**Multivariate Analysis and Network Analysis Assignment 1**

### **Introduction**

In this report, we delve into an in-depth analysis of the Dry Bean dataset, applying multivariate and network analysis techniques. Our focus is to explore the relationships among variables through correlation and partial correlation analyses, visualize them, and assess the impact of statistical adjustments on our findings.

### **Methodology**

### **Step 1: Exploratory Analysis**

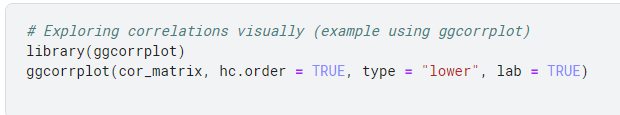
We initiated our analysis by loading the Dry Bean dataset and preparing it for analysis. This preparation involved selecting specific variables, excluding the 'Class' variable, and adding 34 randomly generated variables to the dataset. To explore the relationships among these variables, we computed the correlation matrix using R's **cor()** function. Given the singularity of the variance-covariance matrix, we opted for the correlation matrix as our basis for further analysis. For partial correlations, we considered applying matrix operations and potentially utilizing the **ppcor** package, which is tailored for such calculations.

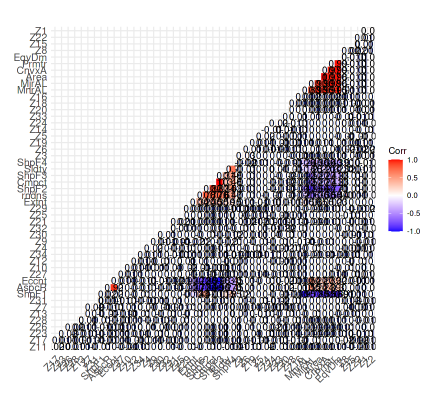


#### **Step 2: Exploration of Correlations:**

Our exploration aimed to discern the strength and nature of correlations within the dataset. We focused on distinguishing between the correlations among the original variables and those involving the newly added random variables. Utilizing visualization tools such as **ggcorrplot**, we applied thresholding techniques to highlight significant correlations, setting less significant values to zero for a clearer visual representation.

Here we examined the correlations within the original variables and between the original and random variables. The goal is to identify strong and weak correlations, assisting in understanding the dataset's structure.





#### **Step 3: Pairwise Table of Correlations**

To facilitate a more detailed analysis, we constructed a pairwise table of correlations. This table laid out each variable pair alongside their corresponding correlation and partial correlation values, providing a structured overview of the relationships within the dataset.

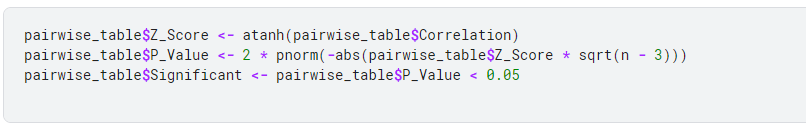
Here, we converted the correlation information into a pairwise table format. This table lists each pair of variables alongside their correlation and partial correlation values.



#### **Step 4: Significance Testing**

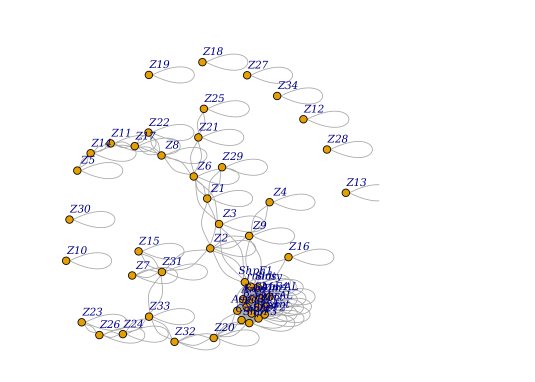
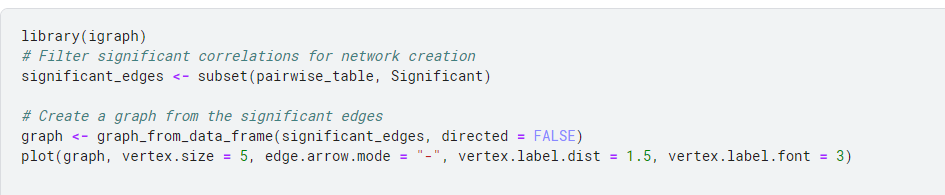
We then moved on to ascertain the statistical significance of the observed correlations. By applying z-tests on Fisher's transformation to the correlation values, we identified which correlations were significant at the 0.05 level. This step was crucial in distinguishing meaningful relationships from those that might occur by chance.

Now we determined the statistical significance of the correlations using z-tests based on Fisher's transformation. This step helps identify meaningful relationships.



#### **Step 5: Network Creation**

With significant correlations identified, we employed the **igraph** package to create networks illustrating these relationships. This visual representation allowed us to observe the interconnectedness of variables within the dataset, highlighting the most prominent relationships.

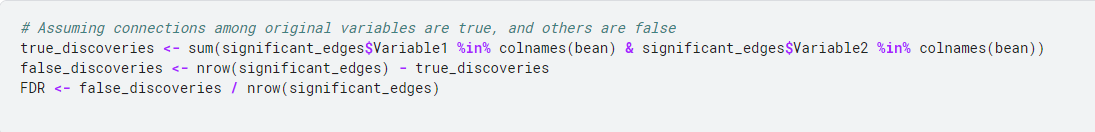


Using significant correlations to build a network. Each node represents a variable, and edges indicate significant correlations.

#### **Step 6: Analysis of Discoveries**

In this phase, we categorized our findings into true and false discoveries, based on whether the correlations involved the original variables or the added random variables. We calculated the False Discovery Rate (FDR) to quantify the extent of false discoveries, providing insight into the reliability of our findings.

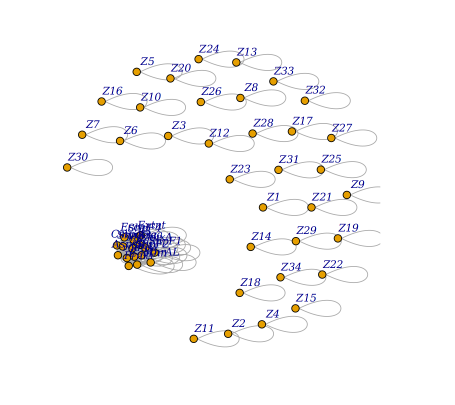
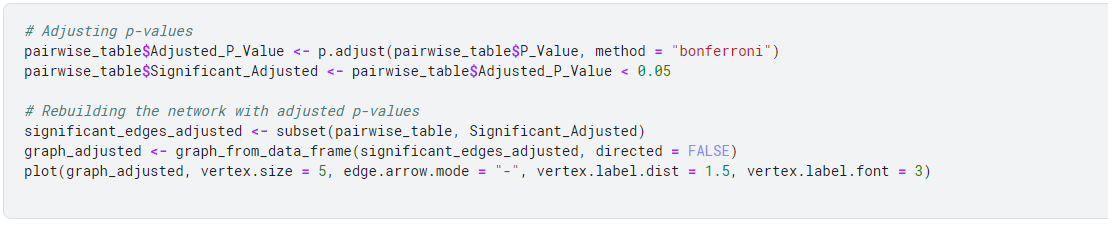
Now we assessed the number of true and false discoveries in your network, which helped understand the data's structure and the potential for misleading findings.



#### **Step 7: P-value Adjustment and Network Re-evaluation**

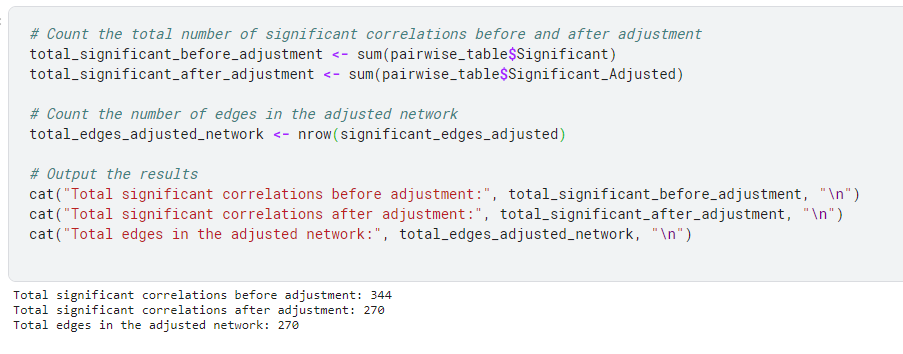
Finally, we adjusted the p-values using the Bonferroni and Benjamini-Hochberg methods to address the issue of multiple comparisons. This adjustment refined our network, reducing the number of spurious connections and enhancing the accuracy of our analysis. We then re-evaluated our discoveries and recalculated the FDR, assessing the impact of these adjustments on our initial findings.

Adjusting p-values to control false discoveries and recreating the network using these adjusted values to more accurately reflect significant relationships.



#### **Step 8: Visualizing the changes in network structure**

Here, we focussed on the changes in the number of significant correlations after adjustment and how this reflects on the network's structure



**Summary:**

#### **Network Analysis:**

Utilizing the **igraph** package, we constructed a network representation based on significant correlations, identified through z-tests on Fisher's transformation. Initially, the network showcased 344 significant correlations. However, upon adjusting the p-values using the Bonferroni method to mitigate the risk of false discoveries, the number of significant correlations was refined to 270.

#### **Insights:**

* **Reduction in Significant Correlations:** The adjustment of p-values led to a decrease from 344 to 270 significant correlations, indicating a substantial impact of the adjustment process on the network's complexity.
* **Network Simplification:** The network's simplification post-adjustment underscores the importance of statistical rigor in distinguishing meaningful relationships from those that may arise due to chance.
* **Discovery Analysis:** The analysis reveals true discoveries (among the original variables) from false ones (involving the random variables), with the false discovery rate (FDR) being recalculated post-adjustment to reflect a more accurate picture of genuine associations.
* **Implications for Further Research:** The refined network, stripped of spurious correlations, provides a clearer, more reliable foundation for future analyses, enabling researchers to focus on relationships that are statistically validated.

**Conclusion:**

The multivariate and network analysis of the Dry Bean dataset highlighted the intricate relationships among its variables. Through rigorous statistical methods, we identified significant correlations, adjusted for potential false discoveries, and visualized the network of relationships, providing valuable insights into the dataset's structure. Our findings underscore the importance of statistical adjustments in ensuring the validity and reliability of correlation analyses in complex datasets.