Homework 2

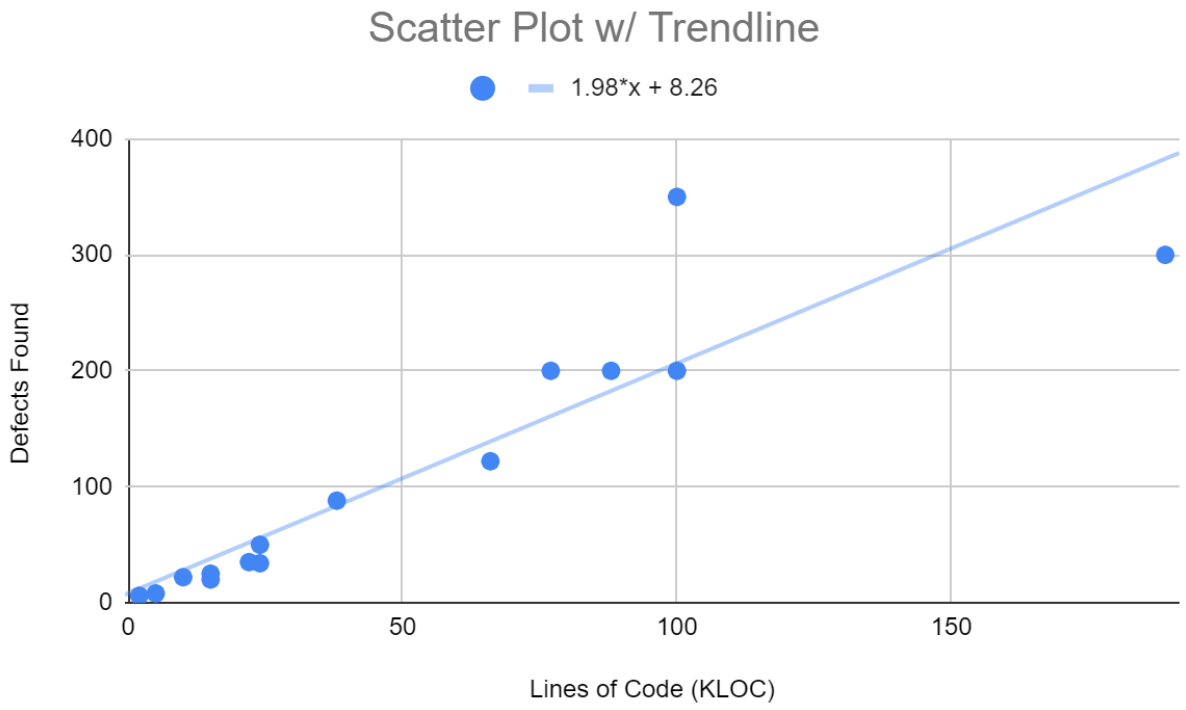
Team 9 - John McFarren, Erica O'Kelly, Devila Bakrania, Matthew Monaco

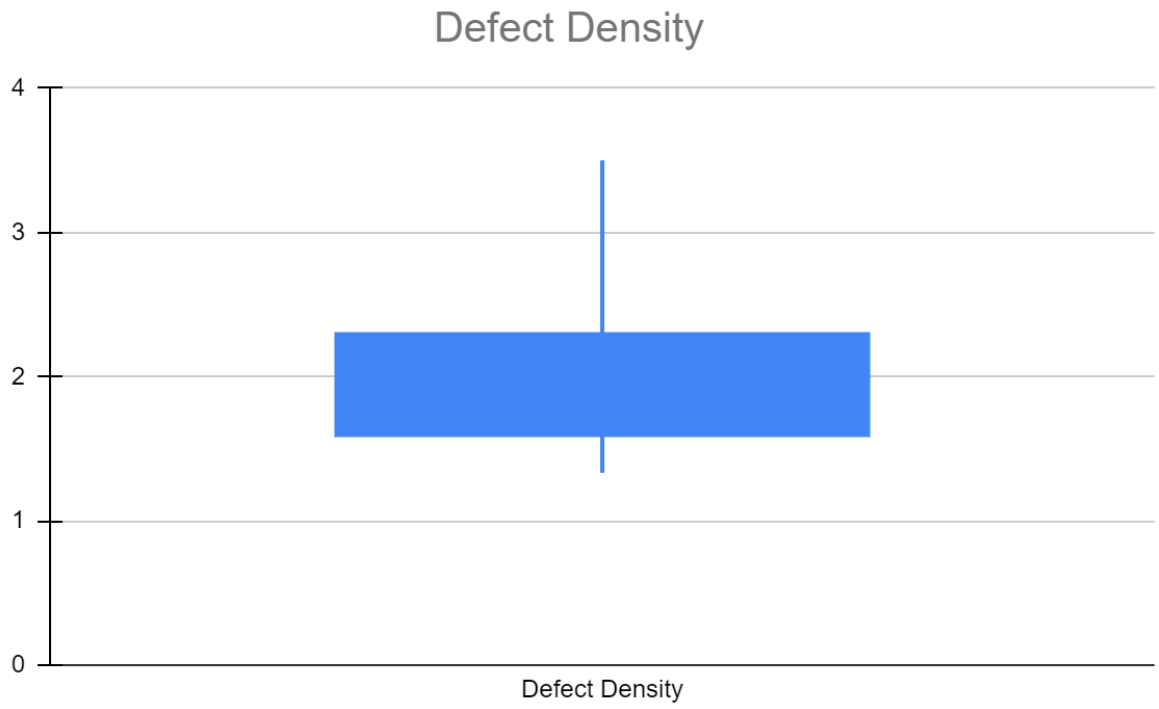
"I pledge my honor that I have abided by the Stevens Honor System." - JM, EO, DB, MM

Summary:

In this assignment, the team was originally tasked with finding whether there was a relationship between KLOC and Defects Found within a given dataset. It was determined that there was a strong correlation between KLOC and Defects Found. Additionally, using a trendline from the scatter plot of the dataset, the team was able to predict the number of defects in a program of a different size, in this case 50 KLOC. With such a correlation, the team found it possible to define a metric that is defect density. Using the normal distribution formula in Google Sheets, the team also plotted the defect density in a bell curve. The team experienced the most difficulty when trying to plot the different charts themselves, and focused mainly on making the graphs intuitive enough to enable the team to make valuable observations. Although this was difficult to do on the software that the team has access to, there is certainly value in the type of data analytics and estimation strategies that this assignment puts forward.

Question 1: Analyze the data above to understand the relationship (or not) between defects and size. Use scatter diagrams, correlation, Box and whisker, and trendline techniques.



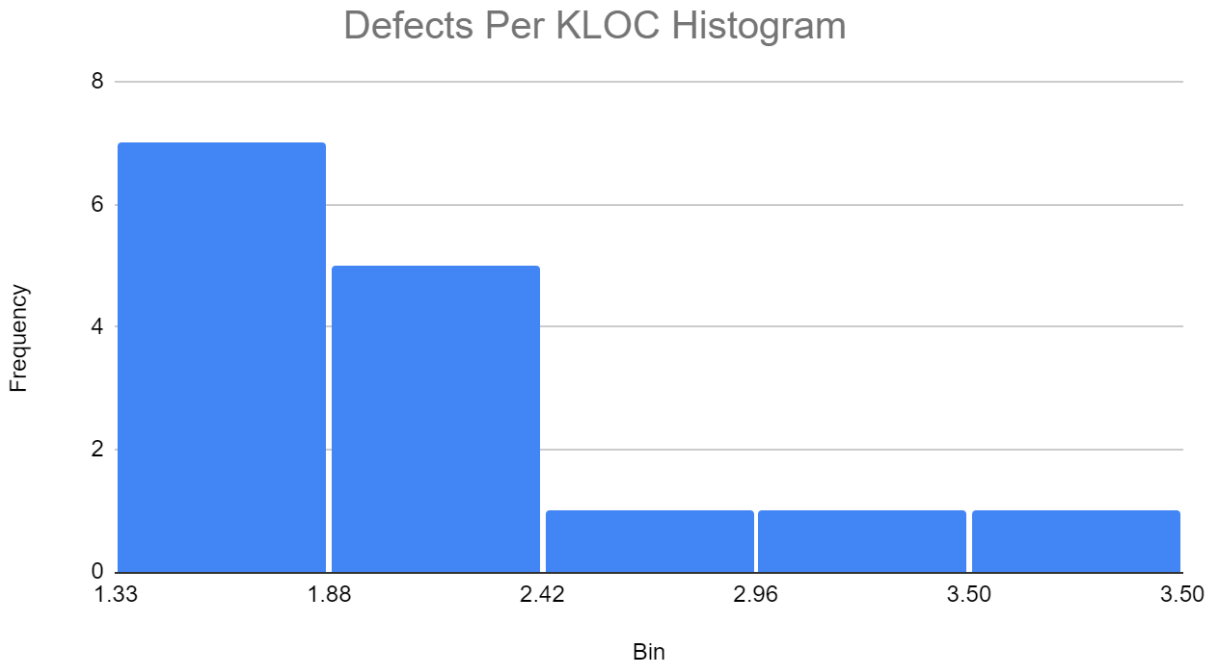


Using the =CORREL() function in Google Sheets, the correlation between LOC and Defects Found comes out to be .9083422813. There is definitely a strong correlation between LOC and Defects Found.

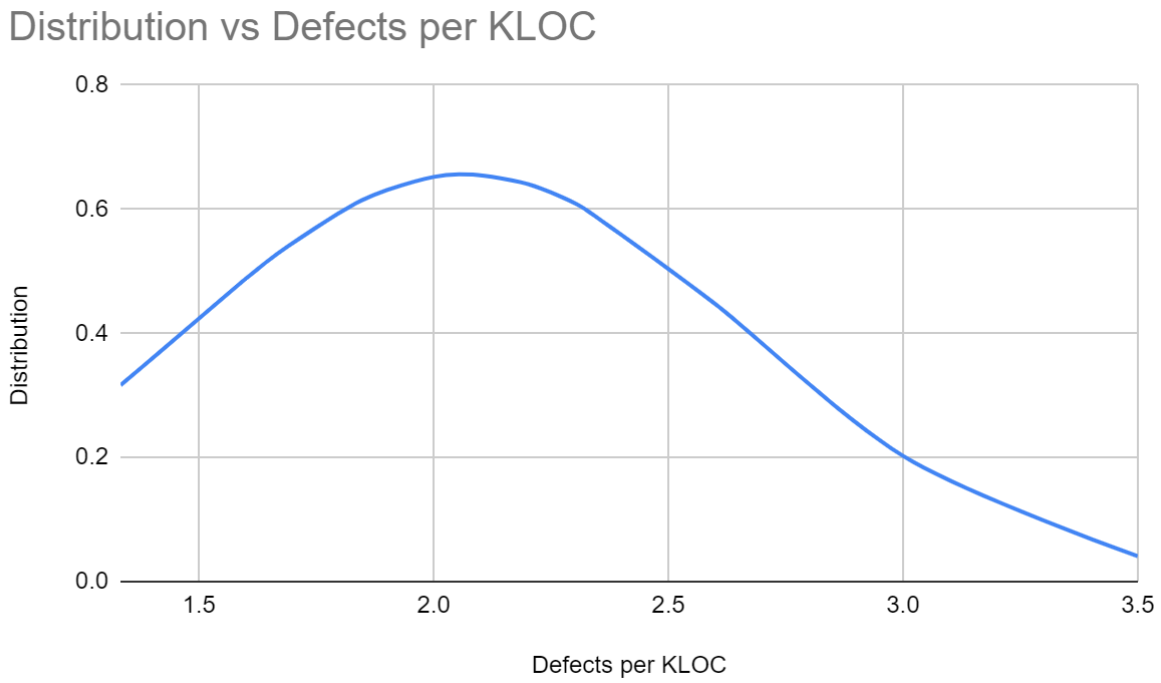
Question 2: Assuming there is, what do you predict as the expected number of defects for your 50K program?

Using the equation provided by the scatter plot's trendline, (1.98*x + 8.26), we can predict that there would be 107.26 (about 107) defects in a program with a size of 50 KLOC.

Question 3: You now define a metric called defect density (defects per KLOC).



Mean	2.067507658
25th Percentile	1.595454545
50th Percentile	2
75th Percentile	2.294258373
Standard Deviation	0.6080439207



Question 4: If you assume a normal distribution for defect density (defects per KLOC), what would be the expected range of number of defects for your 50K system, 95% of the time?

The range the group calculated was between 42.57 and 164.18. This was found by first finding the mean of the defects per KLOC which was 2.067. From there we then found the variance of the data to be .3697. After we took the square root of the variance to find the standard deviation of .608. We multiplied this by 2 to get the second standard deviation of 1.216. From there we scaled the values up by 50 in order to account for our 50K system. This gave us 103.38 as our mean and 60.8 as our standard deviation.

Question 5: Do you think a normal distribution is reasonable or not for the data given?

We believe that it is reasonable to think of a normal distribution for the given data. The curve of the distribution in question 3 looks relatively normal. Additionally, our predicted number of defects of 107 falls within the range stated in question 4. While the histogram in question 3 appears skewed to the right, I believe that this is more attributed to bin size and small sample size.