Data Analysis using R

Veeral Patel

Regis University

March 26th, 2021

1. **Introduction**

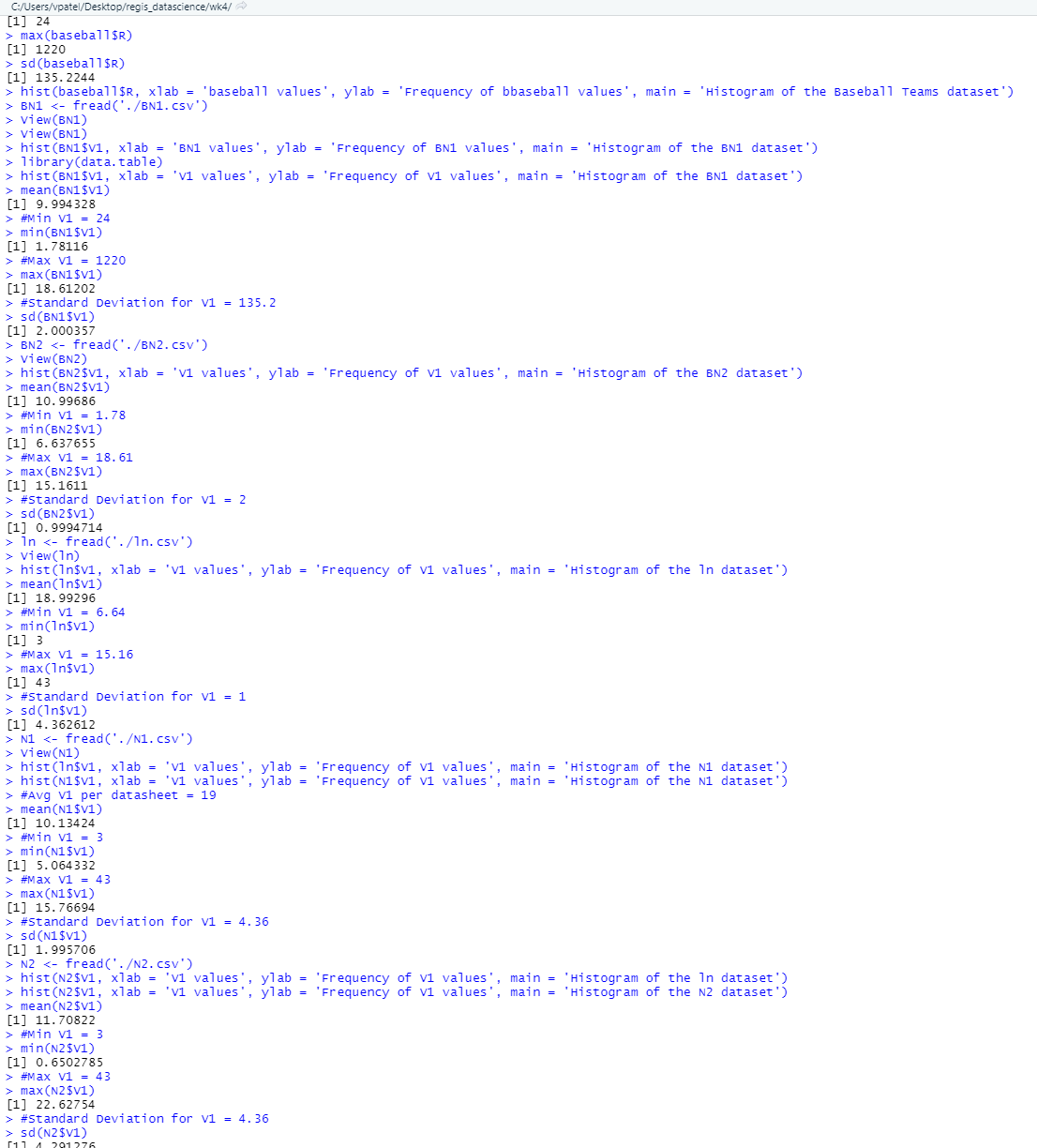
For this project, we were expected to download R and R Studio in order to be able to open data files, and use R for data analysis. We were using a dataset that was included in one of the R packages. This dataset included species of plants and their petal measurements along with other measurements of different species of plants. We were to use R in order to perform mathematical operations on that dataset which helps us gather more insight on the data. We worked with many datasets for this homework, 6 in total.

1. **Methods/Code/Screenshots**

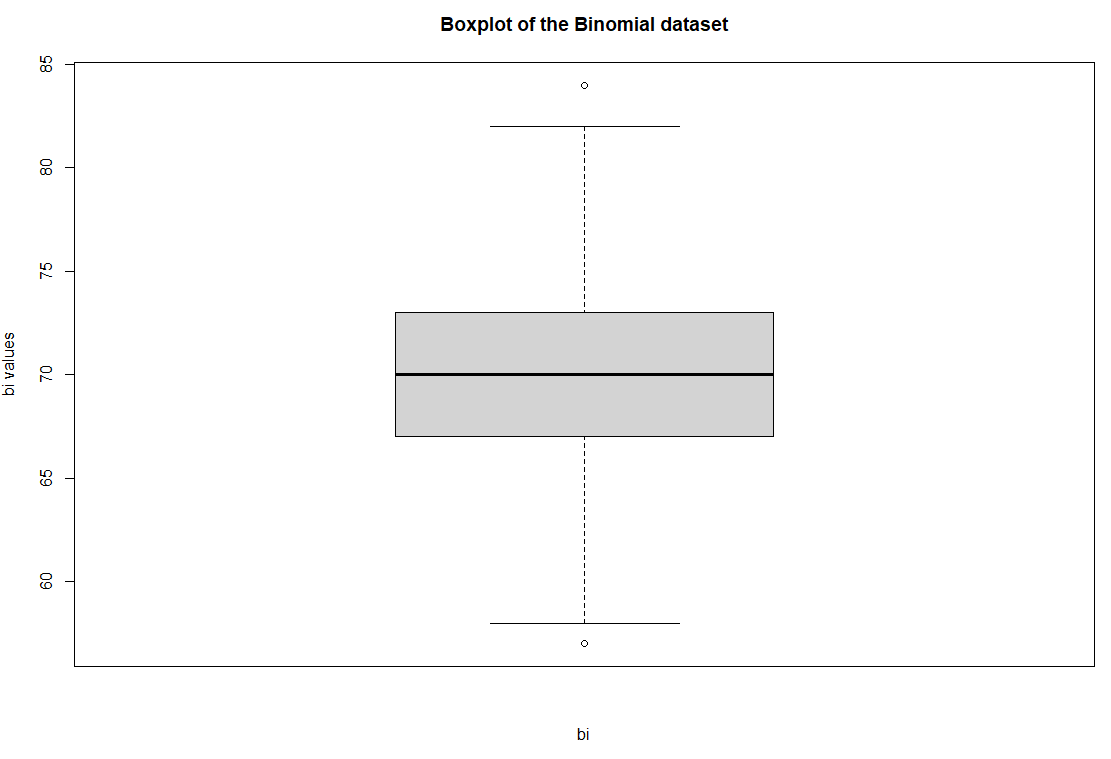
I used a windows computer and I had to install R, R Studio (IDE) and all of the dependency modules that were needed to perform the data analysis. I have worked with R before but it was very brief and it was in a science class. I had only used R before to calculate averages of large sets of data that was in an .rda format which is like a really compressed file. This was a new adventure for me to work with R Studio with more data analysis. I watched a few quick tutorials on YouTube in order to assist me with getting everything installed on my computer. Most of the code in this assignment was supposed by the professor. I was in charge of coming up with a few calculations that included standard deviation, mean, box plot, histogram, and answering questions about the dataset. I also found it generally helpful to actually understand the data by just opening it up in R studio in a data frame before getting started with the assignment. I also used the head() function in order to get a quick glimpse of the data within R Studio.

1. **Results/Analysis**

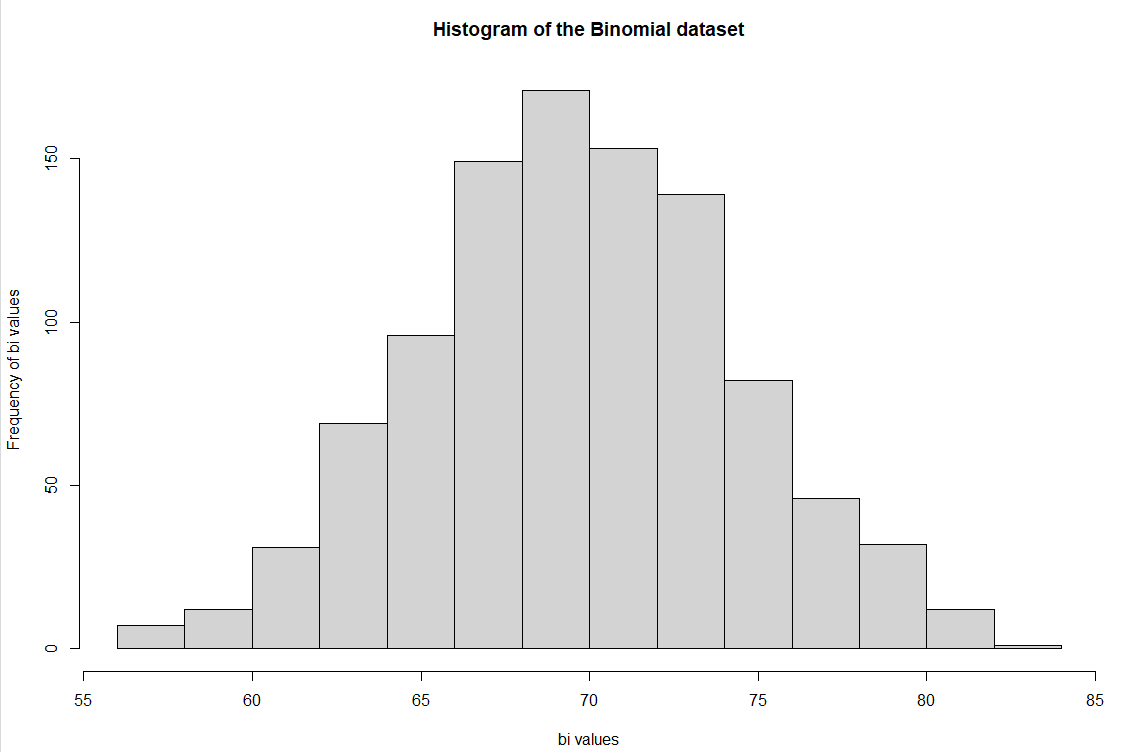
Summary Statistics of datasheets:



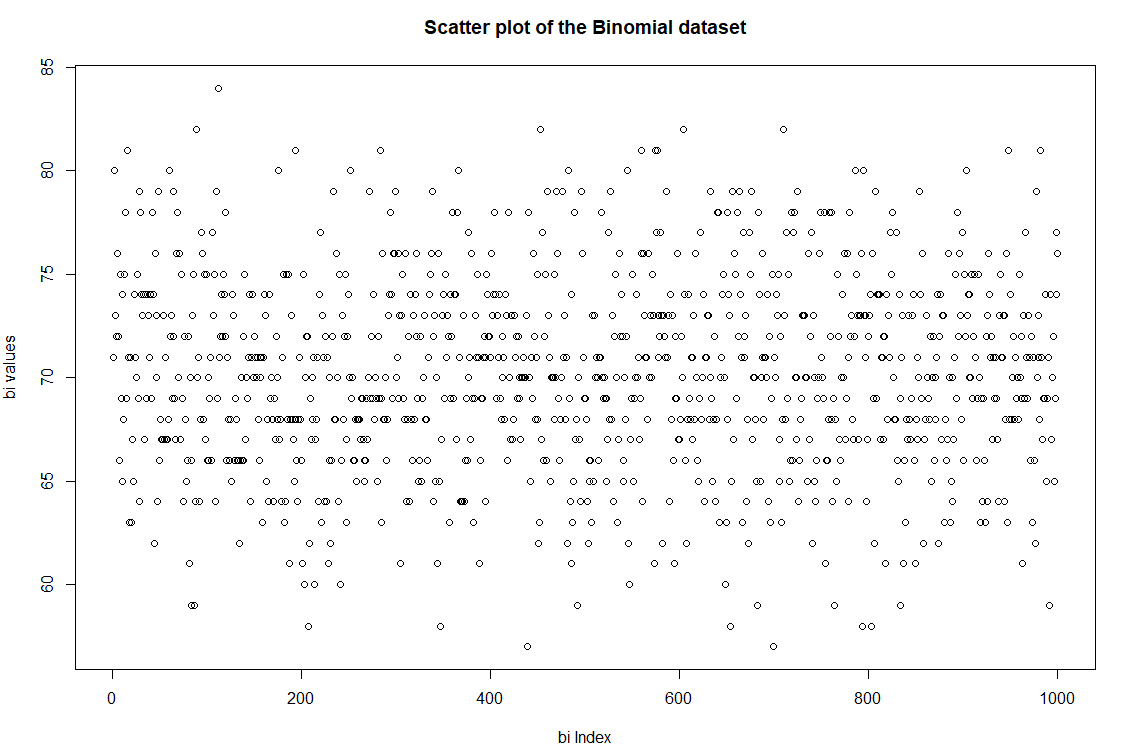
Binomial Data Box Plot:



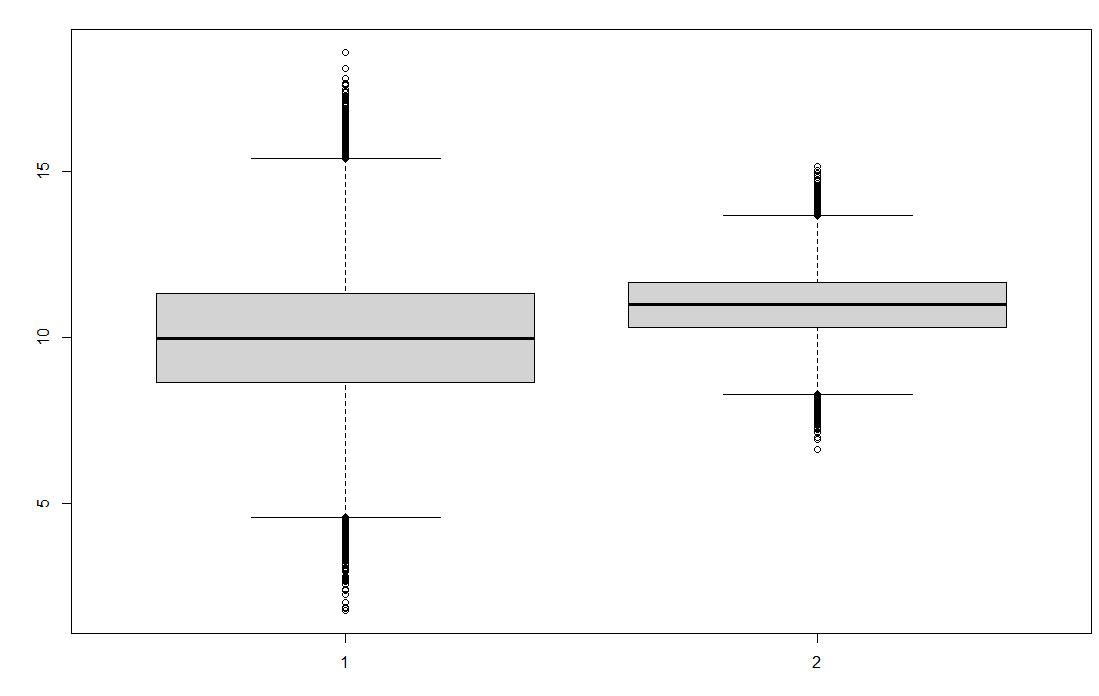
Binomial Data Histogram:



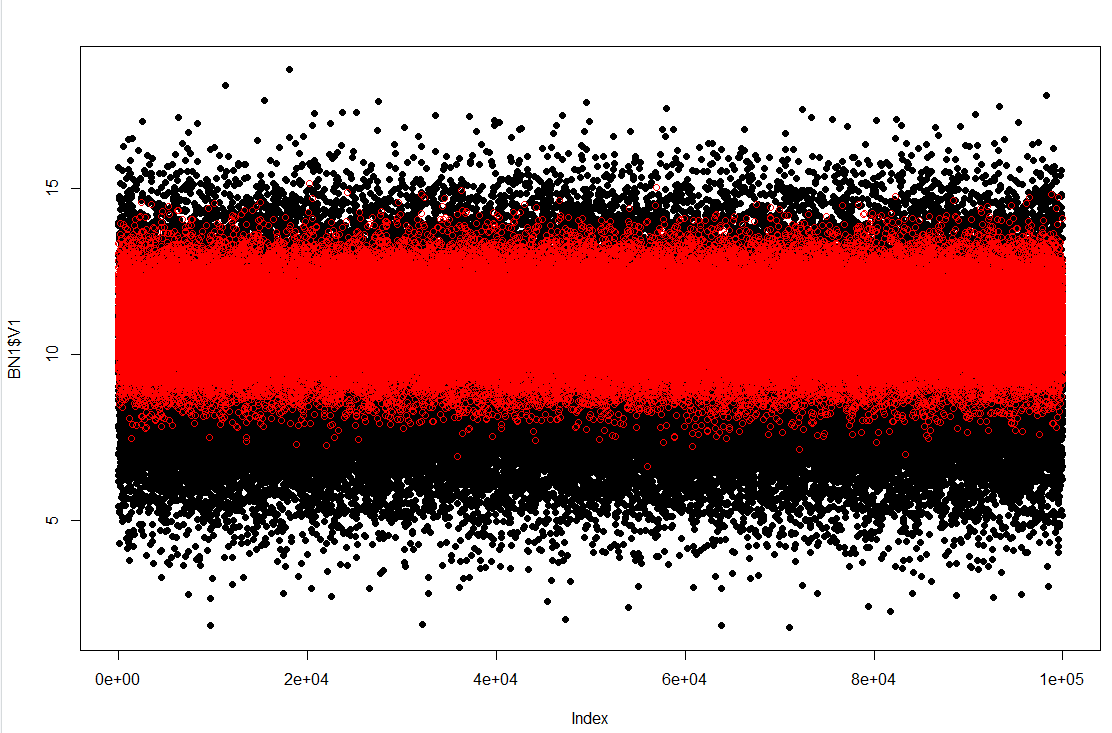
Binomial Data Scatter Plot:



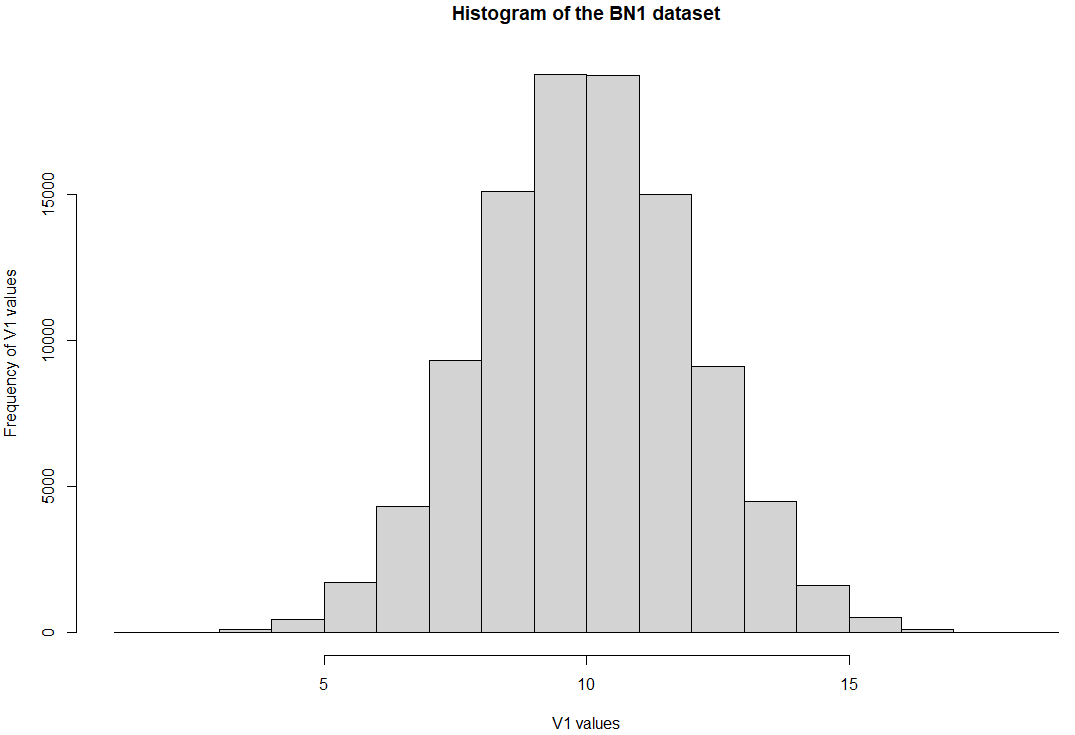
BN1 + BN2 Box Plots:



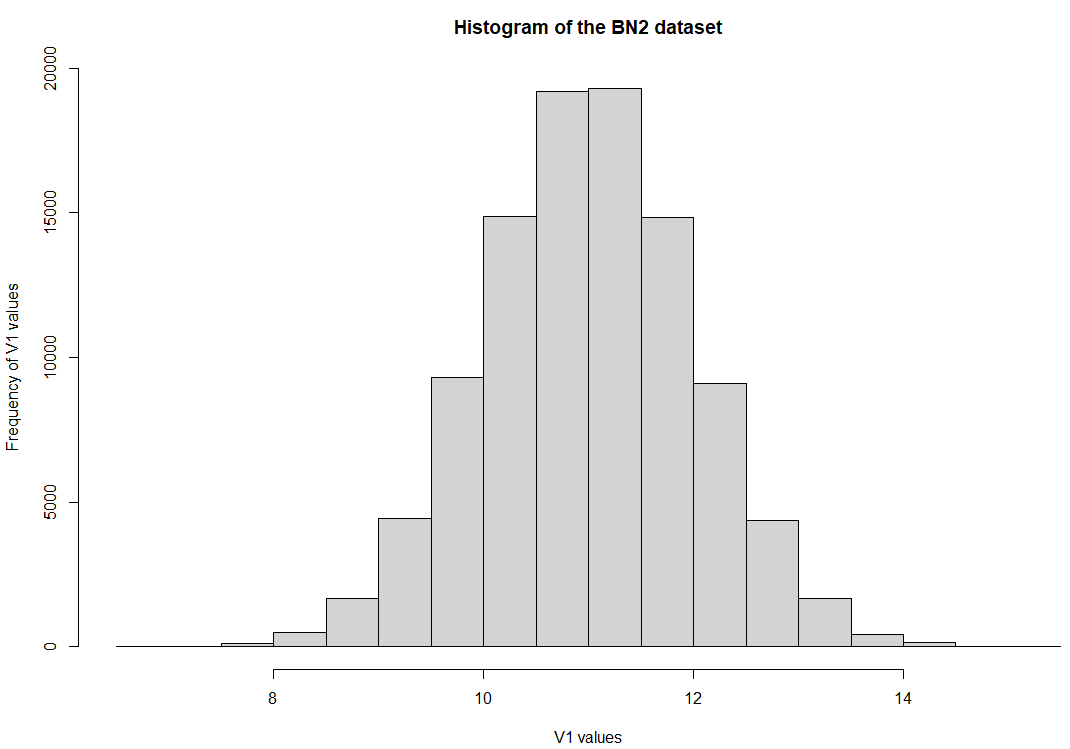
BN1 + BN2 Scatter Plots:



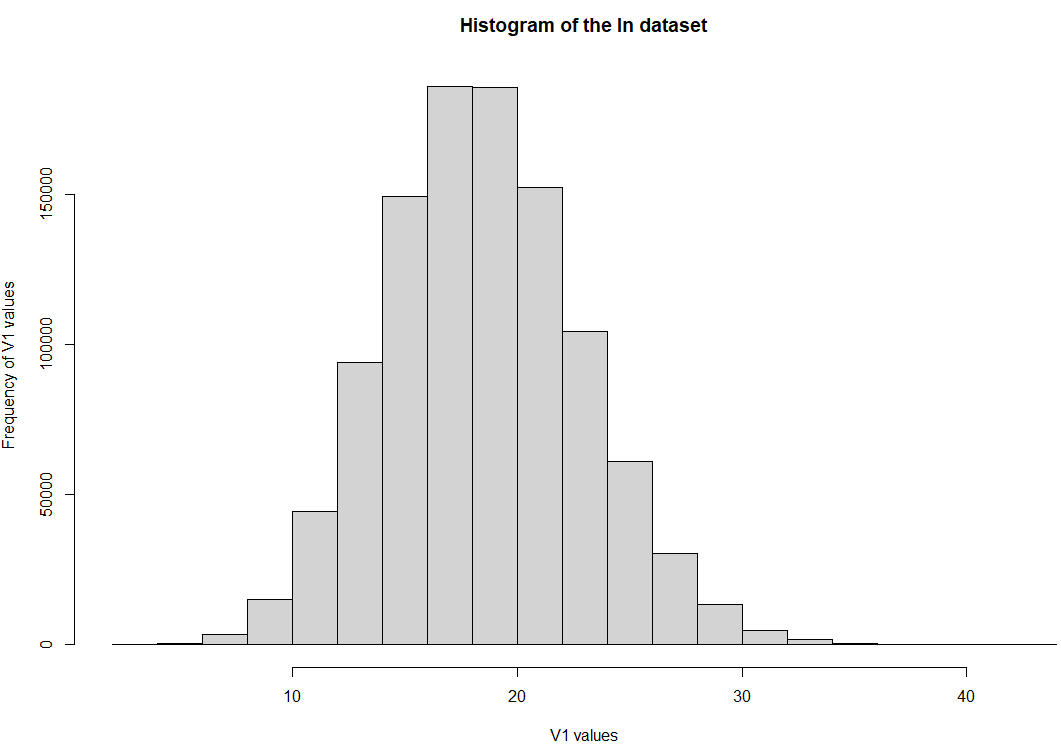
Bn1 V1 Data Histogram:



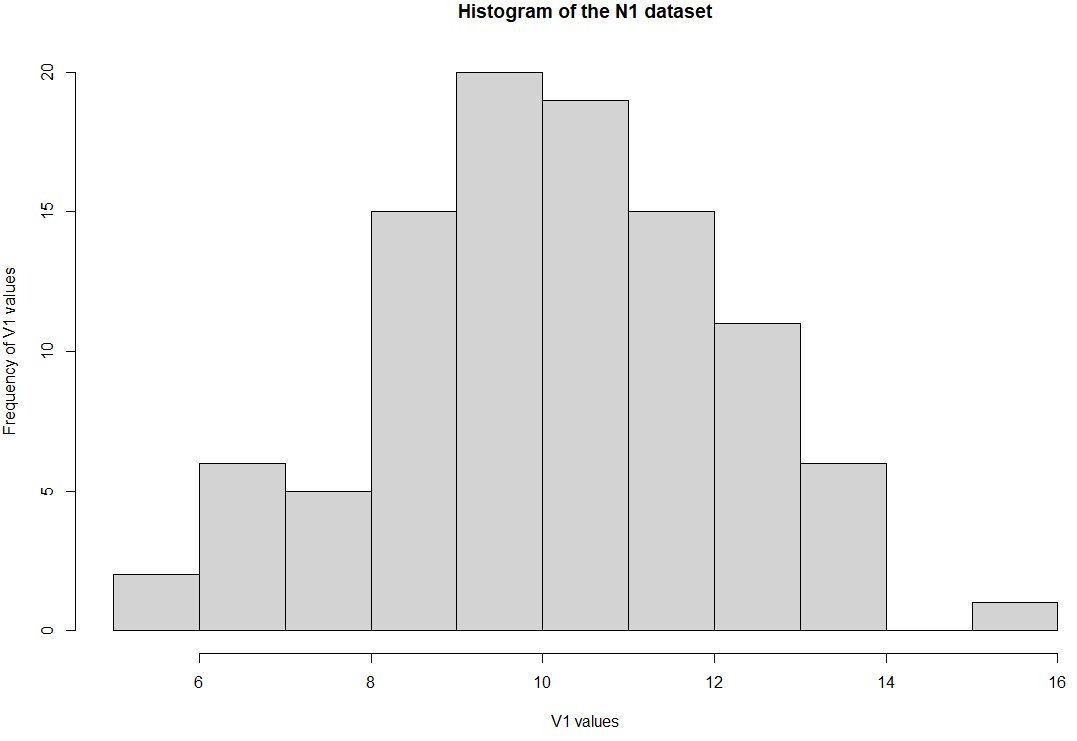
BN2 V1 Data Histogram:



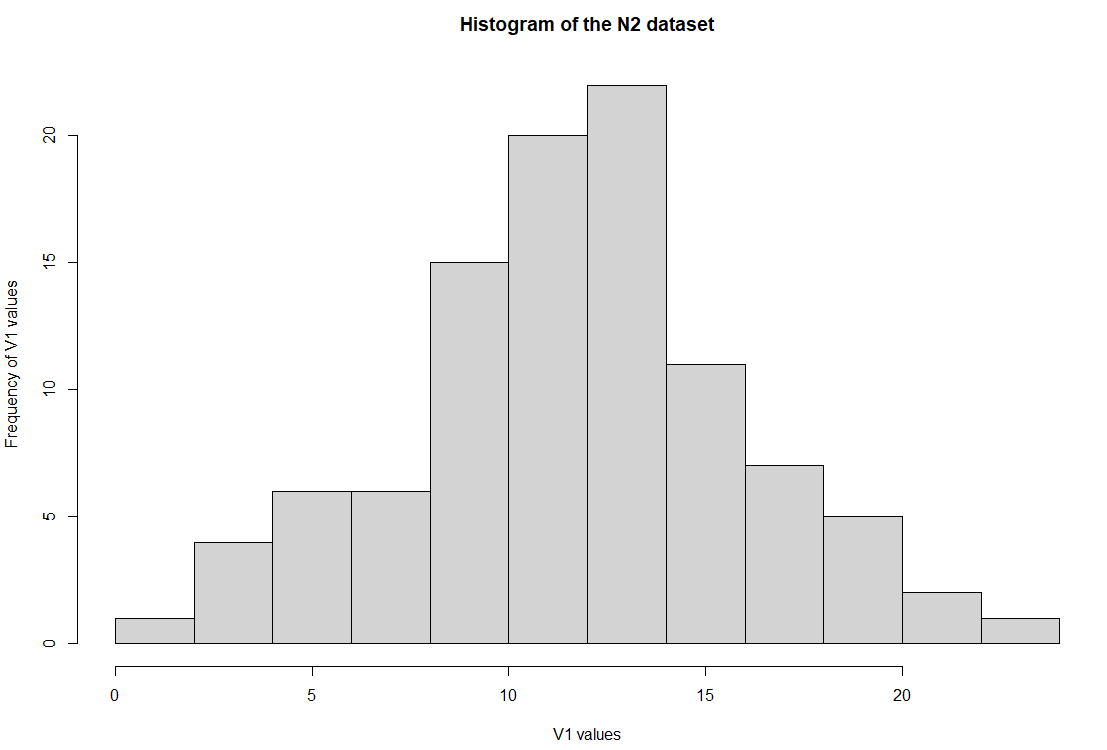
Ln Data Histogram:



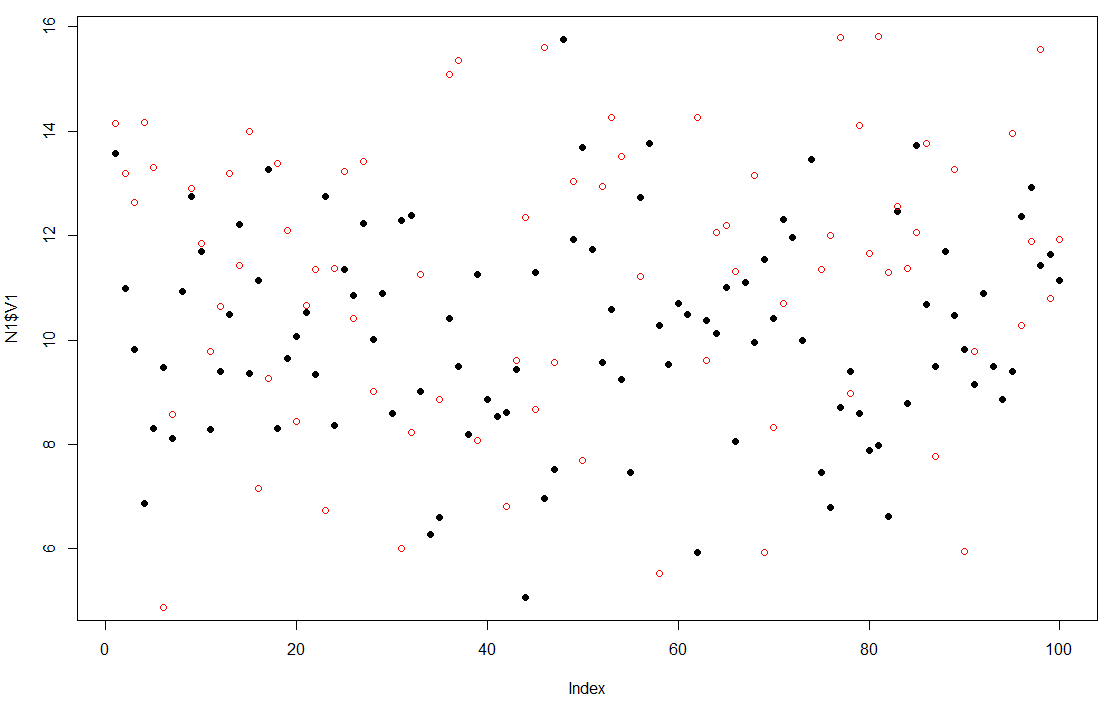
N1 V1 Data Histogram:



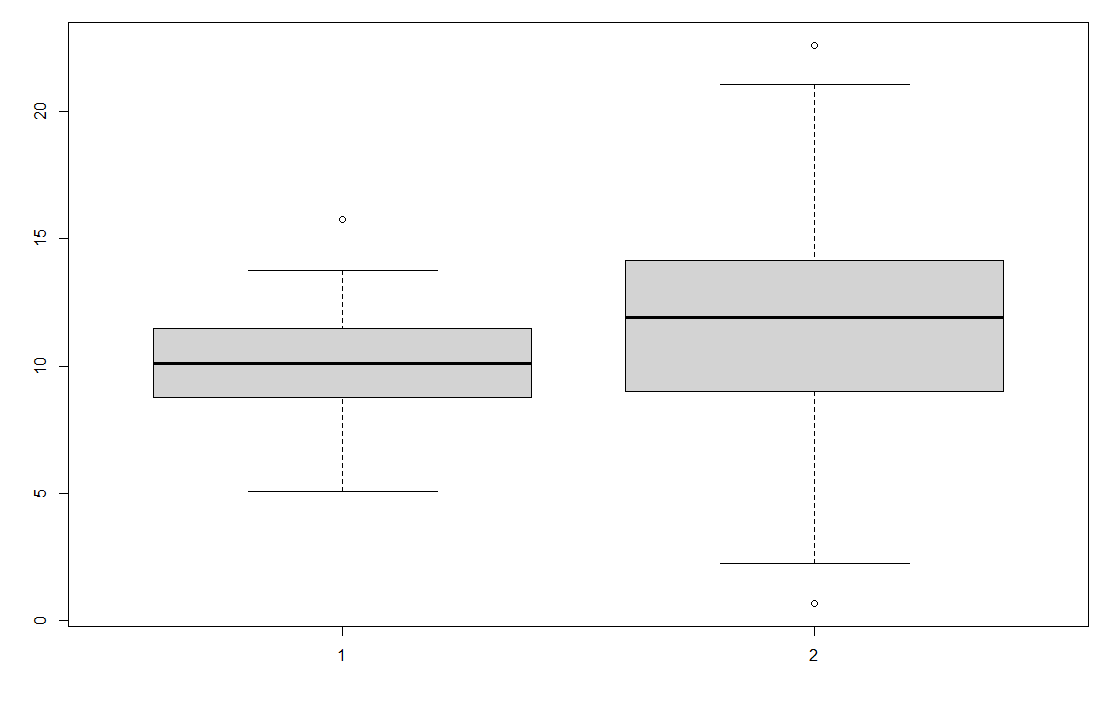
N2 V1 Data Histogram:



N1+N2 Scatter Plot:



N1 + N2 Box Plots:



1. **Analysis of Results**

In order to analyze the results, we must answer that questions that were given to us first.

***What, if anything, can be said about the differences between N1 & N2?***

When looking at the scatter plots it is actually very difficult to discern the difference between the two. However, you can really start seeing the differences when you look at the standard deviation as well as the box plots. You can tell that by the standard deviation, that N1 has a much lesser standard deviation than N2 does. This tells us that N2 V1 values are much more spread apart. This can also be confirmed by looking at the box plots for each data set. N1 data is much more squished within the quartiles when looking at the box plot. N1 also has a much smaller gap between the min and the max.

***What, if anything, can be said about the differences between BN1 & BN2?***

When looking at the differences between BN1 and BN2, the scatter plots actually show a lot of what is going on. You can quickly see that BN1 has scatter points that are in a much wider range compared to the BN2 dataset. This is also confirmed by looking at the standard deviation, BN2 has a much smaller standard deviation, meaning that the dataset for BN2 is much less spread apart. Again, the boxplot also confirms this analysis too. The BN2 box plot is much more squished, meaning that the quartiles are lesser apart and that is where the majority of the data lies.

***What are the differences in the distributions, how do the outliers differ, what are the differences in the means, range, etc?***

The min and max are much more of an outlier in the N2 dataset. In the N2 dataset, there is a difference of 22 between the min and the max. In the N1 dataset, the difference between the min and the max in 10.77. This shows that the range of values for N2 is larger. Also the mean for N1 is and N2 is very similar, while the standard deviation for N2 is larger, meaning the data is more spread out.

The min and max are much more of an outlier in the BN1 dataset. In the BN1 dataset, there is a difference of 17 between the min and the max. In the BN2dataset, the difference between the min and the max in around 9. This shows that the range of values for BN1 is larger. Also the mean for BN1 is and BN2 is very similar, while the standard deviation for BN1 is larger, meaning the data is more spread out.

***When considering the vector individually, check if the distribution is normal or not -- which stats would you use to determine this?***

I can tell that they distribution is normal by looking at the histogram. The histogram shows us that the data is following a bell shaped pattern that indicates a normal distribution.

1. **Conclusion**

I thought that this was a really good example of being able to use simple arithmetic with a dataset in R to learn the basics of data science in R. This homework consisted of more than 1 dataset, which I thought was a really good way to learn how to work with more data as well as graph data together that is coming from different sources. I liked the baseball dataset because we were able to decide by ourselves what statistics we wanted to see from what data metric. I chose to use Runs for my analysis that it was fun to see the wide range of runs that were the min, max, average over the years. It was also fun to see how the runs over the years kept going up, meaning that the game was changing.

1. **Reference:**

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O’Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 1-7.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Torgo, L. (2016). *Data mining with R: learning with case studies*. CRC press.