Assignment #1

[redacted]The assessment is based on five sections; assignments, quizzes, labs, midterms, and final exam. There are two midterms and one final exam. Final exam has 30% of total grade. Midterms is 30% which means each of them is equal to 15%. Assignments have 20% of the total grades which are worth 5% each. The contribution of the labs are 10% of total values for 16 lab works. Quizzes will be taken in class by iClicker device which are totally equal to the 10% of the grade.

[redacted]The data loaded from DDT.csv:

Q#2) (a) The following figure is the output of R:

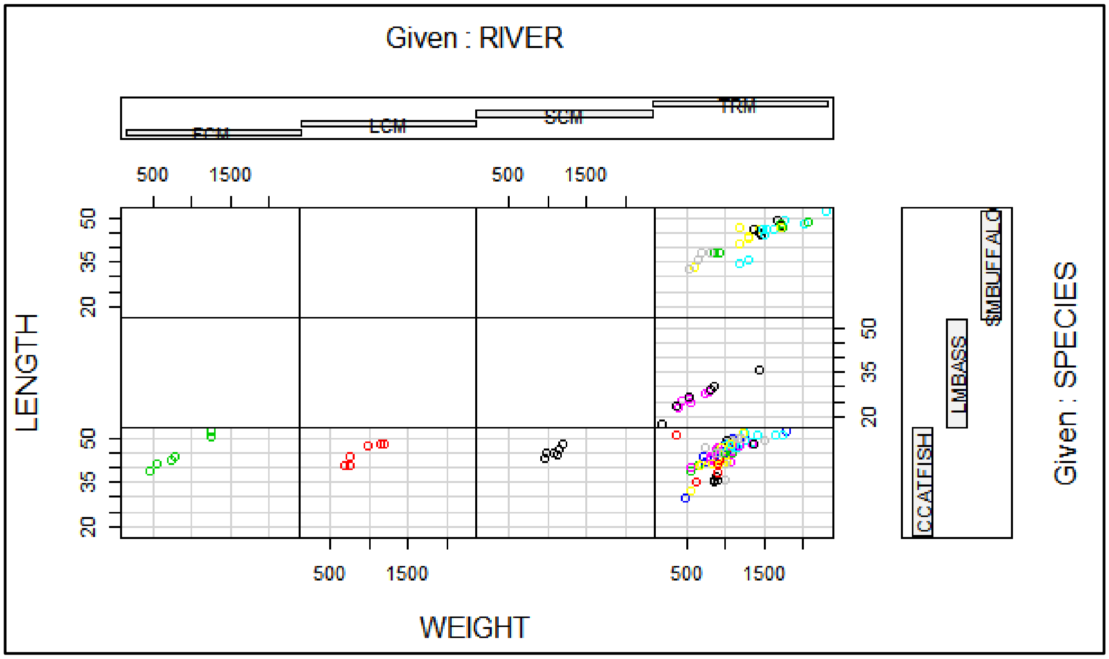


Figure 1. Coplot for length vs. weight

Q#2) (b) The lower three left plots show the length vs. weight plots for the "Catfish" species where caught on 1) FCM, 2) LCM, and 3) SCM rivers.

Q#2) (c) It sorts the emerged values for "MILE" variable. It levels the discrete number that appeared in this variable. We will have a vector of the values in order.

Q#2) (d) Based on the previous section which we sorted the factors of variable miles, in this section, it makes a vector for variables which shows the order of each element.

Q#2) (e) The species of "LMBASS" and "SMBUFFALO" do not exist in "FCM", "LCM", and "SCM" rivers.

Q#2) (f) For this part the hint code shows us a list of the catfishes from FCM River. The following code could be added to calculate the mean value. The result is 45.

```
Q1f<-ddt[ddt$RIVER=="FCM" & ddt$SPECIES=="CCATFISH",]
Q1fddt<-Q1f["DDT"]
Q1fmean = mean(unlist(Q1fddt))
Q1fmean
```

[redacted]MS 1.14

Q#3) (a) Quantitative
Q#3) (b) Quantitative
Q#3) (c) Qualitative
Q#3) (d) Quantitative
Q#3) (e) Qualitative
Q#3) (f) Quantitative
Q#3) (g) Qualitative

[redacted]MS, Pages 12, 13

Q#4) (a) Simple random sampling, stratified random sampling, cluster sampling, and systematic sampling.

Q#4) (b) In simple random sampling, a number assign to each element. The sample is chosen based on random numbers. In stratified sampling, the population divides into several groups with common attributes called as strata. Samples are chosen based on random sampling from strata. In cluster sampling, samples are chosen from natural group called as clusters. For example, some computer codes will be checked out of different codes available entirely instead of choosing code lines from whole codes available. The last one is systematic sampling which used frequently in production lines; chose every kth element in line to test.

[redacted]MS 1.15: The wells are: 167, 115, 189, 84, and 141.

Q#5) (a) Additional questions:

(i) mtbeo = na.omit(mtbe)
(ii) 56.45357

[redacted]MS 1.16: The code is available in R file.

Q#6) (a) Answers:

(i) Plot for earthquake magnitude:

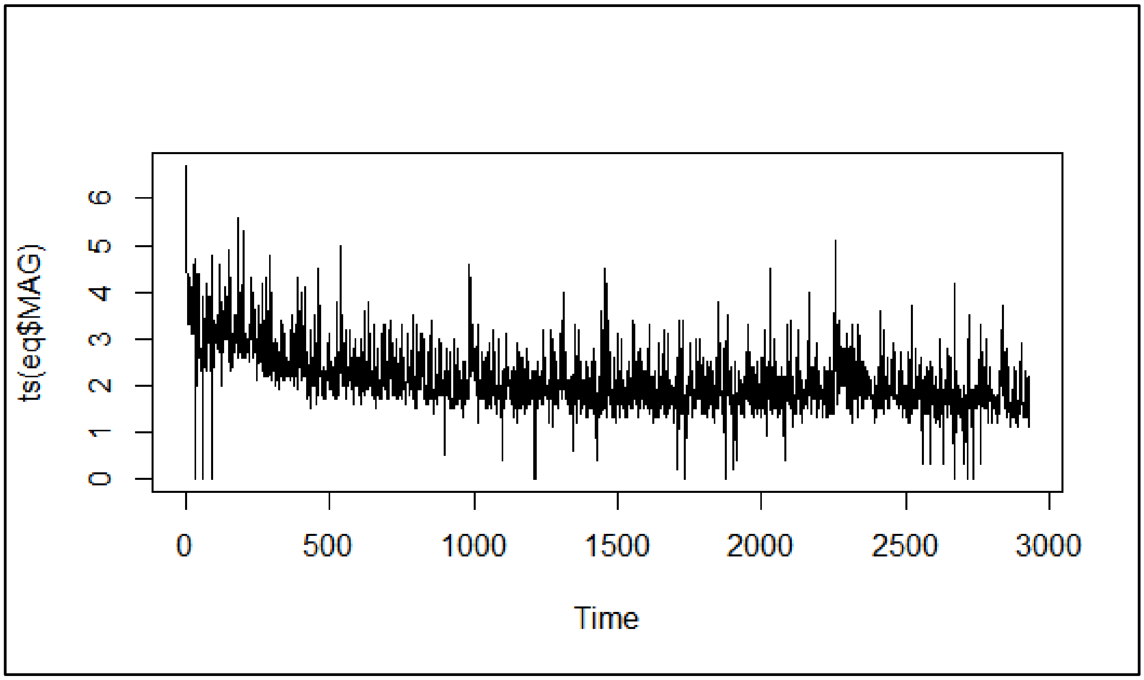


Figure 2. Earthquake magnitude plot in Los Angles area

(ii) Code is available in R file. The answer was 2.

[redacted]MS, page 18:

Q#7) (a) The data collection method is actually a designed experiment, one involving a stratified sample.

Q#7) (b) All fish in the Tennessee River and its tributaries.

Q#7) (c) The location of capture and species are qualitative variables.

[redacted]MS 2.1:

Q#8) (a) Bar graph.

Q#8) (b) Built design if they have wheels or legs. I can call it drive option.

Q#8) (c) Legs only.

Q#8) (d) The total number of robots is 106. The frequency for different types are: 15, 8, 63, and 20. The percentage would be 15/106=14.15%, 8/106=7.55%, 63/106=59.43%, and 20/106=18.87%.

Q#8) (e) The Pareto bar plot:

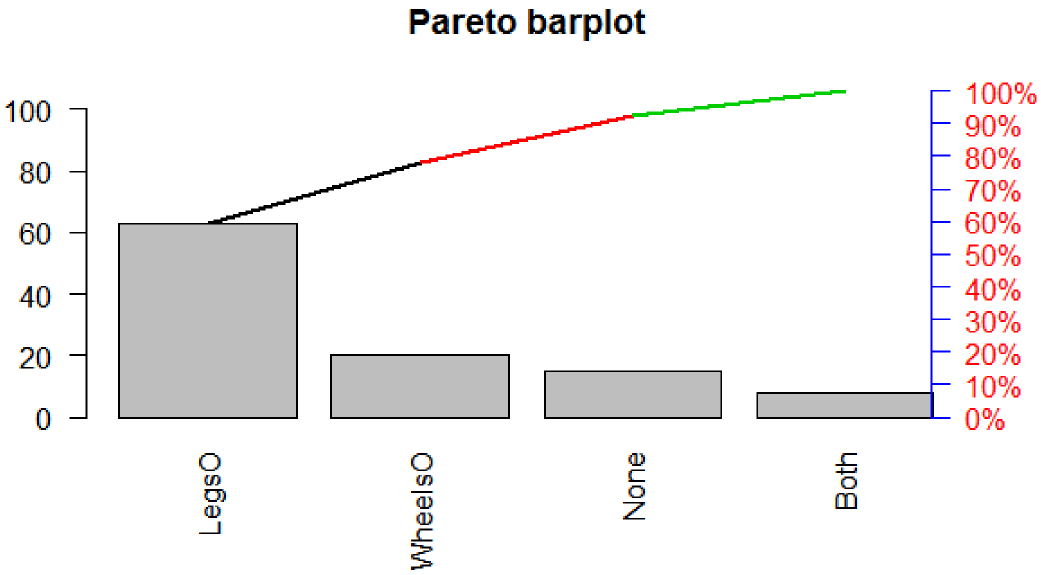


Figure 3. Pareto bar plot for question 8

[redacted]MS 2.4:

Q#9) (a) Explorer had had the lowest proportion of security issues in 2012.

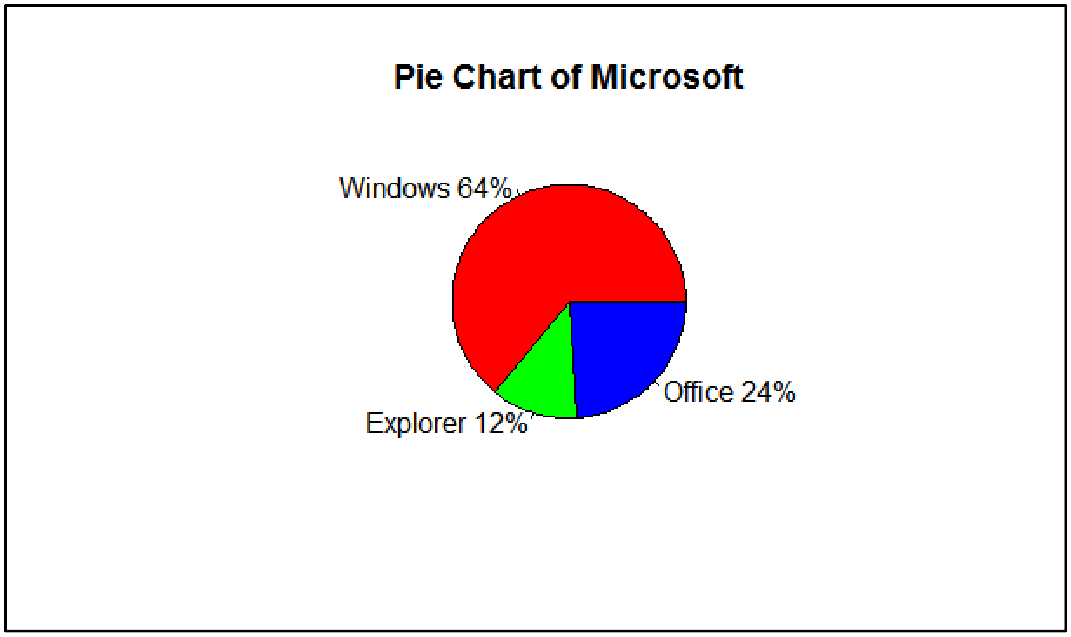


Figure 4. Microsoft security issues in 2012

Q#9) (b) Windows has a serious problem in security:

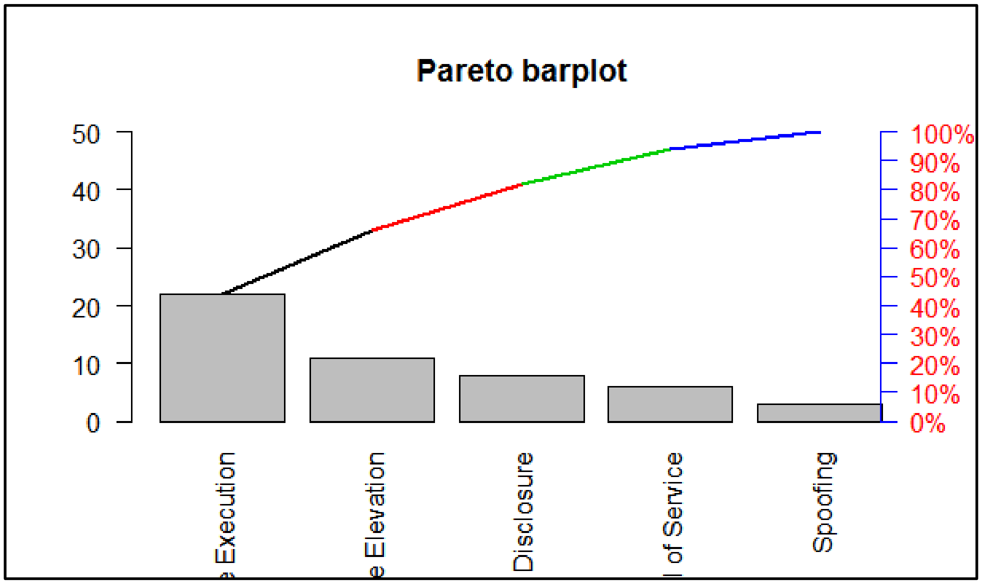


Figure 5. Pareto plot for Microsoft issues in 2012.

[redacted]MS 2.10: the following chart shows that the chance of defection is about 10%. In the other words, 90% of the modules are OK.

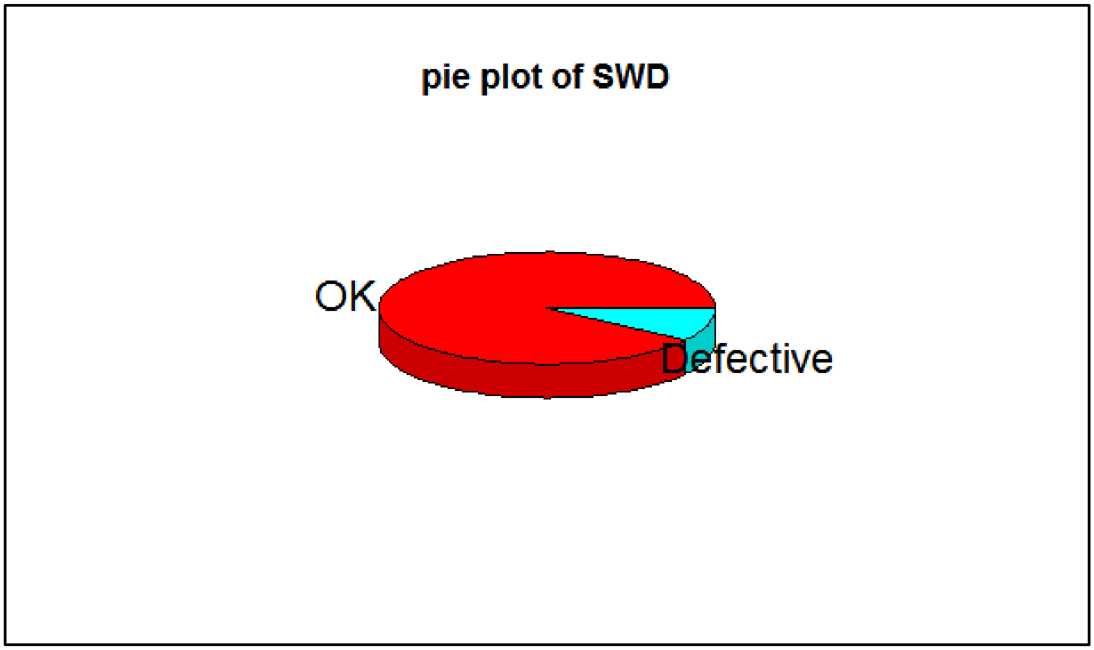


Figure 6. The 3D pie chart showing defection of modules of software.

[redacted]M 2.72:

Q#11) (a) The filled table for old and new locations:

Table 1. The histogram data for old location

Class	Class Interval	Data Tabulation	Frequency	Relative Frequency
1	8.0000 - 8.2889	|	1	0.0333
2	8.2889 - 8.5778		0	0.0000
3	8.5778 - 8.8667	|||	3	0.1000
4	8.8667 - 9.1556		0	0.0000
5	9.1556 - 9.4444		0	0.0000
6	9.4444 - 9.7333	|||	3	0.1000
7	9.7333 - 10.0222	||||| ||||| |||	13	0.4333
8	10.0222 - 10.3111	||||| ||||	9	0.3000
9	10.3111 - 10.6000	|	1	0.0333

Table 2. The histogram data for new location

Class	Class Interval	Data Tabulation	Frequency	Relative Frequency
1	8.0000 - 8.2889		0	0.0000
2	8.2889 - 8.5778	|	1	0.0333
3	8.5778 - 8.8667	||||| |	6	0.2000
4	8.8667 - 9.1556	|	1	0.0333
5	9.1556 - 9.4444	||||| ||	7	0.2333
6	9.4444 - 9.7333	||||| ||	7	0.2333
7	9.7333 - 10.0222	|||	3	0.1000
8	10.0222 - 10.3111	|||||	5	0.1667
9	10.3111 - 10.6000		0	0.0000

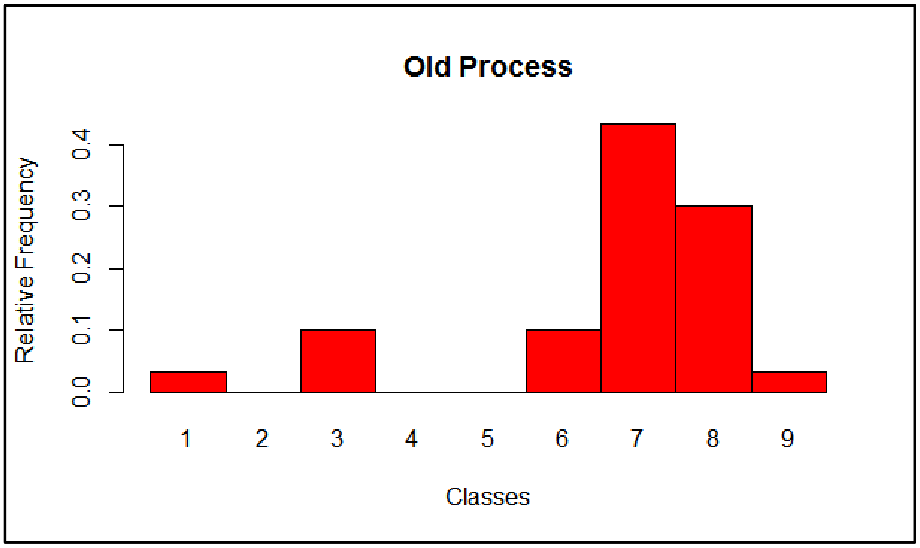


Figure 7. Relative frequency histogram for the voltage reading of old process

Q#11) (b) The decimal point is 1 digit(s) to the left of the |

```
 8 | 1
 8 | 778
 9 |
 9 | 6778888999
10 | 000000011122333
10 | 6
```

Since there are not so many elements in data frame the stem and leaf display provides more accurate data.

Q#11) (c) The bar chart for new process is:

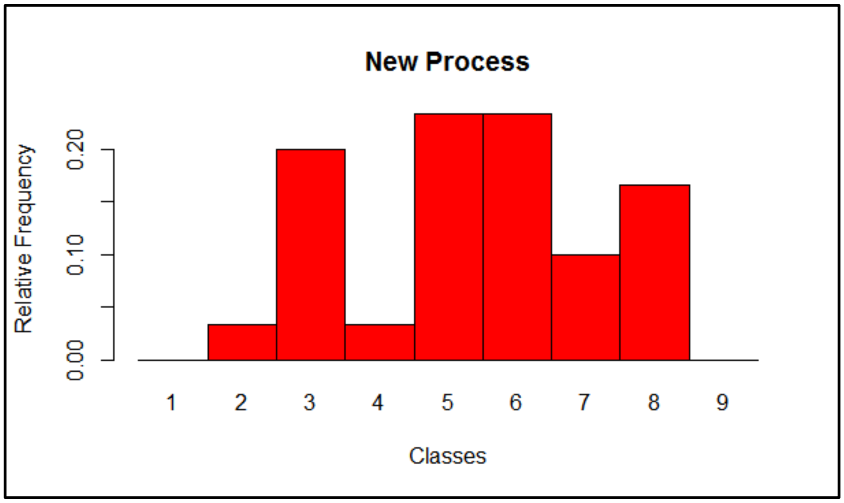


Figure 8. Relative frequency histogram for the voltage reading of new process

Q#11) (d) Histogram for both process:

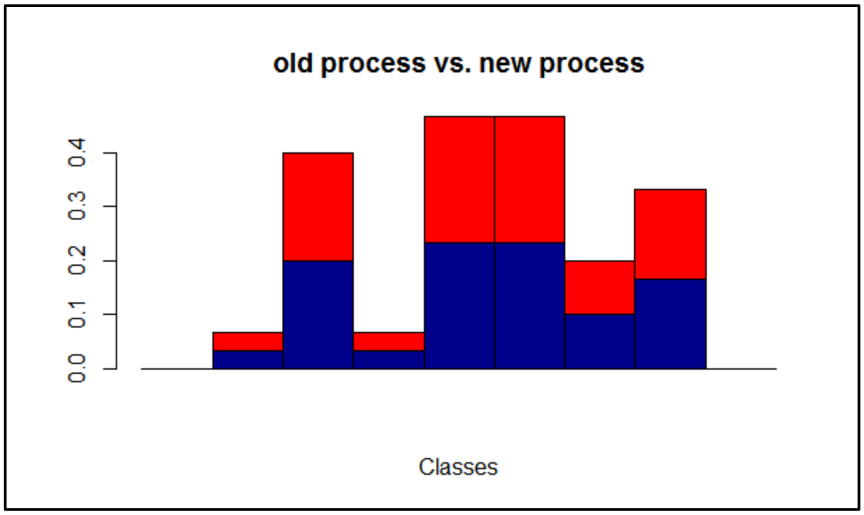


Figure 9. Relative frequency histogram for the voltage reading of both processes

Q#11) (e) Old Location:

```
> mold = mean(OVOLT$VOLTAGE)
> mold
[1] 9.803667
> meold = median(OVOLT$VOLTAGE)
> meold
[1] 9.975
> moold = mfv(OVOLT$VOLTAGE) #mfv is from modeest package
which computes mode
> moold
[1] 8.72  9.80  9.84  9.87  9.98 10.05 10.15 10.26
> sold = sd(OVOLT$VOLTAGE)
> sold
[1] 0.5409155
```

From seeing the histogram, we can say the data is more skewed. So, median is a better measure of the central tendency. If the data is more skewed, then the data is not concentrated at a same value and is distributed heavily from the mean so mean leans towards a side whereas median divides the whole data into two equal parts. So, for this case median is better measure of central tendency.

New Location:

```
> mnew = mean(NVOLT$VOLTAGE)
> mnew
[1] 9.422333
menew = median(NVOLT$VOLTAGE)
```

```
> menew
[1] 9.455
> monew = mfv(NVOLT$VOLTAGE)
> monew
[1] 8.82
> snew = sd(NVOLT$VOLTAGE)
> snew
[1] 0.4788757
```

For this case, from the histogram we can see the data is not very skewed and mean and median both divide the data closely into two equal parts. So, for actual better central tendency **mean** gives a better value. When data is not much skewed mean is better choice of central value.

Q#11) (f) find z-score for a voltage reading of 10.5 at old location

```
> zold = (10.5 - mold)/sold
> zold
[1] 1.287324
```

Q#11) (g) find z-score for a voltage reading of 10.5 at new location

```
> znew= (10.5 - mnew)/snew
> znew
[1] 2.25041
```

Q#11) (h) Z- score of 10.5 is more likely to occur at old location as it is just 1.2878 standard deviations away from the mean, whereas it is 2.25 standard deviations away from the new location. The closer the value to the mean the more likely it is to occur.

Q#11) (i) (i). `> boxplot(OVOLT$VOLTAGE)`

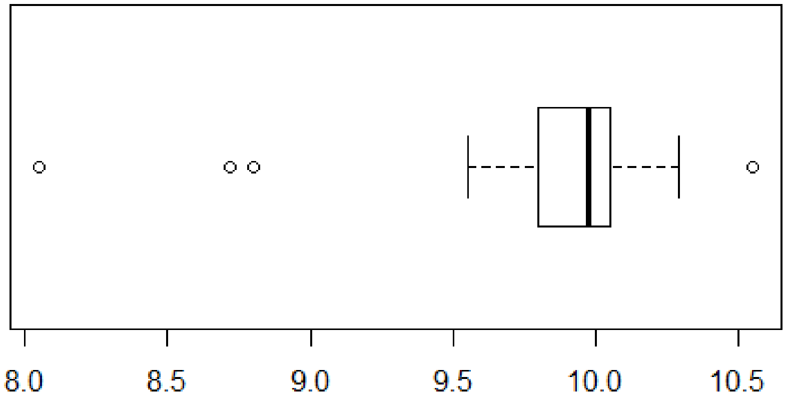


Figure 10. Box plot for old process

Yes, I detect some outliers in this box plot any value that is less than lower quartile and higher than Upper quartile are the outliers, which can be clearly seen in the above box plot.

Q#11) (j) By Method of z-scores, if any data lies outside 3 z-score or 3 standard deviations from mean it is an outlier.

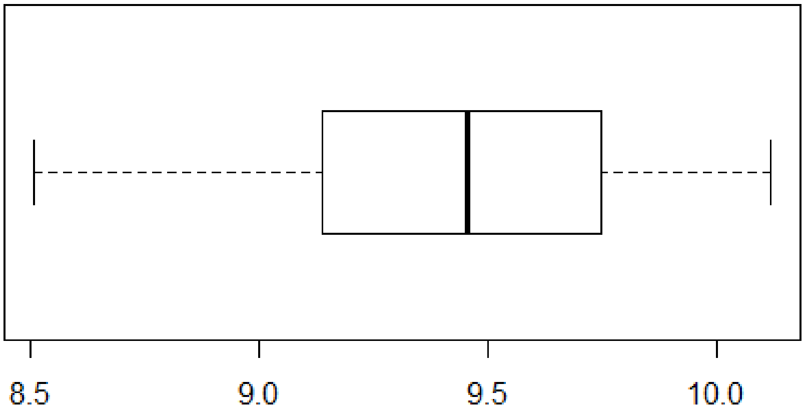
```
> OVOLT$VOLTAGE

 [1]  9.98 10.26 10.05 10.29 10.03  8.05 10.55 10.26  9.97  9.87 10.12 10.05
9.80 10.15 10.00

[16]  9.87  9.55  9.95  9.70  8.72  9.84 10.15 10.02  9.80  9.73 10.01  9.98
8.72  8.80  9.84
> mold - 3*sold
[1] 8.18092
> mold + 3*sold
[1] 11.42641
```

We can see that some data in old location voltages are less than the range of 8.18 and 11.42. so, the data below 8.18 above are outliers in old location voltages.

Q#11) (k)



No, I cannot detect any outliers in the new location box plot.

Q#11) (l)

```
> NVOLT$VOLTAGE
 [1]  9.19  9.63 10.10  9.70 10.09  9.60 10.05 10.12  9.49  9.37 10.01  8.82  9.43 10.03  9.85
[16]  9.27  8.83  9.39  9.48  9.64  8.82  8.65  8.51  9.14  9.75  8.78  9.35  9.54  9.36  8.68
```

```
> mnew - 3*snew
[1] 7.985706
> mnew + 3*snew
[1] 10.85896
```

We can see that no data lies outside the range of 7.985 and 10.858. So, there are no outliers in the new location data.

Q#11) (m) By seeing the above box plots we can say that at old location we have better voltages which are more than 9.2 volts as compared to new location. So, the new process is not as good as old process.

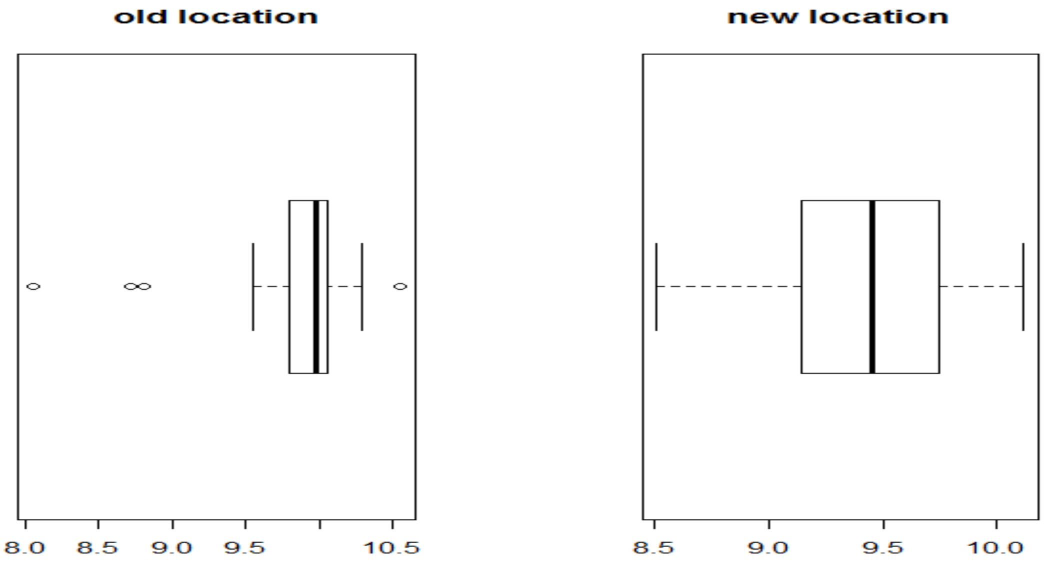


Figure 11. box plot for both

[redacted]MS 2.73: The interval for 95% confidence level is 0.8332 and 2.9288.

[redacted]MS 2.80:

Q#13) (a) Mean value is 12.82 which is arithmetic average for whole data set. The median is 5. If the data was sort, the value which is located in the middle is median. The value of mode which has a high occurrence in the data set is 4.

Q#13) (b) Two big numbers exist in the data set which distract the mean value and may make a huge misleading. Mode is not a center value. It appears just because of the frequency which might be misleading. I prefer median in this case. It is not totally suitable too but makes more sense at this case.

Q#13) (c) Plant coverage percentage for five dry steppe sites, mean = 40.4, median = 40, and mode = 40.

Q#13) (d) Plant coverage percentage for Gobi desert, mean = 28, median = 26, and mode = 30;

Q#13) (e) The plant coverage for these two areas is different. It is obvious; one of them id steppe and another is desert which have different plant coverages.

[redacted]MS 2.84:

Q#14) (a) Histogram:

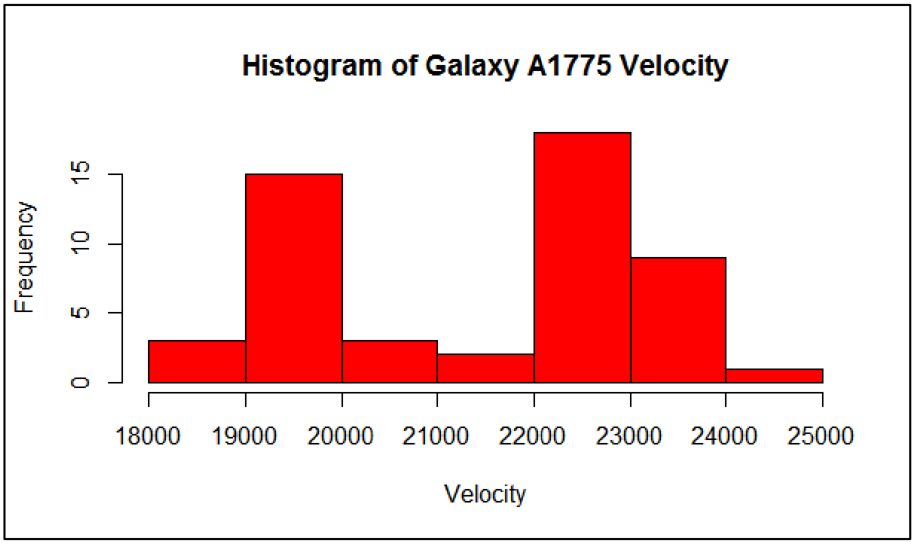


Figure 12. The velocity distribution for galaxy A1775

Q#14) (b) Two picks on distribution show the possibility of dual galaxy. They have different velocities.

Q#14) (c) According to Figure 10, I divided the velocities into two categories for galaxy A1775A and A1775B. The value of 21000 km/s is considered as interface value. Therefore; the mean value for galaxy A is 19,462 km/s with standard deviation of 532, and the mean value for galaxy B is 22,838 km/s with standard deviation of 561.

Q#14) (d) If velocity is 20,000 km/s, the velocity is more likely belong to galaxy A1775A.