Analysis and Data Mining two large data sets

using various Visualization Techniques

by

Rohit A.Khadse*(L20407353)*

Saurabh Mahajan *(L20\_\_\_\_\_)*

Dataset

We analysed two data sets from the link [<http://catalog.data.gov/dataset/campus-safety-and-security-survey-2013>]

1. OnCampusArrest (*oncampusarrest101112.xls*)
2. NonCampusArrest (*noncampusarrest101112.xls*)

We used R Programming and MySQL for this and tried to find where Arrest Rate is higher.   
OnCampusArrest data set has 24 attributes and 11064 tuple entries while as NonCampusArrest data set has 24 attributes and 11064 tuple entries. OnCampusArrest data set and NonCampusArrest data set shows the tabular data of on campus and non campus arrests that took place in various colleges. Following is data format in both data sets[1] :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables in Creation Order | | | | | | |
| # | Variable | Type | Len | Format | Informat | Label |
| 1 | UNITID\_P | Num | 8 |  |  | Unitid\_plus |
| 2 | INSTNM | Char | 93 | $93. | $93. | Institution Name |
| 3 | BRANCH | Char | 89 | $89. | $89. | Branch Name |
| 4 | Address | Char | 92 | $92. | $92. |  |
| 5 | City | Char | 28 | $28. | $28. |  |
| 6 | State | Char | 2 | $2. | $2. |  |
| 7 | Zip | Char | 14 | $14. | $14. |  |
| 8 | sector\_cd | Num | 8 |  |  |  |
| 9 | sector\_desc | Char | 36 | $36. | $36. |  |
| 10 | men\_total | Num | 8 |  |  | Total Men |
| 11 | women\_total | Num | 8 |  |  | Total Women |
| 12 | Total | Num | 8 |  |  | Grand Total |
| 13 | Weapon10 | Num | 8 |  |  | Weapons: carrying, possessing, etc. 2010 |
| 14 | Drug10 | Num | 8 |  |  | Drug Law Violations 2010 |
| 15 | Liquor10 | Num | 8 |  |  | Liquor Law Violations 2010 |
| 16 | Weapon11 | Num | 8 |  |  | Weapons: carrying, possessing, etc. 2011 |
| 17 | Drug11 | Num | 8 |  |  | Drug Law Violations 2011 |
| 18 | Liquor11 | Num | 8 |  |  | Liquor Law Violations 2011 |
| 19 | Weapon12 | Num | 8 |  |  | Weapons: carrying, possessing, etc. 2012 |
| 20 | Drug12 | Num | 8 |  |  | Drug Law Violations 2012 |
| 21 | Liquor12 | Num | 8 |  |  | Liquor Law Violations 2012 |
| 22 | FILTER10 | Num | 8 |  |  | Data\_year = 2010 (FILTER) |
| 23 | FILTER11 | Num | 8 |  |  | Data\_year = 2011 (FILTER) |
| 24 | FILTER12 | Num | 8 |  |  | Data\_year = 2012 (FILTER) |

We are going to analyze the two data sets and interpret the areas where the arrest rate is high i.e. either On Campus or Non Campus. We will also use different visualization techniques to analyse datasets.

Preprocessing and Implementation

We had to preprocess data since most of Arrests/Violation related rows were blank (and not numeric). We imported files in MySQL database into OnCampus and NonCampus table using built in import tool in MySQL workbench. We applied DML queries to modify and clean both tables so we can query and perform analysis and visualization with it.

Results

Scatter Plot

|  |  |
| --- | --- |
|  |  |

Above is Scatter Plots for Total Number of Arrests vs Total number of People who live on-campus and non-campus. After comparing both scatterplots we can say that distribution of number of arrests and total number of people is similar. Hence Data set is ideal for comparison of Arrest Rate between two datasets.

Histogram

|  |  |
| --- | --- |
|  |  |

Histogram shows us distribution of combined women and men population in both data sets. We can deduce that most of colleges and universities have 2000~5000 population of men and women.

Bar Graph

|  |  |
| --- | --- |
|  |  |

Arrest Rate can calculated as follows : [2]  
= Total Number of Arrests/Total Population \* 100000  
For example. If arrest rate is 678, then we say that on an average 678 people got arrested **per 100,000** people.   
  
Above Bar Graph shows year wise (for three years: 2010, 2011, 2012) on-campus and non-campus arrest rate in datasets. We can clearly see on-campus arrest rate is significantly higher than non-campus arrest rate.

Pie Chart

|  |  |
| --- | --- |
|  | |

Pie Chart show arrest rate difference for On - Campus and Non-Campus

K-means Plot Graph

|  |  |
| --- | --- |
|  |  |

K-means plot graph for on-campus:   
As Arrest Rate clusters towards positive zero we can conclude that On campus arrest rate is higher where population is low. Two Dividing clusters clearly shows that small private and public institutions (with low on-campus population) lack

security infrastructure and planning and hence on-campus arrest rate is higher in less populated colleges.

K-means plot graph for Non-campus:   
As Arrest Rate clusters towards positive infinity we can conclude that Non campus arrest rate is higher when population is high. Two Dividing clusters clearly shows that private and public institutions (with large non-campus population) lack enough security infrastructure and planning to handle large population and hence Non-campus arrest rate is higher in largely populated colleges.

Heat Map

|  |  |
| --- | --- |
|  |  |

Heat map of On-campus Arrest Rate shows West Virginia, South Dakota and Wyoming has highest on-campus arrest rate in US. While as heat map of Non-campus Arrest Rate shows North Dakota, Vermont and Wyoming has highest Non-campus arrest rate in US.

|  |
| --- |
| heatmap.png |

Above heatmap shows us that North Dakota, Vermont and Wyoming has highest Non-campus arrest rate in US.

Conclusion

From above visualized data diagrams , we can safely conclude that on-campus arrest rate is higher compared to non-campus arrest rate.

References

[1] <http://catalog.data.gov/dataset/campus-safety-and-security-survey-2013>

[2] <https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/persons-arrested/persons-arrested>