Exploratory Data Analysis of Airline Delay dataset for the year 2008.

Gaurav Gosavi (5009 6832)

Ameya Patil (5009 7850)

Data Intensive Computing CSE 587 Spring 2014

Contact: gauravgo@buffalo.edu ; ameyapat@buffalo.edu

DATA SOURCE : http://stat-computing.org/dataexpo/2009/the-data.html

(657 MB)

Abstract

In this project we have attempted to perform Exploratory Data Analysis on a distinct exhaustive dataset comprising of flight details (including delays) for the year 2008. From this exploratory data analysis we have tried to infer meaningful patterns and conclusions.

Exploratory data analysis here has been performed using the statistical programming language R. In R we have used some standard packages like maps, maptools, ggplot and sqldf for performing the analysis.

From our analysis we have plotted several histograms, bar plots and line charts for presentation. We have inferred many interesting and cool patterns and observations from the same.

Project Objectives

This project "Exploratory Data Analysis in R on Flight Details Dataset" will meet the following objectives:

- Learning what Exploratory Data Analysis really means.
- Learning the Programming language R
- Learning and Understanding the Statistical modelling Concepts
- Understanding Aviation Statistics and problems usually encountered
- Laying the foundation of Data Analysis
- Team work

Project Approach

In this project we have scrutinized and understood Chapter 2 of Doing Data Science book. We have solved the sample problems given on pages 36-44. We have then applied the knowledge gained from these problems to actually perform Exploratory Data Analysis in our real world data.

As stated earlier we are using the statistical programming language R for carrying out our EDA as it is a very elegant and powerful language capable of handling large amounts of data. Our data size is approximately 600 Megabytes, and R can comfortable handle our dataset and perform analysis on it. We use an IDE named R Studio for development and coding in R in our project.

New York Times Data Set analysis with questions and Answers.

Q.1 Create a new variable, age_group, that categorizes users as "<18",

"18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

Ans.

```
d1 <- read.csv("nyt1.csv")

d1$agecat <-cut(d1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))

age_group <- d1$agecat
```

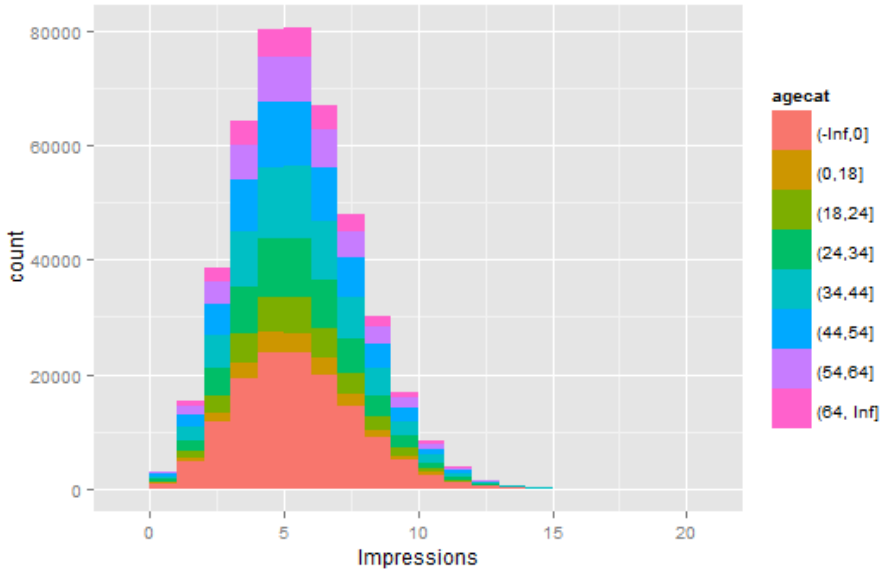
Q2. For a single day,

1. Plot the distributions of number impressions and click through-rate (CTR=# clicks/# impressions) for these six age categories.

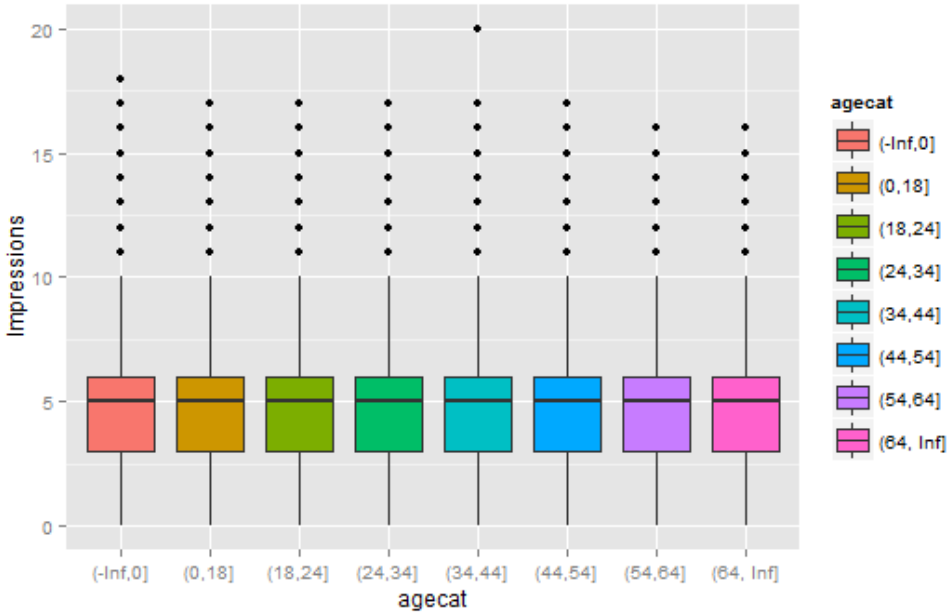
Ans.

```
library(ggplot2)

ggplot(d1, aes(x=Impressions, fill=agecat))+geom_histogram(binwidth=1)

ggplot(d1, aes(x=agecat, y=Impressions, fill=agecat))+geom_boxplot()
```

Output:




```
ggplot(d1, aes(x=agecat, y=Impressions, fill=agecat))+geom_boxplot()
```



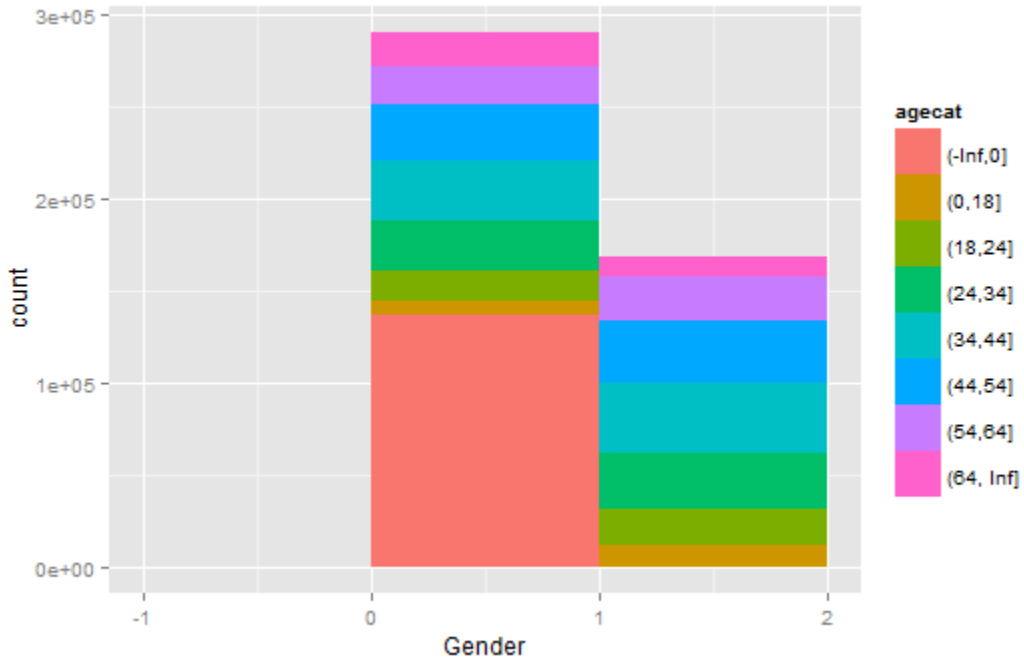
2. Define a new variable to segment or categorize users based on their click behavior.

Ans:

```
d1$ctrcat <-cut(d1$Clicks/d1$Impressions,c(-Inf,0,0.3,0.6,0.9,1.2,1.5,2,Inf))

d1$scode[d1$Impressions==0] <- "NoImps"

d1$scode[d1$Impressions >0] <- "Imps"

d1$scode[d1$Clicks >0] <- "Clicks"
```

3. Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example).

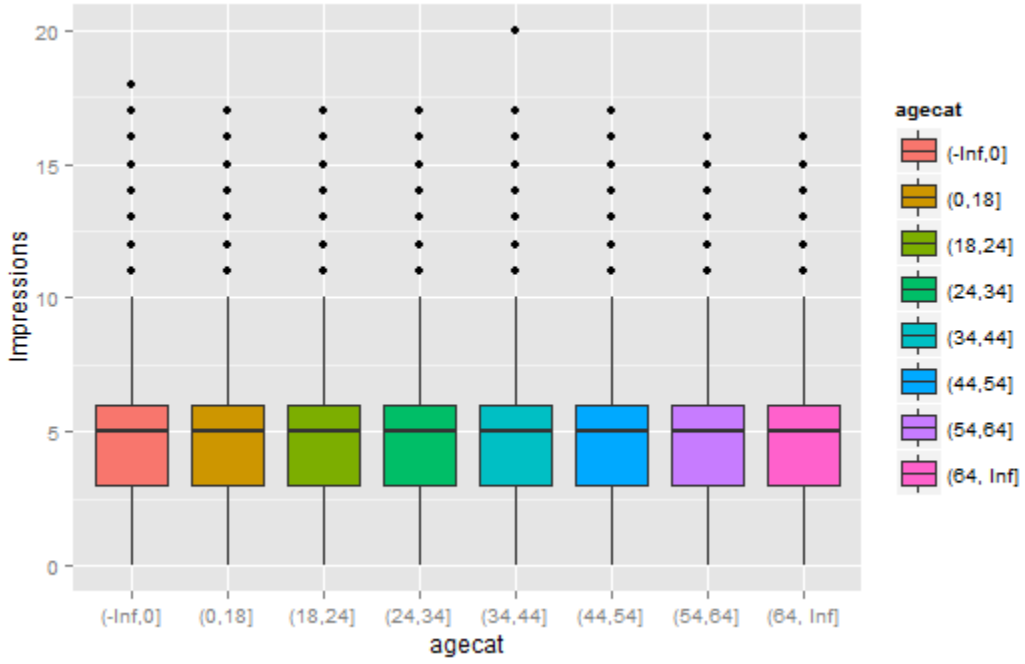
ggplot(d1, aes(x=Gender, fill=agecat))+geom_histogram(binwidth=1)



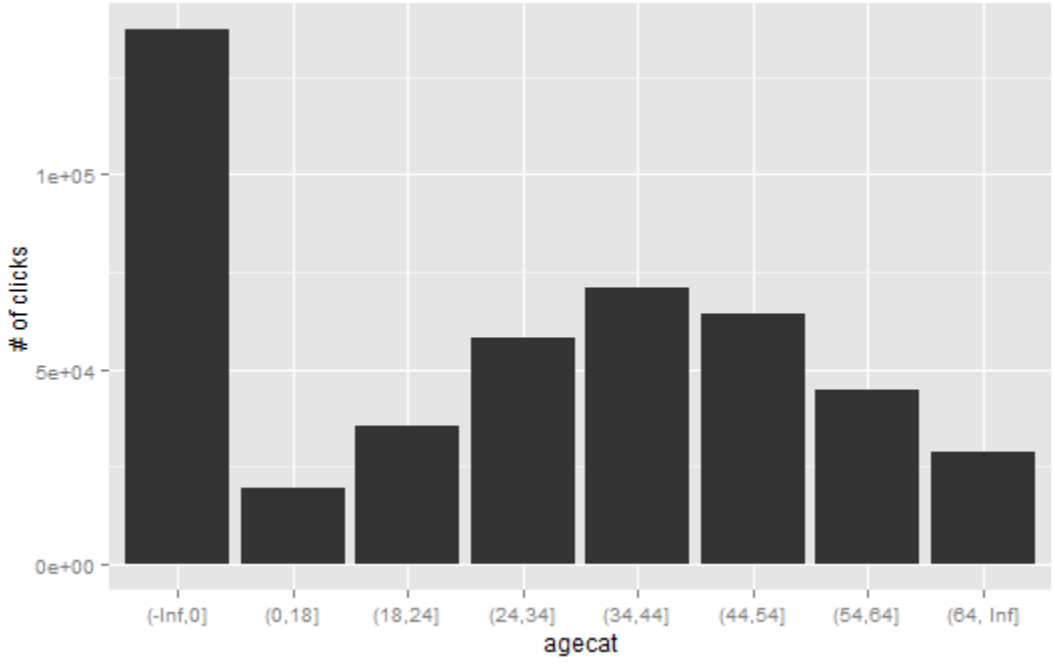
Number of persons of each gender with their age category demarcated.

4. Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quintiles, mean, median, variance, and max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time—what will compress the data, but still capture user behavior.

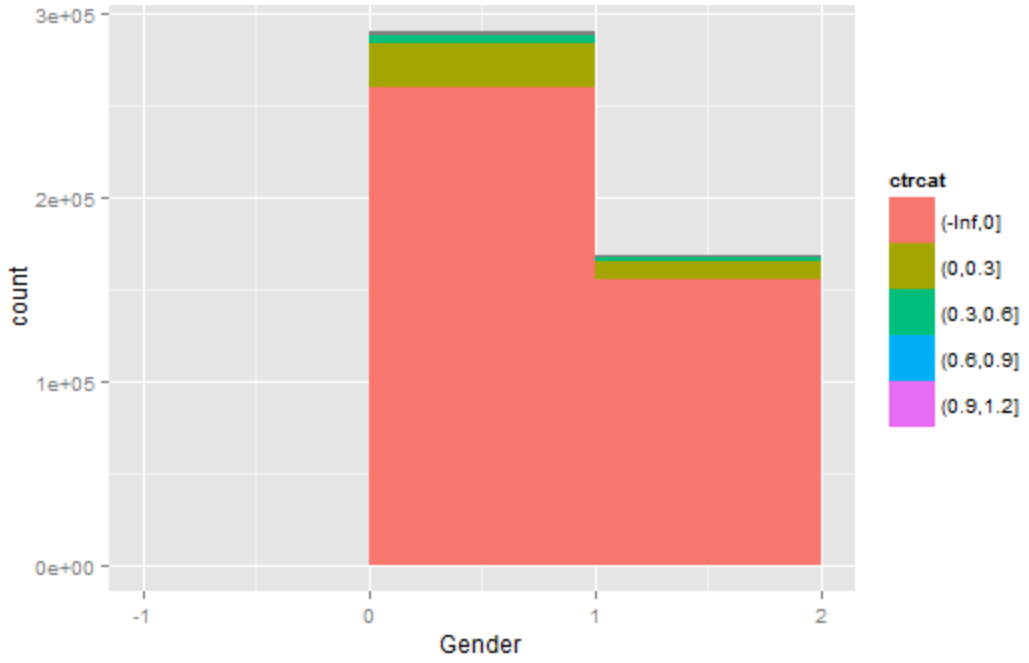
Ans. Box plot of Age Categories vs Impressions



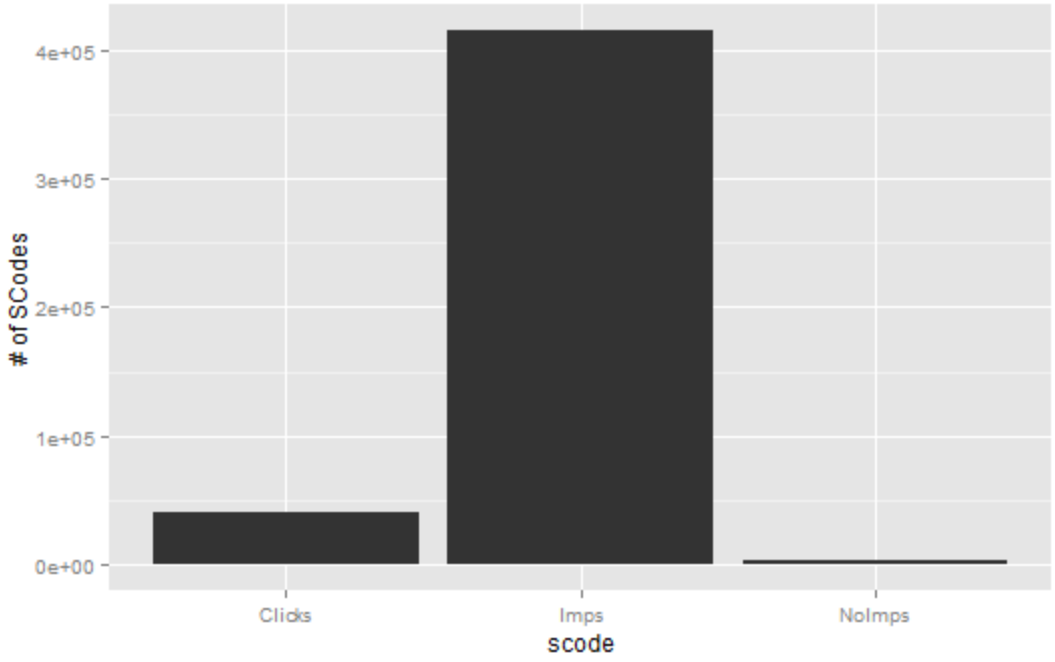
Bar chart of age category vs number of users in that category



Histogram of Click through rate classified according to gender:



Bar plot of Number of Scodes



Q3. Now extend your analysis across days. Visualize some metrics and Distributions over time.

Ans. From the first 2 questions we can see how the exploratory data analysis is done. To extend this analysis over all the days we will just loop over all the csv data files. This can be done with the help of the following piece of code.

```
myFiles <- list.files(path="C:\\Ny times", pattern = "*\\.csv$")

for (file in myFiles) {

  EdAnalysis(file)  }
```

We have named our total computation as a function names (EdAnalysis).

Q4. Interpretations from the patterns

We have drawn 3 graphs above. We infer that maximum data is about Imps (with impressions>0 and clicks=0), as the Imps column is the highest.

Also it can be noted that the click through rate amongst Gender 0 (male) is very high compared to Gender 1.

We can also note that maximum portion of the data is for people whose age is specified as 0, i.e. nonexistent customers.

Real Direct Case Study Answers (Questions on page 48-49)

1) 1. As Real Direct deals with the real estate industry, it would probably be a good idea to have our engineers log data about square feet area of the property, area in which it is located, price of the property per square feet, specifications about the property, like number of bed/baths, etc. and some added features of the property like parking, pool, gym, etc. We can also include attributes like number of similar units sold, asking price and actual selling price. A column on user clicks on a particular unit would also be useful.

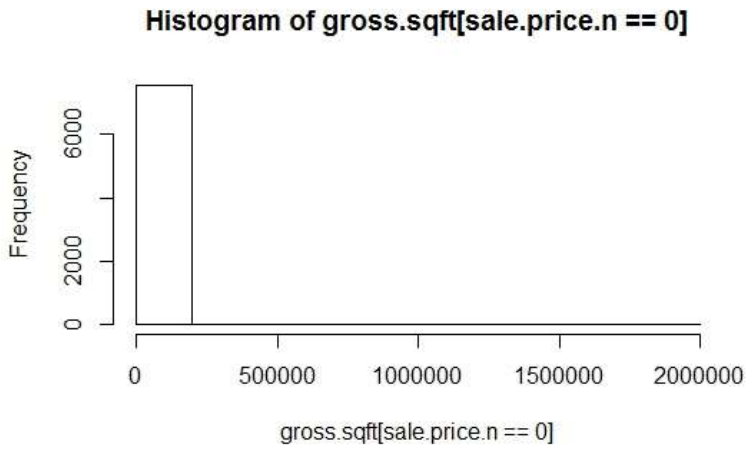
   2. The last attribute describes the amounts of property units of the particular sold previously, this attribute will help monitor the product usage by telling us how much property is being sold of a particular type. We can use the click rate on our property units to determine our hot property units and accordingly pay attention to them.

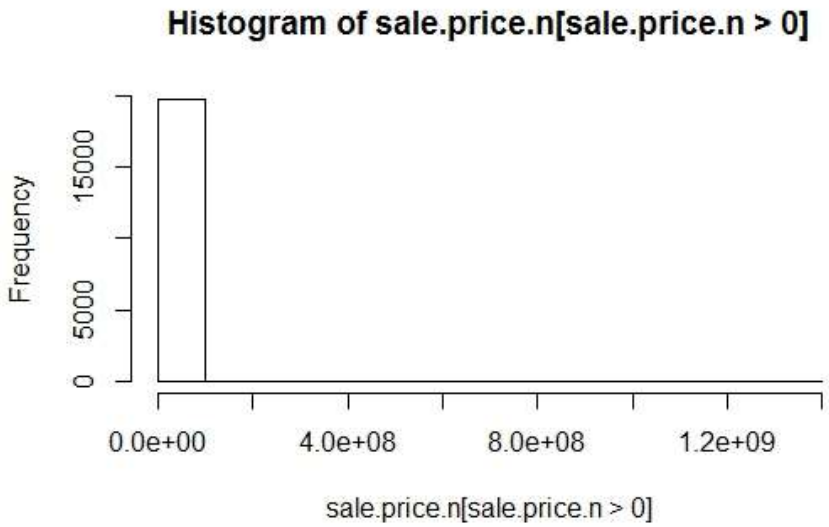
   3. In this case again, we can use number of units sold as an advertisement of the Real Direct, signifying how much market dominance and competency our company currently possess. Using our clicks number we can determine the popularity of our units.

2) Please Refer attached file “Realdirect-manhattan.R”

3) For gathering data and extracting useful messages from it, along with the data scientist, we should talk with the architects, buyer, estate agent, and other service men regarding the property. Thus, we can also determine and optimize our maintenance cost of that property.
4) 1. We got an insight of how data statistics can be applied to the field of real estate. As there were hardly any completely “out of the blue” term used as a variable in our csv dataset it wasn’t too difficult to comprehend and visually understand the dataset.
   2. No, not really.
5) Looking at real direct, we conclude that it is a good strategy to maintain consistent and pre-cleaned data sets spaced over regular intervals. An exploratory data analysis of these data sets by a good data scientist can bring about new patterns from which we can infer conclusions. These inferred conclusions can help us reshape but business operations and planning with a view of optimization.

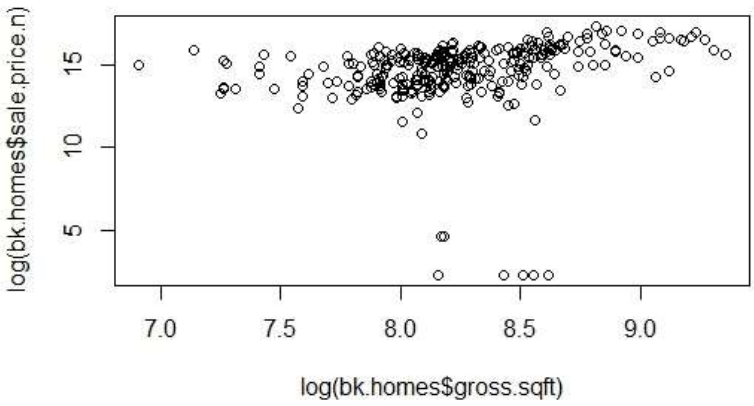
Real Direct Outcomes

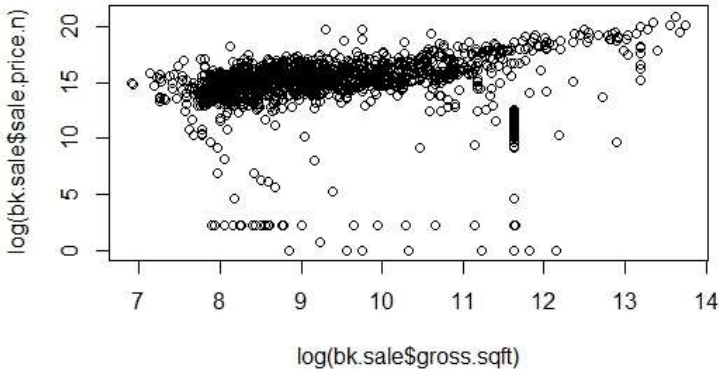




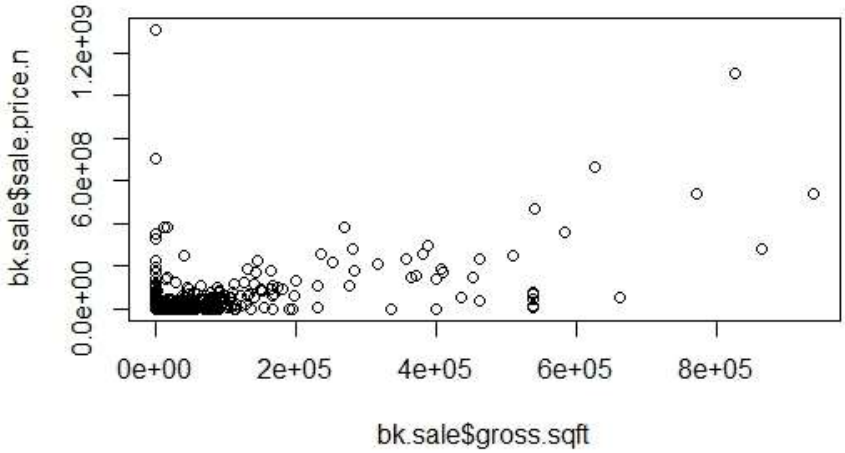
Making sure there are no unexpected sales price values.

Keeping only the actual sales and removing false positives.

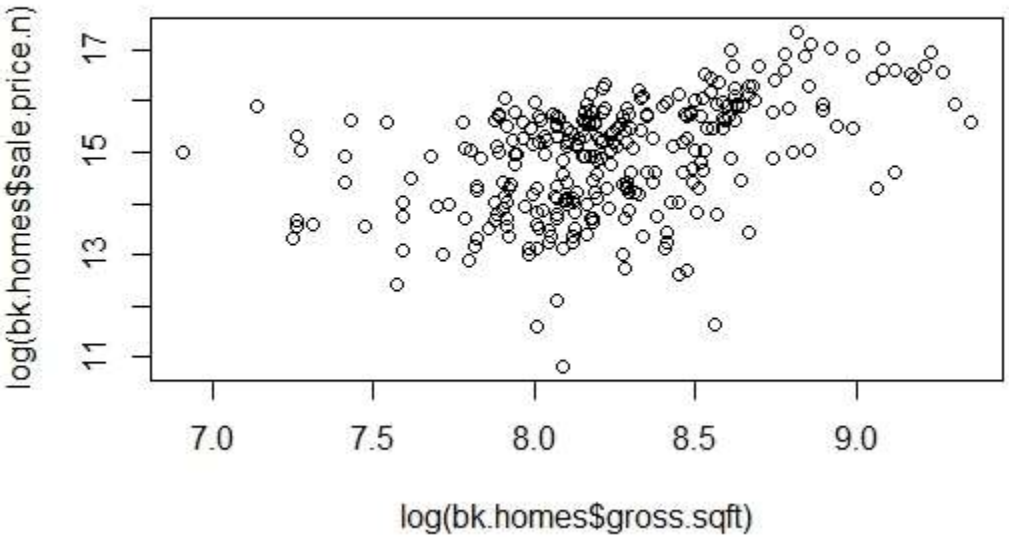




Plot for 3 Family Homes sales



Removing outliers which weren't actual sales.



Our Data – Airline Delay Statistics in 2008

Our dataset consists of stats giving delay of airlines from different destinations in the US. The main aim is to find interesting inferences from EDA which will be useful to a passenger (say Mr. X) using flight as a mode of transport.

1) Firstly we need to get an airline brand which is reliable and has minimum delays over 2008. We found :

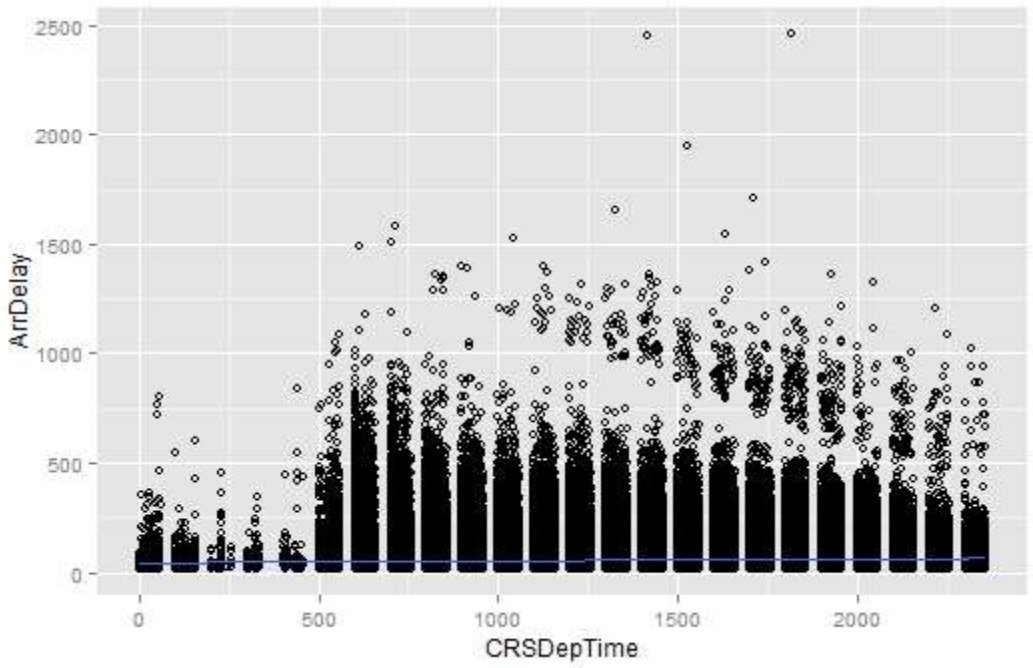
	Description
1	Mesa Airlines Inc.

Is the one which offers minimum delay over the year in general.

2) Okay one airline company we should definitely avoid which has maximum delays associated with it is :

1	Northwest Airlines Inc.

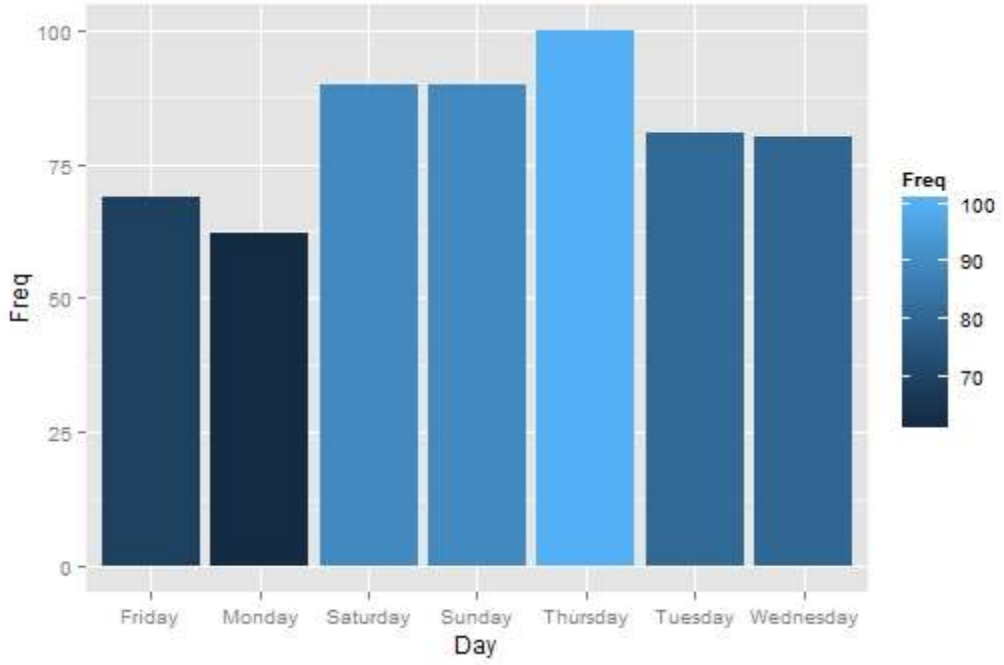
3) In 2008 let's find the scatter plot of Scheduled Departure time vs Arrival Delay



INFERENCE : We can infer that Arrival delay and Scheduled Departure time are close in early morning and dawn. There is larger difference between them during the day from 6am to 3pm. The gaps narrows during the night.

Hence, Mr. X should try to book his flight for early morning or in the night.

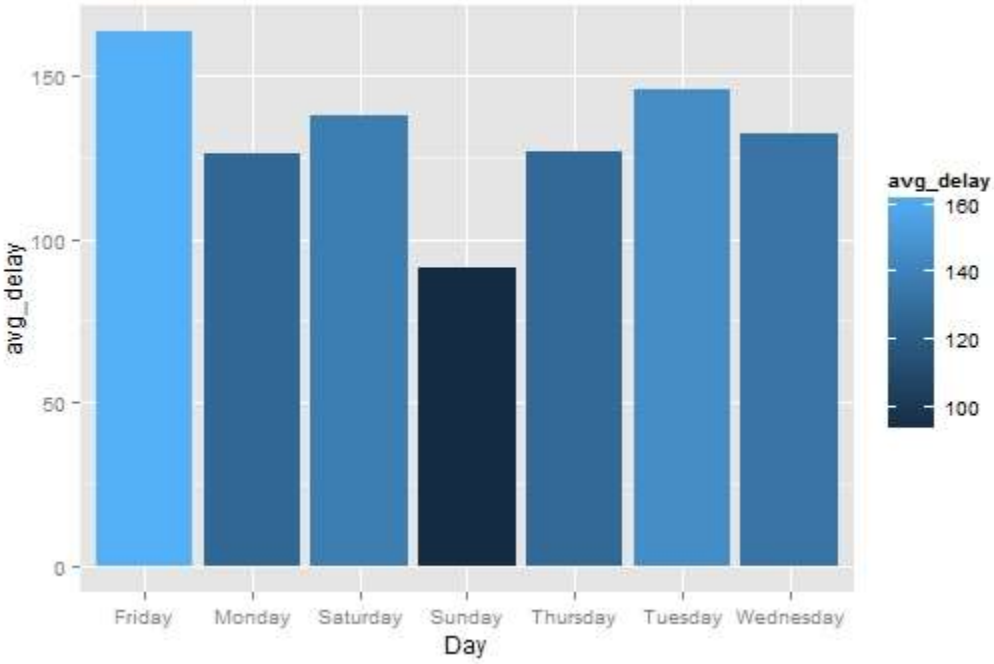
4) No. of departures from LA (LAX) to Boston(BOS) throughout the week histogram



INFERENCE: Thursdays and Saturdays have greater number of departures in most of the major airports. Attached is example from LA to Boston. Hence with greater no. of flights on these days there is greater number of expected **Departure** delay.

Mr. X should try to avoid Saturdays or Thursdays when to fly from major airports due to departure delays.

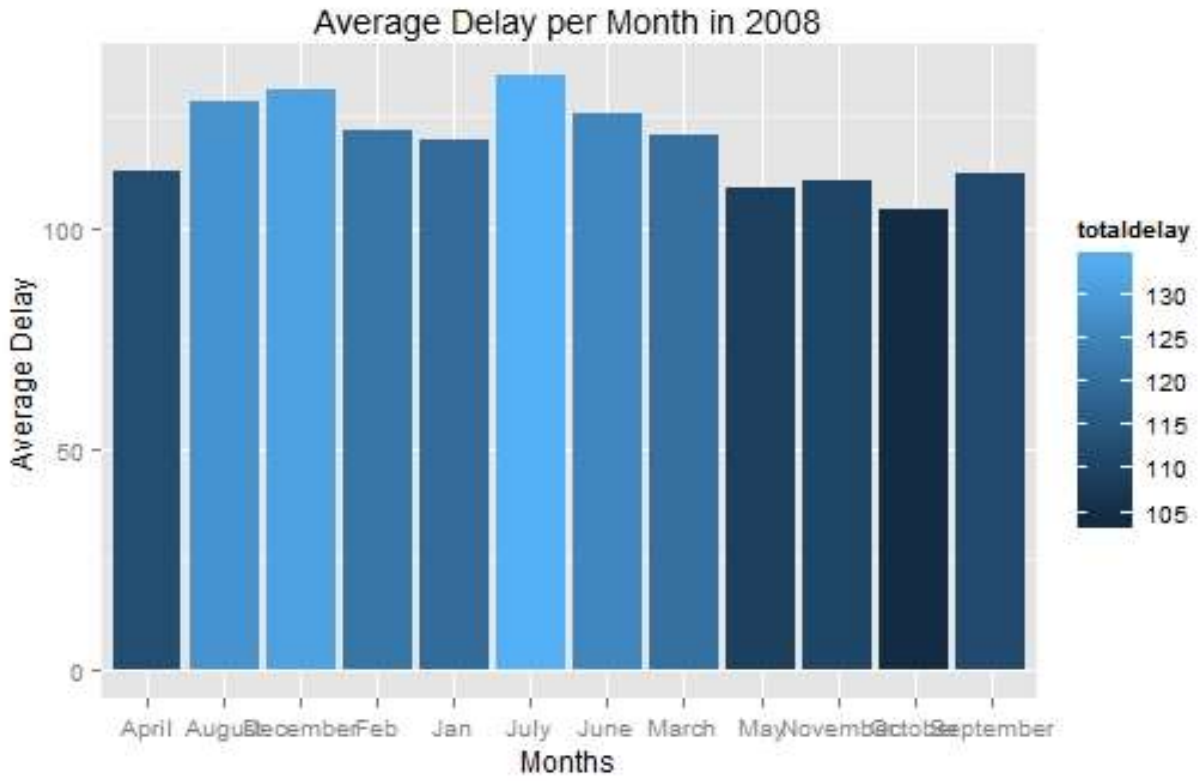
5) Average Delay Throughout the week from LA to Boston



INFERENCE : Average total Delay is greater on Friday, Tuesday and Saturday. Note jump in delay from Sunday to Monday indicating impact of working businesses influencing delays.

Mr. X should avoid Saturday or late Friday if he doesn't want delays in general. Sunday has overall the least delays as it is a business holiday.

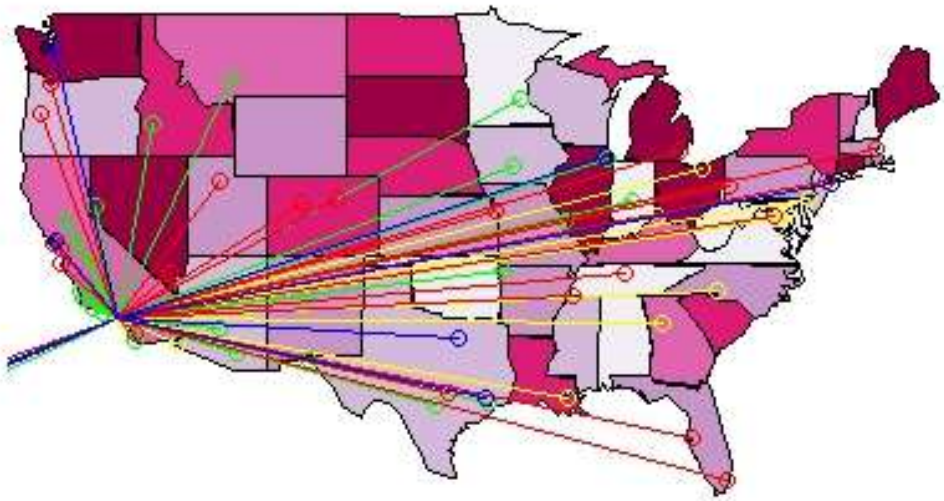
6) Average Delay per month in 2008.



INFERENCE: We can see that there is greater average delay during August, December and July. December has greater delay as it is the holiday season. July and August has greater delays as most like to travel in the US during peak of the summer season. The delay jumps significantly from November to December as the holiday season starts.

Mr. X should expect greater delays during July, August and December. Late September, October and May are good times to travel.

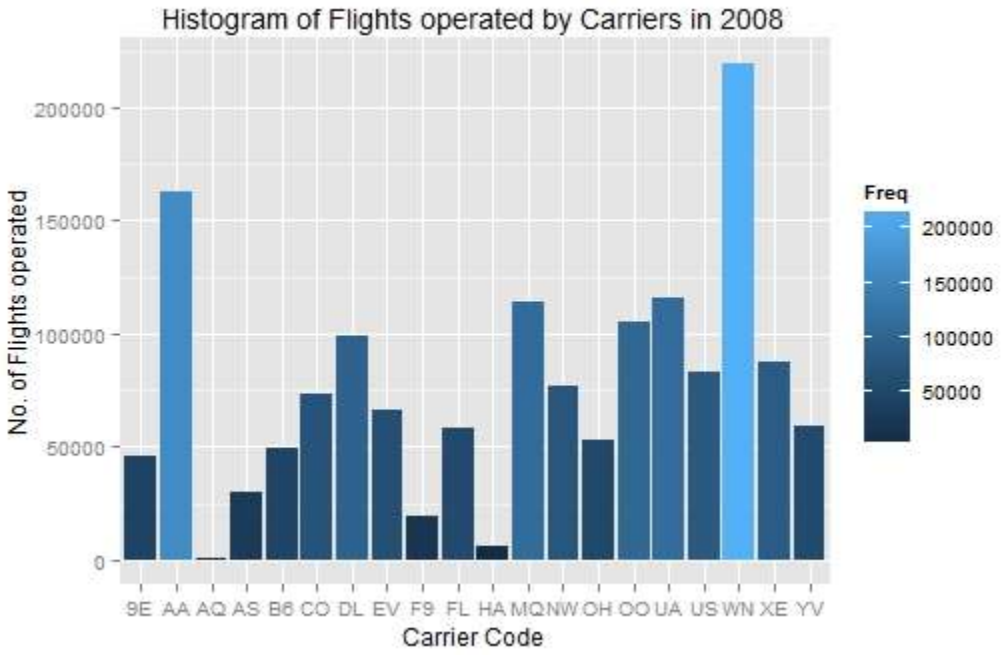
7) Delay Distribution of Flights from LA between 4pm and 6pm.



	-50 < Total Delay < 50
	150 < Total Delay < 500
	500< Total Delay < 1500
	Total Delay > 1000

INFERENCE: Delays are greater for destinations which are major cities, compared to other destinations. Delays are lesser to smaller cities comparatively.

8) Histogram of Airline Carrier Companies operating flights in 2008.



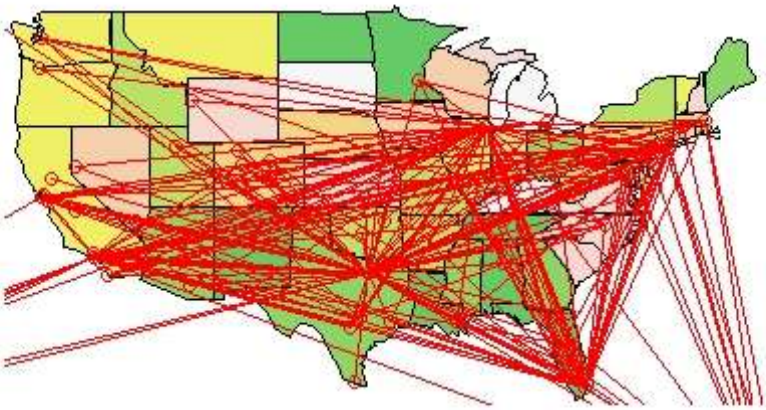
INFERENCE: (WN) Southwest Airlines is a major Airline company with the most number of flights operated through the year. Another major airline company is (AA) American Airlines.

Mr. X should use these major airlines if he desires brand service. Local airlines like Air Tran or Continental Airlines offer cheaper alternatives for short distance commutes.

9) Let’s look at Flight tracks of Southwest Airlines which is a major nationwide airline company.

INFERENCE: We can see that Southwest airlines is truly a nationwide Airline with services present in almost all the states connecting every major city. It has a wide mesh network.

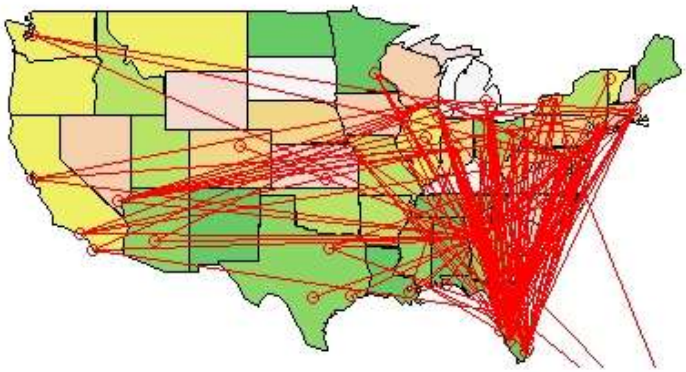
Now let’s look at another major airline company American Airlines.



INFERENCE: We can see that it is not as widely spread as South west airlines but it has hubs in major cities of major states acting as connectors.

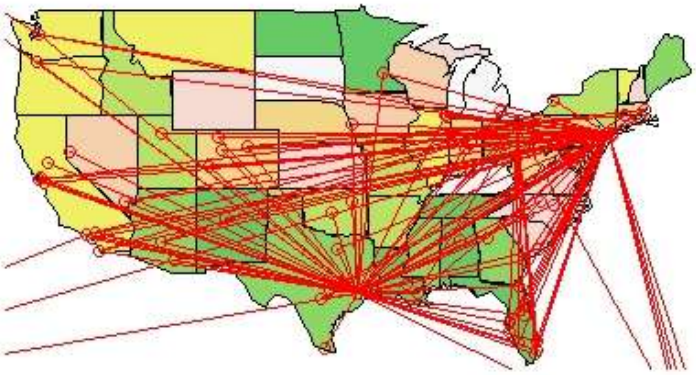
10) Let's look at flight maps of local airline companies.

Flight Map of AirTran Airways.



INFERENCE: AirTran Airways is an American low-cost airline operating locally mostly on the East coast. This is a local airline providing services on the East coast with only some flights to major cities on the west coast.

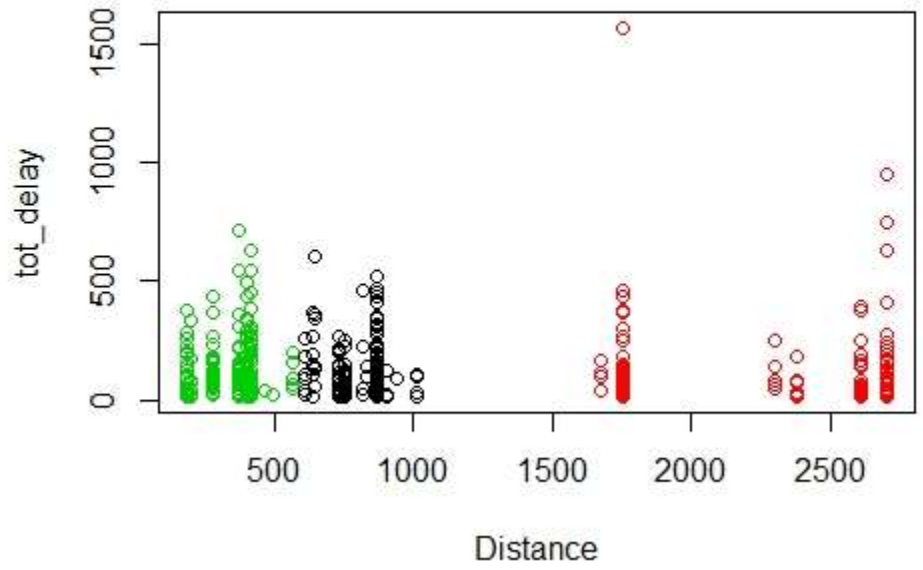
Flight Map of Continental Air Lines Inc.



INFERENCE: Another locally operated airline company, Continental Airlines is a U.S. airline, founded in 1934 and headquartered in Houston, Texas. We can see it uses its headquarters as a hub in Houston and NYC.

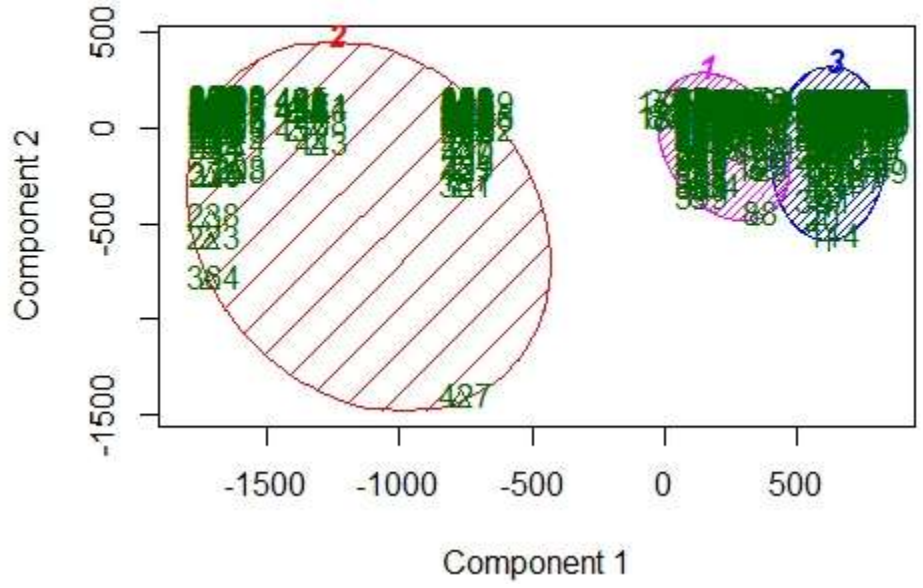
11) K Means Analysis

K means done on Airline data for destination as Boston.



INFERENCE: We have used k=3 in the analysis. We can see the delay in three clear separate clusters showing the relation between total delay and Distance of commute to Boston.

elect comp.Distance,comp.tot_delay from comp where Dest



INFERENCE: The three clusters shown in bounded boundaries.

Lessons Learned

From this project we have learned how to perform exploratory data analysis using R. We have also observed and we now know of some common patterns and phenomenon occurring in the aviation industry. We have also learnt statistical modelling concepts and their use and implementation using R programming language. Importance of EDA before big data analysis and what one might look for or design methodology for big data analysis.

Along with this we also learnt certain non-technical skills such as multi-tasking and team work