DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory_Data_Analysis

Laboratory 5-8

Overview

This report analyzes the students_employability.csv dataset (2,982 undergraduate students from the Philippines) using exploratory data analysis, PCA, and clustering techniques to support decision-making on student employability. The analysis pipeline includes data loading & cleaning, descriptive statistics, visualization, correlation analysis, PCA, hierarchical and k-means clustering, and final interpretation & recommendations.

```r
students <- read.csv("students_employability.csv")
head(students)
```

```
##   Name.of.Student GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION
## 1       Student 1                  4                  5                  4
## 2       Student 2                  4                  4                  4
## 3       Student 3                  4                  3                  3
## 4       Student 4                  3                  3                  3
## 5       Student 5                  4                  4                  3
## 6       Student 6                  4                  4                  3
##   MENTAL.ALERTNESS SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS
## 1                5               5                        5
## 2                4               4                        4
## 3                3               3                        3
## 4                2               3                        3
## 5                3               4                        4
## 6                3               3                        3
##   COMMUNICATION.SKILLS Student.Performance.Rating          CLASS
## 1                    5                          5     Employable
## 2                    3                          5     Employable
## 3                    2                          5 LessEmployable
## 4                    3                          5 LessEmployable
## 5                    3                          5     Employable
## 6                    3                          5     Employable
```

Data exploration

Dimension

```
dim(students)      # Number of rows and columns
```

```
## [1] 2982   10
```

Variables and Types

Variable	Type	Description
Name.of.Student	character	Name of the student
GENERAL.APPEARANCE	integer	Score for general appearance in the interview
MANNER.OF.SPEAKING	integer	Score for how the student speaks
PHYSICAL.CONDITION	integer	Score for physical fitness/presentation
MENTAL.ALERTNESS	integer	Score for alertness and focus
SELF.CONFIDENCE	integer	Confidence score in the interview
ABILITY.TO.PRESENT.IDEAS	integer	Ability to convey ideas clearly
COMMUNICATION.SKILLS	integer	Overall communication skills
Student.Performance.Rating	integer	Overall performance rating in the interview
CLASS	factor/character	Employability class (e.g., Employable, LessEmployable)

- The dataset has **2,982 rows** (students) and **10 columns**.
- Most attributes are numeric scores (integer) from mock interviews.
- `CLASS` is the categorical target representing employability.

Dataset summary statistics

```
summary(students)  # Summary statistics
```

```
##  Name.of.Student    GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION
##  Length:2982        Min.   :2.000      Min.   :2.000      Min.   :2.000
##  Class :character   1st Qu.:4.000      1st Qu.:3.000      1st Qu.:3.000
##  Mode  :character   Median :4.000      Median :4.000      Median :4.000
##                     Mean   :4.247      Mean   :3.885      Mean   :3.972
##                     3rd Qu.:5.000      3rd Qu.:4.000      3rd Qu.:5.000
##                     Max.   :5.000      Max.   :5.000      Max.   :5.000
##  MENTAL.ALERTNESS SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS COMMUNICATION.SKILLS
##  Min.   :2.000    Min.   :2.000   Min.   :2.000            Min.   :2.000
##  1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.000            1st Qu.:3.000
##  Median :4.000    Median :4.000   Median :4.000            Median :3.000
##  Mean   :3.963    Mean   :3.911   Mean   :3.814            Mean   :3.525
##  3rd Qu.:5.000    3rd Qu.:5.000   3rd Qu.:4.000            3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.000   Max.   :5.000            Max.   :5.000
##  Student.Performance.Rating    CLASS
##  Min.   :3.000              Length:2982
##  1st Qu.:4.000              Class :character
##  Median :5.000              Mode  :character
##  Mean   :4.611
##  3rd Qu.:5.000
##  Max.   :5.000
```

Descriptive Statistics by Employability Class

```
interview_attr <- c("GENERAL.APPEARANCE", "MANNER.OF.SPEAKING", "PHYSICAL.CONDITION",
         "MENTAL.ALERTNESS", "SELF.CONFIDENCE", "ABILITY.TO.PRESENT.IDEAS",
         "COMMUNICATION.SKILLS", "Student.Performance.Rating")

numeric_df <- students[,interview_attr]
by(numeric_df, students$CLASS, summary)
```

```
## students$CLASS: Employable
##  GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION MENTAL.ALERTNESS
##  Min.   :3.000      Min.   :3.000      Min.   :2.000      Min.   :2.000
##  1st Qu.:4.000      1st Qu.:3.000      1st Qu.:4.000      1st Qu.:4.000
##  Median :4.000      Median :4.000      Median :4.000      Median :4.000
##  Mean   :4.314      Mean   :4.012      Mean   :4.076      Mean   :4.098
##  3rd Qu.:5.000      3rd Qu.:5.000      3rd Qu.:5.000      3rd Qu.:5.000
##  Max.   :5.000      Max.   :5.000      Max.   :5.000      Max.   :5.000
##  SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS COMMUNICATION.SKILLS
##  Min.   :2.000   Min.   :2.000            Min.   :2.00
##  1st Qu.:3.000   1st Qu.:3.000            1st Qu.:3.00
##  Median :4.000   Median :4.000            Median :4.00
##  Mean   :4.008   Mean   :3.888            Mean   :3.61
##  3rd Qu.:5.000   3rd Qu.:4.000            3rd Qu.:4.00
##  Max.   :5.000   Max.   :5.000            Max.   :5.00
##  Student.Performance.Rating
##  Min.   :3.00
##  1st Qu.:4.00
##  Median :5.00
##  Mean   :4.61
##  3rd Qu.:5.00
##  Max.   :5.00
## ------------------------------------------------------------
## students$CLASS: LessEmployable
##  GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION MENTAL.ALERTNESS
##  Min.   :2.000      Min.   :2.000      Min.   :2.000      Min.   :2.000
##  1st Qu.:4.000      1st Qu.:3.000      1st Qu.:3.000      1st Qu.:3.000
##  Median :4.000      Median :4.000      Median :4.000      Median :4.000
##  Mean   :4.154      Mean   :3.709      Mean   :3.828      Mean   :3.777
##  3rd Qu.:5.000      3rd Qu.:4.000      3rd Qu.:4.000      3rd Qu.:4.000
##  Max.   :5.000      Max.   :5.000      Max.   :5.000      Max.   :5.000
##  SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS COMMUNICATION.SKILLS
##  Min.   :2.000   Min.   :2.000            Min.   :2.000
##  1st Qu.:3.000   1st Qu.:3.000            1st Qu.:3.000
##  Median :4.000   Median :4.000            Median :3.000
##  Mean   :3.777   Mean   :3.712            Mean   :3.409
##  3rd Qu.:4.000   3rd Qu.:4.000            3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000            Max.   :5.000
##  Student.Performance.Rating
##  Min.   :3.000
##  1st Qu.:4.000
##  Median :5.000
##  Mean   :4.611
##  3rd Qu.:5.000
##  Max.   :5.000
```

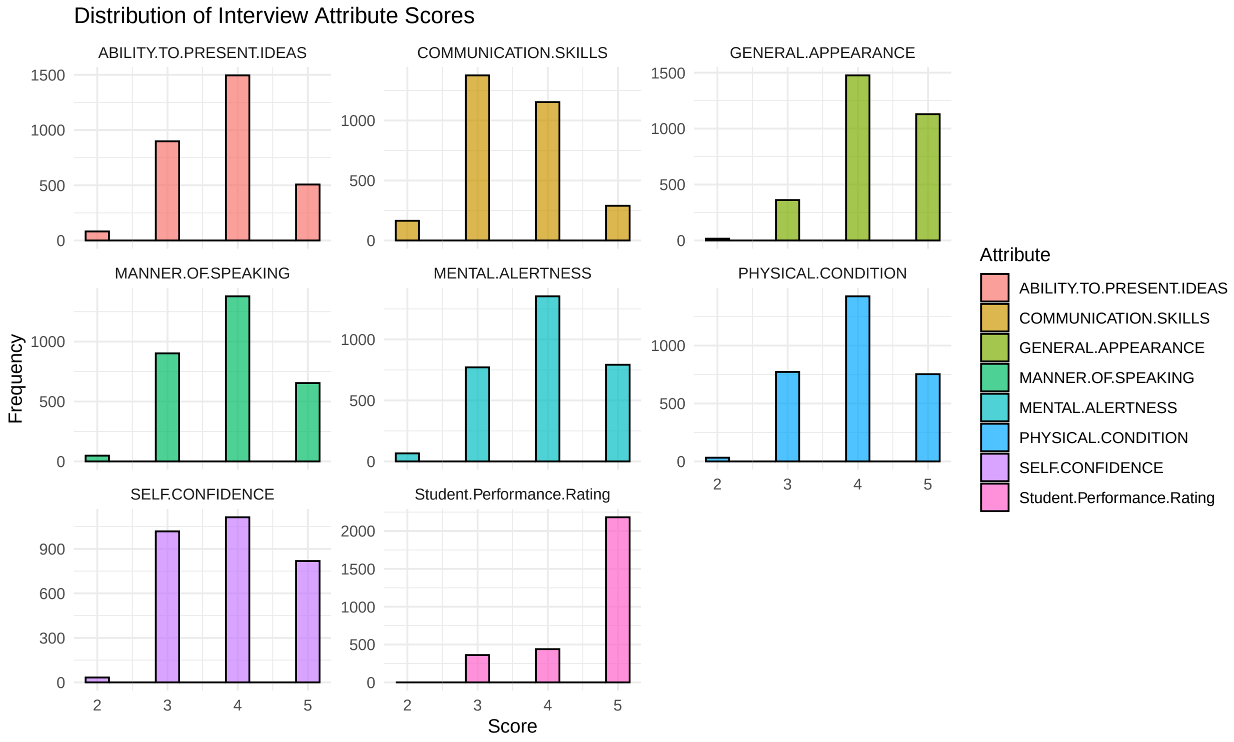
Observations

- `Employable` students generally score higher than the overall average across all attributes.
- `LessEmployable` students score slightly lower, particularly in `MANNER.OF.SPEAKING, SELF.CONFIDENCE, ABILITY.TO.PRESENT.IDEAS, and COMMUNICATION.SKILLS`.
- **Median** values are generally aligned with means, showing balanced distributions without extreme skew.

- This comparison helps identify the attributes most relevant for employability decisions, guiding later analyses like **PCA** or **clustering**.

Graphical Exploration

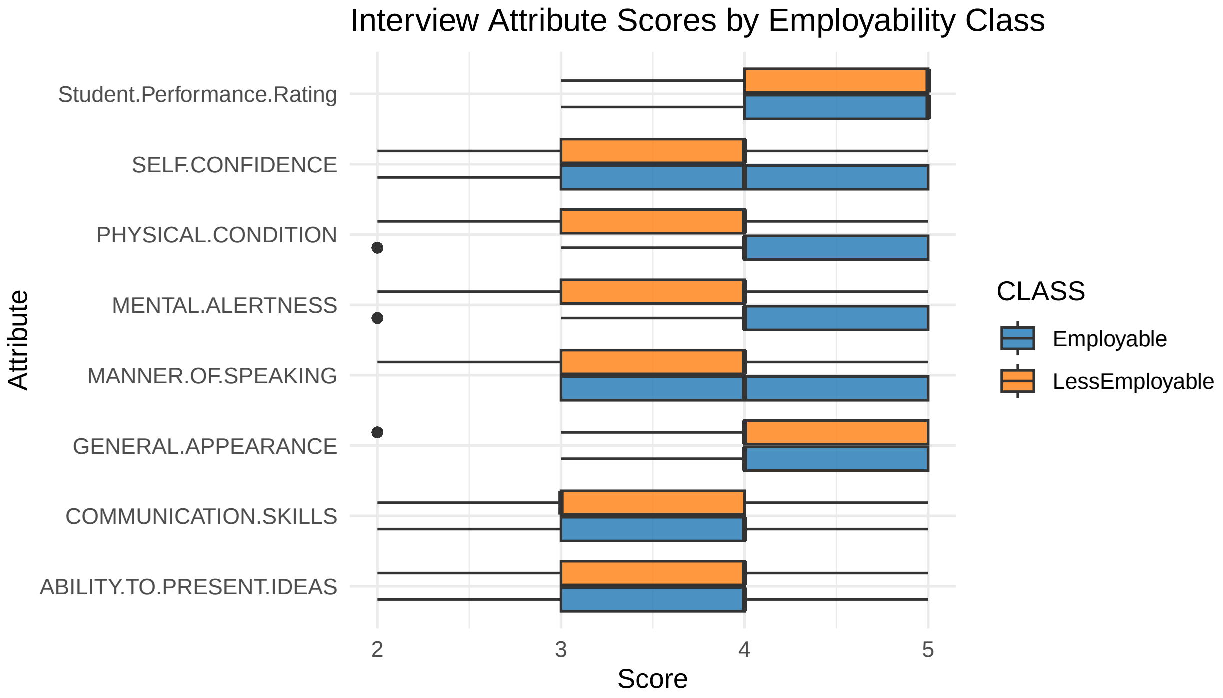
Histogram



Interpretation:

- Most attributes display approximately normal or slightly left-skewed distributions, with score frequencies peaking around 3 to 4. This pattern indicates that the majority of students perform at an average to above-average level across all assessed skills. Only a small number of students receive very low scores (1–2), suggesting that poor performance is relatively rare in this sample. Attributes such as Communication Skills, General Appearance, and Self Confidence show higher concentrations at the upper end (scores 4–5), implying that these traits are more consistently developed among the students. In contrast, attributes like Ability to Present Ideas and Mental Alertness show slightly wider variability, hinting that these areas may differentiate stronger candidates from weaker ones. Overall, the histograms suggest a generally competent student population, with minor gaps in expressive and cognitive-related attributes.

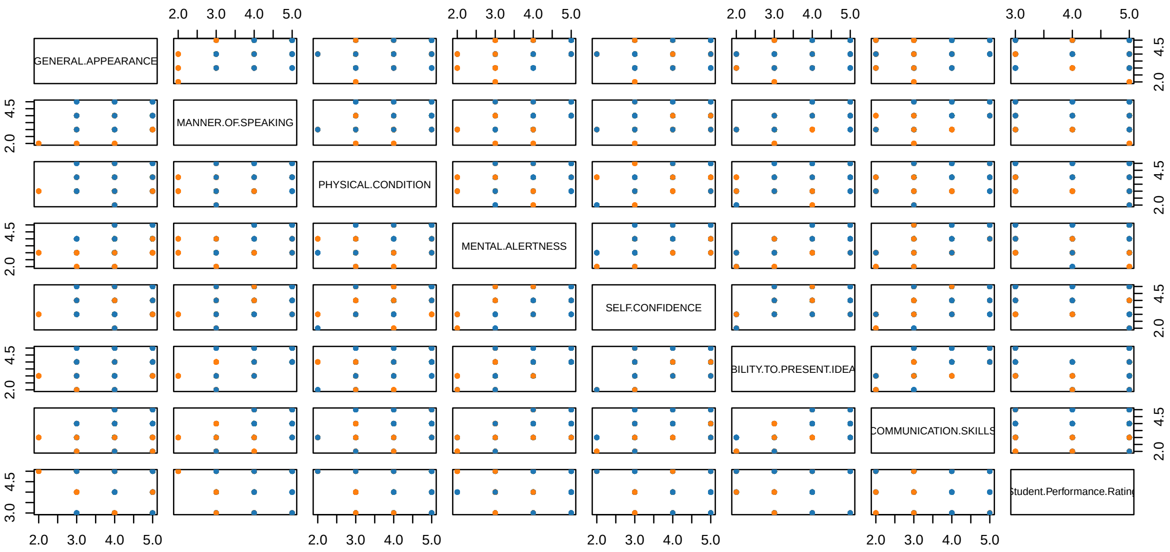
Boxplots by Employability Class



Interpretation:

- Across nearly all interview attributes, Employable students (blue) exhibit higher median scores and narrower interquartile ranges, indicating more consistent and stronger performance than their Less Employable (orange) counterparts. The most prominent gaps are seen in Communication Skills, Self Confidence, and Ability to Present Ideas, where the Employable group's medians are noticeably higher. These attributes strongly relate to how effectively a student conveys ideas, interacts with interviewers, and projects confidence — all crucial indicators of professional readiness. Meanwhile, attributes like General Appearance and Physical Condition show smaller differences between the two classes, suggesting that visual or physical presentation alone does not decisively determine employability. This visualization reinforces that interpersonal, expressive, and confidence-based traits are the most discriminative factors separating employable students from less employable ones.

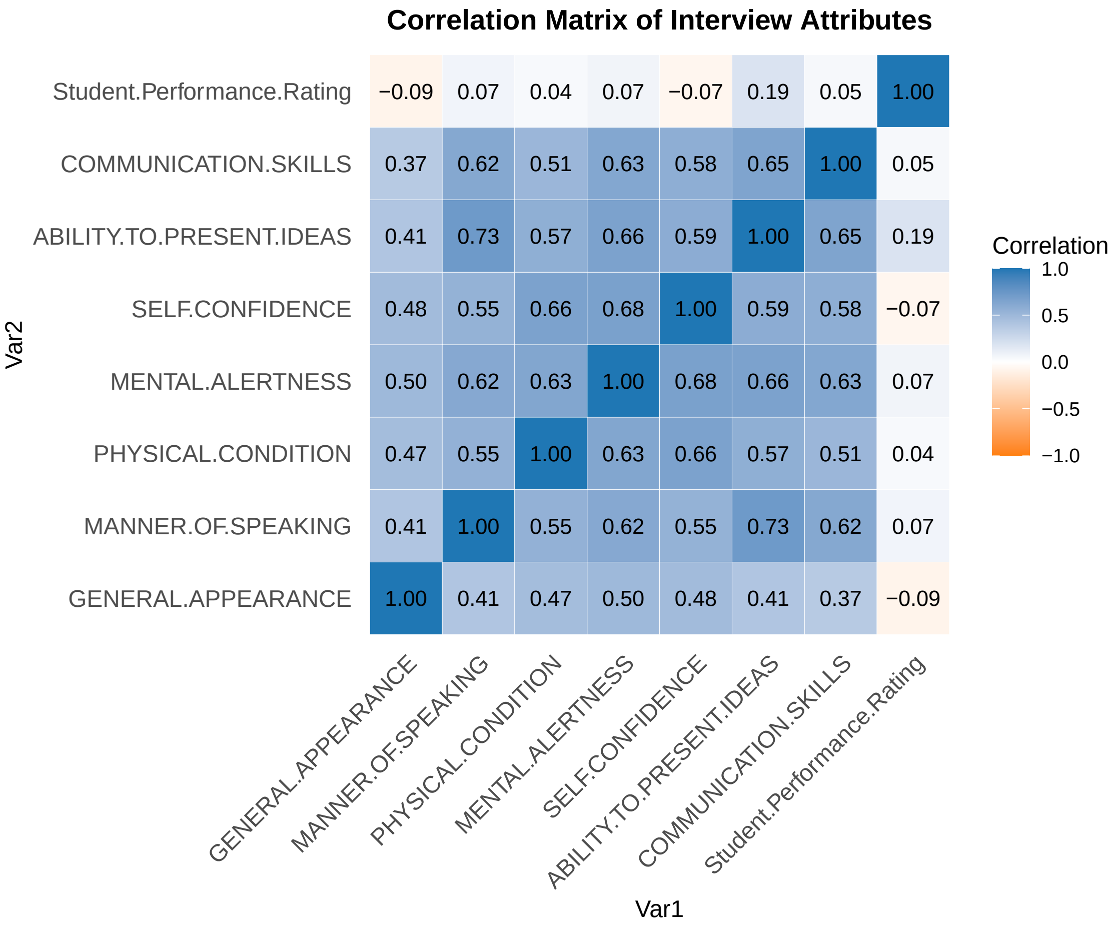
Pairs Plot of Interview Attributes



Interpretation:

- The pairs plot illustrates pairwise relationships among all interview attributes and the overall performance rating. The general pattern shows that most variable pairs are positively associated — students who perform well in one area tend to score high in others. For example, Self Confidence, Communication Skills, and Ability to Present Ideas display visibly aligned positive relationships, implying that confident students also communicate ideas more effectively. Similarly, Mental Alertness and Manner of Speaking show synergy, as alert and focused students tend to articulate themselves more clearly. Moreover, the Blue points (Employable students) tend to cluster toward the upper-right regions of each subplot, indicating consistently high attribute scores across the board. In contrast, orange points (Less Employable students) are dispersed in lower ranges, confirming that their performance lags in multiple dimensions simultaneously. This suggests that employability is not determined by a single standout skill, but rather by a balanced and well-rounded performance across all key attributes such as communication, confidence, and presentation ability.

Correlation Analysis



Correlation Matrix of Interview Attributes

Observation

The correlation matrix reveals generally moderate to strong positive relationships among most interview attributes. For instance, Ability to Present Ideas shows strong correlations with both **Communication Skills (r = 0.73)** and **Self Confidence (r = 0.66)**, suggesting that students who are confident communicators are also more capable of effectively presenting their ideas.

Similarly, **Mental Alertness** correlates positively with **Self Confidence (r = 0.68)** and **Physical Condition (r = 0.63)**, indicating that alert and engaged students tend to maintain strong self-assurance and professional composure during interviews.

However, **Student Performance Rating** exhibits relatively weak correlations (r values around 0.05 to 0.19) with other attributes, implying that this performance rating may not directly mirror the interview sub-scores or might capture additional factors beyond these attributes.

Overall, the correlation analysis highlights a clear interdependence among the communication-related and confidence-based traits, which collectively reflect a student's readiness and professionalism—key indicators of employability.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed on the **scaled interview attributes and performance rating** to uncover underlying dimensions in the employability dataset. The analysis extracted eight principal components (PCs), corresponding to the eight numeric variables included.

```r
pca_model <- prcomp(numeric_df, center = TRUE, scale. = TRUE)
pca_model$rotation # View loadings (contribution of each attribute to each PC)
```

```
##                                  PC1          PC2         PC3         PC4
## GENERAL.APPEARANCE         0.2986819 -0.314045198  0.72830463  0.50964113
## MANNER.OF.SPEAKING         0.3862490  0.105706682 -0.29622907  0.35546014
## PHYSICAL.CONDITION         0.3775706 -0.062675424  0.20807761 -0.56242960
## MENTAL.ALERTNESS           0.4071354  0.009349947  0.04114044 -0.12853044
## SELF.CONFIDENCE            0.3889570 -0.184815817  0.02488112 -0.44824896
## ABILITY.TO.PRESENT.IDEAS   0.3992139  0.227313926 -0.18513495  0.21049049
## COMMUNICATION.SKILLS       0.3761373  0.063247507 -0.39144391  0.17950916
## Student.Performance.Rating 0.0338160  0.892388551  0.38562875 -0.06008296
##                                    PC5          PC6         PC7          PC8
## GENERAL.APPEARANCE          0.08297897 -0.088977258 -0.08472807 -0.004658197
## MANNER.OF.SPEAKING         -0.54312835  0.007892364 -0.01023179  0.574785934
## PHYSICAL.CONDITION         -0.35549684 -0.556963239  0.21175288 -0.111640127
## MENTAL.ALERTNESS            0.25807111  0.505942445  0.69841172  0.074721935
## SELF.CONFIDENCE             0.16519456  0.354234045 -0.64752828  0.202349957
## ABILITY.TO.PRESENT.IDEAS   -0.23656885  0.204830256 -0.16524735 -0.764950013
## COMMUNICATION.SKILLS        0.63926630 -0.507955105 -0.03582171  0.033251257
## Student.Performance.Rating  0.11503009 -0.019477412 -0.11052372  0.156054072
```

```r
summary(pca_model) # View summary to see explained variance
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1026 1.0506 0.81477 0.72649 0.63718 0.57334 0.54458
## Proportion of Variance 0.5526 0.1380 0.08298 0.06597 0.05075 0.04109 0.03707
## Cumulative Proportion  0.5526 0.6906 0.77358 0.83955 0.89030 0.93139 0.96846
##                            PC8
## Standard deviation     0.50231
## Proportion of Variance 0.03154
## Cumulative Proportion  1.00000
```

The **first principal component (PC1)** alone explains **55.3%** of the total variance in the dataset, while the **first two components (PC1 + PC2)** together explain about **69.1%** of the overall variance. This indicates that most of the variation across students' interview performance can be summarized by just **two underlying dimensions**, making PCA an effective tool for dimensionality reduction and visualization.

Interpretation of the Loadings

PC1 — "General Interview Competence"

- High positive loadings:
  - Mental Alertness (0.41)
  - Ability to Present Ideas (0.40)
  - Self Confidence (0.39)
  - Communication Skills (0.38)
  - Physical Condition (0.38)
  - Manner of Speaking (0.39)
  - General Appearance (0.30)

Interpretation: PC1 represents a composite measure of overall interview competence, combining communication, self-confidence, attentiveness, and physical presentation. Students scoring high on PC1 tend to perform well across nearly all attributes, reflecting strong overall employability potential.

PC2 — "Performance-Specific Dimension"

- High positive loadings:
  - Strong positive loading: Student Performance Rating (0.89)
  - Moderate negative loading: General Appearance (-0.31) and Self Confidence (-0.18)

Interpretation: PC2 primarily captures variation in the overall performance rating independent of general appearance or self-confidence. This component isolates how well the interviewer's overall rating aligns (or diverges) from sub-skill scores — possibly reflecting subjective evaluation or overall impression.

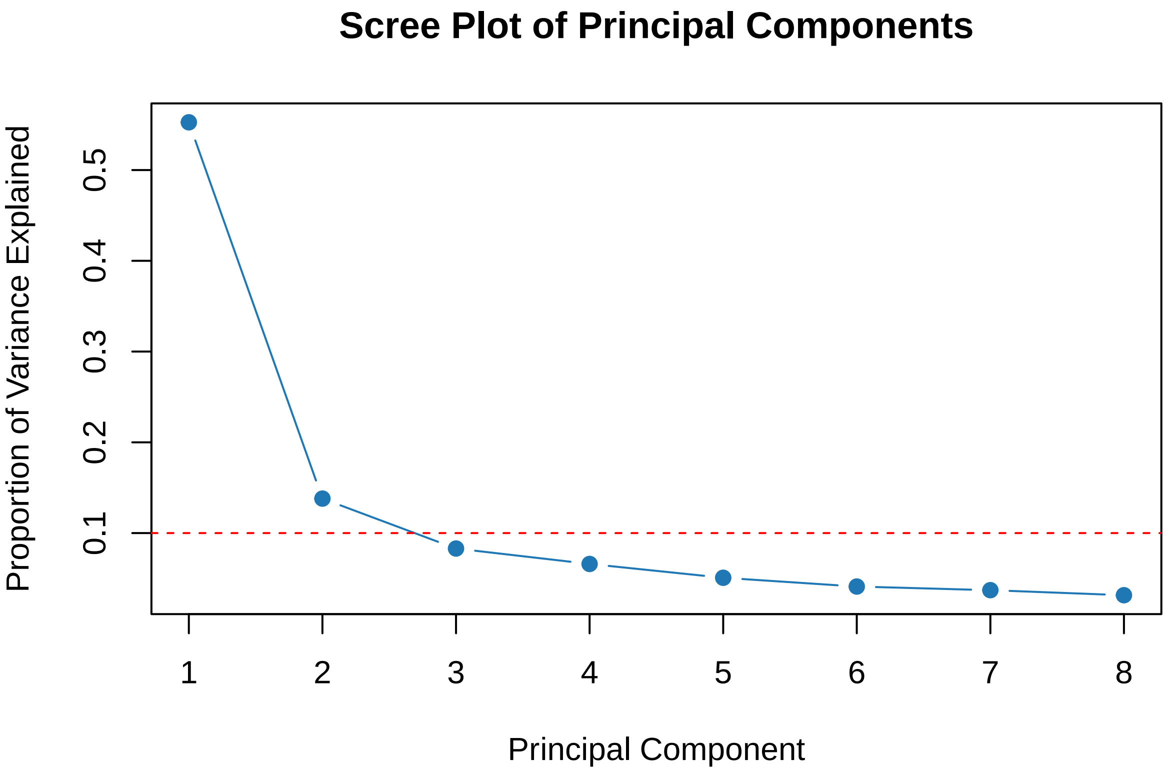
PC3–PC4 — Minor Components

- High positive loadings:
  - PC3 is influenced by General Appearance (0.73) and Student Performance (0.39) — suggesting a minor "appearance/presentation" factor.
  - PC4 loads on Physical Condition (-0.56) and Self Confidence (-0.45) — reflecting a subtle "poise and composure" dimension. However, their individual variance contributions (8% and 6.6%) are comparatively small.

Summary of Findings

- The dataset's variance is highly concentrated in the first principal component (55%), meaning students' interview attributes tend to rise and fall together — those who score well in one area often perform well across others.
- The second component (14%) represents variation primarily driven by the overall performance rating, distinguishing it from specific skill scores.
- Together, PC1 and PC2 effectively summarize most of the variability in employability characteristics, supporting the use of these components for visualization and cluster exploration in later analysis.

Scree Plot(Eigenvalues)



Interpretation of the Scree Plot

The scree plot above visualizes the proportion of total variance explained by each principal component (PC). The steep decline from the first to the second component indicates that most of the data's variability is captured by the first few components, while the remaining ones contribute relatively little additional information.

Specifically, PC1 accounts for approximately 55% of the total variance, and PC2 contributes an additional 14%, bringing the cumulative variance explained to about 69%. This sharp "elbow" at the second component suggests that a two-component solution provides a good balance between simplification and information retention. Beyond PC2, the plot flattens noticeably, meaning that components PC3 through PC8 each explain less than 10% of the variance and mostly capture residual noise or minor patterns.

Therefore, based on the scree plot and the Kaiser criterion (eigenvalues > 1), it is reasonable to retain two principal components for interpretation and visualization. These components effectively summarize the underlying structure of the dataset:

- **PC1 represents the dominant "General Interview Competence" dimension that aggregates performance across multiple attributes.**
- **PC2 represents the "Overall Performance Rating" dimension, capturing assessor-level variation.**

These two components explain nearly 70% of total variance, confirming that they sufficiently describe the main patterns of employability characteristics among students.

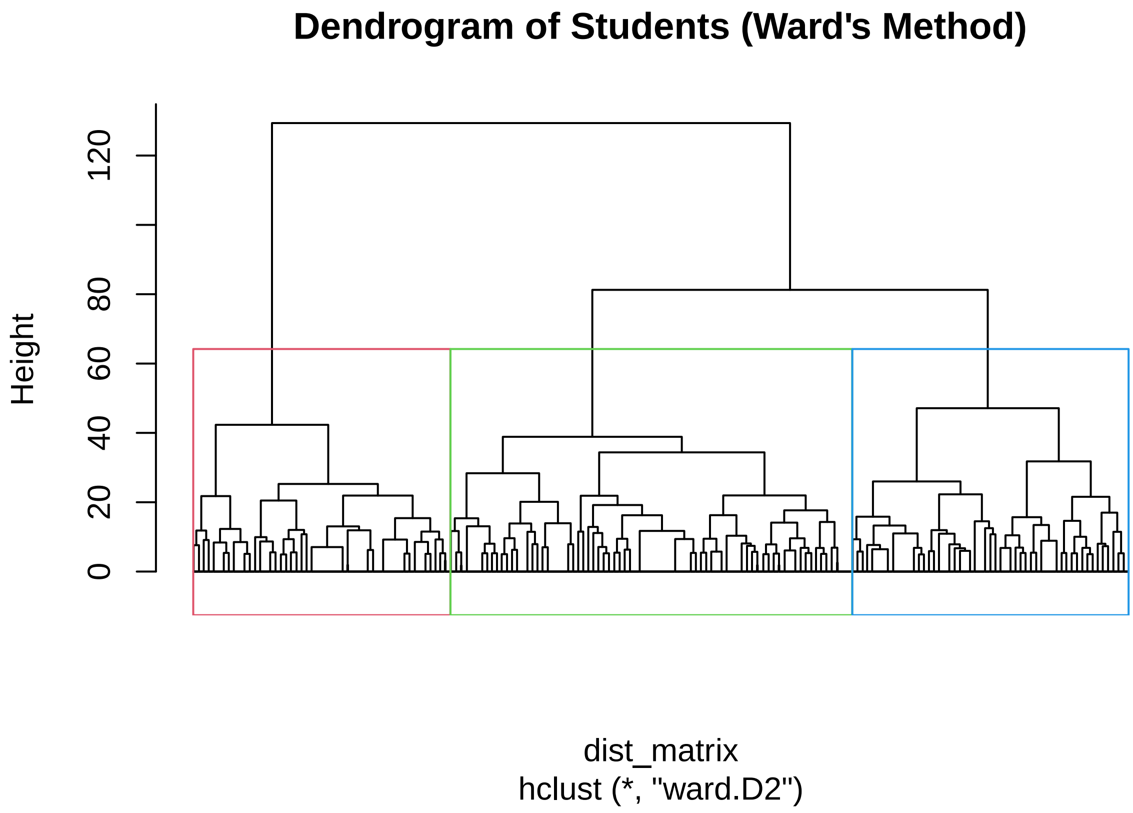
Cluster Analysis

Hierarchical Clustering

```
scaled_df <- scale(numeric_df)  # Scale numeric attributes
dist_matrix <- dist(scaled_df, method = "euclidean")

hc_ward <- hclust(dist_matrix, method = "ward.D2")

plot(hc_ward, labels = FALSE, main = "Dendrogram of Students (Ward's Method)", hang = -1)
rect.hclust(hc_ward, k = 3, border = 2:(3+1))
```



```
# Cut the tree into 3 clusters
clusters <- cutree(hc_ward, k = 3)

students$Cluster <- clusters
table(students$Cluster)
```

```
##
##    1    2    3
##  820 1281  881
```

```
table(Cluster = students$Cluster, EmployabilityClass = students$CLASS)
```

```
##        EmployabilityClass
## Cluster Employable LessEmployable
##       1        627            193
##       2        661            620
##       3        441            440
```

```r
cluster_summary <- aggregate(numeric_df, by = list(Cluster = clusters), mean)
round(cluster_summary, 2)
```

```
##   Cluster GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION
## 1       1               4.86               4.64               4.74
## 2       2               4.12               3.98               3.92
## 3       3               3.86               3.04               3.33
##   MENTAL.ALERTNESS SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS
## 1             4.86            4.80                     4.50
## 2             3.87            3.74                     3.99
## 3             3.26            3.33                     2.92
##   COMMUNICATION.SKILLS Student.Performance.Rating
## 1                 4.31                       4.58
## 2                 3.47                       4.80
## 3                 2.87                       4.37
```

Interpretation of clusters

Cluster 1 consists primarily of students with high scores across all dimensions, particularly in Communication Skills, Ability to Present Ideas, and Self Confidence. This cluster largely corresponds to the Employable class, representing strong overall performance and well-rounded soft skills.

Cluster 2 contains students with moderate scores across attributes. They show adequate Physical Condition and Mental Alertness but slightly weaker Communication and Presentation abilities. This group may represent students with potential but inconsistent performance—borderline employable depending on context.

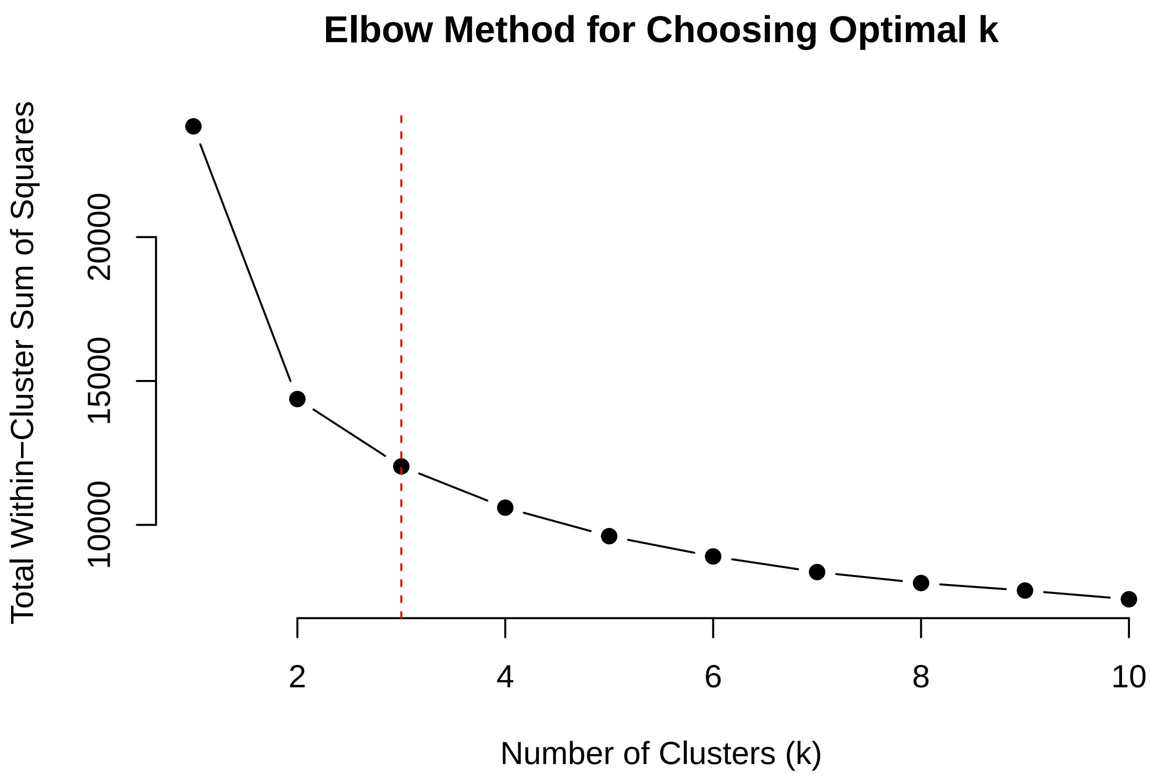
Cluster 3 includes students with lower scores on most attributes, especially Self Confidence and Ability to Present Ideas. This group is dominated by Less Employable students, suggesting limited readiness in key interpersonal and presentation skills.

The dendrogram and cluster summaries reveal a clear stratification among students: those who communicate confidently and present ideas effectively tend to cluster together and align strongly with the Employable category. Meanwhile, weaker communication and confidence scores form a distinct cluster that correlates with lower employability ratings.

K-Means Clustering

```r
set.seed(42)
wss <- sapply(1:10, function(k){
  kmeans(scale(numeric_df), centers = k, nstart = 20)$tot.withinss
})

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster Sum of Squares",
     main = "Elbow Method for Choosing Optimal k")
abline(v = 3, col = "red", lty = 2)
```



The plot shows that there is a sharp decline in WSS from k = 1 to k = 3, indicating that increasing the number of clusters significantly improves the compactness of the clusters up to that point. After k = 3, the curve starts to flatten, meaning that adding more clusters results in only a small reduction in WSS.

This bend or "elbow" at k = 3 marks the point of diminishing returns — where increasing the number of clusters no longer provides substantial gains in explaining the data's variation.

Therefore, **k = 3** is chosen as the optimal number of clusters, as it balances model simplicity and effective partitioning of the data.

```
set.seed(42)
kmeans_model <- kmeans(scale(numeric_df), centers = 3, nstart = 20)
students$KCluster <- kmeans_model$cluster

kmeans_model$size # View cluster sizes
```

```
## [1]  885  933 1164
```

```
round(kmeans_model$centers, 2) # View cluster centroids
```

```
##   GENERAL.APPEARANCE MANNER.OF.SPEAKING PHYSICAL.CONDITION MENTAL.ALERTNESS
## 1               0.84               0.96               1.01             1.14
## 2              -0.56              -1.05              -0.86            -0.92
## 3              -0.18               0.11              -0.08            -0.13
##   SELF.CONFIDENCE ABILITY.TO.PRESENT.IDEAS COMMUNICATION.SKILLS
## 1            1.07                     0.90                 1.00
## 2           -0.74                    -1.17                -0.85
## 3           -0.22                     0.25                -0.08
##   Student.Performance.Rating
## 1                       0.00
## 2                      -0.36
## 3                       0.28
```

```
table(KMeans_Cluster = students$KCluster, EmployabilityClass = students$CLASS)
```

```
##               EmployabilityClass
## KMeans_Cluster Employable LessEmployable
##              1        675            210
##              2        475            458
##              3        579            585
```

Interpretation

The K-Means clustering algorithm was applied to the standardized interview attributes using k = 3 clusters, as suggested by the elbow method. Each cluster represents a distinct student profile based on similar attribute patterns.

Cluster 1 (High Performers): This cluster includes students with **high centroid values across all variables**, particularly in Communication Skills, Self Confidence, and Ability to Present Ideas. The majority of students in this cluster belong to the Employable class, suggesting strong alignment between high interpersonal and presentation skills and employability outcomes.

Cluster 2 (Moderate Performers): Students in this group display **average to slightly above-average scores** in most attributes but do not excel in any particular area. They typically show solid Mental Alertness and Physical Condition but moderate Confidence and Speaking Skills. The group represents students with potential who might benefit from focused training to improve communication and confidence.

Cluster 3 (Low Performers): This cluster is characterized by **lower centroid** scores across all attributes, especially Communication Skills and Ability to Present Ideas. It predominantly includes Less Employable students, indicating deficiencies in core soft skills crucial for professional readiness.

The K-Means results corroborate the hierarchical clustering findings: employability strongly correlates with a cluster of communication-oriented and confidence-based attributes. Students who exhibit consistent strengths in these areas form distinct, high-performing groups, whereas those lacking these skills cluster separately.

Interpretation and Recommendations

The multivariate analyses conducted in this study—spanning descriptive exploration, correlation assessment, principal component analysis (PCA), and clustering—collectively reveal important insights into the determinants of student employability and performance patterns.

1. Key Attributes Influencing Employability

From the exploratory and correlation analyses, attributes such as **Communication Skills, Self-Confidence, Ability to Present Ideas,** and **Mental Alertness** showed consistently strong relationships with the Student Performance Rating. These variables demonstrated positive correlations both with each other and with employability class. Students who performed well in these interpersonal and cognitive attributes were far more likely to be categorized as Employable, suggesting that employers value the combination of articulate communication, confident self-presentation, and critical thinking.

2. Insights from Principal Component Analysis (PCA)

PCA provided a condensed view of the underlying structure among the interview attributes:

- **PC1** (explaining ~55% of variance) represents a composite of overall interview strength, with large positive loadings from Communication Skills, Self-Confidence, Ability to Present Ideas, and Mental Alertness. This component captures the general dimension of professional readiness.
- **PC2** (adding ~14% variance) is dominated by Student Performance Rating, distinguishing actual performance evaluation from other self-reported or observed qualities.
- Together, PC1 and PC2 explain nearly 70% of total variability, showing that employability can be effectively summarized by two latent dimensions—general interpersonal competence and performance outcomes.

The scree plot confirmed that beyond these two components, subsequent dimensions contributed minimally, supporting the interpretability and dimensional reduction achieved through PCA.

3. Cluster Analysis Findings

Both hierarchical and K-Means clustering revealed three meaningful groups of students:

- **Cluster 1:** High-scoring individuals with strong communication, presentation, and self-confidence. This cluster predominantly comprises Employable students.
- **Cluster 2:** Students with average scores across most attributes. They show potential but require targeted development to reach employable standards.
- **Cluster 3:** Low scorers in core interpersonal and presentation skills, mostly Less Employable students.

The PCA-based cluster visualization showed that these clusters align strongly with the employability class labels. Employable students cluster in regions of high PC1 (overall readiness), while less employable students are concentrated in the opposite direction, confirming the coherence between statistical clusters and real-world classification.

4. Integrated Understanding and Recommendations

Combining these multivariate results leads to several important conclusions and actionable points:

- Communication, confidence, and clarity of presentation are the dominant factors determining employability outcomes. These should be focal areas for training and evaluation.
- Cluster profiles can be used for early identification of students needing intervention. Those in Cluster 2 and 3 may benefit from workshops on professional communication, critical thinking, and personal confidence building.
- PCA and clustering together demonstrate that student employability is multidimensional yet quantifiable—highlighting the feasibility of predictive models or targeted support systems in future assessments.
- Employability enhancement programs should integrate behavioral skill development alongside technical training, ensuring balanced readiness for job interviews.

Overall Summary

The study demonstrates that **multivariate methods provide a holistic framework** for understanding student employability. Through quantitative reduction (PCA) and pattern discovery (clustering), we identified that interpersonal and communication-related attributes most strongly separate employable from less employable candidates. This integrated approach not only clarifies the structure of employability factors but also offers a data-driven foundation for designing effective student development initiatives.

References

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065)

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley.

Villanueva, J. (2021). *Students' Employability Dataset – Philippines* [Dataset]. Kaggle.