USING EDA TO FIND MEANINGFUL INFORMATION FROM DATA

VIVEKANANDH VEL RATHINAM, JERRY ANTONY AJAY

AFFILIATION: CSE587, UNIVERSITY AT BUFFALO, vvelrath@buffalo.edu, jerryant@buffalo.edu

1. Abstract

This project deals with finding meaningful information from three sets of data, i.e., the New York Times dataset, Real Direct dataset and Airlines information dataset for the year 2007. In the New York Times dataset, we identify the usage patterns of people belonging to different age groups. Using the Airlines dataset, few statistical inferences were made that deal with distribution of flying patterns, etc. The Real Direct dataset was used to get valuable information from Real Estate business in Manhattan.

2. Project Objectives

The objectives of determining the major airport hubs in the USA are:-

- Learn and explore statistical modeling.
- Learn R Language for data analysis.
- Learn K-means clustering and EDA procedures.

3. Project Approach

We worked on the examples given in Chapter 2 of Doing Data Sciences and applied the knowledge to an outside dataset that has data of all the flight that flew in the year 2007. RGui as well as RStudio was used to analyze this data and meaningful information was gathered and bar graphs, histograms and k-means graphs were plotted.

As one of our analysis, the origin and destination values were extracted from the dataset. This data was then stored in a table that extracted the unique airport IDs from the respective columns. These unique airport IDs were then sent to 'geocode' which outputted the latitude and longitude information of the respective airports. This information was once again stored in a vector. The frequency of the number of the unique airports encountered in the tuples were saved in a table. This frequency information was then plotted on a US map having the airports represented by a dot with a cex value of 0.3. The more the frequency of airports, the denser the plot becomes.

As another part of our data analysis using EDA approach, we plotted a histogram that depicts the information between the numbers of arrivals to the number of departures. The more the difference, the more probable the airport is a hub/maintenance center.

4. NY Times Dataset

a. Create a new variable that categorizes users as "<18","18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

```
head(data1)
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
```

b. For a single day: Plot the distributions of number impressions and clickthrough-rate (CTR=# clicks/# impressions) for these six age categories.

```
data1$hasimps <-cut(data1$Impressions,c(-Inf,0,Inf))
summaryBy(Clicks~hasimps, data =data1, FUN=siterange)
ggplot(subset(data1, Impressions>0), aes(x=Clicks/Impressions,
colour=agecat)) + geom_density()
ggplot(subset(data1, Clicks>0), aes(x=Clicks/Impressions,
colour=agecat)) + geom_density()
ggplot(subset(data1, Clicks>0), aes(x=agecat, y=Clicks, fill=agecat)) + geom_boxplot()
ggplot(subset(data1, Clicks>0), aes(x=Clicks, colour=agecat))
+ geom_density()
```

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

```
data1$scode[data1$Impressions==0] <- "NoImps"
data1$scode[data1$Impressions >0] <- "Imps"
data1$scode[data1$Clicks >0] <- "Clicks"
```

d. 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).

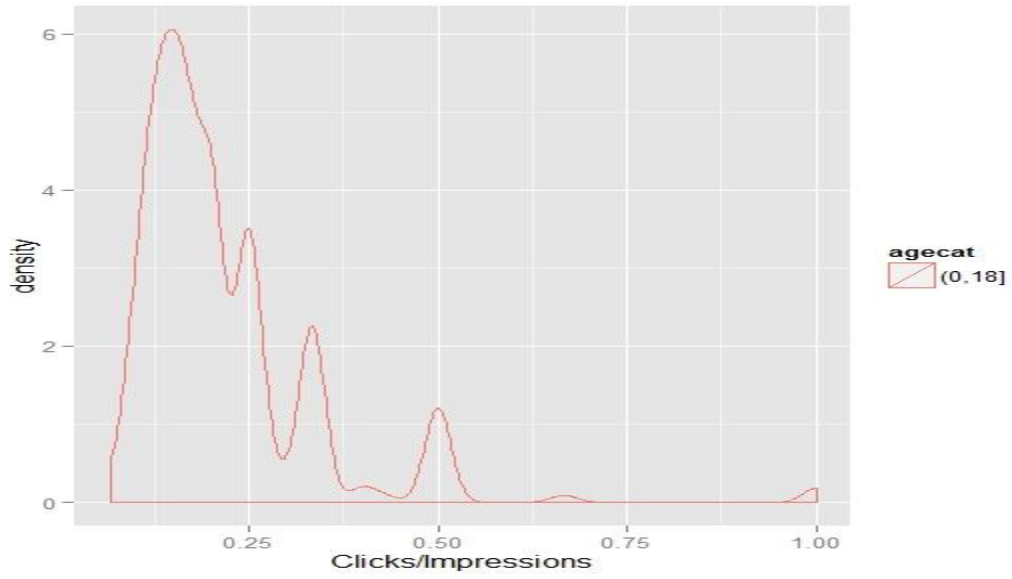

Figure 1 Males less than 18

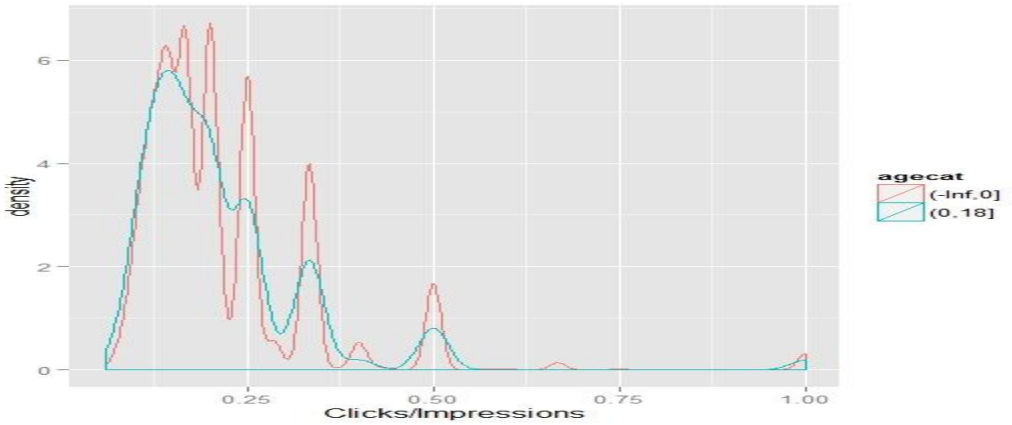

Figure 2 Females less than 18

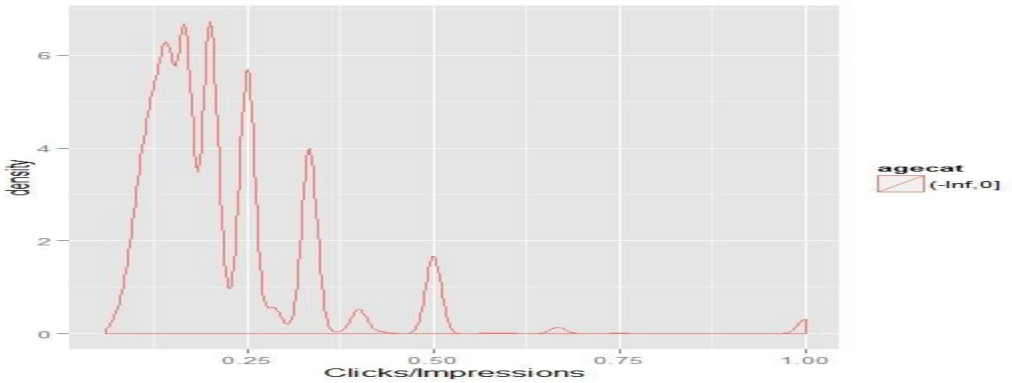

Figure 3 Not logged in users

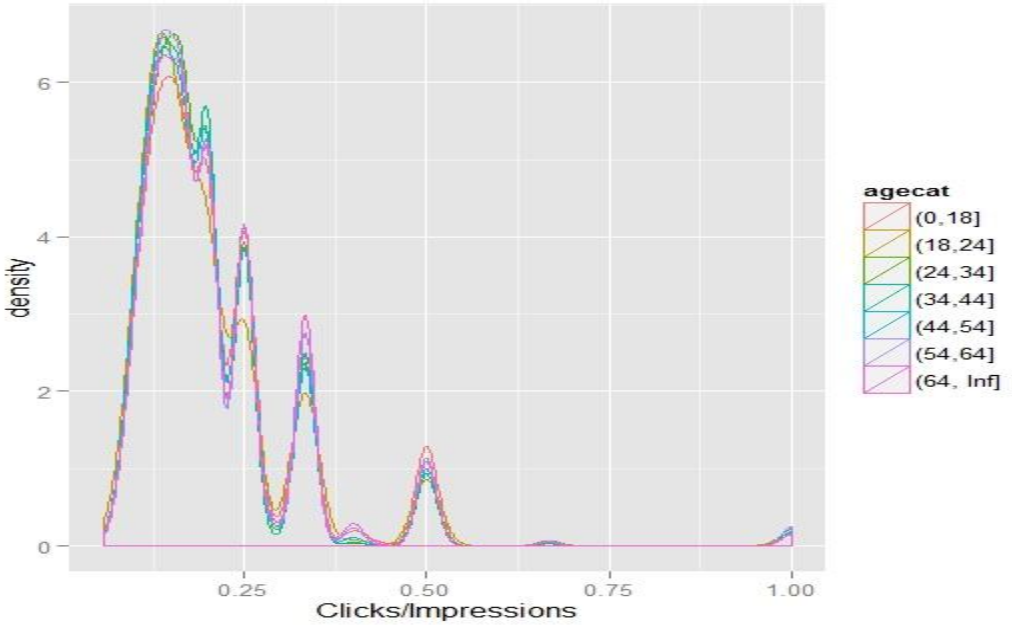

Figure 4 Logged in users

e. Create metrics/measurements/statistics that summarize the data.

```
install.packages("doBy")
library("doBy")
siterange <- function(x){c(length(x), min(x), mean(x), max(x), var(x))}
#Summarizing across Age categories
summaryBy(Gender+Signed_In+Impressions+Clicks~agecat,data =data1)
#Summarizing across Signed in users and Not Signed In Users
summaryBy(Gender+Impressions+Clicks+Age~Signed_In,data =data1)
```

3. Now extend your analysis across days. Visualize some metrics and distributions over time.

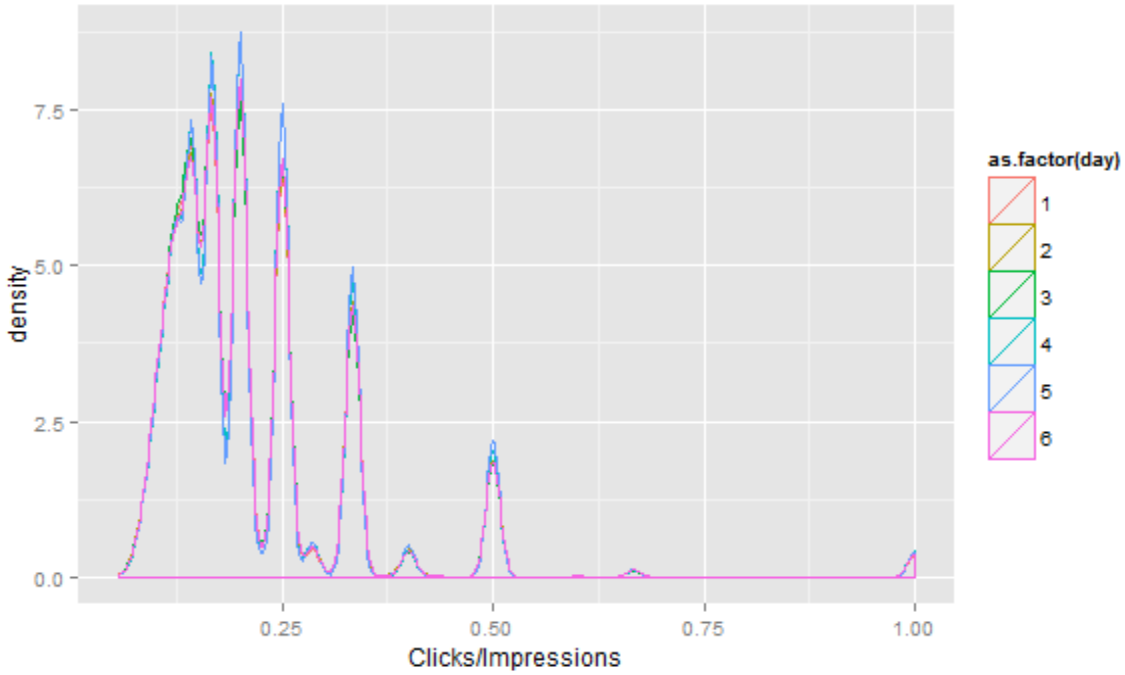

Figure 5 Analysis across six random days

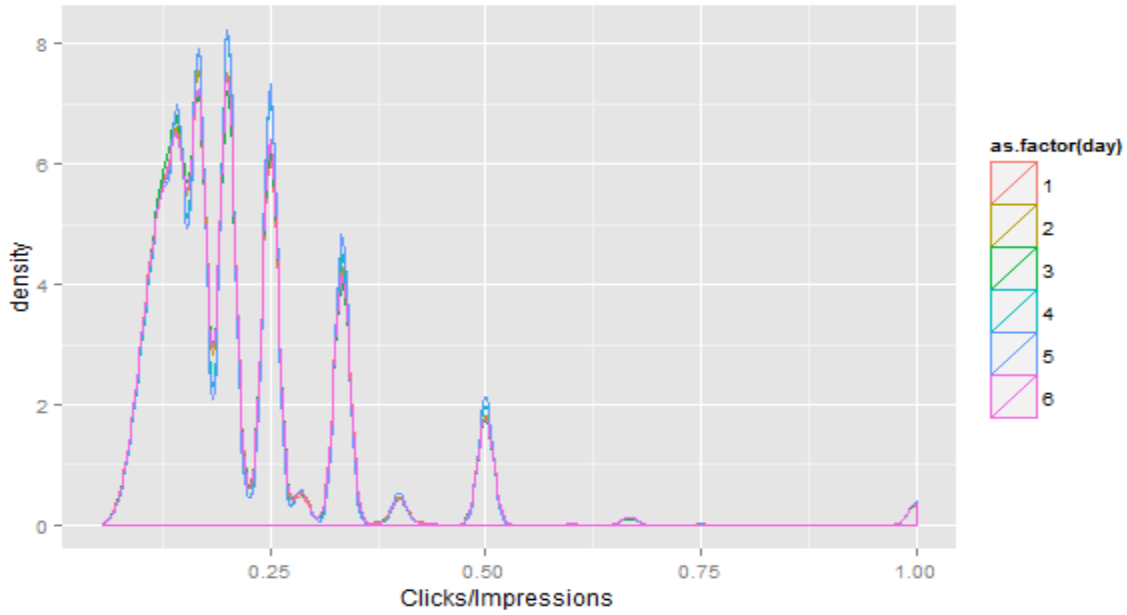

Figure 6 Analysis across six random days (females)

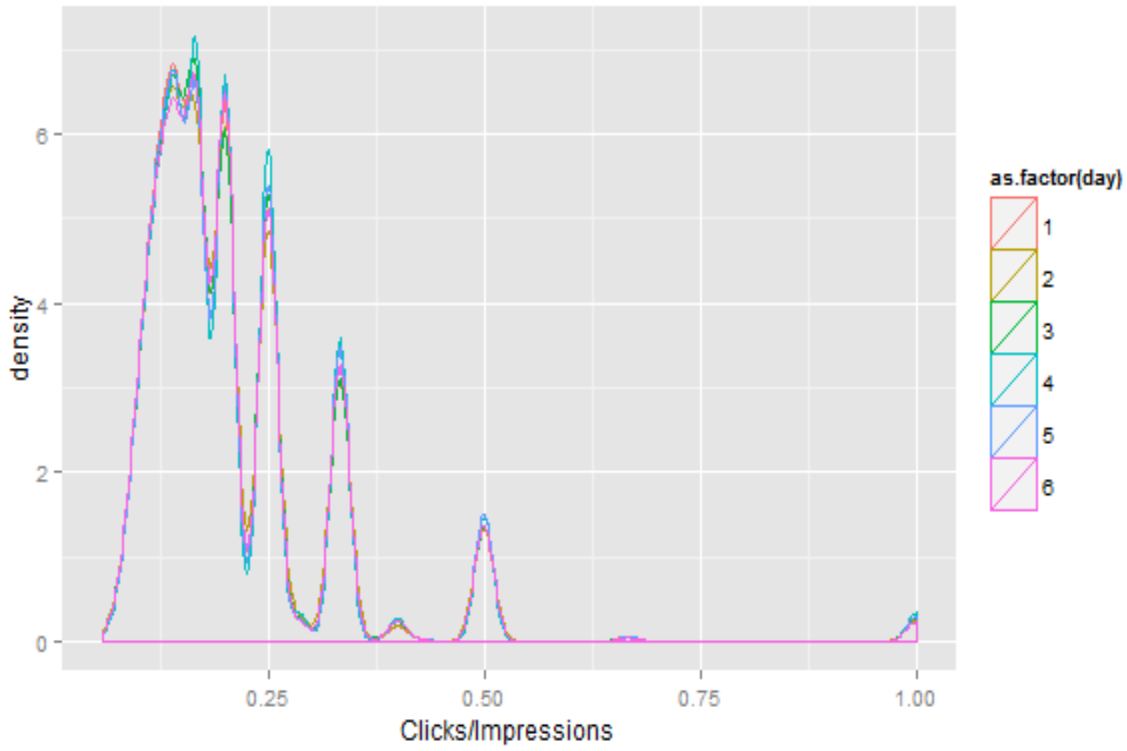

Figure 7 Analysis across six random days (males)

4. Describe and interpret any patterns you find.

We find that the across days the Clicks/Impressions vs density peaks at around 0.17 both for males and females. This symbolizes that users click on the ad 17% of the time no matter whether they are females or males.

We also find that people of the age group 24-34 followed the trend more closely irrespective of the gender compared to people of other ages.

5. RealDirect Data

Explore its existing website, thinking about how buyers and sellers would navigate through it, and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of RealDirect user-behavior data could be used to inform decision-making and product development.
Come up with a list of research questions you think could be answered by data:

- What data would you advise the engineers log and what would your ideal datasets look like?
- How would data be used for reporting and monitoring product usage?
- How would data be built back into the product/website?

Some of the research questions that could be asked to get meaningful information is to first identify the location of maximum number of listings in a city and then analyze about how to increase the number of listings from this particular locality. Further questions can be asked about the reason of high selling rate in a locality, the distribution of wealth in various localities, etc.

Summarize your findings in a brief report aimed at the CEO.

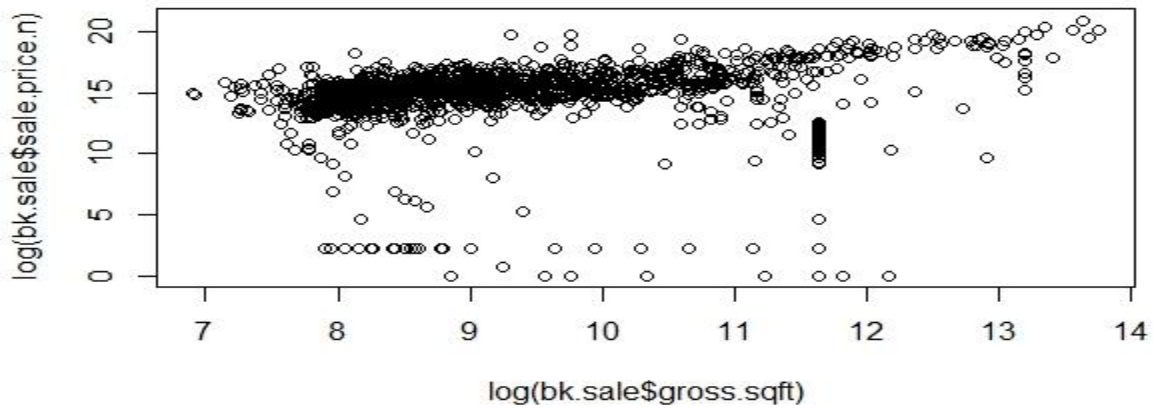

Figure 8 Price vs Area (in logarithmic scale)

The information that can be gathered from the previous plot is that majority of the houses being sold are in the range 512- 1500 sq. feet.

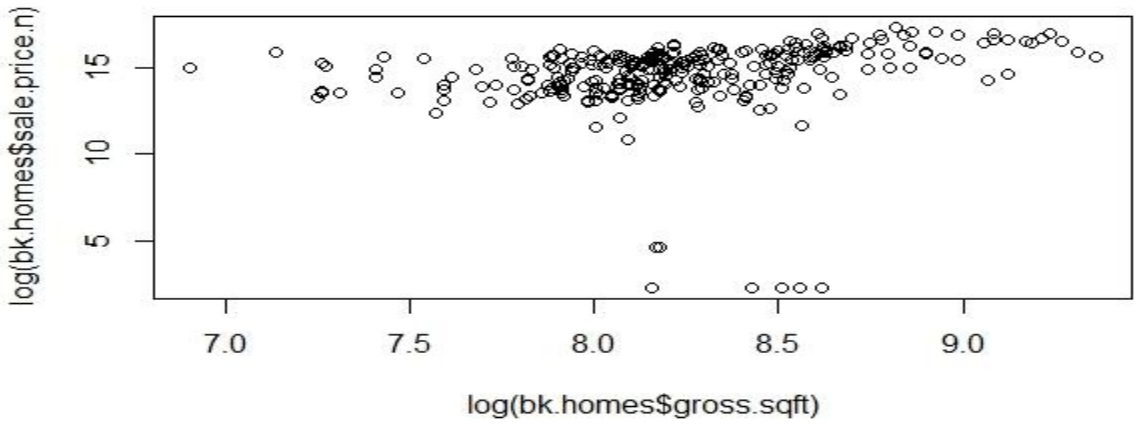

Figure 9 Price vs Area (only for family homes) in logarithmic scale

In this figure it can be observed that most of the family homes in Manhattan are around 256 sq. feet and sell around a price of $32000.

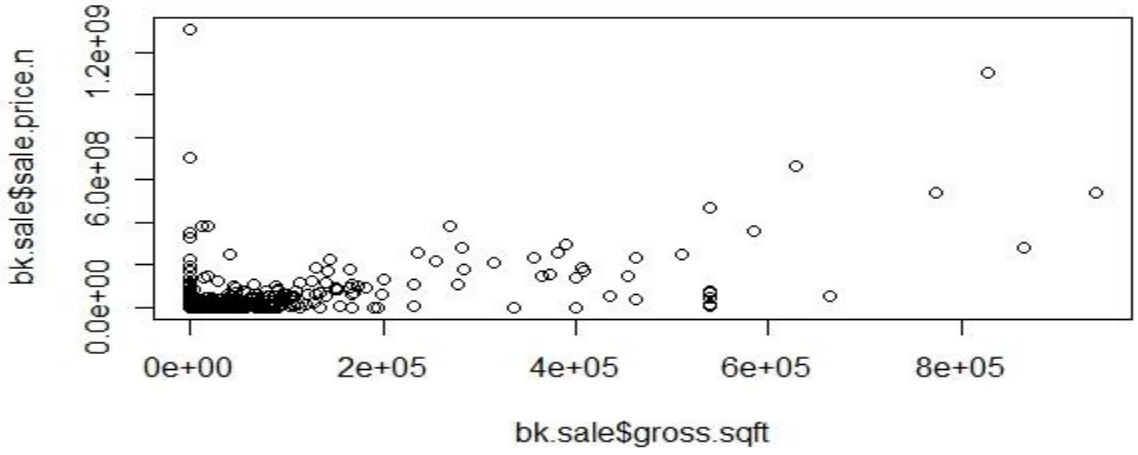

Figure 10 Gross vs Sale price

This graph portrays the outliers pretty well. We find that an outlier situated in the top left corner of the graph sold for a price disproportionate of the size. The points plotted on the lower right corner signifies that the price of these houses are way lesser than the average cost/unit area.

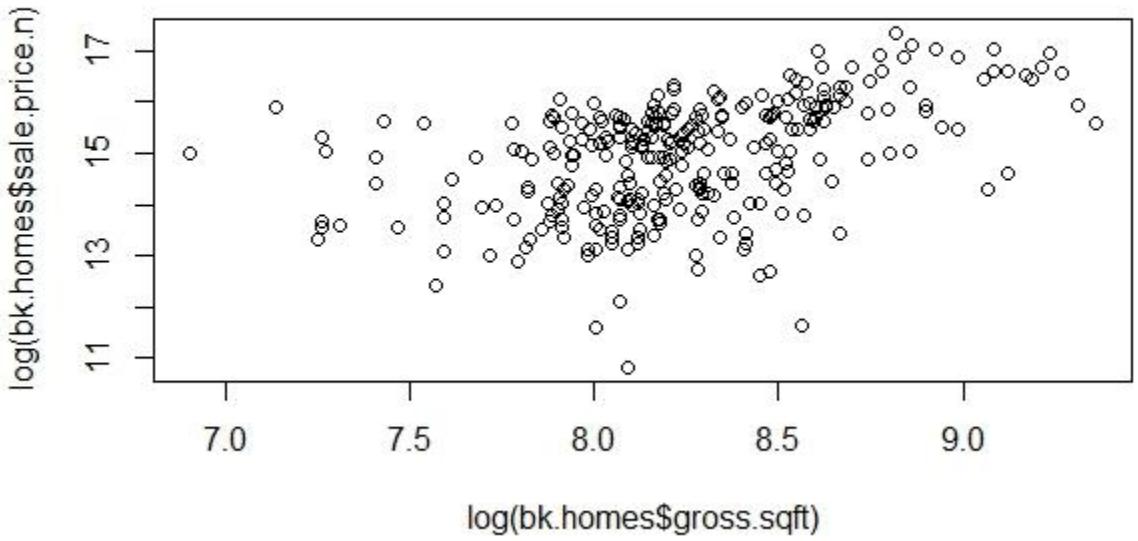

Figure 11 Gross vs Sales price outliers removed

This graph is similar to Figure 8 where the logarithmic value of the price is compared to the logarithmic value of the area. We observe here as well that houses around 512-1200 sq feet sell around $32000.

Being the "data scientist" often involves speaking to people who aren't also data scientists, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?

The engineers maintaining the sites. The information held by them could help us analyze from where most of the network traffic comes from. Secondly, talking to the sales representatives would enable us to understand the mentality of a buyer/seller.

Most of you are not "domain experts" in real estate or online businesses.

· Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insight into how you do it in your own field?

Yes, it does. The procedures followed in data procuring is also followed by software engineers who work in the area of Big Data analytics.

6. Flight data

6.1. *Dataset details*:

Name: 2007.csv

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

6.2. Experiments, plots and interpretations

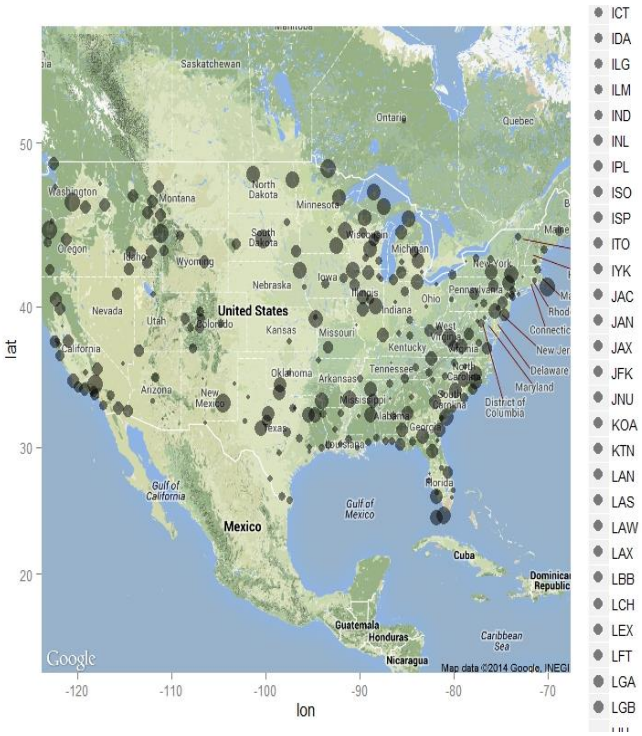

Figure 12 Air Traffic Density Plot

Figure 8 shows the plot of airports on the US mainland based on the number pf flights. This pictorial representation conveys to us the whereabouts of the airports which have high activity rate. Some of the airports with high activity rate are Los Angeles, Chicago, New York, etc. whereas some of the low activity airports are Albury, King Salmon, etc.

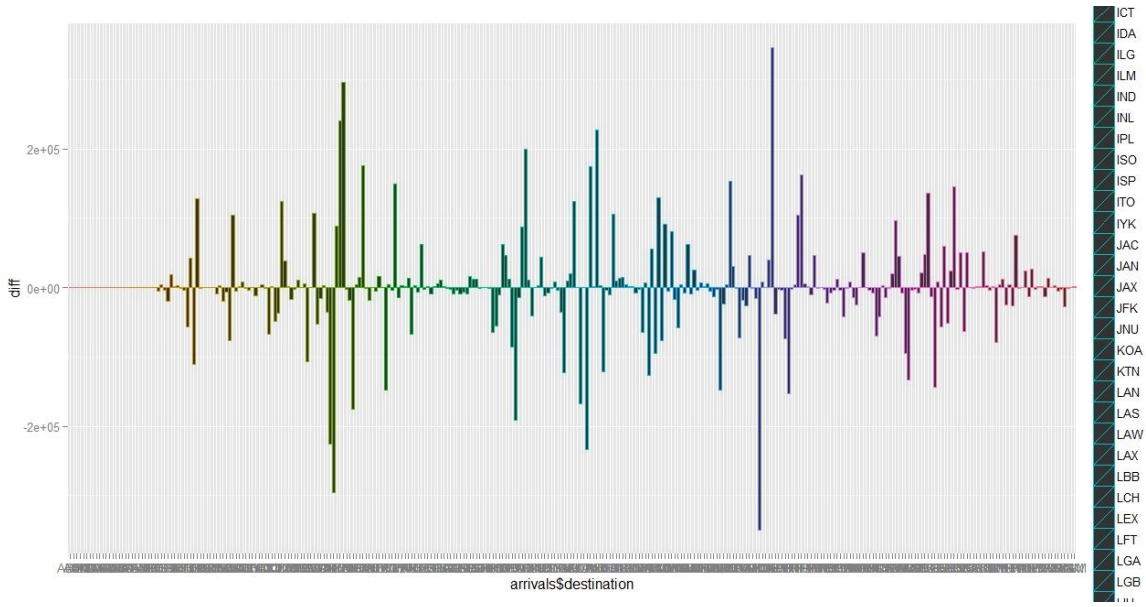

Figure 13 Difference between #Arrivals and #Departure

This graph depicts the difference between the number of arrivals to the number of departures. The bars on the negative side shows that these airports have high arrival dates and the airports that have bars on the positive side have high departure rates. This information is used to identify whether an airport is a hub/maintenance center.

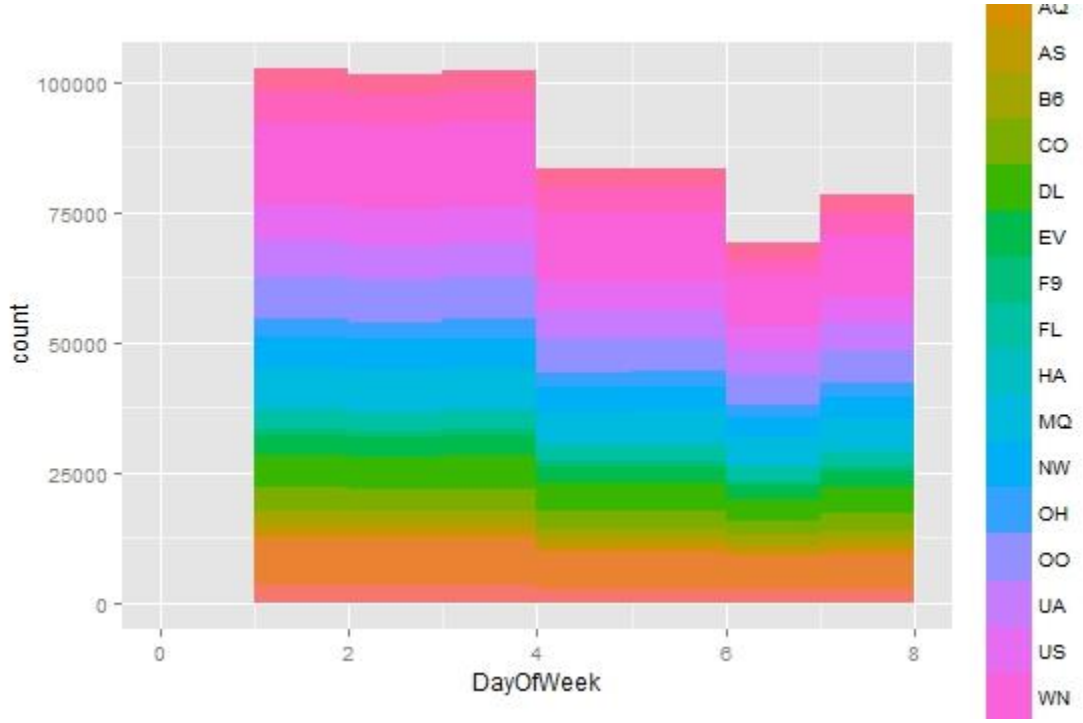

Figure 14 Services offered by airlines by count

This graph shows that the Southwestern Airlines(WN) has the highest number of flights that operate. The next is followed by US Airways and United Airlines. The least number of flights are operated by Aloha airlines.

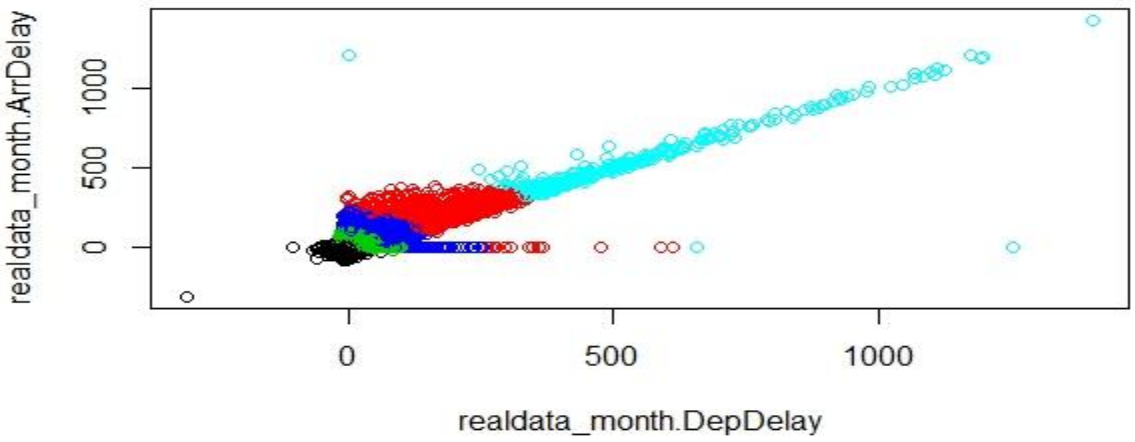

Figure 15 K-means plot of Arrival delay vs Departure delay

This graph shows the relationship between the Arrival delay to the Departure delay. The linear nature of this graph symbolizes that the greater the arrival delay, the greater the departure delay. This information is clustered into 5 groups.

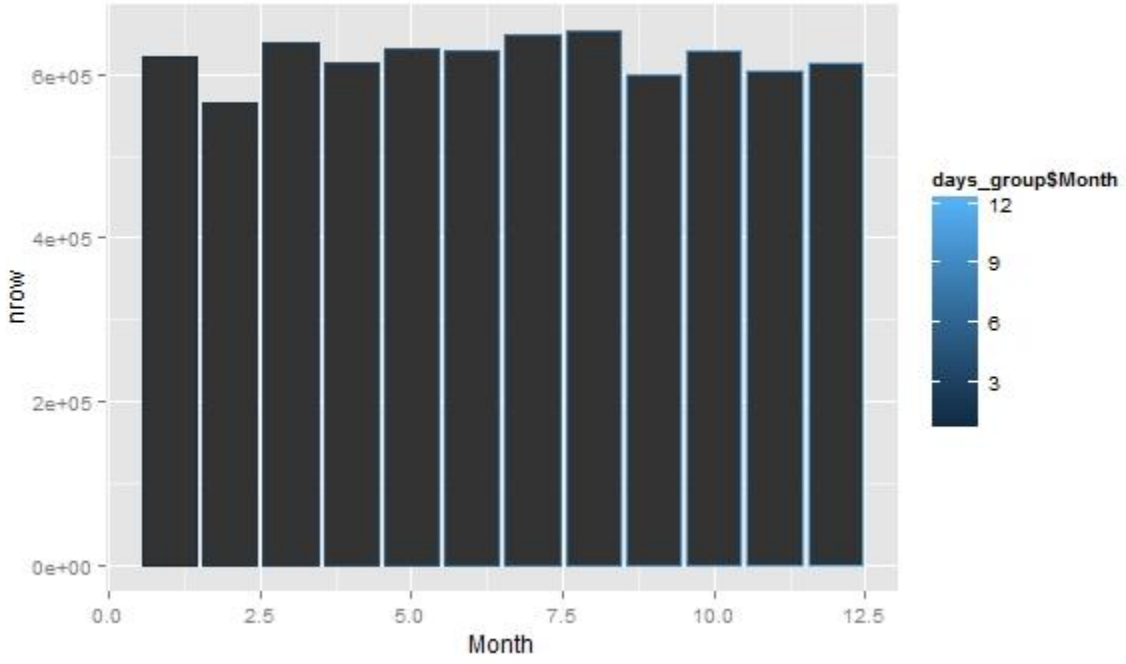

Figure 16 Total number of flights vs month

The graph show above describes the total number of flights operating in a month. We observe that the month of March has the highest number of operating flights whereas the month of February has the least number of operating flights. This information is used to identify holiday season/dry season thereby enabling airlines to offer discounts at appropriate time.

7. Lessons Learned

- Increase in comfort level of handling R.
- Ability to work with maps and different types of plots such as K-means, histogram and ggplots.

8. References

1. Page 39 in Doing Data Sciences Book for New York Times data analysis.
2. Page 49 in Doing Data Sciences Book for Real Direct data analysis.
3. Online R tutorials.
4. Help session document airport.R