# EXPLORATORY DATA ANALYSIS USING R

## CSE 587 - DATA INTENSIVE COMPUTING

## UNIVERSITY AT BUFFALO

**Authors:**

**HARISH MANGALAPALLI**

**hmanagala@buffalo.edu**

**RAJARAM RABINDRANATH**

**rajaramr@buffalo.edu**

# Contents

ABSTRACT ..... 3
OBJECTIVE ..... 3
APPROACH ..... 3
NEW YORK TIMES ANALYSIS ..... 3
*User Distribution* ..... 4
*Impressions and Click distribution* ..... 6
*Time Series* ..... 9
*Learnings* ..... 11
REALDIRECT CASE STUDY ..... 11
*Comparison of freshness of neighborhood (Fraction of new homes in neighborhood)* ..... 12
*Comparison of neighborhoods by percentage contribution of new homes to total* ..... 13
*Improper distribution of Sale Price within each half year* ..... 14
*Distribution of sale prices across time (half-yearly) for various property categories* ..... 15
CENSUS DATA ANALYSIS ..... 17
*Education Analysis* ..... 18
*Employment Analysis* ..... 19
*Personal Income Analysis* ..... 21
*K Means Clustering* ..... 22
*Results* ..... 23
Cluster centroids from k-means for the years 2001-2003 ..... 23
Cluster centroids from k-means for the years 2004-2007 ..... 23
Cluster centroids from k-means for the years 2008-2010 ..... 23
*Within Group Sum of Squares for k between 2 and 15* ..... 24
*3 Dimensional Visualization of the Clusters for 08-10* ..... 26
*Learnings* ..... 30
APPENDIX ..... 31
*NEW YORK TIMES ANALYSIS* ..... 31
*CENSUS DATA EDA* ..... 31
*CENSUS DATAFRAME SCREEN SHOT* ..... 32
REFERENCES ..... 33

# ABSTRACT

Data Science or data-intensive computing process involves a phase where preliminary understanding of the characteristics of the data is explored. This phase is commonly known as exploratory data analysis. In this project we take on three datasets from varied domains and apply standard EDA to those datasets. We pose questions to the datasets and then infer results from the summary statistics that we generate. We use the R statistical language to achieve our goals.

# OBJECTIVE

The Objective of the project is to get a good feel for Exploratory Data Analysis, and a hands on experience in R. Understand the need and requirement for EDA irrespective of the domain being analyzed.

# APPROACH

We pose questions and query the data for answers. We use R statistical language along with all its packages to get the answers that we want. We use R Studio to enhance the programming experience. We learn about the different techniques in EDA by working out the examples given in the "Doing Data Sciences" book and apply the knowledge to analyze the dataset of our choice.

# NEW YORK TIMES ANALYSIS

**Dataset Description**

The key variables that we used for the New York Times dataset are the following:

- Simulated New York Times data
  - Age (Information is available only for users who have signed in)
  - Sex (-"-)
  - Impressions → this information is available for all users
  - Clicks → this information is available for all users
  - SignedIn → A Binary variable indicating if the user has signed in or not

**Derived Variables**

- Age Group → This variable was derived from the Age column, using the following buckets
  - 0 and below, 0 – 18, 18 – 24, 24 – 34, 34 – 44,
  - 34 – 44, 44 – 54, 54 – 64, 64 and above
- Row level Click through rate (CTR) – row level metric calculated using = Clicks/Impressions
- Click Behaviour
  - This variable is a categorical variable derived from the row level CTR variable
  - Classifies users as: using quartiles of "***Row Level CTR***"
    - Not targetted → Straigh Forward from data
    - Non-clickers → Users who have impressions but do not click
    - Low-clickers → First Quartile

- Ok-clickers → Second and Thrid Quartile
- Hi-clickers → Fourth Quartile

*Note :*

***Age groups:*** *There is a Category "Inf--0" for users who have not signed in and therefore there is not age data*

***Gender:*** *Since gender information is only avaiable for logged in users, we have labelled all non-logged in users as "Not logged In" in the gender category*

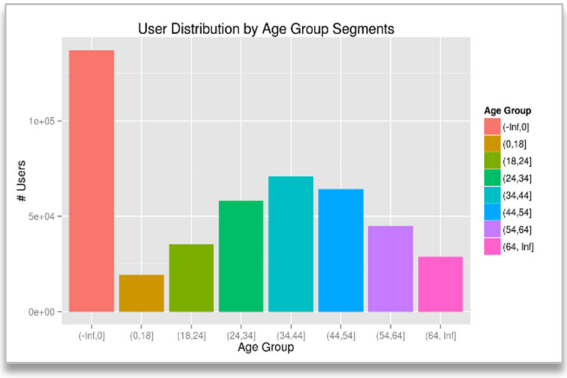
Equipped with these additional variables and derived metrices we went ahead and answered a some of the questions that popped into out head. Following are the questions that we endeavoured to answer by means of exploratory data analysis

We used the following variables as pivots for out analysis:

- Age
- Gender
- Click Bheaviour

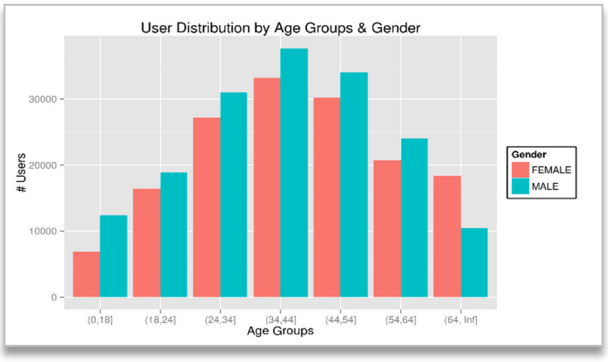
## *User Distribution*

- ***Who is visiting the NYT site? What is the User distribution with respect to pivot variables ?***



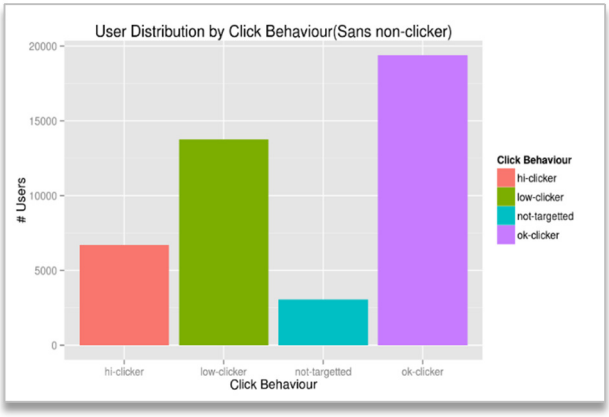

It is quite evident form the chart above that among the valid age groups, the NYT site has more visitors with age between 24 – 54 than any other age group.

- ***How is the distibution split across genders in each of these age group categories ?***



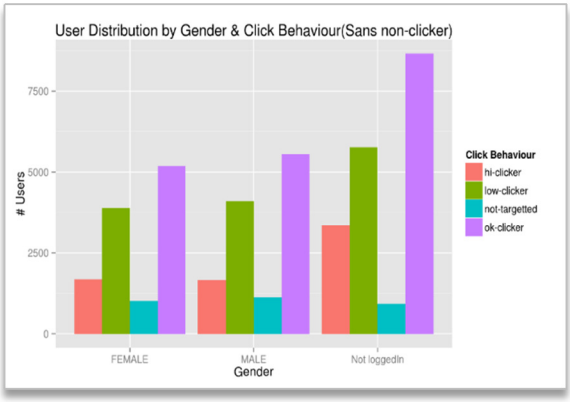

It is quite evident from the data that amongs all age groups the NYT site has more male visitors than female visitors. This can be inferred as NYT site being more popular amongst men of any of the age category under consideration

- ***How are the users distributed by Click Behaviour?***




The chart above is normal in terms of the distribution (considering only low-clicker, ok-clicker and hi-clicker). The results therefore are along expected lines. We plotted this chart to verify if our assumptions were indeed right. A non-normal distribution (an abnormal one) would have signaled something analolous about the underlying data, which would call for deeper analysis.

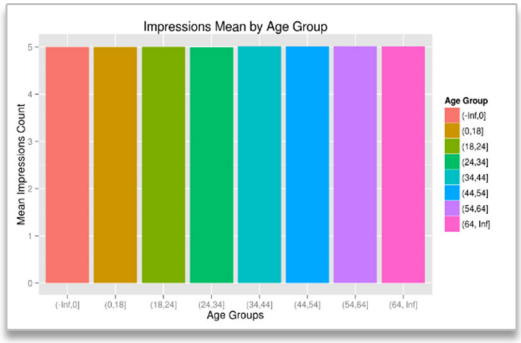
- ***How are users distributed by Gender and Click behaviour? ( more details analysis with respect to the previous question)***




*Note: in the above graphs we have not considered the "non-clicker" category. The reason being that group in general is very large and it ends up dominating the graph. We have therefore added it to the appendix*

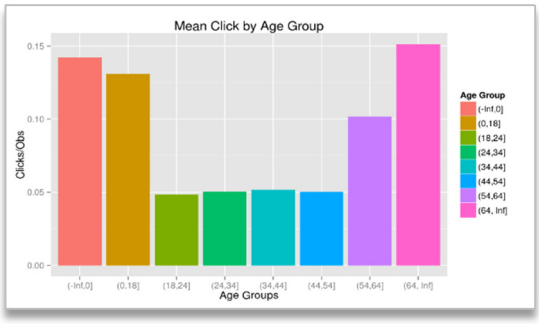
## *Impressions and Click distribution*

- ***Does NYT's ad engine exhibit bias in term of impressions shown to users ? In other words is there a specific segment of the population that the ad-engine targets***




The answer is pretty emphatic there is no Impression bias that we could discern form the data. The only inference that we can make is that the ad-engine does not discriminate (neither positive nor negative) based on age groups. An egalitarian ad-engine is not got for business though

- ***Well then how do users segments respond to the egalitarian ad-engine? Lets plot Mean Click by Age Group***

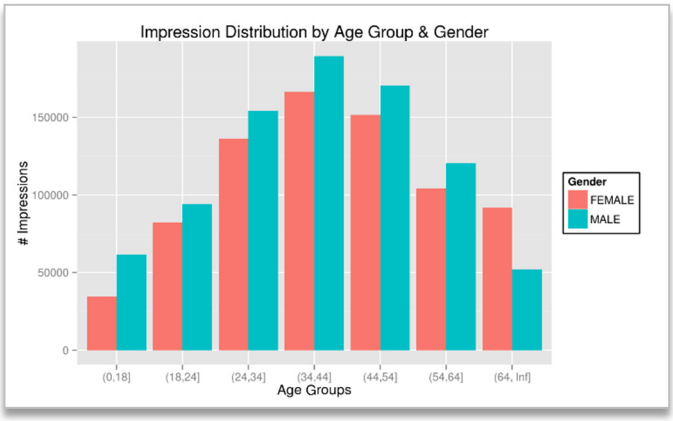



Well users do not reciprocate in an egalitarian fashion. There is indeed a pattern here; it is quite evident that users between 24 – 54 years of age do not on an average click as much as the other age groups. Older visitors and younger visitors click more on an average when compared to the age group highlighted in above. One could argue that based on user distribution insights we have more users from the 24 – 54 age category visiting the website, but we have accounted for that by taking the mean clicks.

- ***Is there some pattern we can find if we split the above chart by gender?***

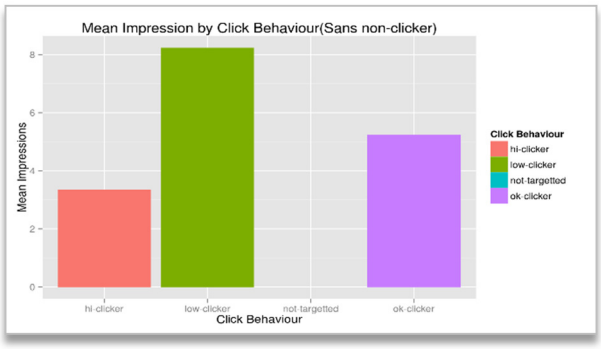
Well, there is no pattern. Men and Women both respond similarly to the egalitarian ad-engine across all age groups.

- ***What is the distribution of impression across the pivot variables ?***

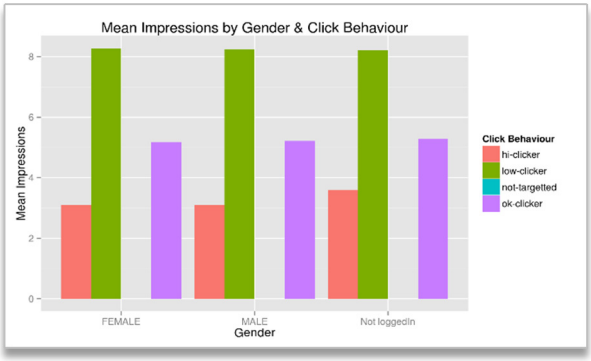



Well this is along explected lines we have more users in the age groups between 24 – 54 there for the aggregate impressions are greater than it is for other. No suprises here.

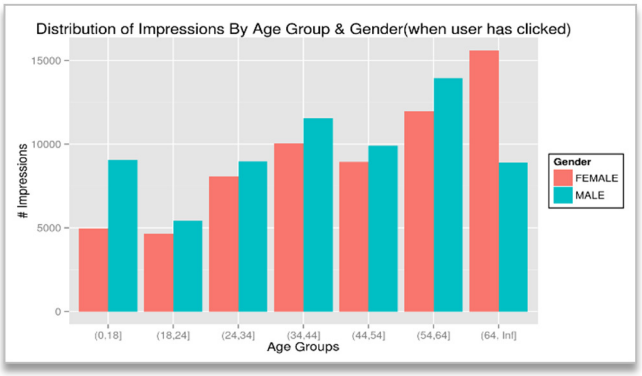
- ***Lets revist our targetting question, does the ad-engine learn from user clicks and from user data to target ads to groups? or could it learn given the data at hand?***




There is clear pattern. We can see that people who have more impressions thrown at them by the ad-engine belong to the low-clicker category, this is a bad return on investment and indicative of lack of algorithmic approach to tragetting users. The ad-engine has a lot to *"learn".* The following chart re-inforces the insight gleaned form the previous chart as users as the "low-clickers" across genders are also bombarded with impressions and do not respond positively to the bombardment.

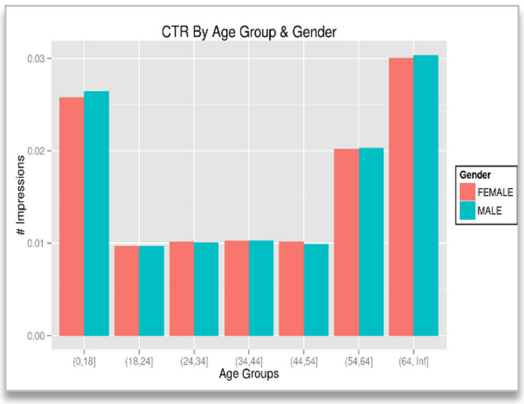


- ***Lets go a step further and ask ourselves what does the distribution of impression with respect to users who have clicked on atleast one ad ?***




We know from our analysis that age groups >64 and 0-18 click on more than rest of the population. They have higher impression count which might lead one to question our inferences we have been making about the ad-engine, however; if one looks at the age group 54-64 which we saw comprises of people who do not click much we see that they also have high number of impressions thrown at them. One can safely conclude that the there is not pattern here and that the ad-engine does not use a robust reinforcement technique to learn from user actions

- ***What is the CTR across Age groups ? We shall also have a gender splice in the bars***

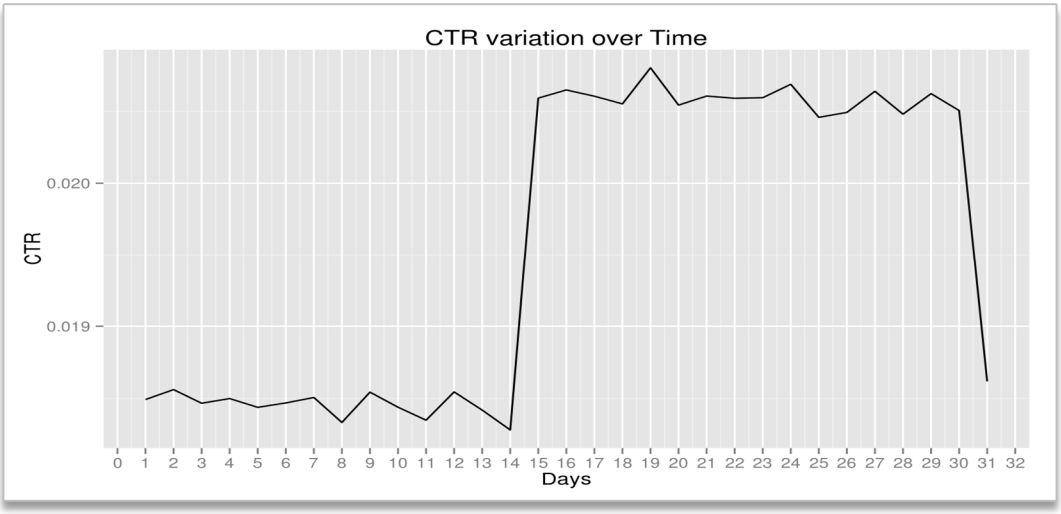



The CTR pattern here follows the Mean Clicks distribution pattern closely.

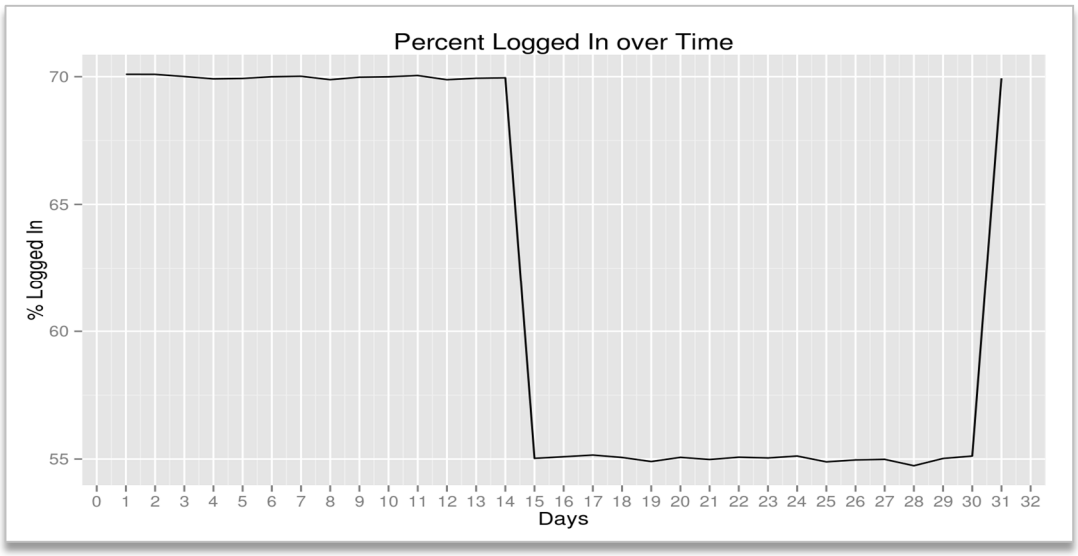
Given that we have 31 days worth of Click data. We could look at user behaviours and other metrics over a time period. Following are the graphs that we plotted to look for patterns in the data across time.

## *Time Series*

- ***What does the CTR variation look like across time ?***

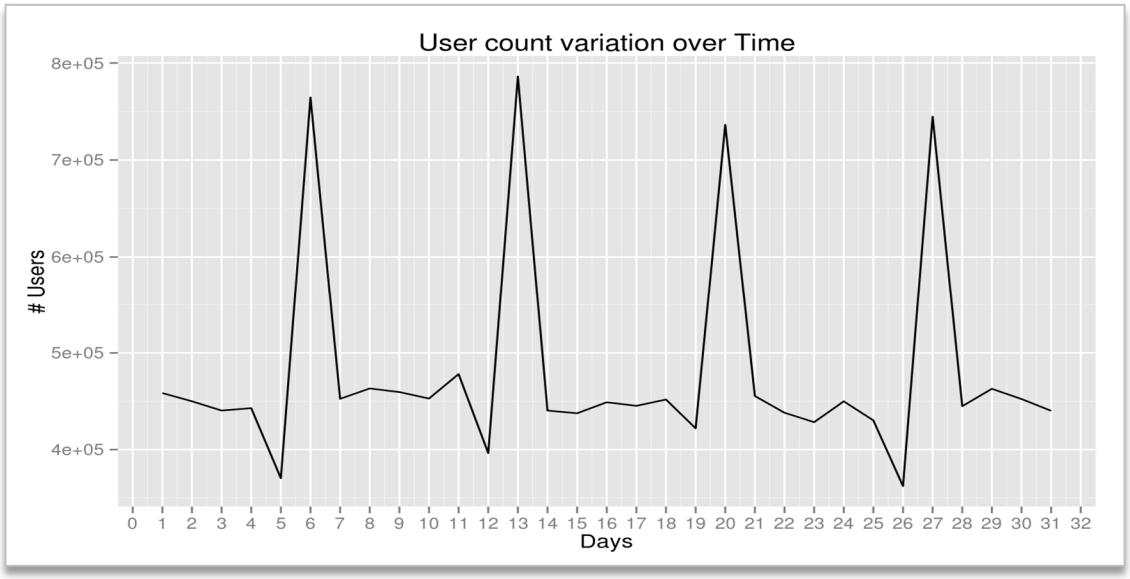


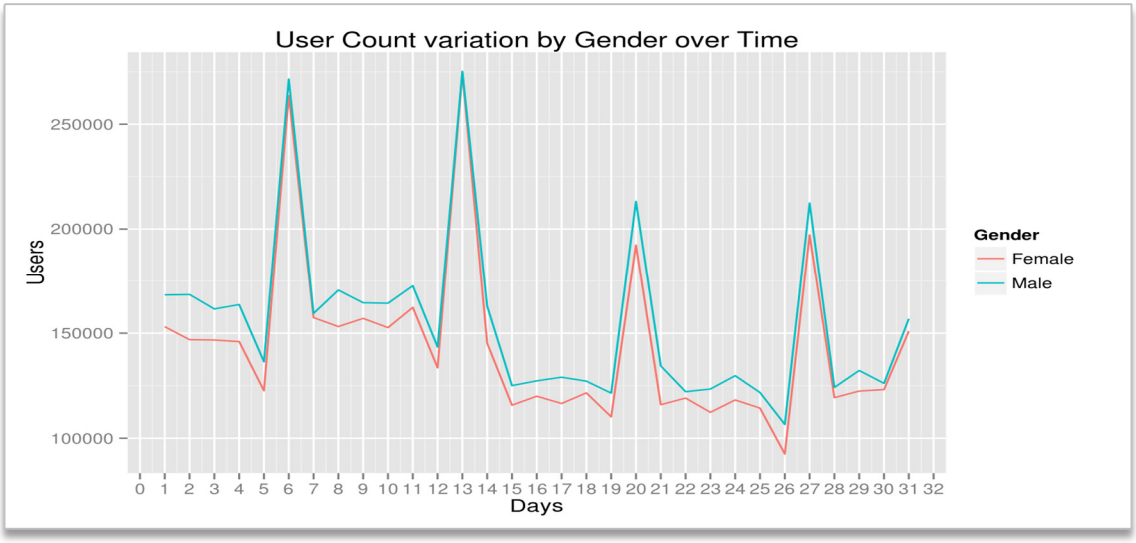

This is quite odd. The CTR climbs up after the first 14 days of the month, what could the reason be. Could this have any relation to the count of user visits to the NYT site ? or Could this have any relationship with the % of users logged in ?

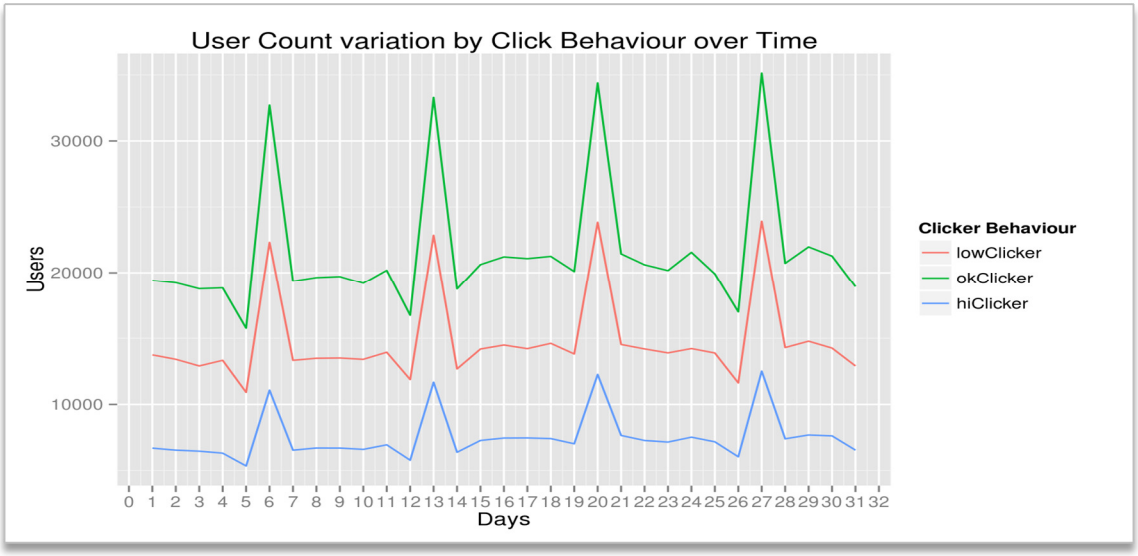



The above chart reveals the reasoning as to why the CTR climbs up after 14$^{th}$. There is clearly a correlation between CTR and the number of user logging into NYT website, in this case the correlation just happens to be negative.

- ***Is there a pattern in the user visits then? (THE NYT HEART BEAT)***


User count variation over Time
# Users
8e+05
7e+05
6e+05
5e+05
4e+05
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Days


User Count variation by Gender over Time
Users
250000
200000
150000
100000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Days
Gender
Female
Male


User Count variation by Click Behaviour over Time
Users
30000
20000
10000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Days
Clicker Behaviour
lowClicker
okClicker
hiClicker

There is clearly a pattern in terms of the number of visits on a daily basis. We see that the user visit cout metric spikes every 6th day (Staturday effect maybe). The other metrics when plotted across time also mimic the heart beat, which is along expected lines.

### *Learnings*

- ggplots
- subsetting data
- summarizing data

## REALDIRECT CASE STUDY

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?***

It is important to understand what users do when they are browsing a particular website. In our case, it would be important to observe the actions of a given user when visiting the RealDirect website in order to understand what types of homes he is looking for, how much time he spends on looking at each of the listings (linger-time).

- ***How would data be used for reporting and monitoring product usage?***

By logging data corresponding to user's requirements indirectly (i.e., via click-through and site navigation analysis), we could recommend listings that would be better suited to his needs. This data would also help us in advertising VIP listings.

We could also classify homes into categories using BHK, Zip code/Neighborhood, Area in sq. ft., Price, Build Date and Unit Category. We could then determine what category of homes a user would prefer based on his listing views.

By recording the linger-time, we could determine how interested the user was in a particular listing and use this to recommend homes later to similar users.

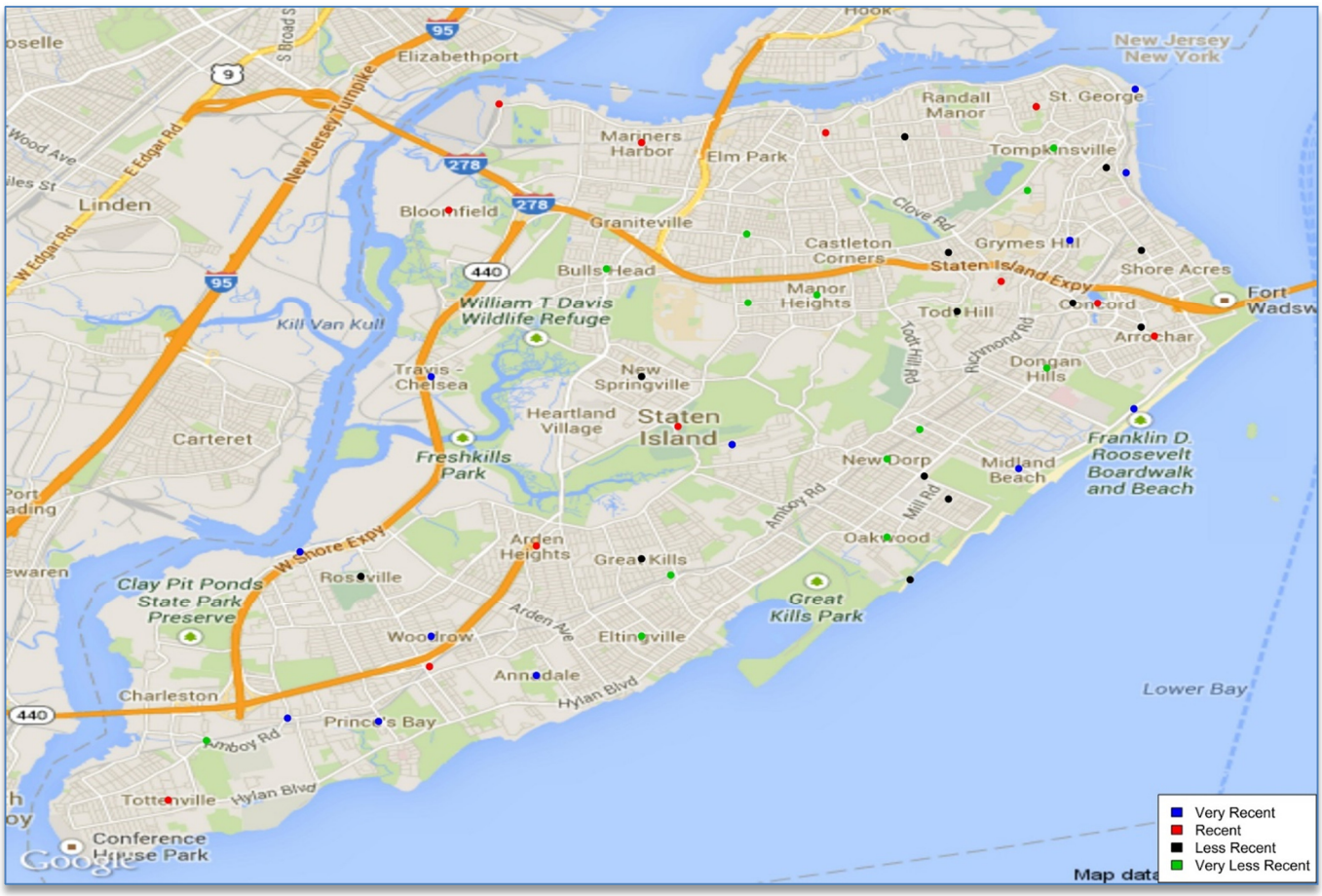
- ***How would data be built back into the product/website?***

The logged data (linger-time, listing views, etc.) can be used in personalized listing recommendation as well as use collaborative filtering to determine

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

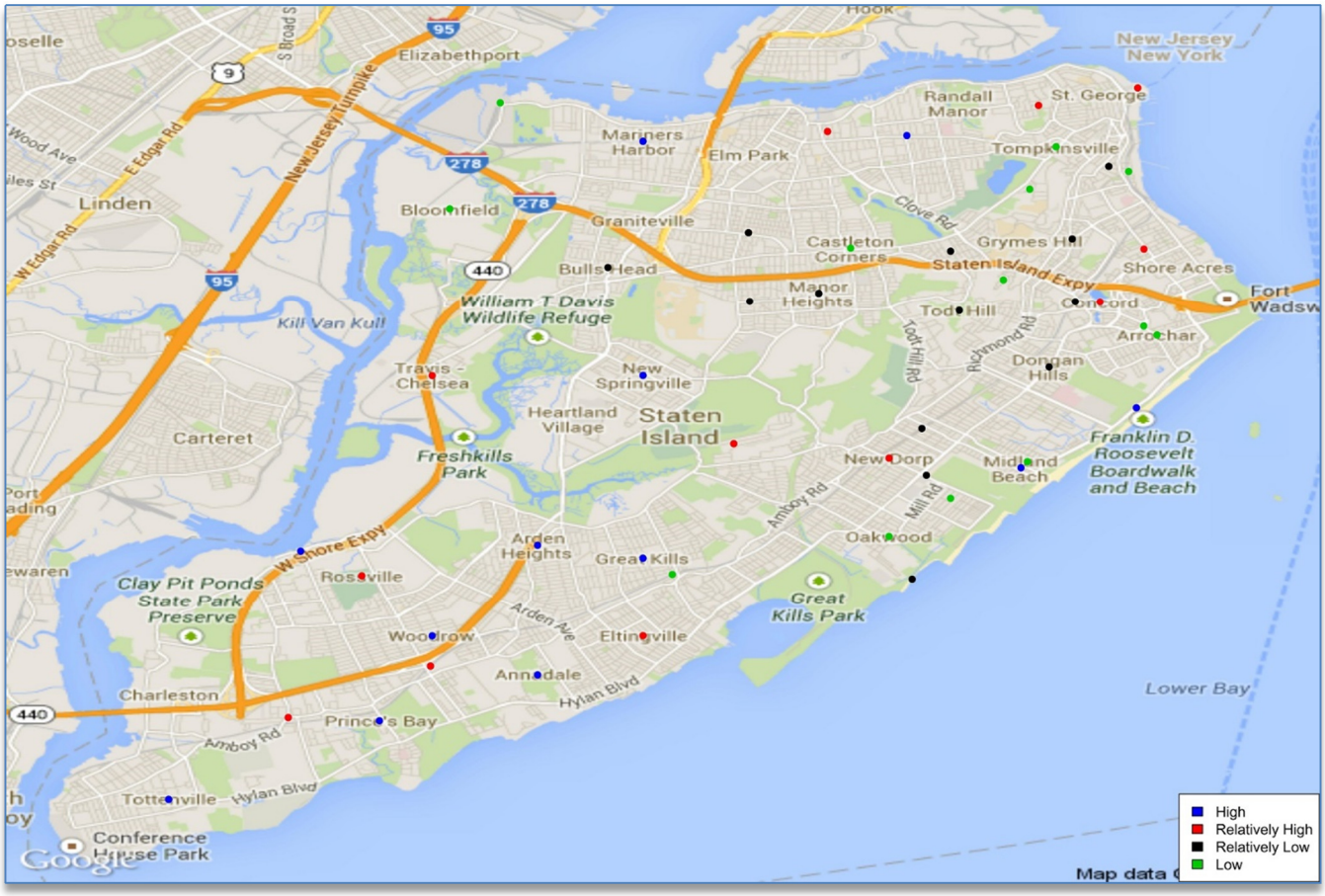
We compared the ages of buildings in the various neighborhoods by introducing a new variable, age, which is the difference between the current date and the year built. To perform this analysis, we cleaned the data by removing records which did not contain a valid value for the year built.

## *Comparison of freshness of neighborhood (Fraction of new homes in neighborhood)*



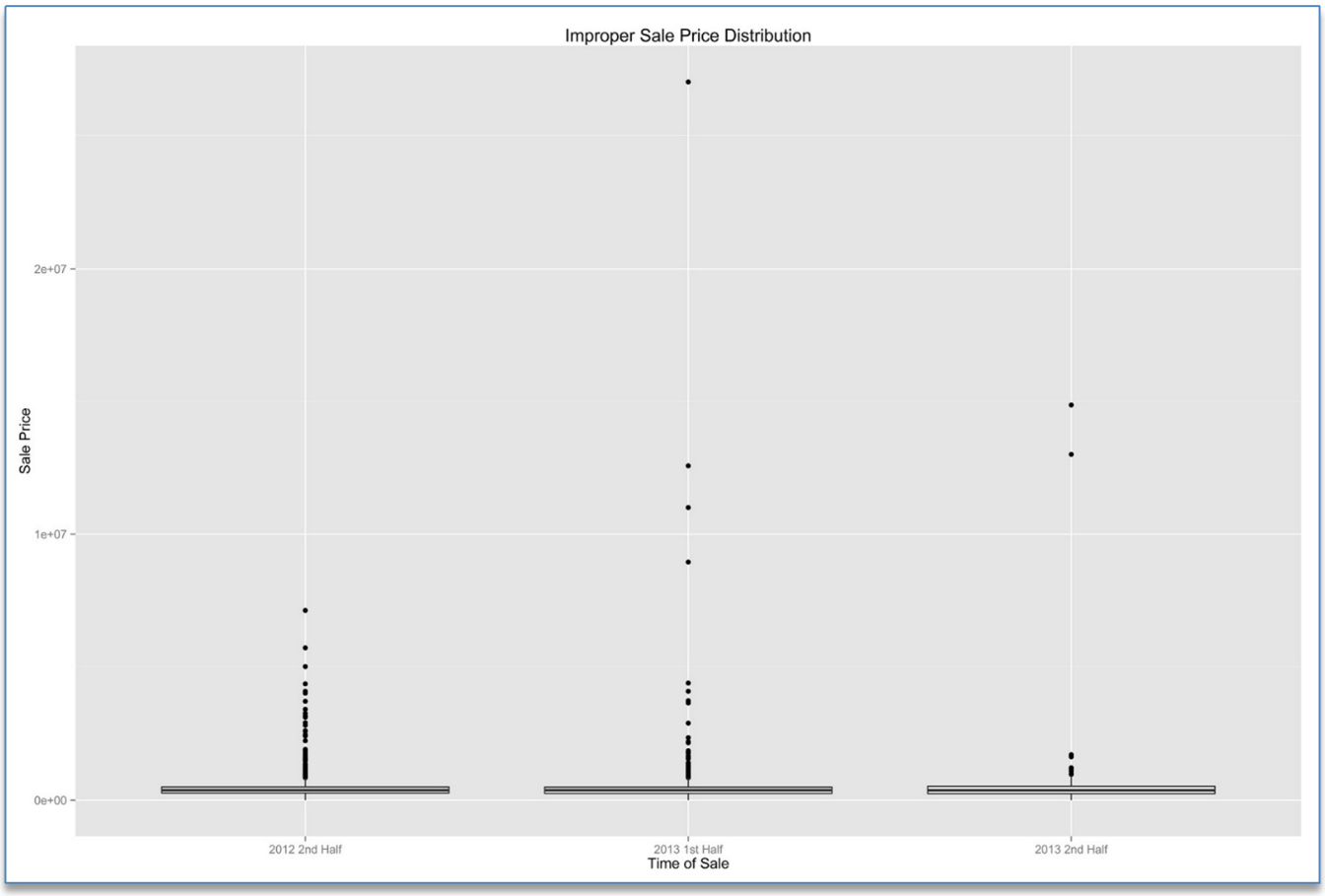

From a comparison of neighborhoods by percentage of new homes in a neighborhood, we found a pattern in the distribution of upcoming neighborhoods that was worth taking note of. Many of the upcoming neighborhoods were located either close to the beach or to the close to Route 440.

## *Comparison of neighborhoods by percentage contribution of new homes to total*



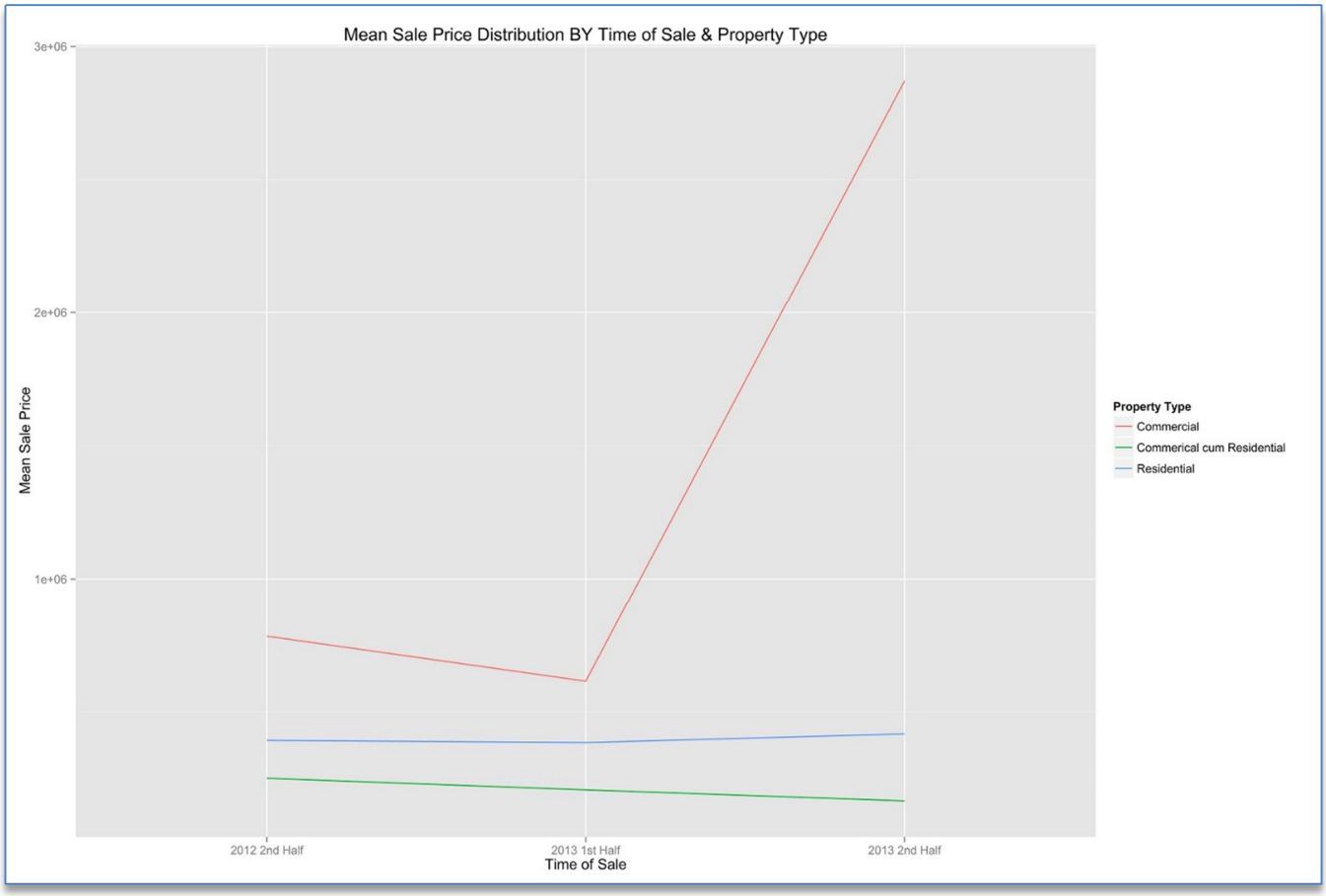

From a comparison of neighborhoods by their percentage contribution to new homes, we found that the neighborhoods with relatively low contribution were located close to Route 278. Also, many of the neighborhoods that had high contribution were located on the south end of Staten Island.

## *Improper distribution of Sale Price within each half year*




We observed that there were too many outliers in the data by generating a box plot of the sale prices grouped by the sale half year. The outliers is a manifestation of the inclusion of very expensive commercial real estate in this data along with residential real estate. Therefore, we removed the records below the 5$^{th}$ and above the 95$^{th}$ percentile and subset the data into 3 classes based on their function (which was supported by the variable “Tax Class” in the dataset).

## *Distribution of sale prices across time (half-yearly) for various property categories*




From the distribution of mean sale price across time, we observe that the mean sale price for "Residential" property remained almost constant in the period for which the data was collected. The mean sale price for "Commercial cum Residential" property decreased very slightly.

However, for "Commercial" property, we see a marginal decrease from the 2nd half of 2012 to the 1st half of 2013, but a sharp increase of close to 4.5 times from the 1st half of 2013 to the 2nd half of 2013. To interpret such findings, we would need more data.

- ***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?***

We would like to talk to potential customers so as to understand what they look for when buying homes (school district, proximity to work place and other civic amenities, etc.) so we can ideally monitor the recommendation system that RealDirect would offer to its customers.

- ***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. Working on a different domain allowed us to look closely at the process we followed while collecting data. We take some things for granted while working in a familiar domain and in the process make it difficult for people not from the same domain to understand the process. Working on RealDirect helped us realize this and ensured that we document the process steps in a detailed sequential manner.

- ***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.***

We had to look up some of the domain vocabulary on the internet. For example, we were not familiar with the existing tax classes of real estate which initially hindered our understanding of the data.

- ***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.***

Step 1 – Understand the problem

Step 2 – Collect and clean data

Step 3 – Perform EDA to understand the data and the variables that impact the business

Step 4 – Use the analysis to improve the strategy used

Perform Steps 2 through 4 iteratively.

# CENSUS DATA ANALYSIS

We chose to do out third analysis on census data. We extracted the data from University of Minnesota's *Minnesota Population Center* site. The data that we extracts that we took had the following important metrics, amongst others:

- Race
- State
- Person Income
- Family Income
- Education information
- Family Size
- Age
- Sex

The dataset has codes for all these (and more) variables, we had to create mapping datasets (in csv formats) and had them imported into R along with the actual dataset for analysis. The education code and Race codes for the census data were at a very fine level (granularity). We had mapping available to us at a broader levels as well. We chose to use broader level categories since they would present us with a reasonable bucket size per category and make the dataset amenable to analysis. For variables which did not have a broader classification category we used our discretion to create the same for those variables; education code being one.

The dataset also spanned many years, 2000 to 2010. We created subsets of the larger data to make analysis more meaningful and understand how the metrics that we were tracking change across time. We created the following time slice subsets

- **2001 – 2003 → referred to in the charts as 01-03**
- **2004 – 2007 → referred to in the charts as 04-07**
- **2008 – 2010 → referred to in the charts as 08-10**

Since the data is rich in information that shall make a social sciences major dizzy in delight. We decided to do our analysis using a social parameter, Race, as pivot variable and analyzed Education, Employment and Income distributions by Race. We later added depth to our EDA by visualizing and summarizing metrics by Regions.

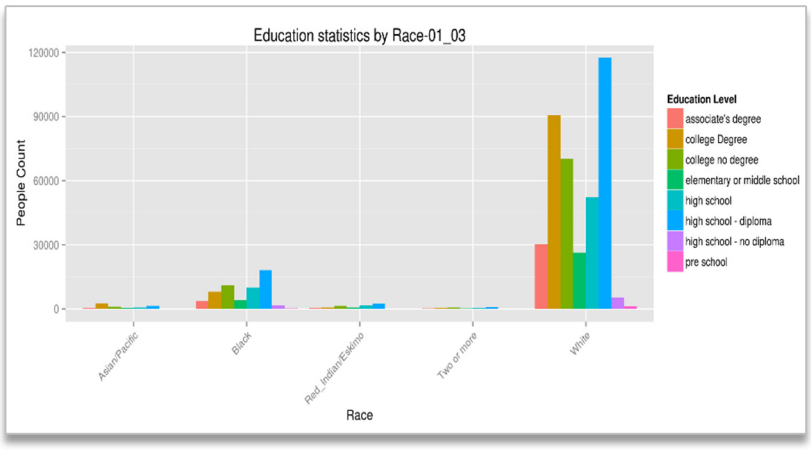
*Note for the sake of convenience while plotting we have made the following adjustments:*

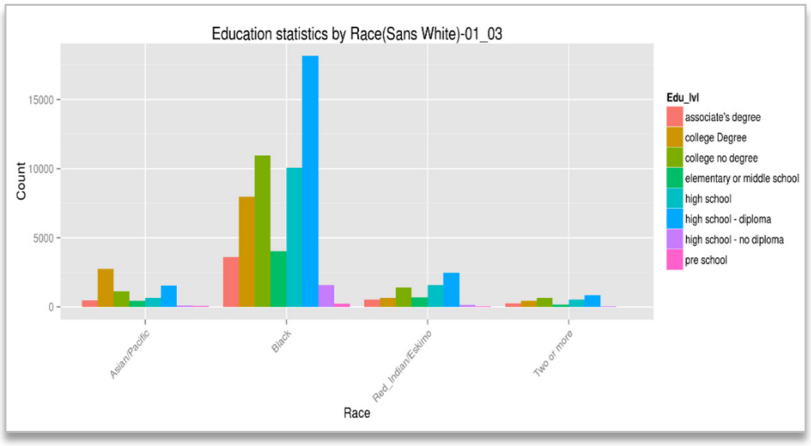
- *American Indian/aleut/Eskimo mapped to Red Indian/Eskimo and*
- *Two or more races (Mixed race) mapped to "Two or more"*

***Data Set used and the format***: Please refer to the appendix for a screen shot of the dataframe. We shall share an RDA an R data object census.rda which when loaded shall include "censusData" dataframe in the global environment. This is already being done in the R script file. Please note that the screen shot in appendix is of a dataframe that has retained only the necessary columns from the loaded rda

## *Education Analysis*

- ***Is there an obvious distribution pattern in Education levels across Race?***



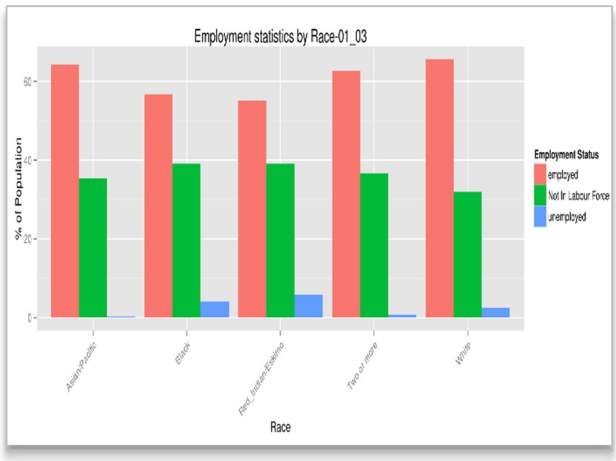

The charts above illustrate that there is indeed a pattern in the education statistics and this pattern holds true for all Races, though the count varies, this is primary due to the nature of the population. It is quite evident that a member of the Asian or White Race is more likely to complete a college degree once enrolled into college when compared to people for other races.

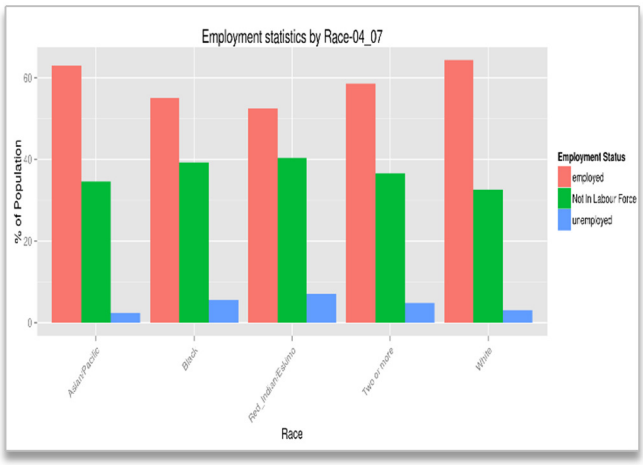
Since this chart deals with Count, an absolute number, we have these vast variation in the size of the manhattans across Race categories, it would make sense to plot the % for each of the populations, which would have made the charts even more striking and conducive for deriving insights. We make this correction when dealing with employment analysis.

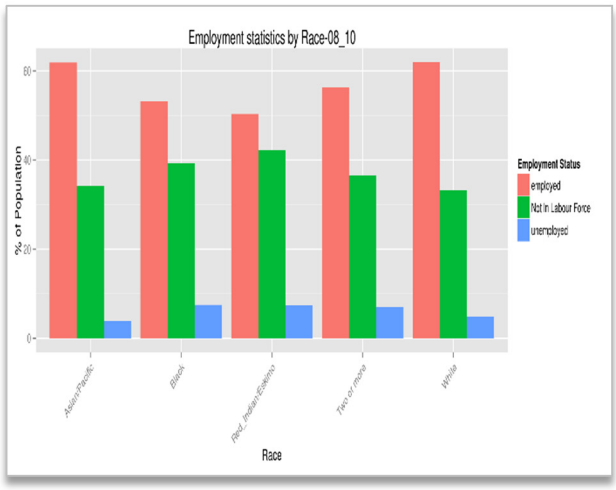
*Note: We have shared charts for rest of the year groups in the Appendix*

## *Employment Analysis*

- ***How does the distribution of employment status across Race look like?***



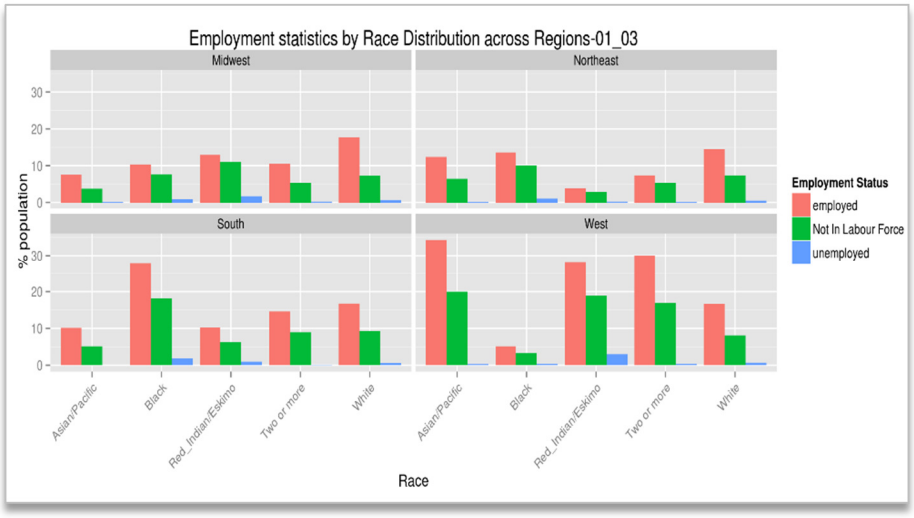



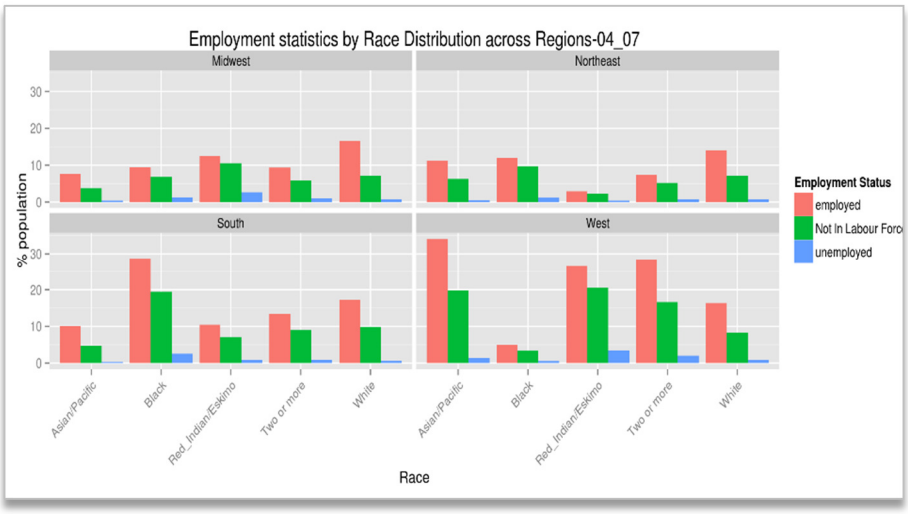

**The charts do confirm the existence of a pattern and this pattern seems to have been unaffected by time, since no variation is displayed between the distributions of two time splices.**
**These charts do not have absolute counts but rather percentages. This makes for better comparison across race categories; absolute count would have made the charts skew the scales.**

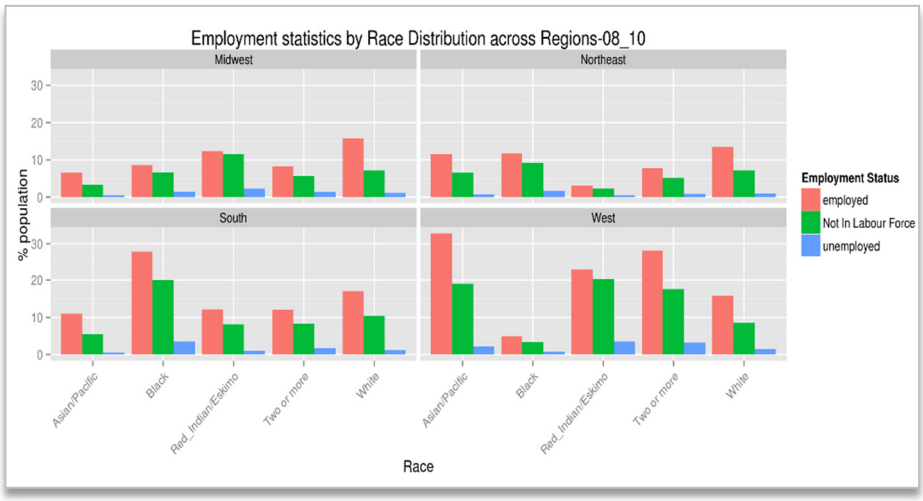
- ***How are the employment numbers for each of the race categories distributed across the major regions in the United States; namely South, North-East, West and Mid-West?***

In the charts that follow, we have display the percentage contribution of each Region to the absolute count of the employment status categories for each of the Race categories. For example if we were to choose the Race category Asians/Pacific and wanted to see which regions contributes more to the overall employment statistics for let's say the category "Employed", then the following charts would be able to answer the question. Now that we have explained how to read the chart, we present the charts below.

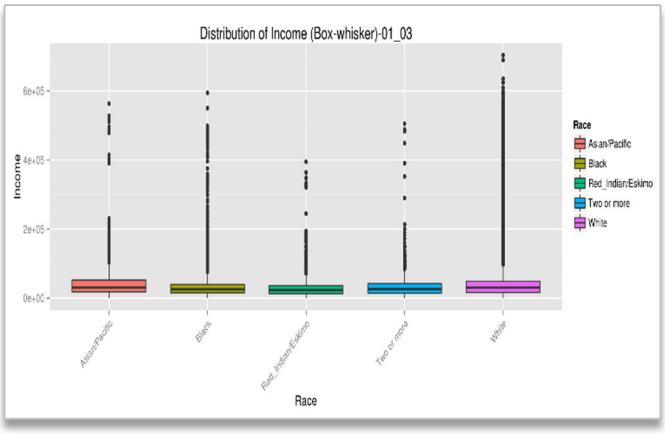
It is clear from the charts if were to use our example in the previous paragraph, we can see that as far as the Asians/Pacific category is concerned the "West" region contributes more to the employment statues "employed" and "unemployed"; this could be because of greater numbers. One could similarly draw insights for other race categories as well.

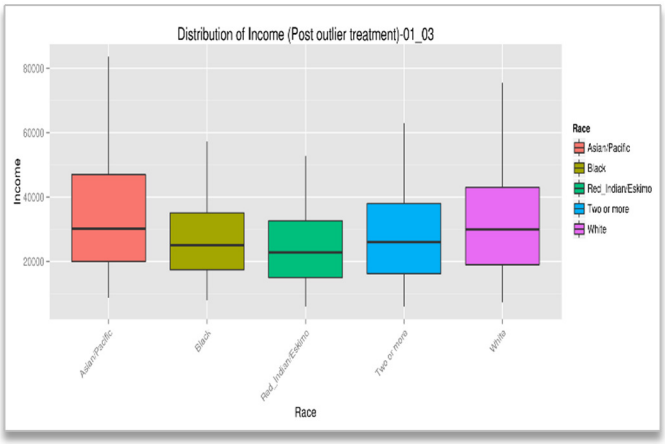

Employment statistics by Race Distribution across Regions-01_03
Midwest
Northeast
South
West
% population
0
10
20
30
Asian/Pacific
Black
Red_Indian/Eskimo
Two or more
White
Race
Employment Status
employed
Not In Labour Force
unemployed


Employment statistics by Race Distribution across Regions-04_07
Midwest
Northeast
South
West
% population
0
10
20
30
Asian/Pacific
Black
Red_Indian/Eskimo
Two or more
White
Race
Employment Status
employed
Not In Labour Force
unemployed


Employment statistics by Race Distribution across Regions-08_10
Midwest
Northeast
South
West
% population
0
10
20
30
Asian/Pacific
Black
Red_Indian/Eskimo
Two or more
White
Race
Employment Status
employed
Not In Labour Force
unemployed

# *Personal Income Analysis*



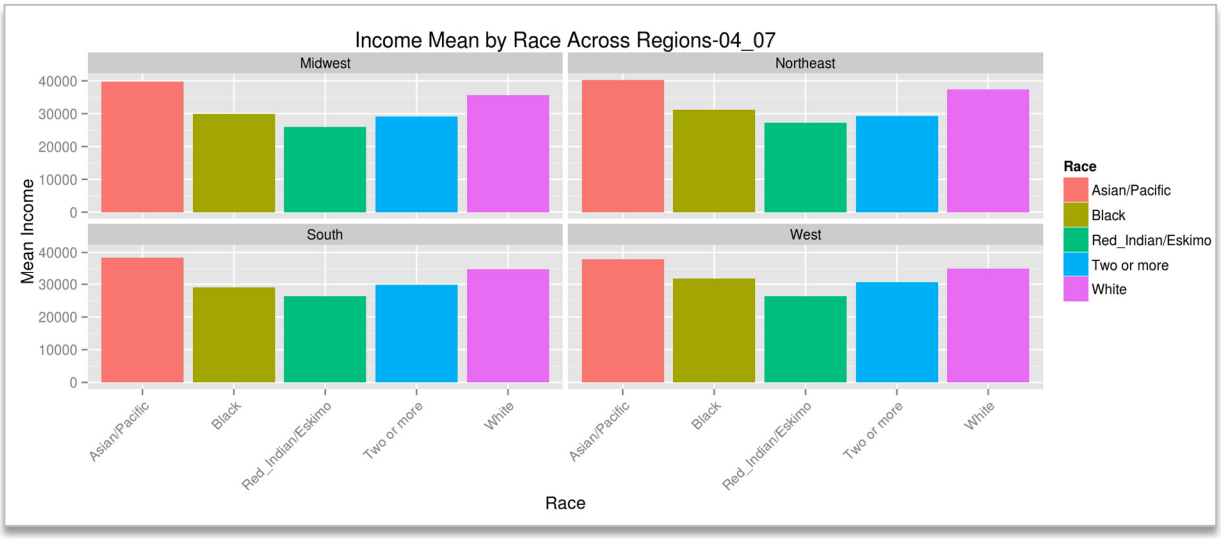




For Income analysis across race categories we chose to plot the mean income across race categories. Prior to the analysis we wanted to see the distribution of income in each category. The plots to the left are the box-whisker for all categories.

It is quite evident from the plots that the data had a lot of outliers. We had to do outlier treatment prior to plotting and analyzing mean Income by Race category.

We purged the records which were less than 10$^{th}$ percentile and above 90$^{th}$ percentile. The resulting data had a box-whisker plot which show that the data was not good for income analysis

➢ ***What is the pattern of mean income by Race across regions?***





From the charts above we see that there is a pattern in the data and this pattern in consistent across all regions of the United States i.e. on an average the Asian and White communities have higher income. Please note that the scales of the "Mean Income" vary across regions and across time.

## *K Means Clustering*

In order to perform K Means Clustering, we selected the three continuous variables that we had in the dataset, Family Income, Family Size and Age of the Family Head. The dataset was cleaned to remove missing values and NA values. The processed dataset was then scaled using the scale(). We ran k-means for values of k between 2 and 15 and plotted the Within Group Sum of Squares against the Number of Clusters. Bayesian Information Criterion (BIC) was used to modify the standard k-means penalty which generated kinks in the plots, indicating the optimal value of k to be used for k means clustering. K-means was run again on the optimal value, 4 in our case, and the fitted model was used to map the clusters

obtained back to the original unscaled dataset. We then ran summary statistics on this dataset grouping people by the state and the cluster they belonged to. We also ran summary statistics by grouping people by the county and the cluster they belonged to. Voting was used to determine the cluster that dominated a particular state or county and the whole state or county was represented using that cluster. These cluster were then plotted as a heat map on the map of USA.

## *Results*

Cluster centroids from k-means for the years 2001-2003

| Family Income | Age | Family Size | Cluster |
|---|---|---|---|
| 38253.57 | 34.10234 | 1.850043 | 1 |
| 43144.82 | 62.44876 | 1.772033 | 2 |
| 71154.19 | 40.74415 | 4.366625 | 3 |
| 339012.54 | 46.79298 | 3.330748 | 4 |

Cluster centroids from k-means for the years 2004-2007

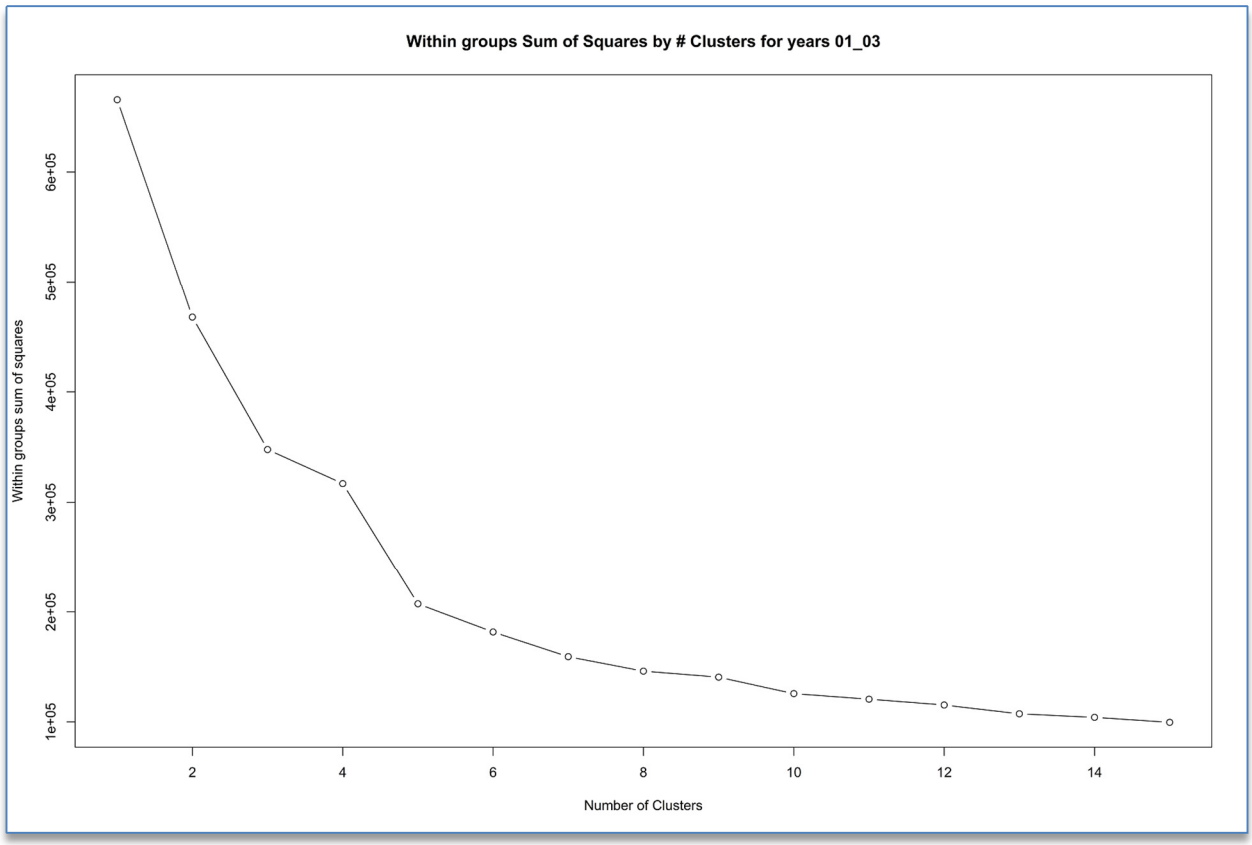
| Family Income | Age | Family Size | Cluster |
|---|---|---|---|
| 40227.14 | 33.72858 | 1.862199 | 1 |
| 49453.8 | 61.76757 | 1.763398 | 2 |
| 80283.86 | 41.16746 | 4.363489 | 3 |
| 420240.76 | 47.64088 | 3.354053 | 4 |

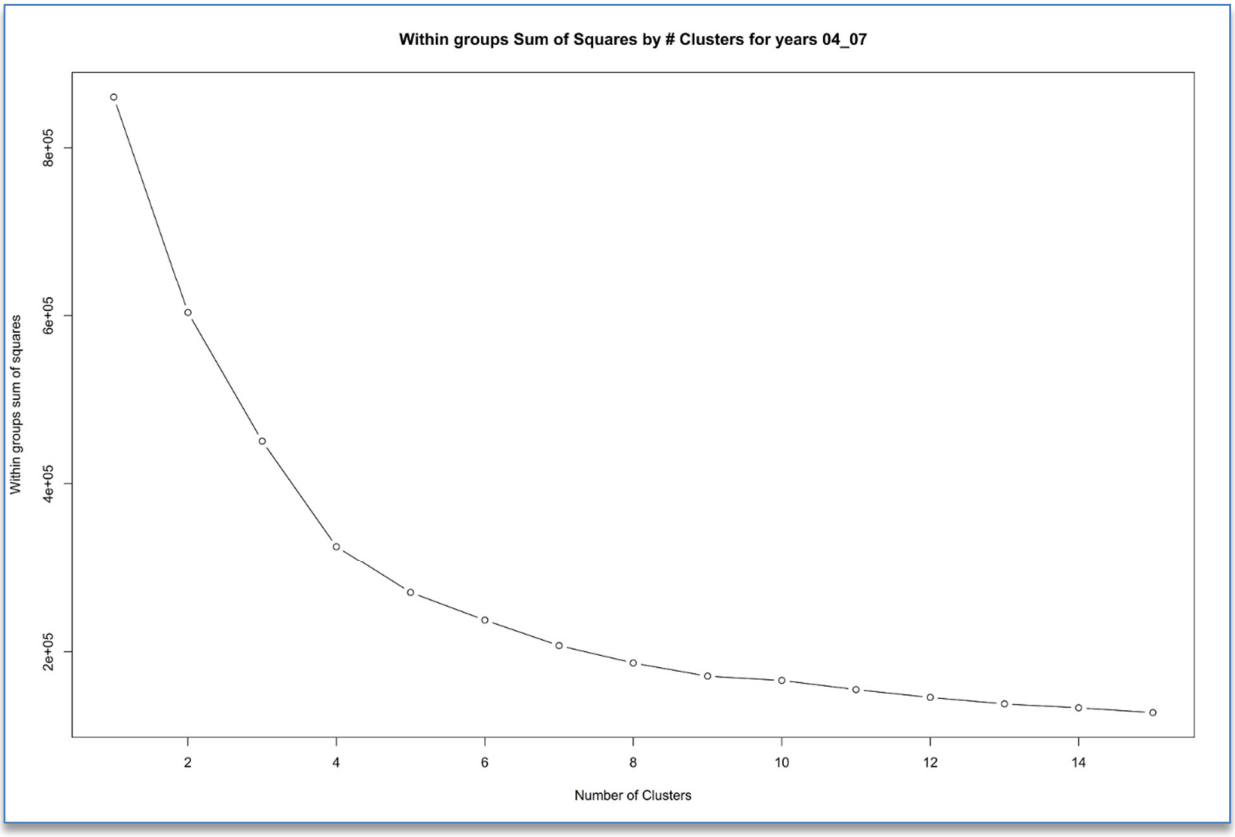
Cluster centroids from k-means for the years 2008-2010

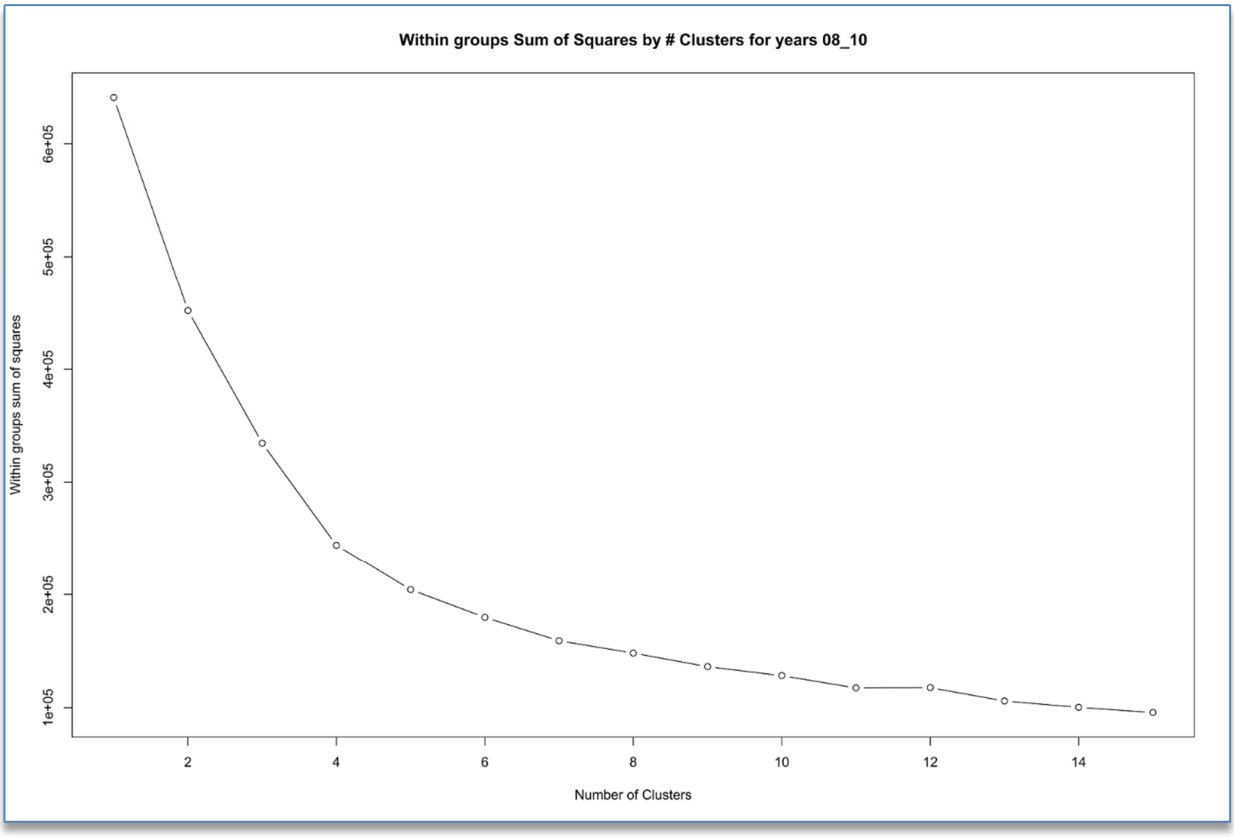
| Family Income | Age | Family Size | Cluster |
|---|---|---|---|
| 44072.11 | 33.51828 | 1.85295 | 1 |
| 54250.45 | 61.56577 | 1.758215 | 2 |
| 86068.9 | 41.66696 | 4.403678 | 3 |
| 393765.15 | 48.9878 | 3.312415 | 4 |

## *Within Group Sum of Squares for k between 2 and 15*

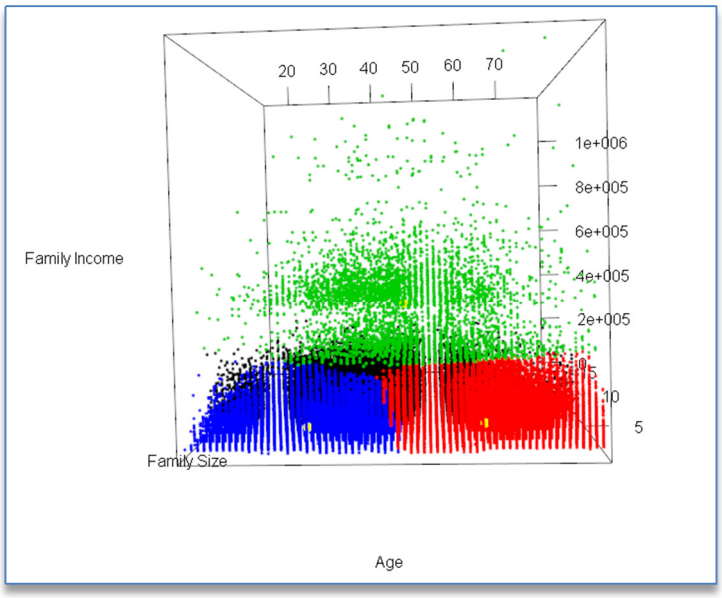
The Within Group Sum of Squares allowed us to visually determine that the optimal value for k was 4. We also tried other values of k neighboring 4 such as 2, 3, 5 and 6. As expected, we got poor results for those values of k. The results shown above also show that the clustering generates very intuitive and discriminative clusters of families by their income, size and age of the head of family.

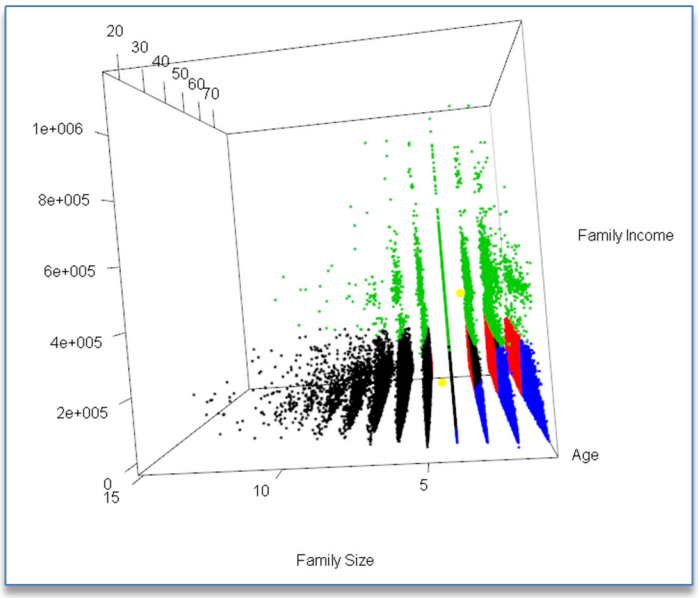


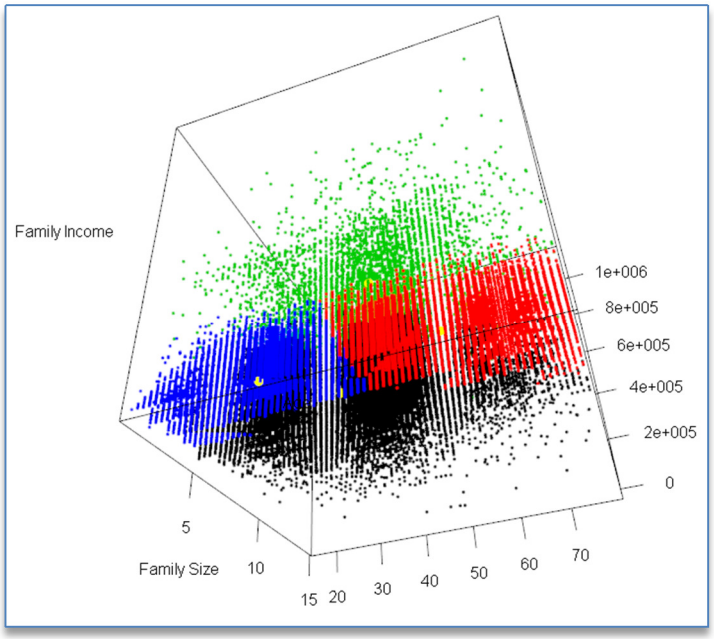

Within groups Sum of Squares by # Clusters for years 04_07
Within groups sum of squares
2e+05
4e+05
6e+05
8e+05
2
4
6
8
10
12
14
Number of Clusters

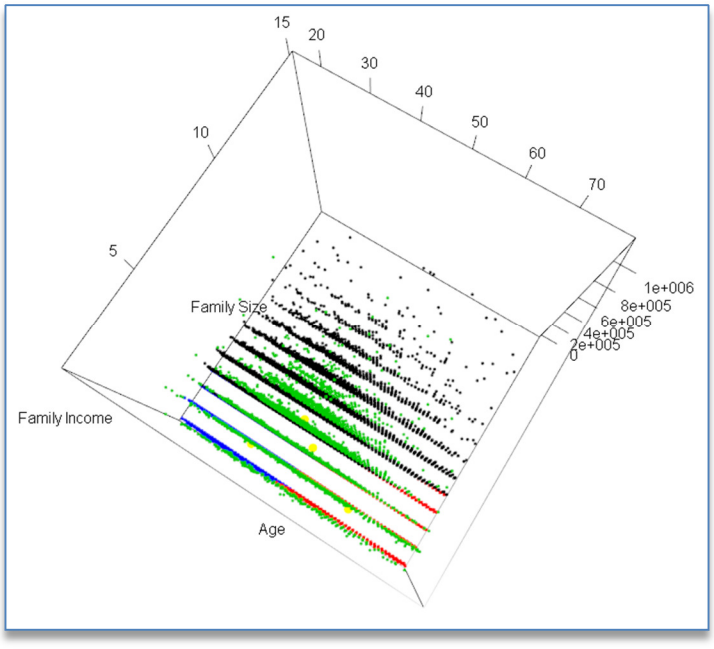

Within groups Sum of Squares by # Clusters for years 08_10
Within groups sum of squares
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
2
4
6
8
10
12
14
Number of Clusters

## *3 Dimensional Visualization of the Clusters for 08-10*

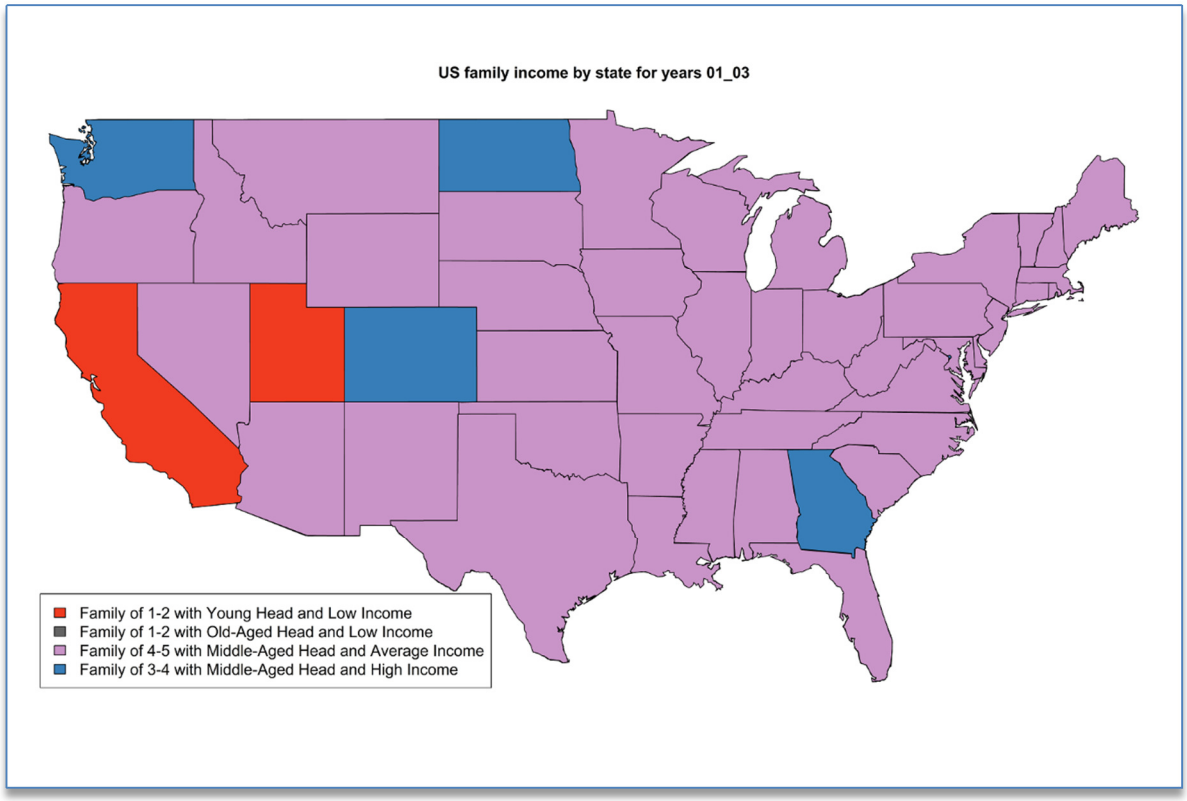


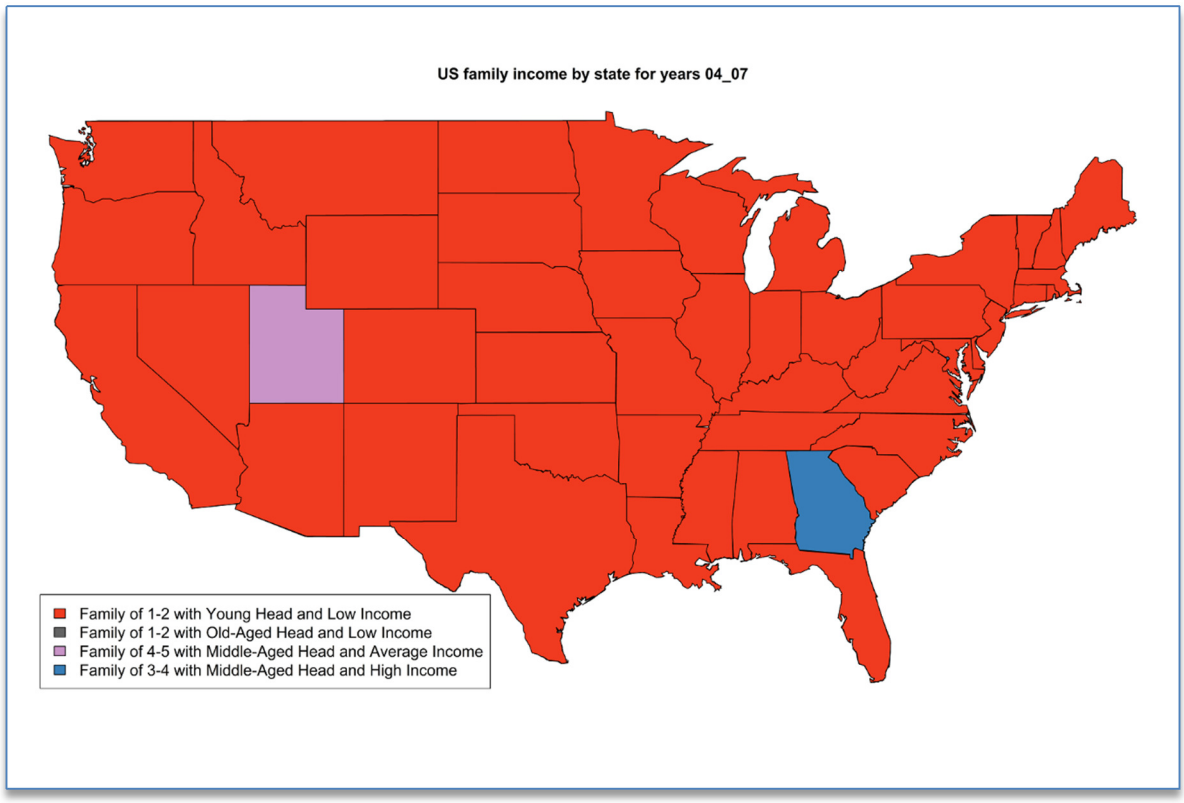



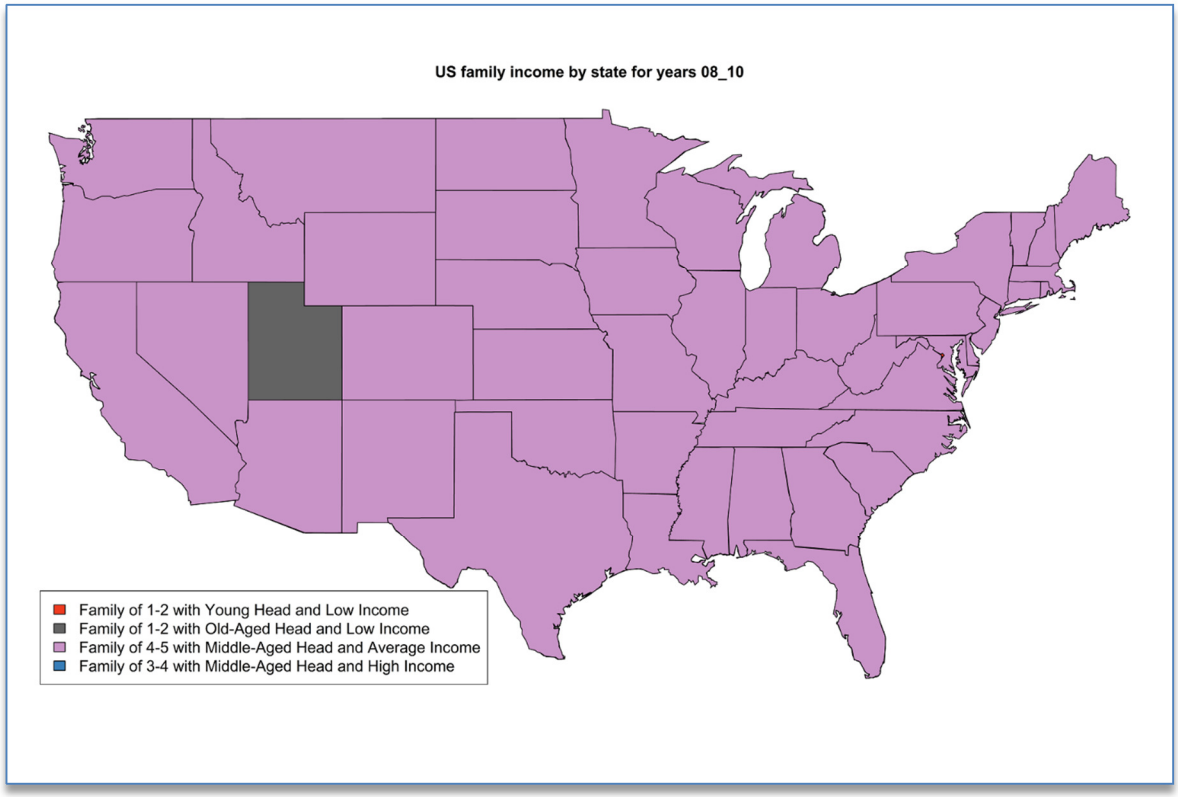



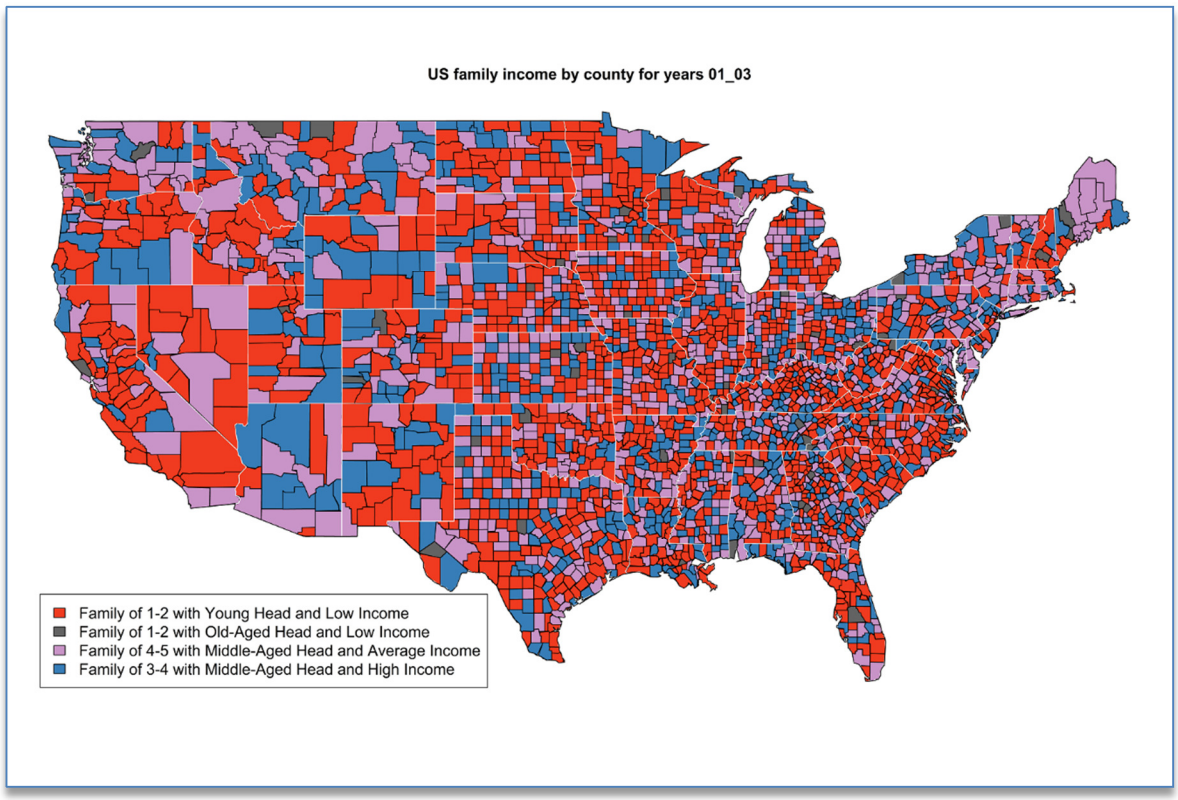



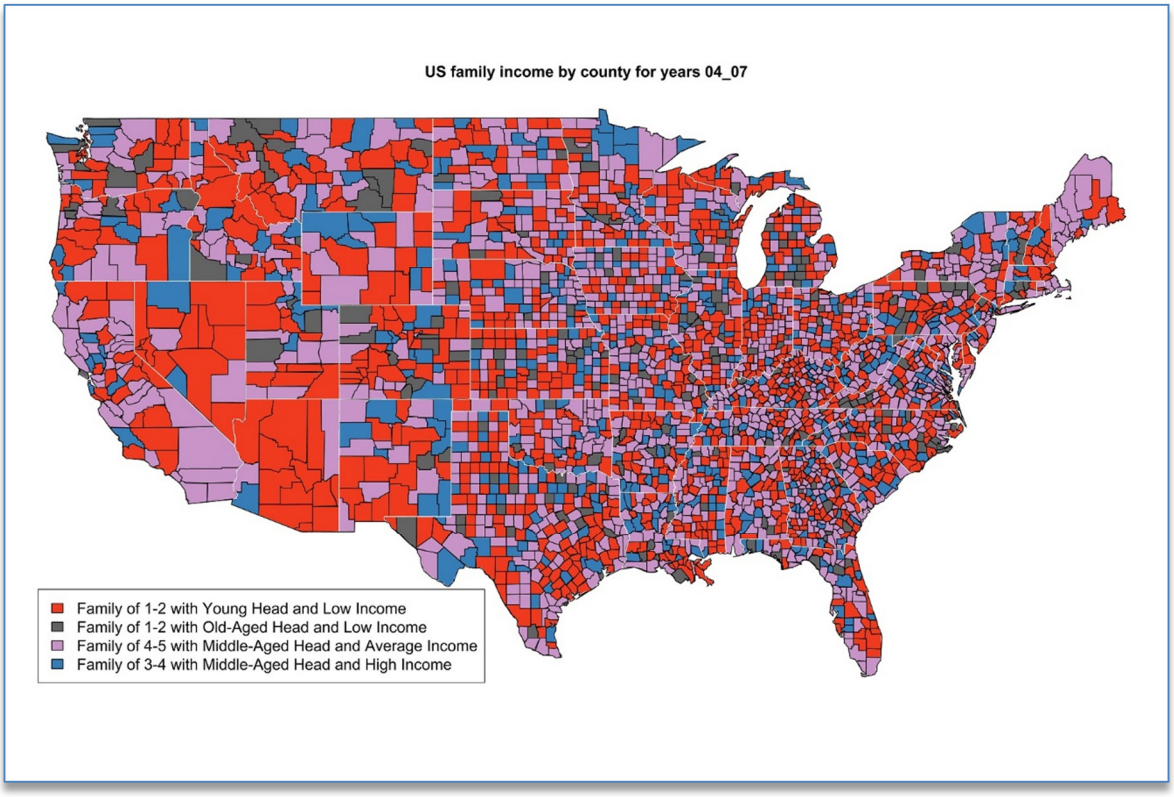
## *Heat maps of family income by state/county*

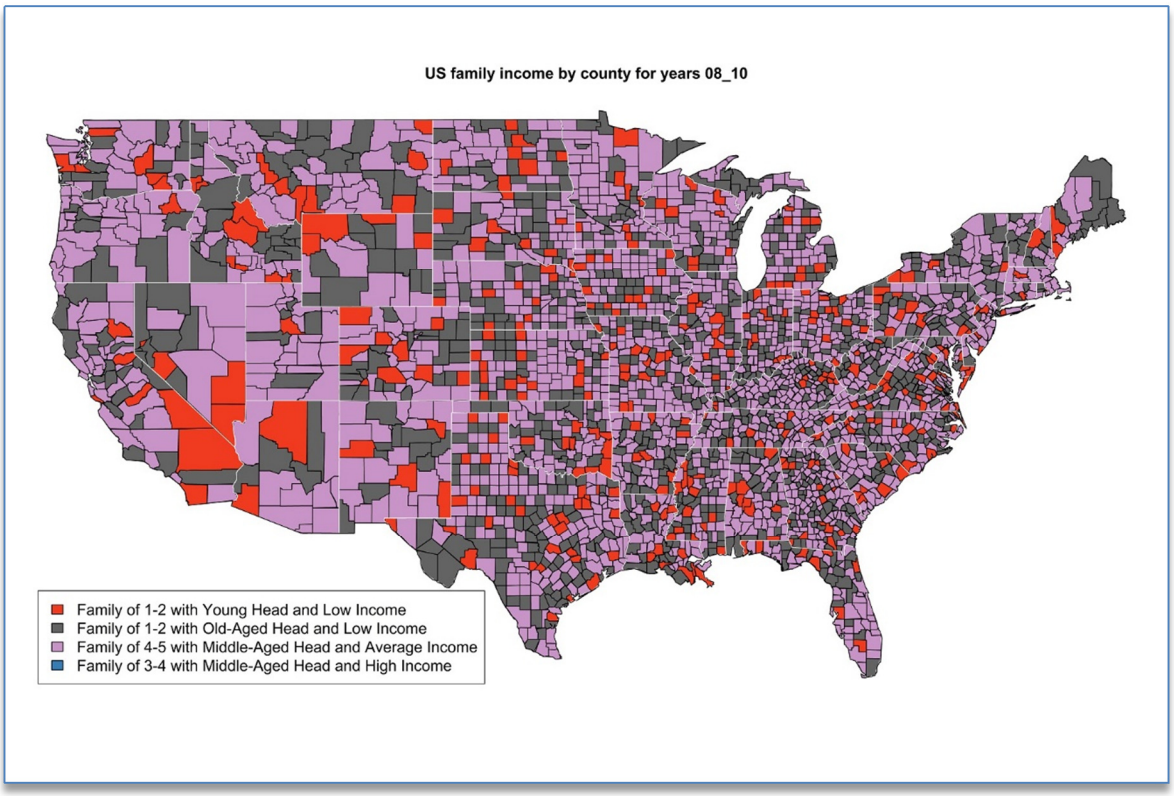



US family income by state for years 08_10
Family of 1-2 with Young Head and Low Income
Family of 1-2 with Old-Aged Head and Low Income
Family of 4-5 with Middle-Aged Head and Average Income
Family of 3-4 with Middle-Aged Head and High Income


US family income by county for years 01_03
Family of 1-2 with Young Head and Low Income
Family of 1-2 with Old-Aged Head and Low Income
Family of 4-5 with Middle-Aged Head and Average Income
Family of 3-4 with Middle-Aged Head and High Income


US family income by county for years 04_07
Family of 1-2 with Young Head and Low Income
Family of 1-2 with Old-Aged Head and Low Income
Family of 4-5 with Middle-Aged Head and Average Income
Family of 3-4 with Middle-Aged Head and High Income


US family income by county for years 08_10
Family of 1-2 with Young Head and Low Income
Family of 1-2 with Old-Aged Head and Low Income
Family of 4-5 with Middle-Aged Head and Average Income
Family of 3-4 with Middle-Aged Head and High Income

For the years of 2001-2003, we observe that most counties are dominated by families of 1 or 2 people with low income having a young family head. A prominent portion of the counties is also dominated by middle-aged families with average income.

For the years of 2004-2007, we observe that most counties are dominated by families of 4 or 5 people with average income having a middle-aged family head. However, a prominent portion of the counties are also dominated by the young families.

For the years of 2008-2010, we observe that most counties are dominated by families of 4 or 5 people with average income having a middle-aged family head. Many counties are dominated by families of 1 or 2 people with low income having an old-aged family head.

The reasons for the observations could be one or more of the following:

- Some of the data does not have county and state mappings.
- External factors might have influenced the distribution of clusters. Some of the factors could be:

***Baby Boomer phenomenon*** – The birth rate in USA has declined over the years and was at its peak from 1947-1964. People born during that period would have been around 50-60 years old during the census of 2008 to 2010. Since there were fewer births post 1964, the population of young people would have been lesser than the population of the older, retired people who seem to dominate most counties in 2008-2010.

***Economic Recession of 2001 and 2008*** – The economic recession of 2001 brought down the employment rates. This could be one of the reasons for the domination of young people with low income during 2001-2003. There was an increase in the US employment ratio from 2004-2007. This might be a reason for the reduction in the number of counties dominated by youth with low income from the 2001-2003 period to the 2004-2007 period.

Due to the economic recession of 2008, many of the low income category people lost their jobs. The major hit would have been taken by the young people with low income (colored red). The older people were aged 60 or above and would have retired by then and living on pension justifying their low income. People without income do not feature on the plot and hence we observed a domination of older families for the years 2008-2010.

***Migration of people over the years due to various factors*** – Mass migration of people due to various reasons could have shifted the dominating clusters from one county to another or possibly to another state as well accounting to the change in the distribution of the clusters.

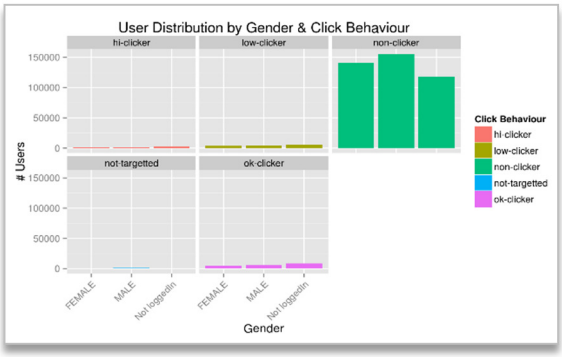
## *Learnings*

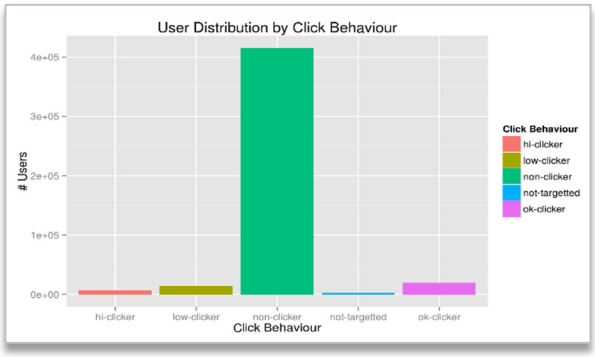
- Plotting
- Analysis
- Outlier treatment

- Kmeans
- Maps

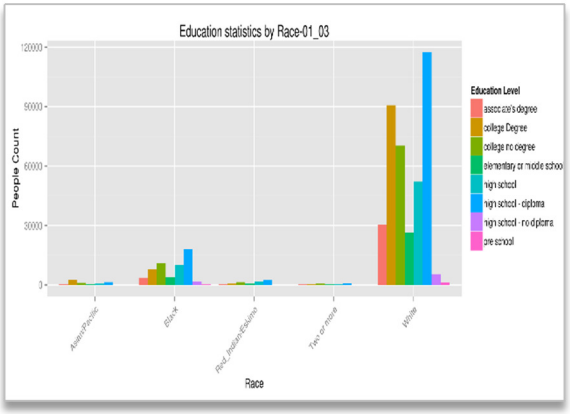
# APPENDIX

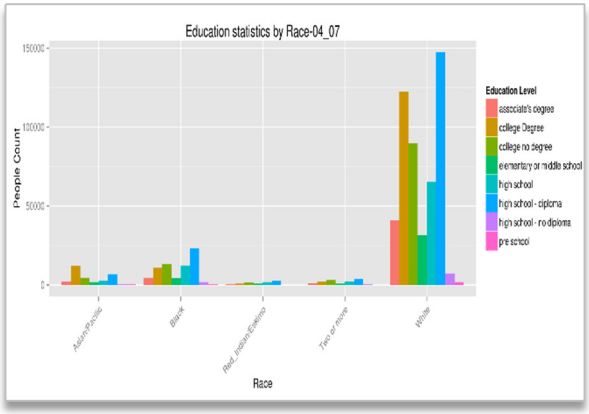
## *NEW YORK TIMES ANALYSIS*

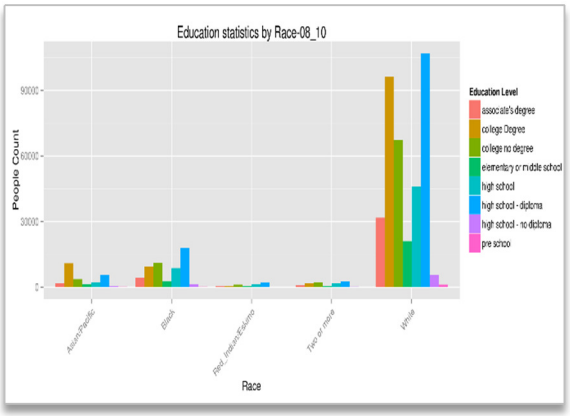


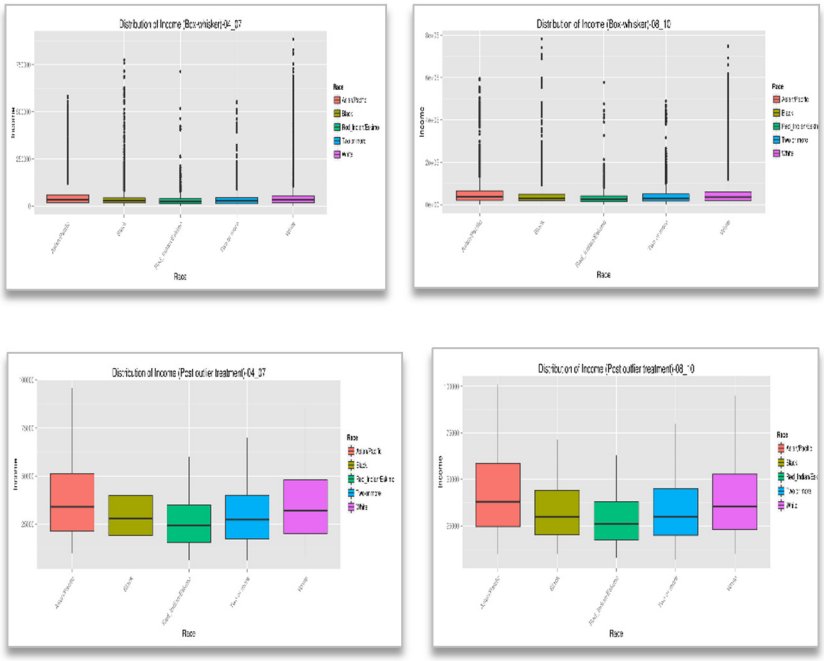




## *CENSUS DATA EDA*





## *CENSUS DATAFRAME SCREEN SHOT*

censusData_EDA.withVarOfInterest

| | yr | stateCode | age | sex | race | citizen_status | nativity | edu_code | emp_status | fam_income | pers_income |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2008 | 23 | 55 | 2 | 100 | 0 | 1 | 81 | 10 | 142162 | 53106 |
| 2 | 2008 | 23 | 58 | 1 | 100 | 0 | 1 | 123 | 10 | 142162 | 89056 |
| 3 | 2008 | 23 | 34 | 2 | 100 | 0 | 1 | 123 | 10 | 90141 | 36651 |
| 4 | 2008 | 23 | 37 | 1 | 100 | 0 | 1 | 123 | 10 | 90141 | 53490 |
| 5 | 2008 | 23 | 3 | 2 | 100 | 0 | 1 | 1 | 0 | 90141 | 99999999 |
| 6 | 2008 | 23 | 1 | 1 | 100 | 0 | 1 | 1 | 0 | 90141 | 99999999 |
| 7 | 2008 | 23 | 37 | 1 | 100 | 0 | 1 | 91 | 10 | 81750 | 35412 |
| 8 | 2008 | 23 | 40 | 2 | 100 | 3 | 5 | 81 | 10 | 81750 | 46338 |
| 9 | 2008 | 23 | 8 | 2 | 100 | 0 | 1 | 1 | 0 | 81750 | 99999999 |
| 10 | 2008 | 23 | 80 | 2 | 100 | 0 | 1 | 73 | 30 | 17923 | 17923 |
| 11 | 2008 | 23 | 34 | 2 | 100 | 0 | 1 | 50 | 30 | 7716 | 0 |
| 12 | 2008 | 23 | 37 | 1 | 100 | 0 | 1 | 60 | 32 | 7716 | 7716 |
| 13 | 2008 | 23 | 14 | 2 | 100 | 0 | 1 | 1 | 0 | 7716 | 99999999 |
| 14 | 2008 | 23 | 8 | 2 | 100 | 0 | 1 | 1 | 0 | 7716 | 99999999 |
| 15 | 2008 | 23 | 41 | 1 | 100 | 0 | 1 | 81 | 10 | 96502 | 91502 |
| 16 | 2008 | 23 | 41 | 2 | 100 | 0 | 1 | 81 | 12 | 96502 | 5000 |
| 17 | 2008 | 23 | 11 | 2 | 100 | 0 | 1 | 1 | 0 | 96502 | 99999999 |
| 18 | 2008 | 23 | 9 | 1 | 100 | 0 | 1 | 1 | 0 | 96502 | 99999999 |
| 19 | 2008 | 23 | 47 | 1 | 100 | 0 | 1 | 73 | 10 | 85000 | 50000 |
| 20 | 2008 | 23 | 41 | 2 | 100 | 0 | 1 | 91 | 10 | 85000 | 30000 |
| 21 | 2008 | 23 | 20 | 1 | 100 | 0 | 1 | 73 | 30 | 85000 | 5000 |
| 22 | 2008 | 23 | 10 | 1 | 200 | 0 | 1 | 1 | 0 | 85000 | 99999999 |
| 23 | 2008 | 23 | 6 | 2 | 200 | 0 | 1 | 1 | 0 | 85000 | 99999999 |
| 24 | 2008 | 23 | 39 | 1 | 100 | 0 | 1 | 81 | 10 | 95450 | 38650 |
| 25 | 2008 | 23 | 39 | 2 | 100 | 0 | 1 | 73 | 10 | 95450 | 56400 |
| 26 | 2008 | 23 | 16 | 2 | 100 | 0 | 1 | 40 | 10 | 95450 | 400 |
| 27 | 2008 | 23 | 80 | 1 | 100 | 0 | 1 | 73 | 30 | 40462 | 33173 |
| 28 | 2008 | 23 | 77 | 2 | 100 | 0 | 1 | 73 | 30 | 40462 | 7289 |
| 29 | 2008 | 23 | 60 | 1 | 100 | 0 | 1 | 92 | 12 | 21350 | 21350 |
| 30 | 2008 | 23 | 65 | 2 | 100 | 0 | 1 | 81 | 10 | 25200 | 25200 |

# REFERENCES

Census Dataset - https://cps.ipums.org/cps/

Baby Boomer Effect - http://en.wikipedia.org/wiki/Baby_boomers

Recession of 2008 - http://en.wikipedia.org/wiki/Great_Recession

US Economy - http://en.wikipedia.org/wiki/Economy_of_the_United_States

http://www.statmethods.net/

http://www.r-bloggers.com/

http://stackoverflow.com/

http://stats.stackexchange.com/

http://blog.revolutionanalytics.com/

http://www.molecularecologist.com/2012/09/making-maps-with-r/

http://stat.ethz.ch/

http://www.inside-r.org/