Exploratory Data Analysis & Airline On-Time Performance

By: Amit Kamat

University at Buffalo

1. Abstract

Exploratory Data Analysis & Airline On-Time Performance gives one a fantastic exposure to a vast domain associated with airlines, flight data, and various related statistics. This specific data set gives us a great overview of delay related indicators within the United States in the year of 2013.

The design of this project will follow a linear progression. First I will familiarize myself with the data set by cleaning and compressing it slightly to a bit more of a manageable size and format. I will then categorize the data as appropriate so that I can explore it using different methods and visualizations. After this I will perform EDA on a single month. This will consist of various summaries and visualizations such as plots and distributions. I will further extend my analysis to the entire year. To do this I will compress the data by collecting and compiling important key values for each month. Once I have the data I need for the entire year, I will explore it through more summaries and visualizations. I will follow up all of this with a nice analysis using K-means and some related informative scatter plots on a map of the United States.

I will play a little more with maps and provide some other cool and interesting visualizations that express the data in a more easy to understand fashion. I will also try and make all summaries and graphs as informative as possible. Last but not least, I will make my code easily accessible by providing a method to run it with as few commands as possible.

2. Project Objectives

The Exploratory Data Analysis & Airline On-Time Performance project, at the very least, will meet the following list of objectives:

- Learn and explore statistical modeling.
- Learn R Language for data analysis.
- Learn how to utilize tools such as R Studio.
- Learn how to structure data into a more useful format.
- Learn how to read and analyze results from a data set.
- Become familiar with a few different domains.
- Get some practice using K-means.
- Get some practice using maps and coordinate systems.
- Get a good understanding of flight delays within the United States.
- Provide useful metrics and visualizations.
- Provide code and work in an easily reproducible fashion.
- Provide detailed comments and descriptions within the code.
- Present results in an easy to understand manner.

3. Project Approach

First and foremost I will read (and re-read) Chapter 2 of *Doing Data Science* until I feel like I understand the material well enough to delve into the project and some actual Exploratory Data Analysis. Once I am ready to proceed I will first finish the NY Times example in Chapter 2 of the book. I will be using R Studio for this example and the rest of the project. From the first example I hope to learn how to explore data from a single file and extend this exploration to multiple files, a skill I will need later on. I also hope to delve into different functionalities of R such as plots, distributions, summaries, and other complicated functions.

Once I have a completed the questions in the New York Times example, I will proceed with RealDirect, where I will apply and reaffirm my new knowledge of statistical analysis using R. I hope to learn some more through the second example and to fully understand the domains I have encountered thus far. It is essential that I pay close attention to the distributions, summaries, and methodologies I am utilizing. I will do my best to streamline and comment the code so that it is easily accessible to a user. I will also play around with R and its functionalities to ensure that I have a good understanding of a large set of practices that will indubitably prove helpful when I start working with my own data set.

Once I begin work on the Airline On-Time Performance data for 2013 I hope to have a good enough understanding of methodologies so that I can efficiently and properly perform Exploratory Data Analysis. This should hopefully be the case after the New York Times and RealDirect examples. Since the data set is rather large and the source site informs us of missing values, I will clean up those missing values. The tables also seem to contain a bunch of redundant and unnecessary columns that I will get rid of. Since our analysis will focus on delays we have to measure those delays by different categories, this is why we will categorize the data set as appropriate. I will then use summaries and visualizations, applying only those that seem most relevant, a skill which should have acquired through earlier parts of the assignment. I will also perform K-means after reading about it in Chapter 3 of the book and fully understanding the concept. I shall express the scatter plots on a map of the United States. Once again I will comment my code well and make it easily accessible.

Last but not least I plan on having some fun while doing this assignment. I shall experiment with the functionalities that R provides in an effort to learn more about its capabilities and fully understanding the material that I absolutely need to know. I also plan on playing a bit more with maps beyond the requirements of the assignment.

4. Chapter 2: New York Time Data Set, Questions, and Outcomes

4.1 - Questions 1, 2, and 3 ask us to perform Exploratory Data Analysis in R.

You can find the complete code for this example in the following file:
"project1/new_york_times/nyt.r".

If you would like to run the code yourself, all you have to do is place the data set .csv files in a directory named "data_set" which must be located in the same directory where "nyt.r" resides. Following this, simply run "source('<path_to_file>/nyt.r')". Doing this is going to set up the environment with everything you will need. You can now run plots, summaries, and any additional parts of the code that you desire.

4.2 - Question 4 asks us to describe and interpret any patterns we find:

- Users, Clicks, and Impressions are generally much higher on Sundays; there is a huge spike in usage every Sunday, near double the average!
- Users, Clicks, and Impressions are noticeably lower on Saturdays.
- Click-Through-Rate (CTR) seems to alternate between around 0.018 and 0.021 every 15 days at least as seen in the data provided.
- CTR over time seems to be the same for females and males.
- Clicks seem to be increasing very slightly over time, but the rate is so low that we may need more data to make sure that this is indeed the case.
- User, User Type, and Impression numbers seem pretty consistent over time.
- On average there are around 5 impressions per user.
- The higher the age-range the more the lick-through-rate probability tends toward certain values.
- When there are clicks, they are rarely more than 1 per user.
- The age groups seems to be normally distributed around 35-44.
- Most of the time there are no clicks, rarely no impressions.
- There are generally more males than females.
- The 65+ group is the only group with more females than males.
- There are significantly more females than males in the 65+ group.
- Around 2/3 of the users are signed in.
- In general there are more than 400000 users per day.

4.3 - Outcomes:

I have gathered a variety of plots, which you can find in:
"project1/new_york_times/plots/"

I have learned plenty about R, R Studio, exploratory data analysis, the domain of the given data set, and interpreting data through summaries, plots, distributions, and other visualizations. I am also now familiar with extending analysis over multiple files, by compressing the data in a useful manner.

5. Chapter 2: RealDirect Questions and Outcomes

5.1 – Question 1 Solution:

I would advice engineers to log as much data as possible. It is certainly true that the more the data, the easier it would be to form a variety of conclusions. To be more specific however, it would be most useful to log data such as sale prices, locations, crime, schools, types of buildings, building age, taxes by type, and perhaps proximity to nearby points of interest.

The way we are going to use the data to monitor product usage could be by analyzing what types of places each user is browsing, and learning what they may or may not like. This is very useful if we want to give users suggestions. We can also log what sort of places are being actually sold so that we can infer the likelihood a listing would have of selling at a certain price range.

We can build this data back into the production website by providing useful statistics, summaries, and visualizations to each user so that they know exactly what sort of realty they are looking into. We have to evaluate what information is useful to the user and what isn't. It is certainly true that if we put too much on the screen a user could get confused rather than informed.

5.2 – Question 2 – This question asks us to perform Exploratory Data Analysis in R.

You can find the complete code for this example in the following file:
"project1/realdirect/realdirect.r"

If you would like to run the code yourself, all you have to do is place the data set .csv files in a directory named "data_set" which must be located in the same directory where "realdirect.r" resides. Following this, simply run "source('<path_to_file>/realdirect.r')". Doing this is going to set up the environment with everything you will need. You can now run plots, summaries, and any additional parts of the code that you desire.

5.3 – Question 3 Solution:

Here are some observation and patterns I have found while looking into the given data set I looked into Manhattan for this exercise:

- Most sales happen in Harlem and the Upper East Side.
- Housing is most expensive in the Upper East Side.
- Housing is least expensive in Harlem.
- Houses have the most gross square feet in Gremich and the Upper East Side.
- Houses seem to have about an equal amount of land square feet.
- The Upper West side has the newest housing.
- Housing was build mostly around the early 1900s.
- Harlem has the newest housing.

- The amounts of sales seem equally distributed around Wednesday.
- December has a lot more sales than any other month.
- The least amount of sales happen in September and October.
- Home sales according to other metrics seem equally distributed.
- Sale prices vary anywhere from $10000 to $34 million, however most sale prices are around $4-5 million.

5.4 – Question 4 Solution:
For this data set it would be useful to talk to realtors, as well as people who manage historical data related to house sales. It may also be useful to talk with some homebuyers in order to get their perspective. Last but not least one should talk to the CEO in order to determine his business strategy.

5.5 – Question 5 Solution:
Steeping out of my comfort zone in terms of domain expertise is certainly challenging however it does give me a clear insight on the sort of data that should be collected, and the general methodology of analyzing it. It is very useful to learn how to read and interpret data and this exercise certainly helps one do just that.

The vocabulary used wasn't really too confusing. If I did not understand something I simply looked it up. It does help to understand the vocabulary correctly if you are going to be working on a certain domain. This is certainly a good lesson I learned from this exercise.

5.6 – Question 6 Solution:
A good data strategy is to collect as much data as possible. The one thing that we should perhaps be more worried about is how we format the data. We should try and avoid outliers and missing values as much as possible. It would also be useful to preformat the data into useful categories. If, as a data scientist, you notice that you keep repeating an operation over and over, just put the output of that operation into a new data column, where appropriate, and you would have a much nicer data set.

5.7 – Outcomes:

I have gathered a variety of plots, which you can find in:
"project1/realdirect/plots/"

This exercise taught me a bit about the domain related to realty. It also solidified my skills in using R and R Studio. I now feel much more confident with Exploratory Data Analysis and ready to undertake my own large data set. I feel like I learned some important lessons in terms of what to look for in the data, how to clean it up appropriately, and the fact that one needs to properly understand the vocabulary of the domain they are working with.

6. Airline On-Time Performance

6.1 Data Set Name and Source

The data set is named "Airline On-Time Performance" and can be found here: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
In order to download the .csv files you need to select the checkbox "Prezipped File" and use the filter to download a file for each month in 2013.

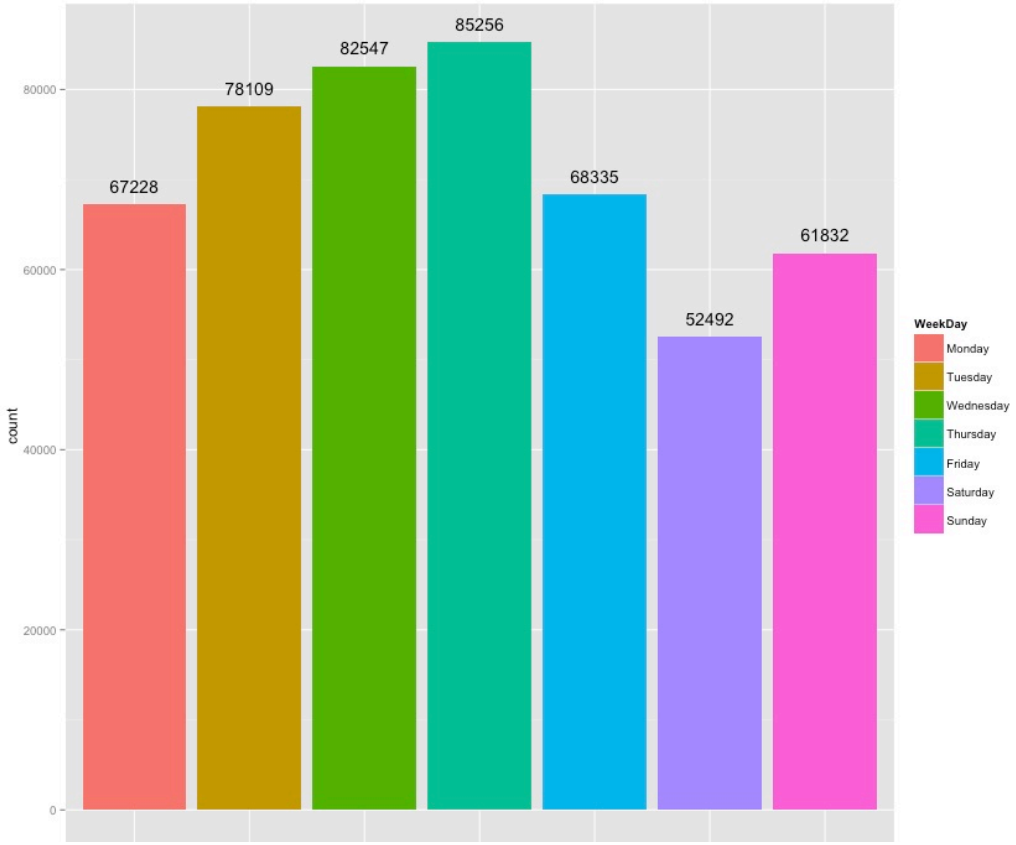
6.2 Experiments, Plots, and Interpretations

The R code for this data set can be found in the following folder:
"project1/airline_on-time_performance/aotp.r"

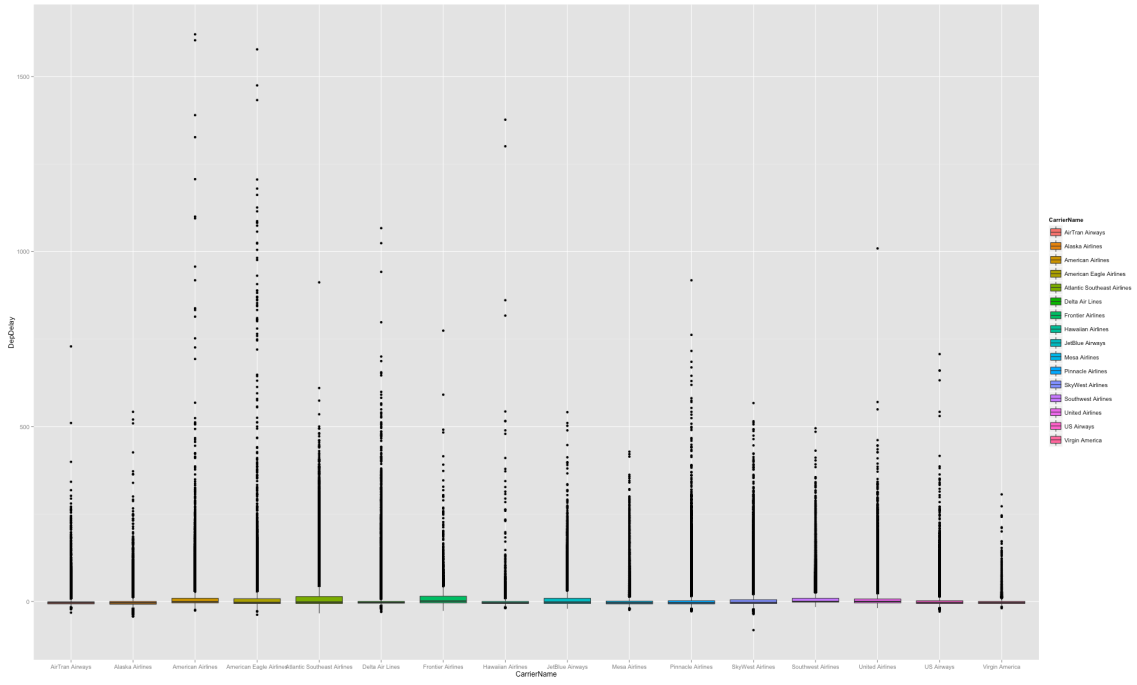
If you would like to run the code yourself, all you have to do is place the data set .csv files in a directory named "data_set" which must be located in the same directory where "aotp.r" resides. Following this, simply run "source('<path_to_file>/ aotp.r')". Doing this is going to set up the environment with everything you will need. You can now run plots, summaries, and any additional parts of the code that you desire.

6.2.1 – Experiments:

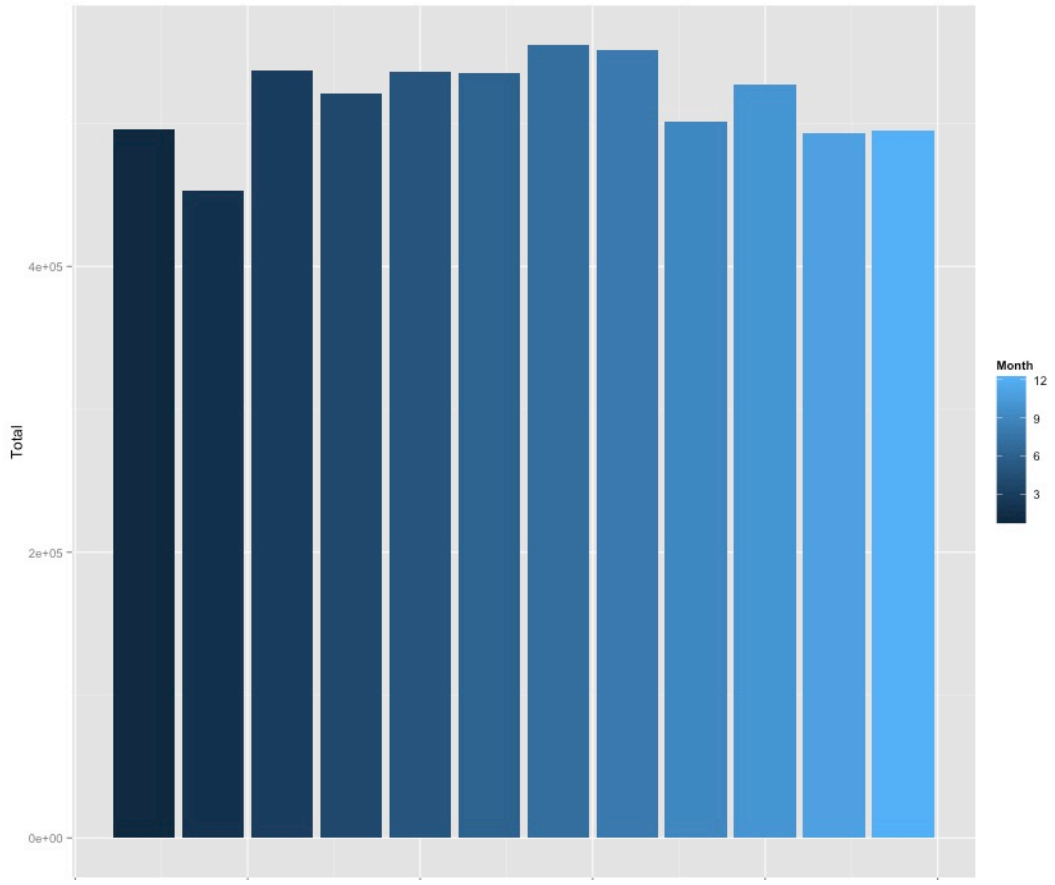
The conducted experiments are initially only on a single month (January). I began the Exploratory Data Analysis by cleaning up and categorized the data. Having a variety of categories I was able to get some interesting statistics. First I explored some raw number (flight number) data. Here is an example of this by weekday:



I also got some plots and summaries of distributions such as the following one showing us a box plot departure delays by carrier:

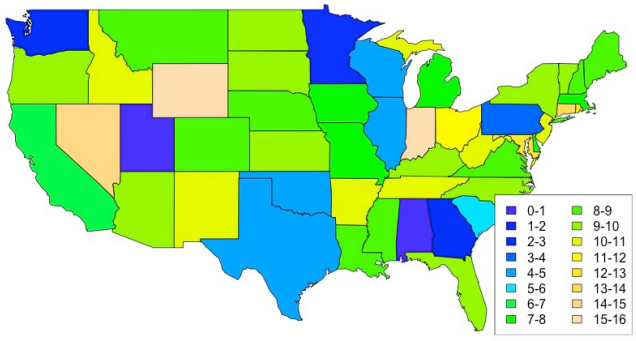


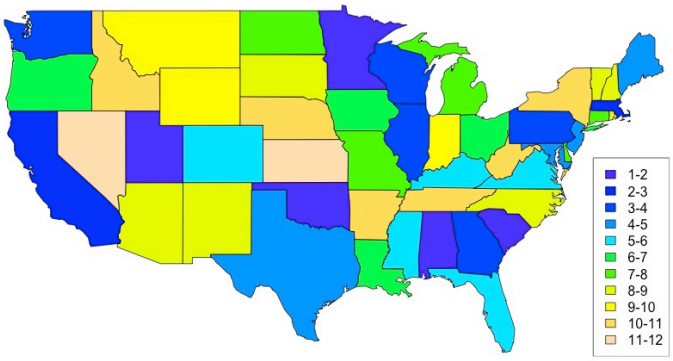
I extended these experiments to similar ones over the entire year like this one showing us number of flights per month:



I continued on with some K-means, which I plotted on a map of the United States. The K-means aimed to express geographical regions for airports within the United States. It turned out that the best K value was 4. This gave us a region for East Coast, Central, West Coast, and a separate one for Alaska and Hawaii. Here is the actual plot:

The plot is much easier to see when you zoom in so I suggest looking at the .png file in the plots directory. I also provided some extra plots for fun, which show the average departure and arrival delays per state for the year on a map.





6.2.2 – Plots:

I have gathered a variety of plots, which you can find in:
"project1/ airline_on-time_performance/plots/"

The plots are divided by month and year, although we do have a separate category for maps. The plots display a variety of data, ranging from raw flight numbers, to a variety of ways to display different categories of flight departure/arrival delay data.

6.3.3 – Interpretations:

Here are my observations, after conducting Exploratory Data Analysis on the Airline On-Time Performance data set:

- There is a linear relationship between arrival delay and departure delay.
- As departure delay increases, arrival delay increases.
- The longer the distance the fewer the flights.
- Southwest Airlines has the most flights, Virgin America the fewest.
- February has the least amount of flights, perhaps due to fewer days.
- July and August have the most flights.
- California and Texas have the most flights.
- Saturday is the least busy day in terms of flights.
- Alaska and Hawaiian Airlines have the least amount of delays.
- Southeast Airlines and JetBlue have the highest departure delays.
- The day with least delays is Saturday.
- The day with most delays is Thursday.
- Illinois and New Jersey have the highest departure delays.
- Arkansas and Minnesota have the least departure delays.
- Southeast Airlines, Frontier Airlines, and JetBlue have the most arrival delays.
- New Jersey has the highest arrival delays.
- Utah has the lowest arrival delays.
- The United States can be nicely divided into 4 regions in terms of flights; East Coast, Central, West Coast, and Alaska and Hawaii
- There aren't many flight arrivals between 12 AM and 7 AM.
- There aren't many flight departures between 11 PM and 5 AM.
- There are around 70000 flights a day except for on weekends.

7. Lessons Learned

The Exploratory Data Analysis and Airline On-Time Performance project has taught me many valuable lessons that I am definitely going to take advantage of. First and foremost I acquired a good understanding of the R language, and R Studio. This is definitely a very valuable skill to have; in fact I have seen a lot of job posting in major companies seeking for such knowledge. I have also gotten some experience with a few different domains, something that is always great. Exploratory Data Analysis is certainly something that doesn't come quite naturally, but now I feel a lot more confident when it comes to it. I also solidified my understanding of K-means.

R is a language I definitely had never used beforehand. In fact I had never used a language designed for statistical analysis, other than perhaps MatLab. The language itself wasn't too difficult to learn but the vast libraries and things one can accomplish with it are certainly a challenge. I am very happy that I now have a very good understanding of analyzing data, graphing, plotting, summarizing, and doing other fun things in R.

The Airline On-Time Performance section of the project has taught me some very useful things regarding flights. Now when I select flights I will certainly be pickier and refer back to my results before purchasing a ticket. There are definitely some companies and days that just aren't all too great for flying.

Last but not least I am a lot more confident now when it comes to working with big data sets, than I was before. This will be very helpful for future projects in the class, and beyond. As I mentioned such skills are definitely something a lot of large tech companies are looking for.