CSE 487/587 Project 1 : R

USA College Crime Analysis 2001- 2013

Authors

✓ Palaniappan Meiyappan

- UB# 50097597 email : palaniap@buffalo.edu

✓ Arun Nagendran A M

- UB# 50096549 email : arunnage@buffalo.edu

Course

■ CSE 587 of State University of New York, University at Buffalo

Abstract

Using R to perform Exploratory Data Analysis on the National Campus Safety and Security Data set provided by US Department of Education, Office of Postsecondary Education. We analyse the data given with the aim of aggregating campus safety measures state wise and sector wise.

Project Objective

1. Lay foundations to learning the area of Exploratory Data Analysis by working on data of a National Scale.
2. Familiarise ourselves with R.
3. Use R to visualise patterns on this data.

Project Approach

We start with understanding in theory, the idea behind exploratory data analysis. Chapter 2 in Doing Data Science[(1)] is a great place to start. Additional information of interest is also available from Wikipedia[(2)], Engineering Statistics Handbook[(3)]. Apart from understanding the idea in theory, trying out the examples problems given in the Doing Data Science book (New York Times Data Set and RealDirect data set) using R studio helps in understanding the use of aggregate functions, filtering functions, clustering functions and to visualise data distribution using graphs and plots.

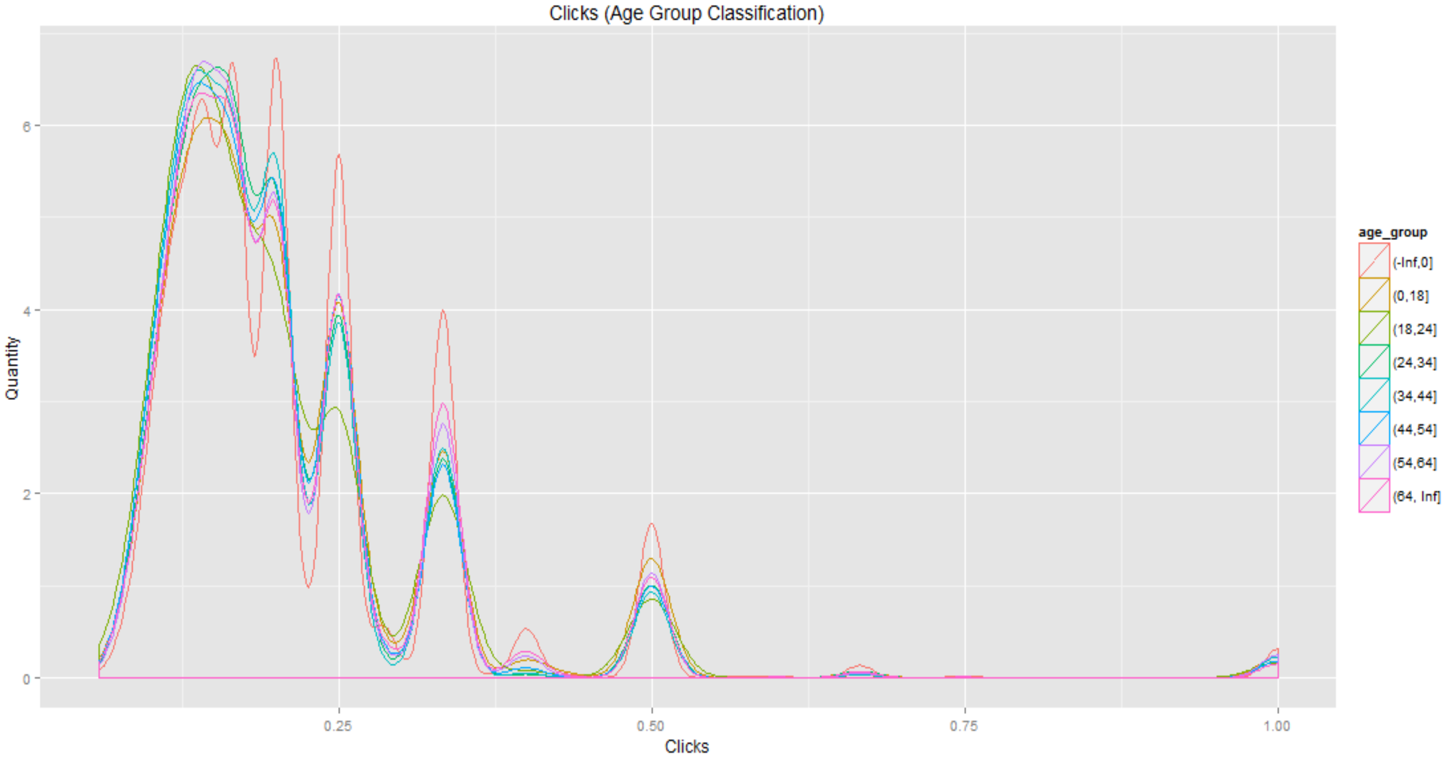
Once we are familiar with the above concepts, we can now move forward to explore data which is not normally distributed. We infer the meanings of the attributes available in this data set and then find patterns in this data set and try to generate meaningful analysis from the data set.

Book Examples - Doing Data Science

Chapter 2: NY Times Data set + questions + outcomes

Data Set : All 31 days of data available from http://stat.columbia.edu/~rachel/datasets/nyt<dayNumber>.csv

i. Create a new variable agecat to categorize the users into different ages [<18, 18-24, 25-44, 55-64, 65+]. Plot the distributions of number impressions and click-through-rate for these size age groups for one day.

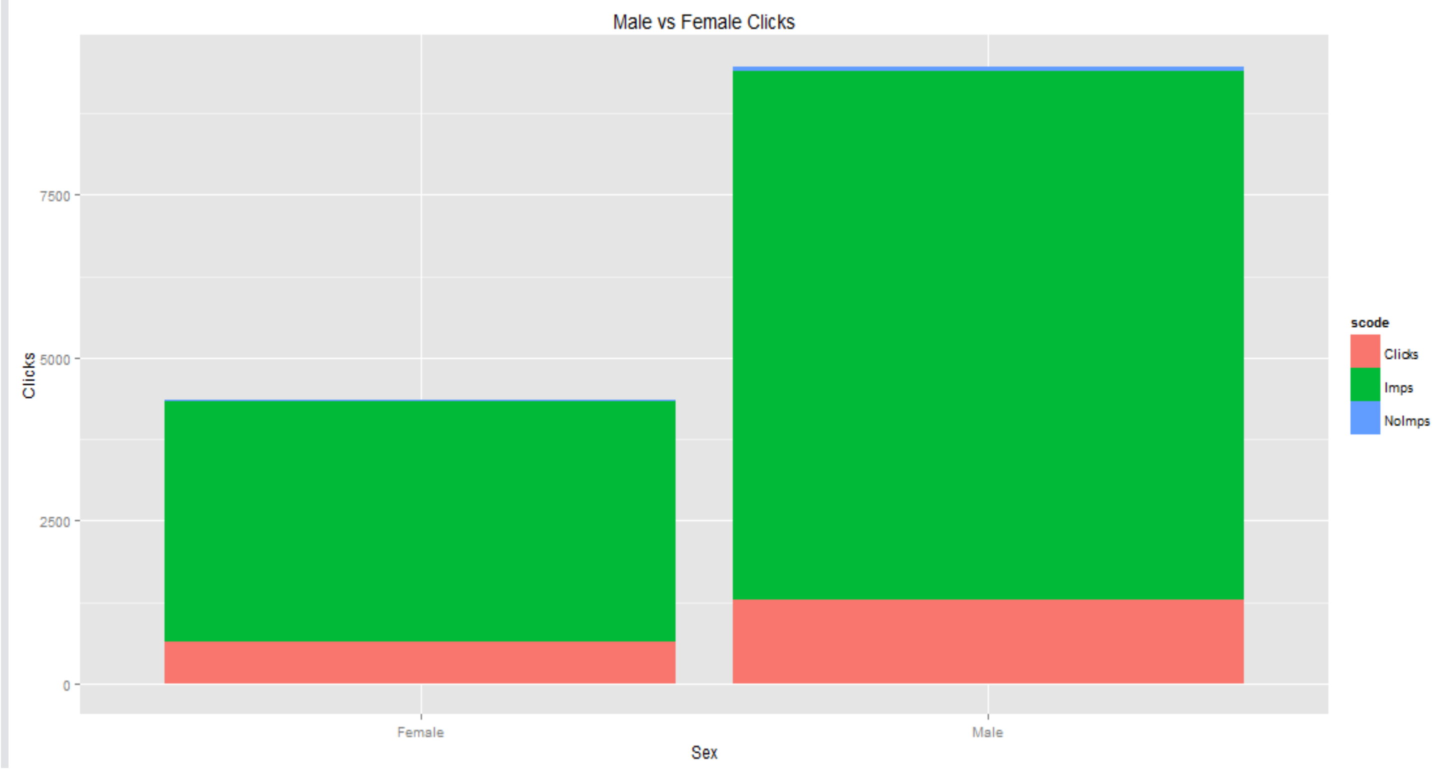


This is the data used for nyt1.csv .

ii. Define a new variable to categorize the users based on the click behavior. 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 vs not logged in).

Summary of the gender based grouping for less than 18 year olds.

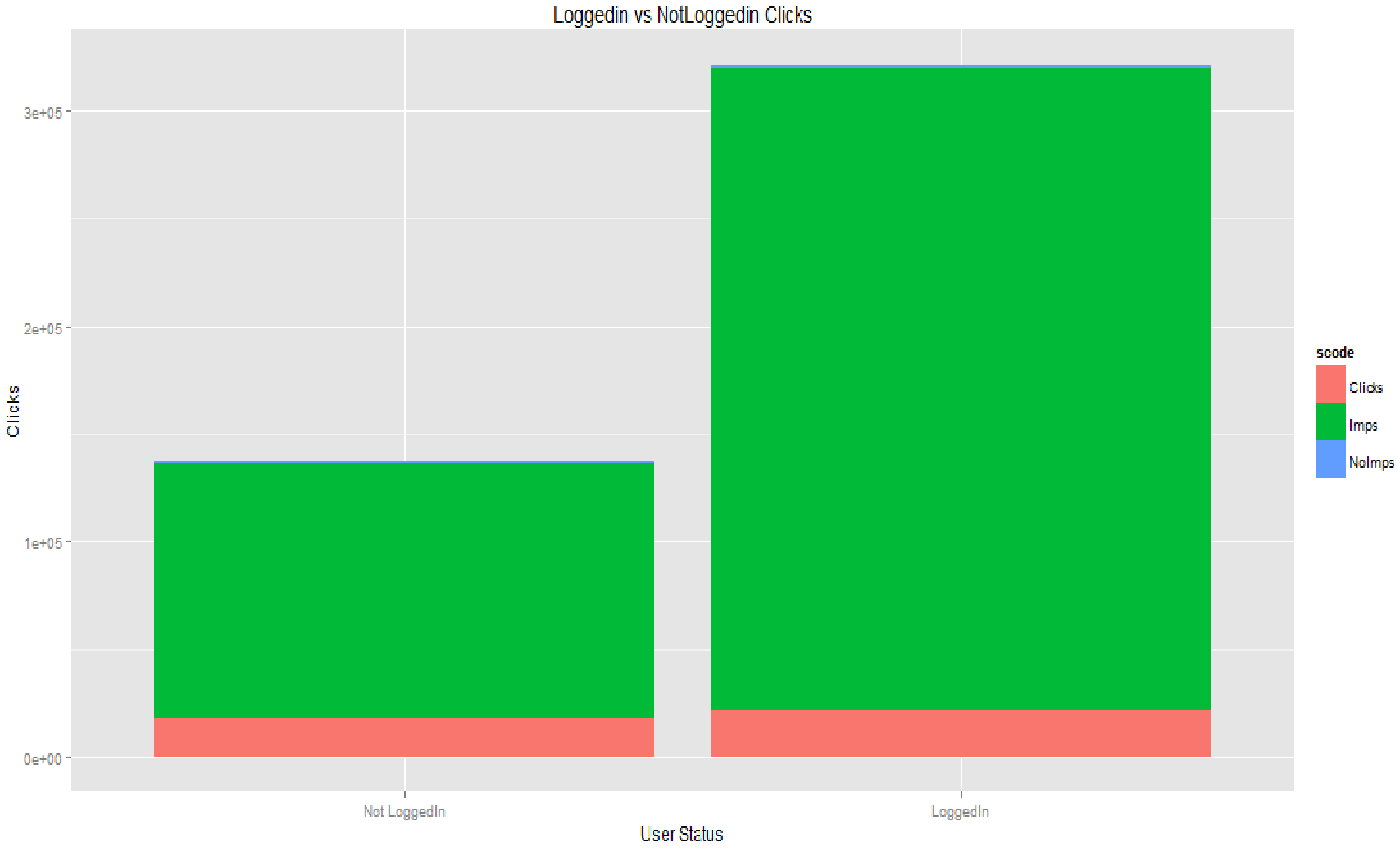
	scode	Gender	age_group	Impressions.clen
1	Clicks	0	(0,18]	648
2	Clicks	1	(0,18]	1285
3	Imps	0	(0,18]	3684
4	Imps	1	(0,18]	8115
5	NoImps	0	(0,18]	26
6	NoImps	1	(0,18]	70



Summary of data for impressions on people who have logged in versus who have not logged in.

	scode	signedInExp	Impressions.clen
1	Clicks	Not LoggedIn	17776
2	Clicks	LoggedIn	22062
3	Imps	Not LoggedIn	118401
4	Imps	LoggedIn	297136
5	NoImps	Not LoggedIn	929
6	NoImps	LoggedIn	2137

Image in the following page.



iii. Extend your findings for all 31 days. Create metrics like CTR, quantiles, mean, median, variance, and max that summarizes the data.

Age wise categories.

	age_group	Age.sum	Age.len	Age.min	Age.mean	Age.median	Age.max
1	(-Inf,0]	0	17776	0	0	0	0
2	(0,18]	37424	2371	7	15.78406	16	18
3	(18,24]	35546	1669	19	21.29778	21	24
4	(24,34]	84800	2870	25	29.54704	30	34
5	(34,44]	141953	3592	35	39.51921	39	44
6	(44,54]	155278	3139	45	49.46735	49	54
7	(54,64]	258454	4337	55	59.59281	60	64
8	(64, Inf]	297681	4084	65	72.88957	72	107

	Age.var	Age.quant.0%	Age.quant.25%	Age.quant.50%	Age.quant.75%	Age.quant.100%	ctr
1	0	0	0	0	0	0	0.1850

							34
2	3.540691	7	15	16	17	18	0.179803
3	2.556956	19	20	21	23	24	0.16946
4	6.778372	25	27	30	32	34	0.17243
5	6.914696	35	37	39	42	44	0.169623
6	6.997898	45	47	49	52	54	0.171423
7	6.959155	55	57	60	62	64	0.175582
8	35.38996	65	68	72	76	107	0.177486

Signed In User Classification

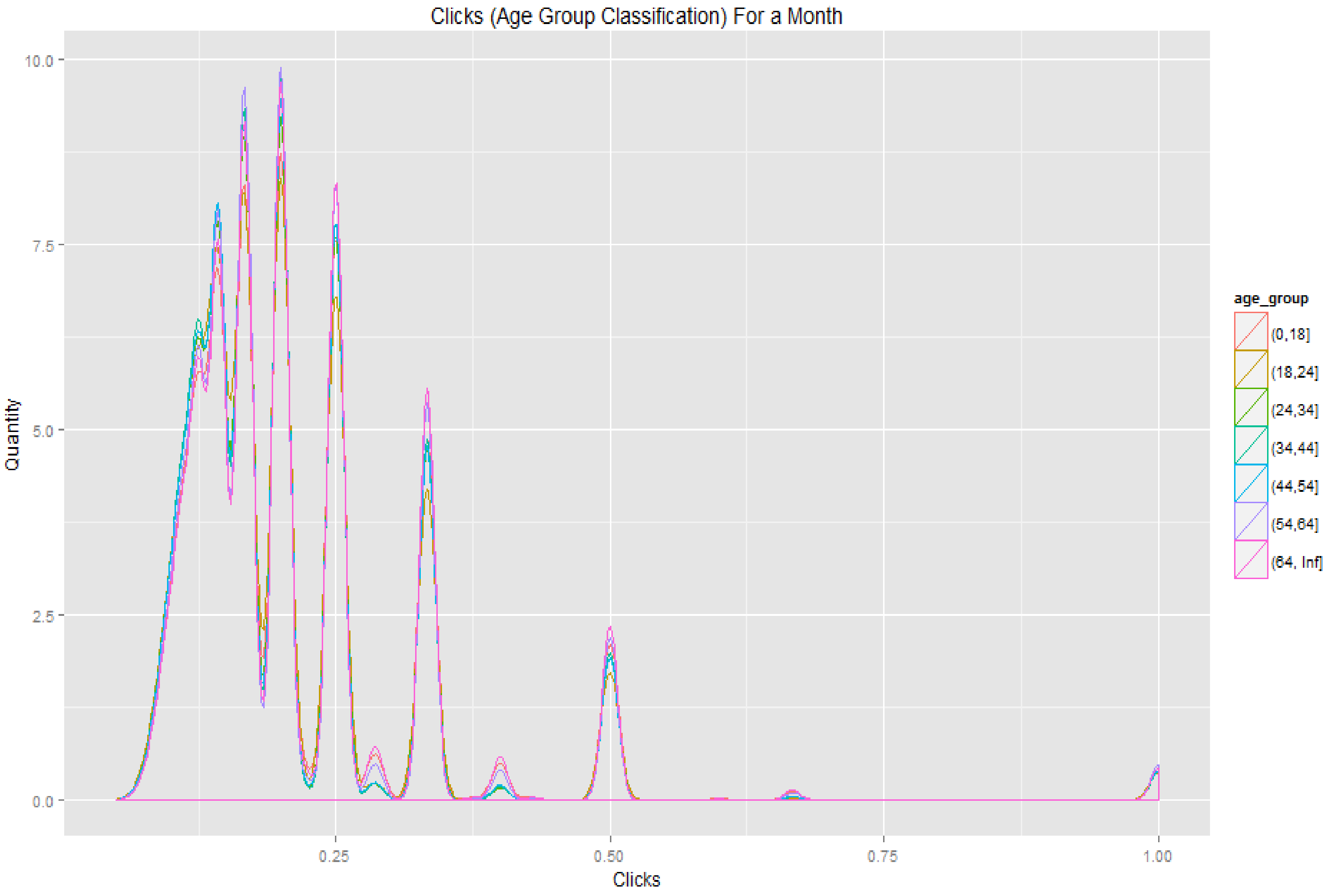
	age_group	Signed_In.sum	Signed_In.len	Signed_In.min	Signed_In.mean	Signed_In.median
1	(-Inf,0]	0	17776	0	0	0
2	(0,18]	2371	2371	1	1	1
3	(18,24]	1669	1669	1	1	1
4	(24,34]	2870	2870	1	1	1
5	(34,44]	3592	3592	1	1	1
6	(44,54]	3139	3139	1	1	1
7	(54,64]	4337	4337	1	1	1
8	(64, Inf]	4084	4084	1	1	1

	Signed_In.max	Signed_In.var	Signed_In.quant.0%	Signed_In.quant.25%
1	0	0	0	0
2	1	0	1	1
3	1	0	1	1
4	1	0	1	1
5	1	0	1	1
6	1	0	1	1
7	1	0	1	1
8	1	0	1	1

	Signed_In.quant.50%	Signed_In.quant.75%	Signed_In.quant.100%	ctr
1	0	0	0	0.185034
2	1	1	1	0.179803
3	1	1	1	0.16946
4	1	1	1	0.17243
5	1	1	1	0.169623
6	1	1	1	0.171423

7	1	1	1	0.175582
8	1	1	1	0.177486

Expansion of Clicks for all age group ideally on a larger data set (Month - Cumulative)



Male vs Females (under 18 year age group)

Male

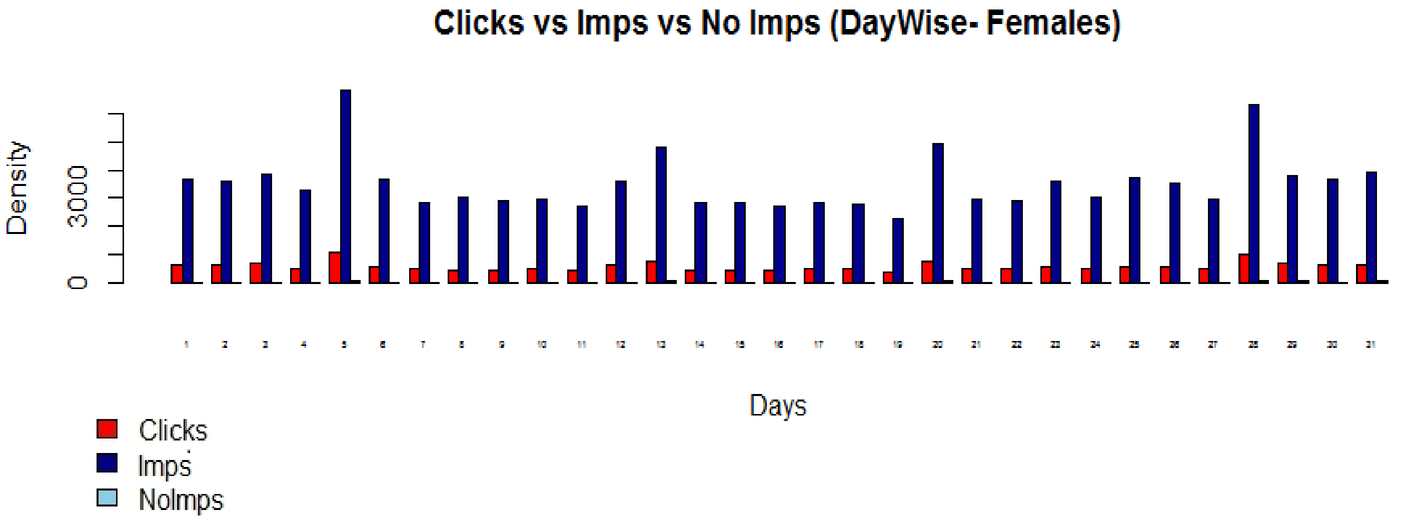
	Gender	age_group	day	clicks	Imp	NoImp
32	1	(0,18]	1	1285	8115	70
33	1	(0,18]	2	1350	7891	72
34	1	(0,18]	3	1338	8464	64
35	1	(0,18]	4	1066	6930	63
36	1	(0,18]	5	2201	13621	99
37	1	(0,18]	6	1236	8004	61

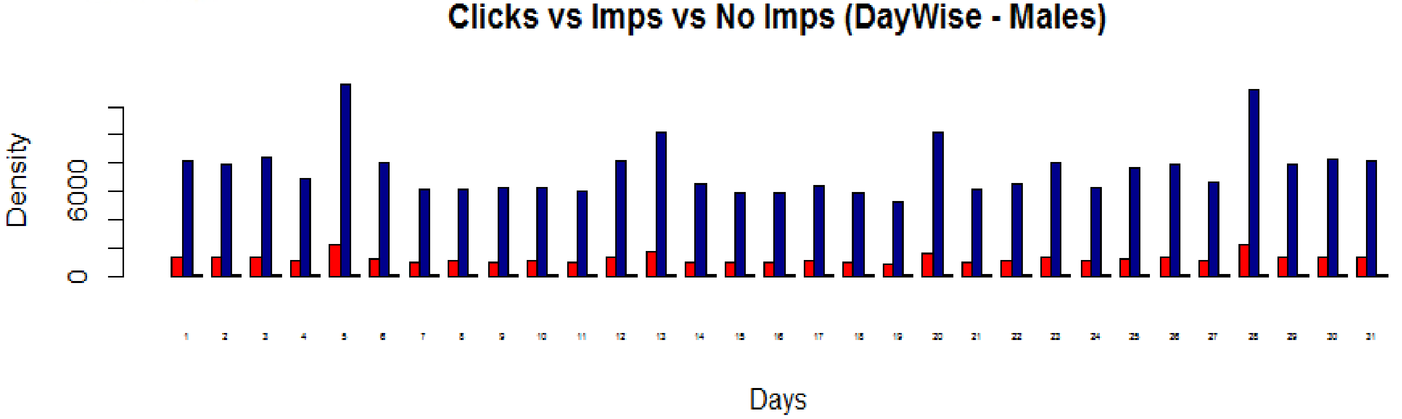
38	1	(0,18]	7	970	6089	38
39	1	(0,18]	8	1034	6165	59
40	1	(0,18]	9	976	6248	50
41	1	(0,18]	10	1022	6281	49
42	1	(0,18]	11	987	5970	48
43	1	(0,18]	12	1364	8184	80
44	1	(0,18]	13	1733	10193	84
45	1	(0,18]	14	982	6457	50
46	1	(0,18]	15	961	5933	43
47	1	(0,18]	16	1003	5832	50
48	1	(0,18]	17	1084	6393	49
49	1	(0,18]	18	964	5868	52
50	1	(0,18]	19	770	5197	34
51	1	(0,18]	20	1605	10196	70
52	1	(0,18]	21	990	6103	47
53	1	(0,18]	22	1083	6541	57
54	1	(0,18]	23	1314	7962	67
55	1	(0,18]	24	1021	6211	41
56	1	(0,18]	25	1194	7657	68
57	1	(0,18]	26	1279	7864	68
58	1	(0,18]	27	1050	6615	48
59	1	(0,18]	28	2162	13244	93
60	1	(0,18]	29	1322	7901	76
61	1	(0,18]	30	1326	8305	81
62	1	(0,18]	31	1346	8146	61

Female

	Gender	age_group	day	clicks	Imp	NoImp
1	0	(0,18]	1	648	3684	26
2	0	(0,18]	2	605	3600	27
3	0	(0,18]	3	679	3874	26
4	0	(0,18]	4	534	3282	27
5	0	(0,18]	5	1060	6794	52
6	0	(0,18]	6	589	3645	27
7	0	(0,18]	7	486	2859	26
8	0	(0,18]	8	454	3005	21
9	0	(0,18]	9	462	2895	27
10	0	(0,18]	10	490	2957	25
11	0	(0,18]	11	450	2719	18
12	0	(0,18]	12	605	3591	31

13	0	(0,18]	13	734	4778	33
14	0	(0,18]	14	444	2838	16
15	0	(0,18]	15	457	2868	16
16	0	(0,18]	16	437	2730	23
17	0	(0,18]	17	483	2807	19
18	0	(0,18]	18	492	2791	22
19	0	(0,18]	19	347	2261	13
20	0	(0,18]	20	786	4918	40
21	0	(0,18]	21	486	2949	21
22	0	(0,18]	22	490	2929	19
23	0	(0,18]	23	566	3570	22
24	0	(0,18]	24	511	2998	22
25	0	(0,18]	25	571	3750	26
26	0	(0,18]	26	560	3538	23
27	0	(0,18]	27	491	2981	19
28	0	(0,18]	28	1035	6331	54
29	0	(0,18]	29	662	3810	33
30	0	(0,18]	30	638	3665	26
31	0	(0,18]	31	626	3899	41





Logged in Versus Not Logged in (31 days)
Summary :

Logged in

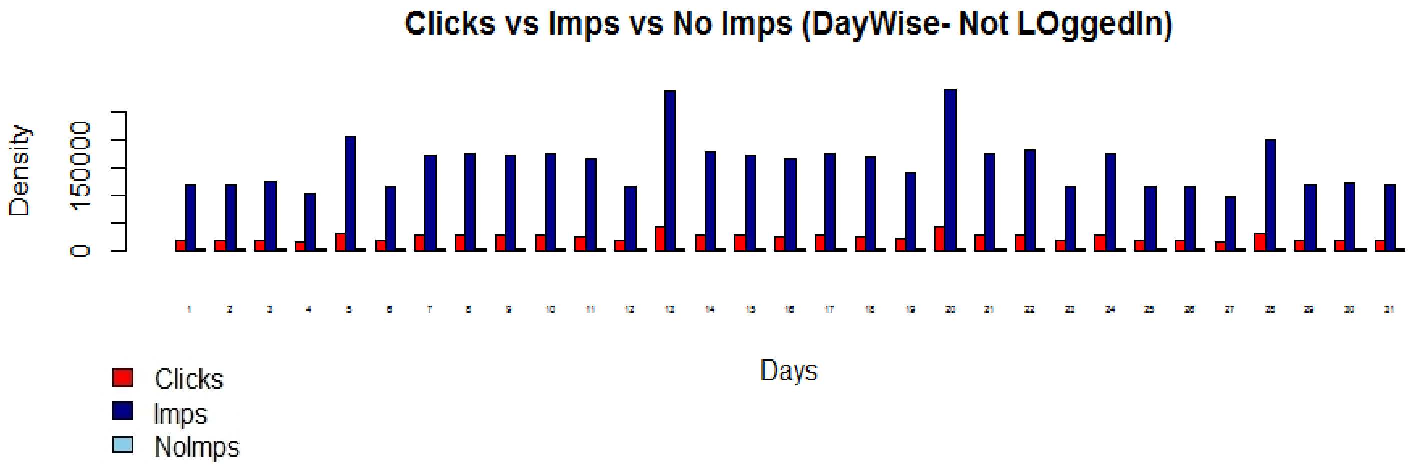
day		ks.sum	p.sum	
	clic	Im	NoIm	p.sum
1	1	22062	297136	2137
2	2	21675	293131	2110
3	3	22562	310064	2238
4	4	18772	256271	1899
5	5	37074	509013	3657
6	6	20717	285265	2057
7	7	16324	222915	1547
8	8	17015	228626	1671
9	9	16553	227385	1636

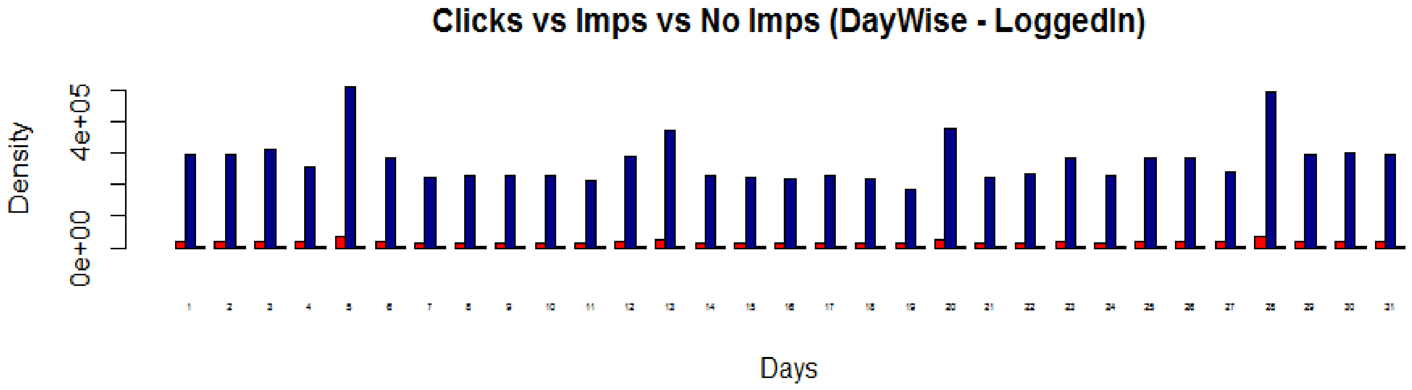
10	10	16979	230048	1727
11	11	15904	214194	1574
12	12	21643	291588	2132
13	13	27532	375233	2736
14	14	17086	231648	1635
15	15	16415	223215	1640
16	16	16170	218120	1479
17	17	17161	229143	1672
18	18	16020	218499	1576
19	19	13475	184345	1282
20	20	27872	378977	2696
21	21	16415	225437	1647
22	22	17256	235738	1678
23	23	20985	285203	2099
24	24	16956	230572	1758
25	25	21009	284738	2047
26	26	21155	286360	2109
27	27	17524	239683	1770
28	28	36301	495283	3557
29	29	21532	293107	2184
30	30	21846	299645	2199
31	31	21876	297519	2139

Not Logged in:

	day	clicks	Imp	NoImp
1	1	17776	118401	929
2	2	17520	117420	910
3	3	18510	123727	965
4	4	15632	102920	814
5	5	30754	203995	1551
6	6	17097	114370	832
7	7	25764	169714	1303
8	8	26147	174125	1333
9	9	26210	172089	1355
10	10	26308	175320	1381
11	11	25032	163994	1276
12	12	17530	116118	924
13	13	42968	285800	2111
14	14	26542	177108	1346
15	15	25664	169821	1337
16	16	25046	166171	1344
17	17	26141	174393	1388
18	18	24952	167825	1254

19	19	21190	140849	1093
20	20	43695	289245	2303
21	21	26005	173904	1435
22	22	27183	179566	1420
23	23	17147	114036	900
24	24	26195	175397	1405
25	25	17362	114055	901
26	26	17315	115079	839
27	27	14496	96120	735
28	28	29809	198004	1556
29	29	17697	117046	927
30	30	17980	120674	852
31	31	18038	118907	993





Classification of age over diverse metrics (Cumulative for a month)

	age_group	Age.sum	Age.len	Age.min	Age.mean	Age.median	Age.max
1	(-Inf,0]	0	729705	0	0	0	0

2	(0,18]	1090225	69092	5	15.77932	16	18
3	(18,24]	1058819	49772	19	21.27339	21	24
4	(24,34]	2411148	81742	25	29.49705	29	34
5	(34,44]	3947568	99933	35	39.50215	40	44
6	(44,54]	4504715	90994	45	49.50563	50	54
7	(54,64]	7361284	123682	55	59.51783	60	64
8	(64, Inf]	8510197	116651	65	72.95434	72	107
	Age.var	Age.quant.0%	Age.quant.25%	Age.quant.50%	Age.quant.75%	Age.quant.100%	ctr
1	0	0	0	0	0	0	0.184399
2	3.546256	5	15	16	17	18	0.180635
3	2.560508	19	20	21	23	24	0.171639
4	6.892414	25	27	29	32	34	0.172045
5	6.913869	35	37	40	42	44	0.171629
6	6.903318	45	47	50	52	54	0.172148
7	6.896509	55	57	60	62	64	0.176989
8	36.29249	65	68	72	76	107	0.181651

Classification of Signed In users over diverse metrics (Cumulative for a month)

	Signed_In.max	Signed_In.var	Signed_In.quant.0%	Signed_In.quant.25%
1	0	0	0	0
2	1	0	1	1
3	1	0	1	1
4	1	0	1	1
5	1	0	1	1
6	1	0	1	1
7	1	0	1	1
8	1	0	1	1
	Signed_In.quant.50%	Signed_In.quant.75%	Signed_In.quant.100%	ctr
1	0	0	0	0.184399
2	1	1	1	0.180635
3	1	1	1	0.171639
4	1	1	1	0.172045
5	1	1	1	0.171629
6	1	1	1	0.172148
7	1	1	1	0.176989

8	1	1	1	0.181651

From the above summary, it can be seen that ctr rates have improved over a month.
And, we have signed in user from all age group.

RealDirect Data Strategy

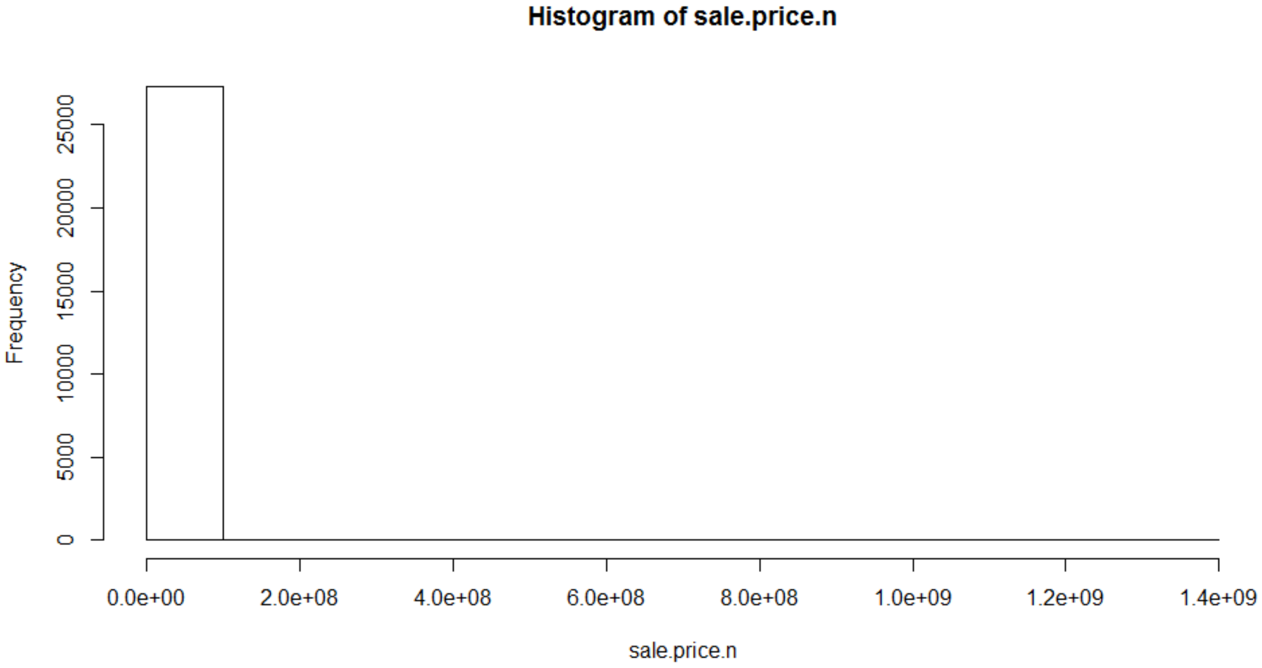
Exercise:
Clean and explore the real direct data for Manhattan, New York.

Cleaning the code involves replacing non numeric characters in numeric values. Converting string represented numeric values to numerics.

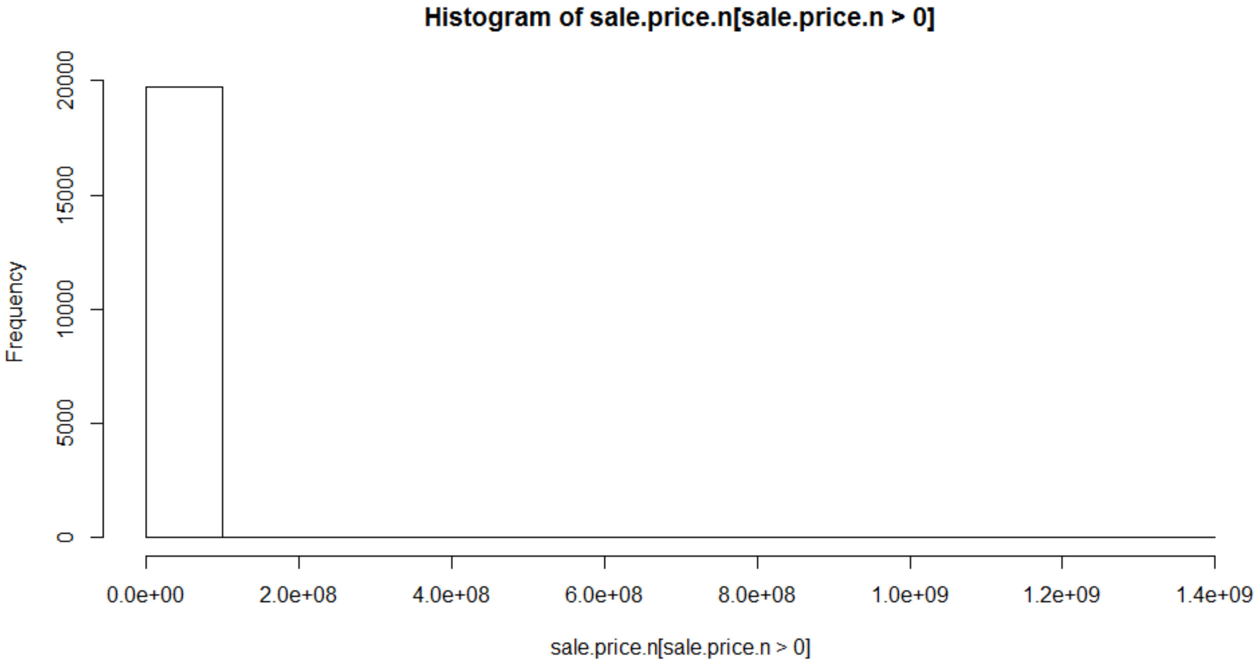
Being data scientist often involves discussing with people who are not data scientist. Can you think of any other people you should talk to?
Bank managers, Software Engineers, Lawyers, CEOs, Ministers of the state and other idiots.
Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insights into how to do it in your own field?
Yes it makes translation the thoughts into my own field even easier. Though I am still wondering why this is an important question for this assignment.

The graphs to be plotted given in the exercise.

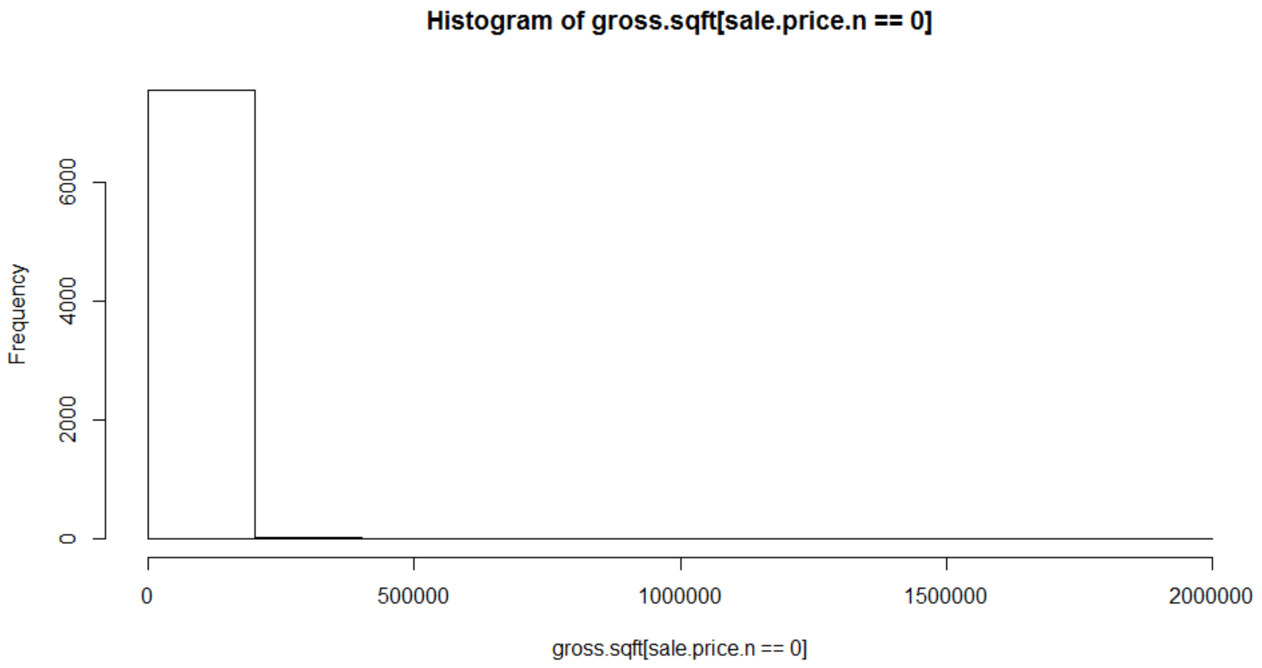
Plot of salesprices and its frequency



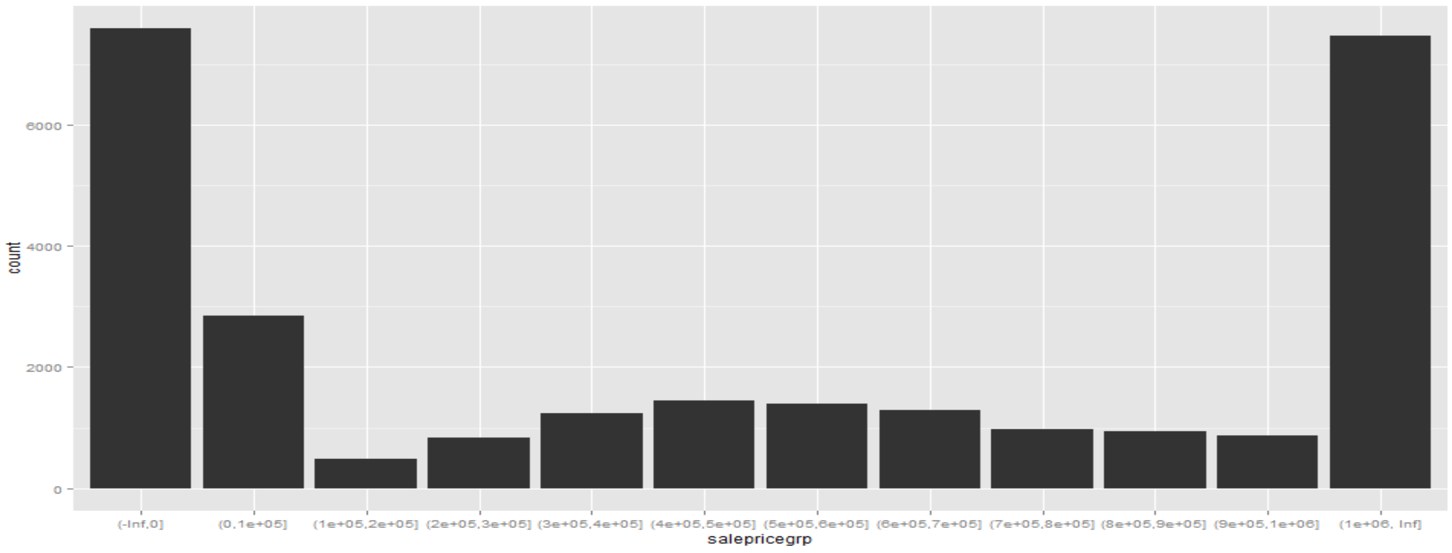
Plot of salesprices (Greater than zero) and its frequency



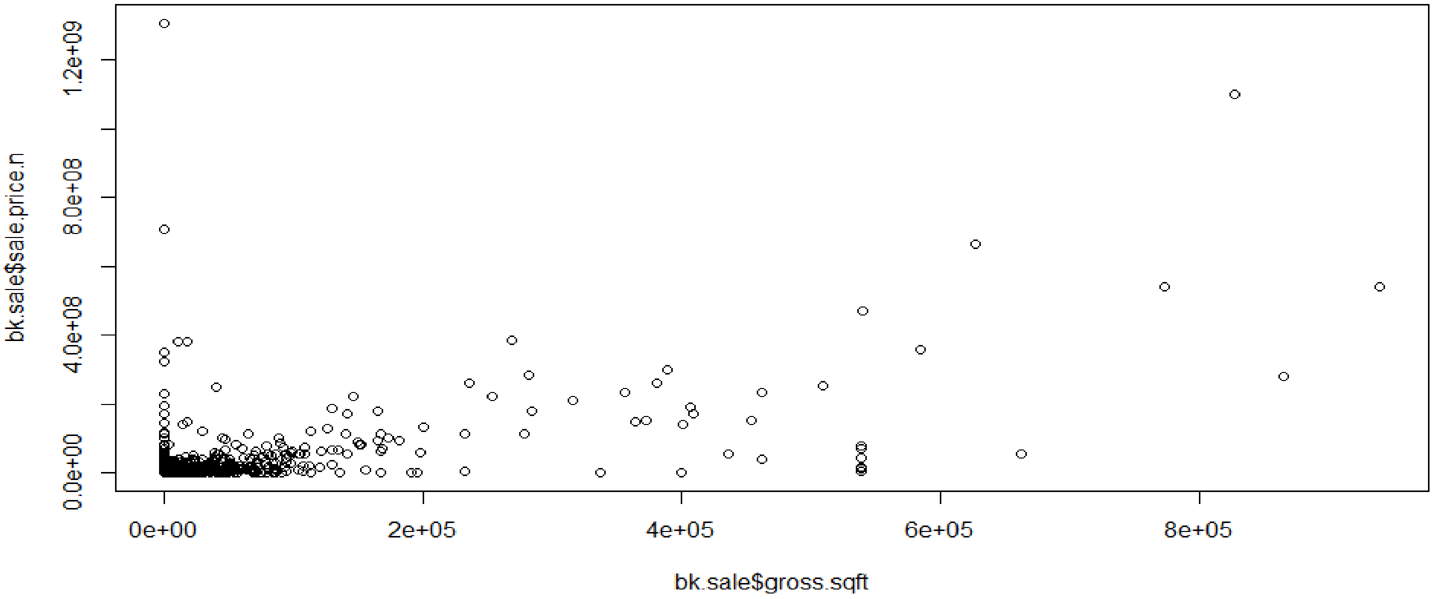
Plot of Gross Sq. Ft (SalesPrice equal to zero) and its frequency



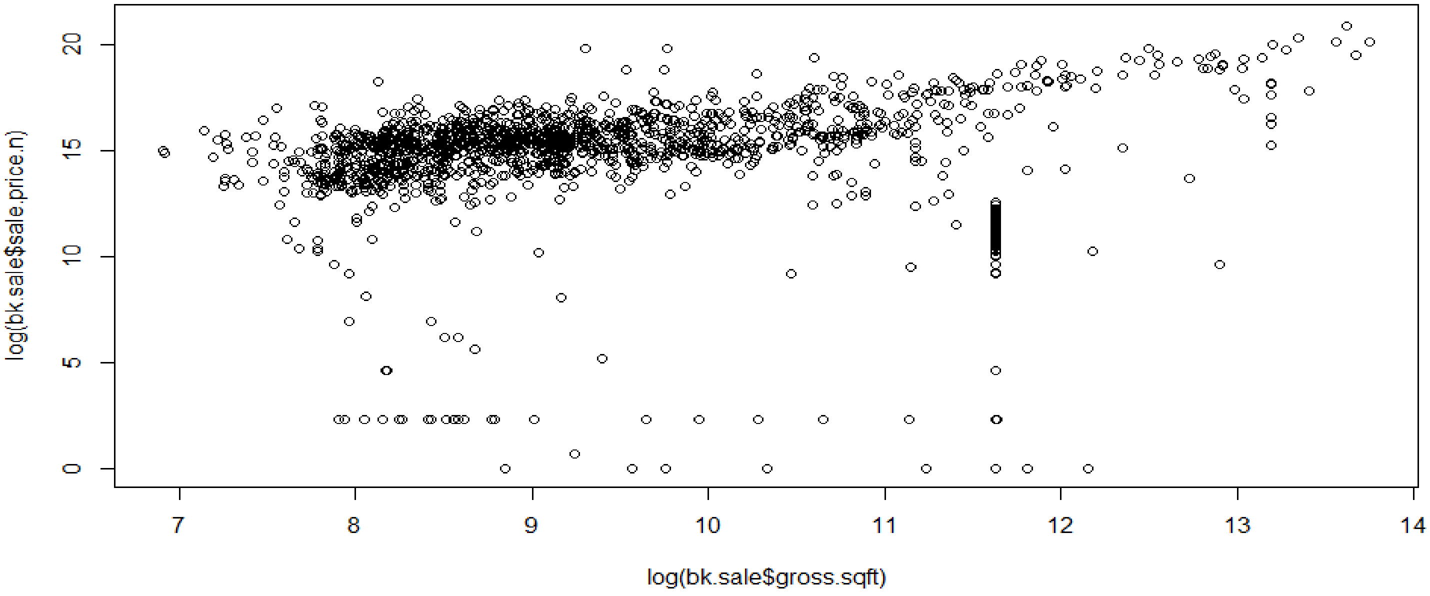
Plot of salesprices (Groupwise slice of sales price) and its frequency



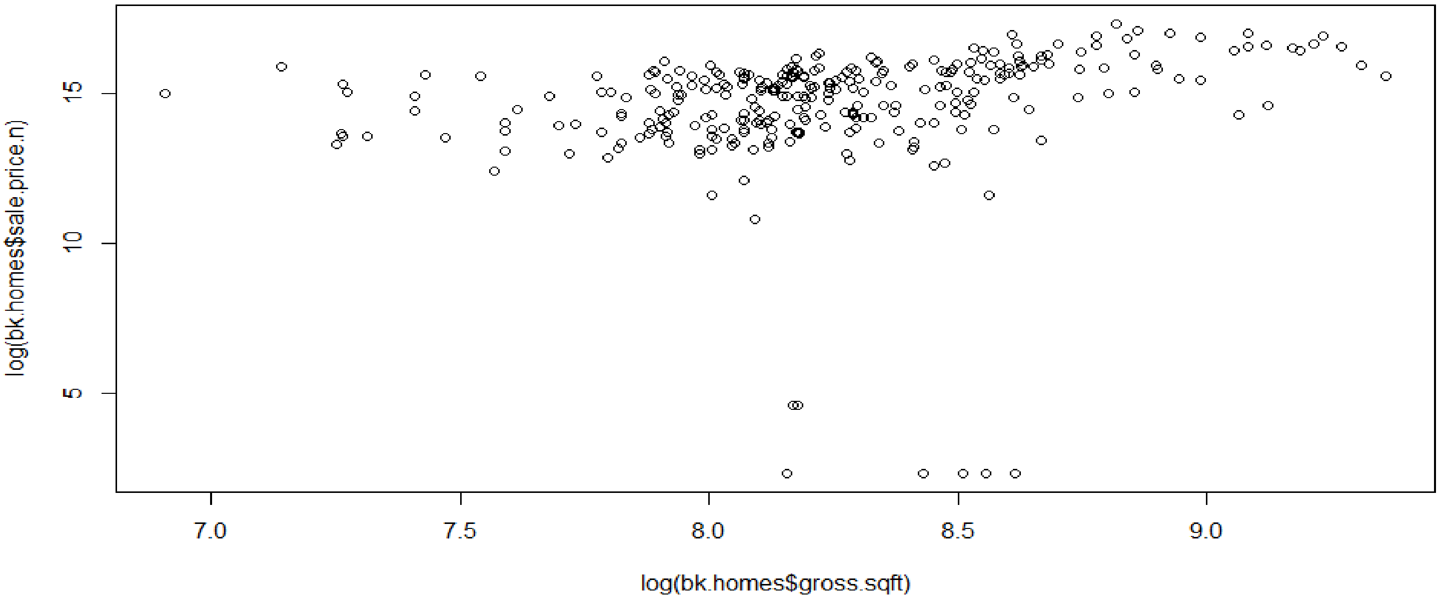
Plot of salesprices vs Gross sq. ft.



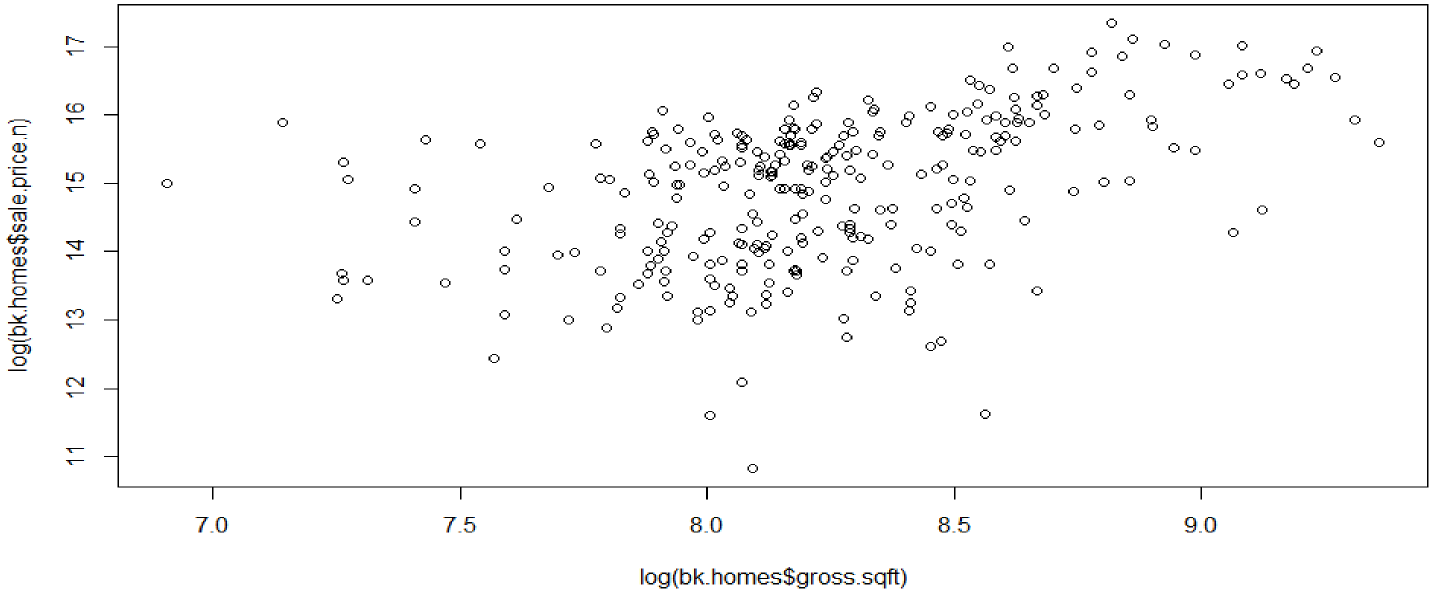
Plot of salesprices vs Gross sq. ft (Logarithmic Expansion).



Plot of salesprices vs Gross sq. ft (Logarithmic expansion only for family type houses)



Plot of salesprices vs Gross sq. ft (Logarithmic Expansion after removing outliers)



Own Data

- *Identifying the data set*

To generate a meaningful analysis from a data set we shall set out to find a data set which has data collected from a national (or international) level distributed over a substantial period of time. data.gov is a common data repository which mirrors data available from all the participating US government entitities. The website has daily updates about the data sets added and also has a very good search engine to find past data sets from the domain of our interest.

We found an interesting dataset, Campus Security[(4)], which has recorded over a period of 11 years, the incidents of crimes, fires recorded in various colleges in USA. We also found the 2011 "*Integrated Postsecondary Education Data System (IPEDS) - Institutional Characteristics - Directory Information - 2011*" [(5)] provided by the National Center for Education Statistics (NCES) .

- *Data Set Description*

About College Directory [(7)]

The directory information for every institution in the 2011 IPEDS universe. Includes name, address, city, state, zip code and various URL links to the institution's home page, admissions, financial aid offices and the net price calculator. Identifies institutions as currently active, institutions that participate in Title IV federal financial aid programs for which IPEDS is mandatory. It also includes variables derived from the 2011 Institutional Characteristics survey, such as control and level of institution, highest level and highest degree offered and carnegie classifications.

About College Crimes and Fires [(6)]

It counts the year wise incidents of various crimes including but not limited to thefts, arson, sex offences, drug violations, liquor violations , fires, deaths due to fires for each campus of each institution. Each university has a unitid which can be mapped with the unitid in college directory to gather college location and college sector details.

- *Experiments and Results*

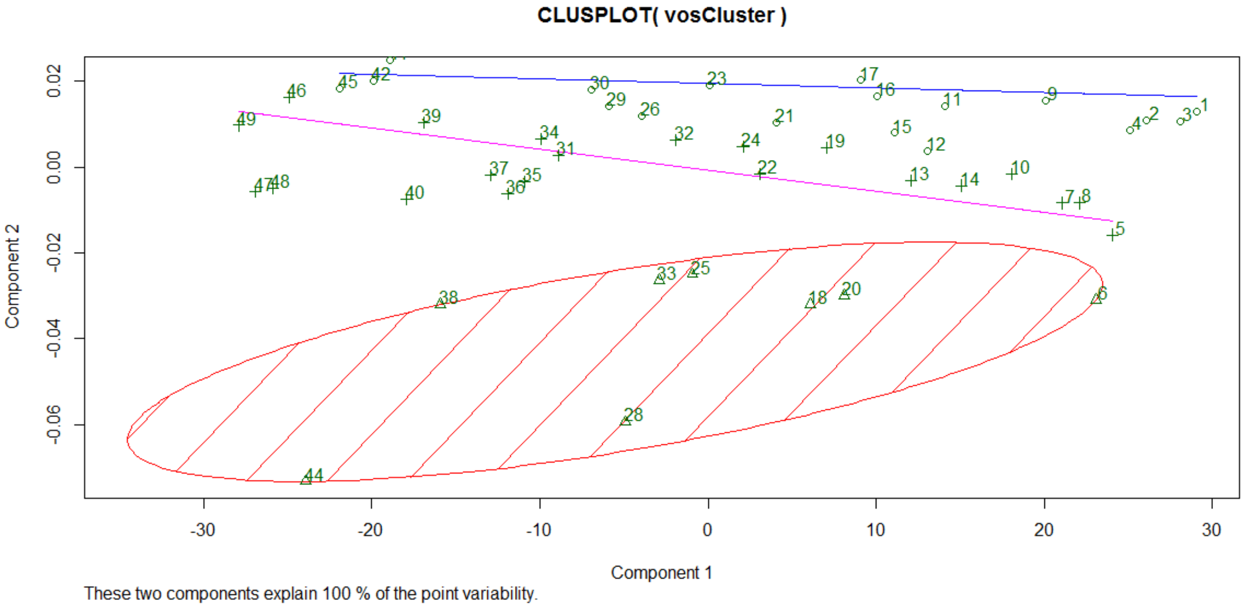
First part of the experiment is to clean up some data by replacing NULLs and NAs to something less erroneous. The description of the data sets given by the data provider is that *some* of the counts which have been marked as 0 could have been due to unavailability of the college to provide that data. *Some* of these 0s themselves could have been replaced with Nulls or NAs. Since we are not likely to find strong information for the

experiments we are about to conduct we have replaced these with 0s.
Also there was not a proper one-one mapping from the crimes data and the campus directory data. Hence we are including only those universities which are present in both these datasets. This in itself is a substantial dataset which is very useful to perform the experiments.

✓ Which state is safest to pursue your degree?

Having a clean data set now our first experiment is to try and classify the 49 states in US mainland on the basis of campus security. We shall use the k-means algorithm implemented in R to make three buckets.
Kmeans needs as its input a dataset which contains only the meaningful attributes. On the basis of these attributes we will cluster this data into various buckets. In our experiment we are trying to classify the states as three categories (Low, Medium and High crime risks). We will group the institution crime data state wise. Group the total crimes recorded for each state giving equal weightage to the different types of crimes. We shall also the group the total number of students enrolled in the different universities in these states. A nice indicator of the crime statistics would be the sum total of crimes per student for each state.

Having run this indicator over the kmeans function in R we get the following cluster plot.



There are three cluster like we wanted . Each of the 49 states in the mainland USA will belong to one of the 3 clusters.

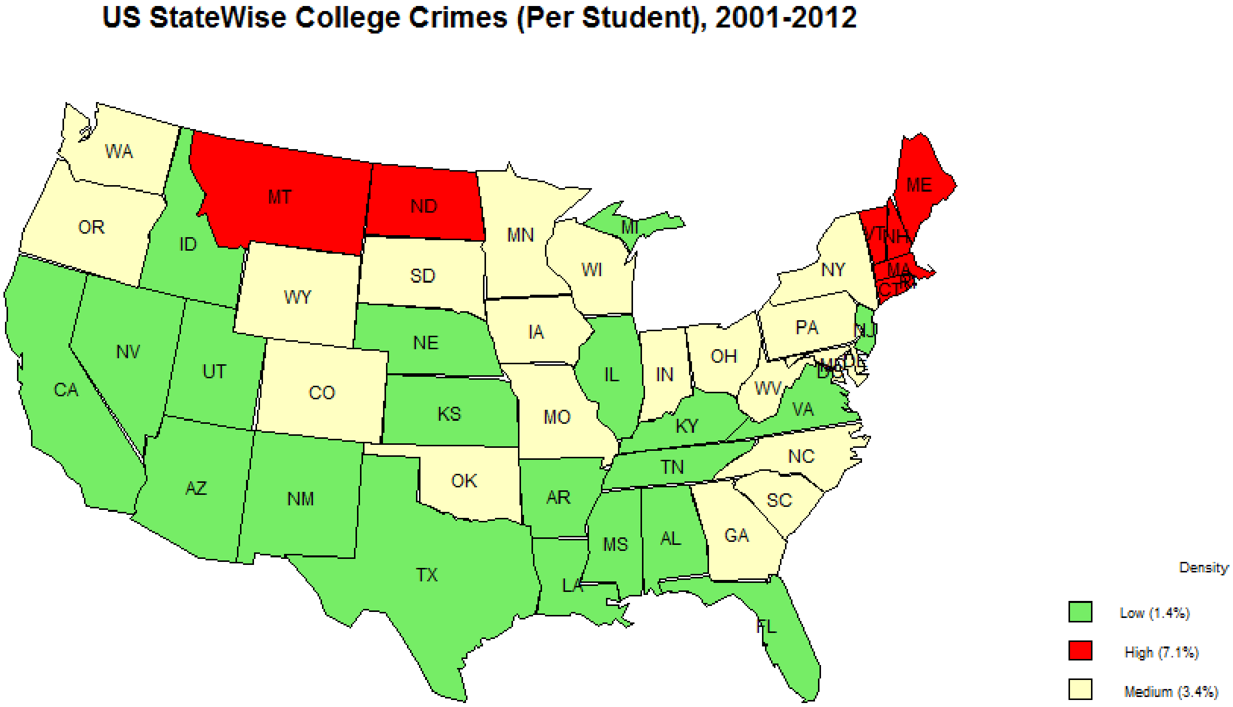
Cluster summary

Cluster 1 : center : 0.014

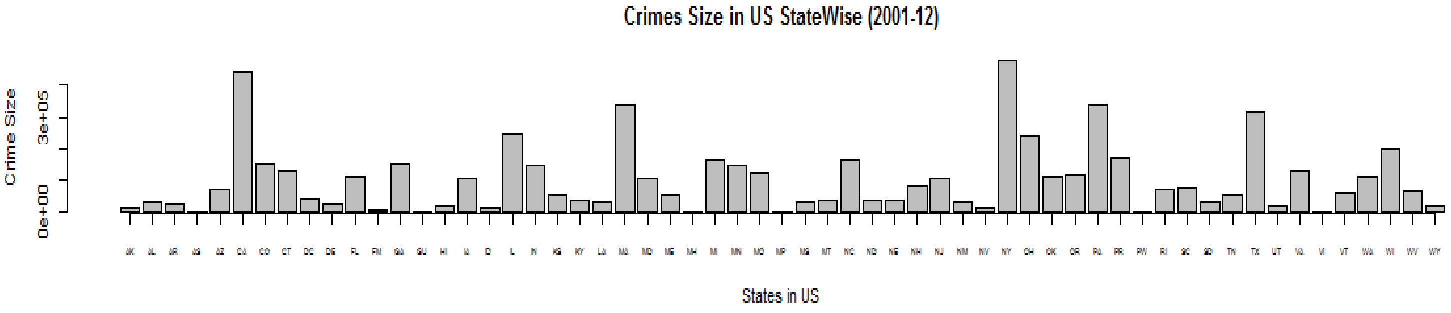
size : 20
Cluster 2 : center : 0.071
size : 8
Cluster 3 : center : 0.034
size : 21

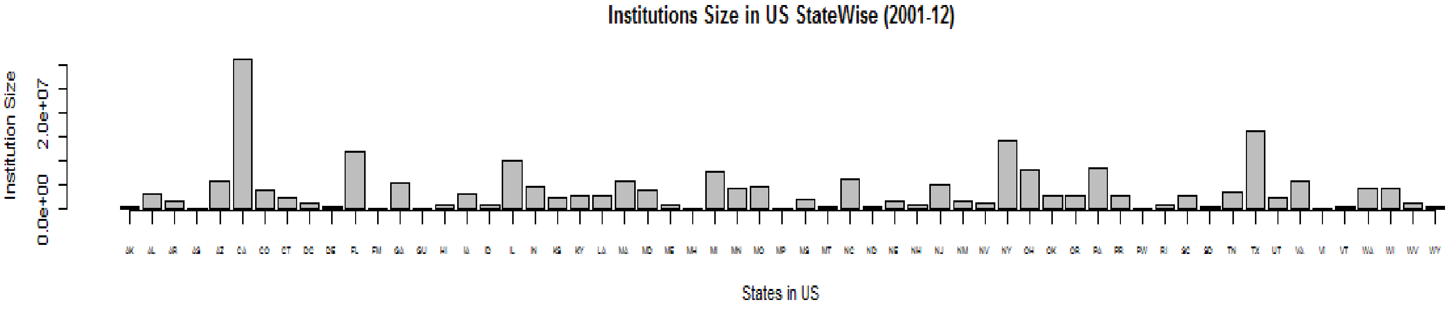
Centers are indication of the per capita crimes recorded. So the greater the per capita crime in a state , it will be allocated a bucket of higher mean. The bucket with the highest mean has so many states. This could be a wake up call for the education and security authorities of these states.

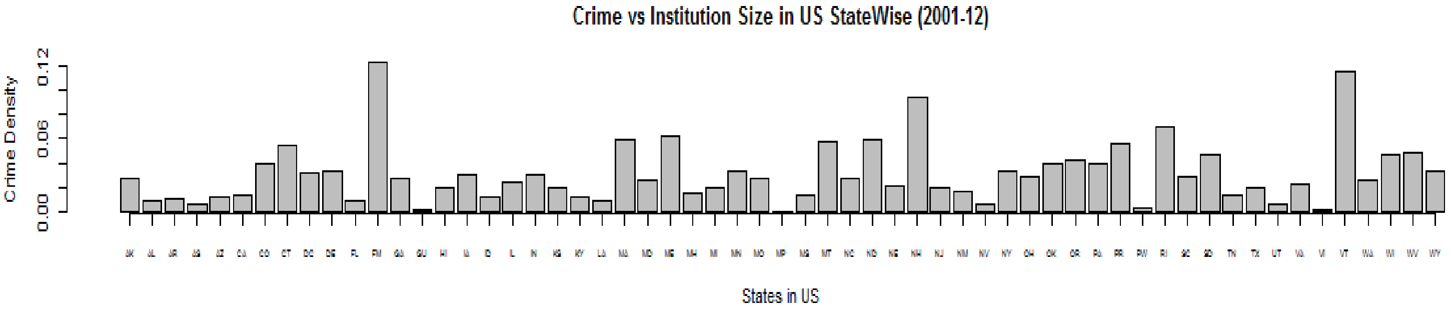
To get a more meaningful visual idea of which states are safer we plot the clusters on a map. The states filled with color Red are the high campus crime risk states. The ones filled with yellow have a medium risk of campus crime. Greens have the least risk on a relative basis.



Here is visual summary of the facts that went behind generating this crime per student choropleth map. The below graph is a bar graph of the distribution of statewise crimes recorded, statewise students enrolled for post secondary education and exact values of crimes per student. The states that have the most number of students are California, New York and Texas. New York records the most campus security violations. Followed by California, Rhode Island and Texas.

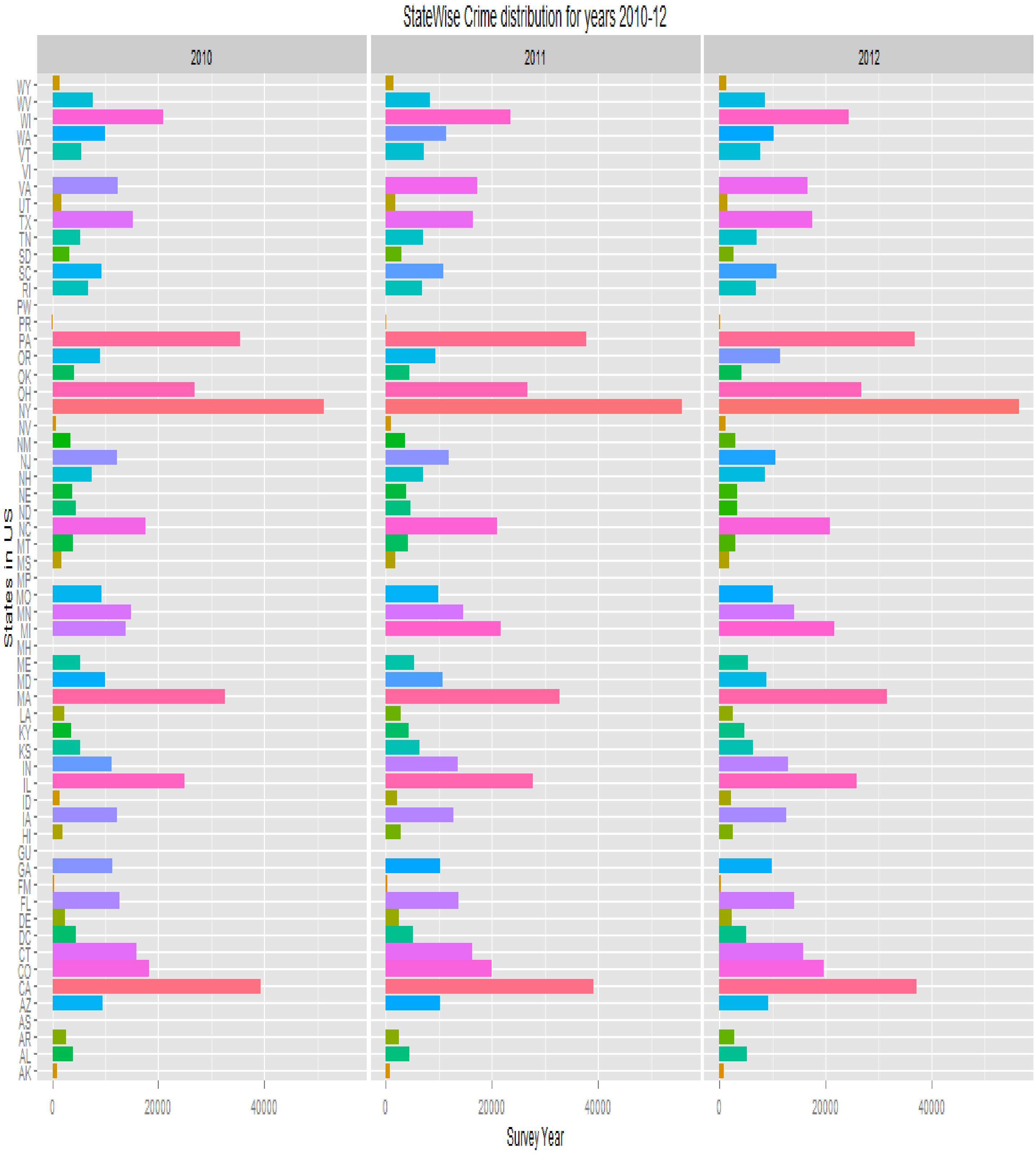






✓ Recent Trends in Campus Security

It would be interesting see how the states have performed over the past 3 years the latest data in security would always be give a nice indication on which state is probably safer to go and study. We will use the statewise grouped data but filter out only the last three years records. We will use a horizontal bar plot to accommodate all 49 states in a single page and compare them year by year for the years 2010 ,2011, 2012. Having plotted the graph we can see that New York state has more recorded crimes over the last three years than any other state. California and Pennsylvania fight for the second and third spot. But given the fact that there are a lot of student enrollments in these three states than any other this statistic is expected.



StateWise Crime distribution for years 2010-12
2010
2011
2012
States in US
WY
WV
WI
WA
VT
VI
VA
UT
TX
TN
SD
SC
RI
PW
PR
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MP
MO
MN
MI
MH
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GU
GA
FM
FL
DE
DC
CT
CO
CA
AZ
AS
AR
AL
AK
0
20000
40000
0
20000
40000
0
20000
40000
Survey Year

✓ Private vs Public Sector Campus Safety

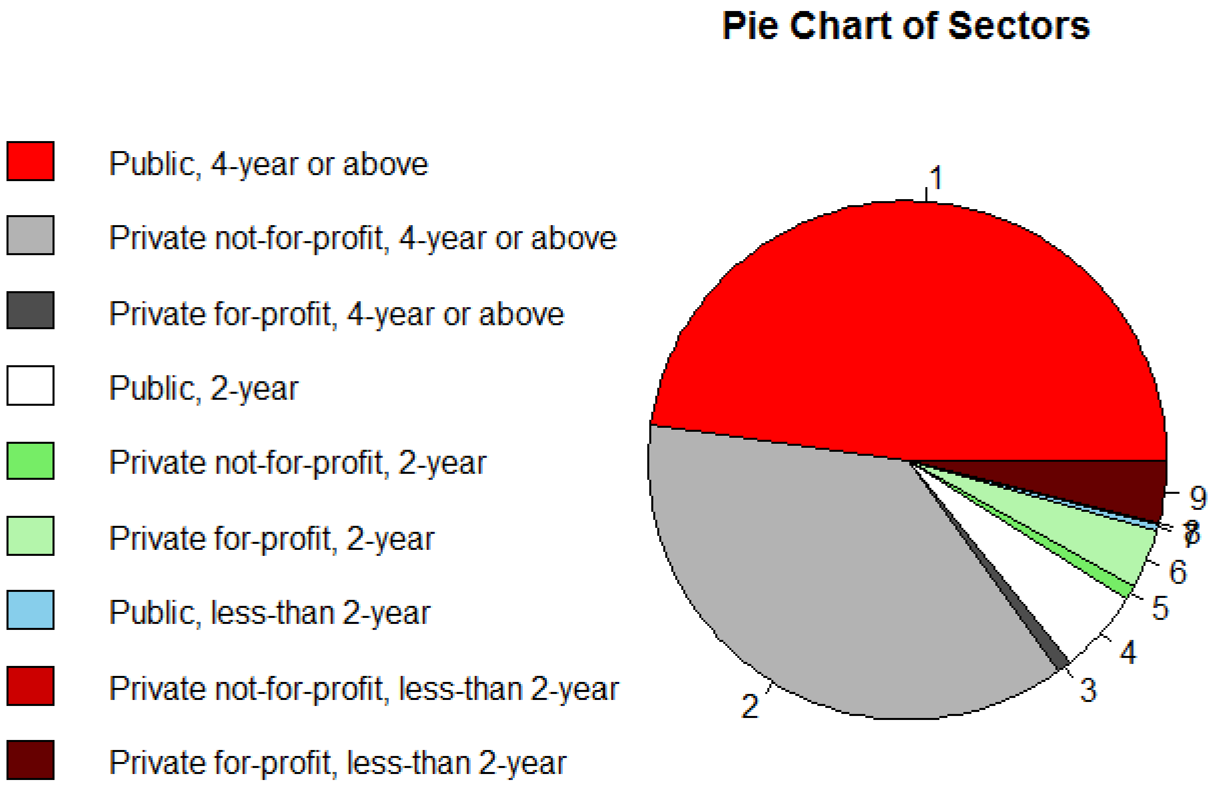
Having found out which states are safer to go to school to, we shall try and find if the going to private school is better than a public school or if a 4 year college has more crimes per student than a college offering a 2 year course.

As with the previous experiment this will also be an aggregate experiment. We will group all the colleges into their 9 different sectors.

3 based on nature of ownership : Public, Private not-for-profit, Private for-profit.

3 based on the duration of the courses offered : 4 year or above, 2 to 4 years, less than 2 years.

Totalling 3*3 =9 categories, in all. Having the data set into 9 categories, we shall find the crime rate per student for the individual categories. Using R's pie chart we shall plot the results to visually summarize our findings.



This result a couple of extremely interesting facts:

1. Private for profit colleges perform exceptionally well in terms of maintaining campus safety. We can infer

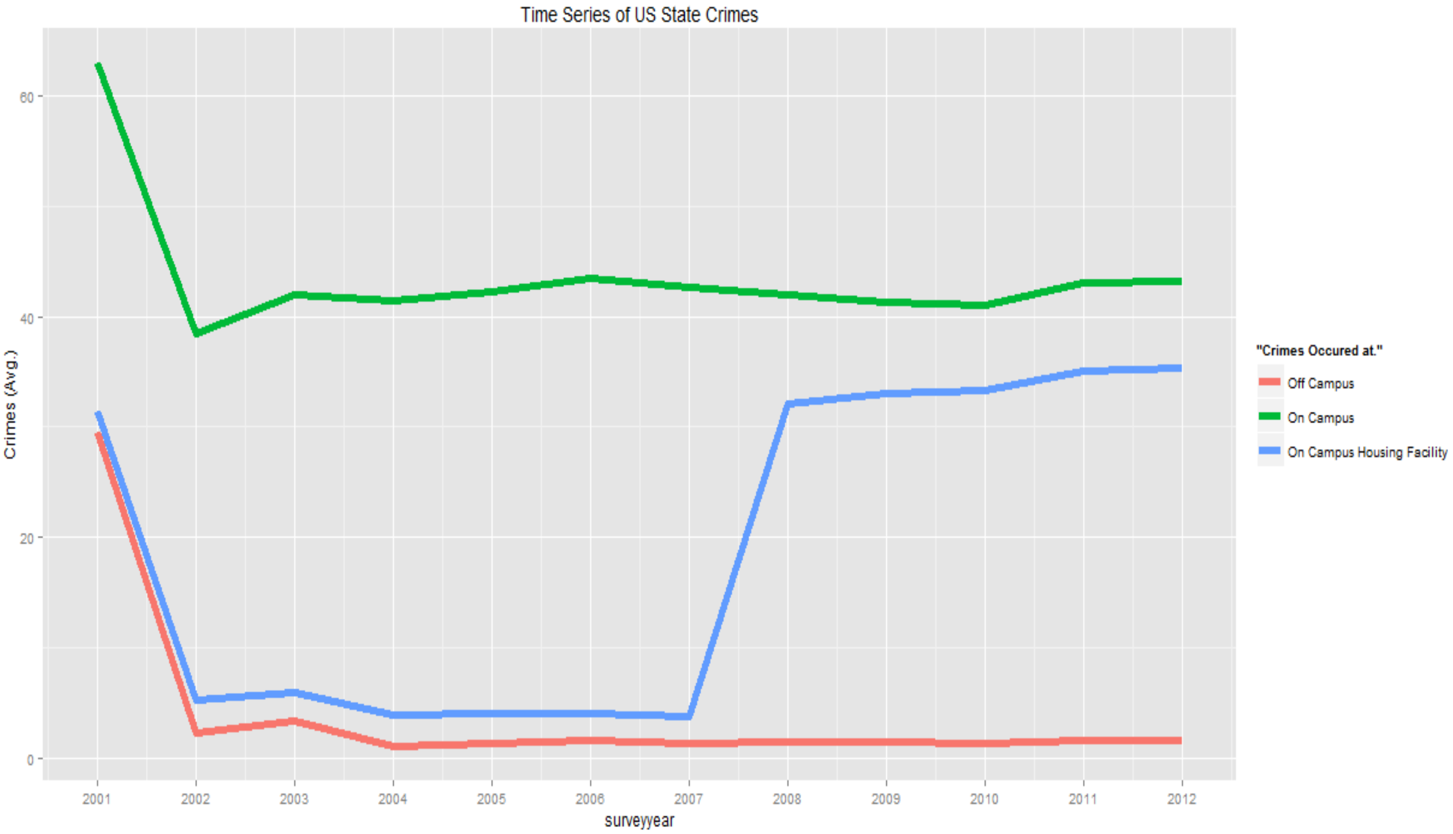
that these colleges spend enough time and money to maintain the high standards of campus safety.

2. There is a significant increase in criminal activities in the 4 year colleges. We can make an educational inference that with the increase in the course duration, students form cliques and eventually start indulging in crimes, big and small.

✓ Should I get a house on campus ?

Colleges offer on campus housing facilities. Let us see if we can see if we can see if these facilities are safer for students than living outside campus.

We shall aggregate the sum of crimes recorded under the following three categories : on-campus, non-campus and housing facilities. We shall group these 3 categories for each of the 12 years for which we have data (2001-2012). We shall now use a line graph to see how campuses in general have performed over the years in terms of on campus and housing facilities security and also get an idea about how crimes are committed non campus.



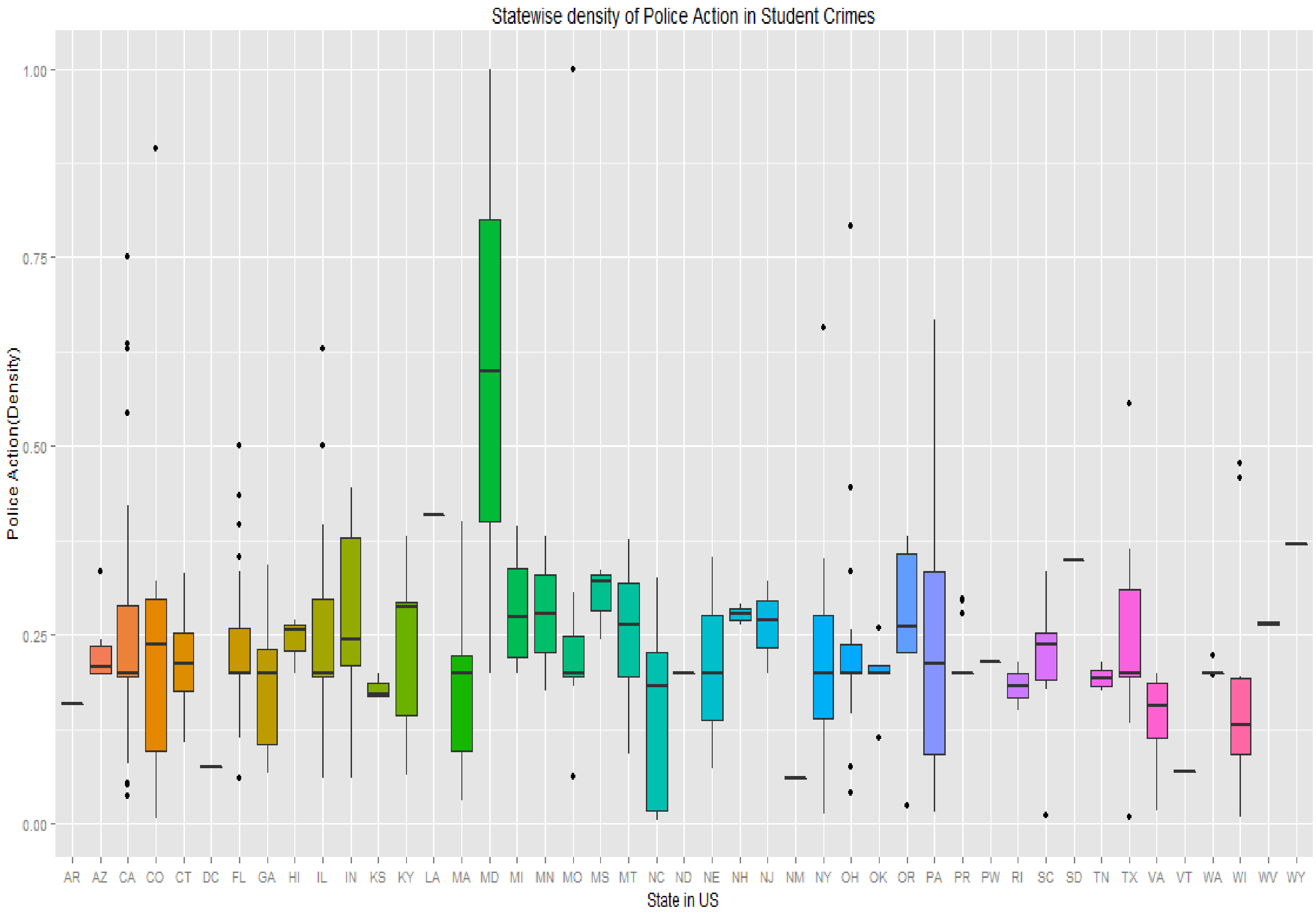
The three line graphs are for the total of crimes per student. Red line is for Non Campus security violations, Green line indicates for us the On campus security violations. Blue line indicates the security violations at the

Housing facility provided on campus. Interestingly **on campus security violations are the most** among the three. It has been easier to commit a crime on campus.

As much as both the on campus housing facilities crime and non campus crimes have been equal till the year 2007, after that the current trend in crimes seems to have seen an **increased affinity towards on campus housing facilities**.

✓ Call 911?

The involvement of police in curbing crimes bolsters the college’s cause to maintain security on and off campus. Since we have a count of the Police Arrested done for each of the campuses, it would be interesting to analyse how much each state police(and university police) involve in the campus security violations. We shall get the police involvement factor by summing up the total police arrests and police records divided by the total crimes for that college in that survey year.

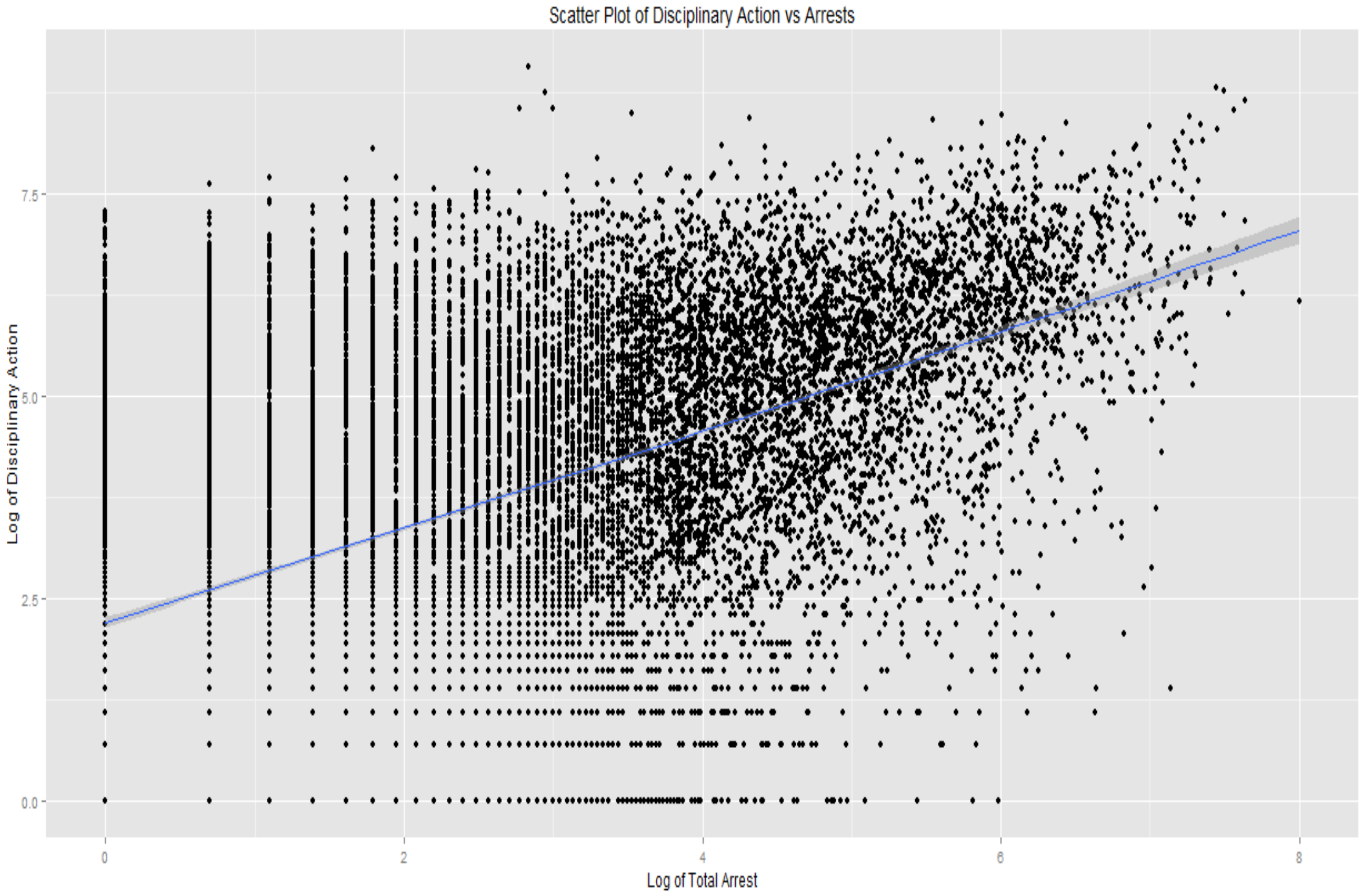


Box plot gives an idea of the distribution of the police factor variable for each of the states. A well distributed data set gives a bigger box length. However, if data points are crowded near a place then the box will shrink towards that place. The middle line these boxes indicates the mean.

Hence we can see that **Madison** police are clearly more involved in the campus crimes than any other state's police. Highly over 50% of all campus crimes do not go to the extent of police being involved.

✓ Fear of arrests

The fear of arrests will help keep the number of crimes committed under check. Do the colleges know this? To get an idea we are going to plot if the number of disciplinary actions that went on to see arrests. We shall take individual campus records and plot the logarithmic total of disciplinary actions take in the campus and the logarithmic total of the arrests that happened for the crimes in that campus.



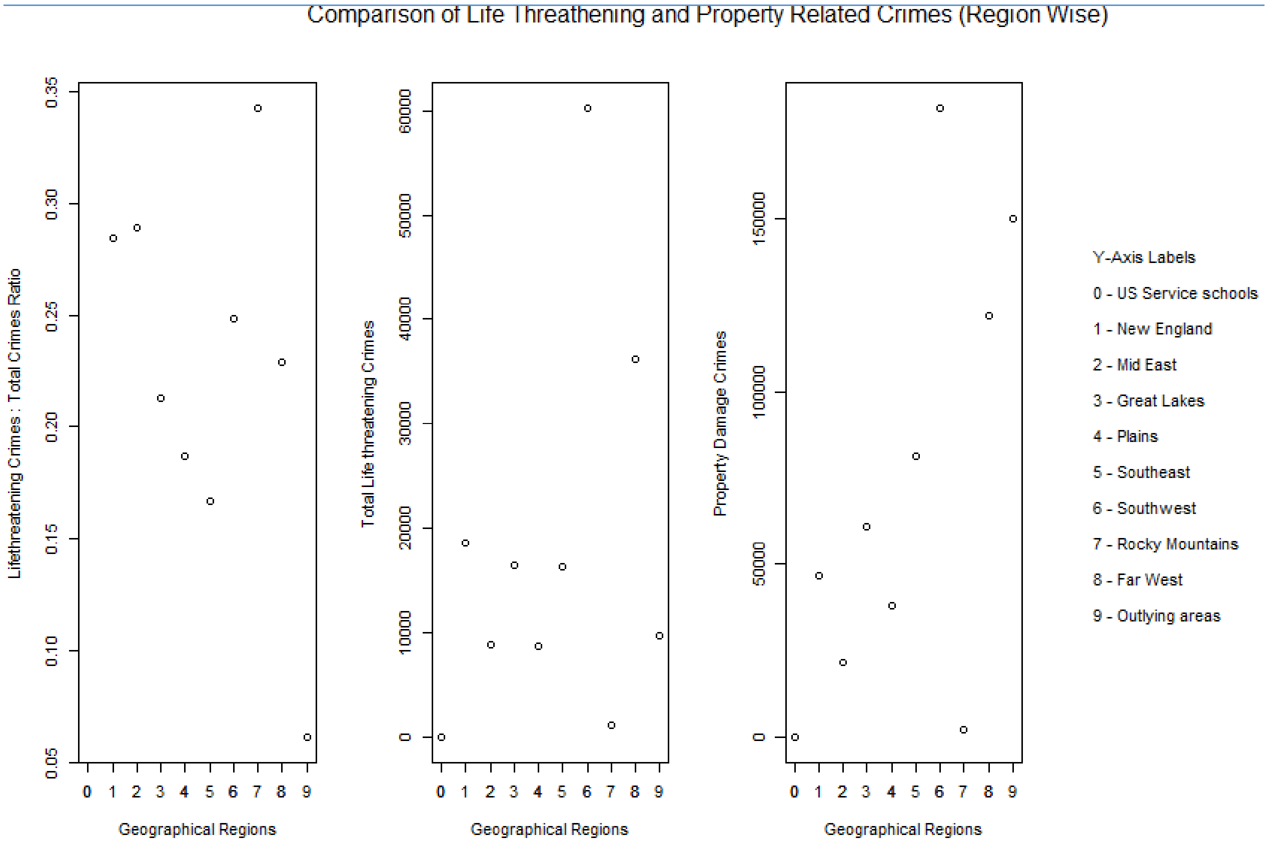
A scatter plot will help us visualise the relationship between Arrests and Disciplinary actions on a case by case basis. The blue line in the middle is the line of linear regression for the two attributes. Since we start with the assumption that with an increase in one linearly affects the other, we can get an idea of how campuses might fall behind our assumption(or sometimes overdo it). As a popular concentration of points, there is lesser arrests over disciplinary actions take in each campus. But they do not fall behind very much, in fact if we can relax the linear line slightly more it is a comforting revelation that **most campuses do not shy away from using arrests as an option to strengthen security measures.**

✓ **Size does matter**

Isn’t it comforting to know that the maximum we could lose is our wallet , you know, instead of being raped or murdered. We will classify the crime counts into two categories based on the intensity of the crime as 1. ‘life threatening’ which will include murder, assault, sex offences and 2. ‘property damage’ which will include thefts, arson and other related violations. Any crime is unforgivable but then size does matter when it comes to intensity of the crime.

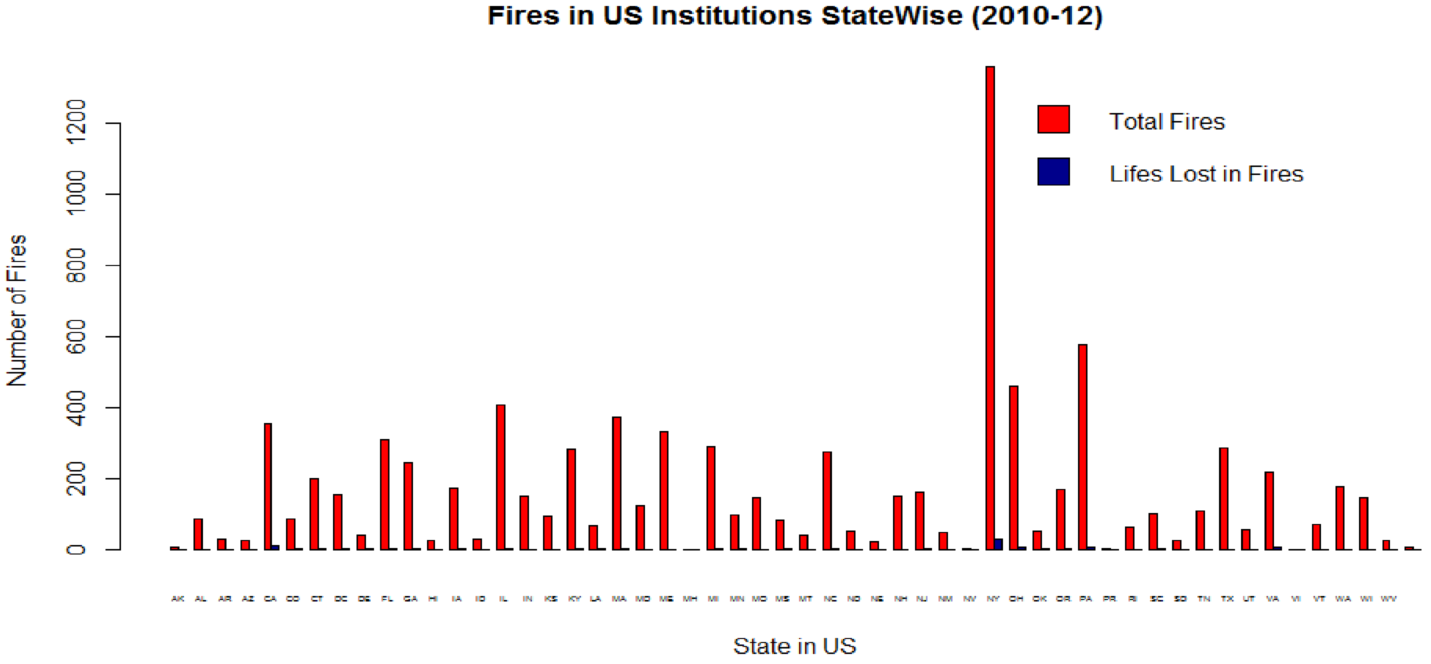
We shall also see which are the regions in the USA that also empathises with us on this sentiment to keep ‘life threatening’ crimes under check. 9 regions of our interest are New England, Mid East, Great Lakes, Plains, Southeast, Southwest, Rocky Mountains, Far West, Outlying Areas. We will plot the total number of life threatening crimes in each region, the total number of property damage related crime in each region and then find out who does better or worse in terms of a ration between these two.

Southwest and Far West lead in Life threatening crimes and property damages. One clearly has to think twice before choosing to study in colleges in these areas. Outlying areas have very less life threatening crimes in comparison to property damages.



✓ FIRE !!!

The importance of fire safety is paramount and needs no introduction. We shall consolidate the total incidents of fire and fire related life loses per state to try and get an idea of how vulnerable institutions in that area are for fire accidents. We will use a bar plot to plot this information.



New York leads the count (by a huge margin) on total fire accidents over the last 3 years. Pennsylvania, Ohio, Illinois and California are the other states which have recorded sizeable fire accidents. It is comforting to note that even though all the states have recorded sizeable counts of fire accidents there has been very little life losses due to fire accidents. This is to show that all campuses across the country maintain high standards in fire fighting systems.

Lessons Learned

<u>Exploratory Data Analysis using R.</u>

R can handle datasets of the order of few 100 MBs to a few GBs with good speed and effectiveness. Being open source, there are a lot of visualization tools in R which help us understand pictorially both the distribution in the dataset and trends in the data set. However, documentation of the functions and the support of technologies leaves a lot to be desired.

1. Looking for data sets
Getting a dataset can be a real challenge if you are not sure where to look for it. Generating data sets is a good option but needs some guidance. Finding a sizeable yet meaningful dataset for us was the tough part of the assignment in itself.

2. Cleaning and formatting the data set
Having found the dataset, cleaning it was not as simple as the exercise problem. Lots of NULLs, duplicates and out of range values had to be fixed. Factors had to be placed on categorizable values like year, state code, states, and sectors.

3. Understanding the variables
Understanding the data set and its attributes gives a nice idea on how to move forward towards experimenting with data analytics on this data.

4. Basic plots
We learnt to make a lot of basic graphs and plots including line, bar, scatter, box plots.

5. Clustering and Maps
Given a dataset with no prior knowledge on how it is going to behave for each record, we could cluster these into various categories by passing the relevant attributes. We could also used maps to plot the choropleth map of such a clustered data.

6. Analysis
Analyzing the plots, graphs and summaries was the biggest lesson of all. To make sense of the data using pictures and numbers to reveal details about a dumb data set was a nice lesson. It would be an interesting challenge to expand analytics for big data and to use Hadoop to do that. We are looking forward to the next project.

References Legend

[1] Doing Data Science www.amazon.com/Doing-Data-Science-Straight-Frontline/dp/1449358659
[2] Wikipedia EDA en.wikipedia.org/wiki/Exploratory_data_analysis
[3] Engineering Statistics handbook www.itl.nist.gov/div898/handbook/eda/eda.htm
[4] Campus Security ope.ed.gov-campus_security
[5] College Directory National-Center-for-Education-Statistics-NCES

[6] Campus Security about Page frequently updated.
[7] College Directory about - about page