**Student ID:** 21127690

**Name:** Ngô Nguyễn Thanh Thanh

**Report: Scikit-learn with PCA & LDA**

**I. Evaluation summary:**

|  |  |  |
| --- | --- | --- |
| **Task** | **Requirement Met(%)** | **Notes** |
| Create CSV file | 100% |  |
| Describe data info in the report | 100% |  |
| Implement basic multivariate analysis | 100% |  |
| Apply PCA & LDA using Scikit-learn | 70% | LDA is not completed |
| List comments for basic multivariate analysis | 100% |  |
| List comments for PCA & LDA implementation | 100% |  |

**II. List of funtion:**

1. **print\_mean\_and\_sd\_by\_group(variables, group\_variable)**
2. **calc\_within\_groups\_variance(variable, group\_variable)**
3. **calc\_between\_groups\_variance(variable, group\_variable)**
4. **calc\_separations(variables, group\_variable)**
5. **calc\_within\_groups\_covariance(variable1, variable2, group\_variable)**
6. **calc\_between\_groups\_covariance(variable1, variable2, group\_variable)**
7. **most\_highly\_correlated(my\_dataframe, num\_to\_report)**
8. **pca\_summary(pca, standardised\_data, out=True)**
9. **scree\_plot(pca, standardised\_values)**
10. **hinton(matrix, max\_weight=None, ax=None)**

**III. Function Summaries and Implementation**

1. **print\_mean\_and\_sd\_by\_group(variables, group\_variable)**
   * **Summary**: Calculates the mean and standard deviation of variables grouped by a specified group variable and prints the results.
   * **Implementation**:
     + Divides the data into groups based on the **group\_variable**.
     + Computes the mean and standard deviation of each variable within each group.
     + Prints the mean, standard deviation, and sample size for each group.
2. **calc\_within\_groups\_variance(variable, group\_variable)**
   * **Summary**: Computes within-group variance for a specific variable.
   * **Implementation**:
     + Divides the data into groups based on the **group\_variable**.
     + Calculates the variance of the variable within each group.
     + Returns the within-group variance.
3. **calc\_between\_groups\_variance(variable, group\_variable)**
   * **Summary**: Computes between-group variance for a specific variable.
   * **Implementation**:
     + Compares the mean values of the variable between the groups defined by the **group\_variable**.
     + Calculates the variance between these mean values.
     + Returns the between-group variance.
4. **calc\_separations(variables, group\_variable)**
   * **Summary**: Calculates separations between groups based on input variables.
   * **Implementation**:
     + Uses **calc\_within\_groups\_variance()** and **calc\_between\_groups\_variance()** functions to compute separation within a variable or between pairs of variables.
     + Returns the separation values for each variable.
5. **calc\_within\_groups\_covariance(variable1, variable2, group\_variable)**
   * **Summary**: Computes within-group covariance between two specific variables.
   * **Implementation**:
     + Divides the data into groups based on the **group\_variable**.
     + Calculates the covariance between the two variables within each group.
     + Returns the within-group covariance.
6. **calc\_between\_groups\_covariance(variable1, variable2, group\_variable)**
   * **Summary**: Computes between-group covariance for two specific variables.
   * **Implementation**:
     + Compares the mean values of the two variables between the groups defined by the **group\_variable**.
     + Calculates the covariance between these mean values.
     + Returns the between-group covariance.
7. **most\_highly\_correlated(my\_dataframe, num\_to\_report)**
   * **Summary**: Computes and returns the most highly correlated variable pairs from a DataFrame.
   * **Implementation**:
     + Calculates the correlation matrix using the **corr()** function.
     + Retrieves the variable pairs with the highest correlation and returns them.
8. **pca\_summary(pca, standardised\_data, out=True)**
   * **Summary**: Generates a summary table for Principal Component Analysis (PCA).
   * **Implementation**:
     + Computes the standard deviation, proportion of explained variance, and cumulative proportion of variance for each principal component.
     + Returns a DataFrame containing the summary statistics.
9. **scree\_plot(pca, standardised\_values)**
   * **Summary**: Plots a scree plot to show the distribution of variance explained by the principal components.
   * **Implementation**:
     + Uses the explained variance ratio from PCA to plot the graph.
10. **hinton(matrix, max\_weight=None, ax=None)**
    * **Summary**: Plots a Hinton diagram to visualize a weight matrix.
    * **Implementation**:
      + Uses size and color to represent the value of each cell in the matrix.

**V. The results:**

***Full image proof:***

- Read data from the given sample CSV file

***A screen shot of a computer screen

Description automatically generated***

A screenshot of a computer

Description automatically generated

- Plot data for visualization A graph with blue dots

Description automatically generated with medium confidence

- Scatterplot with the data points labelled by their Group

A screen shot of a computer screen

Description automatically generated

- Profile plot, used to shows the variation in each of the variables, by plotting the value of each of the variables for each of the samples

A screen shot of a graph

Description automatically generated

- Calculating summary statistics for multivariate data

Output: → Mean

A screenshot of a computer

Description automatically generated

→ Standard deviation

A screenshot of a computer screen

Description automatically generated

→ Max

A screenshot of a computer screen

Description automatically generated

- Means and variances per group

A screenshot of a computer

Description automatically generated

- Standardising variables

A screenshot of a computer

Description automatically generated

Check the summary of PCA results

Output:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

- Check how many principal components to retain

A graph with a line

Description automatically generated

- Calculate the values of the first principal component

Output:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

- Obtain the loadings for the second principal component

Output:

A black screen with numbers

Description automatically generated

- Visualize scatterplots of the principal components

A graph with colored dots

Description automatically generated

- Get mean, standard deviations, and sample sizes

Output:

A screenshot of a computer

Description automatically generated

-Scatter plot of LDA components

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated