Business Analytics –Assignment 2

Saha Debanshee Gopal
U101113FCS074

AIM:

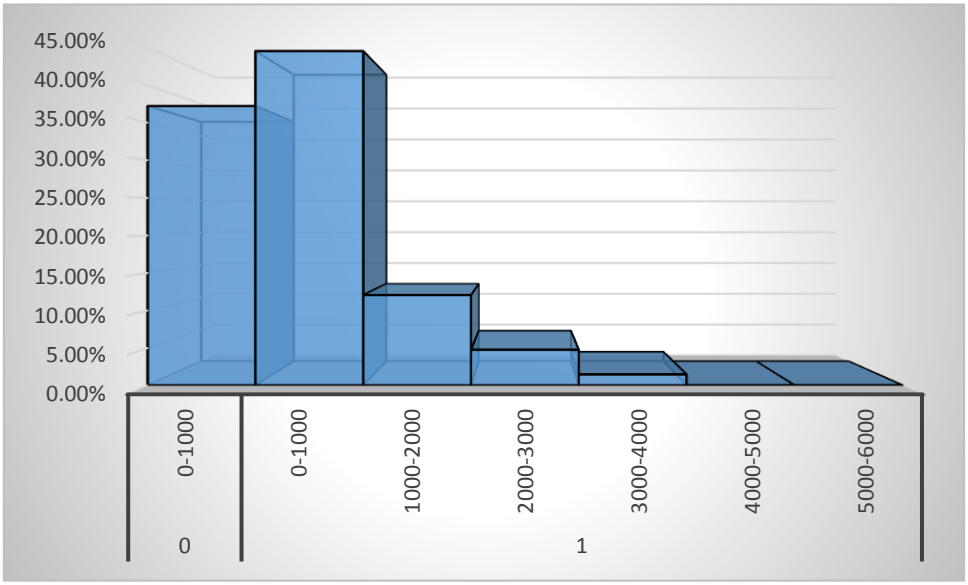
- To analyse the given data and find out the Z-score and Box plot to understand how the factors affect the data better.
- To treat the 0 values as missing values and analyse the data.
- To take 10 SRSWR samples and 10 SRSWOR samples and analyse the data.
- To take one SRSWOR sample using stratified sampling method of size 100 from Non-churners and 100 from Churner group and analyse the data.

SECTION 1:

In the table 1.a given below, we analyse how the acquisition of rental equipment effects the churn.
Here,
0= don't have rental equipment
1= have rental equipment
The sections are further divided according to the equipment rental over tenure.
The table shows that the people who don't have rental equipment churn 37.23%, all of them belonging to the 0-1000 equipment over tenure group.
It also shows that the people who have rental equipment churn the most, 62.77%, out of which the people belonging to 0-1000 equipment over tenure group churn 44.53% followed by the people belonging to 1000-2000 equipment over tenure group who churn 12.04% .

Table 1.a

Equipment over tenure	% churn
0	**37.23%**
0-1000	37.23%
1	**62.77%**
0-1000	44.53%
1000-2000	12.04%
2000-3000	4.74%
3000-4000	1.46%
4000-5000	0.00%
5000-6000	0.00%
Grand Total	**100.00%**



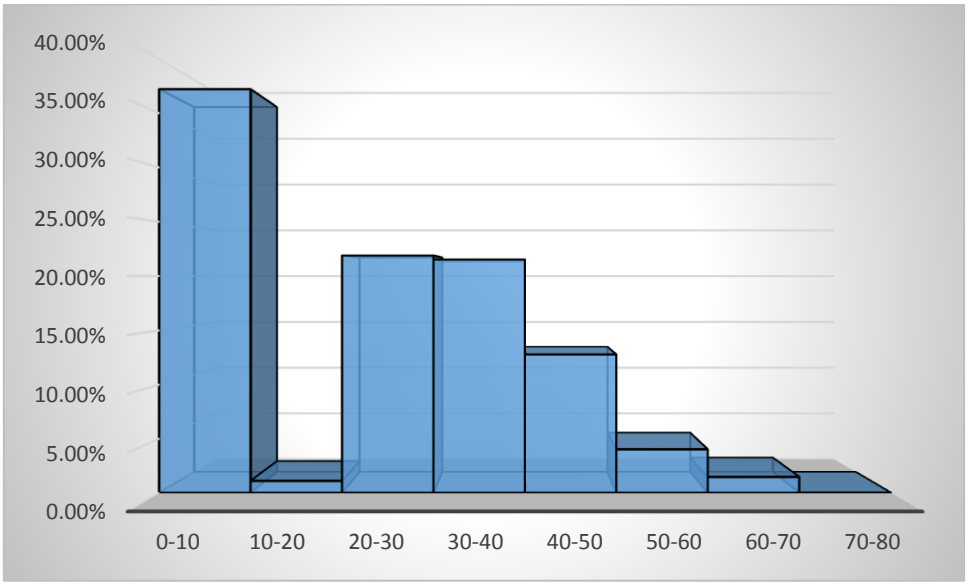
Histogram 1.a
(Rental Equipment over tenure vs Percentage Churn)

Histogram 1.a shows that the data is right skewed or positively skewed which signifies that the mass of the distribution is concentrated towards the left.

The data of equipment rental over tenure can also be further divided into equipment rental over last month, to analyse the cause of churn better. The analysis of the data shows that 37.23% churn was generated by the people who had rental equipment over last month in group 0-10 followed by 21.90% and 21.53% belonging to group 20-30 and 30-40 respectively.

Table 1.b

Equipment Rental over Last Month	% Churn
0-10	37.23%
10-20	1.09%
20-30	21.90%
30-40	21.53%
40-50	12.77%
50-60	4.01%
60-70	1.46%
70-80	0.00%
Grand Total	**100.00%**

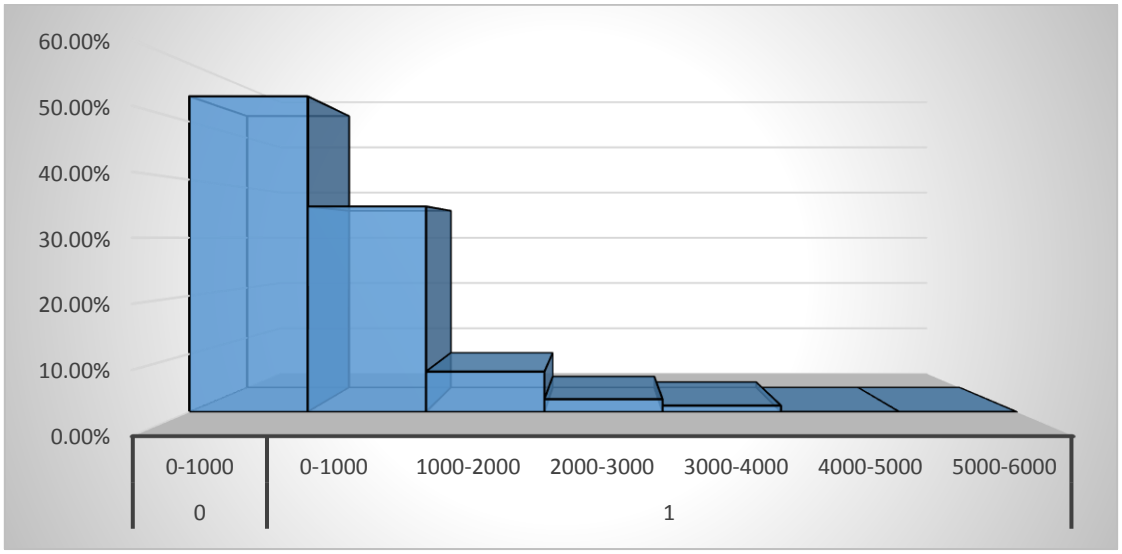


Histogram 1.b
(Rental Equipment over last month vs Percentage Churn)

In the table 2.a given below, we analyse how the toll free services effects the churn.
Here,
0= don't have toll free services
1= have toll free services
The section is further divided according to the toll free services over tenure.
The table shows that the people who don't have toll free services churn the most, 54.38%, all of them belonging to the 0-1000 toll free services over tenure group.
It also shows that the people who have toll free services churn 45.62%, out of which the people belonging to 0-1000 toll free services over tenure group churn 35.40% followed by the people belonging to 1000-2000 equipment over tenure group who churn 6.93% .

Table 2.a

Toll Free Services	%Churn
0	**54.38%**
0-1000	54.38%
1	**45.62%**
0-1000	35.40%
1000-2000	6.93%
2000-3000	2.19%
3000-4000	1.09%
4000-5000	0.00%
5000-6000	0.00%
Grand Total	**100.00%**



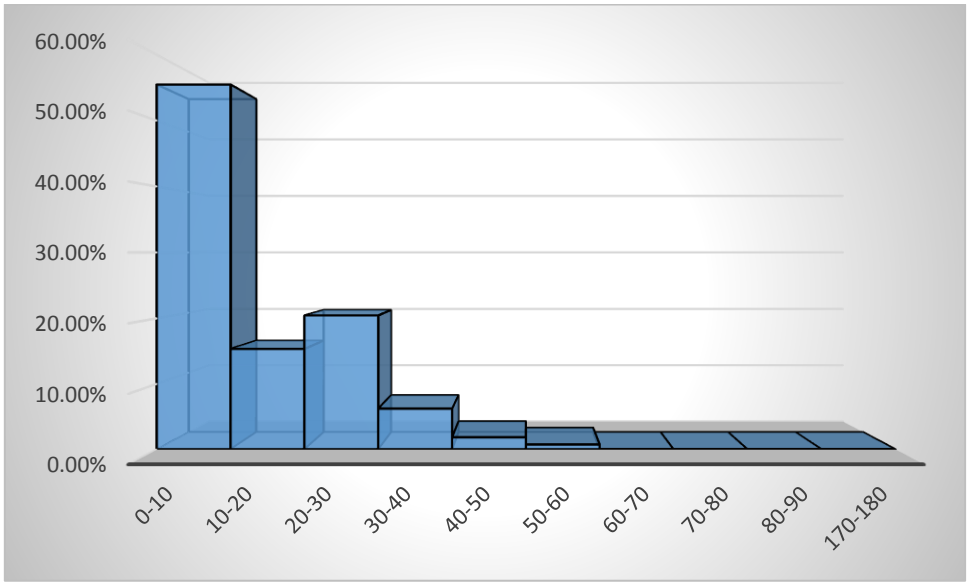
Histogram 2.a
(Toll Free Services over tenure vs Percentage Churn)

Histogram 2.a shows that the data is right skewed or positively skewed which signifies that the mass of the distribution is concentrated towards the left.

The data, toll free services over tenure, can further be analysed as toll free services over last month. Analysing tollmon or the toll free services over last month gives us a better understanding of the no of people who had toll free services in the last month and decided to churn. The data shows that the people belonging to the group 0-10 churned the most, 55.47% followed by the people belonging to the group 20-30% who churned 20.44%.

Table 2.b

Toll free Service last month	%churn
0-10	55.47%
10-20	15.33%
20-30	20.44%
30-40	6.20%
40-50	1.82%
50-60	0.73%
60-70	0.00%
70-80	0.00%
80-90	0.00%
170-180	0.00%
Grand Total	**100.00%**



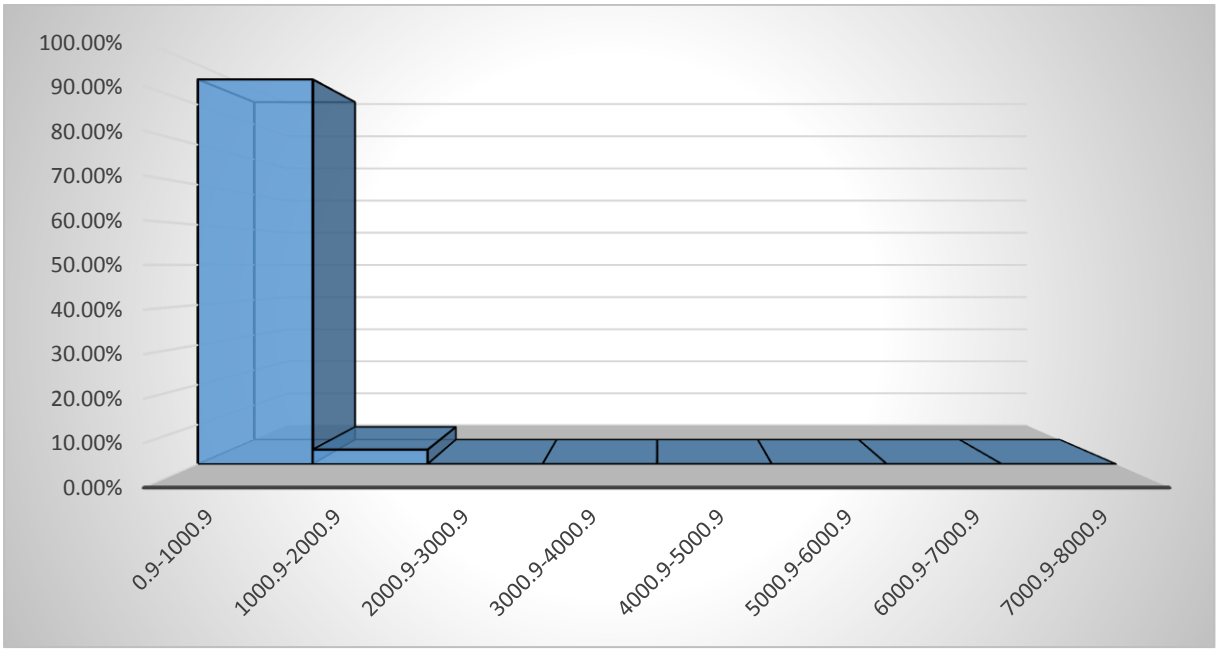
Histogram 2.b
(Toll Free Services over last month vs Percentage Churn)

Histogram 2.b also shows that the data is right skewed or positively skewed which signifies that the mass of the distribution is concentrated towards the left.

In table 3.a given below, we analyse how the long distance calls effect the churn.
The section is further divided according to the long distance calls over tenure.
The table shows that the people belonging to the long distance call over tenure group 0.9-1000.9 churn the most, 96.35% followed by the people belonging to 1000.9-2000.9 group who churn 3.65%

Table 3.a

Long Distance call over tenure	% Churn
0.9-1000.9	96.35%
1000.9-2000.9	3.65%
2000.9-3000.9	0.00%
3000.9-4000.9	0.00%
4000.9-5000.9	0.00%
5000.9-6000.9	0.00%
6000.9-7000.9	0.00%
7000.9-8000.9	0.00%
Grand Total	**100.00%**



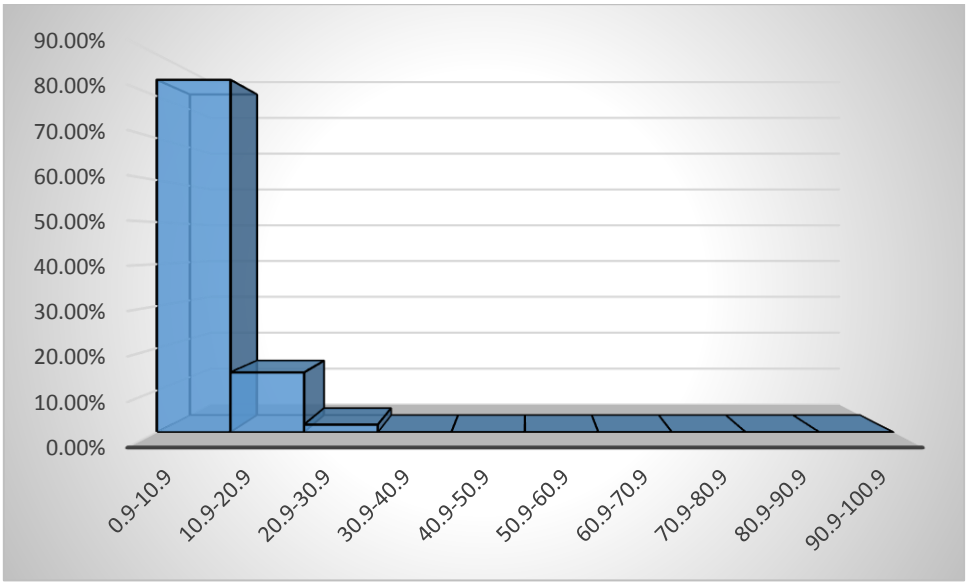
Histogram 3.a
(Long Distance Calls over tenure vs Percentage Churn)

Histogram 3.a shows that the data is right skewed or positively skewed which signifies that the mass of the distribution is concentrated towards the left.

The data, long distance calls over tenure, can further be analysed on the basis of the people who made long distance calls in the last month and churned. The data shows that 83.94% of the people who churned belonged to the 0.9-10.9 group followed by 14.23% belonging to the 10.9-20.9% group.

Table 3.b

Long distance call over last month	%churn
0.9-10.9	83.94%
10.9-20.9	14.23%
20.9-30.9	1.82%
30.9-40.9	0.00%
40.9-50.9	0.00%
50.9-60.9	0.00%
60.9-70.9	0.00%
70.9-80.9	0.00%
80.9-90.9	0.00%
90.9-100.9	0.00%
Grand Total	**100.00%**



Histogram 3.b
(Long Distance Calls over last month vs Percentage Churn)

Histogram 3.b shows that the data is right skewed or positively skewed which signifies that the mass of the distribution is concentrated towards the left.

Box plots characterize a sample using the 25th, 50th and 75th percentiles, also known as the lower quartile (Q1), mean (*m* or Q2) and upper quartile (Q3), and the interquartile range (IQR = Q3 − Q1), which covers the central 50% of the data. Quartiles are insensitive to outliers and preserve information about the center and spread. Consequently, they are preferred over the mean and s.d. for population distributions that are asymmetric or irregularly shaped and for samples with extreme outliers. In such cases these measures may be difficult to intuitively interpret: the mean may be far from the bulk of the data, and conventional rules for interpreting the s.d. will likely not apply. Since the above analysis showed that the data is asymmetric, plotting a box plot is more efficient to analyse the data better.

The table below gives the important box plot data- min, Q1, average, Q3 and max.

Table 4.

	Equipmon	Tollmon	Longmon
Min	0.00	0.00	0.90
Q1	0	0	5.2
Mean	14.22	13.27	11.72
Q3	31.475	24.25	14.4125
Max	77.70	173.00	99.95

The figure below consists of 3 boxplots where the 1st boxplot (green) represents the rental equipment over the last month (equipmon) while the 2nd (red) and 3rd (yellow) boxplots represent the toll free services over last month (tollmon) and long distance calls over last month (longmon) respectively.

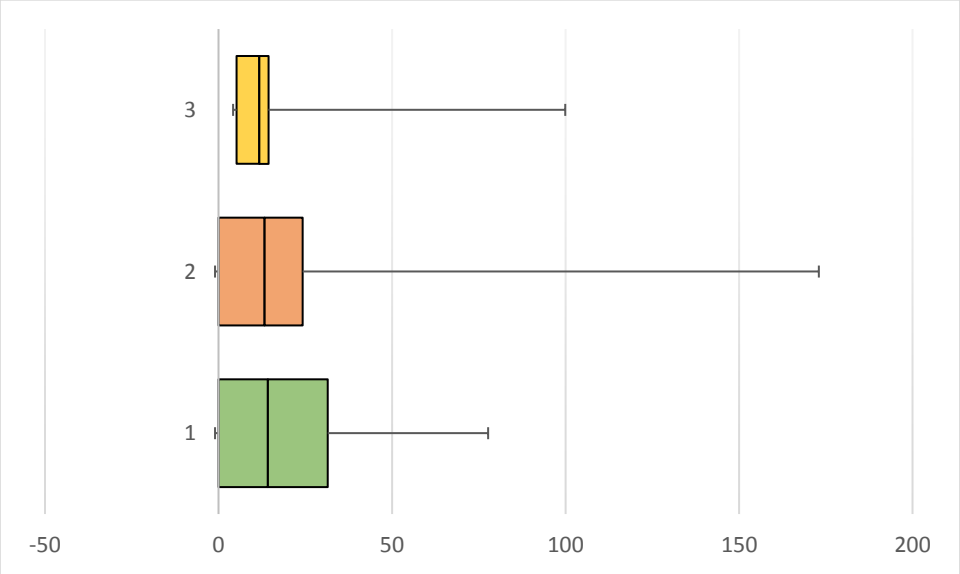
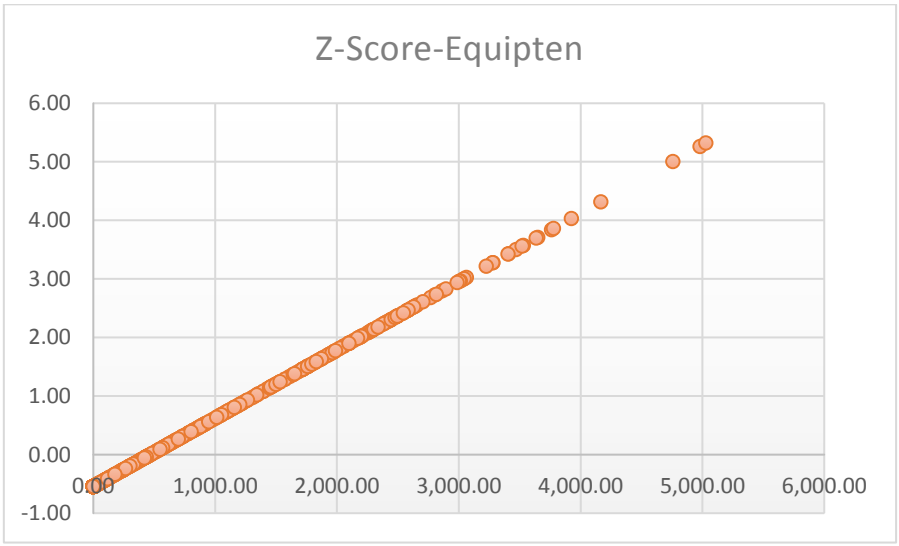
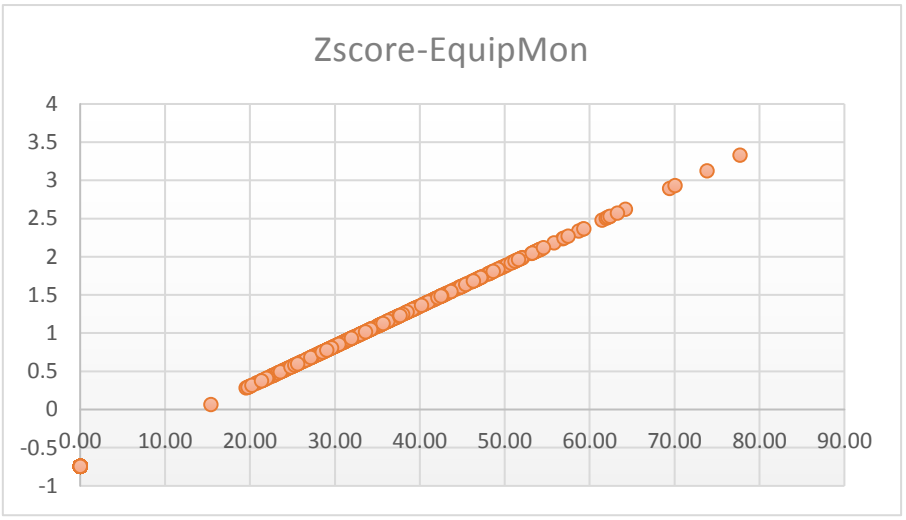


Figure 4.

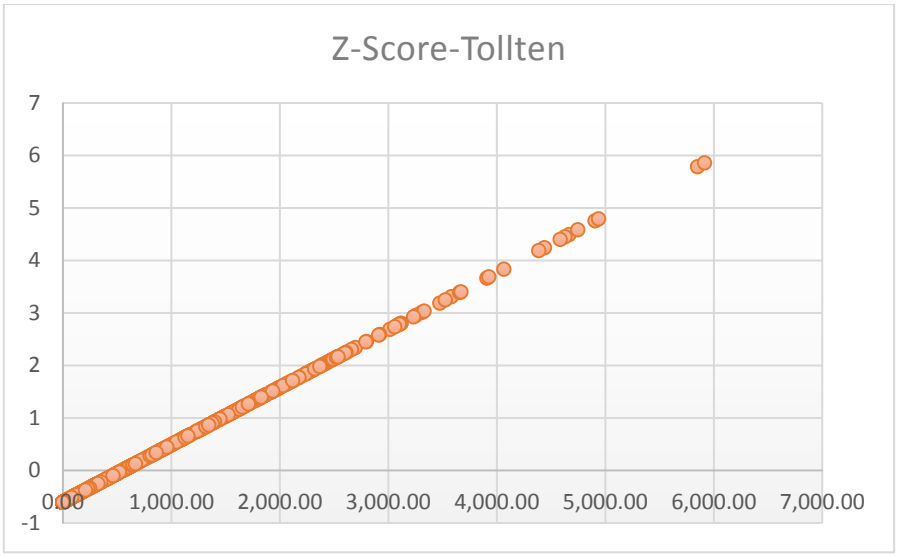
The standard score (more commonly referred to as a z-score) is a very useful statistic because it allows us to calculate the probability of a score occurring within our normal distribution and enables us to compare two scores that are from different normal distributions. The standard score does this by converting (in other words, standardizing) scores in a normal distribution to z-scores in what becomes a standard normal distribution.

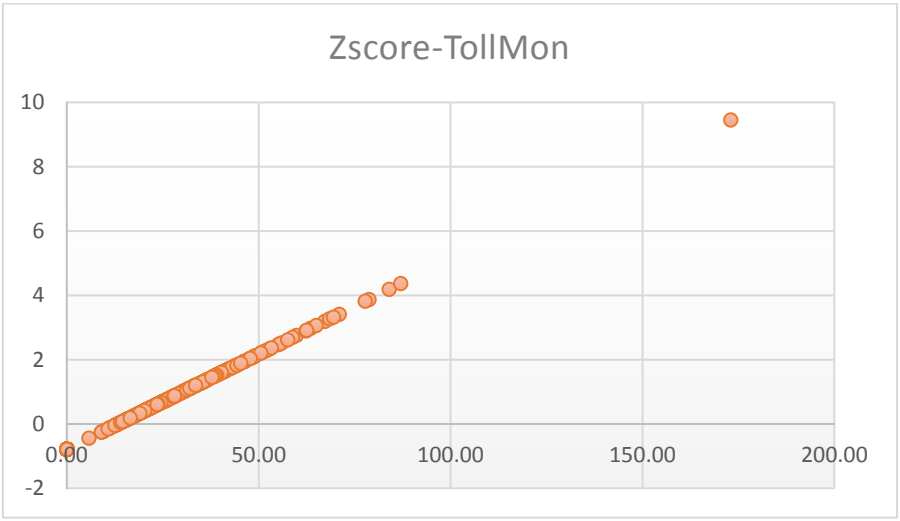
We know that 99.7% of the data lies within 3 standard deviations of the mean. The scatter plot on the data against its Z-score shows that there are data points which lie outside 3 standard deviations as well.

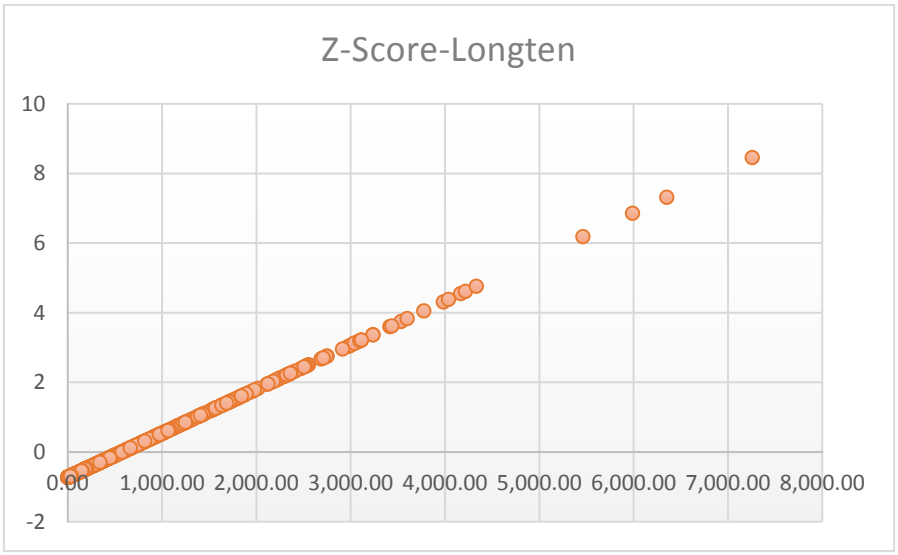


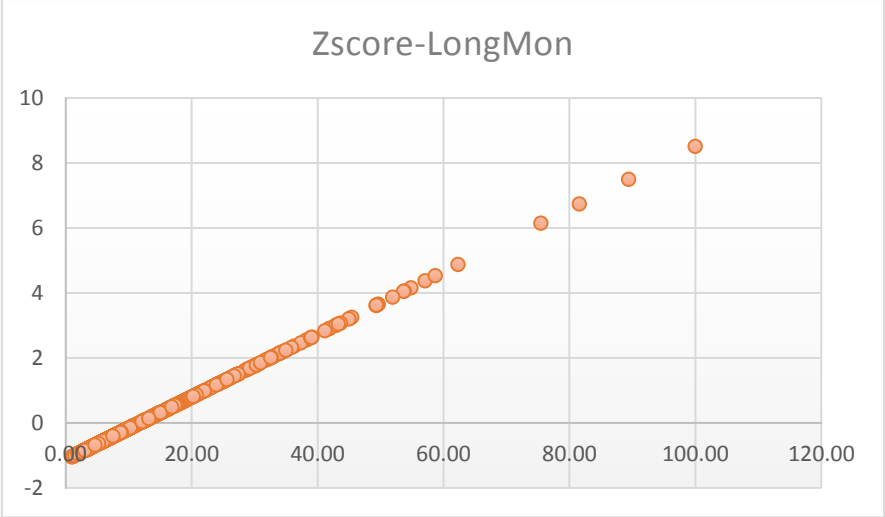


The scatter plots show that there are outliers. These outliers effect the final result of the analysis and need to be taken into consideration.


Z-Score-Tollten
7
6
5
4
3
2
1
0
-1
0.00
1,000.00
2,000.00
3,000.00
4,000.00
5,000.00
6,000.00
7,000.00


Zscore-TollMon
10
8
6
4
2
0
-2
0.00
50.00
100.00
150.00
200.00


Z-Score-Longten
10
8
6
4
2
0
-2
0.00
1,000.00
2,000.00
3,000.00
4,000.00
5,000.00
6,000.00
7,000.00
8,000.00



All the scatter plots and the box plots show that the data is right skewed and has outliers which make it difficult to represent the population.

The outliers need to be taken into consideration when analysing the data. The sample without outliers will represent the population better as compared to the sample which consists of outliers as outliers lead to extreme values.

SECTION 2:

In the previous section, we analysed the data with the assumption that 0 stands for people who do not use the service (do not have rental equipment/ don't make long distance calls/ don't have toll free services).

In this section we treat 0 as a missing value. To draw conclusions which give correct understanding of the data, we do not consider the values which have 0 as their field entry. We use the rest of the value to draw the conclusion.

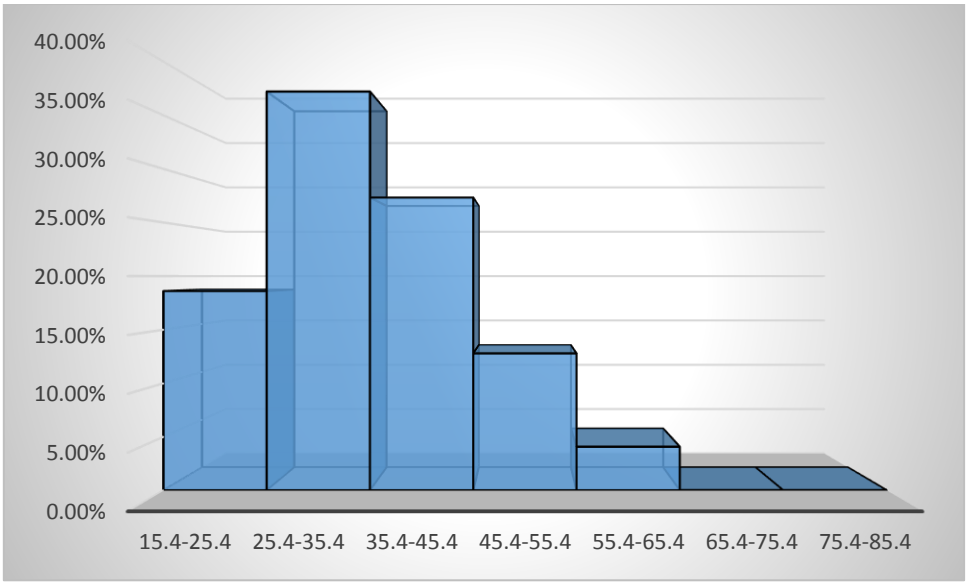
The conclusions drawn after eliminating the missing values are more accurate and can give us a better understanding of the situation as compared to the conclusion drawn using the missing as well as valid data. Based on the conclusions drawn after eliminating the missing values, further steps taken to reduce churning would be more effective.

The proactive steps taken on the basis of conclusions drawn from missing and valid data may or may not be correct.

Table 5 shows that out of the people who used rental equipment, people belonging to the 25.4-35.4 equipment over last month churned the maximum, 37.21% followed by the people belonging to the group 35.4-45.4 group who churned 27.23% which differs from the analysis made earlier which showed that 37.23% churn was generated by the people who had rental equipment over last month of group 0-10 followed by 21.90% and 21.53% belonging to group 20-30 and 30-40 respectively.

Table 5.

Equipment over last month	%churn	Sum of churn
15.4-25.4	18.60%	32
25.4-35.4	37.21%	64
35.4-45.4	27.33%	47
45.4-55.4	12.79%	22
55.4-65.4	4.07%	7
65.4-75.4	0.00%	0
75.4-85.4	0.00%	0
Grand Total	**100.00%**	**172**



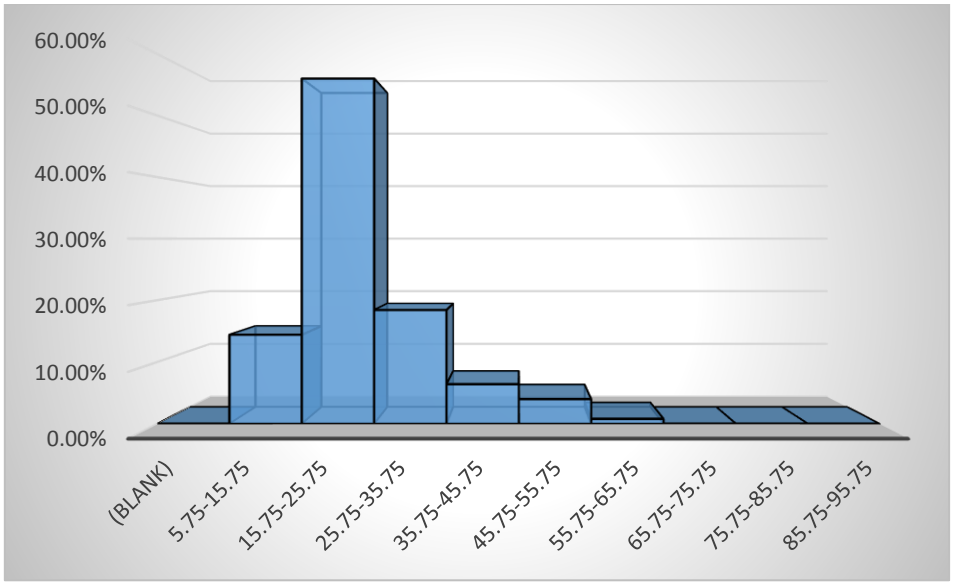
Histogram 5
(Equipment over last month vs Percentage churn)

The histogram shows a slight right skewness or positive skewness which signifies that the mass of the distribution is slightly concentrated towards the left.

Table 6 shows the percentage churned by the people who used toll free services over the last month. The analysis shows that 56.00% was churned by people belonging to the group 15.75-25.75 followed by 18.40% churned by 25.75-35.75.
This analysis is different than the previous once made using all the data- valid and missing which showed that the people belonging to the group 0-10 churned the most, 55.47% followed by the people belonging to the group 20-30% who churned 20.44%.

Table 6.

Toll over the last month	%churn	Sum of churn
5.75-15.75	14.40%	18
15.75-25.75	56.00%	70
25.75-35.75	18.40%	23
35.75-45.75	6.40%	8
45.75-55.75	4.00%	5
55.75-65.75	0.80%	1
65.75-75.75	0.00%	0
75.75-85.75	0.00%	0
85.75-95.75	0.00%	0
Grand Total	**100.00%**	**125**



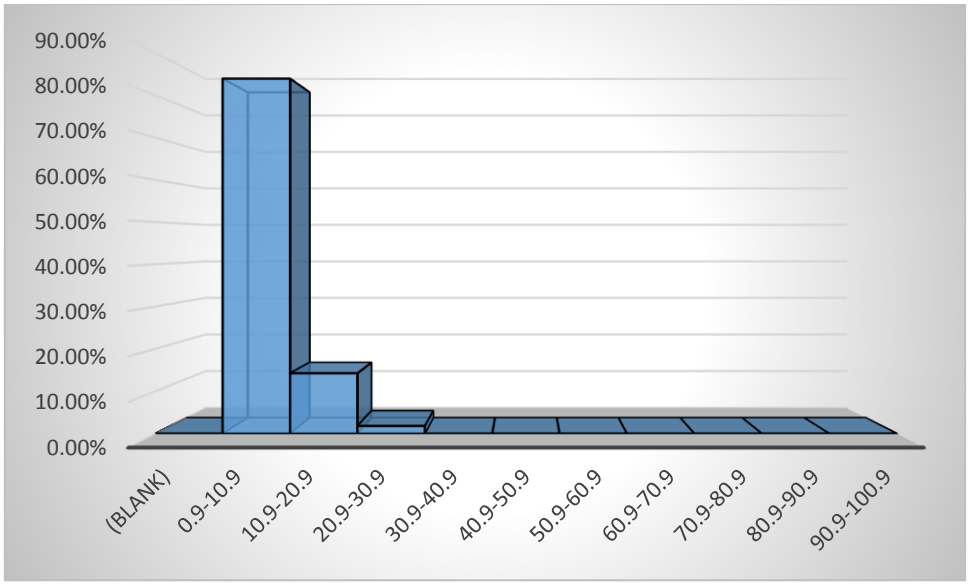
Histogram 6
(Toll free services over last month vs percentage churn)

The histogram shows a slight right skewness or positive skewness which signifies that the mass of the distribution is slightly concentrated towards the left.

Table 7 shows the analysis of data made without considering the missing values. The table shows that 83.94% of people who churned belonged to 0.9-10.9 group. The analysis made after eliminating the missing data is the same as before that showed 83.94% of the people who churned belonged to the 0.9-10.9 group followed by 14.23% belonging to the 10.9-20.9% group. We can conclude that there were no missing data and long distance calls is a goof factor which influences the churn.

Table 7.

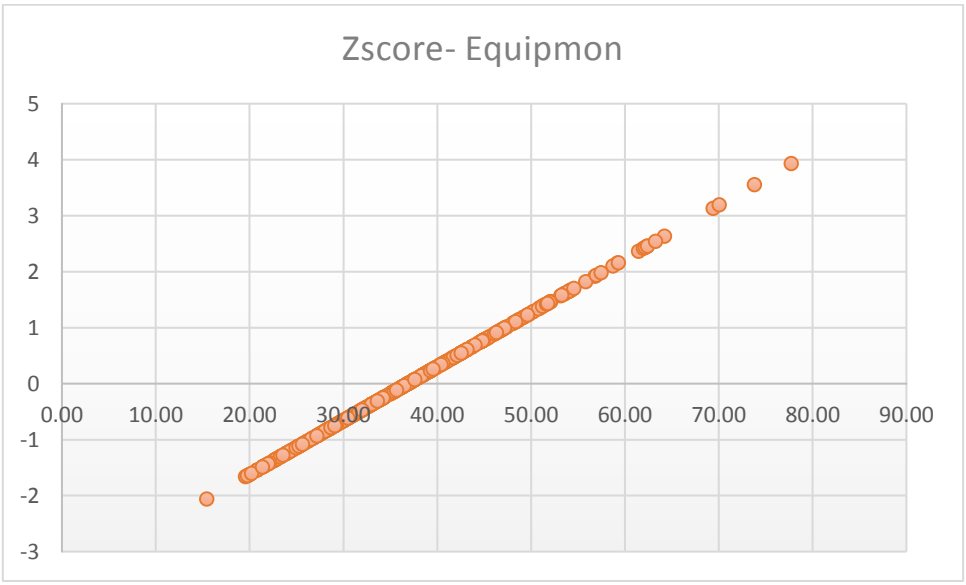
Long Distance call over last month	Sum of churn	%churn
0.9-10.9	230	83.94%
10.9-20.9	39	14.23%
20.9-30.9	5	1.82%
30.9-40.9	0	0.00%
40.9-50.9	0	0.00%
50.9-60.9	0	0.00%
60.9-70.9	0	0.00%
70.9-80.9	0	0.00%
80.9-90.9	0	0.00%
90.9-100.9	0	0.00%
Grand Total	**274**	**100.00%**

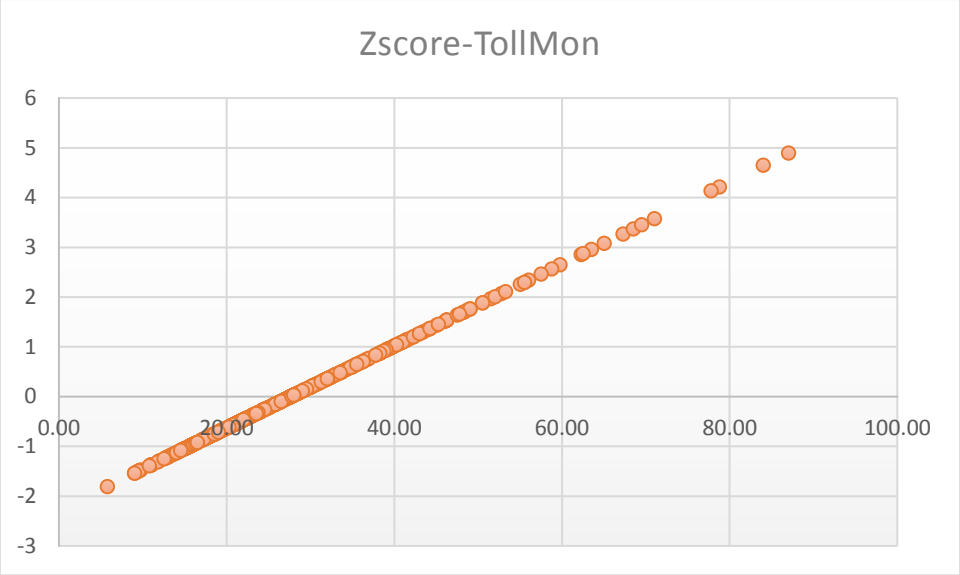


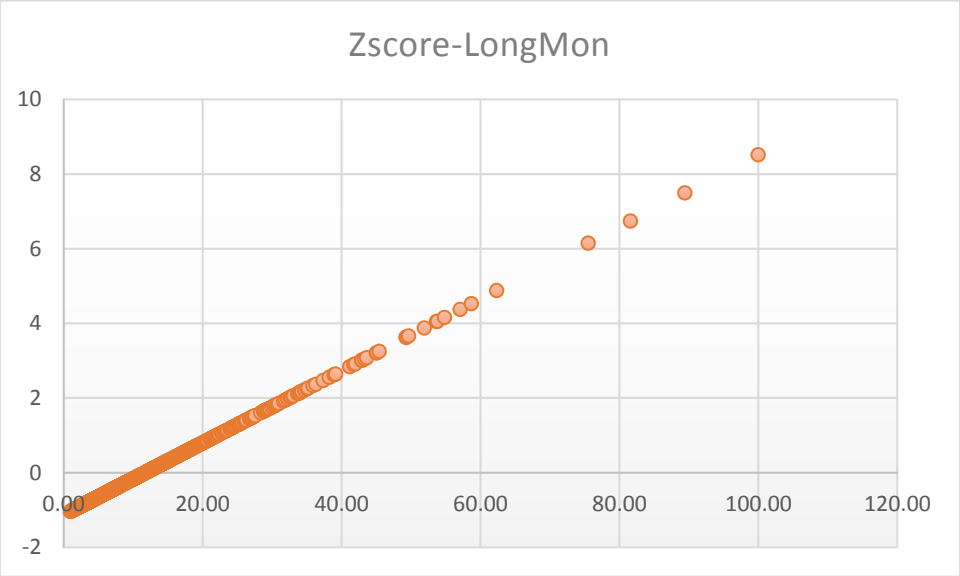
Histogram 7.
(Long distance calls over last month vs percentage churn)

We know that 99.7% of the data lies within 3 standard deviations of the mean. The scatter plot on the data against its Z-score shows that there are data points which lie outside 3 standard deviations as well. But the points lying outside 3 standard deviations are lesser in these scatter plots as compared to the scatter plots in the previous section which shows that there are lesser number of outliers.

Since the no of outliers are lesser, the chances of data representing the actual reason is higher.




Zscore-TollMon
6
5
4
3
2
1
0
-1
-2
-3
0.00
20.00
40.00
60.00
80.00
100.00

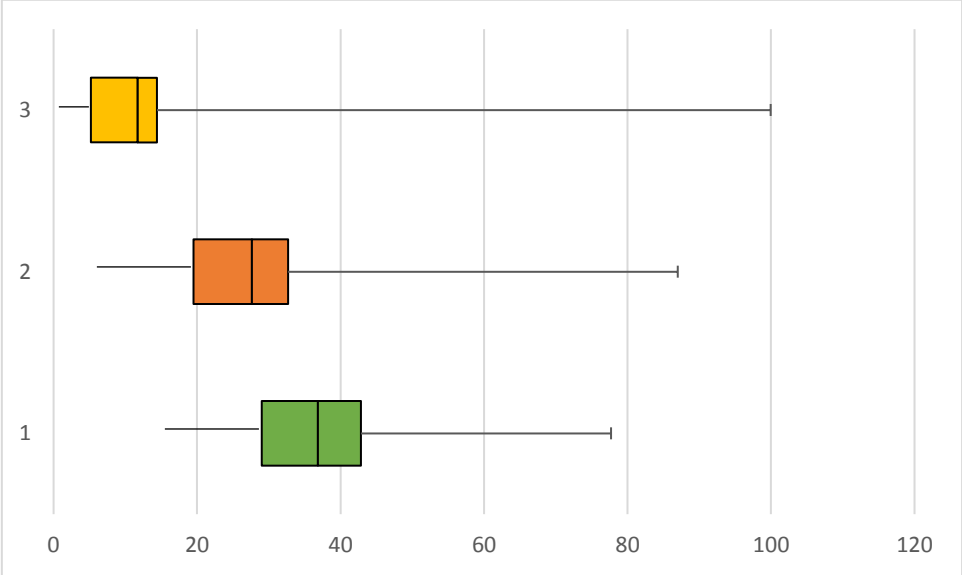

Zscore-LongMon
10
8
6
4
2
0
-2
0.00
20.00
40.00
60.00
80.00
100.00
120.00

The box plot of the data analysed after eliminating the missing values shows slight skewness which means that the data is a better representation of the population and that using this data will give us better results as compared to results computed on the basis of missing and valid values.

Table 8.

	Equipmon	Tollmon	Longmon
Min	15.40	5.75	0.90
Q1	29.0125	19.5	5.2
Mean	36.84	27.64	11.72
Q3	42.8375	32.6875	14.4125
Max	77.70	87.00	99.95

Figure 2.



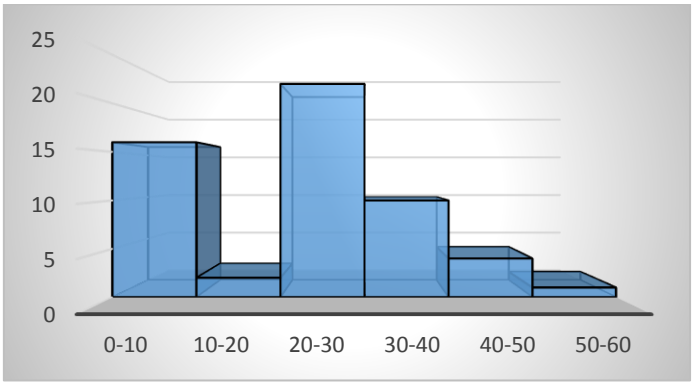
SECTION 3:

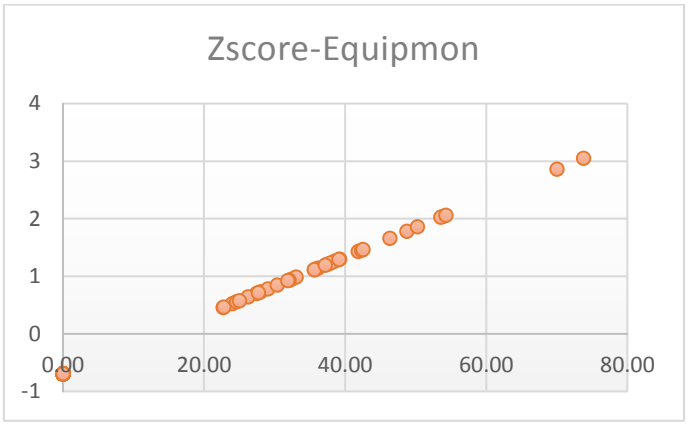
SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT:

When 10 samples are chosen using simple random sampling without replacement of 100 data values, each sample is giving similar graph and statistics.

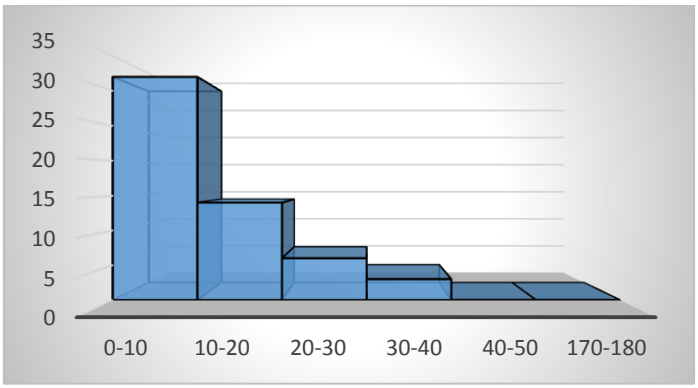
Thus, we can conclude based on the graphs, Zscore and boxplot that the samples created using simple random sampling without replacement represents the population well.

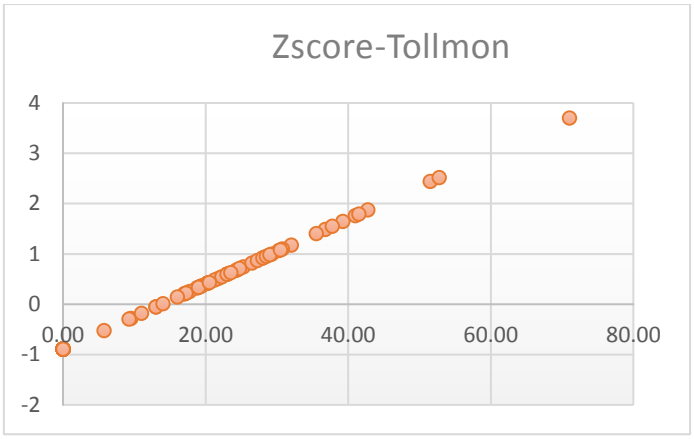
1) **Equimon vs Churn:**



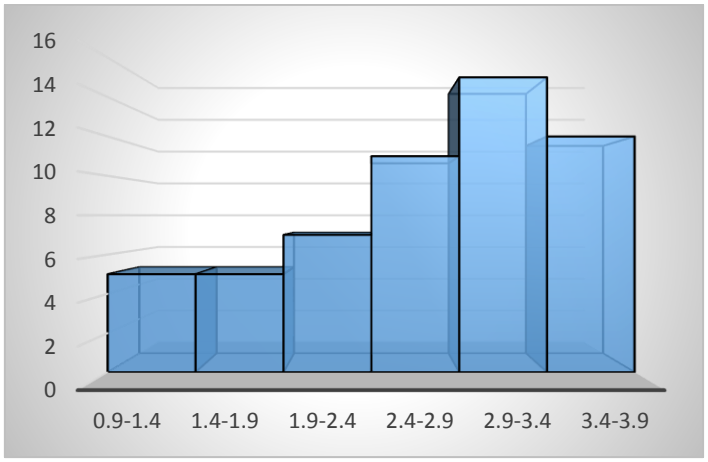


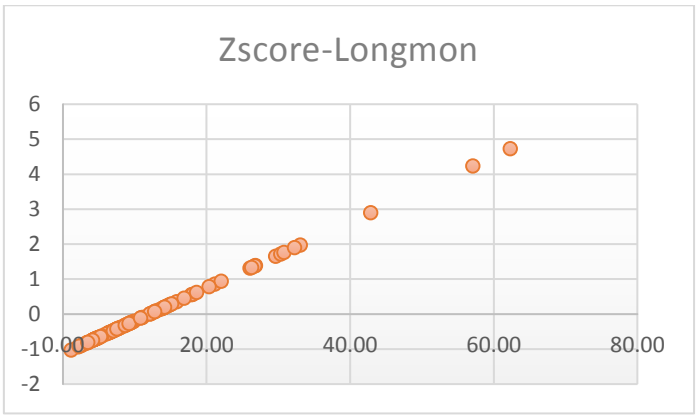
2) **Tollmon vs Churn**





3) Longmon vs Churn

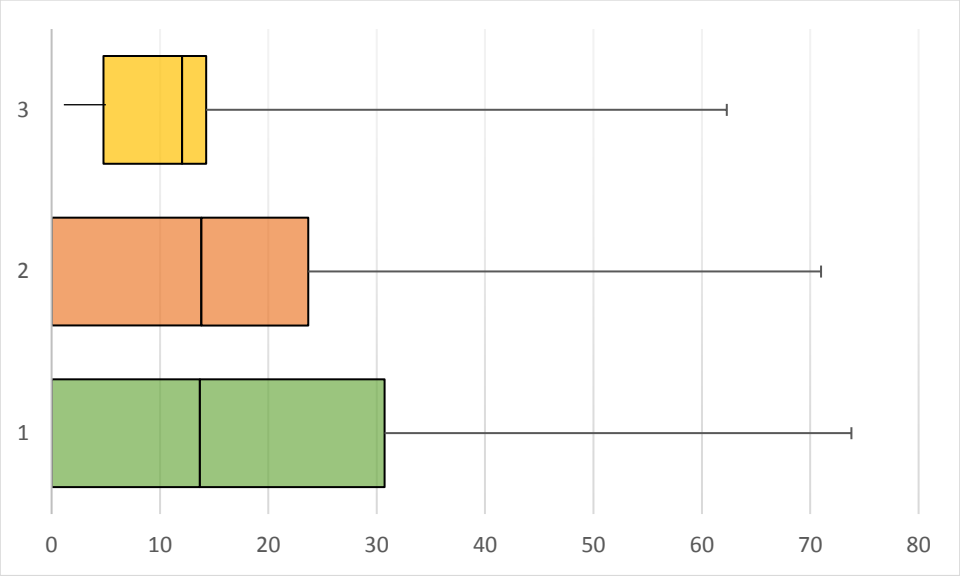




Supporting Data:

The boxplot obtained is similar to the boxplot we obtained in section 1. Thus, we can say that the samples obtained represent the population correctly.

	Equipmon	Tollmon	Longmon
Min	0.00	0.00	1.10
Q1	0	0	4.7875
Mean	13.68	13.81	12.04
Q3	30.725	23.6875	14.2625
Max	73.80	71.00	62.30



SIMPLE RANDOM SAMPLING WITH REPLACEMENT:

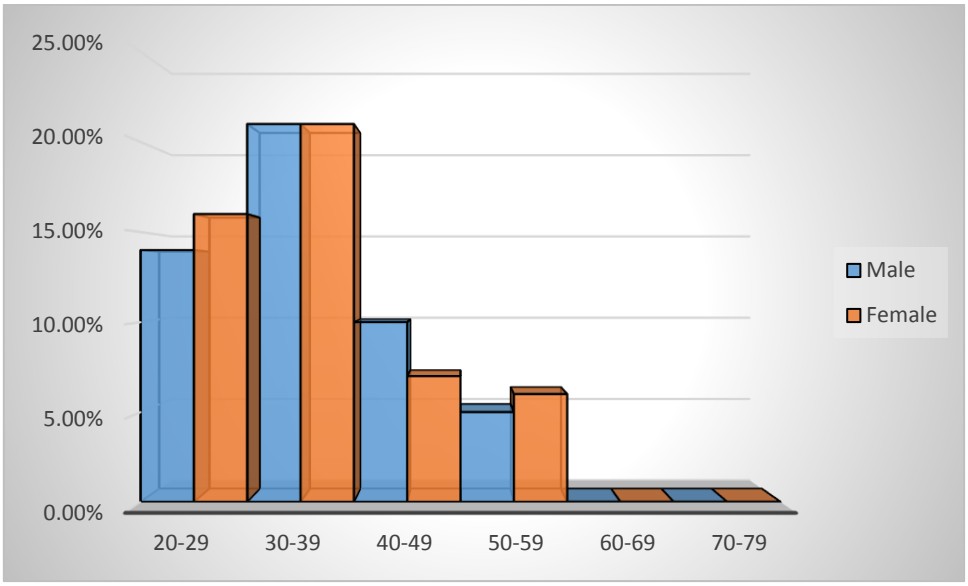
When 10 samples are chosen using simple random sampling with replacement of 100 data values, each sample is giving different graph and statistics.
Thus, we can conclude based on the graphs, Zscore and boxplot that the samples created using simple random sampling with replacement does not represents the population well. The chances of outliers getting picked increases which in turn effects the value of average, Standard deviation and Zscore. It also effect the histograms and boxplots. The boxplots obtained where dissimilar- some were positively skewed while some were negatively skewed. Thus, we can say that simple random sampling with replacement does not represent the population as well as simple random sampling without replacement.

SECTION 4:

In this section, we collect a sample of 100 churners and 100 non churners and analyse the data. The analysis shows that people belonging to the age group 30-39 churn the most, 42% which can be divided into male and female churners, 21 % each, which is also the individual gender highest.

Table 9.

% Churn	Gender		
Age	Male	Female	Grand Total
20-29	14.00%	16.00%	30.00%
30-39	21.00%	21.00%	42.00%
40-49	10.00%	7.00%	17.00%
50-59	5.00%	6.00%	11.00%
60-69	0.00%	0.00%	0.00%
70-79	0.00%	0.00%	0.00%
Grand Total	**50.00%**	**50.00%**	**100.00%**



Histogram 9.

Age and Gender vs Percentage Churn

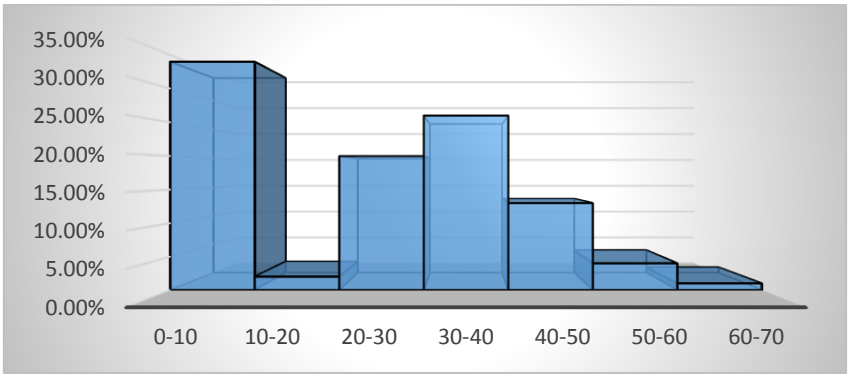
Analysing the data also shows us that the percentage churn or equipmon, tollmon and longmon gives us similar output/statistics as given by the entire population in section 1.

Thus, we can conclude that a stratified sample represents population better than simple random sampling with replacement.

Supporting Data:

Table 10

Equipmon	%churn
0-10	34.00%
10-20	2.00%
20-30	20.00%
30-40	26.00%
40-50	13.00%
50-60	4.00%
60-70	1.00%
Grand Total	**100.00%**



Histogram 10.

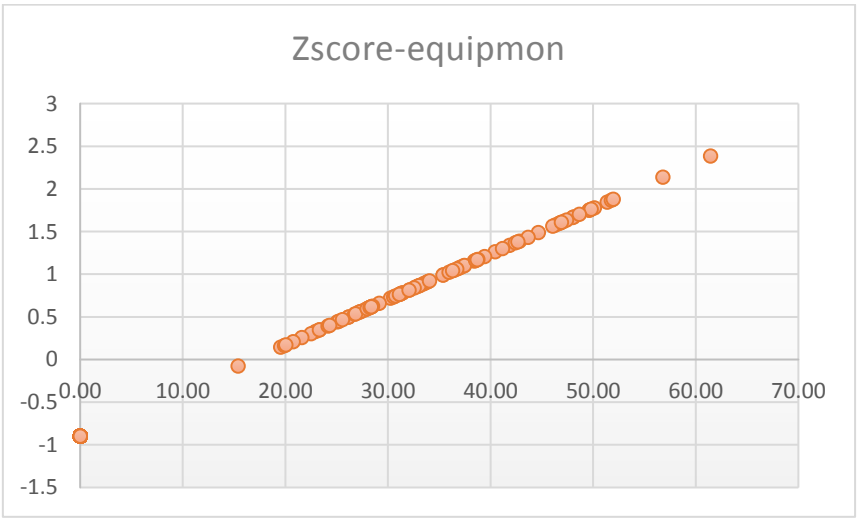
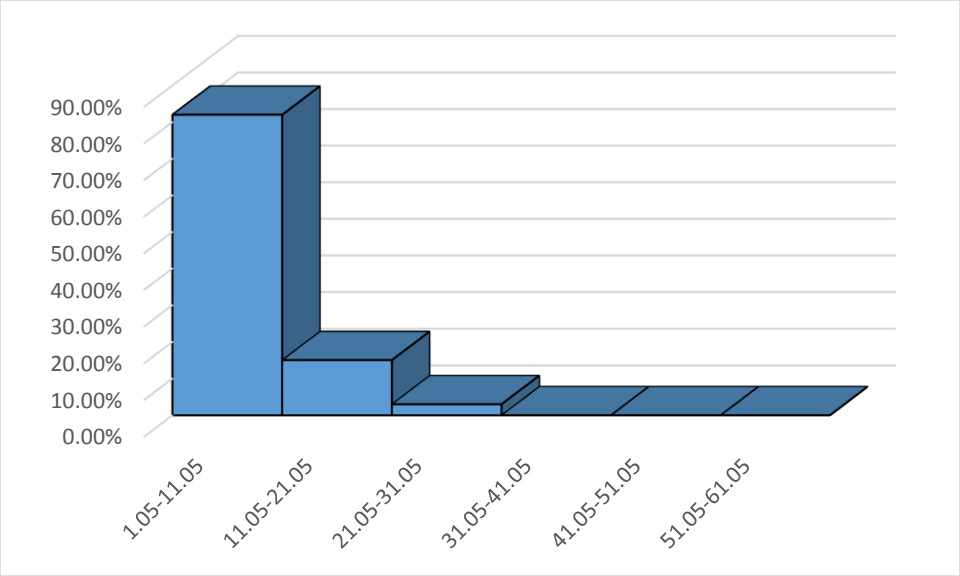


Table 11.

Longmon	% Churn
1.05-11.05	82.00%
11.05-21.05	15.00%
21.05-31.05	3.00%
31.05-41.05	0.00%
41.05-51.05	0.00%
51.05-61.05	0.00%
Grand Total	**100.00%**



Histogram 11.

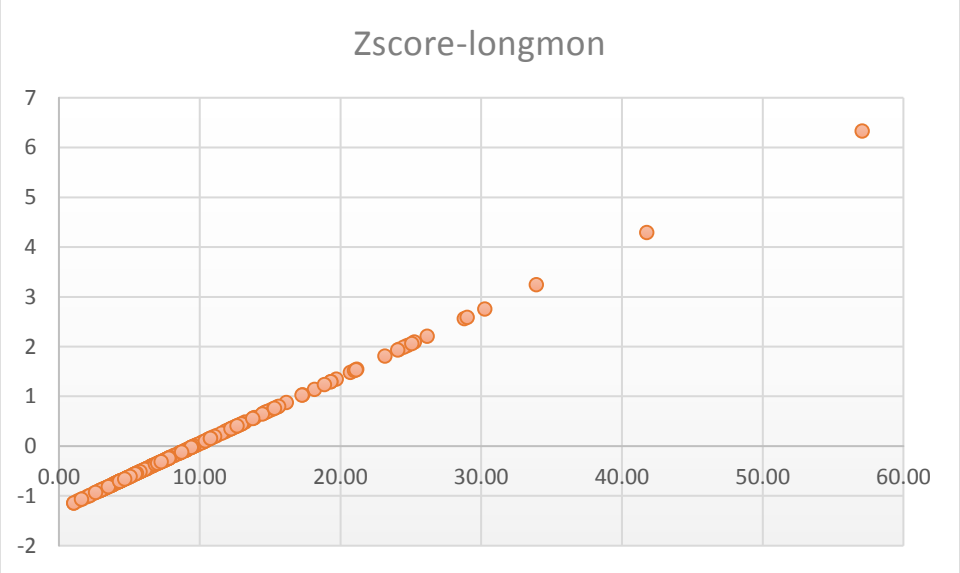
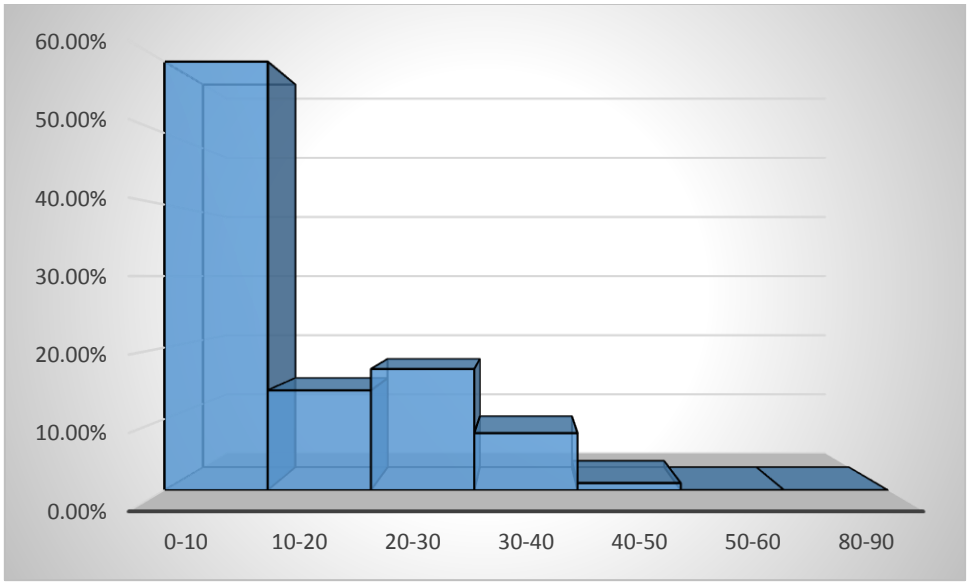
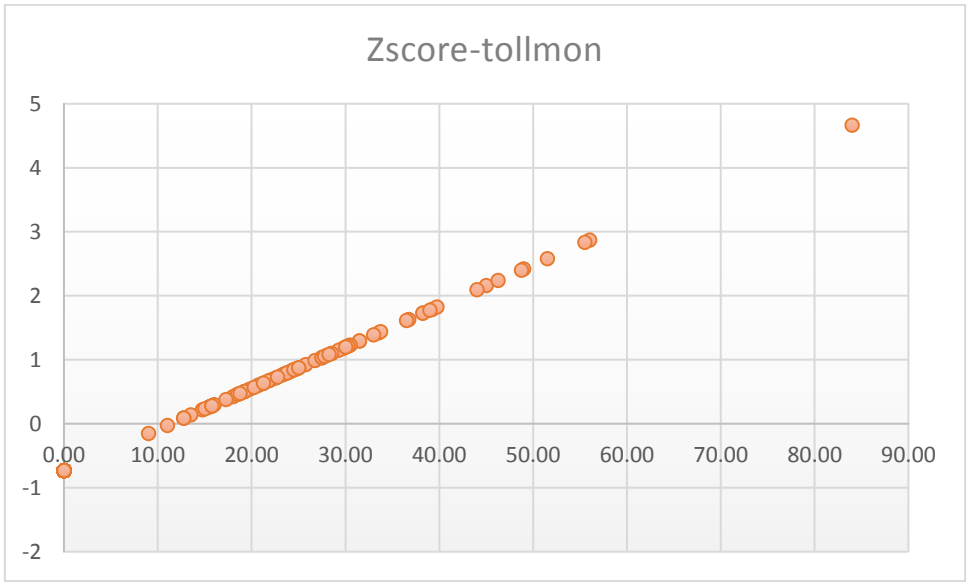


Table 12.

Tollmon	% Churn
0-10	60.00%
10-20	14.00%
20-30	17.00%
30-40	8.00%
40-50	1.00%
50-60	0.00%
80-90	0.00%
Grand Total	**100.00%**



Histogram 12.



Thus, we can conclude that SRSWOR and stratified sampling represents the population well while SRSWR does not represent the population.