INT 234 PROJECT REPORT

Data Analysis & Visualization on Penguins Dataset

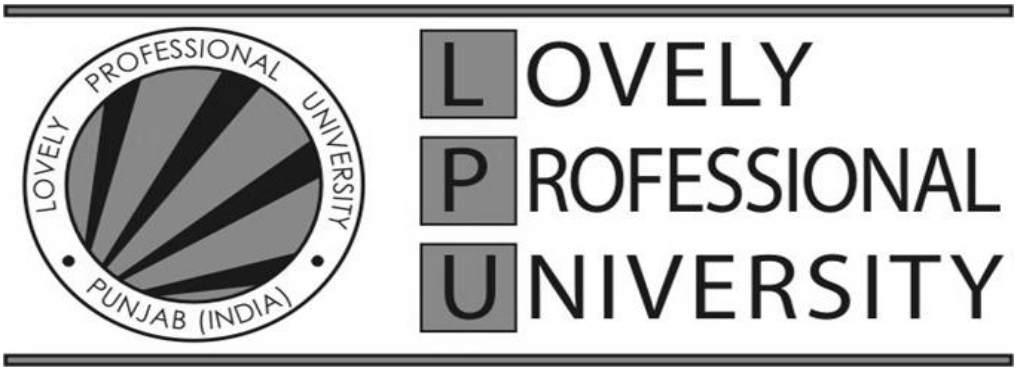
Submitted by

Vijay Deshmukh

1201959

K20SP INT 234

Under the guidance of Ms. Baljinder Kaur: 27952
Lovely Professional University, Phagwara



DECLARATION

I, **Vijay Deshmukh**, student of Computer Science Discipline at, Lovely Professional University, Punjab, here by declare that all the information on furnished in this project report is based on my own intensive work and is genuine.

Date: 01-11-2023

Registration No.12019859

Signature

Vijay Deshmukh

Introduction

This comprehensive analysis explores the penguins dataset, a rich source of information on penguin species, their physical attributes, and habitat. The process begins with data installation and library loading, followed by an initial examination of the dataset's structure and contents. Missing values are addressed through data cleaning, ensuring robustness in subsequent analyses.

Categorical variables such as species, island, and sex are thoroughly investigated to gain insights into their distributions. Subsequently, a series of exploratory data analysis (EDA) techniques are employed. Histograms provide detailed distributions of numerical variables including bill length, bill depth, flipper length, and body mass. Boxplots delve into the relationships between species and sex, as well as island and sex.

A scatterplot matrix sheds light on potential correlations between the numerical attributes. Furthermore, a correlation matrix quantifies the relationships, providing a nuanced understanding of the dataset's interdependencies. Visualizations, such as bar charts, are employed to showcase the distribution of species and islands, offering valuable contextual insights.

To enhance understanding, scatter plots illustrate relationships between numerical variables, such as bill length vs. bill depth and flipper length vs. body mass. The analysis culminates with a heatmap displaying correlation strength among numerical variables.

Acknowledgement

I would like to extend my sincere thanks to the creators of the datset for providing access to the invaluable Penguins dataset. This dataset formed the cornerstone of our analysis, enabling us to gain insights into the behavior and characteristics of penguins.

I would also like to express my gratitude to the developers of the R packages class, gmodels, ggplot2, rpart, rpart.plot, neuralnet, tm, wordcloud,factoextra, and cluster. These packages played a pivotal role in visualizing, modeling, and exploring the data, enhancing the depth and breadth of our analysis.

Furthermore, I would like to acknowledge the R community for their collective efforts in developing and maintaining these packages. Their dedication and contributions have significantly enriched the R ecosystem, providing researchers and analysts with powerful tools for data exploration and analysis.

OBJECTIVES

- **Exploratory Data Analysis (EDA)**
- **Apply Knn**
- **Word Cloud**
- **Decision Tree**
- **Linear Regression**
- **k-means Clustering**
- **Neural Network**
- **Hierarchical Clustering Dendrogram**

Source of Dataset

The dataset used in this analysis was the penguins dataset, which contains information about various penguin species, including their physical attributes and habitat details. The dataset is a valuable resource for conducting exploratory data analysis and gaining insights into penguin biology and ecology. The dataset is available on GitHub and can be accessed using the following link:

Ml_With_R_Project/ML With R Project/penguins.csv at main · VijayDeshmukh12/Ml_With_R_Project(github.com)

ETL PROCESS

The ETL (Extract, Transform, Load) process is a common method for preparing data for analysis. An ETL process for the penguins dataset:

1. Extraction:

The penguins dataset .csv file is downloaded.

The penguins dataset is loaded using the read.csv function.

2. Transformation:

Initial exploration of the data is performed to understand its structure and content.

Data is viewed using the head and View functions to get a glimpse of the first few rows.

The str function is used to understand the structure of the data (e.g., data types, number of observations).

Missing values are checked using is.na and the total count of missing values is calculated using sum(is.na(data)).

Rows with missing data are removed using na.omit(data) to ensure data integrity.

3. Loading:

The cleaned data is stored in the variable data for further analysis.

Analysis on dataset (for each analysis)

I. Introduction

In this analysis, we explore the 'penguins' dataset. Our goal is to extract valuable insights from the data. We begin by loading necessary packages, examining data structure, and performing data cleaning to handle missing values. Categorical variables like species, island, and sex are investigated.

We delve into relationships between numerical variables using visualizations like histograms, boxplots, scatterplot matrices, and correlation matrices. Additionally, we examine the distribution of species and islands to understand categorical patterns.

Duplicates are addressed, enhancing data accuracy. We set a random seed for result reproducibility and split the data for model building. The k-Nearest Neighbors (kNN) algorithm is used for classification, with performance evaluated through confusion matrices and accuracy metrics.

Text mining techniques are employed, including the creation of a corpus, term-document matrix calculations, and a word cloud visualization. A decision tree model sheds light on species classification. We also fit a linear regression model to explore the relationship between body mass and flipper length, visualized through a regression line.

Principal Component Analysis (PCA) is used for dimensionality reduction, followed by k-means clustering. The clustered data is visualized. A neural network model is built to predict body mass based on flipper length and bill depth.

Finally, hierarchical clustering reveals potential subgroups within the dataset via a dendrogram. This multi-faceted analysis aims to uncover meaningful insights and relationships in the penguins dataset, employing a range of exploratory techniques and machine learning approaches.

II. General Description

Data Loading and Initial Exploration:
The code begins by loading the dataset and data("penguins").
The first six rows of the dataset are displayed using head(penguins).
The dataset is assigned to the variable data for further processing.
The structure of the dataset is examined using str(data) to understand variable types and data integrity.
Missing values are checked using is.na(data) and the total count of missing values is calculated with sum(is.na(data)). Any rows with missing data are removed using data <- na.omit(data).
Exploring Categorical Variables:
The frequency distribution of categorical variables (species, island, sex) is examined using table().
Exploratory Data Analysis (EDA):

- Distribution of Numeric Variables:

Four histograms are plotted for bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g.

- Boxplots for Categorical Variables:

Boxplots are created to visualize the distribution of species by sex and island by sex.

- Scatterplot Matrix for Numeric Variables:

A scatterplot matrix is generated to explore relationships between numeric variables.

Correlation Analysis:

The correlation matrix is calculated for the numeric variables (bill_length_mm, bill_depth_mm, flipper_ length_mm, body_mass_g) using cor().

Visualization of Species and Islands:

Bar plots are created to visualize the distribution of species and island variables.

Visualizing Relationships:

Scatterplots are generated to visualize relationships between numerical variables (bill_length_mm vs bill_depth_mm and flipper_length_mm vs body_mass_g) based on the species.

Heatmap for Correlation:

A heatmap is generated to visually represent the correlation matrix, providing insights into the relationships between numeric variables.

Handling Duplicate Rows:

Duplicate rows in the dataset are identified and removed.

Data Splitting for Machine Learning:

The dataset is split into training and testing sets for machine learning model training and evaluation.

k-Nearest Neighbors (kNN) Classification:

kNN algorithm is applied to predict the species based on bill_length_mm and bill_depth_mm. The optimal k value is set to 5.

Confusion Matrix and Accuracy:

A confusion matrix is created to evaluate the performance of the kNN model, and accuracy is calculated.

Text Mining and Word Cloud:

A term-document matrix is generated to analyze word frequencies based on the columns species, island, and sex. A word cloud is created to visualize the most frequent terms.

Decision Tree and Visualization:

A decision tree model is built to predict species based on other variables. The resulting tree is visualized.

Linear Regression Model:

A linear regression model is fitted to predict body_mass_g based on flipper_length_mm. The regression line is visualized.

Principal Component Analysis (PCA) and k-Means Clustering:

PCA is performed for dimensionality reduction, and k-means clustering is applied using the first two principal components. Clusters are visualized.

Neural Network Model:

A neural network model is created to predict body_mass_g based on flipper_length_mm and bill_depth_mm. The neural network is visualized.

Hierarchical Clustering Dendrogram:

Hierarchical clustering is applied using average linkage, and a dendrogram is plotted to visualize the clustering structure.

Cluster Assignments:

The data points are assigned to clusters based on the hierarchical clustering results. Rectangles are added around the clusters in the dendrogram.

III. Specific Requirements, functions and formulas

Exploratory Data Analysis (EDA)
Data Loading and Preliminary Analysis:
Specific Requirements: None
Functions and Formulas:
data("penguins"): Loads the penguins dataset.
head(penguins): Displays the first few rows of the dataset.
data <- penguins: Assigns the dataset to a variable named data.
View(data): Opens a data viewer to explore the dataset.
str(data): Provides the structure of the dataset.
is.na(data): Checks for missing values in the dataset.
sum(is.na(data)): Calculates the total number of missing values.
data <- na.omit(data): Deletes rows with missing data.
Exploring Categorical Variables:
Specific Requirements: None
Functions and Formulas:
table(penguins$species): Counts the occurrences of each species.
table(penguins$island): Counts the occurrences of each island.
table(penguins$sex): Counts the occurrences of each sex.
Exploring Relationships Between Variables:
Specific Requirements: None
Functions and Formulas:
par(mfrow=c(2,2)): Sets the layout for multiple plots.
hist(): Generates histograms for numeric variables.
boxplot(): Creates boxplots for categorical variables.
pairs(): Generates a scatterplot matrix for numeric variables.
cor(): Calculates the correlation matrix.
Visualizing Data Distributions:
Specific Requirements: None
Functions and Formulas:
ggplot(): Creates various types of plots using the Grammar of Graphics.
Heatmap for Correlations:
Specific Requirements: None
Functions and Formulas:
heatmap(): Generates a heatmap to visualize the correlation matrix.
Data Preprocessing and Cleaning:
Handling Duplicate Rows:
Specific Requirements: None
Functions and Formulas:
duplicates <- sum(duplicated(data)): Counts duplicate rows.
data <- data[!duplicated(data),]: Removes duplicate rows.
Data Splitting:
Specific Requirements: None
Functions and Formulas:
set.seed(123): Sets a seed for reproducibility.
indexes <- sample(1:nrow(data), size = 0.7 * nrow(data)): Generates random indexes for splitting data.

train_data <- data[indexes,] and test_data <- data[-indexes,]: Splits data into training and testing sets.

Normalization Function:

Specific Requirements: None

Functions and Formulas:

normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }: Defines a function for normalization.

k-Nearest Neighbors (kNN) Classification:

Specific Requirements: None

Functions and Formulas:

install.packages("class") and library(class): Installs and loads the class package for kNN.

knn(): Applies k-Nearest Neighbors algorithm for classification.

Model Evaluation:

Confusion Matrix and Accuracy:

Specific Requirements: None

Functions and Formulas:

CrossTable(): Generates a cross table to analyze classification results.

table(): Creates a confusion matrix.

Calculation of accuracy using sum(diag(conf_matrix)) / sum(conf_matrix).

Text Mining and Wordcloud:

Text Corpus Creation:

Specific Requirements: None

Functions and Formulas:

install.packages("tm") and library(tm): Installs and loads the tm package for text mining.

Corpus() and TermDocumentMatrix(): Create a term-document matrix from text data.

Word Frequency and Wordcloud:

Specific Requirements: None

Functions and Formulas:

install.packages("wordcloud") and library(wordcloud): Installs and loads the wordcloud package for wordcloud creation.

wordcloud(): Generates a wordcloud visualization.

Decision Tree, Linear Regression, and Clustering:

Specific Requirements: None

Functions and Formulas:

Functions and formulas specific to decision trees, linear regression, and clustering algorithms.

Neural Network Modeling:

Specific Requirements: None

Functions and Formulas:

install.packages("neuralnet") and library(neuralnet): Installs and loads the neuralnet package for neural network modeling.

neuralnet(): Creates a neural network model.

IV. Analysis Results

Exploratory Data Analysis (EDA):

The analysis begins with the installation of the palmerpenguins package and loading the necessary libraries. The penguins dataset is then loaded.

A glimpse of the dataset is provided using the head() function, displaying the first few rows.

The dataset is then assigned to the variable data.

The structure of the dataset is examined using str(data), giving insights into variable types and data structure.

Handling Missing Values:

Missing values are checked using is.na(data).

The total number of missing values is calculated with sum(is.na(data)).

Rows with missing data are removed using na.omit(data).

Exploring Categorical Variables:

Frequencies of categorical variables (species, island, and sex) are computed using the table() function.

Exploring Relationships Between Variables:

The exploratory data analysis continues with a series of visualizations to understand relationships between variables.

Distribution of numerical variables is displayed using histograms.

Boxplots are used to compare categorical variables.

A scatterplot matrix is created to visualize relationships between numerical variables.

A correlation matrix is computed to understand the relationships between numeric variables.

Visualization:

The distribution of species and islands are visualized using bar plots.

Scatter plots are created to visualize relationships between numerical variables.

Heatmap for Correlations:

A heatmap is generated to visualize the correlation matrix between numerical variables.

Handling Duplicate Rows:

Duplicate rows are checked for using duplicated(data).

Total duplicate rows are printed.

Data Splitting for Modeling:

The data is split into training and testing sets using a random seed for reproducibility.

k-Nearest Neighbors (kNN) Classification:

The kNN algorithm is applied for species classification based on bill_length_mm and bill_depth_mm.

A confusion matrix is created to compare predicted labels with actual labels.

Accuracy is calculated based on the confusion matrix.

Text Mining and Wordcloud:

A text corpus is created from the species, island, and sex columns.

A term-document matrix is generated, and word frequencies are calculated.

A wordcloud is generated based on word frequencies.

Decision Tree Modeling:

A decision tree model is built for species classification using all available features.

Linear Regression Modeling:

A linear regression model is fitted to predict body mass (body_mass_g) based on flipper length (flipper_length_mm).

Principal Component Analysis (PCA) and K-Means Clustering:

Principal Component Analysis is performed for dimensionality reduction.

K-Means clustering is applied to the first two principal components.

Neural Network Modeling:

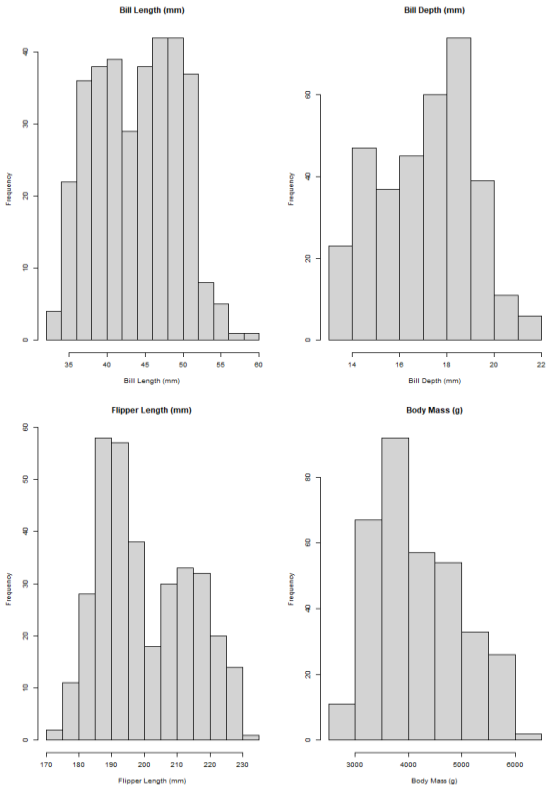
A neural network model is created to predict body mass based on flipper_length_mm and bill_depth_mm.

Hierarchical Clustering:

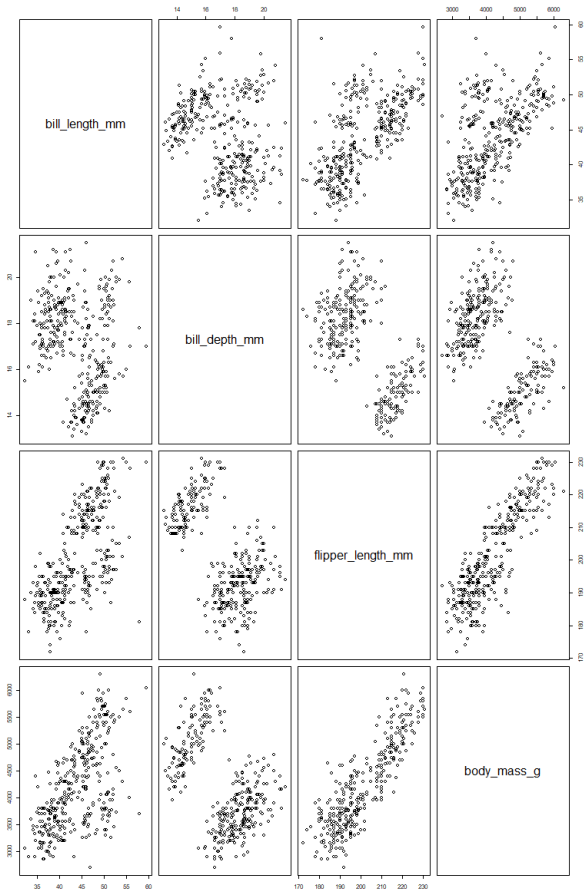
Hierarchical clustering is performed on the cleaned data using average linkage.

V. Visualization

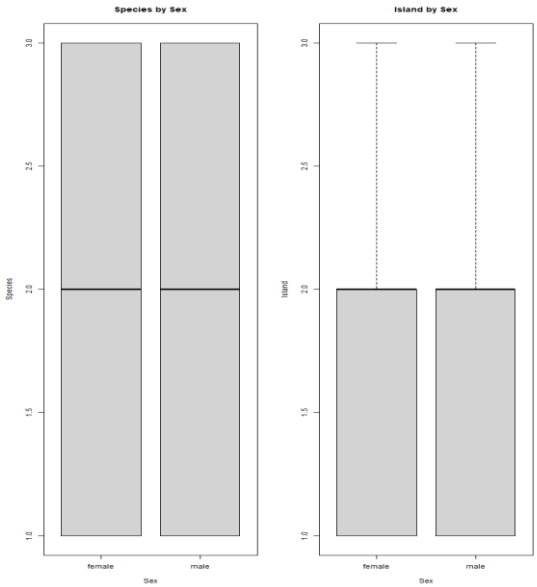
#Exploratory Data Analysis (EDA)
#Distribution of numerical variables



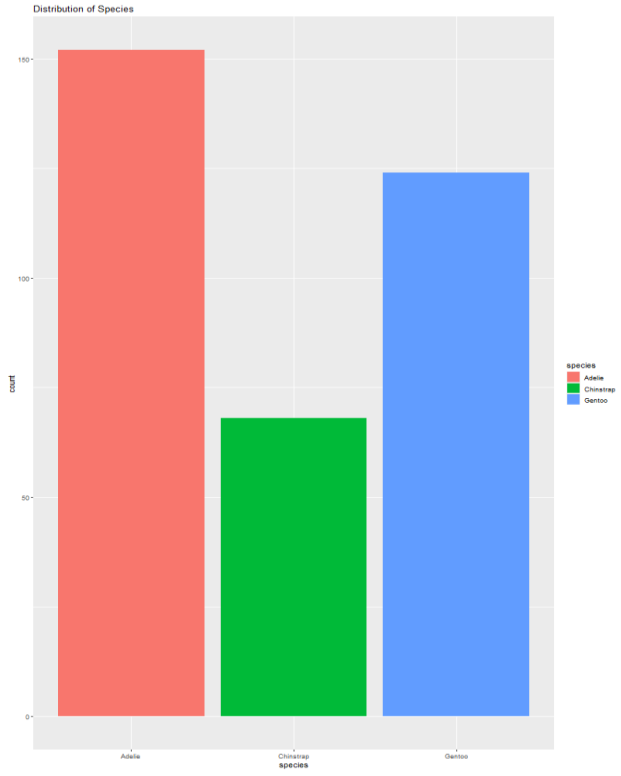
Scatterplot matrix for numerical variables



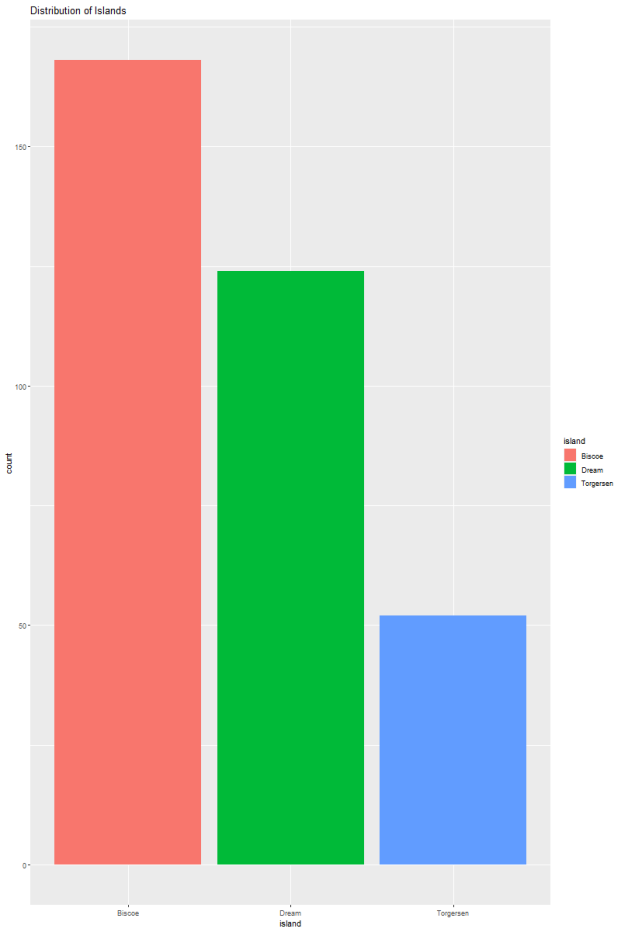
Boxplots for categorical variables



#Visualize the distribution of species

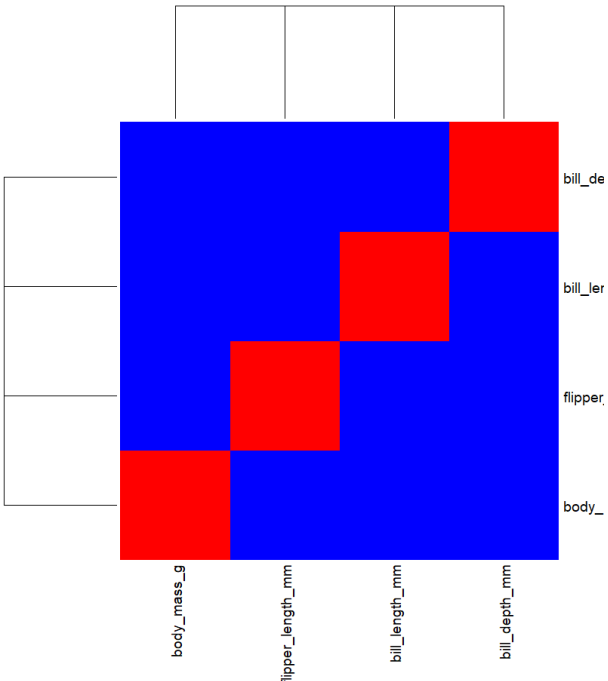


Visualize the distribution of islands

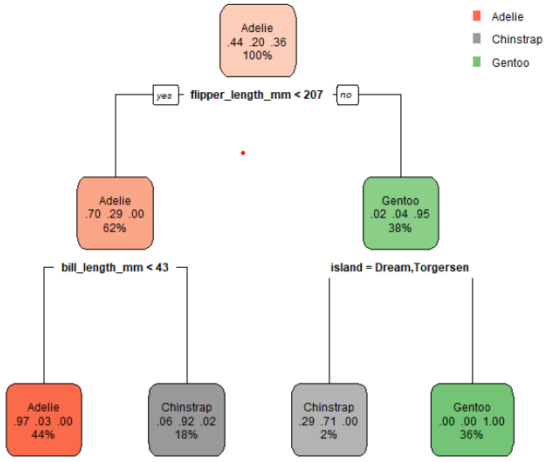


#WordCloud

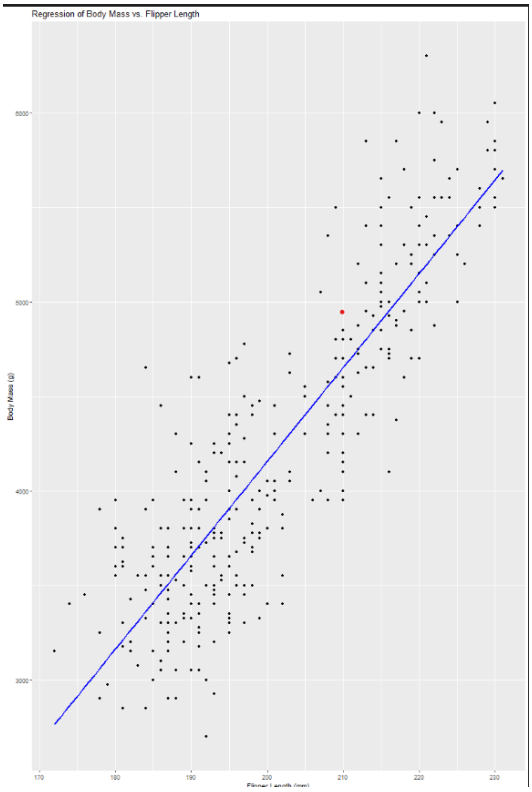
#Heatmap for correlations



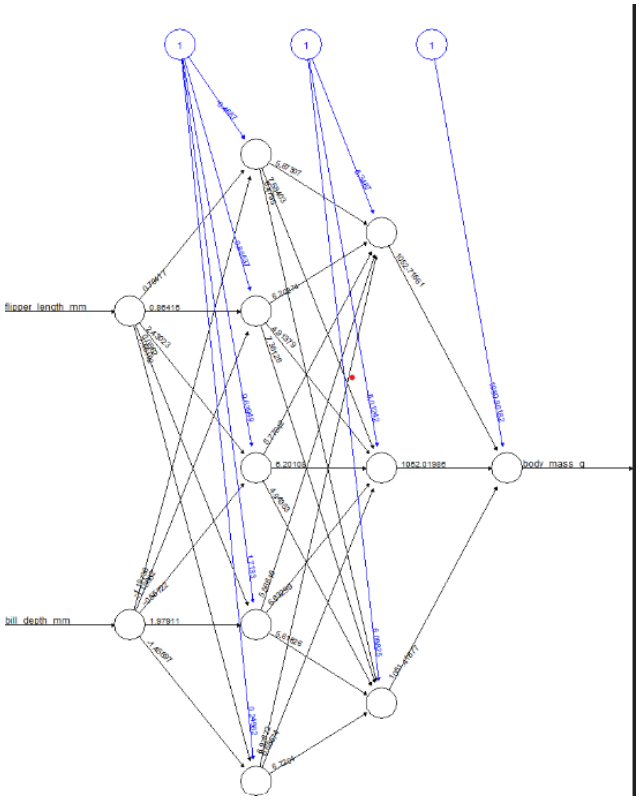
#Decision Tree



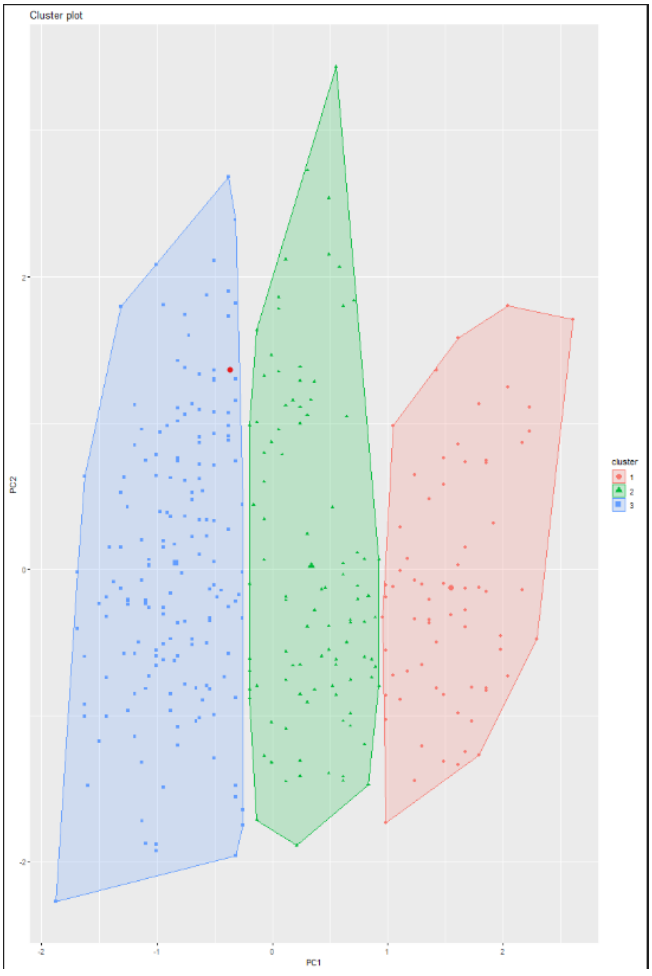
#Linear Regression



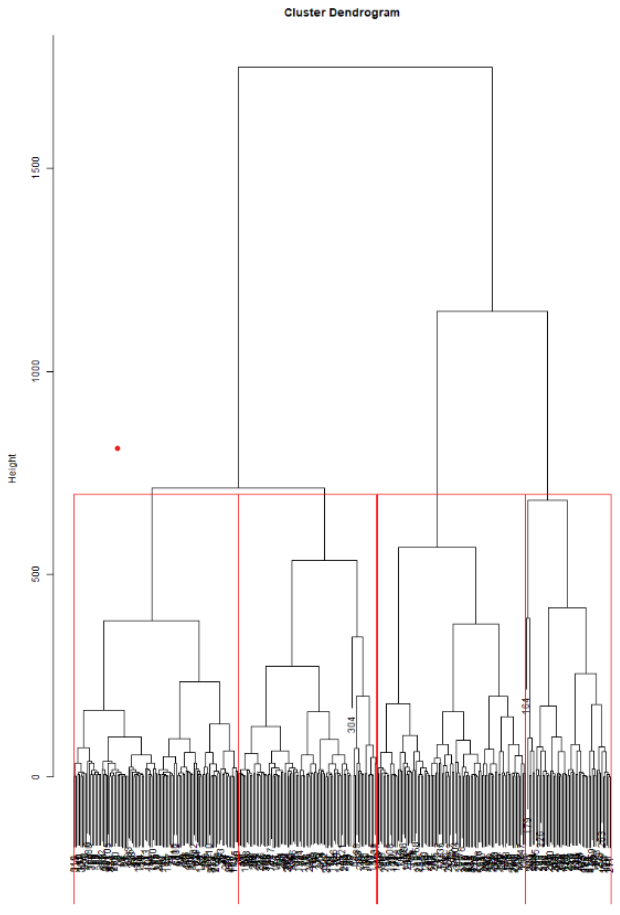
Visualize the neural network



#Visualize the clusters



#Hierarchical Clustering Dendrogram



List of Analysis with Results

Data Preparation and Exploration:
Installed and loaded the required package.
Loaded the necessary libraries and imported the penguins dataset.
Checked the first few rows of the dataset using head.
Viewed the dataset using View.
Examined the structure of the dataset with str.
Checked for missing values and removed them using na.omit.
Exploring Categorical Variables:
Conducted an analysis of categorical variables including species, island, and sex.
Exploratory Data Analysis (EDA):
Conducted a comprehensive EDA which included:
Distribution of numerical variables through histograms.
Boxplots for categorical variables comparing species and sex, and island and sex.
Scatterplot matrix for numerical variables.
Correlation matrix for numerical variables.
Visualized the distribution of species and island using bar plots.
Explored relationships between numerical variables through scatterplots.
K-Nearest Neighbors (kNN) Classification:
Applied kNN classification using the knn function from the class package.
Evaluated the results using a confusion matrix and calculated accuracy.
Text Analysis:
Conducted a basic text analysis using the tm and wordcloud packages.
Created a term-document matrix and generated a word cloud based on species, island, and sex columns.
Decision Tree Model:
Built and visualized a decision tree classification model using the rpart and rpart.plot packages.
Linear Regression:
Fitted a linear regression model to predict body_mass_g based on flipper_length_mm.
Principal Component Analysis (PCA) and K-Means Clustering:
Conducted PCA for dimensionality reduction and used the first two principal components for clustering.
Applied k-means clustering to identify potential clusters in the data.
Visualized the clusters.
Neural Network Model:
Created a neural network model to predict body_mass_g based on flipper_length_mm and bill_depth_mm.
Visualized the neural network.
Hierarchical Clustering:
Performed hierarchical clustering on the clean dataset using Euclidean distance and average linkage.
Plotted the dendrogram and identified 4 clusters.
Added rectangles around the clusters for better visualization.

References

Dataset URL :-
https://github.com/VijayDeshmukh12/Ml_With_R_Poject/blob/main/ML%20With%20R%20Project/penguins.csv

Thank You..!