DATA INTENSIVE COMPUTING

CSE 587

PROJECT 1

PROBLEM SOLVING & EXPLORATORY DATA ANALYSIS

ANKIT JAIN #50097432

MILKY SAHU #50096350

TABLE OF CONTENTS

Problem Solving and Exploratory Data Analysis Using R

Abstract

The purpose of this project is to familiarize us with options for exploring, screening and making adjustments to our data – in other words, methods for exploratory data analysis and graphics in R.

Plan to design and implement

For the first and second part of the question, we have first worked on the R solution provided on the data set provided in the book to learn the basic functionalities of R.
Then we have worked on an independent dataset (World Development Indicators Data) and used all the functions implemented in the first two parts.

During our implementation, we have worked on the following plots,

Single variable distributions Plots:

Histograms
Box-and-whisker plots
Normal quantile-quantile plots.

Relationships between pairs of variables

Scatterplots
Line Graphs
Bar Plots
Pie Charts
Maps

Uniqueness and Creativity

We have worked on a large number of "proof of concept" visualizations that illustrated the power of R for compiling and analyzing disparate datasets. Then we have analyzed different indicators among each other to come out to a meaningful interpretation.
Ex, Comparison of trends of growth of Mobile Users and Internet Users for United States over a period of 10 years.

Project Objectives

Problem Solving and Exploratory Data Analysis using R will meet the following objectives:

- Most common methods for data exploration, screening and adjustment in R.
- Know our data set and, in particular, screen for irregularities (e.g., outliers) and transform and/or standardize data if appropriate.
- Statistical and graphical summaries, including extremely advanced methods for graphical display of data.

Project Approach

We have used RStudio for our analysis of three main Datasets for our Project:

- **New York Times Dataset:**
  There are 31 datasets named nyt1.csv, nyt2.csv,..., nyt31.csv, which can be found here: https://github.com/oreillymedia/doing_data_science.
  Each one represents one (simulated) days' worth of ads shown and clicks recorded on the New York Times home page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged in.

- **Real Direct Dataset** –
  We have performed statistical analysis via plots and charts and infer useful information from the different tables, for e.g. Staten Island.

- **Analysis of World Development Indicators Data by UN –**
  We did the analysis on whole data, then we analyzed this data to reach a conclusion"
  1) We used the data of 10 countries for one year say 2011 and find out which country is leading in the development front. (India, United States, China, Canada, Sweden, Kuwait, Namibia, Saudi Arabia', 'Singapore', 'Australia)
     Indicator: Agricultural Land (% of land Area)
  2) Next, to analyze the data clearly, we have analyzed the data for one country i.e. US for year range 2001-2011 to comment on its development record for various indicators like GDP Growth, Agricultural Land available, Internet and Mobile Users. We have chosen United States for our analysis.
  3) Finally, we used the maximum value data for 10 countries and the corresponding year in which they had the maximum value.

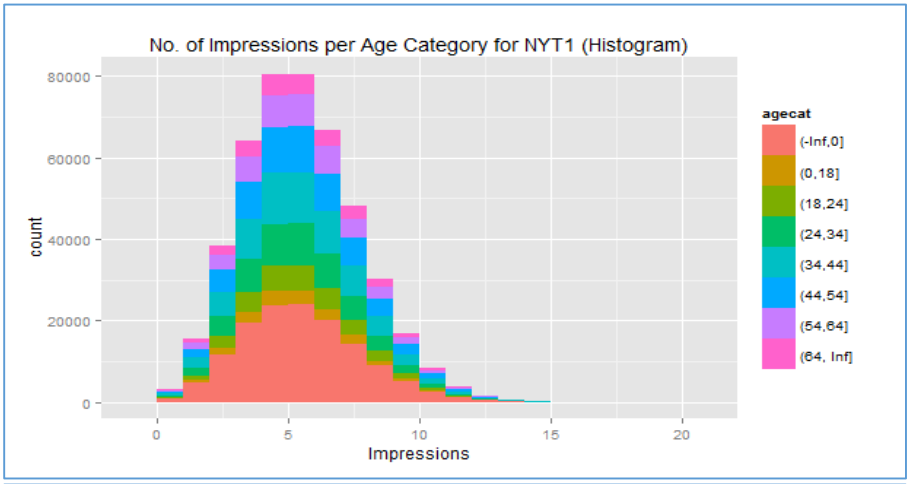
New York Times Data Set

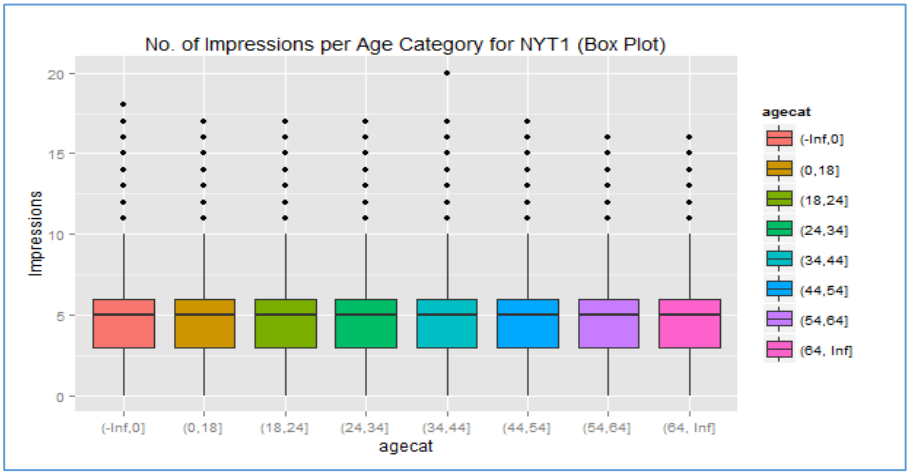
- Data Source: http://stat.columbia.edu/~rachel/datasets/nyt(x).csv
- Questions and Outcomes:

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

```
## ---- Categorizing data on the basis of age categories ----
## ============================================================
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
data2$agecat <-cut(data2$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
data3$agecat <-cut(data3$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
data4$agecat <-cut(data4$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
data5$agecat <-cut(data5$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
```

2. **i) For a single day:**
   Plot the distributions of number of impressions and click through-rate for these six age categories. (CTR=# clicks/# impressions)





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

```
data1$CTRatio <- data1$Clicks/data1$Impressions
data2$CTRatio <- data2$Clicks/data2$Impressions
data3$CTRatio <- data3$Clicks/data3$Impressions
data4$CTRatio <- data4$Clicks/data4$Impressions
data5$CTRatio <- data5$Clicks/data5$Impressions
```

iii) **Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, 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.**

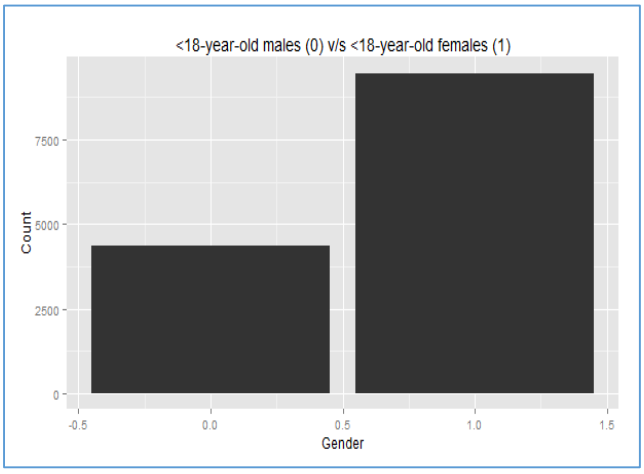
```
## calculating quantiles
## ============================================================
M1 <-length(data1$CTR[(data1$CTR > quantile(data1$CTR,0.95))])
M2 <-length(data2$CTR[(data2$CTR > quantile(data2$CTR,0.95))])
M3 <-length(data3$CTR[(data3$CTR > quantile(data3$CTR,0.95))])
M4 <-length(data4$CTR[(data4$CTR > quantile(data4$CTR,0.95))])
M5 <-length(data5$CTR[(data5$CTR > quantile(data5$CTR,0.95))])
M6 <-length(data6$CTR[(data6$CTR > quantile(data6$CTR,0.95))])
```

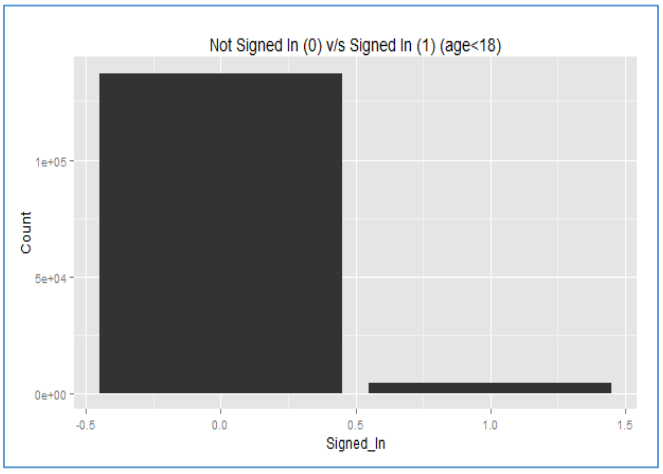
```
## defining siterange function to calculate length,sum,min,mean,max
## ============================================================
siterange <- function(x){c(length(x),sum(x),min(x),mean(x),max(x))}
```

```
## Calculating Clicks/Impressions for every data frame
## ================================================
data1$CTR <- data1$Clicks/data1$Impressions
data2$CTR <- data2$Clicks/data2$Impressions
data3$CTR <- data3$Clicks/data3$Impressions
```

Important to track over time: Click through Ratio (CTR)
Click through ratio will compress the data but will still capture user behavior.

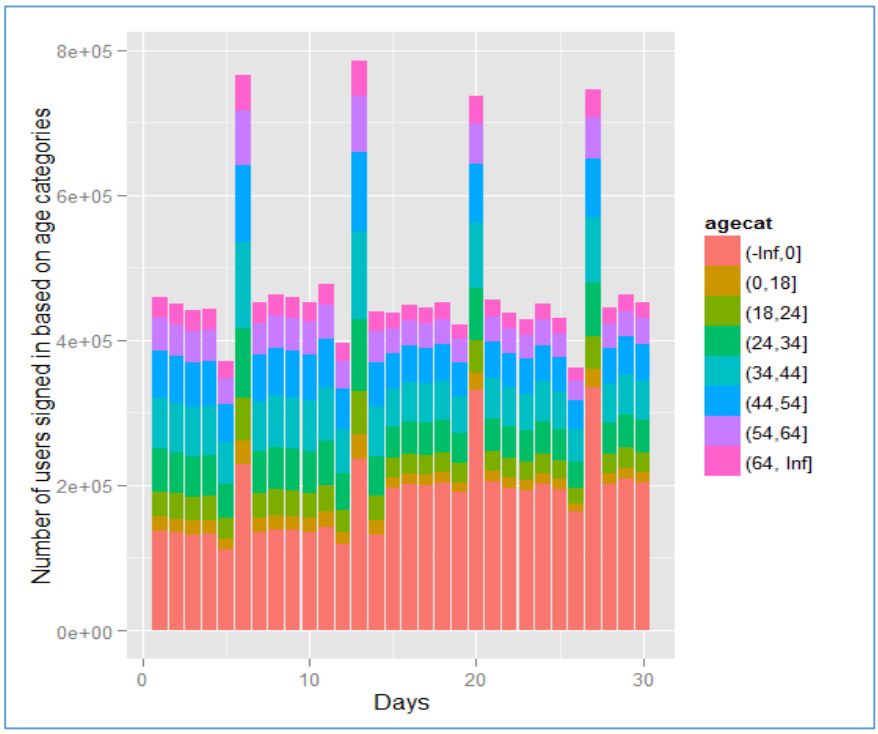
i) Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example).

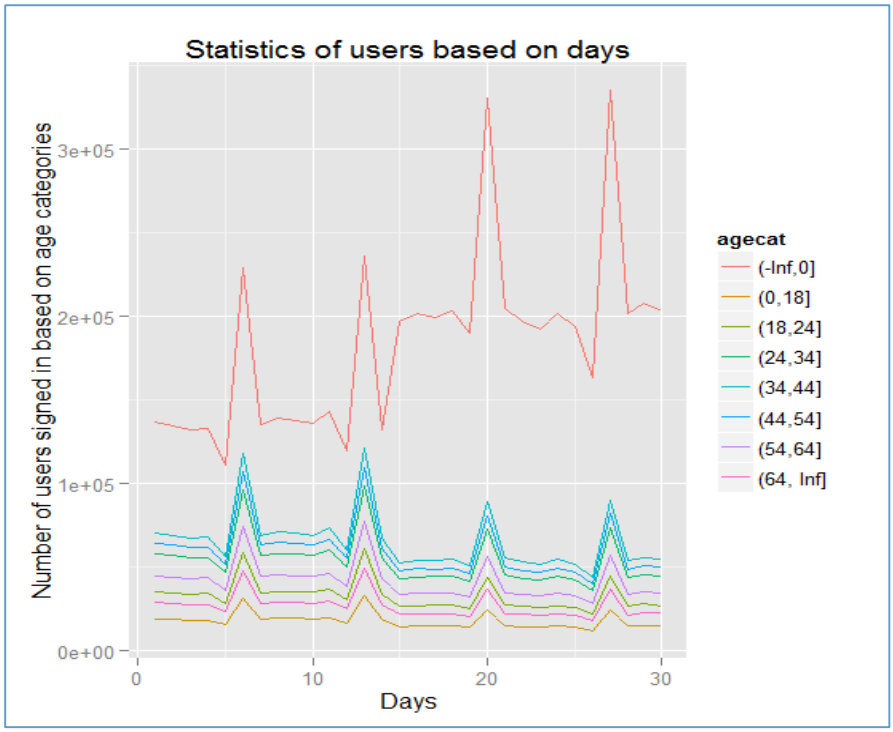




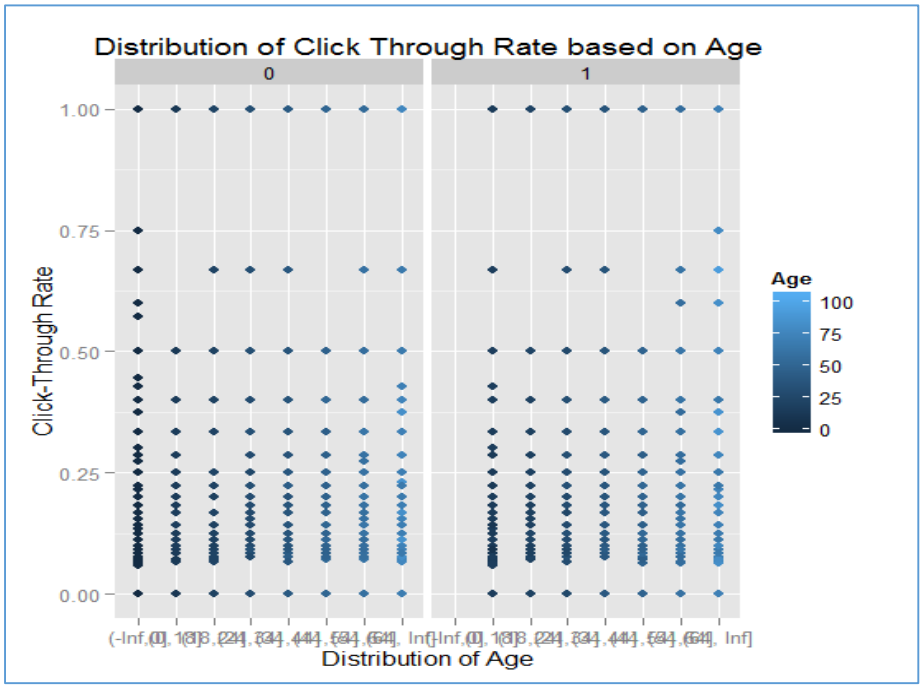
The number of females, aged less than 18, who logged in were greater than the number of males who logged in and aged less than 18. There was a big number of people who visited the website but did not sign in.

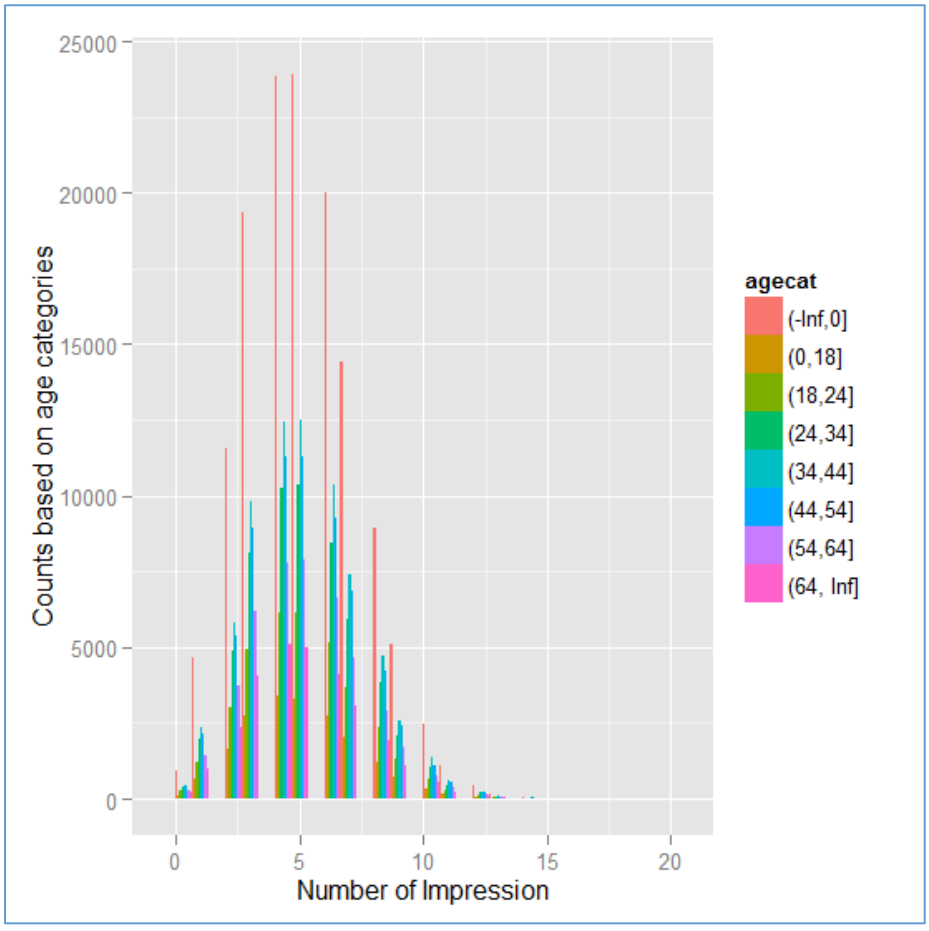
3. **Now extend your analysis across days. Visualize some metrics and distributions over time. Plot for 30 days for number of users signed in based on age categories**

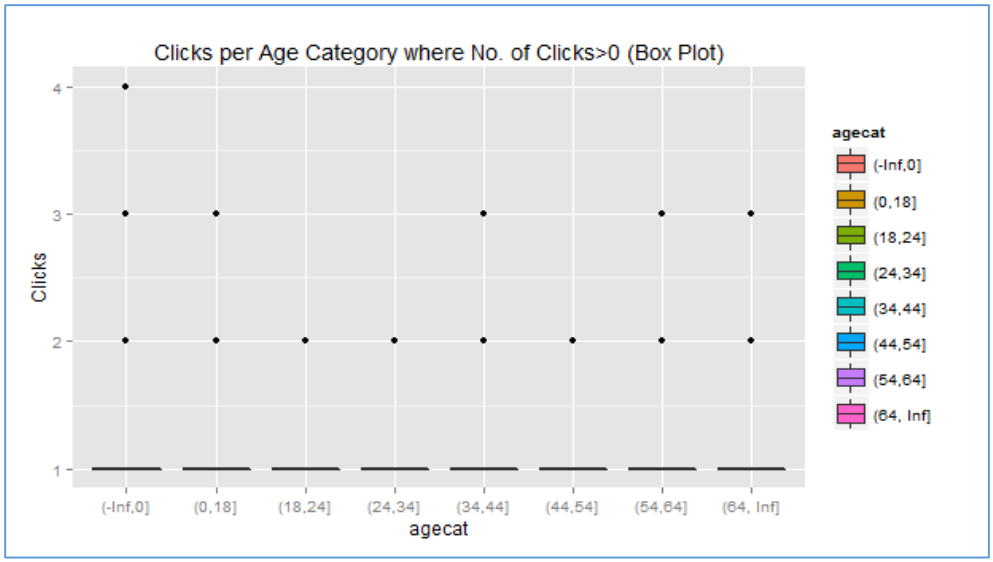


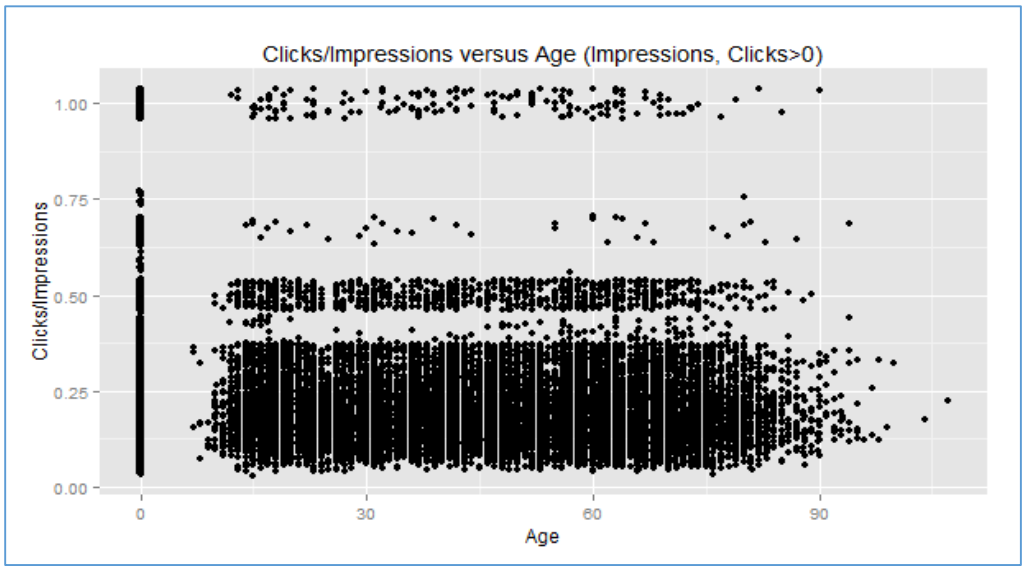


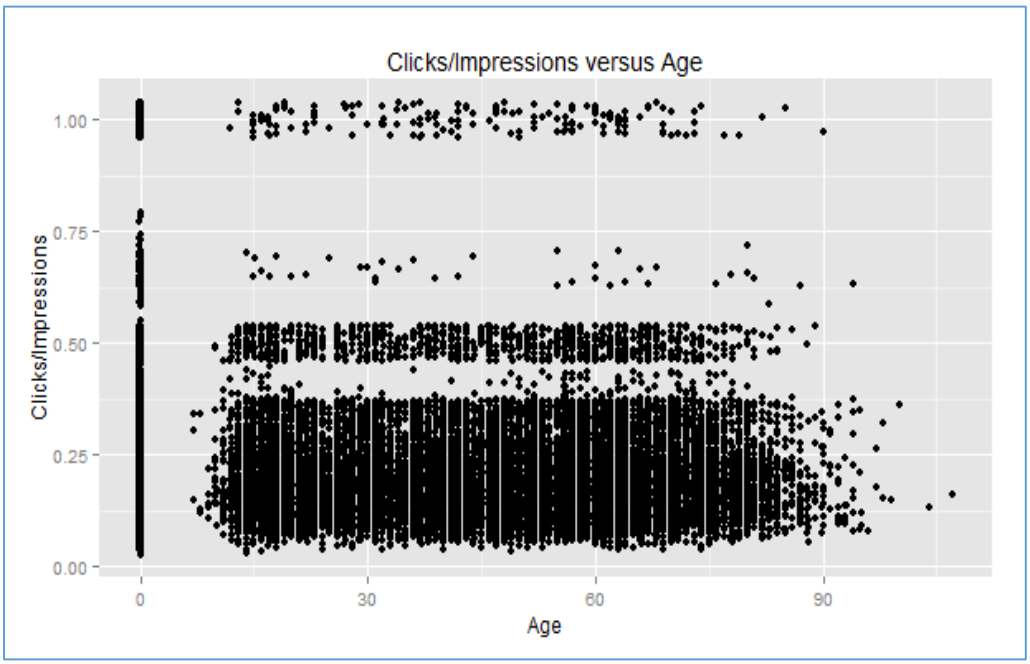
There is a huge number of users who did not fill in their age as the age group [-Inf, 0] has the highest number of signed in users.

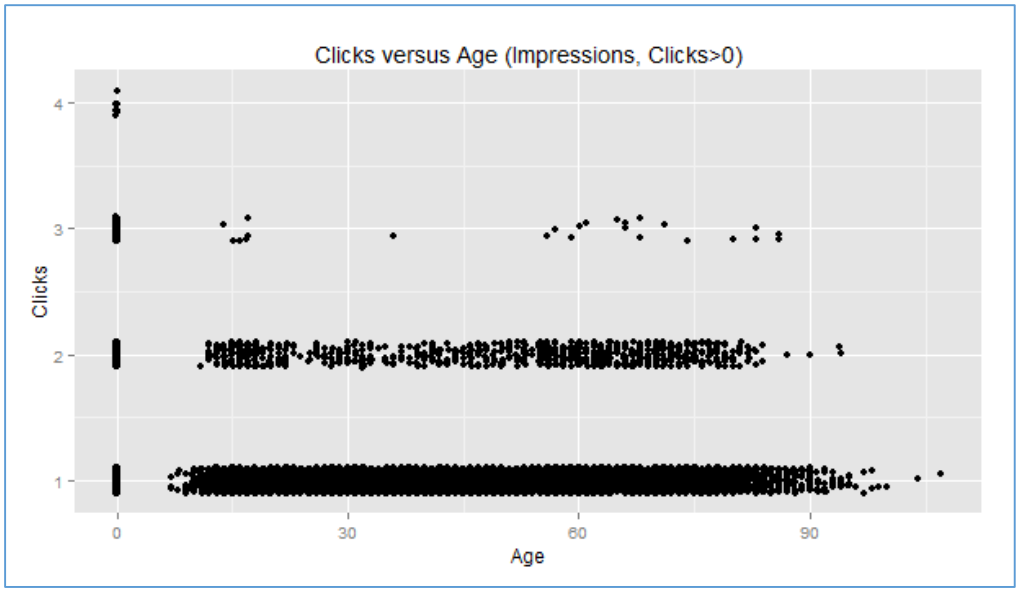

Distribution of Click Through Rate based on Age
0
1
1.00
0.75
0.50
0.25
0.00
Click-Through Rate
Age
100
75
50
25
0
Distribution of Age

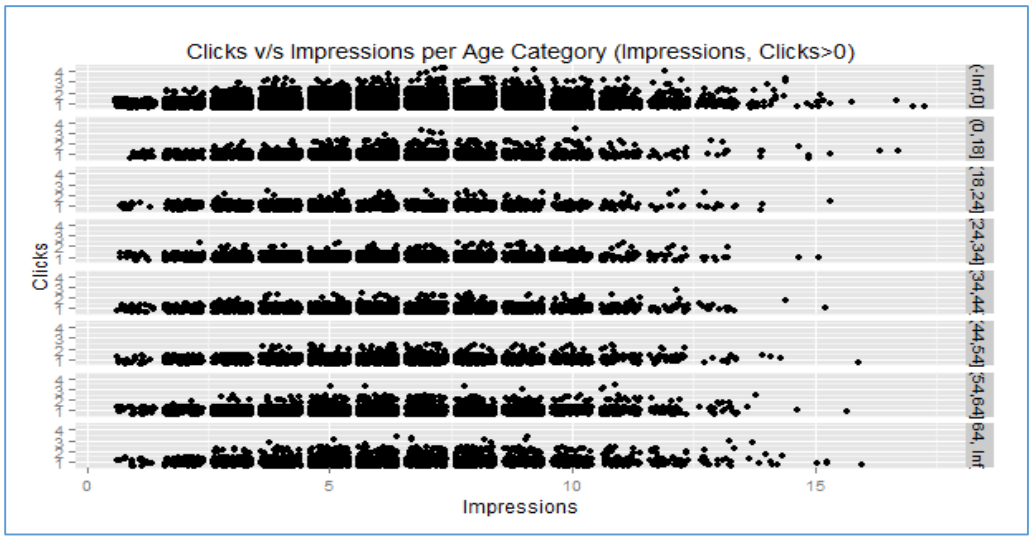

25000
20000
15000
10000
5000
0
Counts based on age categories
0
5
10
15
20
Number of Impression
agecat
(-Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64, Inf]

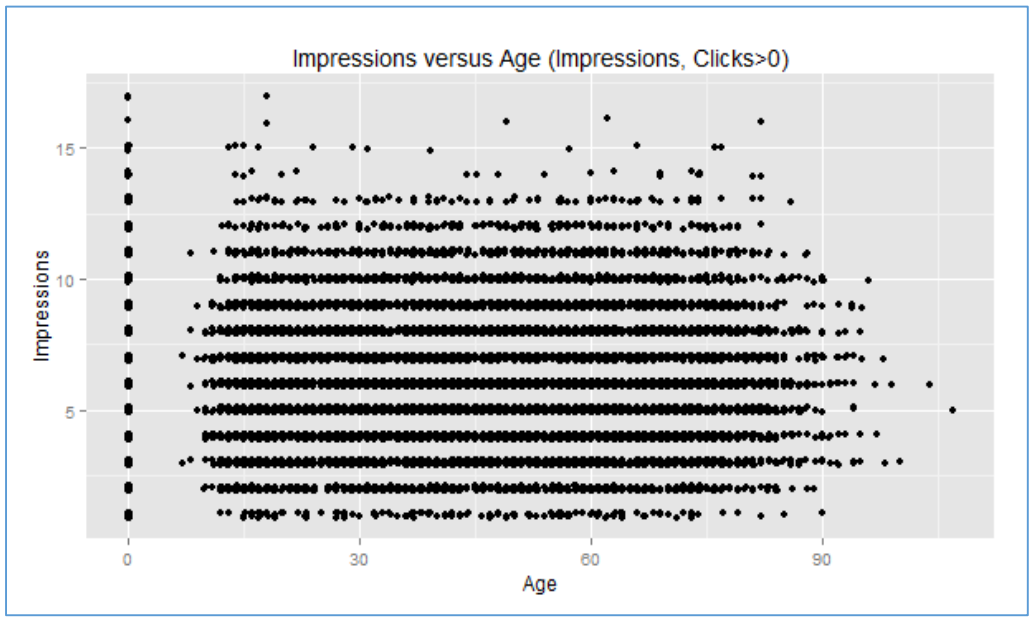

Clicks per Age Category where No. of Clicks>0 (Box Plot)
Clicks
agecat
(-Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64, Inf]

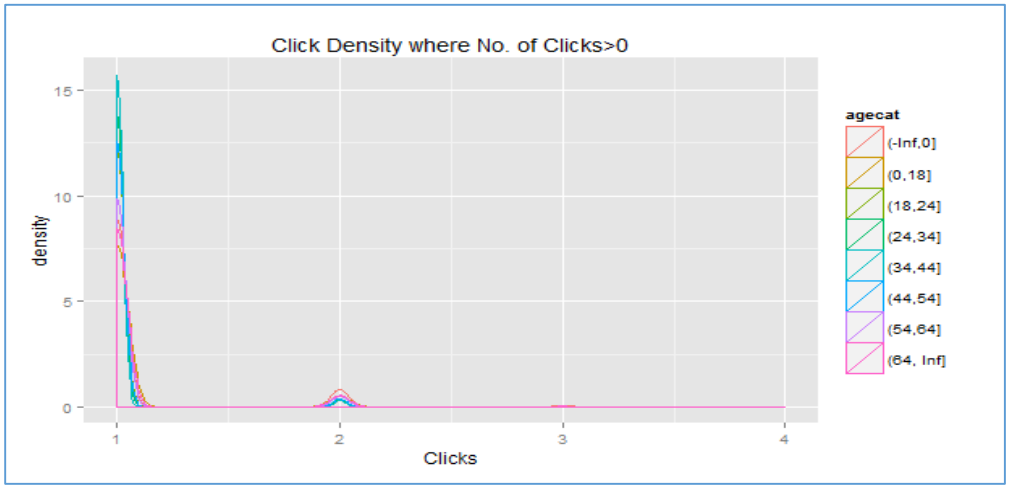

Clicks/Impressions versus Age (Impressions, Clicks>0)
Clicks/Impressions
Age


Clicks/Impressions versus Age
Clicks/Impressions
Age

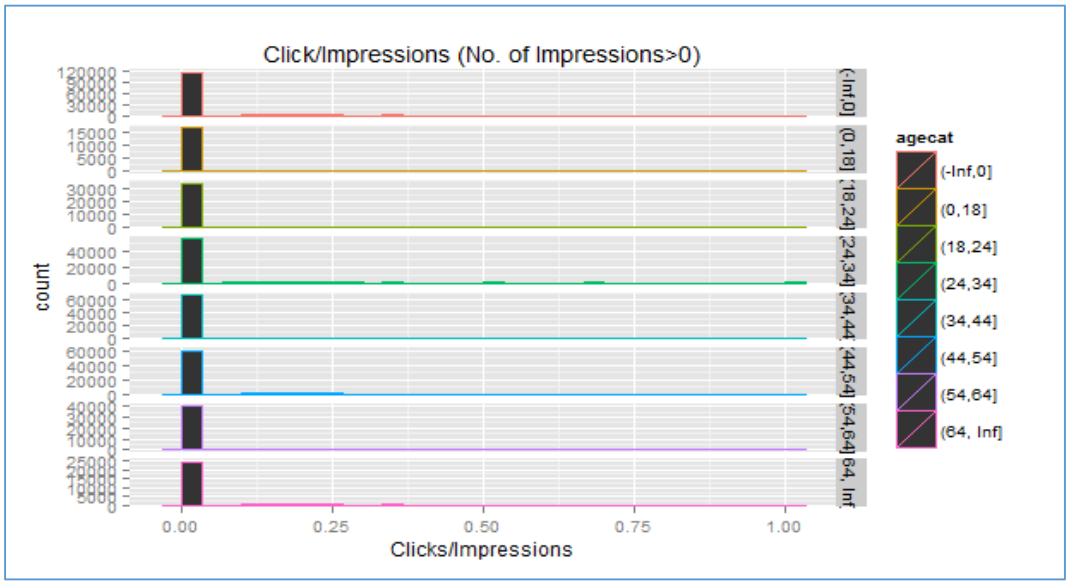

Clicks versus Age (Impressions, Clicks>0)
Clicks
Age
0
30
60
90
1
2
3
4

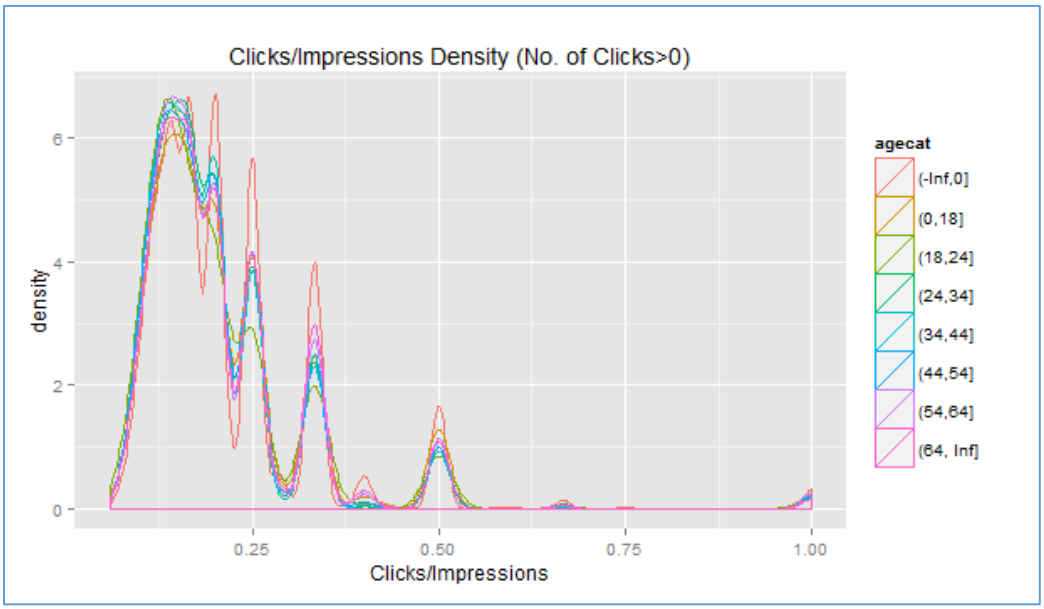

Clicks v/s Impressions per Age Category (Impressions, Clicks>0)
Clicks
Impressions
0
5
10
15
(-Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
64, Inf

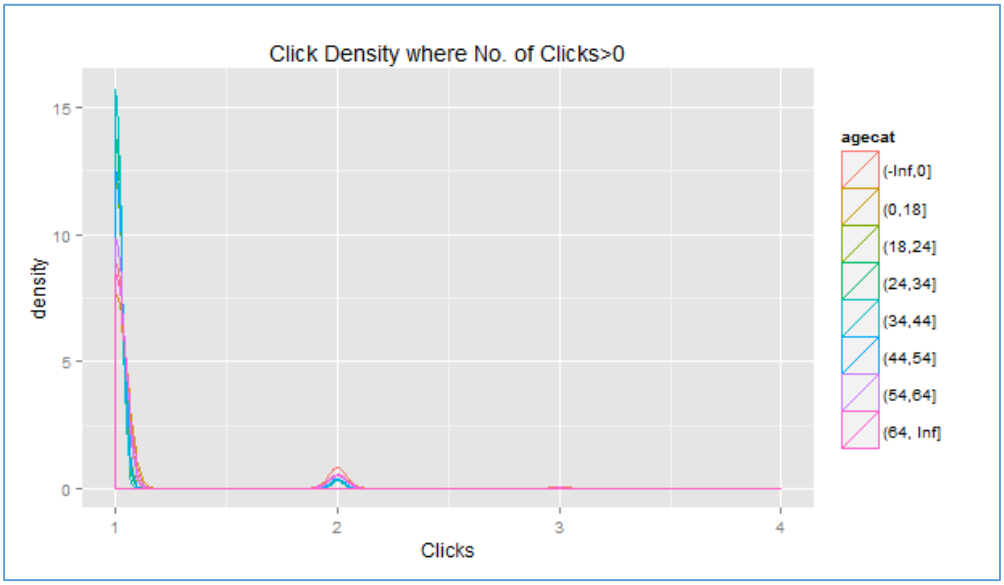

Impressions versus Age (Impressions, Clicks>0)
Impressions
Age
0
30
60
90
5
10
15

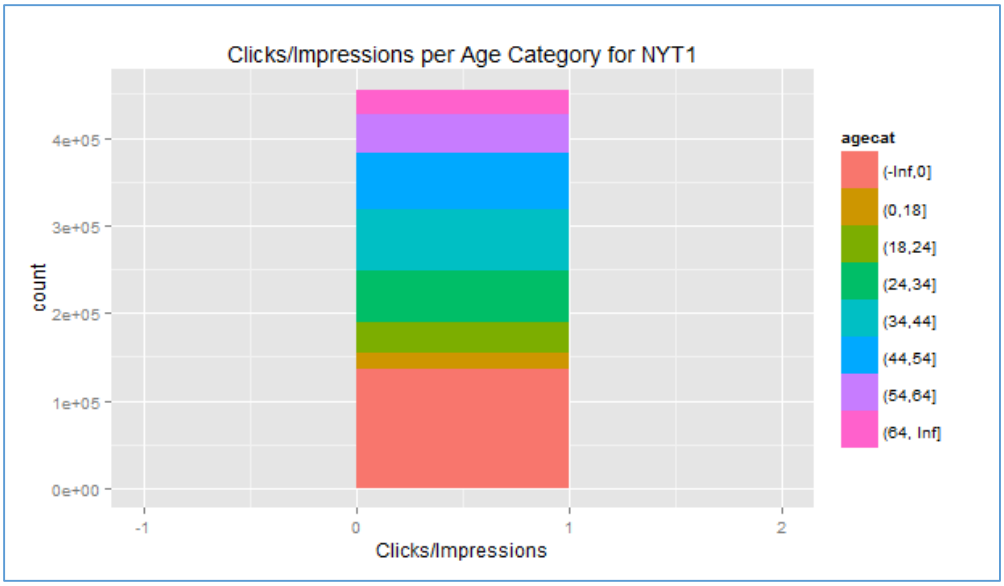


This plot shows the density of Clicks over the different Age Categories, where No. of Clicks>0









4. Describe and interpret any patterns you find.

In the above plots we see a number of observations which are listed as follows:

- The largest number of users do not sign in to the website and hence their information is not available with the website.
- Most of the users only open the website pages but do not click on the links. The pages might have been opened because of the advertisements also. Most of the users are not interested in viewing the ads or some pages of the website.
- The age groups [0,18] and [64, Inf] have the least number of users. As a matter of fact, these users either do not have access to internet or are not used to it.
- Clicks/Impressions ratio is very low in all the age groups.
- Several pages have zero number of clicks and zero number of impressions as well.
- The website has content which does not please a wide range of users.

RealDirect.com Data Set

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

   The report aims at finding the rates and explaining the types of houses of different cities and which are available at realdirect.com. The combined study of the cities will give a bigger picture and avail the users to make a clear cut comparison between the features of the houses of different cities. The study will show a better place to invest money and move away from the areas where there is no or very less profit. This study of this project leads us to involve the CEO of Real Direct to analyze the findings which will give him an overall scenario of the situation in the market for Real States.

2. **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?**

   A data scientist is a job title for an employee or business intelligence (BI) consultant who excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge. She deals with big data which is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, duration, and storage. This Big Data is a collection of the information of millions of users being collected every day. It is very easy to explain laymen what big data is because it is they who have created this enormous data.

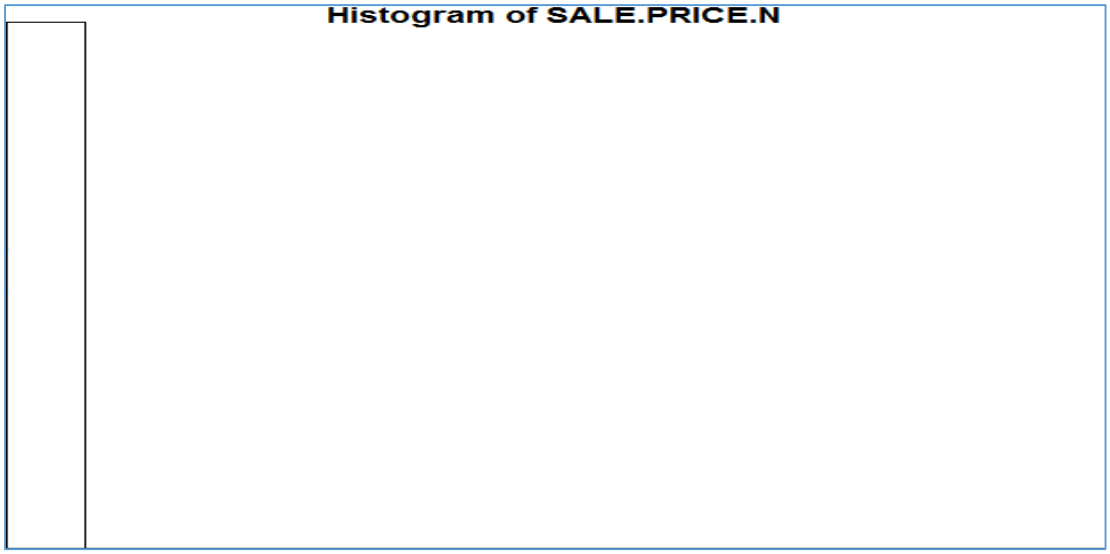
3. **Most of you are not "domain experts" in real estate or online businesses. Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give your insight into how you do it in your own field? Sometimes "domain experts" have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn't understand ("comps," "open houses," "CPC")? Sometimes if you don't understand vocabulary that an expert is using, it can prevent you from understanding the problem. It's good to get in the habit of asking questions because eventually you will get to something you do understand. This involves persistence and is a habit to cultivate**

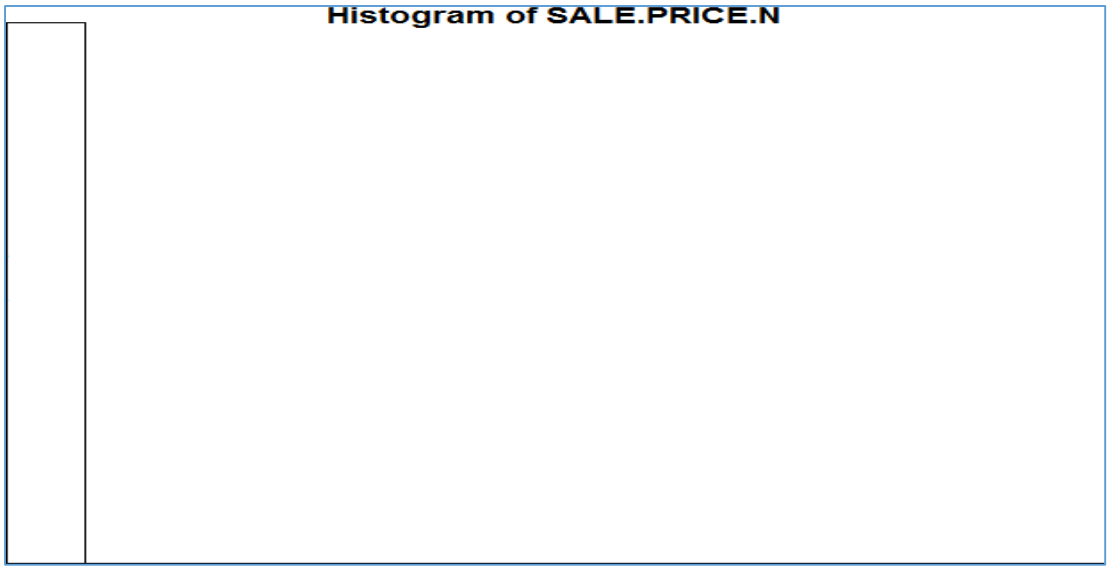
   Stepping out of one's domain and exploring new avenues should be a matter of interest for anyone who takes it as challenge and wants to give it her best. Working in new settings can be a lucrative experience and help one broaden her domains.

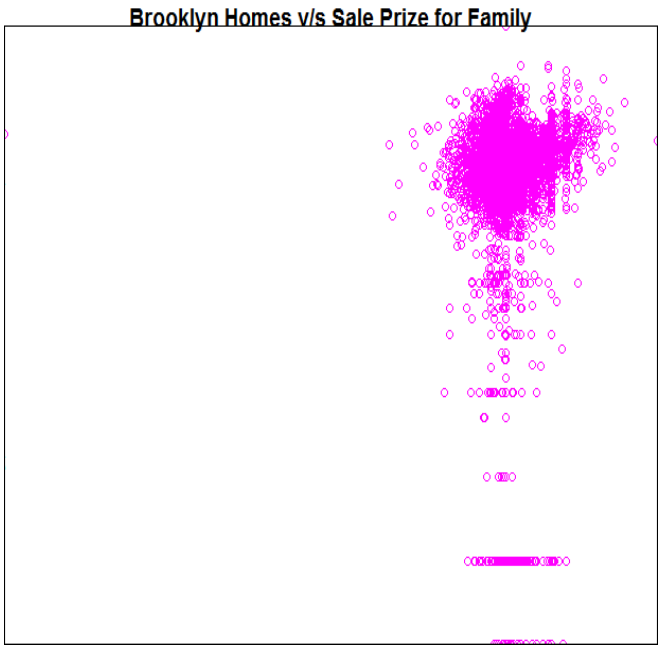
4. **Doug mentioned the company didn't necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain**

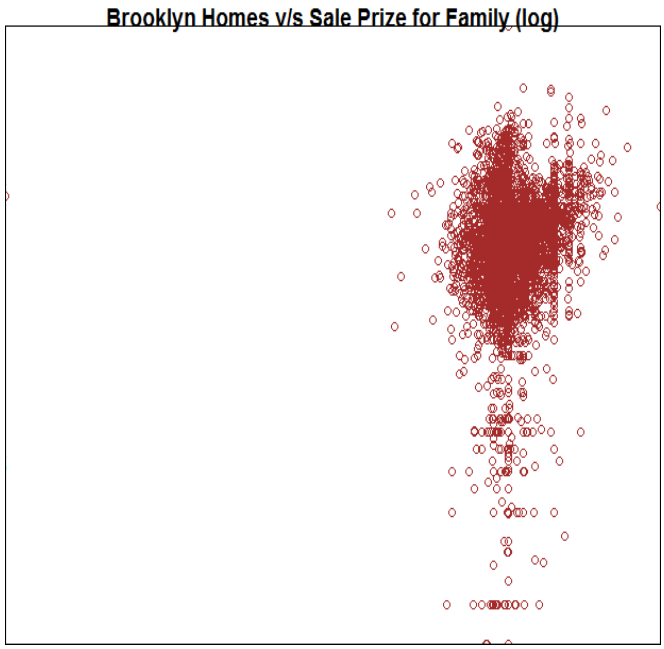
   The data should be rich with information. It should cover as many aspects as possible for a comprehensive analysis. It should be easily readable and easy to use. It should not be condensed over 1 -2 big large file but distributed over small datasets for security and portability.

Analysis of Brooklyn Homes

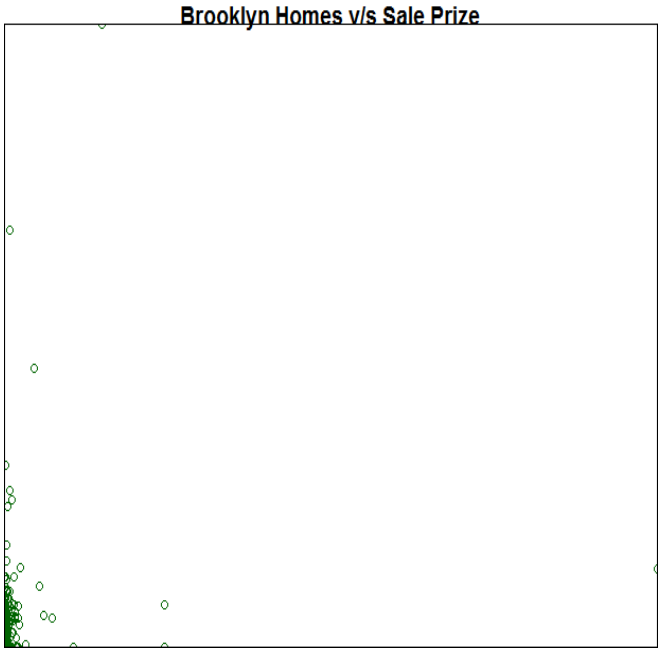


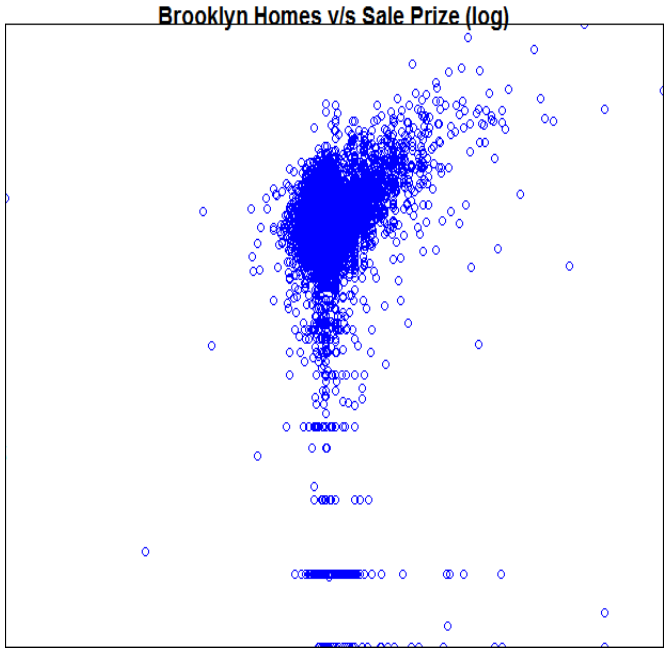






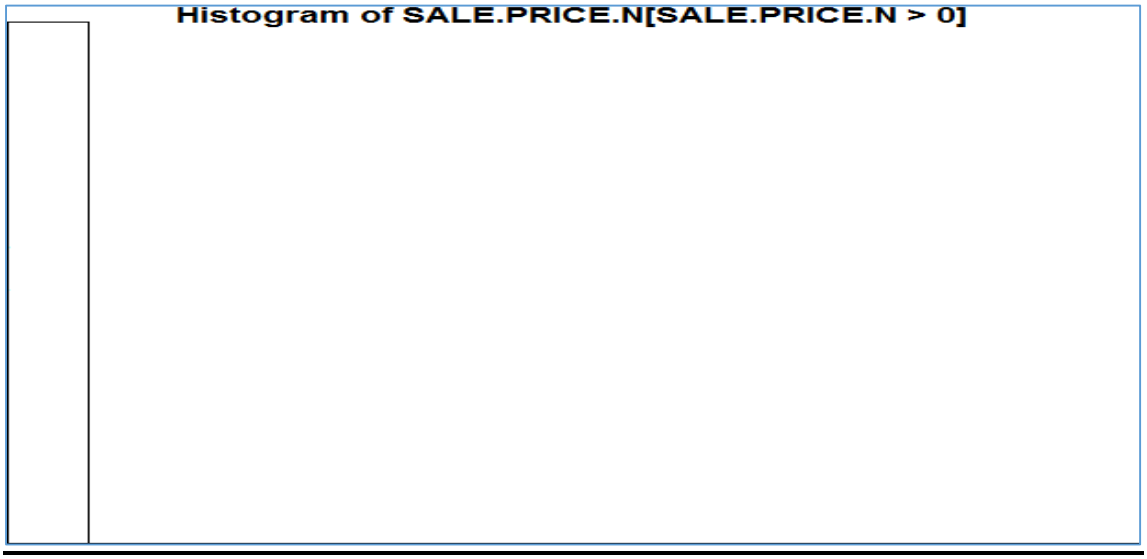
Brooklyn Homes v/s Sale Prize for Family

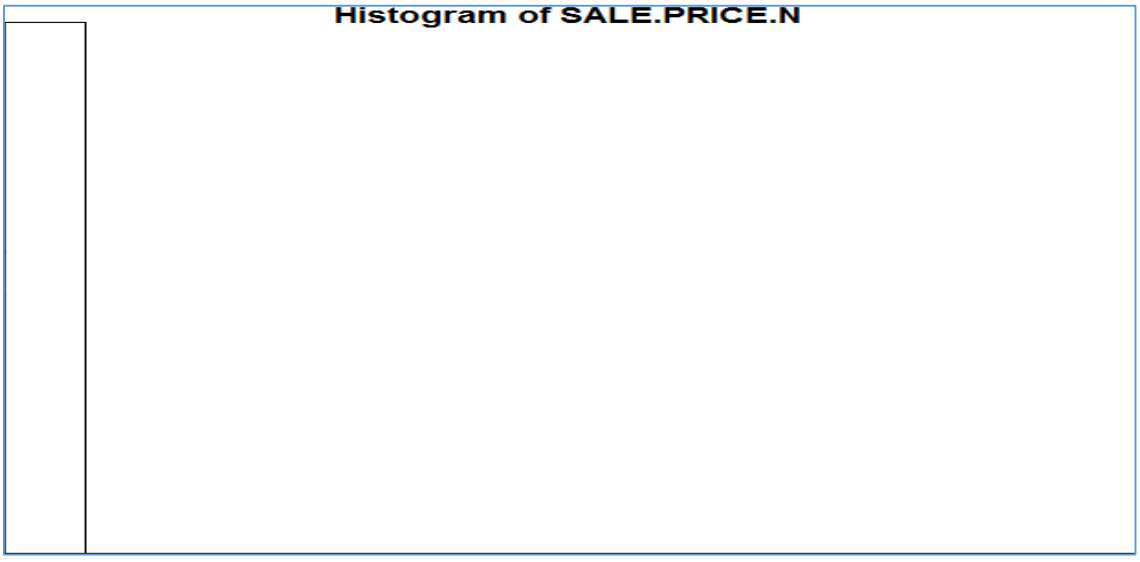


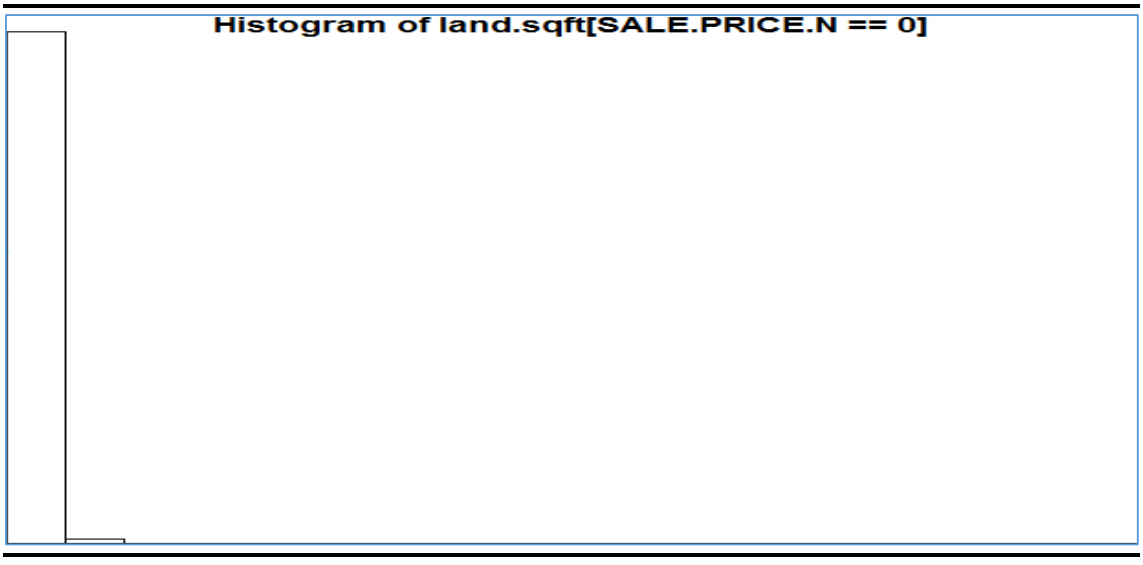


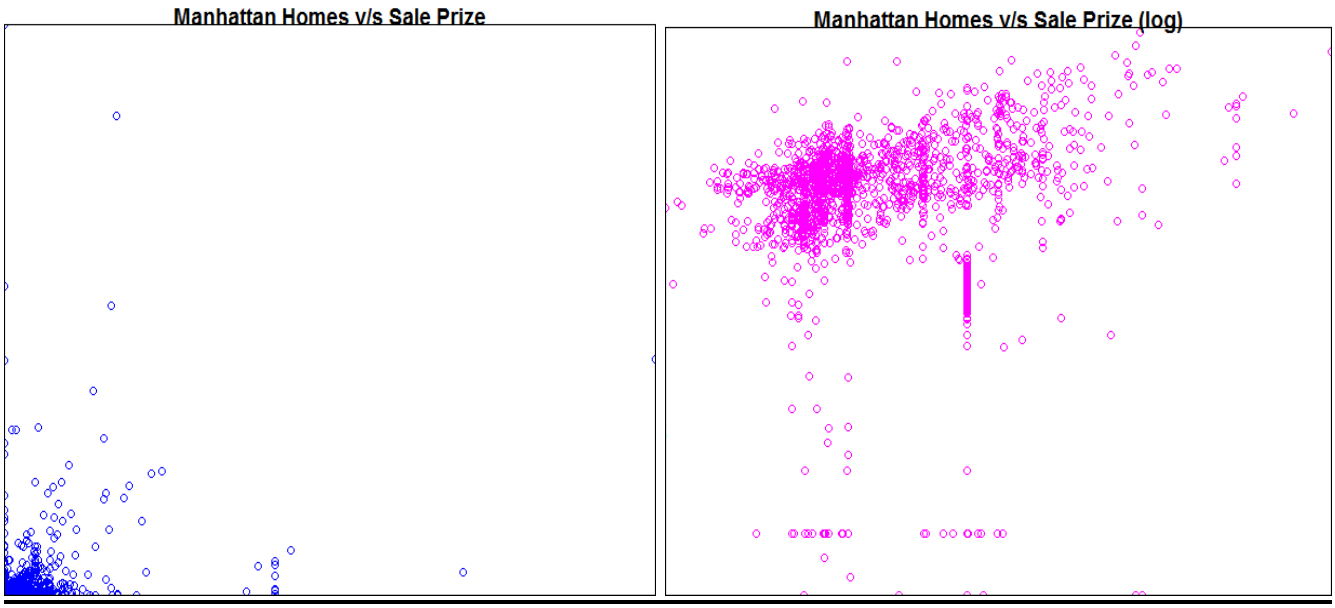
Brooklyn Homes v/s Sale Prize for Individual

Analysis of Manhattan Homes

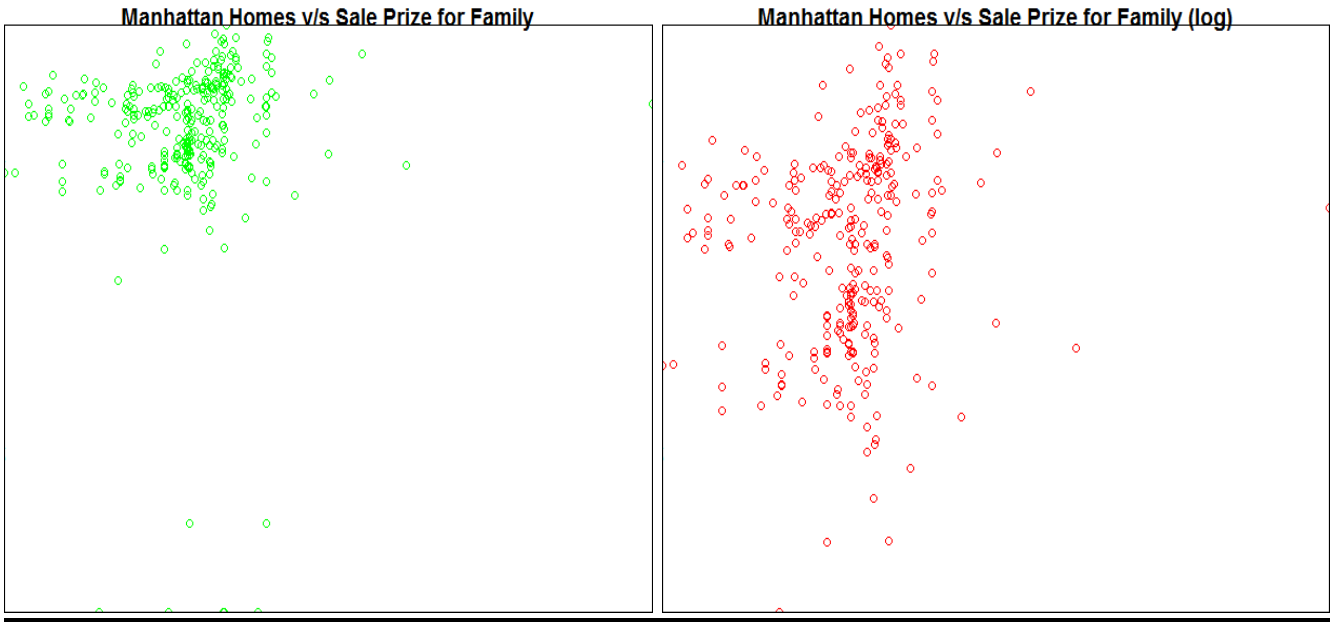






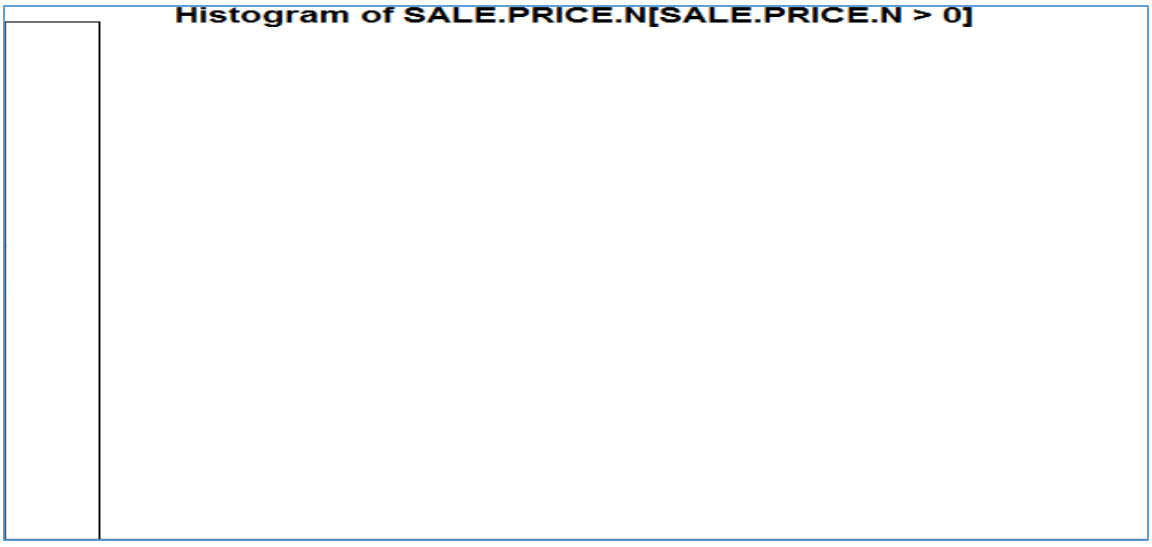


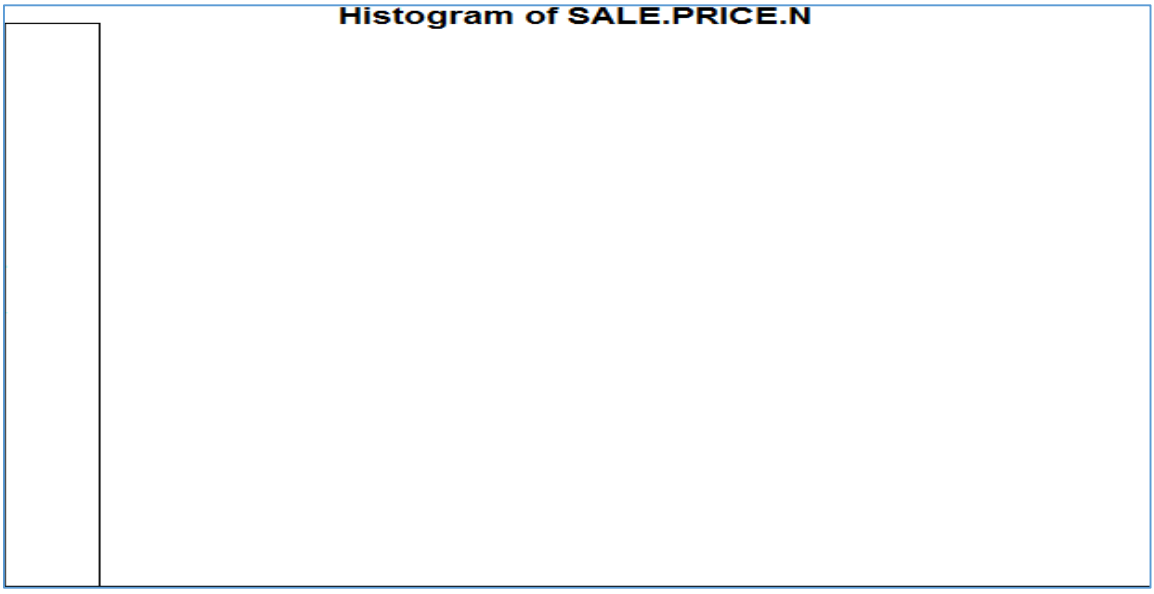
Manhattan Homes v/s Sale Prize for Family

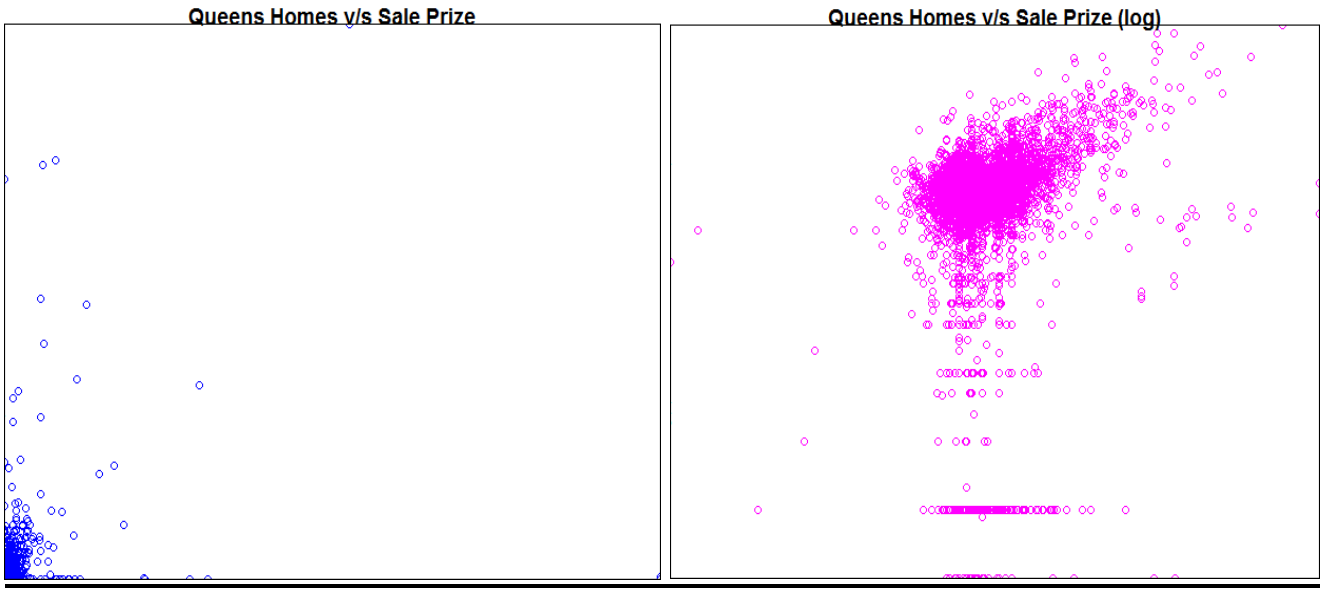


Manhattan Homes v/s Sale Prize for Individual

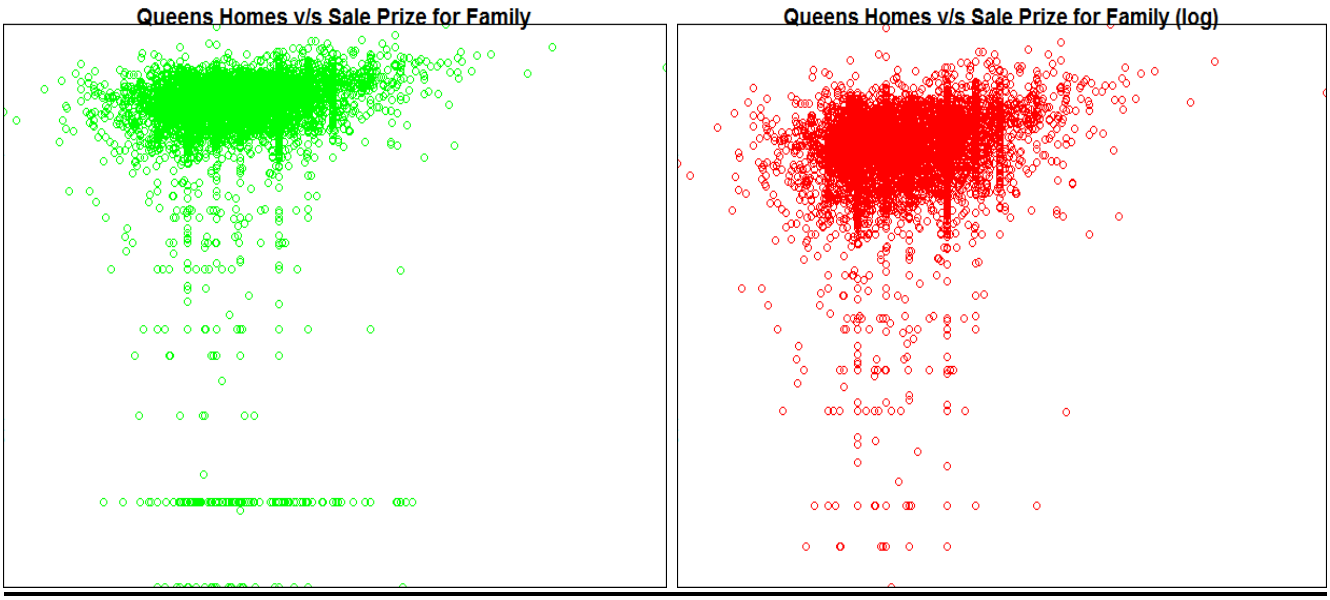
Analysis of Queens Homes





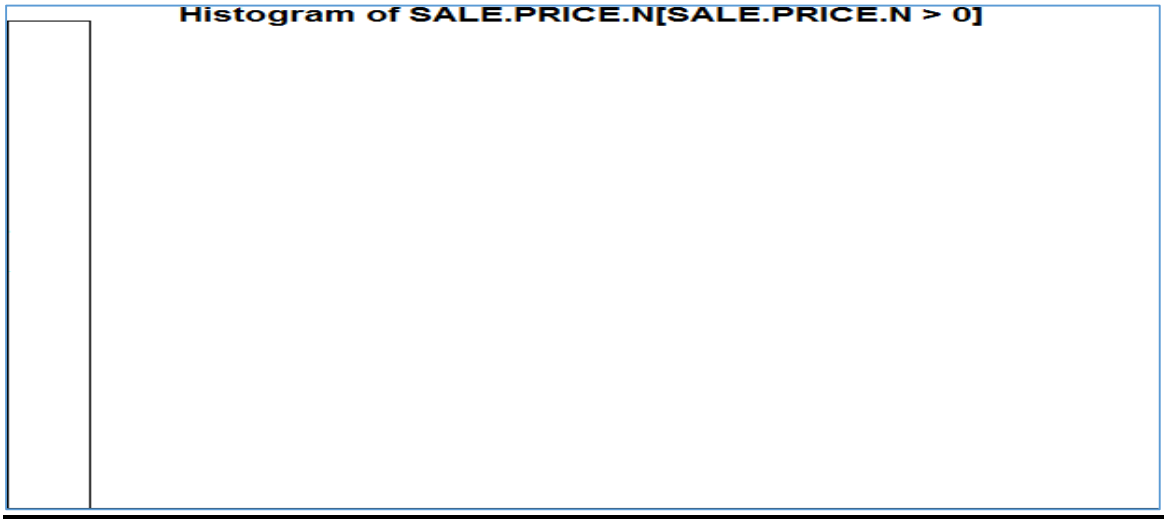


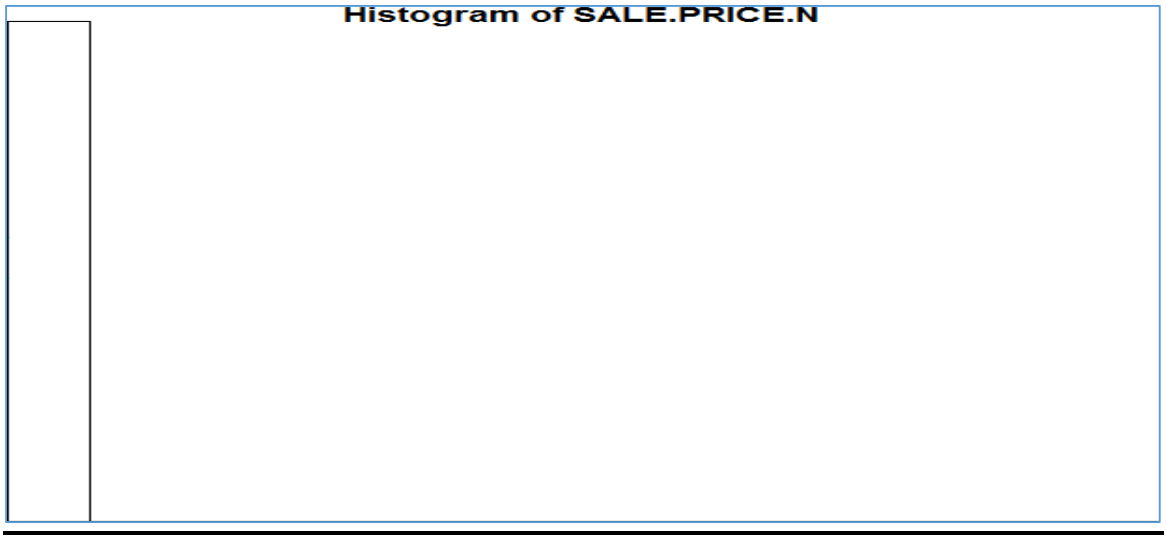
Queens Homes v/s Sale Prize for Individual

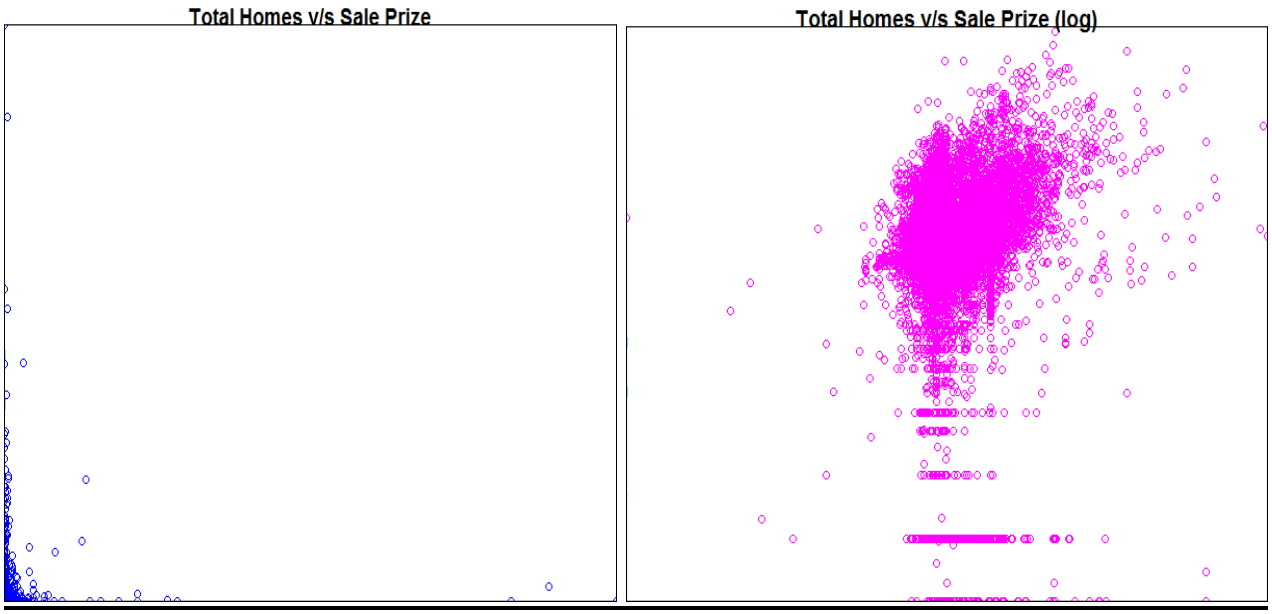


Queens Homes v/s Sale Prize for Family

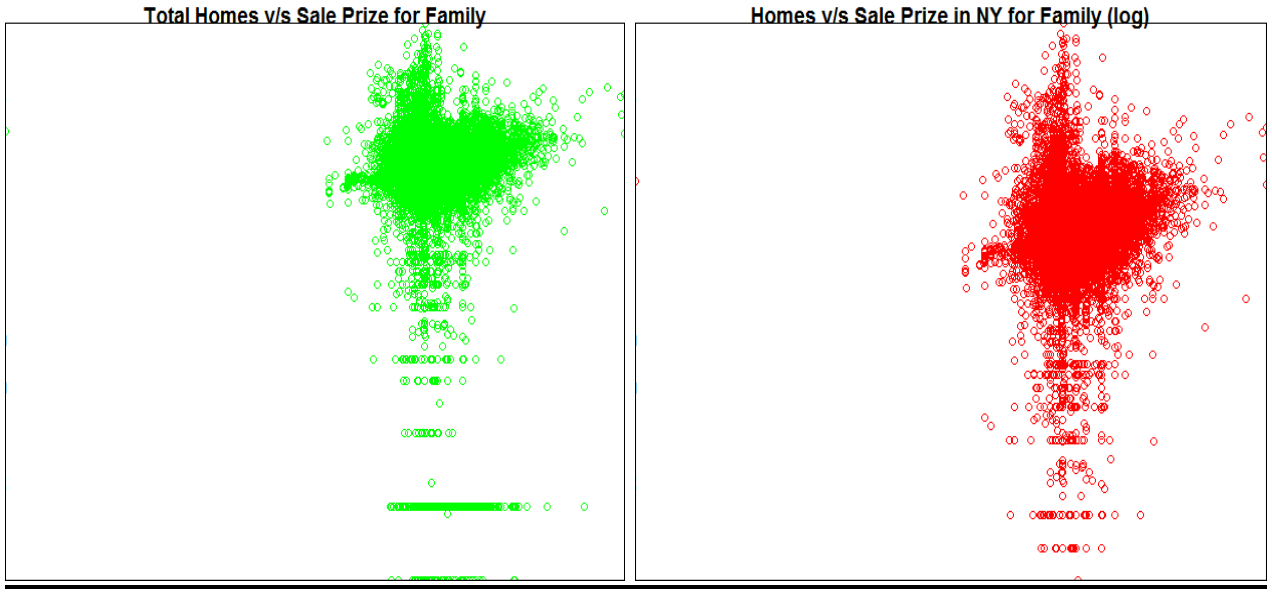
Analysis of Total Homes







Total Homes v/s Sale Prize for Individual



Total Homes v/s Sale Prize for Family

World Development Indicators Data (Own data)

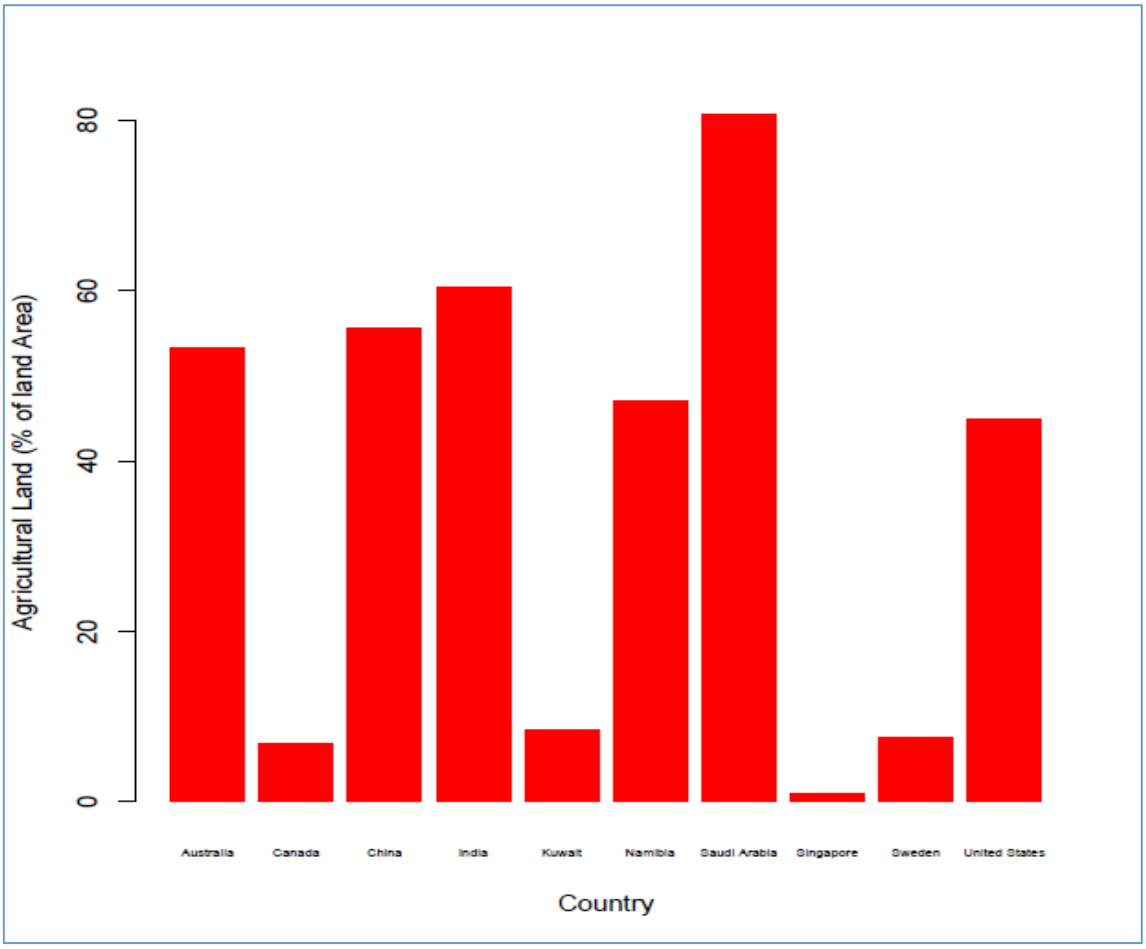
We have used the World Development Indicators (WDI) data for different countries by UN. The World Development Indicators (WDI) is the statistical benchmark that helps measure the progress of development. The WDI provides a comprehensive overview of development drawing on data from the World Bank and more than 30 partners. The complete WDI database includes more than 1,200 indicators.

Data Set Name: World Development Indicators Data by UN
Resource: http://data.un.org/Browse.aspx?d=19

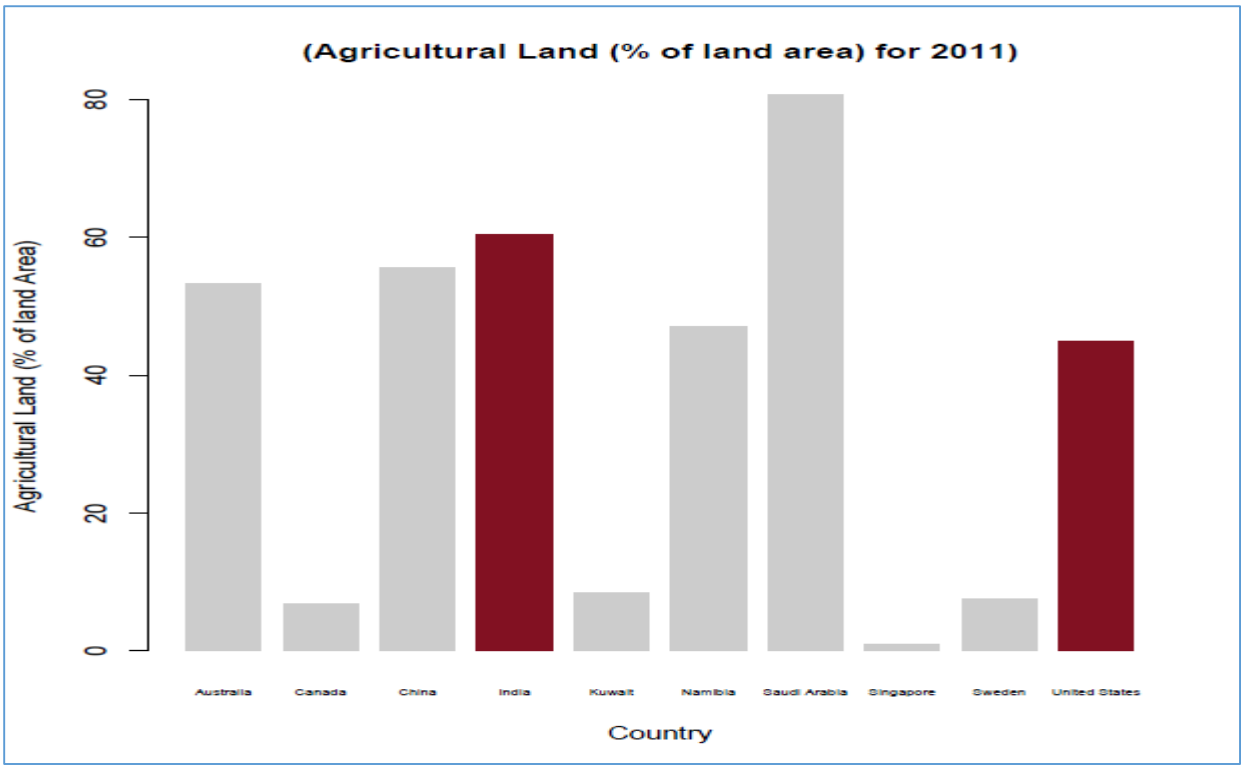
Experiments, Plots and Interpretations

For some cases we show the Indicators of all the countries of the world. But to be more clear and specific we pick 10 countries to show their comparisons on the basis of various indicators.
We have chosen indicators like: Agricultural Land available, GDP Growth, Electric Power Consumption, % of boys and girls enrolled in primary and secondary education etc.
On the basis of these indicators we have also plotted some indexes such as Environmental Performance Index (EPI) and Biodiversity.
After this we show the parameters of US alone for years 2002 and 2011 (taking a gap of 10 years).

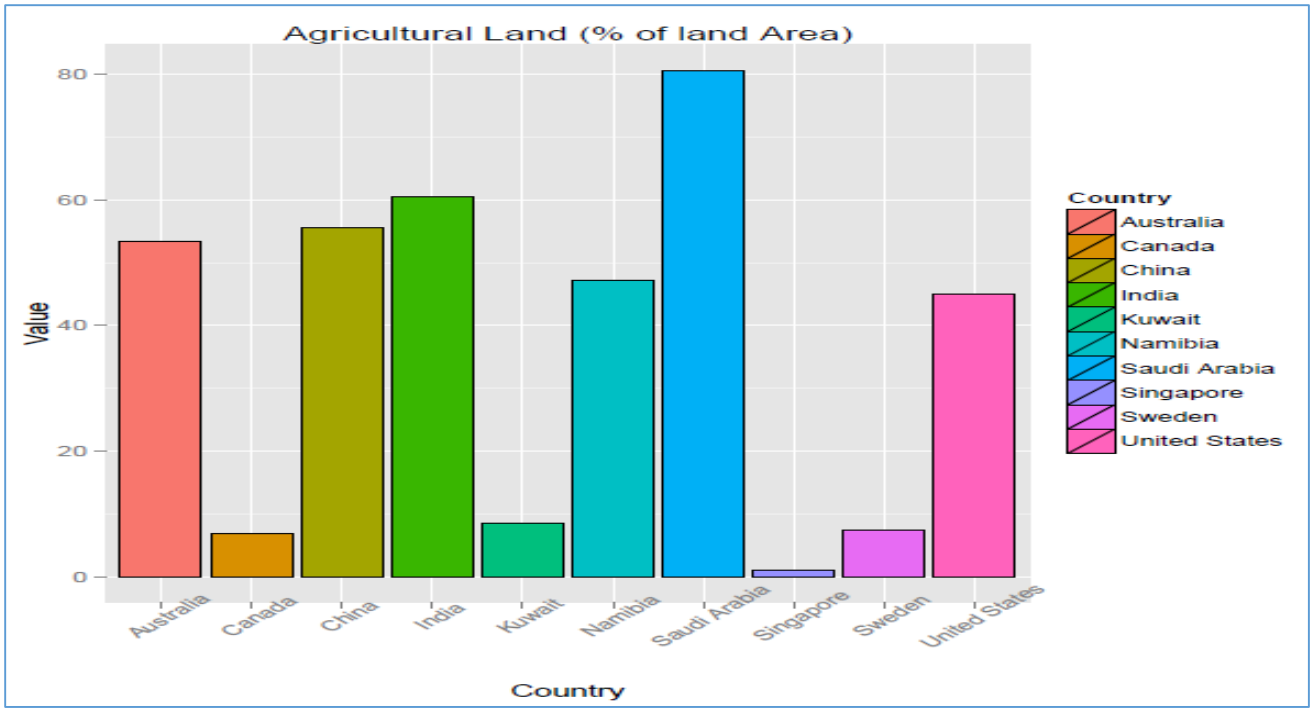
Case 1: Data of 10 countries for 2011 to find out which country is leading in the development front.
10 Countries, 2011: India, United States, China, Canada, Sweden, Kuwait, Namibia, Saudi Arabia', 'Singapore', 'Australia
Indicator: Agricultural Land (% of land Area)
Here we show the various plots on the same indicator so that we are able to find a relation in the plots and easily compare them.



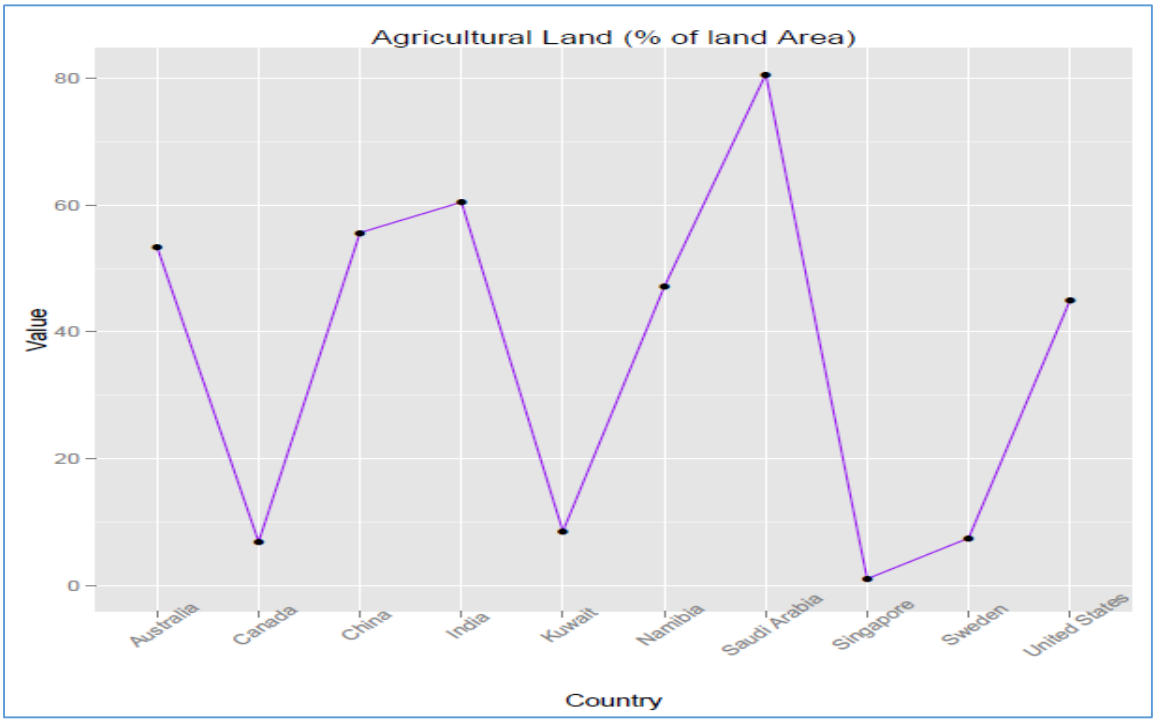
Bar Plot showing that Saudi Arabia has highest proportion of agricultural land area for year 2011



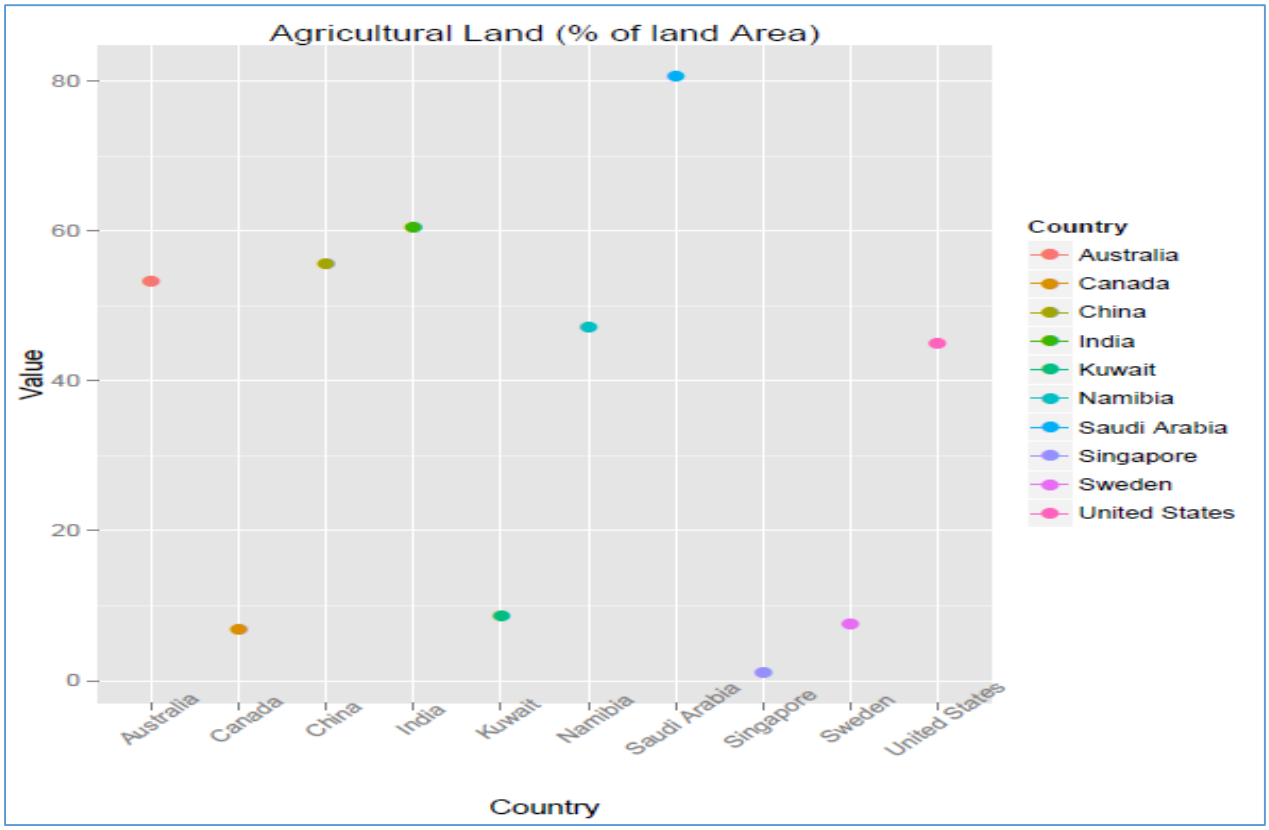
Bar Plot showing that India's agricultural land area is greater than United States for year 2011.



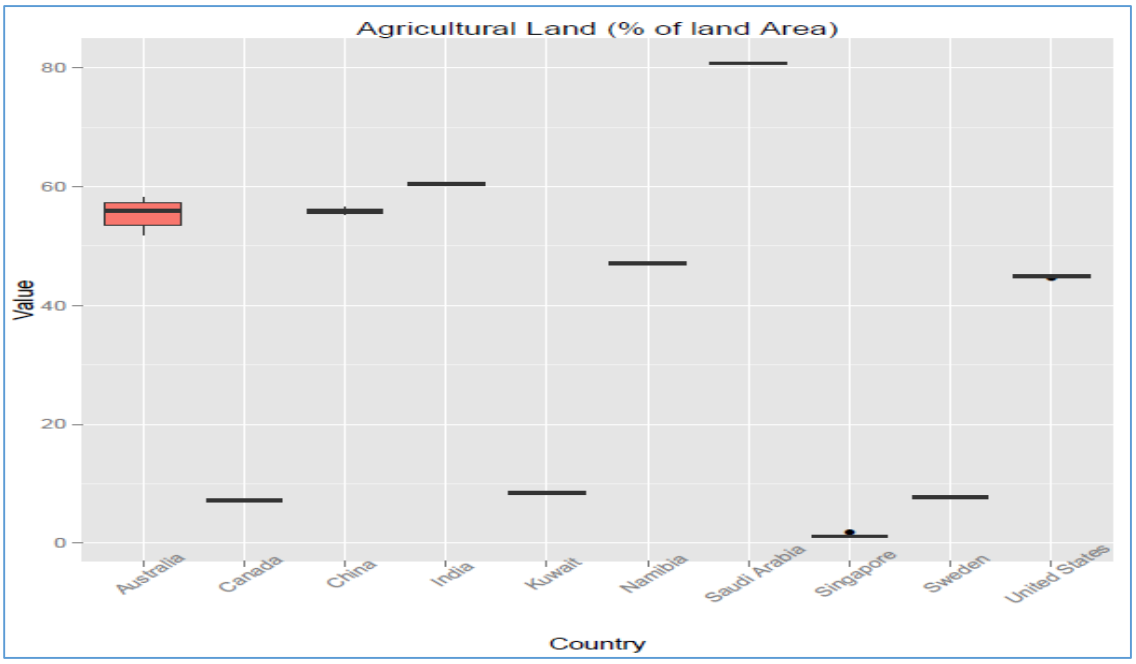
Bar Plot showing Agricultural land share for 10 countries for year 2011



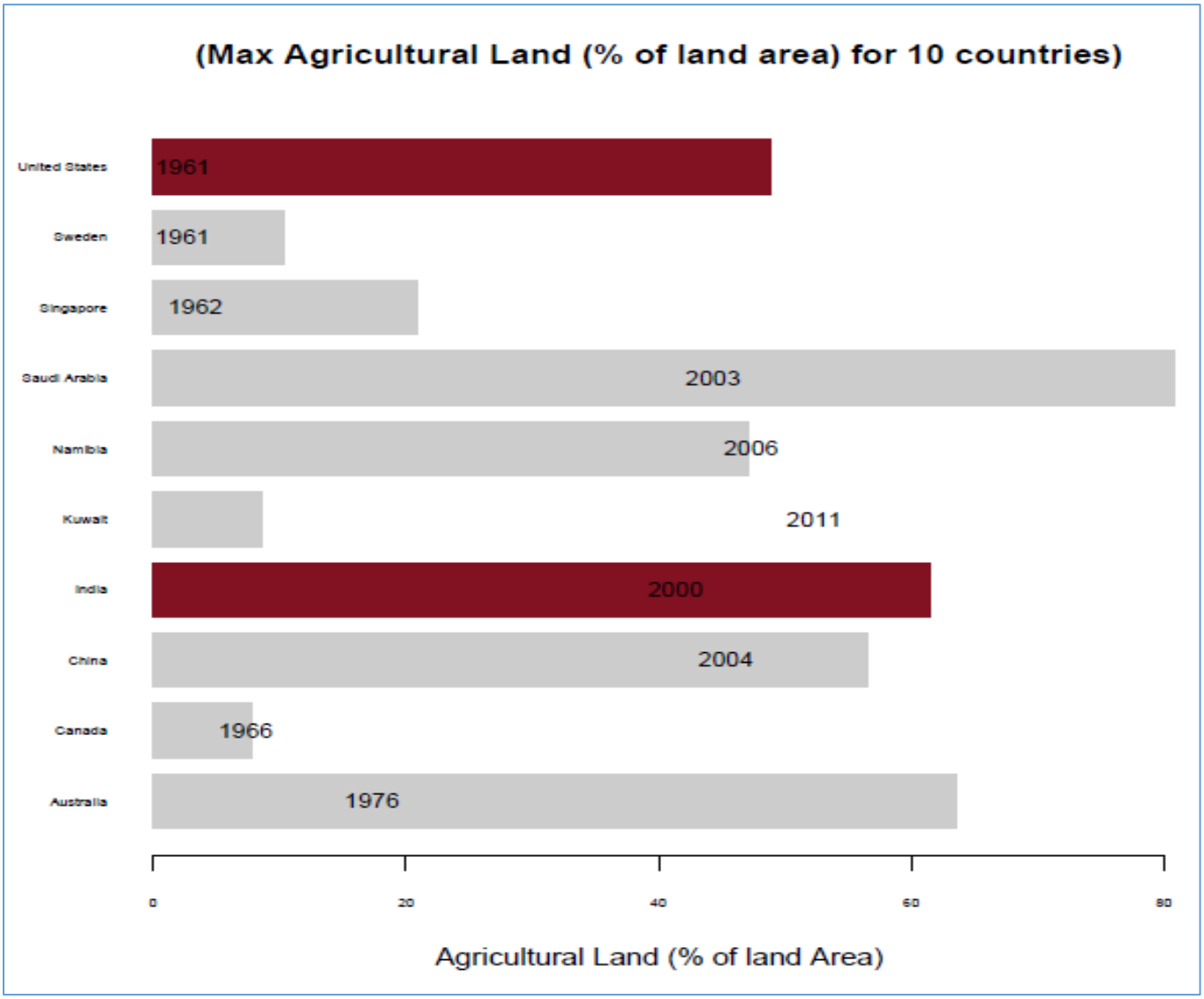
Line Graph showing Agricultural land share for 10 countries for year 2011



Scatter Plot showing agricultural land area for 10 different countries for year 2011



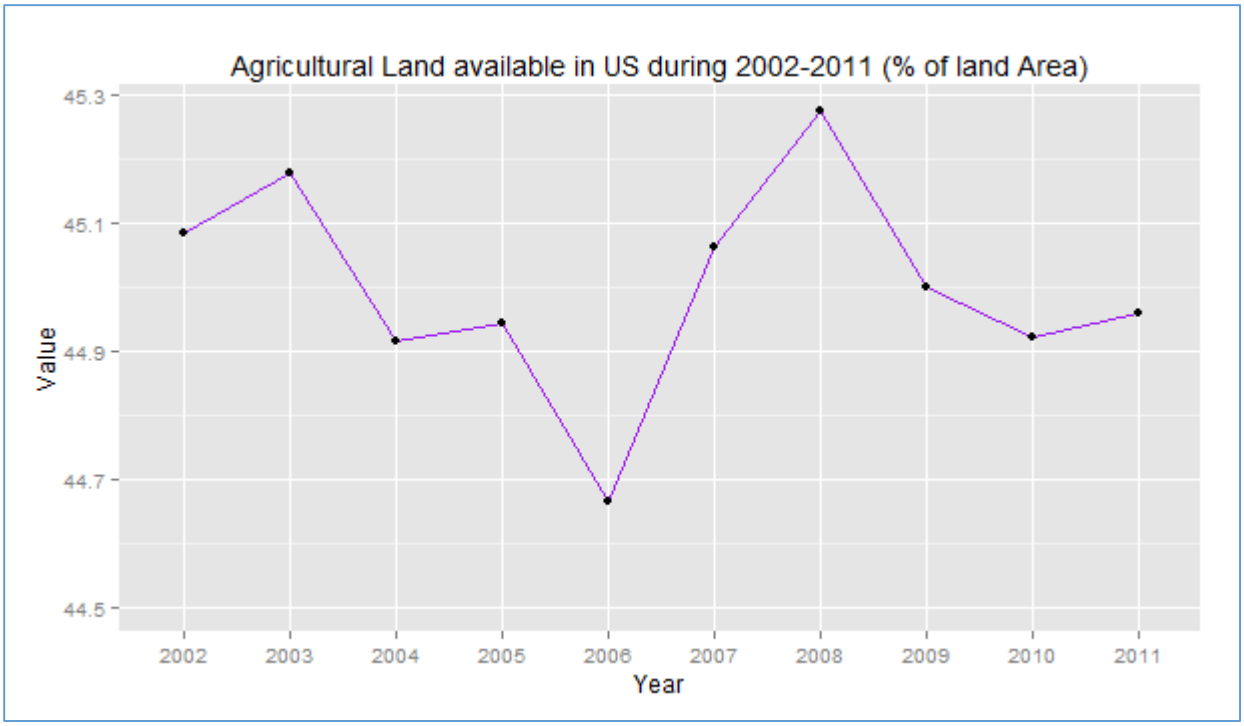
Box Plot showing agricultural land area for 10 different countries for year 2007-2011



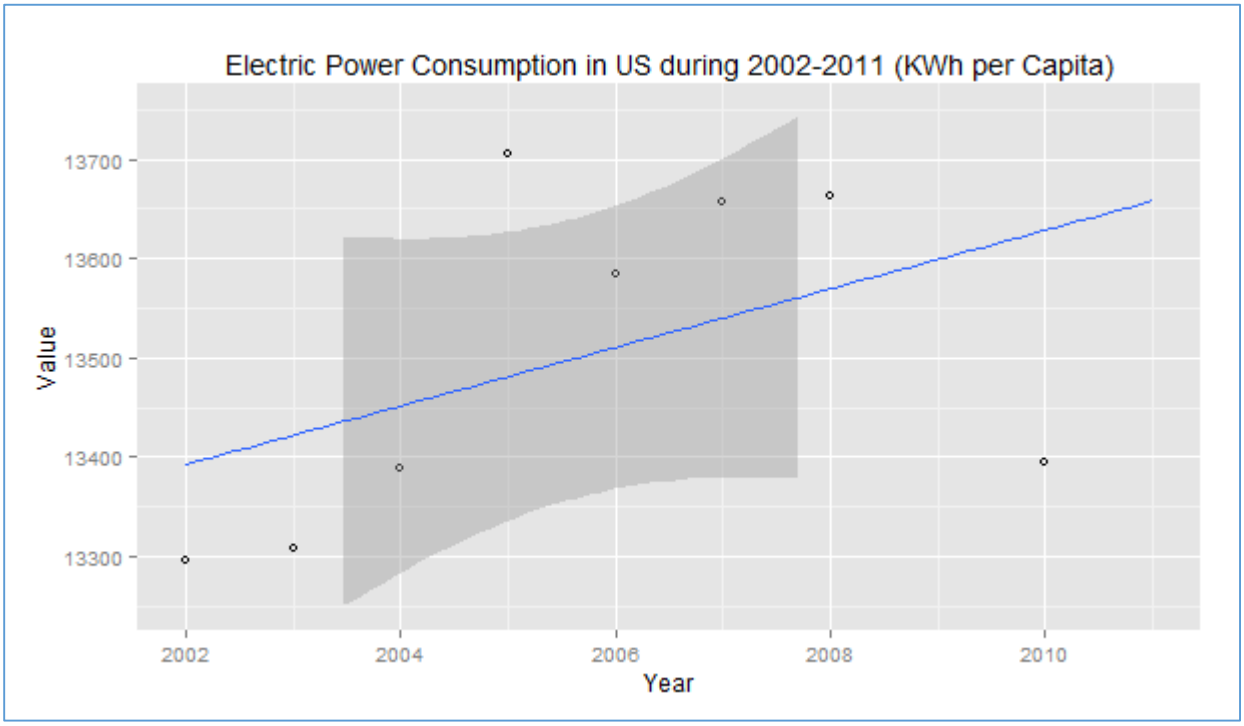
Maximum value data for the 10 countries and the corresponding year in which they had the maximum value.

Result: Out of the chosen 10 countries, Saudi Arabia has the highest percent of agricultural land from 1961-2011

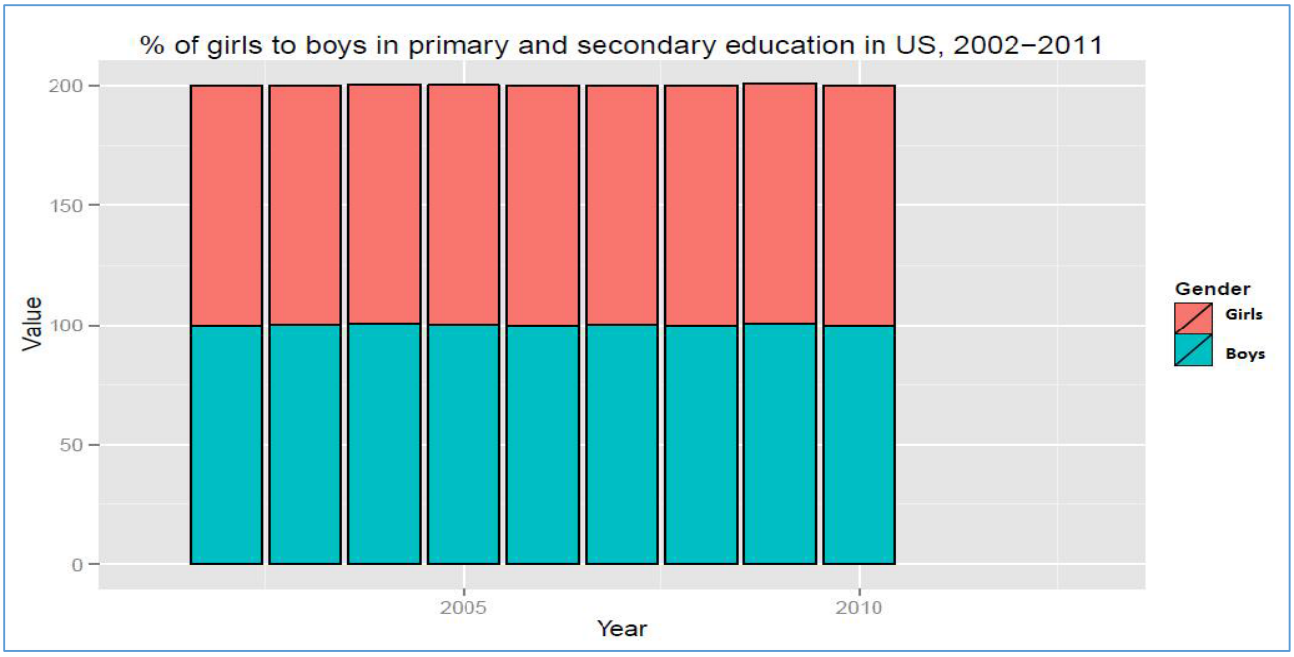
Case 2: Data of USA for 2002 – 2011 based on different development indicators to comment on its development record.



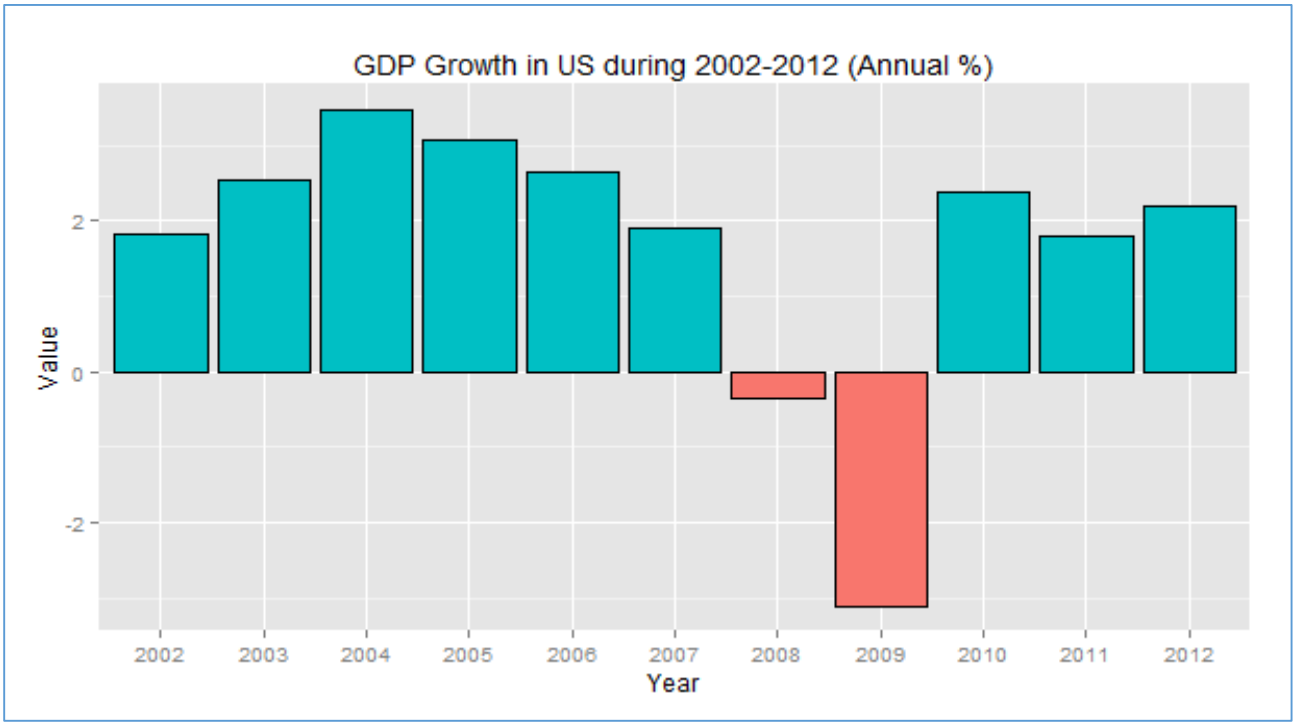
This plot depicts the % of land available for agriculture during 2002 to 2011. Though the variation in value is very slight over the years, we can see that in 2006 the land available was least and in 2008 the available land was the highest. Reasons behind variation in land could be shift in industries and expansion of housing.



Electric Power Consumption in US over 2002 to 2011 ranges between 13,300 KWh per Capita to 13750 KWh per Capita which is a sign of stability and consistency in power usage.

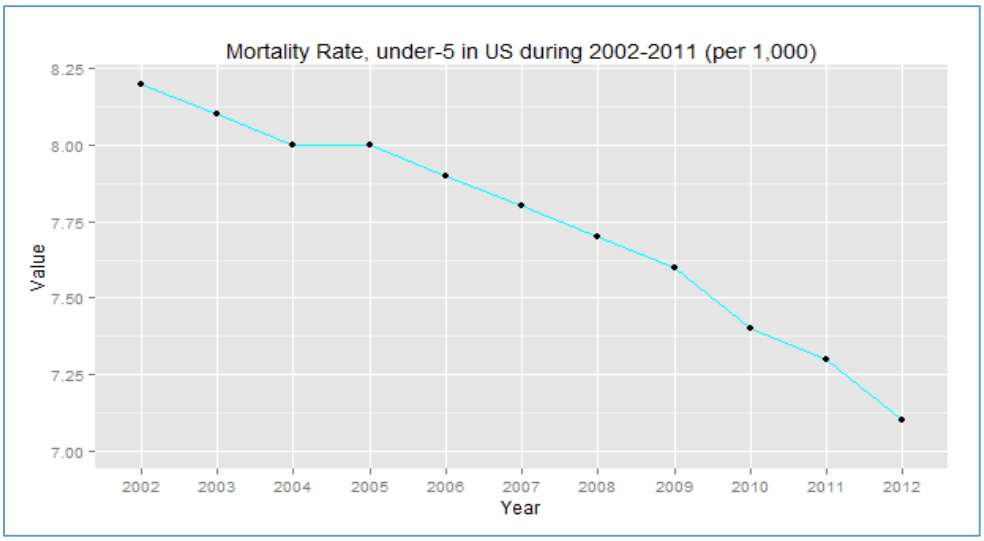


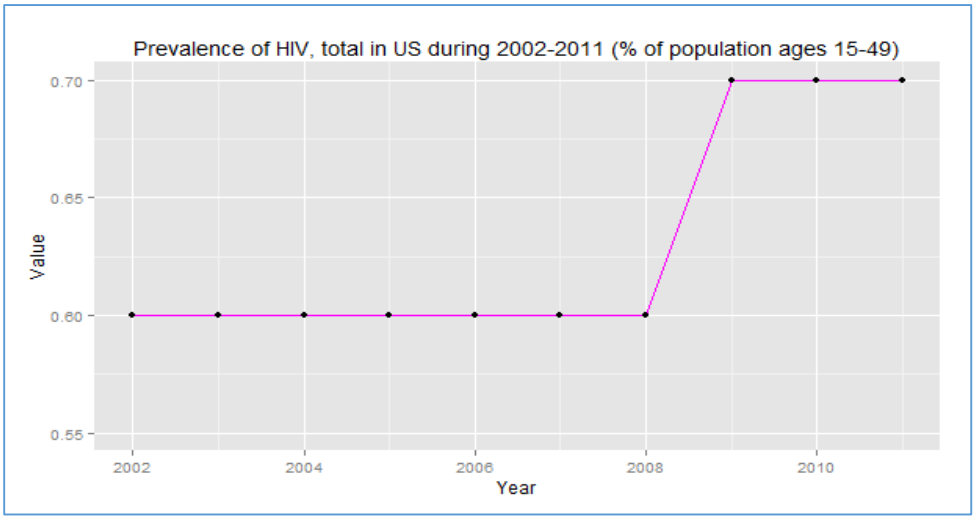
In this graph, the bars in blue color show percentage of boys and those in red color show percentage of girls. The plot is over depiction of the ratio between % of girls to boys enrolled in primary and secondary education in US over 10 years i.e. 2002-2011. We can see that US has maintained a good ratio over the years and we can conclude that US does not support gender bias.

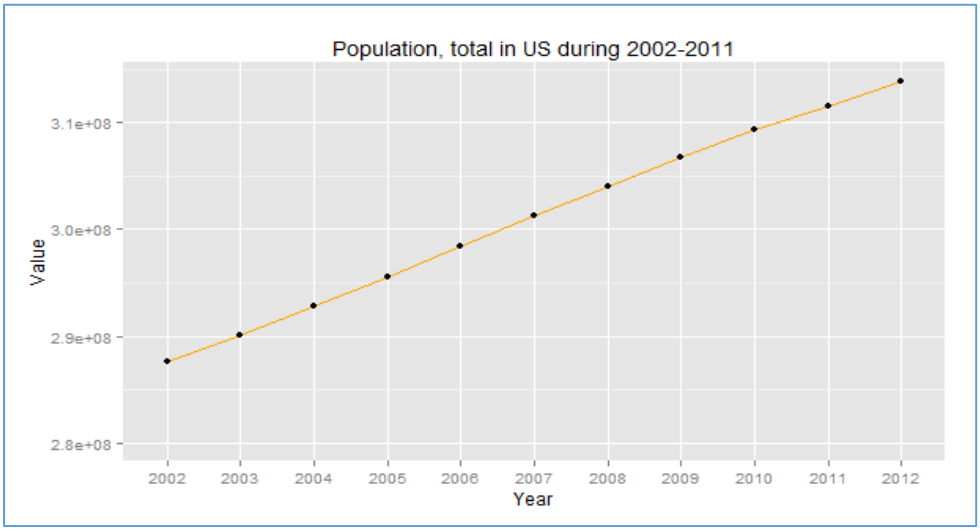


The graph shows that there had been a negative growth in GDP in US during 2008 and 2009. The reason could be the recession which gripped the entire nation during those years. The impact is evident on the GDP growth. But from 2010 the nation quickly gained the growth rate.

Health Indicators



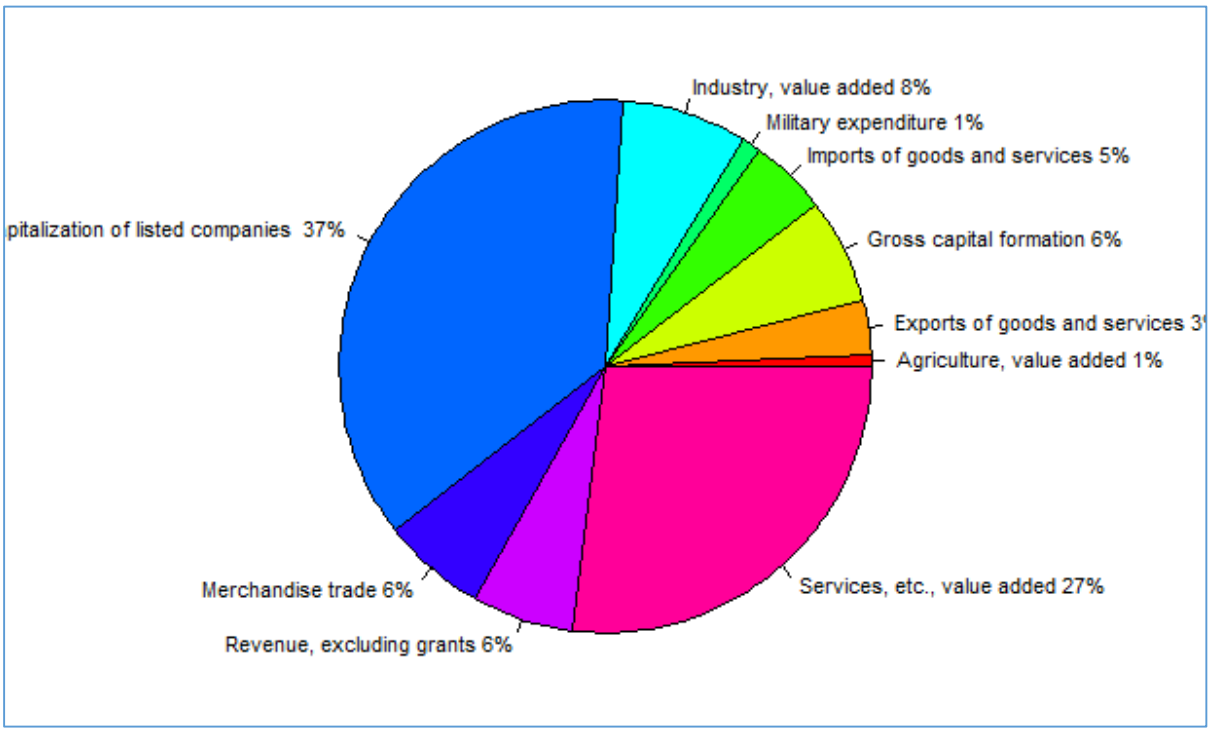




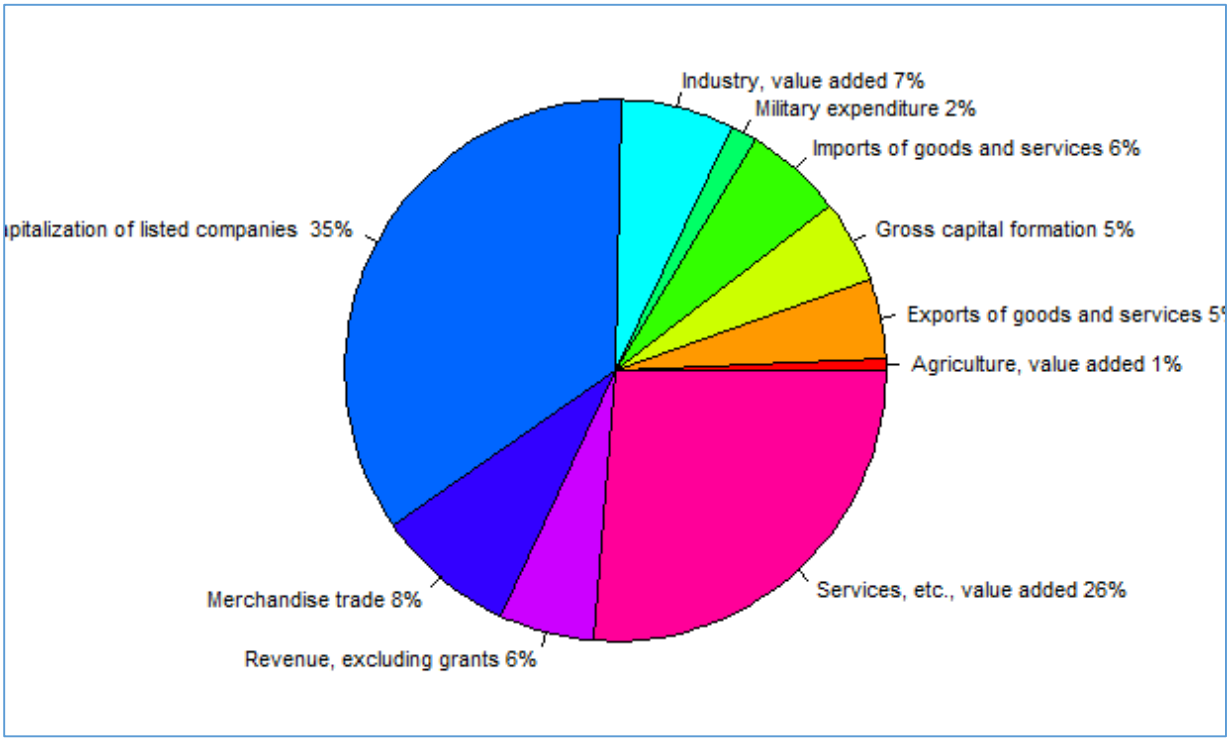
These indicators indicate the health and medical progress of US over the 10 years. The nation has shown a constant decline in mortality rate which is a good sign, though prevalence of HIV showed a hike in 2009. The population has grown steadily in 10 years.

Contribution of Indexes towards GDP of US

Contribution of Indicators in the Year 2002

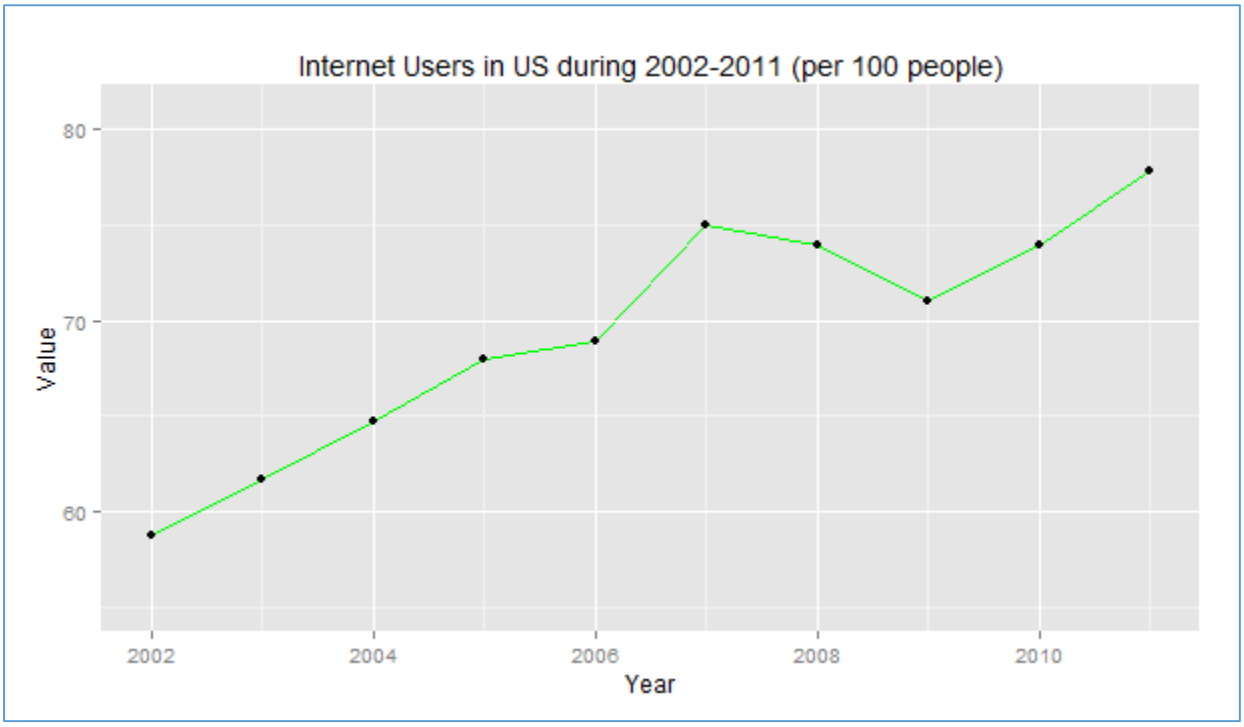


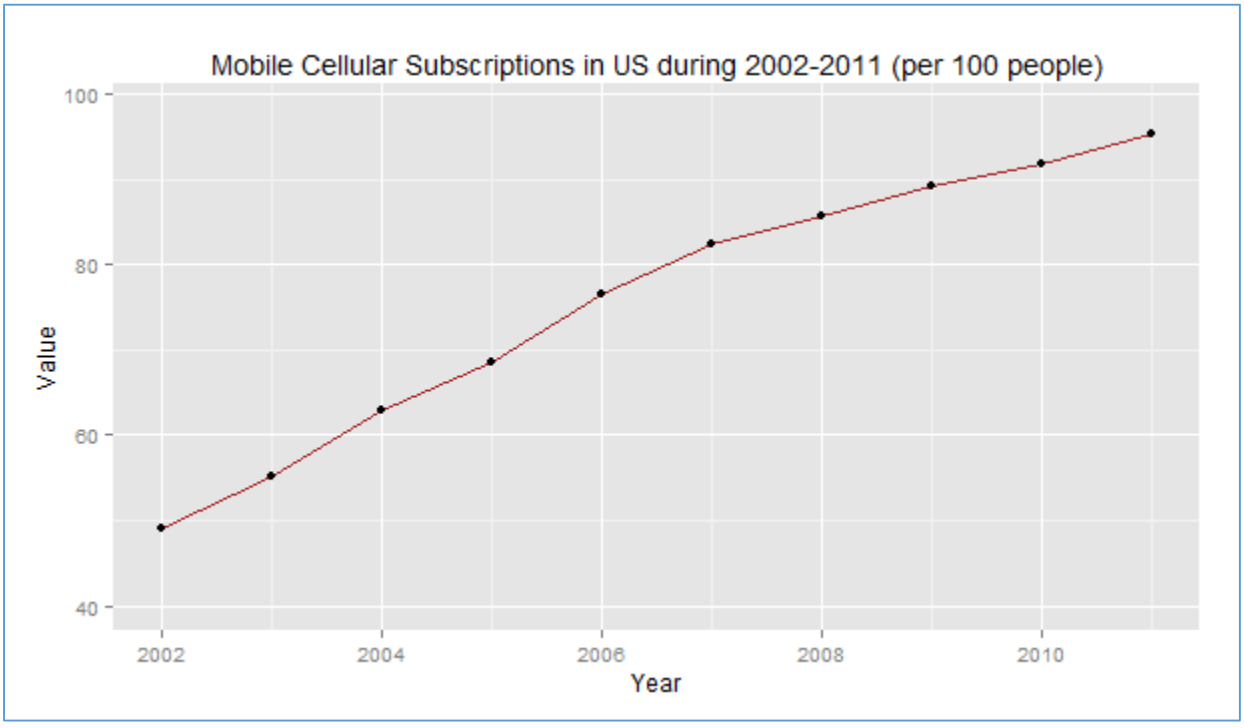
Contribution of Indicators in the Year 2011



These pie charts show contribution of various World Development Indicators towards GDP of US and give a comparison of their contributions in the year 2002 and 2011. It seems there hasn't been much change in the shares of indicators over the 10 years.

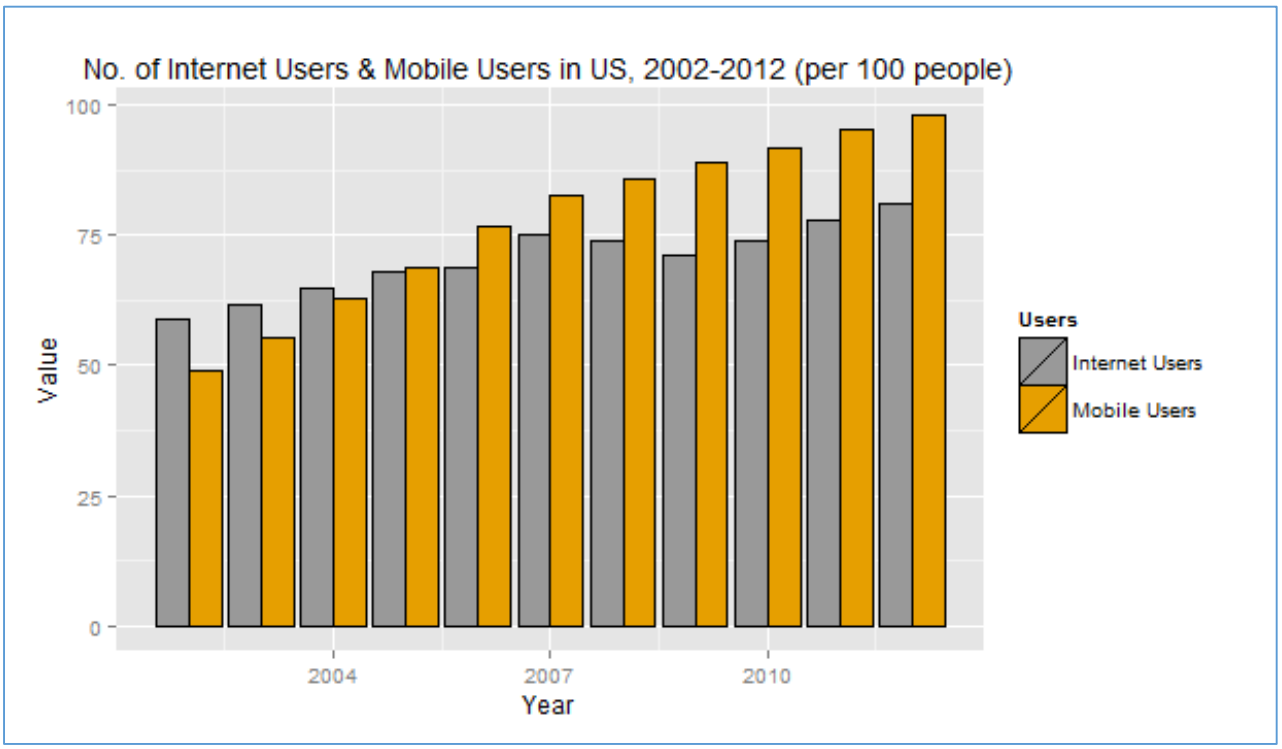
Growth in Technological Advances in US

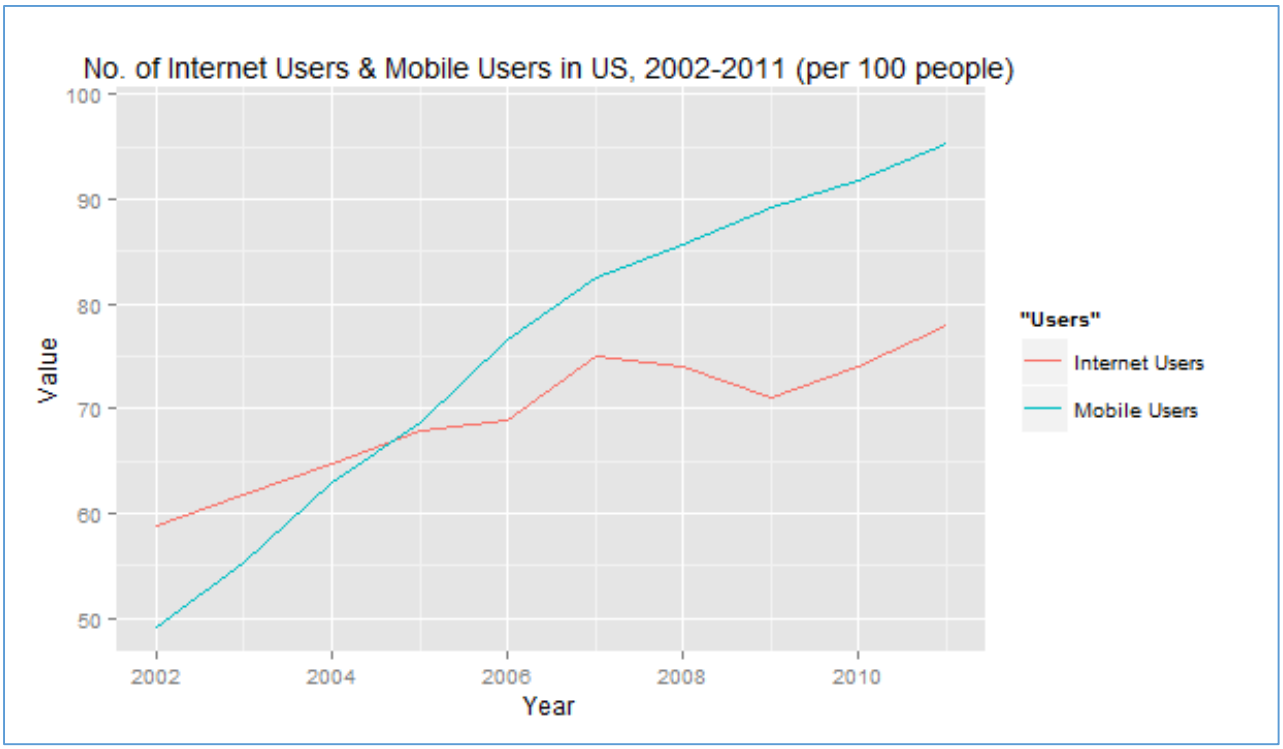




There has been a constant rise in Mobile Cellular Subscriptions in US over the 10 years. Internet Users show some lows in 2006 and 2009 but by the year 2011, US showed an increase in Internet Users as well. This shows that the country has been thriving in technological advances as well.

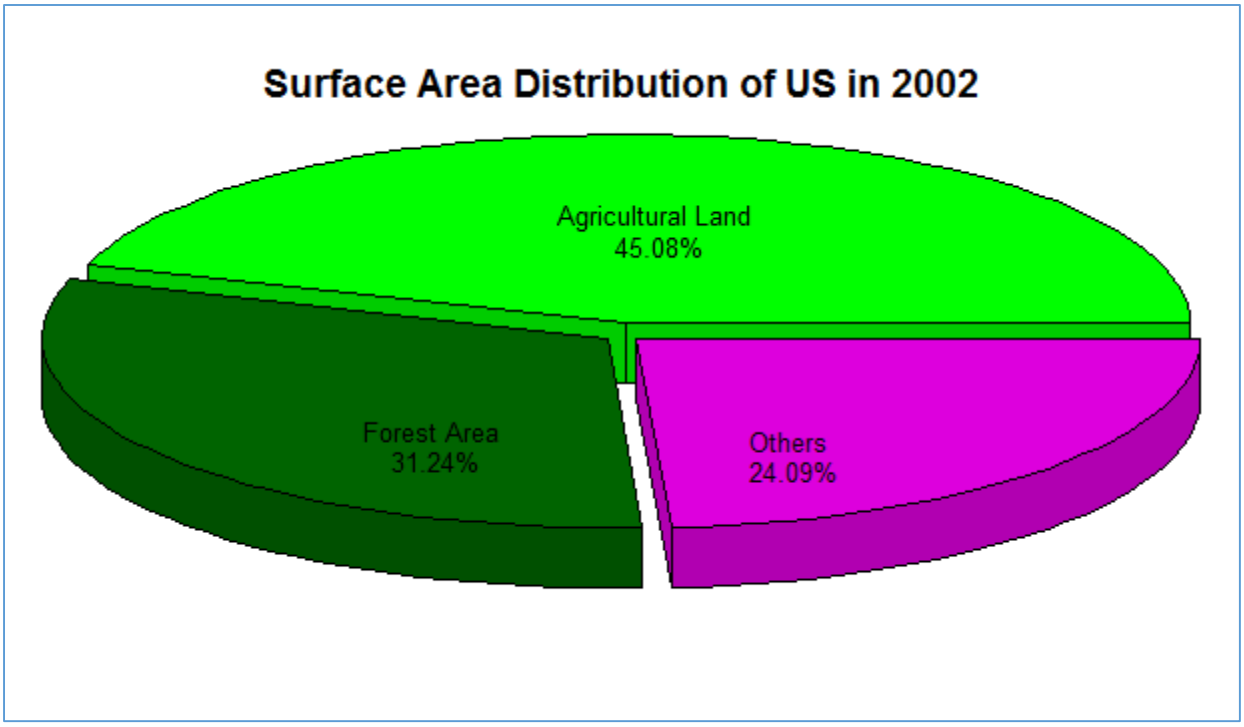
Comparison between Internet and Mobile Users

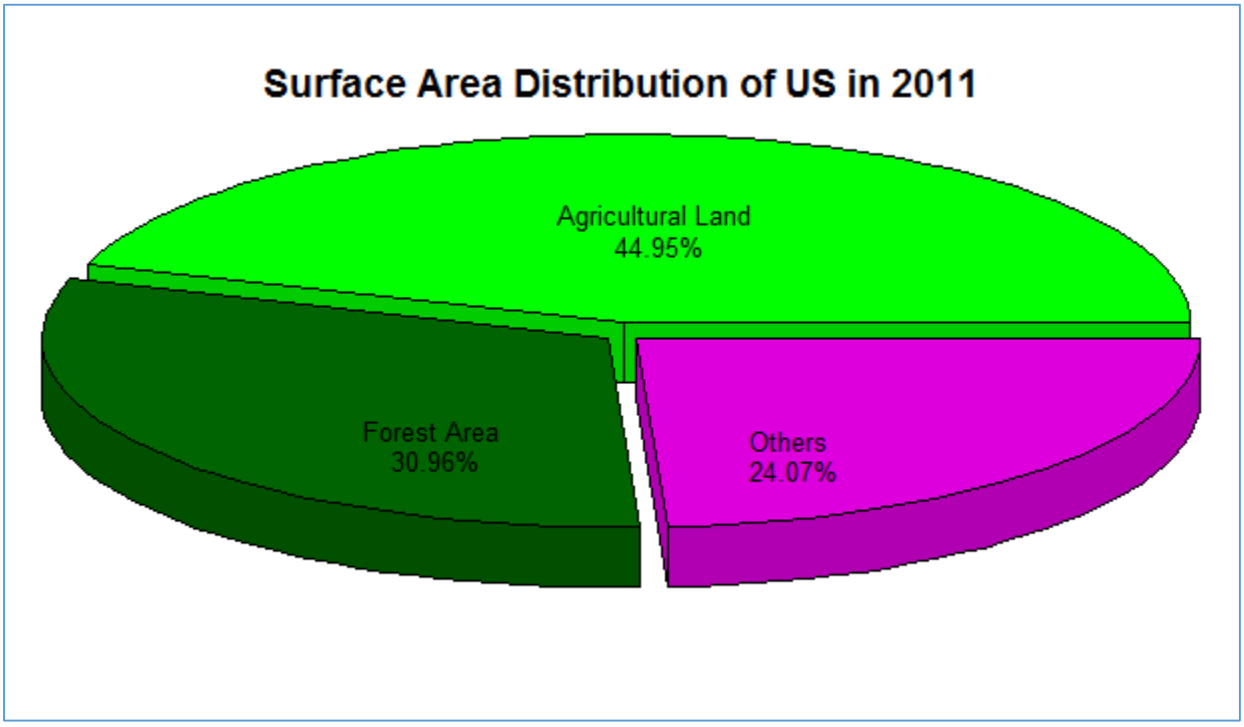




Initially internet connections were leading over mobile connections. But with the time, the mobile subscriptions surpassed internet connections with a fairly large margin, reaching 96 users per 100 users which is a remarkable figure.

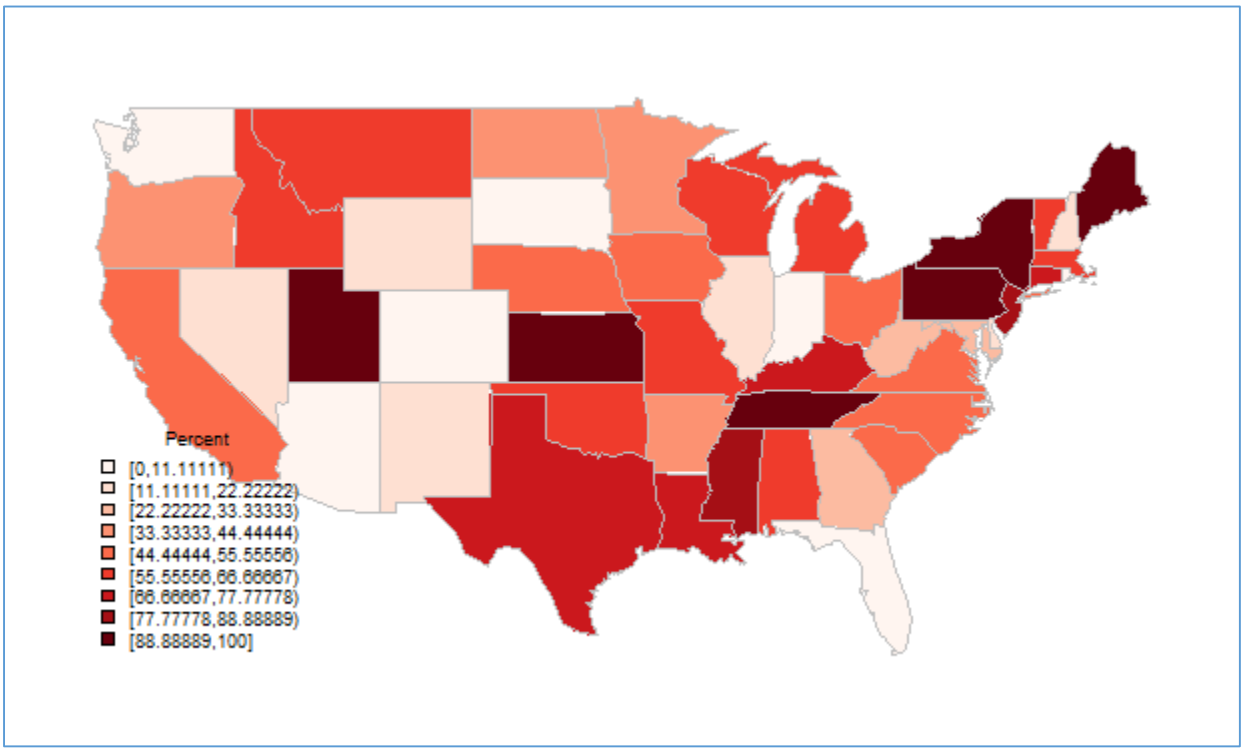
Comparison of Land Distribution of US in 2002 and 2011





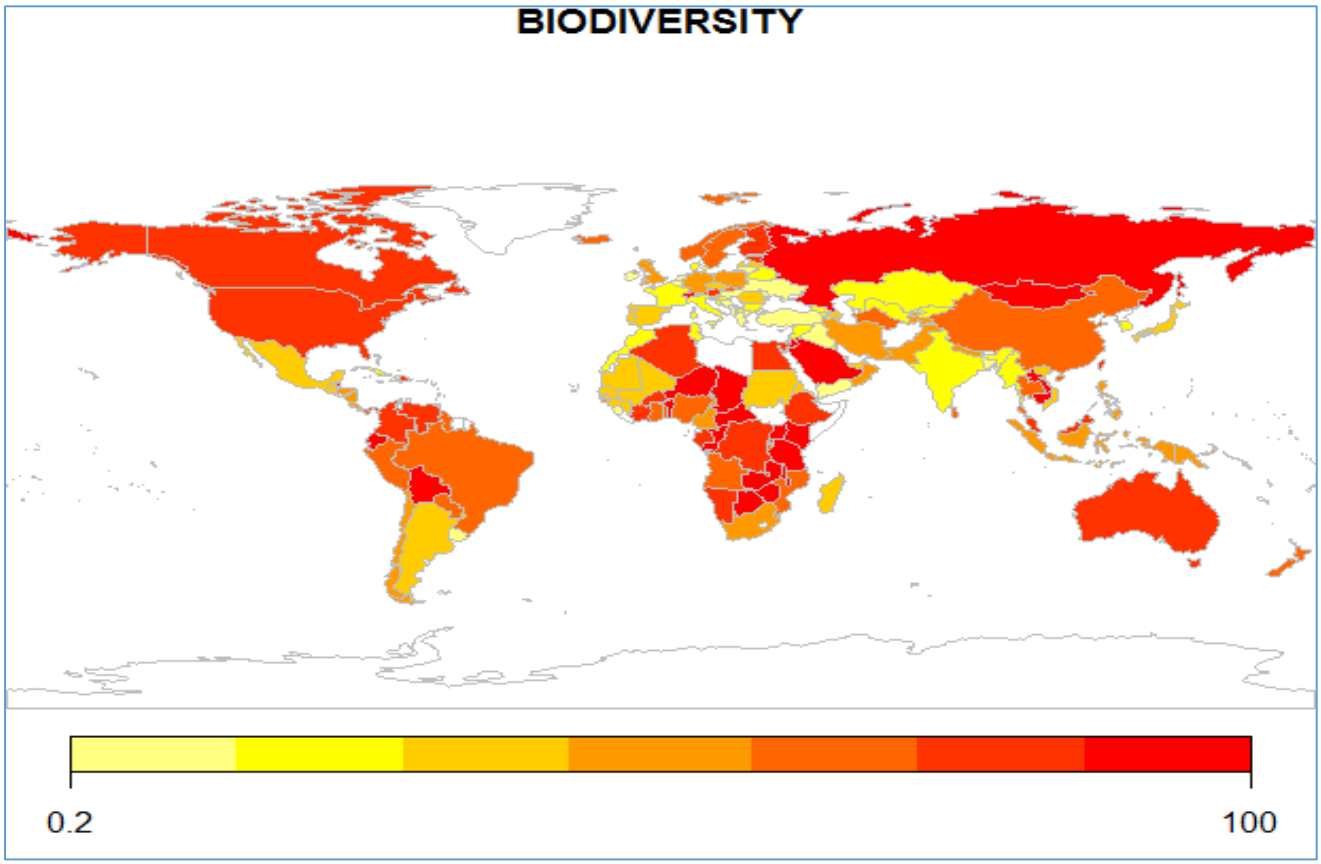
The pie chart depicts the land distribution in US over 2002 and 2011. The distribution remains more or less the same over the 10 years out of which agricultural land occupies the highest percentage of land, followed by forest area and then housing, industrial and other purposes.

Coal Power Concentration in US, 2011



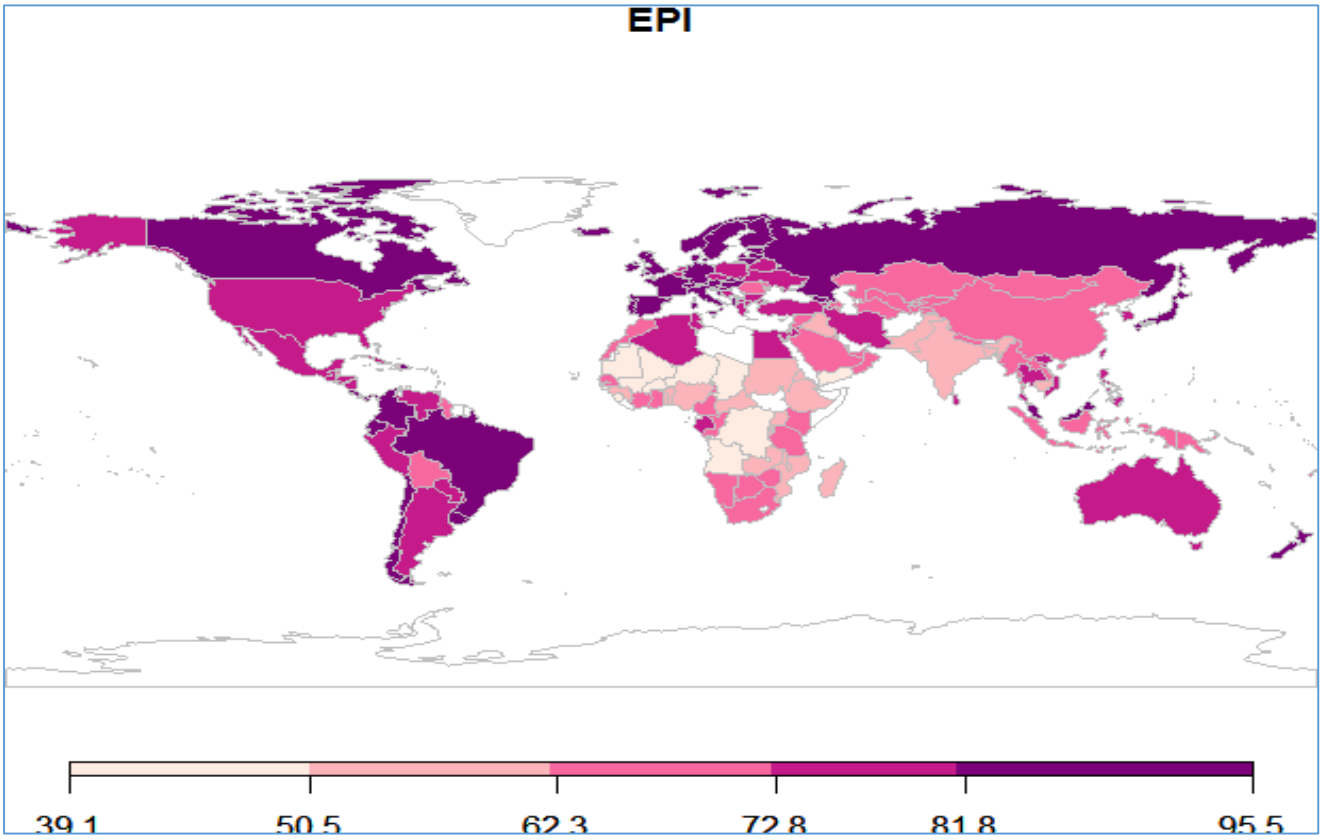
The map shows the distribution of coal power in US during 2011. The regions in darker shade show higher percentage of coal power concentration and vice-versa with light shades. States like Utah, Kansas, Tennessee, New York, Pennsylvania show the highest concentration of coal power whereas states like Arizona, Colorado, S. Dakota show the lowest concentration.

Some Indicators in a Larger Picture



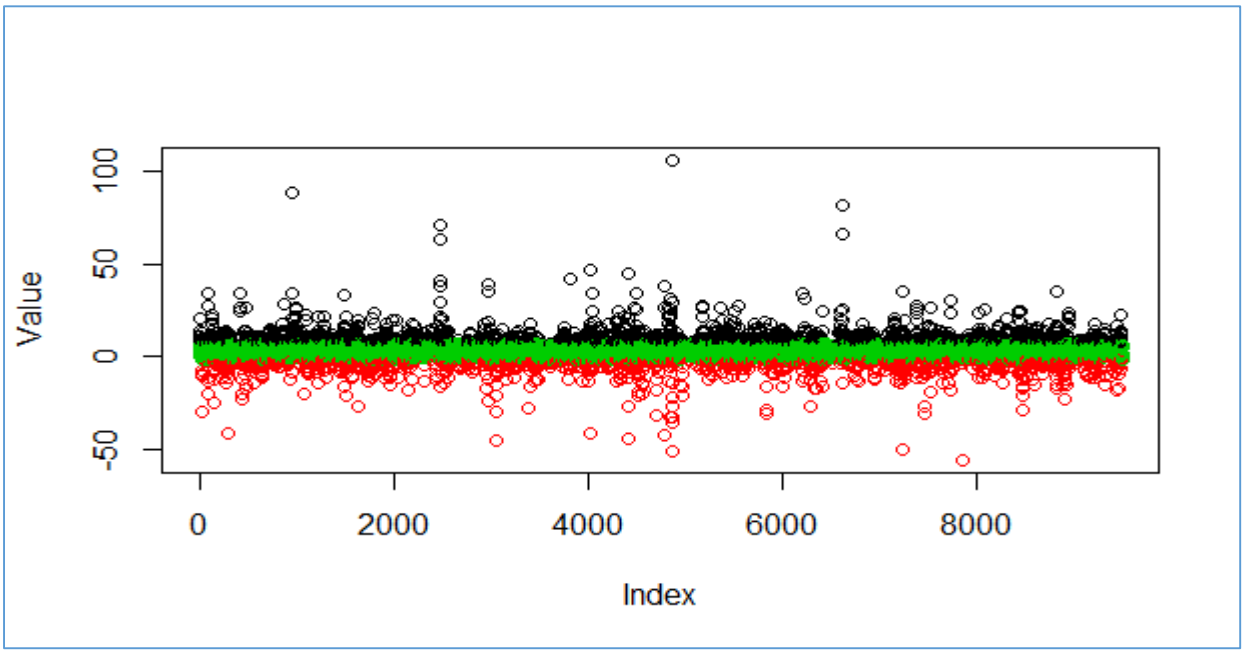
Biodiversity is the variety of life in the world or in a particular habitat or ecosystem. This plot depicts the variation in biodiversity across the countries of the world ranging between 0.2 and 100. We can see that majority of the countries with greater surface area support greater diversity and those with lesser surface area support lesser diversity.

Environmental Performance Index



The Environmental Performance Index (EPI) ranks how well countries perform on high-priority environmental issues in two broad policy areas: protection of human health from environmental harm and protection of ecosystems. We see that countries like Russia, Canada, USA and Australia.

K means Clustering of GDP Growth of US, 2011



k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In this plot we have plotted the GDP Growth of US in the year 2011 against its values for the range of indexes.

Lessons Learned

- Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch.

- This project has introduced us to various methods for data exploration, screening and adjustment using R.

- R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

- As Data Science is an emerging field, this project gave us a thorough insight in what Data Science can be all about. Working on real world data such as that of New York Times, RealDirect.com and World Development Indicators introduced us to how these data are collected and concocted to form a big data.

- The project taught us various skills in R programming and forced us to think of new ways and strategies to gather and organize data so that it is in its best form to use and analyze.