Data Analysis of Airline On-Time Performance

Akshay Mandke
College of Information Studies
University of Maryland, College Park
amandke@umd.edu

Krishnesh Pujari
College of Information Studies
University of Maryland, College Park
kpujari@umd.edu

Mohit Juneja
College of Information Studies
University of Maryland, College Park
mjuneja@umd.edu

I. BACKGROUND AND PROBLEM DESCRIPTION

In the airline industry, it is very common that the airlines are finding it difficult to get plane to the gate on time. The major challenge the industry has been facing, is to improve the quality of the airline on-time performance. Thus, our goal with this research study is to determine which factor most impacts the airline on-time performance, which airline industry the customer should choose and lastly, the report suggests the time and day of the week at which the customer must travel to avoid longer delays. The analysis is important because it will increase the Nation's understanding of airline transportation statistics and enable the people to better informed decisions about their travels. Additionally, this analysis can be used by the airline industries to work upon their performance by looking into the root causes of the various factors affecting it.

II. DATA PREPARATION

Data Description: For our study, we have used the airline on-time data obtained from the Research and Innovative Technology Administration (RITA). The dataset contains huge information about the airline statistics across the United States, including 38 variables like Airline ID, Flight Number, Flight Date, Arrival time, and Departure time and so on. The scope of the data under our study is from January 2014 through December 2014. The selected time range has over 5 million records across 38 variables, which are good enough to obtain satisfying outcomes. For our study, the descriptive list of useful variables is as follows:

- DAY_OF_WEEK: Day of the Week, 1 refers to Monday and likewise, 7 refers to Sunday
- AIRLINE_ID: An identification number assigned by US DOT to identify a unique airline (carrier)
- CRS_DEP_TIME: Scheduled Departure Time (hhmm)
- ARR_DELAY: Difference in minutes between the scheduled and the actual arrival time, with early arrivals as negative numbers
- ARR_DELAY_NEW: Difference in minutes between the scheduled and the actual arrival time, with early arrivals set to 0
- CARRIER_DELAY: Carrier Delay, in Minutes
- WEATHER_DELAY: Weather Delay, in Minutes
- NAS_DELAY: National Air System Delay, in Minutes
- SECURITY_DELAY: Security Delay, in Minutes
- LATE_AIRCRAFT_DELAY: Late Aircraft Delay, in Minutes

Data Processing: Since the data obtained from the Bureau of Transport Statistics has monthly statistics, thus we first need to merge the datasets.

```
library(rJava)
library(XLConnect)
library(XLConnectJars)
#read the names of files in dataframe 'dataFiles'
dataFiles <- list.files()
for(i in 1:length(dataFiles))
{
  #create merged dataset, if it doesn't exists
  if (!exists("dataset"))
  {
    dataset<-read.csv(dataFiles[i], header = TRUE)
  }
  #if the merged dataset exists, append to it
  if (exists("dataset"))
  {
    temp_dataset <-read.csv(dataFiles[i], header=TRUE)
    dataset<-rbind(dataset, temp_dataset)
    rm(temp_dataset)
  }
}
#write dataframe 'dataset' in human readable csv format
write.csv(x=dataset,file="OnTime_Airline.csv",row.names=FALSE)
```

Next, our analysis needs a well-structured dataset, so we filter out the NA values. Additionally, due to the limited capability of the computer processor to handle large datasets, we decided to create random samples of 500,000 observations across 38 variables. This has been done using R and we believe that the sample data is enough to obtain good data analysis results.

```
#generate random samples
sampleData <- dataset[sample(1:nrow(dataset), 500000, replace=FALSE),]
```

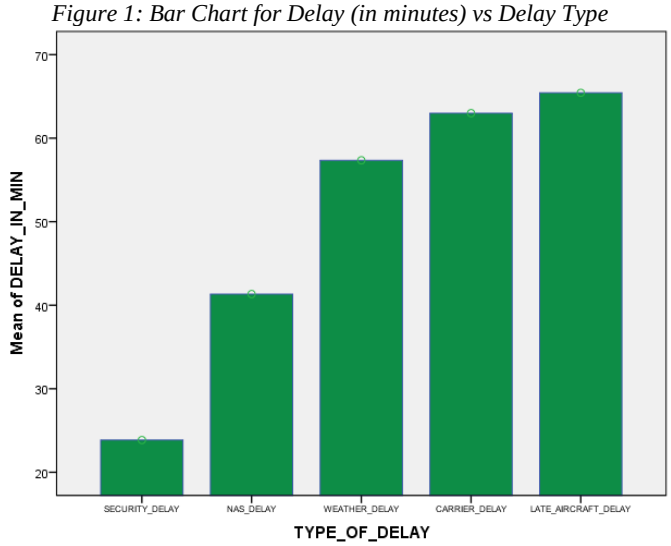
Lastly, SPSS has been used to analyze the data samples, after applying the various descriptive statistics techniques as well as statistical tests, to reach the conclusions.

Reliability and Validity: We performed initial checks on the dataset in SPSS, to analyze the reliability and validity of data. From our analysis we found out that the data collected is valid across required variables with respect to the coded values. The population from which the samples are obtained, is positively skewed and violates the normality assumption of the ANOVA test. However, this violation does not have any profound effect on the results since the sample size is very large. Furthermore, the data comes from populations that have equal amount of variability and the observations are independent i.e. the results obtained from one sample won't affect the others.

III. ANALYSIS AND FINDINGS

We carried out the analysis of the dataset using various statistical analysis techniques of One-Way ANOVA, Two-Way ANOVA, Power calculations. Using these tests we have identified certain dependencies in the dataset and factors impacting the airline on-time performance. The analysis involved finding out the factor that has most impact on the airline's on-time performance. We performed One-Way ANOVA since the study design involved one dependent variable (DELAY_IN_MIN), which is the delay encountered by the airline, in minutes and one independent variable (TYPE_OF_DELAY) which has five categories of delay i.e. carrier, weather, NAS, security, and delay due to late aircraft arrival & departure. On performing One-way ANOVA test on the above two variables, we observed that the significance value for the test is less than significance level, $\alpha = 0.05$ for our test. We further performed Tukey's Post Hoc test to conclude that there is a significance difference between the delay values for each group of delay. Using t-tests to compare 5 groups of delays would need performing ten t-tests and thus, it was avoided.

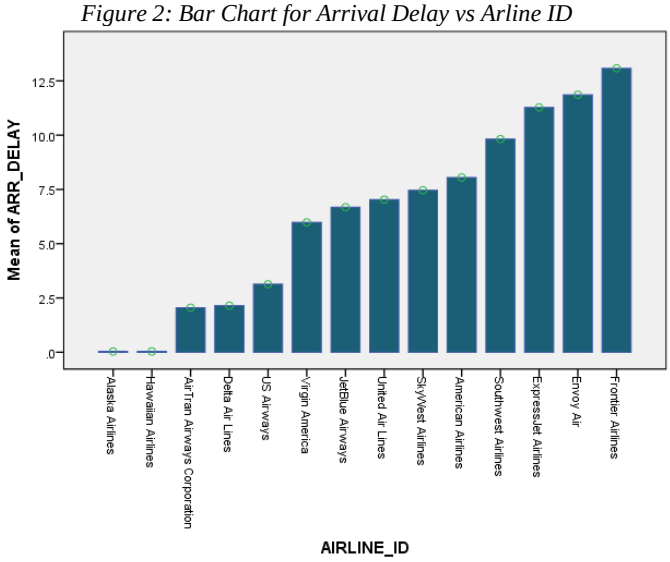
Figure 1: Bar Chart for Delay (in minutes) vs Delay Type



As it is evident from the above bar chart, plotted for mean of delays in minutes across the different delay types, there is significant difference in mean value of each delay. The delay that most impacts the airplanes on-time performance is the Late Aircraft Delay type.

Our next analysis involved identifying the airline that has experienced maximum delays over the year 2014. We used One-Way ANOVA since we identified two variables from the available dataset i.e. (ARR_DELAY) which is the dependent variable and (AIRLINE_ID) the independent variable. The AIRLINE_ID contains 14 categories for which we compared the mean delay values between each categories. The significance value calculated for this test was less than the significance level assumed as $\alpha = 0.05$. Tukey's Post Hoc test concluded that there is a significance difference between the delay values for each group of delay. Use of t-test was avoided, as to compare 14 groups of Airline Company we would have to perform large number of t-tests. We also calculated the power of our test which came out to be 100% i.e. the test performed on the available data is powerful.

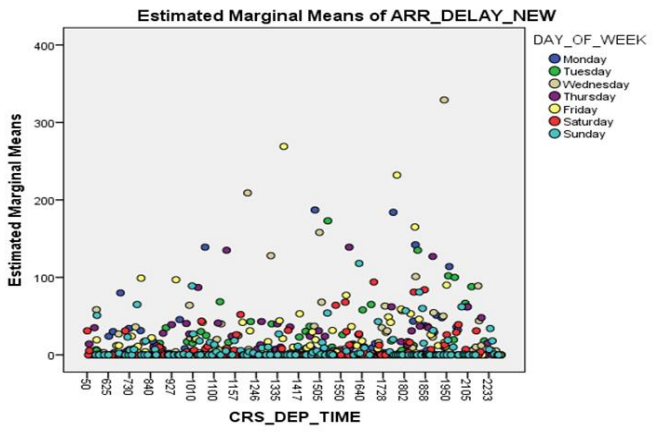
Figure 2: Bar Chart for Arrival Delay vs Arline ID



From the above bar chart, plotted for mean of arrival delay across the 14 different categories of airline companies it is evident that there is a significant difference in the delay value of each airline company. Frontier Airlines has experienced maximum delay and Alaska Airlines had minimal delays in 2014.

Our third analysis involved determining the impact of departure time of the airline and the day of the week on the on-time performance of the aircraft. We conducted Two-Way ANOVA test on the departure time, day of the week which are independent (categorical) variables and arrival delay which is a (continuous) dependent variable. After analyzing the results, we conclude that the significance value at the intercept of the two independent variable is less than the significance level assumed to be $\alpha = 0.05$. Using the statistical test we identified the exact day of the week and time within a day where there is maximum delay experienced by the airline companies.

Figure 3: Scatterplot for Arrival Delay vs Departure time, plotted across Day of the Week



The scatterplot shows the delay value of the aircrafts across the 24 hours' time scale for each day of the week. From the above scatterplot, it is evident that maximum delay experienced by the airline companies is for the flights which are scheduled on Friday evening, and minimum delay is experienced by the flights which depart on Sunday.

IV. CONCLUSION

We analyzed what factors most affect the airline on time performance. The major factor which impacts the airlines on-time performance is the Late Aircraft delay. The delay values identified from our dataset are significantly different across different airline carriers. Then we deduced which airlines has the most amount of delays and our results lead us to conclude that Frontiers Airline Company has experienced maximum delays in the past year while Alaska and Hawaiian Airlines has the highest on-time arrival rate. So, this analysis will help the customers in making decision while choosing the airline for their travel. It will also provide statistics to the airline company to work on their problems and provide better service. Finally, we used our results to answer which day and time of the week to choose in order to avoid delays. Our analysis concluded that Friday evening is the time period when maximum delays occur for departure of flights while there are less delays encountered on Sunday. This analysis can help customers to take decisions on flight timings while making reservations. It will also assist the Airport authorities to take decisions on handling the airplane schedules to improve airline on-time performance.

V. LIMITATIONS

There are a few limitations associated with the analysis conducted for the research study. Firstly, the airline statistics of only 2014 is used for the research study. This dataset is huge, and contains 5,819,811 observations, so we selected a part of it, owing to the limited capability of the computer to process large scale data, and loaded the dataset into SPSS. The random samples were generated using the random function in R. Another limitation is that, the study only focuses on the delays experienced by the airlines. However, there may be other interesting relationship among other variables.

VI. REFERENCES

1. Bureau of Transportation Statistics (n.d.). Retrieved from February 17, 2015. http://www.transtats.bts.gov/Fields.asp?Table_ID=236
2. GraphPad Statistics Guide. (n.d.). Retrieved April 23, 2015, from http://www.graphpad.com/guides/prism/6/statistics/index.htm?f_ratio_and_anova_table_(one-way_anova).htm
3. StatCrunch (n.d.). Retrieved May 02, 2015, from https://www.statcrunch.com/5.0/example.php?example_id=71
4. Field, A. (2012). Discovering Statistics Using IBM SPSS Statistics. And Sex and Drugs and Rock'n'Roll. Pflege, 430-430. Retrieved May 05, 2015, from http://www.sagepub.com/upm-data/52063_00_Field_4e_SPSS_Prelims.pdf