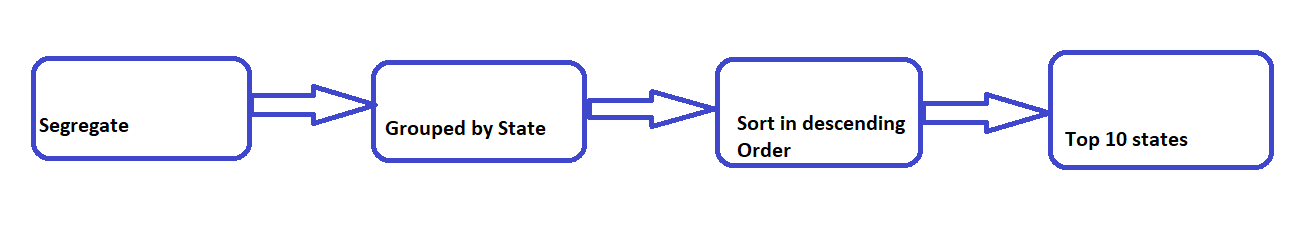
**INTRODUCTION**

A total County Migration and Population for 2017 is provided in the data using mainly two sources US Census bureau and other with the governing website. Counties of all the 50 states are included along with the District of Columbia. Migration rate for domestic, international net domestic as well as international migration rate along with the natural change I given. In the below examples we have covered highest population, frequency distribution for various migration rate, descriptive statistics, cumulative frequency, outliers, correlation etc. with the help of histogram, scatter plot, pareto chart. This can be done using various method but here R has been used because of its efficiency and ease of use. Below we will cover all the examples one by one in descriptive manner.

**ANALYSIS**

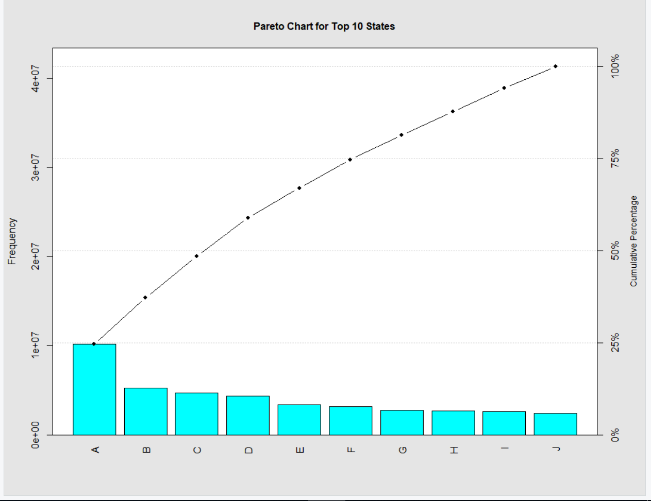
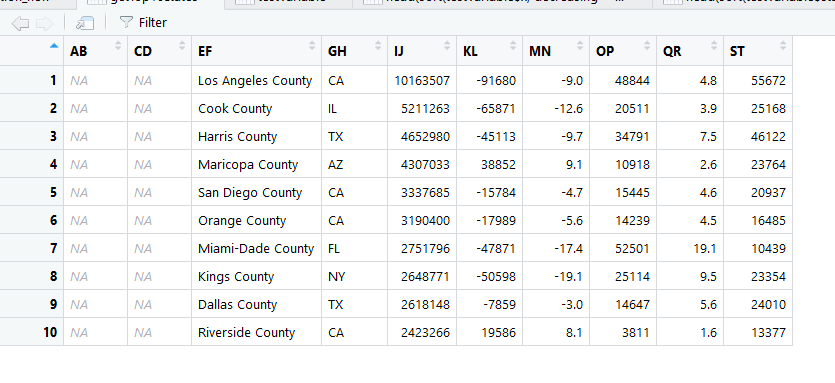
**Problem Statement 1: -**

In the first example elaboration of Pareto chart is given in which we are identifying population of top 10 states in USA. A step by step approach is required here to cleanse the data as below: -



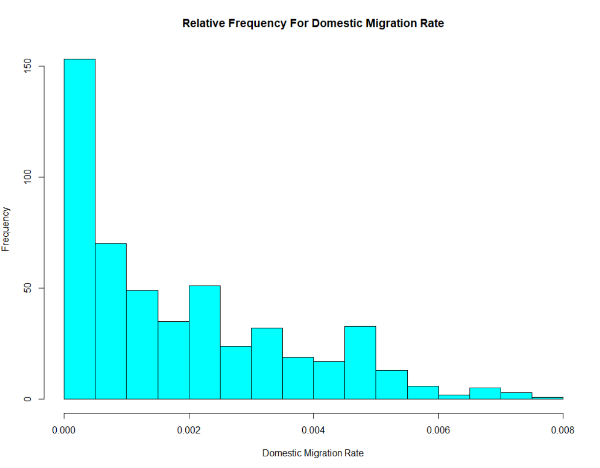
Pareto chart provides the combination of bar and line graph for different states. The line gives an ease in finding the cumulative total for the state. Initially the counties are divided based on the respective states and then group by is applied

so we get the population as per each state. Once the data is grouped as per the states it is easy to segregate top 10 states from it. Here we are using aggregate function for group. Once the data is sorted pareto chart is created. As per the chart it can be concluded that CA has the highest population in 2017.

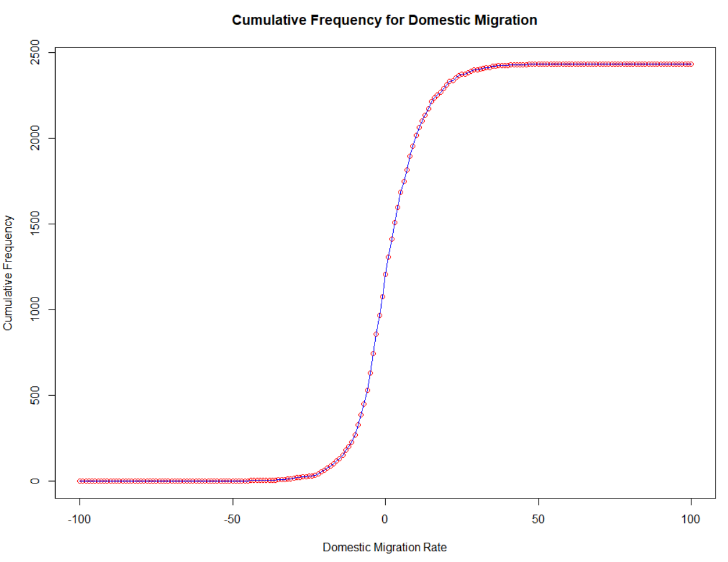
**Problem Statement 2: -**

The function to create histogram in R is predefined once the data is defined implementation can be done directly in R. Matrix function is used to derive the relative frequency and then it is graphically represented by using histogram. As from the diagram it can be observed that the highest number of times the domestic migration occurred was in the beginning for all counties and the depletion is done at the end of the year. The rate is decreased drastically after the first occurrence and then the changes are gradual.



**Problem Statement 3: -**

Cumulative frequency is in simple words running total of all frequencies. A cumulative frequency graph will give a graphical representation of the frequency distribution of Domestic migration rate for all the counties. Taking in consideration the range from -100 to 100 as keeping margin for 20+ on both negative and positive rate is defined by interval of 1. The rate of frequency is more towards the lowest and highest point of graph then compare to middle layer.

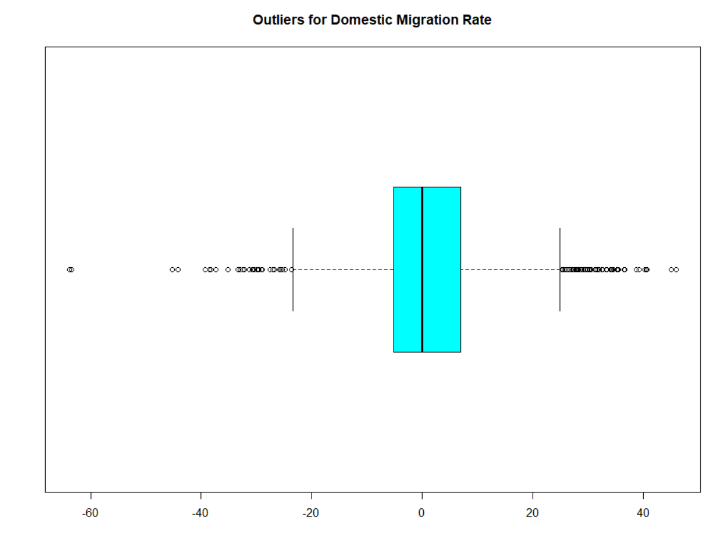


**Problem Statement 4: -**

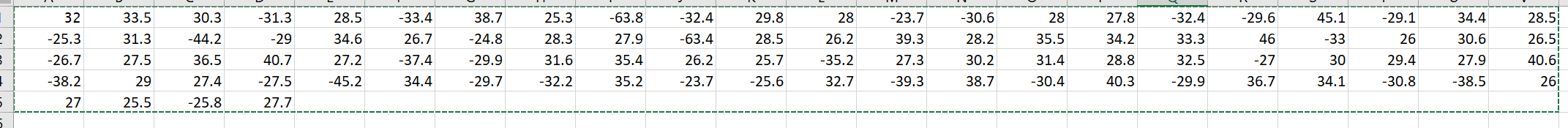
In the numerical descriptive statistics set of data is defined by numerical methods such as mean, median etc. R contains wide range of functions for obtaining the descriptive statistics. Two of the methods are used here. One of them is **summary** and other one is **sapply.** We can get Mean, Median, Variance, Range, Standard deviation etc. Both the methods will work efficiently in order to give the required output. The output is added below in the comparison of Domestic versus Intenational.

**Problem Statement 5: -**

Outlier is an observation point that is distant from other observations. It may indicate experimental error which can cause serious issues in statistical analysis. Here box plot is used to identify the outliers. And the outliers of all the counties are listed. Here I have used excel to determine the states in outlier after defining the outlier through R



Outlier Data:-



**Problem Statement 6: -**

Same tasks from second to 5th is repeated for International Migration rate. Below is the comparison point by point for Domestic and international Rate: -

* For the Relative frequency the coverage of International rate is broader as compared to the Domestic one as well the depletion is more drastic then domestic from first interval to the second. We can conclude from the graph that

initially the rate of migration is more in international as compared to domestic even though the gradual decrement is almost the same.

* In the cumulative frequency graph, the curve for International rate is steeper then domestic graph. We can conclude that the cumulative frequency for migration rate is very less in the neutral position. Though for both the frequency is more in the rage of 0 to -100 and 0 to -100 still there is considerable amount of difference in the curve of the graph.
* Regarding the descriptive statistics below is the example of both rates: -

Domestic Migration Rate: -

|  |  |
| --- | --- |
| **Min** | -63.8000 |
| **1st Qu.** | -5.2000 |
| **Median** | 0.0000 |
| **Mean** | 0.9188 |
| **3rd Qu** | 7.0000 |
| **Max** | 46.0000 |

International Migration Rate: -

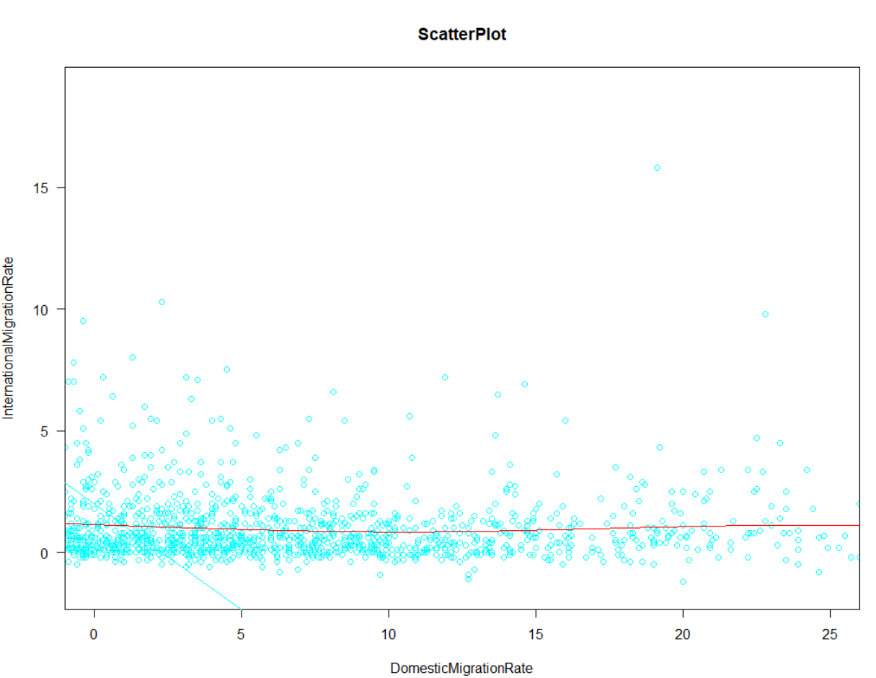
|  |  |
| --- | --- |
| **Min** | -1.500 |
| **1st Qu.** | 0.200 |
| **Median** | 0.600 |
| **Mean** | 1.244 |
| **3rd Qu** | 1.600 |
| **Max** | 19.100 |

The difference in the Mean, Median, Min, Max etc. is significant. For example, Mean for domestic is 0.9188 and for International rate is 1.244 which clearly indicates that the rate of migration for international is higher than domestic even though the maximum value of domestic rate is double than the International.

* The outliers are inclined on both the sides of box plot for domestic rate which ranges from -60 to -30 and 30 to 50. But for the international one the outliers are there only in one range i.e. 10 to 25. As compared to Domestic the removal of outliers international is easier but the number of outliers are more in the international plot.

**Problem Statement 7: -**

The Scatter plot is more efficient way to graphically display the data on 2 dimensions. Each dot represents observation with Domestic rate on X-axis and International on Y -axis. It is helpful in observing the relationship between two variables. We are defining correlation between this by using line. Here we can see that correlation exists between domestic and international rate as the Domestic rate is decreased there is decrease in the International rate as well. We can observe the rate is more less in the range of 10 to 15. However, the general tendency that Domestic rate decreases with the international is unquestionably present.



**PS**: - The input file used is uspopulation\_segregated in which the segregation of state from county is done in excel.

**CONCLUSION**

As from the above countless number of example we can observe that a single problem can be demonstrated using various methods. Even though we have number of software’s available for the usage but ‘R’ acts very efficiently here with its predefined functions. A relationship can be defined very effectively using the statistical analysis. Here we have observed the graphical representation of Migration rate for both International and domestic is presented in such a way that it is easy to make conclusion from it. Here the usage of the methods was limited and kept simple but we have various option to design the graph in more presentable format.