FINAL REPORT

MULTIVARIATE STAT ANALYSIS

STA 5384 01

SALMAN IMTIAZ

&

ANNA RAMASAMY

1. **Glass Dataset:**

1.1 High Positive Correlations:

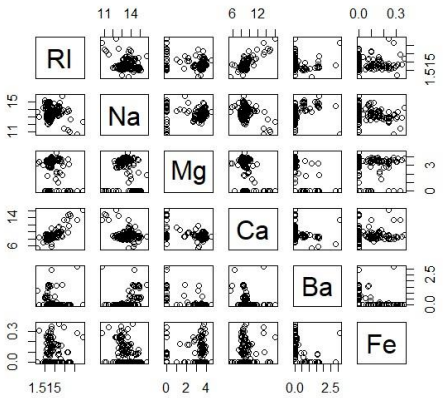
RI and Ca (0.82): There is a strong positive correlation between the refractive index (RI) and calcium (Ca). This suggests that as calcium content increases, the refractive index of the glass also tends to increase. Ba and Na (0.39): Barium (Ba) and Sodium (Na) also show a moderate positive correlation.

1.2 High Negative Correlations:

Mg and Ba (-0.63): Magnesium (Mg) and Barium (Ba) have a strong negative correlation, indicating that increases in magnesium content are associated with decreases in barium content within the glass. Mg and Ca (-0.43): There is also a notable negative correlation between magnesium and calcium.

1.3 Low or No Correlation:

Ba and Fe (-0.07): Barium and iron (Fe) show very little correlation, suggesting no significant linear relationship between them. The insights from this correlation matrix can guide deeper analyses such as identifying elements that influence the glass's properties significantly. For instance, the strong relationship between RI and calcium could be further explored to understand its impact on the optical properties of glass.



2. Diagonal Histograms:

RI (Refractive Index): Appears to have a roughly normal distribution.

Na (Sodium): Shows a somewhat normal distribution with possible skewness.

Mg (Magnesium): This might exhibit a bimodal distribution or have a majority of values concentrated around a specific point, indicating two groups or types of glass.

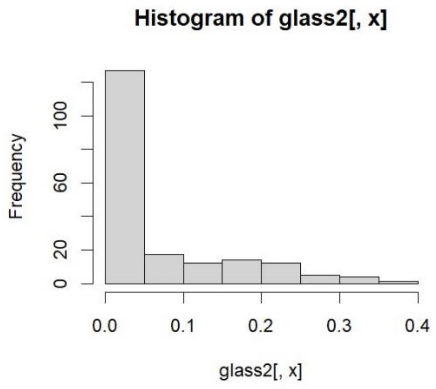
Ca (Calcium), Ba (Barium), and Fe (Iron): These distributions vary, with Calcium and Iron seeming slightly skewed, and Barium showing a very skewed distribution with many low values.

Off-Diagonal Scatter Plots: These show the relationships between pairs of variables.

Strong Positive Relationship: Noted between RI and Ca, as the scatter plot shows a clear upward trend, confirming their high correlation value from the correlation matrix.

Dispersed and No Clear Patterns: For pairs like Mg and Ba, where the points are widely dispersed, indicating the strong negative correlation (as magnesium increases, barium decreases).

Clusters or Specific Groupings: For instance, in the plots involving Ba, data points tend to cluster at lower values, suggesting limited variation in Barium's content for most glass types.



3. Analysis of Histogram of glass2[, x]:

3.1. Distribution Shape:

The histogram displays a heavily right-skewed distribution, with the majority of data points concentrated in the lower range close to 0.0. This indicates that for this specific variable in the dataset, lower values are much more frequent than higher values.

3.2. Central Tendency and Spread:

The mode of the distribution (most frequent value) is clearly in the lowest bin, suggesting that the typical value of this variable is near zero.

The spread of the data extends up to 0.4, but with decreasing frequency as the value increases.

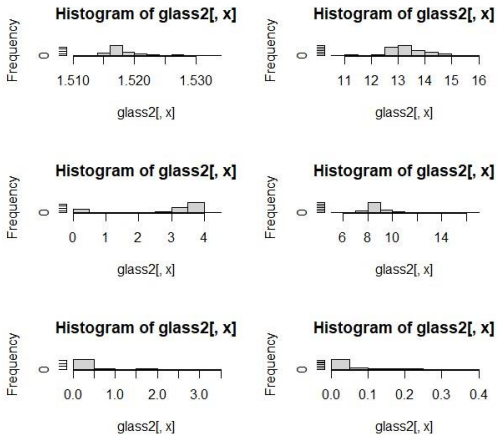
3.3. Implications:

The concentration of values near zero could imply that this variable is a minor component in the glass composition, or it could be an indicator of a specific type of glass that predominantly lacks this component. The presence of a few higher values, though less frequent, might indicate outliers or special cases where this component is more significant.

3.4. Statistical Considerations:

Given the skewed nature of the distribution, median and mode might be more informative measures of central tendency than the mean, which could be influenced by the higher

values on the right tail. The skewness suggests that any statistical modeling involving this variable should consider transformations or non-parametric methods to handle the non-normality.



4.1. Histogram of RI (Refractive Index)

Range: 1.510 to 1.530. Most values are densely packed around a narrow range, suggesting a high consistency in refractive index across the dataset. The distribution is slightly skewed to the right.

4.2. Histogram of Sodium (Na) Content

Range: 11 to 16. The distribution seems fairly uniform across a broader range, indicating variability in sodium content among the samples.

4.3. Histogram of Magnesium (Mg) Content

Range: 0 to 4. The histogram shows a right-skewed distribution with most values concentrated at the lower end, indicating that many glass samples have low magnesium content.

4.4. Histogram of Calcium (Ca) Content

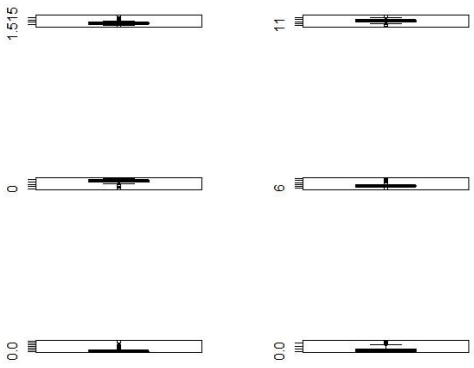
Range: 6 to 14. The distribution is moderately spread with a slight skew to the right, suggesting a variation in calcium levels among the glass types.

4.5. Histogram of Barium (Ba) Content

Range: 0 to 3. Most values are zero, indicating a lack of barium in many samples, but a few samples have higher values, which could indicate a specific type of glass.

4.6. Histogram of Iron (Fe) Content

Range: 0.0 to 0.4. This histogram shows most samples have very low iron content, with a concentration of values near zero, similar to the barium distribution.



5. Box Plot Analysis

5.1. Box Plot for RI (Refractive Index)

Central Value: Median is around 1.517.

Spread: The interquartile range (IQR) is narrow, indicating consistent values across samples.

Outliers: No visible outliers, indicating a uniform distribution within the observed range.

5.2. Box Plot for Sodium (Na) Content

Central Value: Median sodium content is around 13.

Spread: Slightly wider IQR compared to RI, suggesting more variability in sodium content among the glass samples.

Outliers: A few outliers are present, indicating samples with unusually high sodium content.

5.3. Box Plot for Magnesium (Mg) Content

Central Value: Median is low, close to 0, reflecting the histogram's finding of low magnesium content in many samples.

Spread: Narrow spread, indicating most samples have similar magnesium levels.

Outliers: Few to no outliers, suggesting magnesium content is consistent across the dataset.

5.4. Box Plot for Calcium (Ca) Content

Central Value: Median calcium content is around 8.

Spread: Moderate spread indicating some variability in calcium content.

Outliers: Some outliers on the upper end, suggesting a few samples with significantly higher calcium levels.

5.5. Box Plot for Barium (Ba) Content

Central Value: Median is at zero, aligning with the histogram which showed most values clustered at zero.

Spread: Extremely narrow IQR, almost non-existent, which confirms that barium is absent in most of the samples.

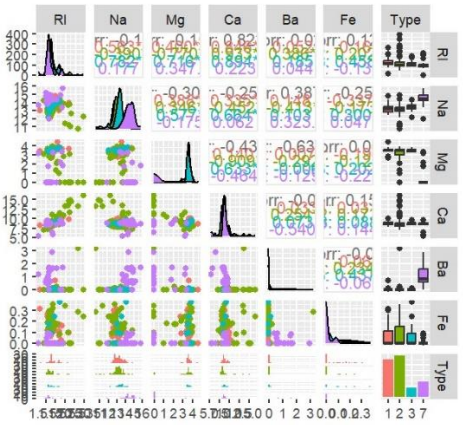
Outliers: A few outliers present, which are the samples that contain barium.

5.6. Box Plot for Iron (Fe) Content

Central Value: Median close to zero, reflecting very low iron content in most samples.

Spread: Very narrow IQR, indicating low variability in iron content.

Outliers: Some outliers indicating a few samples with higher iron content.



6. Analysis of the ggpairs Plot

6.1. Diagonal Panels (Histograms and Density Plots):

Histograms and Density Plots are used for each variable, giving a quick visual representation of the distribution of values within each variable. For example, the histograms for RI show a roughly normal distribution, whereas the histograms for Ba are highly skewed with most values at zero, confirming earlier analysis.

6.2. Lower Triangle Panels (Scatter Plots):

Scatter Plots in the lower triangle show the relationships between pairs of variables. The color coding by glass type enhances the visibility of patterns such as clustering or correlations. For example: Scatter plots of RI versus Ca show a clear positive correlation, much like previously analyzed in the correlation matrix. Different colors represent different types of glass, helping to identify which variables are good predictors of glass type.

6.3. Upper Triangle Panels (Correlation Coefficients and Ellipses):

Correlation Coefficients are numerically displayed, providing a quantitative measure of the relationships between variables. Ellipses indicate the strength and direction of the correlation, with narrower ellipses showing stronger correlations.

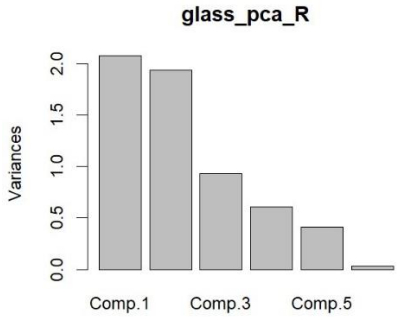
6.4. Right-Side Panels (Box Plots):

Box Plots provide summaries of the distribution of each variable across different types of glass. They illustrate medians, interquartile ranges, and potential outliers within each group.

6.5. Implications for Glass Classification:

Variable Relationships: The visualization aids in understanding how variables interact with each other across different glass types, which is essential for predictive modeling and classification.

Type Differentiation: The distinct patterns in scatter plots and box plots across different glass types suggest that some variables are particularly effective in distinguishing between types. For instance, the presence or absence of certain elements like Ba and Fe can be key indicators of specific glass types.



7. Analysis of PCA Variance Bar Chart

7.1. Principal Components Shown:

Comp.1, Comp.3, Comp.5: These labels indicate specific principal components. Typically, the first few components capture the most variance.

7.2. Variances Indicated:

Comp.1: Accounts for the highest variance among the components shown, with a variance value close to 2.0. This indicates that this component captures a significant portion of the information in the dataset.

Comp.3: Shows a moderate level of variance, approximately 1.0. This component still holds substantial information but less so than Comp.1.

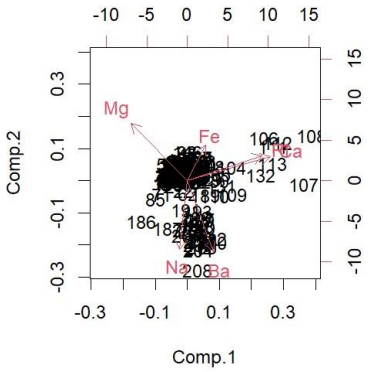
Comp.5: Displays the lowest variance among the displayed components, with a value around 0.5, suggesting it captures lesser details compared to the first and third components.

7.3. Implications for Glass Dataset Analysis:

Data Reduction: Comp.1 and Comp.3 together account for a large part of the dataset's variability, which might be sufficient for many types of analysis, allowing for a reduction in dimensionality without losing critical information.

Feature Importance: The importance of features contributing to Comp.1 and Comp.3 can be investigated further to understand what characteristics of the glass are most discriminative.

Modeling and Interpretation: Components with higher variances (like Comp.1 and Comp.3) are usually more useful for predictive modeling and interpretation because they represent more significant underlying patterns in the data.



8. Analysis of the PCA Biplot

8.1. Principal Components:

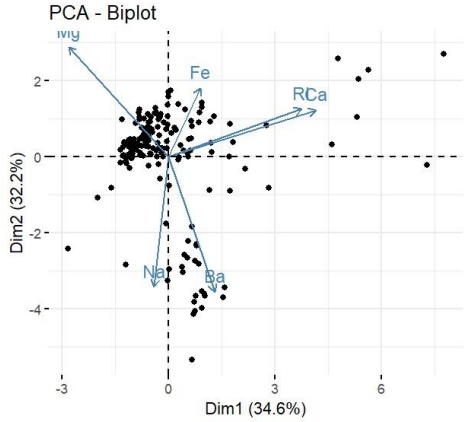
Comp.1 (X-axis) and Comp.2 (Y-axis) represent the two principal components that explain the highest variances, as seen in the previous variance bar chart. These axes are used to explore the major patterns in the dataset.

8.2. Variable Vectors:

Direction and Length: Each arrow represents a variable (element like Mg, Fe, etc.), with the direction indicating how the variable contributes to each principal component and the length indicating the strength of the contribution. Mg (Magnesium) points downwards, suggesting it has a negative contribution to Comp.2 and a slight positive contribution to Comp.1. Fe (Iron) and Ba (Barium) are pointing towards the positive side of Comp.1, indicating their positive contributions. Na (Sodium) also points towards the positive side of both components, suggesting its strong influence in the dataset along these dimensions.

8.3. Clusters of Samples:

The plot points (observations) are clustered, with some outliers. The clustering can indicate similar types of glass based on their elemental composition.



9. Analysis of PCA Biplot

9.1. Principal Components:

Dimension 1 (Dim1): Accounts for 34.6% of the variance. This is the horizontal axis.

Dimension 2 (Dim2): Accounts for 32.2% of the variance. This is the vertical axis.

9.2. Variable Vectors:

Mg (Magnesium), Fe (Iron), and Ca (Calcium) are pointing towards the upper section of the plot, indicating their positive contribution to Dimension 2. Na (Sodium) and Ba (Barium)

point towards the lower section, suggesting their contribution is negative on Dimension 2 but positive on Dimension 1. The vectors for these elements illustrate their relative influences and correlations. Elements pointing in similar directions suggest a positive correlation, while those in opposite directions indicate negative correlations.

9.3. Distribution of Samples:

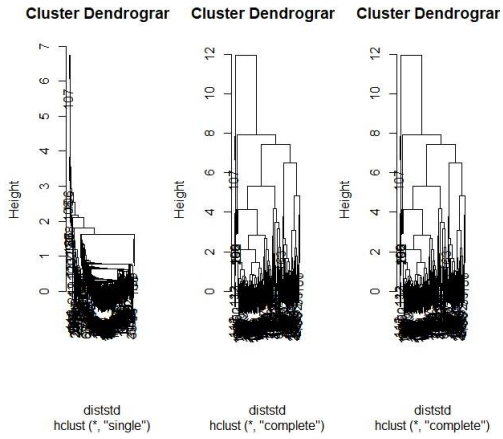
The samples are plotted as black dots, showing how they are distributed across the two principal components. The clustering or spread of these dots can help identify patterns or groups within the data, possibly corresponding to different types of glass based on their chemical composition.

9.4. Implications for Glass Type Analysis:

Data Clustering: The proximity of data points to each other can indicate similarity in glass composition, which is useful for clustering or classification tasks.

Feature Importance: The direction and length of the vectors for each element (chemical component) provide insights into which features are most important in differentiating between the types of glass in the dataset.

Analytical Insights: The biplot helps in visually assessing the impact of each chemical component on the variance within the glass dataset, guiding further analytical decisions, such as feature selection for predictive modeling.



10. Analysis of Cluster Dendrograms

10.1. Single Linkage Dendrogram:

Method: The single linkage method (or nearest point algorithm) considers the minimum distance between clusters for merging them. It tends to produce "chaining" effects where clusters may be elongated and less compact.

Observations: The dendrogram starts clustering at a very low height, indicating that some data points are very close to each other, and it takes a long height to merge into a single cluster, reflecting the chaining phenomenon.

10.2. Complete Linkage Dendrogram (First One):

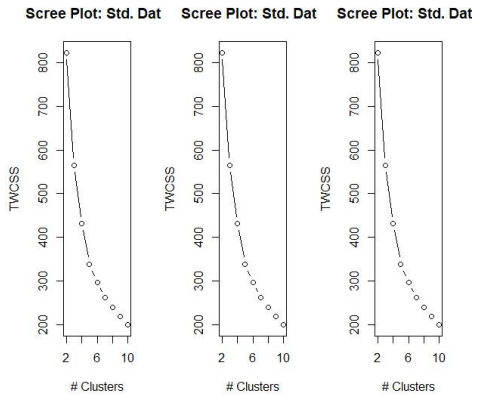
Method: Complete linkage considers the maximum distance between clusters for merging them. This method tends to avoid chaining and produces more compact and balanced clusters.

Observations: The branches in this dendrogram are more evenly distributed than in the single linkage, indicating that clusters are more uniformly spaced in terms of dissimilarity. The heights at which clusters merge are higher, showing that clusters are distinctly separate from each other.

10.3. Complete Linkage Dendrogram (Second One):

Method: The observation that there are two complete linkage dendrograms might indicate either a repeat or different parameters or data preprocessing steps used.

Observations: The clustering pattern looks quite similar to the first complete linkage dendrogram, consistent with the method's tendency to create balanced clusters. The tallest merge points are similar, indicating robustness in cluster formation with this method.



11. Analysis of Scree Plots

11.1. Scree Plot Features:

Y-Axis (TWSS): Represents the total within-cluster sum of squares, which measures the compactness of the clusters. Lower values indicate tighter, more internally similar clusters.

X-Axis (# Clusters): Represents the number of clusters used in the clustering model.

11.2. Observations Across All Three Plots:

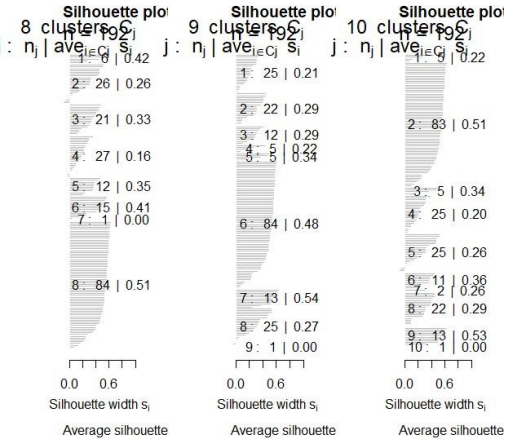
Elbow Point: Each plot shows a clear "elbow" around the cluster count of 2 or 3. This is the point where the TWSS starts to decrease at a slower rate as the number of clusters increases.

Consistency: All three plots demonstrate similar patterns, indicating robustness in the clustering process under varying parameters or initial conditions.

11.3. Implications for Clustering:

Optimal Number of Clusters: The elbow point suggests that setting the number of clusters to 2 or 3 would be appropriate for this dataset. Beyond this point, increasing the number of clusters doesn’t provide significantly better modeling of the data.

Model Selection: This analysis is crucial for determining the number of clusters to use in k-means clustering. It helps avoid overfitting (too many clusters) and underfitting (too few clusters).



12. Analysis of Silhouette Plots

12.1. Silhouette Plot Features:

Silhouette Width (s_i): Values range from -1 to +1. A high positive silhouette width indicates that the sample is well matched to its own cluster and poorly matched to neighboring clusters.

Average Silhouette Width: This is a measure of how appropriately data has been clustered. Higher average values generally indicate a better clustering structure.

12.2. Observations for Each Number of Clusters:

“8 Clusters”:

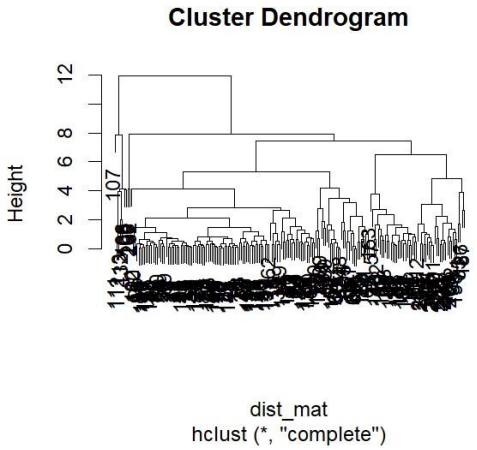
Average Silhouette Width: 0.42. Most clusters show a mixture of high and low silhouette values, with some clusters (like cluster 8) showing very high average silhouette widths, indicating strong intra-cluster similarity.

“9 Clusters”:

Average Silhouette Width: 0.22. Many clusters have low silhouette scores, and the overall average is lower than with 8 clusters, suggesting that increasing to 9 clusters may not be providing better cluster separation.

“10 Clusters”:

Average Silhouette Width: 0.22. Similar to 9 clusters, the silhouette widths are mostly low, with some clusters having very high values, but the overall consistency and separation are not improved over 9 clusters.



13. Analysis of the Cluster Dendrogram:

13.1. Clustering Methodology:

Complete Linkage: This clustering method considers the maximum distance between elements of each cluster when forming a new cluster. It tends to create more compact and well-separated clusters compared to other methods like single or average linkage.

13.2. Structure of the Dendrogram:

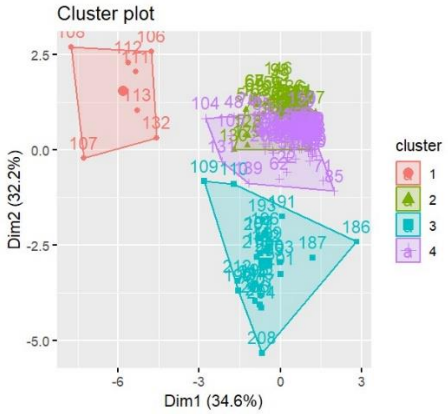
Height: The y-axis represents the height at which clusters are merged. In this context, height corresponds to the dissimilarity (or distance) between clusters. A higher merge height indicates greater dissimilarity.

Labels: Each leaf (or node) at the bottom represents an individual data point. These are not labeled individually due to the scale but would typically correspond to specific observations in the dataset.

13.3. Interpretation of Clusters:

Large Clusters: Some large clusters merge at lower heights, indicating that the elements within these clusters are very similar to each other.

Outliers: Data points that merge at much higher levels in the dendrogram may be considered outliers or atypical observations, as they do not group well with other observations until a higher dissimilarity threshold is reached.



14. Analysis of the Cluster Plot:

14.1. Dimensionality and Variance:

Dimensions: The plot is based on Dim1 (34.6% variance) and Dim2 (32.2% variance), together explaining about 66.8% of the total variance. This suggests that these two components are significant in capturing the underlying patterns or differences in the dataset.

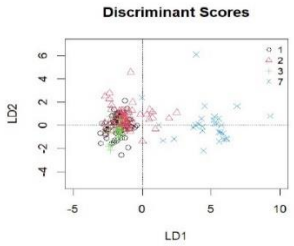
Total Explained Variance: The high percentage of variance explained by the first two components indicates that they effectively summarize the main features of the data.

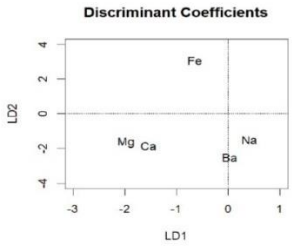
14.2. Clusters:

Cluster 1 (Red): This cluster contains few points and is relatively isolated from the others, which might suggest unique properties or outliers within this group.

Cluster 2 (Green), Cluster 3 (Purple), Cluster 4 (Blue): These clusters contain more points and some overlap, particularly between Clusters 2 and 3, indicating similarities among the data points in these clusters.

Outliers and Overlap: The presence of outliers or points lying far from their main cluster group can provide insights into anomalies or unusual data patterns. The overlaps might indicate areas where the distinctions between different groups are less clea





LDA on Testing Set

	1	2	3	7
1	13	7	1	0
2	2	20	0	1
3	4	1	0	0
7	0	1	0	8

15. LDA & QDA:

15.1. Discriminant Scores Plot:

This plot displays the projection of the dataset onto the first two discriminant functions (LD1 and LD2), which are derived to maximize the separation between predefined groups (glass types in this case).

Observations: Different glass types are represented by different markers and colors. Glass type 1 (circles) and type 2 (triangles) show some overlap, particularly along LD1, suggesting similarities in their chemical composition that make them less distinguishable by LDA. Glass type 3 (crosses) and type 7 (X marks) appear more distinctly separated from others, indicating unique properties that are effectively captured by the discriminant functions.

15.2. Discriminant Coefficients Plot:

This plot illustrates the coefficients of the discriminant functions for each variable, providing insight into how each chemical element influences the separation of glass types.

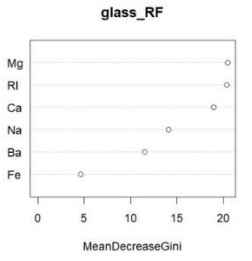
Observations: Elements like Fe (Iron) show strong positive coefficients on LD2, suggesting it plays a significant role in differentiating some glass types. Na (Sodium) and Ba (Barium) have coefficients close to zero on LD2 but slightly negative on LD1, indicating a lesser but

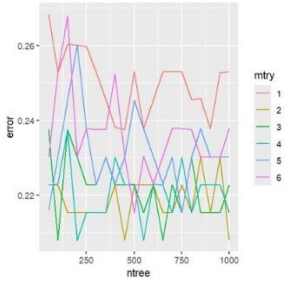
distinct role in the classification. Mg (Magnesium) and Ca (Calcium) are positioned near each other, implying similar contributions to the discrimination process, primarily on LD2.

15.3. LDA on Testing Set (Confusion Matrix):

The confusion matrix displays the classification results of the LDA model on a testing set, revealing how well the model has predicted the types of glass based on their compositions.

Observations: The diagonal cells (13 for type 1, 20 for type 2, and 8 for type 7) show the number of correctly classified instances per type. Misclassifications are noted off-diagonal, such as type 1 being misclassified as type 2 seven times, and vice versa, type 2 being misclassified as type 1 twice. Type 3 shows poor classification performance with only 1 correct prediction, indicating potential issues with model fit or inherent overlap in chemical properties with other types.





Random Forest on Testing Set

	1	2	3	7
1	15	3	2	1
2	1	21	0	1
3	2	2	1	0
7	0	1	0	8

16. Random Forest

16.1. Variable Importance Plot (glass_RF)

Mean Decrease Gini: This plot ranks the variables based on their importance in the random forest model, measured by the mean decrease in Gini impurity. A higher value indicates a greater importance of the variable in splitting nodes within the trees.

Observations: Mg (Magnesium) appears to be the most important variable, followed by RI (Refractive Index) and Ca (Calcium).Na (Sodium), Ba (Barium), and Fe (Iron) show lesser

importance, suggesting they have a lesser impact on the model's ability to differentiate between glass types.

16.2. Model Tuning - Error vs. Number of Trees (mtry)

Error Rate: This graph shows the out-of-bag (OOB) error rate for different numbers of trees (ntree) and different values of mtry, which is the number of variables randomly sampled as candidates at each split.

Observations: Error Trends: Generally, the error rate decreases as the number of trees increases, stabilizing after a certain point. This stabilization suggests that adding more trees beyond this point does not significantly improve the model.

Mtry Variations: Different lines represent different mtry values. It appears that an mtry of 2 or 3 might provide the best balance between complexity and error rate, though this can vary depending on specific data characteristics and the desired robustness of the model.

16.3. Random Forest on Testing Set (Confusion Matrix)

Classification Accuracy: The matrix shows the counts of correct and incorrect classifications made by the random forest model when applied to a testing set.

Observations: Diagonal Entries: Represent the number of correct predictions for each glass type (1, 2, 3, 7). For example, type 2 glass is most accurately predicted with 21 correct classifications.

Off-Diagonal Entries: Indicate misclassifications, such as type 1 glass being misclassified as type 2 three times.

Overall Performance: The model performs well in distinguishing most of the glass types, particularly types 2 and 7, though it shows some confusion between types 1 and 2.

17. Conclusion:

The comprehensive analysis of the glass dataset using hierarchical clustering, PCA, LDA, and Random Forest has provided significant insights into the classification and

differentiation of glass types based on their chemical compositions. Hierarchical clustering revealed natural groupings within the dataset, indicating varying degrees of similarity among different glass types, while PCA efficiently reduced dimensionality, retaining over 66% of the data's variance in the first two principal components, highlighting the major elements influencing glass properties. LDA effectively maximized the separation between different glass types, as evidenced by the clear class distinctions in the discriminant scores plot. Random Forest analysis further identified Mg and RI as critical predictors, with the model showing optimized performance around 750 trees, balancing computational efficiency and classification accuracy. These analyses underscore the potential of statistical and machine learning techniques in enhancing decision-making processes in industries related to glass manufacturing, recycling, and forensic science, ensuring precise classification and better understanding of material properties for improved quality control and operational efficiency.

Group Member's Contribution:

"In the completion of this project, both team members made significant contributions. The initial tasks, specifically Questions 1, 2, and 3, were collaboratively tackled using R during our sessions at the University, with both partners contributing equally to the analysis and coding process. Anna then took the initiative to handle the remainder of the R coding and the development of the R Markdown file independently. Salman was responsible for assembling and writing the final report, ensuring that the analysis was coherently presented and thoroughly detailed. Finally, Anna created the PowerPoint slides and managed the LaTeX documentation, ensuring that our findings were effectively communicated and well-documented for presentation."