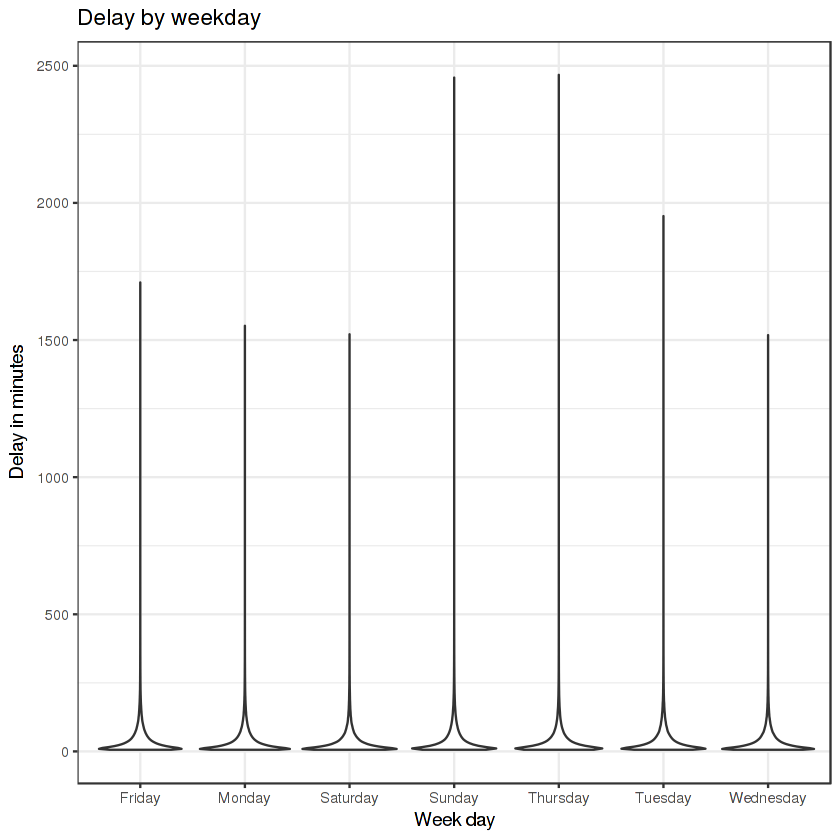
Data Exploration:

The data exploration of the delayed flights dataset yielded the following results:



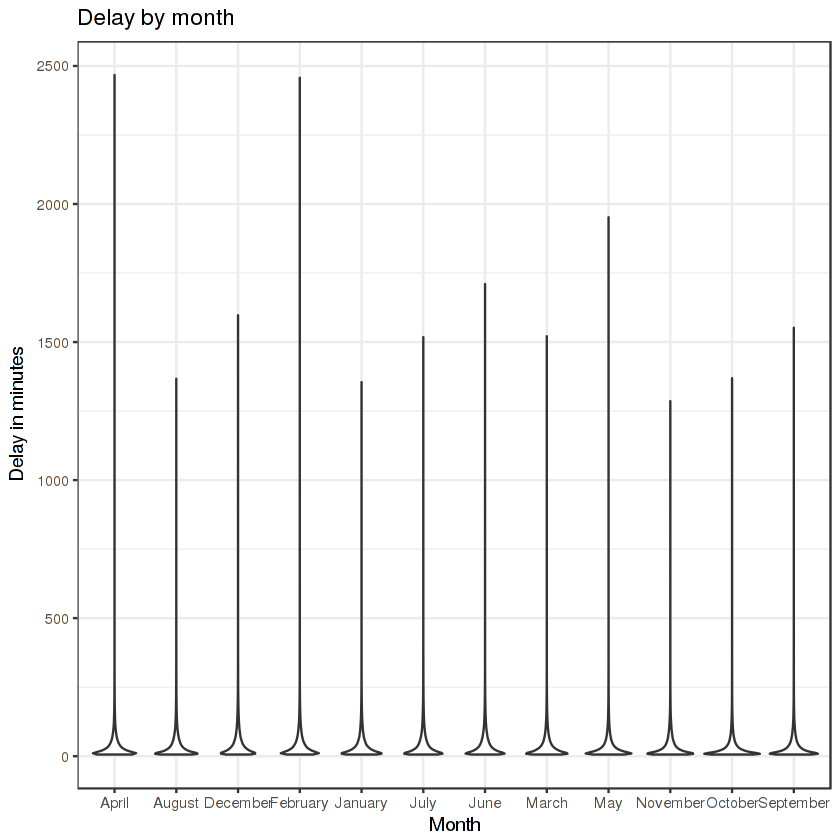


Delay by DayofWeek:



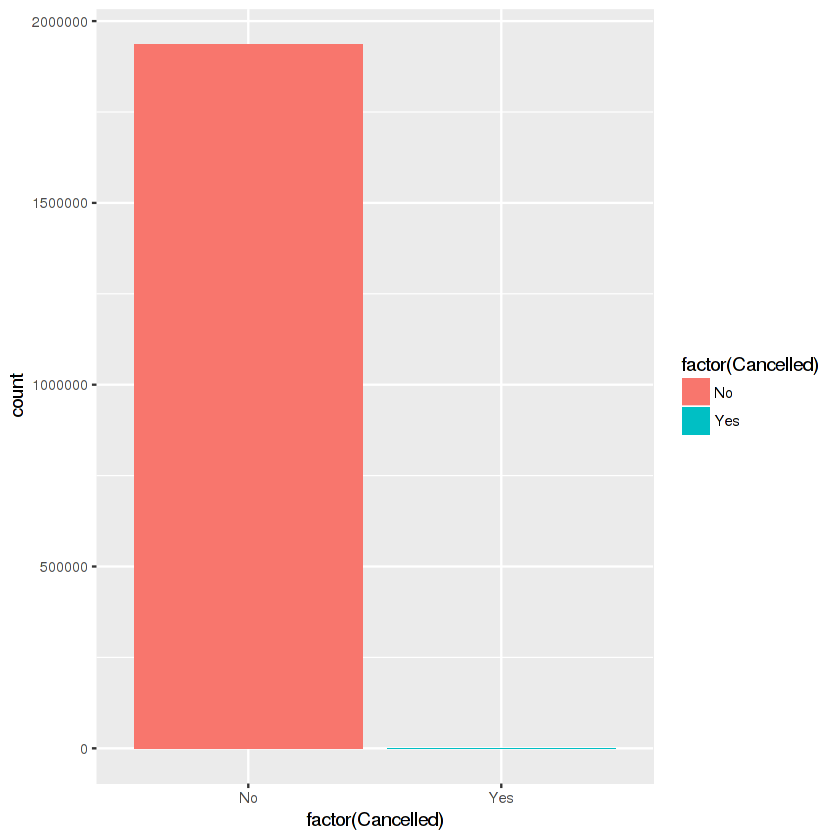
The most common day for flight delays is on friday, which isn't surprising as it's also one of the most common days to fly.

Delay by Month:

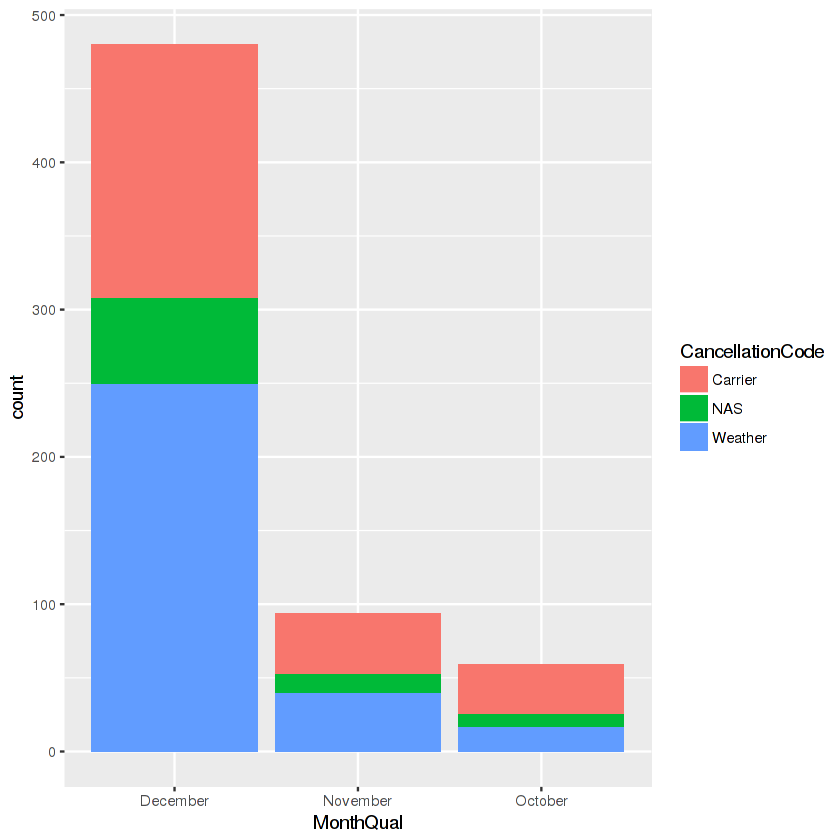


The most common months to fly are December and April which is the holiday season, which arises more delays.

Total Cancelled flights:



Major reason of cancellation:



Weather is the most common reason for a flight cancellation, and there are no instances of a security related cancellation in the dataset. The majority of cancellations are in November and December. Considering the timing of the cancellations, it appears that some cancellations related to the carrier could also be weather related.

Summary of Delays:



From the summary statistics in the first few cells, it appears that around 68,000 observations have NA values for the cause of the delay. Judging by that and the fact that the smallest departure delay in the entire dataset was 6 minutes, each observation that's 0 must be due to the delay being for a reason outside the specific category. Because of that, to get a better idea of how long delays in each category last, it is best to filter out each 0 instance.

Weather Delay by Month:



Although there's substantially more weather delays in December, the average delay was much longer in June. Five of the six months that have the longest average delays do not frequently see snow in the US, which could potentially mean that delays caused by snowy conditions don't last as long. The amount of cancellations in November and December also could suggest that when weather conditions become too poor in the winter, the flights are cancelled instead of waiting for weather conditions to improve, which could decrease the average delay time in those months.

Data Cleaning and Visualization:

New Variable: We add a 31st variable using mutate for last five delay variables, where whenever there is any type of delay(weather, carrier, etc), it will show a 1 or “TRUE” in the variable column.

Missing Values: All the missing values in the delay columns are replaced by ‘0’, indictaing the cause is not valid. Also, if Arrtime/Deptime is missing, it has been replace with its CRS i.e. ScheduleTime.

Cleaning/ Removing unwanted variables:

First we can eliminate columns Carrierdelay, Weatherdelay, NASdelay, SecurityDelay, LateAircraft as we have added a new variable ‘delay’ in our dataset which records a ‘1’ for every record in all the five columns.

We remove X(1st variable) as it is just the serial number of the record and will not really help us in our analysis.

Year is removed as the dataset involves only one value of ‘2008’ for the whole dataset.

Day of month is not necessarily important and hence eliminated as there is no significant help provided from the day of travelling of any particular month.

DepartureTime, CRSDepartureTime, ArrivalTime, CRSArrivalTime, TaxiIn, TaxiOut are removed as have actual elapsed time and CRS elapsed time, which is actually the sum of all those times.

FlightNumber, TailNumber are unnecessary as our purpose is to find delays during particular days, months, and destinations. Flight numbers would not really help us.

AirTime is unnecessary and is just and additional variable which wont help us in analysis.

Cancellation code is being eliminated now as we are performing predictive analysis on delaye flights.

DepDelay mentions the amount of time a flight is delayed at any airport. But, in many cases it can be seen that even though there is a delay on the departure airport, the flight arrives on time at its destination. This gives rise to different analysis which we aren’t concerned in our study. So we decide to not include DepDelay.

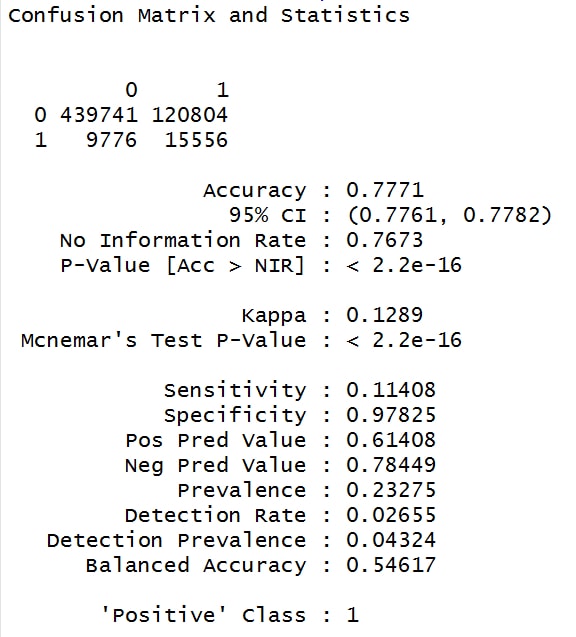
Data Mining Techniques and Implementation:

Model 1: Regression Tree:

Training Dataset: The training dataset for the case study is a 1162054 i.e. 60% of 1936758 consisting of 15 variables.

Validation Dataset: The training dataset for the case study is a 585877 i.e. 40% of 1936758 consisting of 15 variables.

The model was first trained with the training dataset and then it was used to predict the values of the validation dataset. The confusion matrix after running the model shows values as:



This confusion matrix indicates that the model has a 77.71% accuracy. This indicates that the model has quite high accuracy and can be used for further analysis.

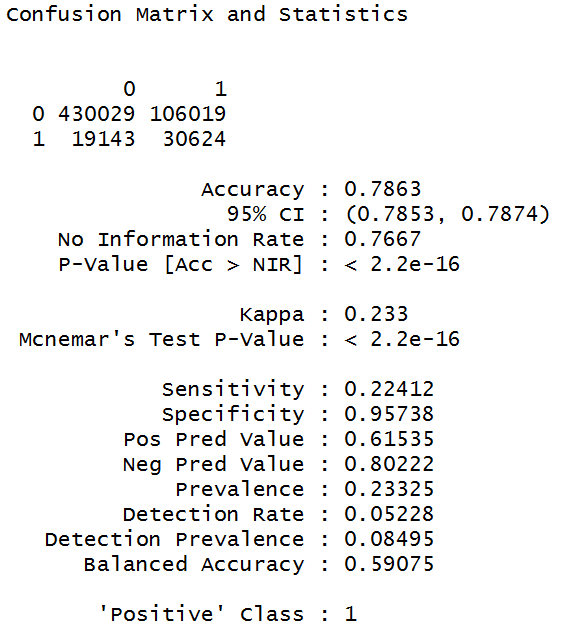
Correlation between predicted values of the validation dataset and the actual values of the validation dataset is 0.9114853. Whereas, correlation between predicted values of the new dataset and the actual values of the new dataset is 0.5445108 which is quite low. One reason for the lower correlation coefficient could be the size of the dataset. Since we are trying to predict the values only, the results may not be as desired.

Model 2: Logistic Regression:

Training Dataset: The training dataset for the case study is a 1162054 i.e. 60% of 1936758 consisting of 15 variables.

Validation Dataset: The training dataset for the case study is a 774704 i.e. 40% of 1936758 consisting of 15 variables.

The model was first trained with the training dataset and then it was used to predict the values of the validation dataset. The confusion matrix after running the model shows values as:



This confusion matrix indicates that the model has a 78.63% accuracy. This indicates that the model has quite high accuracy and can be definitely used for further analysis. It can prove to be the best model available.

Model 3: Random Forest:

Training Dataset: The training dataset for the case study is a 1162054 i.e. 60% of 1936758 consisting of 15 variables.

Validation Dataset: The training dataset for the case study is a 774704 i.e. 40% of 1936758 consisting of 15 variables.

The model was first trained with the new training dataset and then it was used to predict the values of the validation dataset. The confusion matrix after running the model shows values as:



This confusion matrix indicates that the model has a 64.38% accuracy. This indicates that the model has quite high accuracy and can be used for further analysis.

Performance Evaluation:

All the three models predict well when they are fit using the 12 most important variables-**Month, DayofWeek, UniqueCarrier, Actualelapsedtime, CRSelapsedtime, ArrivalDelay, Origin, Destination, Cancelled, Diverted and delay.** Hence, only these variables are considered for our model selection and prediction.

Of all the 3 models mentioned above, i.e. Regression tree, Logistic Regresion and Random Forest. The accuracy for Logistic Regression is highest among the three models. It has an accuracy of 78.63%. indicating that the second model fits the data better than others. Hence, selecting model 2 (Logistic Regression) seems to be the best choice for prediction and any other future analysis.

Discussion and Recommendation:

From all the analysis done, we have found some results such as, there have been lots of flight delays for some or the other reason. The main reason being weather delays but some also being carrier and late departure delays caused by airline or due to fault of the airport. We found that almost 52% of the fights were cancelled because of the weather delays, whereas security related cancellations are very low. Also, nearly 65% of the flights have some or the other delay and almost 1 out of every three aircrafts has arrival delays.

Most common day of the week is Friday for delays as it is the most preferred day for travel. More number of flights and passengers give rise to more delays which need to further improved. Further, winter months of December to February face large number of delays which are mostly weather related. Now since, weather changes are unpredicted, we need to deploy serious measures to reduce any other types of delays that occur, so the rate of delays is reduced.

Although there's substantially more weather delays in December, the average delay was much longer in June. Five of the six months that have the longest average delays do not frequently see snow in the US, which could potentially mean that delays caused by snowy conditions don't last as long. The amount of cancellations in November and December also could suggest that when weather conditions become too poor in the winter, the flights are cancelled instead of waiting for weather conditions to improve, which could decrease the average delay time in those months.

Summary:

The case study has helped us in understanding the different methods used for predictive analysis. It also helps us in understanding their performance in general and for this dataset most importantly. The regressive trees or logarithmic regression or the random forests helped us identify the most important predictor variables for our study. The random forests also demonstrated a better predictive accuracy as compared to other models used in our study. Through this study, we also learnt that multicollinearity in the dataset degrades the predictive accuracy of the random forest and hence should be avoided. Regression models are best suited of this type of dataset. Thus, the model we build to predict the delays in aircrafts works well and can be used to further to predict for any given year in the future.

Appendix: R code for study:

Exploratory analysis

> library(readr)

> airdelay <- read\_csv("C:/Users/Admin/Desktop/Data Mining/Datasets/airlinedelaycauses\_DelayedFlights.csv")

> summary(airdelay)

> str(airdelay)

> airdelay$WeekDay <- airdelay$DayOfWeek

> airdelay$MonthQual <- airdelay$Month

> airdelay$WeekDay[ airdelay$WeekDay == 1] = 'Monday'

> airdelay$WeekDay[ airdelay$WeekDay == 2] = 'Tuesday'

> airdelay$WeekDay[ airdelay$WeekDay == 3] = 'Wednesday'

> airdelay$WeekDay[ airdelay$WeekDay == 4] = 'Thursday'

> airdelay$WeekDay[ airdelay$WeekDay == 5] = 'Friday'

> airdelay$WeekDay[ airdelay$WeekDay == 6] = 'Saturday'

> airdelay$WeekDay[ airdelay$WeekDay == 7] = 'Sunday'

> airdelay$MonthQual[ airdelay$MonthQual == 1] = 'January'

> airdelay$MonthQual[ airdelay$MonthQual == 2] = 'February'

> airdelay$MonthQual[ airdelay$MonthQual == 3] = 'March'

> airdelay$MonthQual[ airdelay$MonthQual == 4] = 'April'

> airdelay$MonthQual[ airdelay$MonthQual == 5] = 'May'

> airdelay$MonthQual[ airdelay$MonthQual == 6] = 'June'

> airdelay$MonthQual[ airdelay$MonthQual == 7] = 'July'

> airdelay$MonthQual[ airdelay$MonthQual == 8] = 'August'

> airdelay$MonthQual[ airdelay$MonthQual == 9] = 'September'

> airdelay$MonthQual[ airdelay$MonthQual == 10] = 'October'

> airdelay$MonthQual[ airdelay$MonthQual == 11] = 'November'

> airdelay$MonthQual[ airdelay$MonthQual == 12] = 'December'

> airdelay%>%group\_by(WeekDay)%>%tally%>%arrange(desc(n))%>%as.data.frame()

> ggplot(data, aes(x=WeekDay, y=DepDelay, fill=DepDelay))+

+ geom\_violin()+

+ theme\_bw()+

+ theme(legend.position="none")+

+ ggtitle("Delay by weekday")+

+ xlab("Week day")+

+ ylab("Delay in minutes")

> airdelay%>%group\_by(MonthQual)%>%tally%>%arrange(desc(n))%>%as.data.frame()

> ggplot(data, aes(x=MonthQual, y=DepDelay, fill=DepDelay))+

+ geom\_violin()+

+ theme\_bw()+

+ theme(legend.position="none")+

+ ggtitle("Delay by month")+

+ xlab("Month")+

+ ylab("Delay in minutes")

> airdelay$Cancelled[airdelay$Cancelled==0]='No'

> airdelay$Cancelled[airdelay$Cancelled==1]='Yes'

> qplot(factor(airdelay$Cancelled), data=airdelay, geom="bar", fill= factor(airdelay$Cancelled))

> airdelay %>% group\_by(Cancelled) %>%tally %>% arrange(desc(n)) %>% as.data.frame()

> airdelay$CancellationCode[airdelay$CancellationCode == 'A'] = 'Carrier'

> airdelay$CancellationCode[airdelay$CancellationCode == 'B'] = 'Weather'

> airdelay$CancellationCode[airdelay$CancellationCode == 'C'] = 'NAS'

> airdelay$CancellationCode[airdelay$CancellationCode == 'D'] = 'Security'

> airdelay %>% filter(CancellationCode != 'N') %>%

+ group\_by(CancellationCode) %>%

+ tally %>%

+ arrange(desc(n)) %>%

+ as.data.frame()

> CancelledSubset = subset(airdelay, CancellationCode != 'N')

> ggplot(CancelledSubset,aes(MonthQual,fill=CancellationCode)) + geom\_bar()

> airdelay %>% filter(CarrierDelay != 'NA',WeatherDelay != 'NA',

+ NASDelay != 'NA', SecurityDelay != 'NA', LateAircraftDelay != 'NA') %>%

+ select(CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay) %>%

+ summarize(CarrierDelay = mean(CarrierDelay),

+ WeatherDelay = mean(WeatherDelay),

+ SecurityDelay = mean(SecurityDelay),

+ LateAircraftDelay = mean(LateAircraftDelay)) %>%

+ gather(var,mean) %>%

+ as.data.frame()

> airdelay %>% filter((CarrierDelay!= 'NA') & CarrierDelay != 0) %>%

+ select(CarrierDelay) %>% summarize(CarrierDelay = mean(CarrierDelay)) %>%

+ gather(var,mean) %>% as.data.frame()

> airdelay %>%

+ filter((CarrierDelay != 'NA') & CarrierDelay!=0) %>%

+ group\_by(MonthQual) %>%

+ summarize(CarrierDelay = mean(CarrierDelay)) %>%

+ arrange(desc(CarrierDelay)) %>% as.data.frame()

> airdelay %>% filter((WeatherDelay!= 'NA') & WeatherDelay != 0) %>%

+ select(WeatherDelay) %>% summarize(WeatherDelay = mean(WeatherDelay)) %>%

+ gather(var,mean) %>% as.data.frame()

> airdelay %>% filter((WeatherDelay != 'NA') & WeatherDelay!=0) %>%

+ group\_by(MonthQual) %>%

+ summarize(WeatherDelay = mean(WeatherDelay)) %>%

+ arrange(desc(WeatherDelay)) %>% as.data.frame()

> airdelay %>% filter((WeatherDelay!= 'NA') & WeatherDelay != 0) %>%

+ group\_by(MonthQual) %>% tally %>% arrange(desc(n)) %>% as.data.frame()