Eric Altenburg - Homework 01
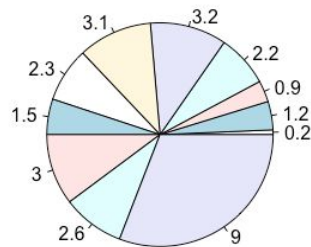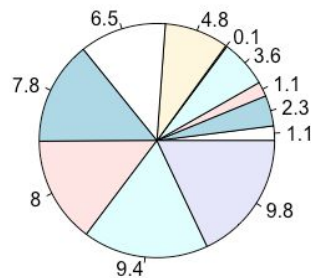9/6/19

Histogram of x

**Histogram of y**

1.

Pie Chart of X

Pie Chart of Y

Distribution of X: Although there is not much data to go off of, there does not seem to be any sort of symmetry here, instead, the histogram is skewed-right as seen by the outlier values in the 8 to 10 range occurring once. The sample is centered around the 2 to 4 range with a probable outlier to the right. The range is from 0 to 10 producing a variation of 10 in the sample.

Distribution of Y: This sample has two peaks in the histogram at 0 to 2 and 6 to 8, though it is not symmetric. The center of the distribution is around 4 to 6, with the variation being 10 once again.

Box-plot of X

Box-plot of Y

2.
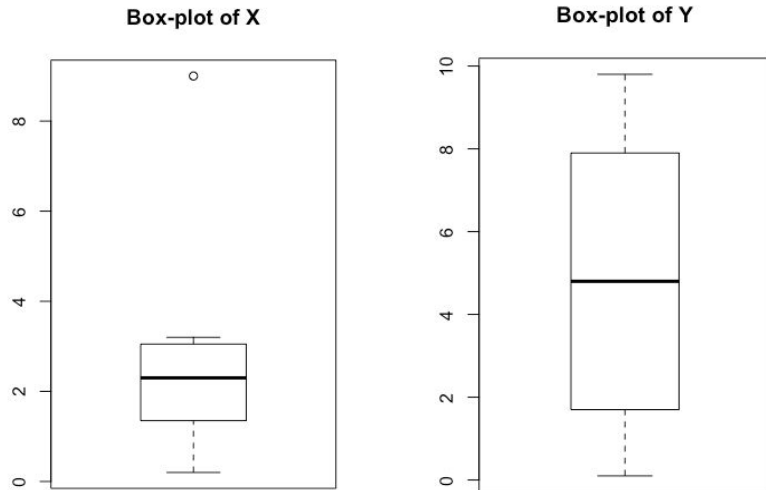
Five number summary of x:

Min.	1st Qu.	Median	3rd Qu.	Max.
0.2000	1.350	2.300	3.050	9.000

Five number summary of y:

Min.	1st Qu.	Median	3rd Qu.	Max.
0.100	1.700	4.800	7.900	9.800

Variance of x: 5.376727

Variance of y: 12.51873

The only outliers are in the sample of x which is 9. This is found by calculating the IQR of X and Y then multiplying it by 1.5 and fitting that to their individual quartiles. In X's case, the final number was 2.55, therefore, if a number was 2.55 above or below quartiles 1 and 3, then it would be an outlier (such as 9). For Y, the final number was 9.3 above and below quartiles 1 and 3, which there was none.

Scatterplot of X and Y
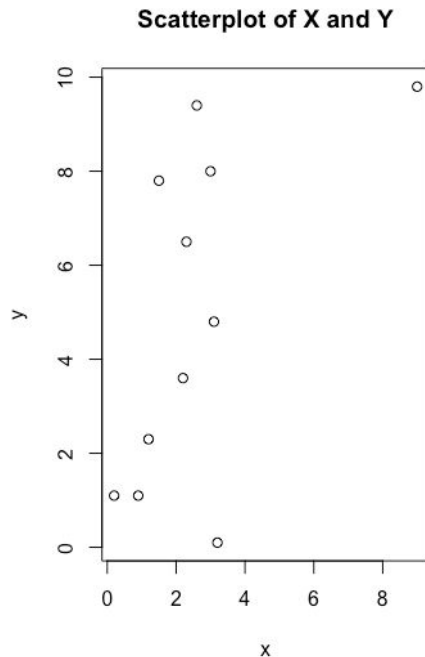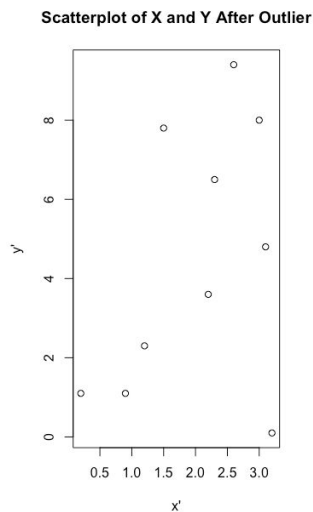

3.   Correlation coefficient: 0.5571167

Linear association: The linear association pictured above is considered to be a weak positive one as the correlation coefficient is not particularly high and resulting in a spread out positive line. But it is positive as it trends to the upper right.

4. Yes, outliers include: (9.0, 9.8)

New correlation coefficient: 0.3873604

5. The new scatter plot has one less data point resulting in:

Scatterplot of X and Y After Outlier

The new plot now has a correlation coefficient of 0.3873604 which is significantly less than what it was previously. A reason for this is because the outlier could have been giving off a false sense of positive linear association when in fact it was skewing the data. After removing the outlier it is clear that there was never really a positive linear association, hense the reduced correlation coefficient.

6. Based on the two normal Q-Q plots for X and Y, it seems as though X is more likely to be of normal distribution than Y as the data points follow more tightly with the normal distribution line with little divergence beside the outlier. The ends of Y's data diverge too greatly for it to be a likely candidate for a normal distribution.
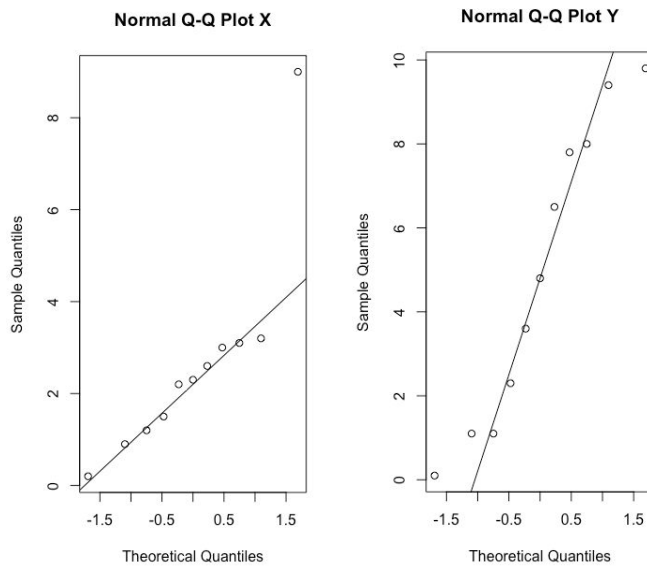
I pledge my honor that I have abided by the Stevens Honor System.
-   Eric Altenburg