|  |
| --- |
| Homicide Analysis Report |
| 2017 Term project |
|  |

By:

*Nitish Gaddam*

*U89827204*

*nitishg@bu.edu*

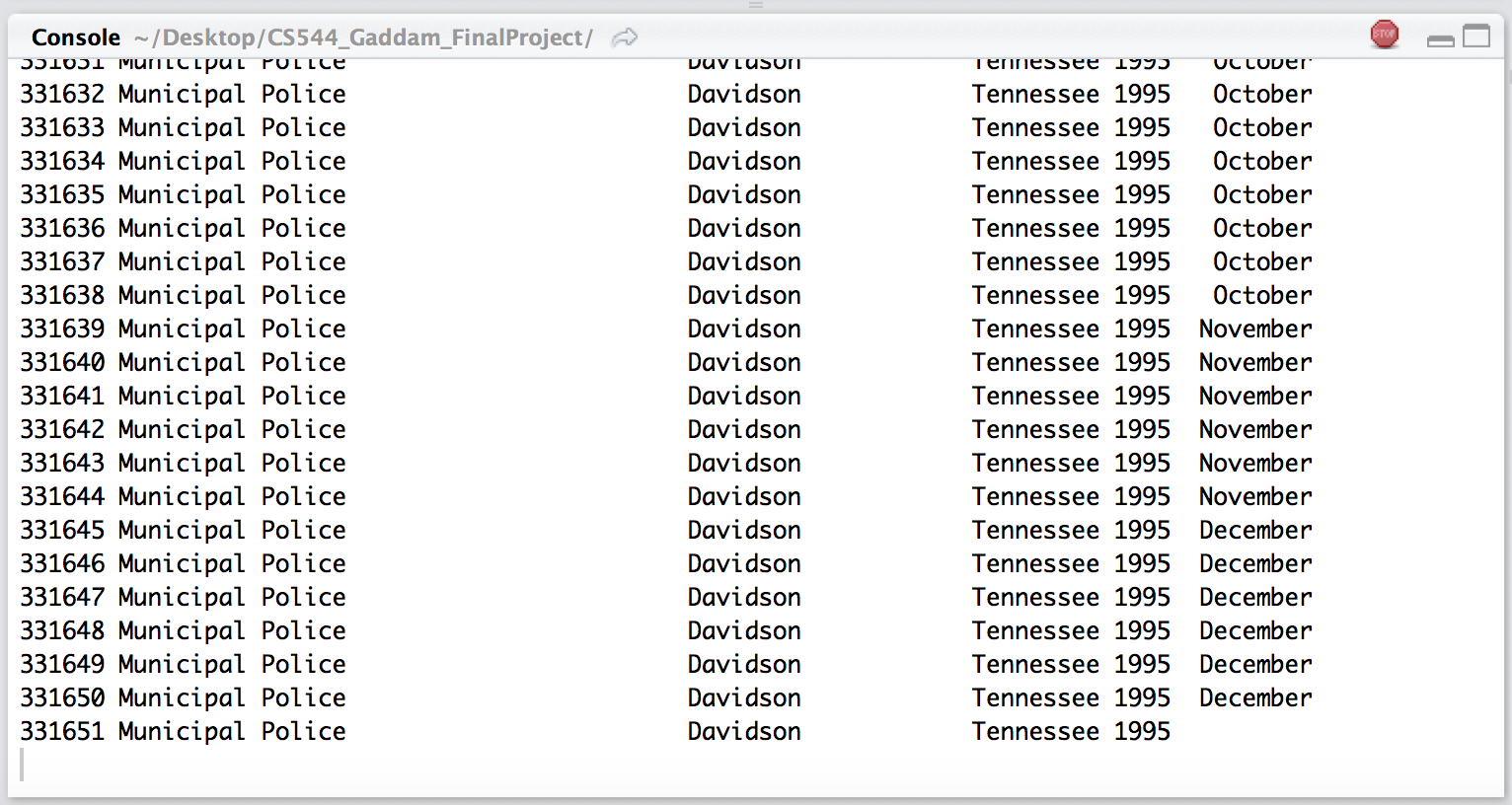
**Introduction**

I picked the **Homicide** dataset from <https://www.kaggle.com/murderaccountability/homicide-reports>. This dataset includes murders from the FBI's Supplementary Homicide Report from 1976 to the present. This dataset includes the age, race, sex, ethnicity of victims, in addition to details such as crime solved status, crime and its crime type was reported state wise.

So, I used R programming and its various techniques to analyze and visualize the data.

**1.** **Installing the libraries and Reading the file**:

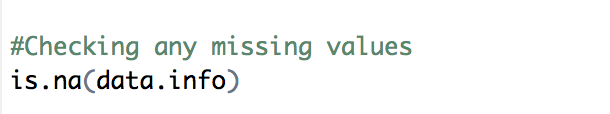




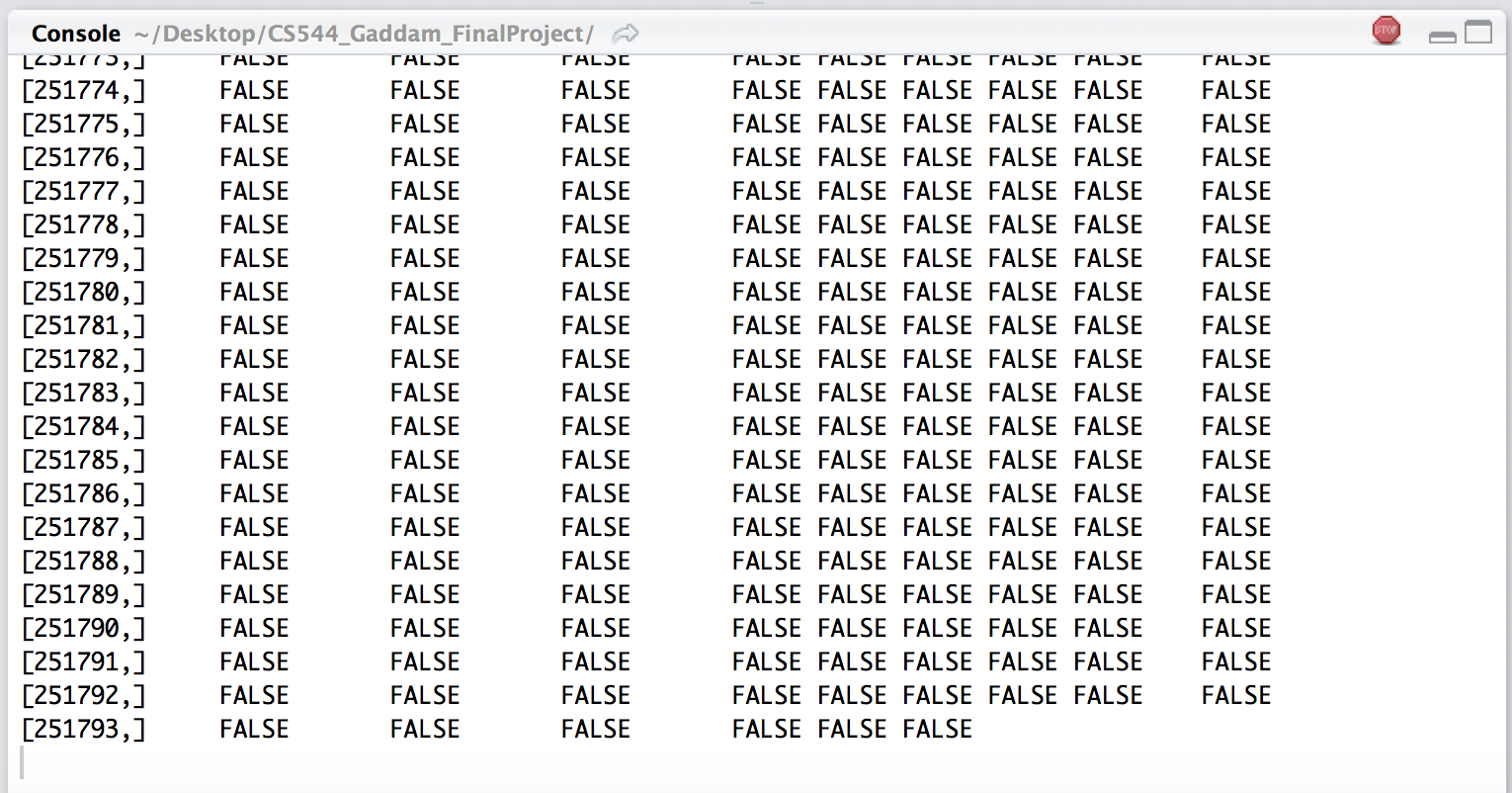
**2. Preparing the data:**

Pre-processing the data prior analyzing is really important if the data contains NA or there are some missing values.

I checked whether there is any missing value or not:



There were ***no*** missing values

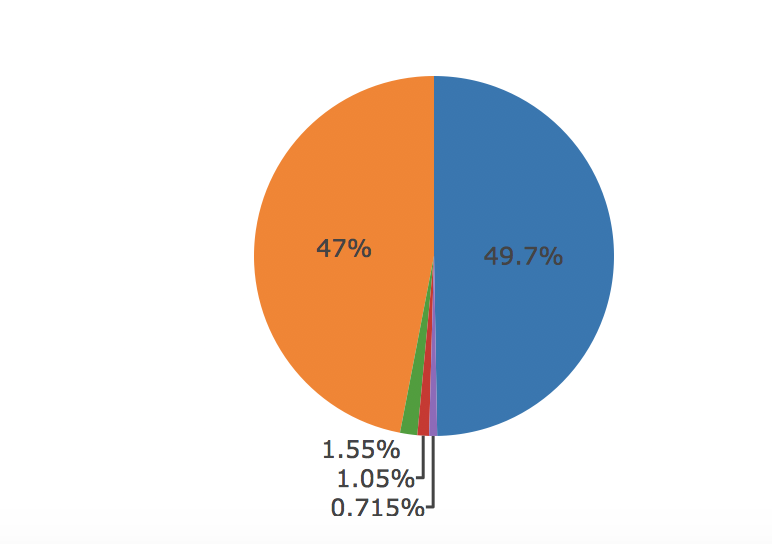


**3. Analyzing the data**

**A) For Categorical Attributes**, I took **State** and **Victim Race** into account.

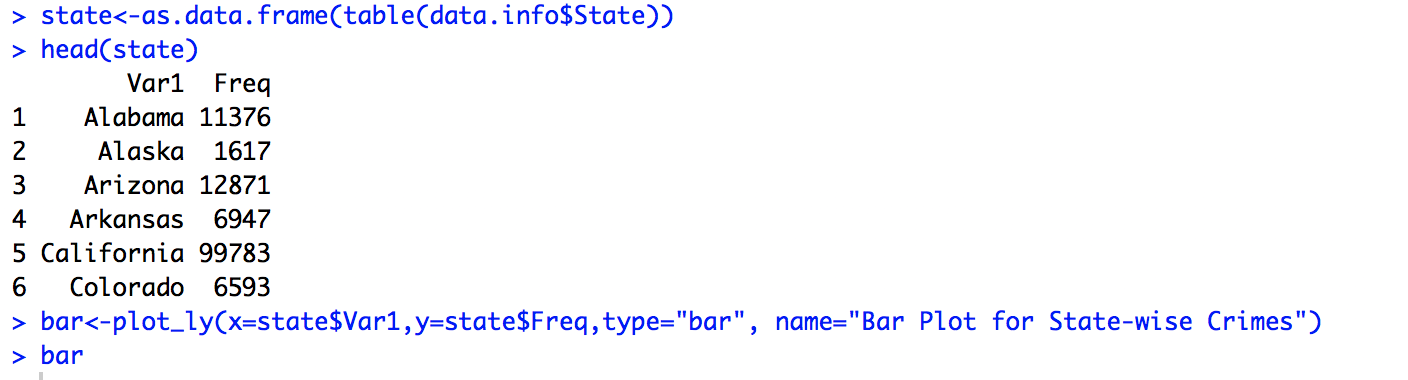
* **Victim Race:** Through pie chart, I analyzed 50% of population which makes ***White race*** is highly victimized. While, ***Native American/Alaska Native*** shares mere 0.7% of population being victimized.

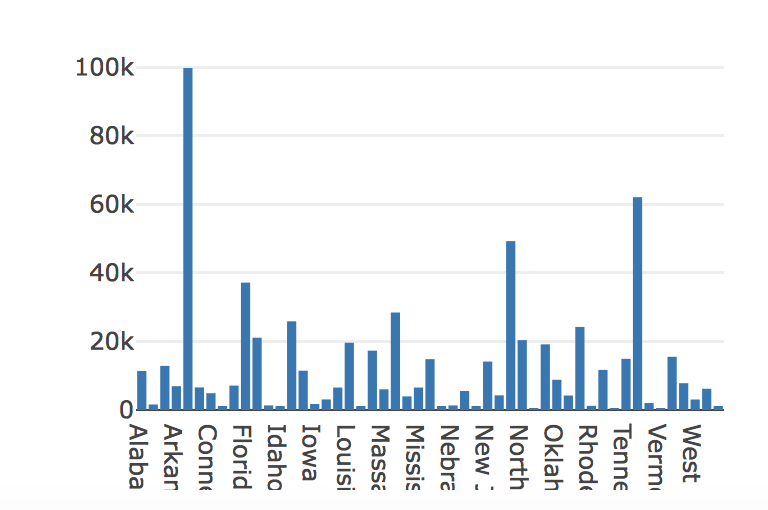




* **State:** Which state has highest homicide rate as reported by FBI is analyzed by Bar Plot.

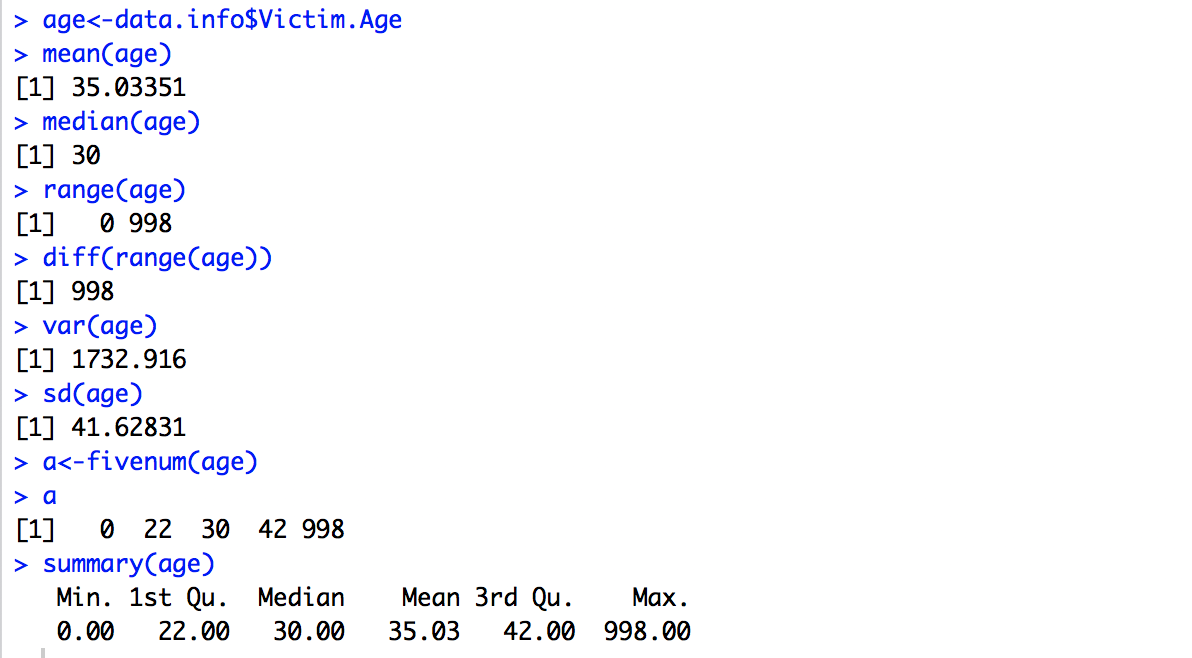
***California*** leads as state with highest homicide rate.





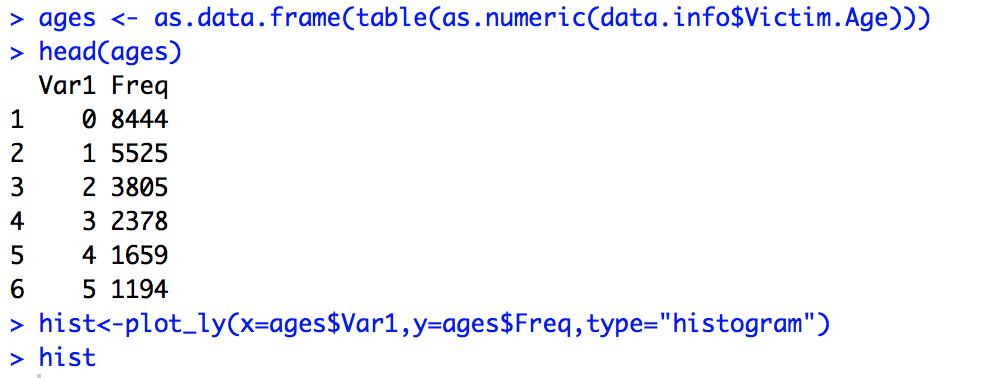
**B) Numerical Attribute**: ***Victim\_Age*** fits for analyzing the numerical attribute.

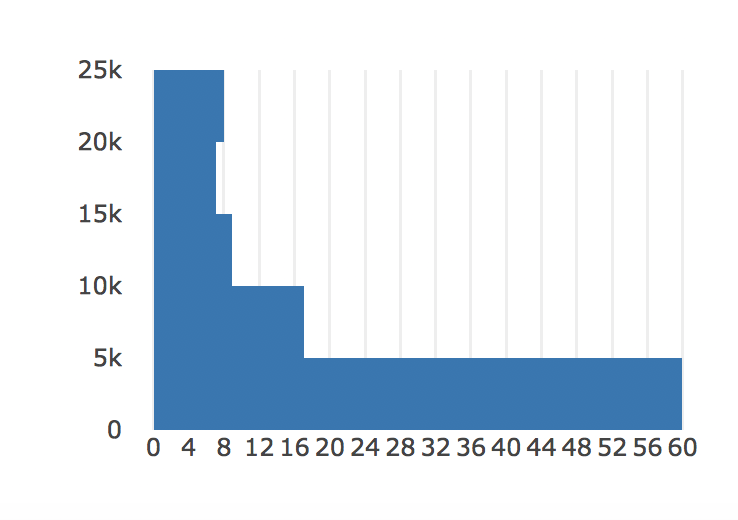
* Visualizing the age of victims through mean, standard deviation and other concepts gives the idea which age is most prone to murders.



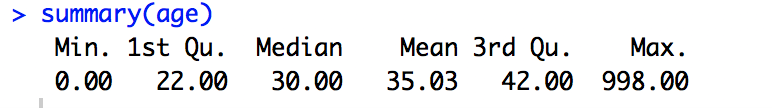
As we see that mean is larger than median, so the distribution has positive skew. Hence, ***age below 30 is more prone to homicide as compared to ages above 30.***

I plotted Histogram to further show the ***distribution of ages*** is skewed to ***right***.

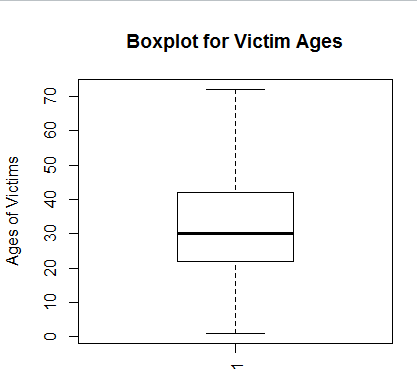




* **Box Plot:** Distribution of Victim Age based on five number summary:



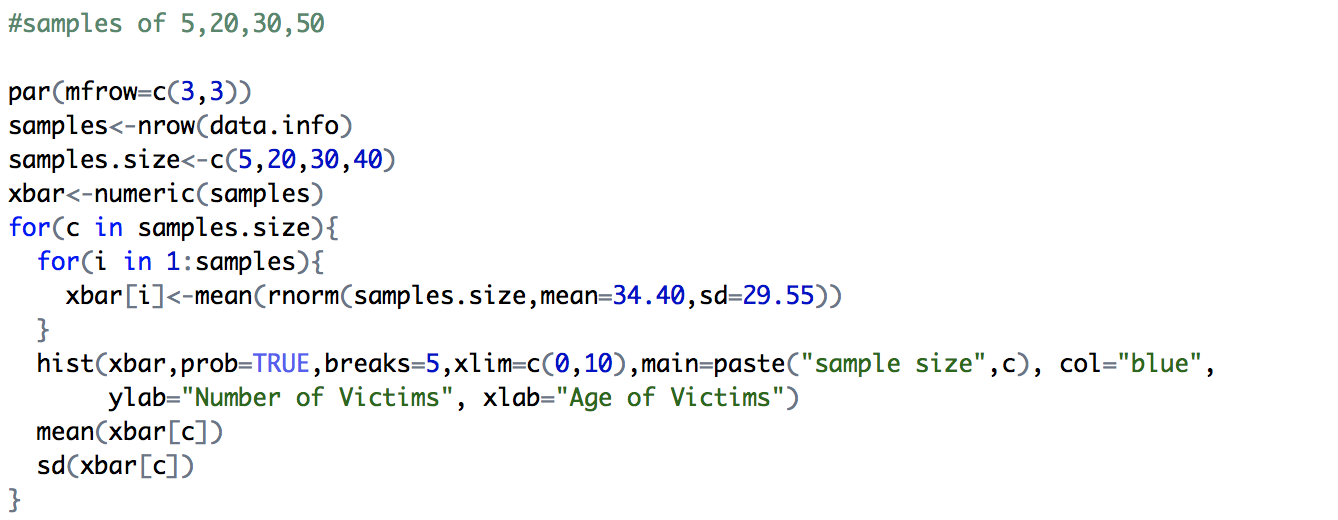


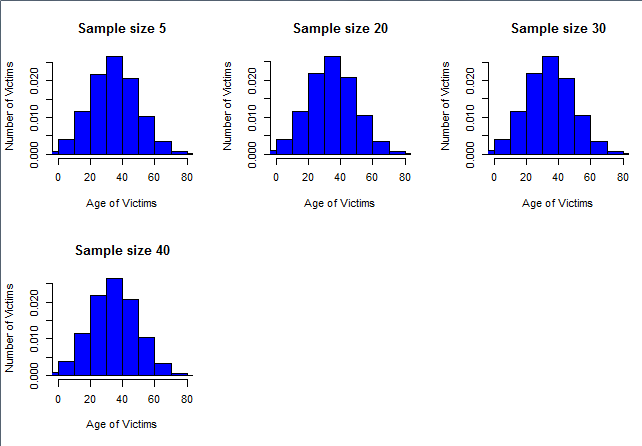


**4) Central Limit Theorem**

The Central Limit Theorem states that the distribution of the sample means for a given sample size of the population has the shape of the normal distribution.

For the sample sizes (5, 20, 30, 40), the mean of all the sample means is equal to the mean of the population (34.41). A histogram of the densities of the distribution of the sample means is shown below.



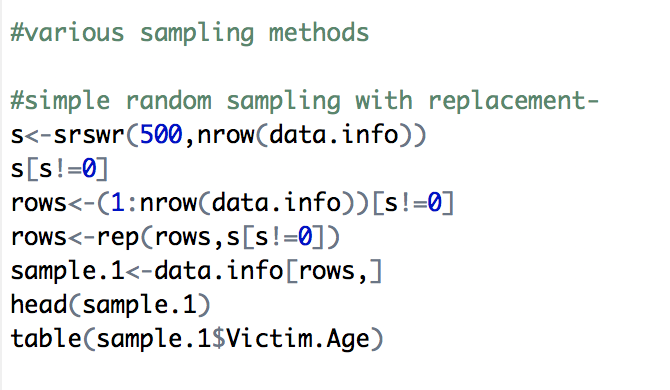


**5.** **Sampling Methods:**

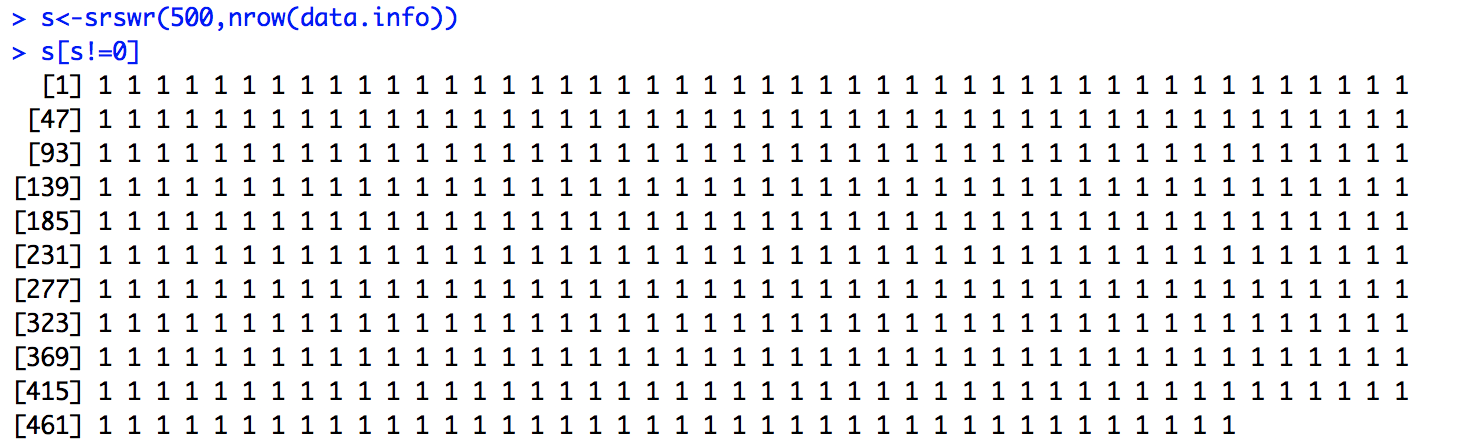
A sample is a portion of the population that is selected for doing the data analysis. The results from this sample are then used to estimate the characteristics of the population.

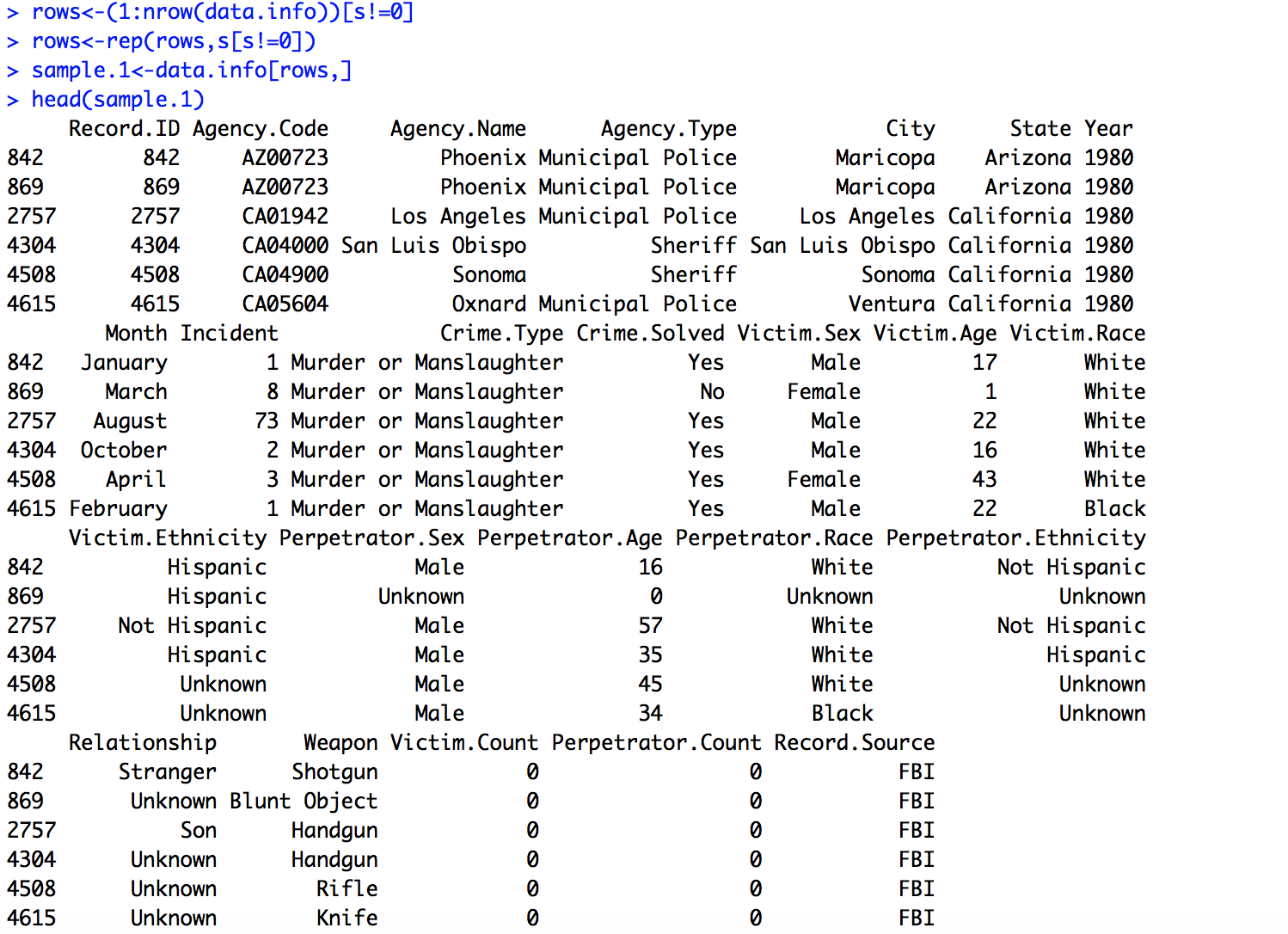
Methods used:

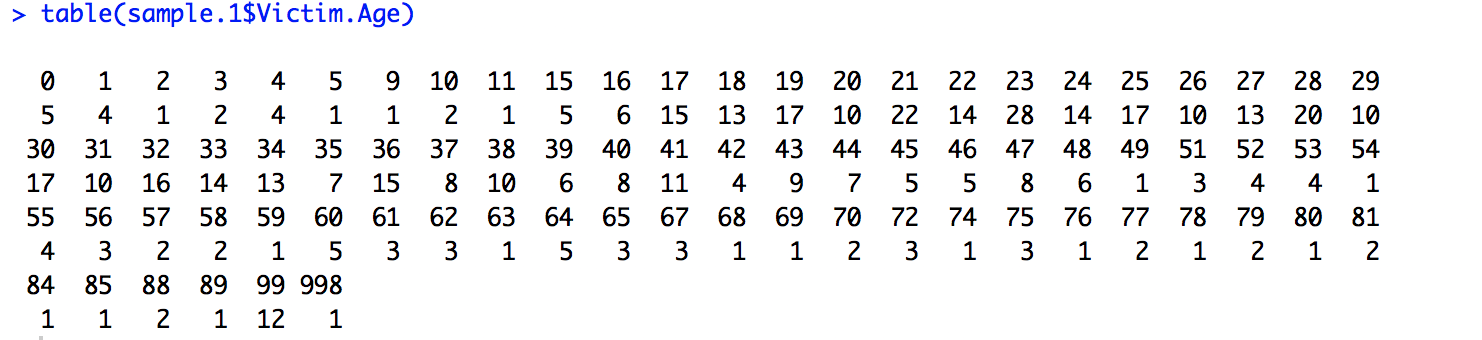
* **Simple Random Sampling with Replacement:** A simple random sample of size 500 is drawn from the population with replacement.



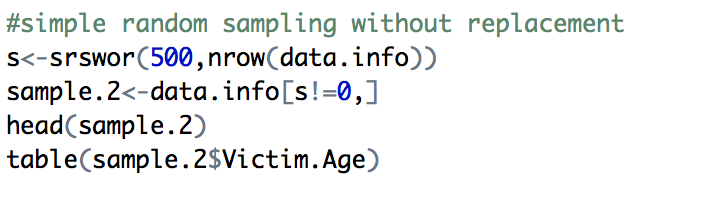
The data of the selected sample and the frequency of Victim age in each state are shown below:



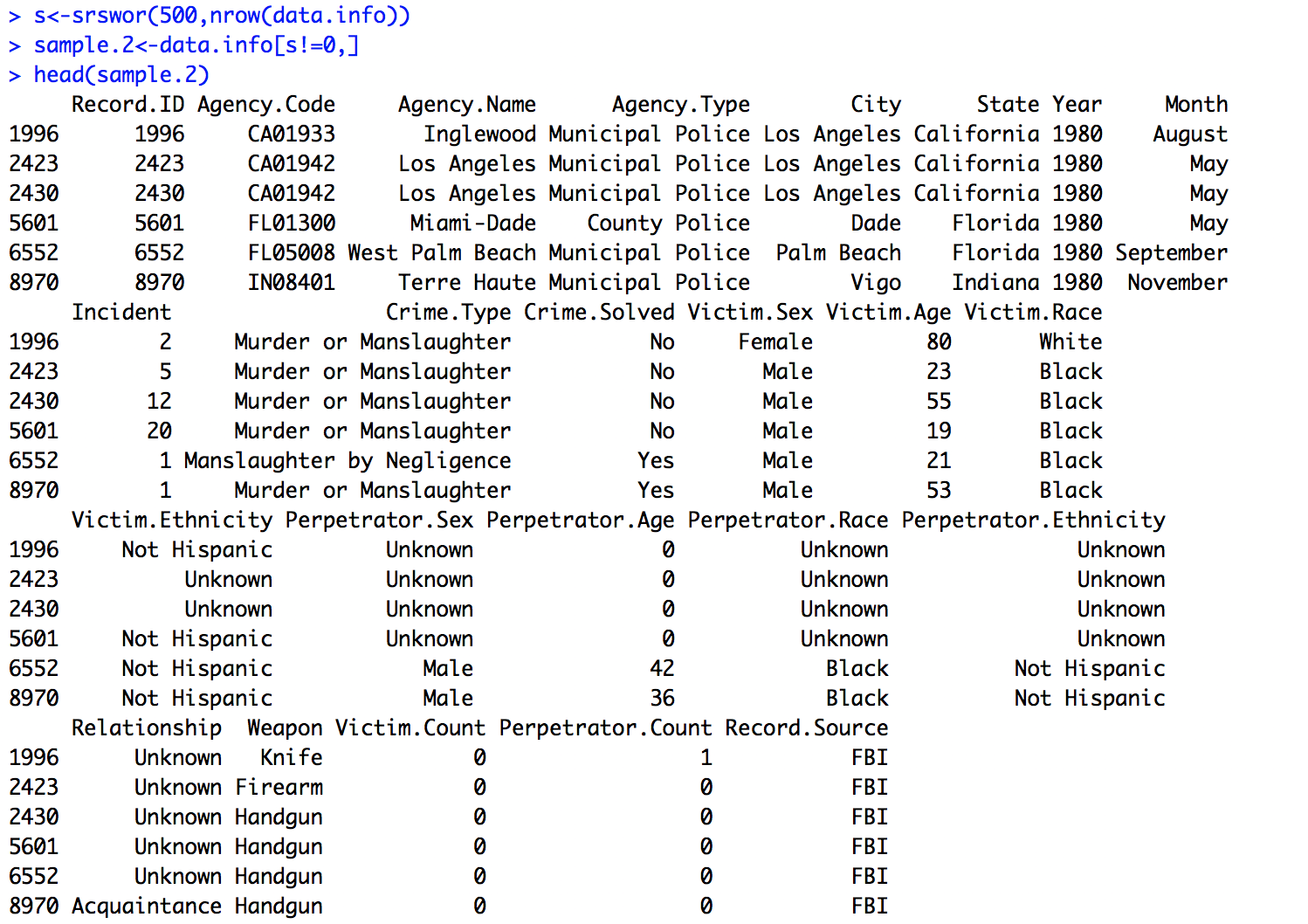


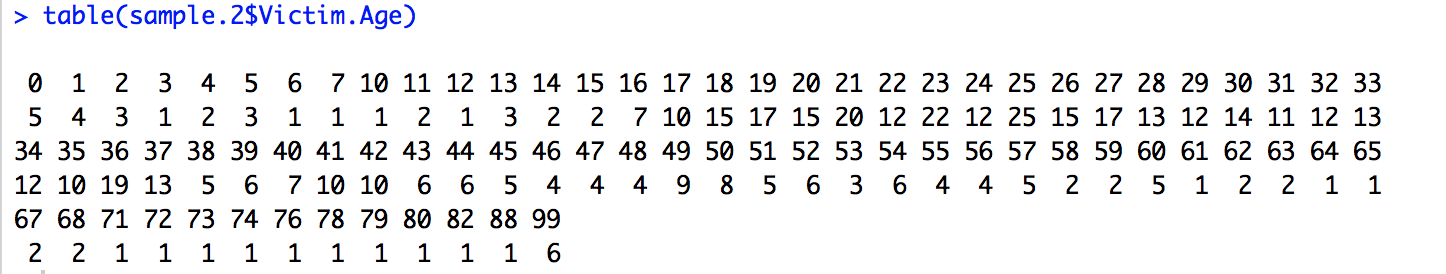


* **Simple Random Sampling without Replacement:** A simple random sample of size 500 is drawn from the population without replacement.



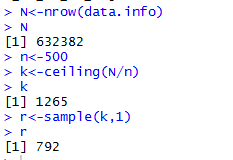
The data of the selected sample and the frequency of victim age in each state are shown below:





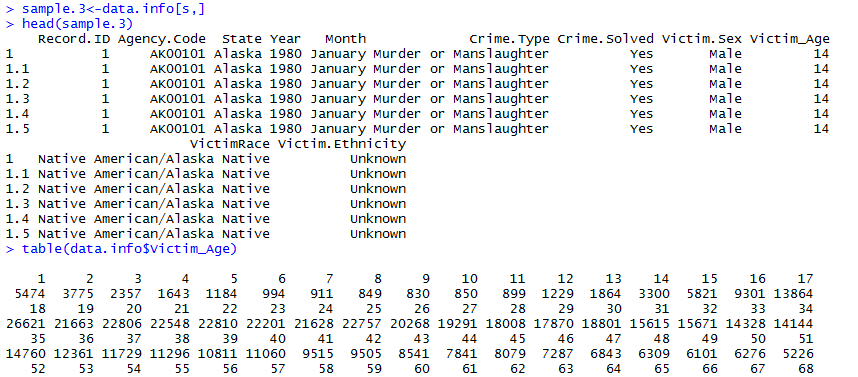
* **Systematic Sampling:** From a population of more than 6lakhs victims, if a sample of size 500 is to be selected, and then data is divided into 1265 groups.

From the first group, a random item is selected:

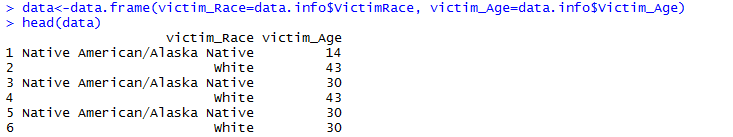


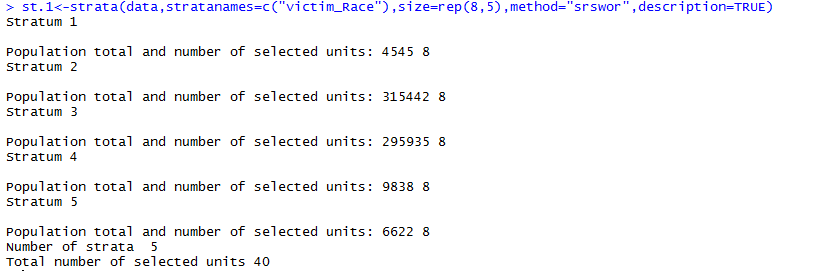
The rows of the systematic sample are now computed by taking every 1265th item. The selected sample is indexed from these rows.

The frequency of victim ages in each state is shown below:



* **Stratified Sampling:** A stratified sample of size 8 from each section is sampled without replacement using the strata function as shown below:



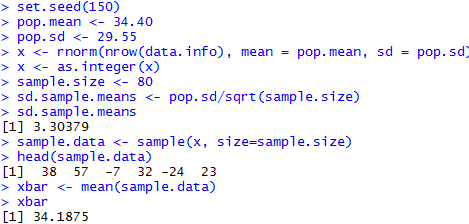


* **Cluster Sampling:** A cluster of size 8 can be selected from data, clustered by Victim age.

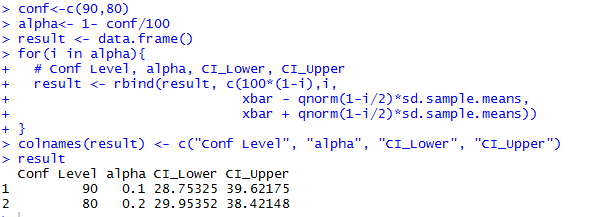


**5) Confidence interval:**

The sample data of size 80 used is shown below:



The confidence intervals for the confidence levels 90% and 80% and are computed as shown below



**The populations mean lies between the ranges of both two confidence intervals.**

The length of the confidence interval provides the precision for the estimate. Short confidence intervals provide good precision, whereas long confidence intervals provide poor precision. Hence, ***90% confidence interval is more precise than 80% confidence.***