**Iris Dataset Exploratory Data Analysis (EDA)**

This repository contains Python code for performing Exploratory Data Analysis (EDA) on the Iris dataset. The analysis includes generating summary statistics, creating histograms, scatter plots, and a correlation heatmap for the numeric variables.

Table of Contents

[Installation 1](#_Toc166994527)

[Usage 1](#_Toc166994528)

[Project Overview 1](#_Toc166994529)

[Directory Structure 2](#_Toc166994530)

[Results 2](#_Toc166994531)

[References 3](#_Toc166994532)

# Installation

To run the code in this repository, you need to have Python 3.x and the following Python packages installed:

ucimlrepo

pandas

seaborn

matplotlib

You can install the required packages using pip:

pip install ucimlrepo pandas seaborn matplotlib

# Usage

Ensure that the output directory is set correctly in the script. Then, execute the script:

python iris\_EDA\_G00305450.py

# Project Overview

**Dataset**

The UCI Machine Learning Repository provides 150 samples of iris flowers with four features: sepal length, sepal width, petal length, and petal width. These samples make up the Iris dataset. One of the three species—Iris setosa, Iris versicolor, or Iris virginica—represents each sample.

**Approach**

The approach to the analysis is as follows:

1. Data Fetching: The dataset is fetched using the ucimlrepo package.
2. DataFrame Creation: The data is converted into a Pandas DataFrame.
3. Summary Statistics: Basic summary statistics for each numeric variable are computed and saved to a text file. Provides an overview of the central tendency, dispersion, and shape of the dataset's distribution
4. Histograms: Histograms are created for each numeric variable to visualize their distributions. Help in understanding the distribution and frequency of numeric variables.
5. Scatter Plots: Scatter plots for each pair of numeric variables are generated to observe relationships. Useful for identifying potential relationships and correlations between pairs of variables.
6. Correlation Heatmap: A heatmap of the correlation matrix is created to show correlations between variables.

# Directory Structure

├── iris\_EDA\_G00305450.py

├── output\_files1/

│ ├── summary\_statistics.txt

│ ├── sepal\_length\_histogram.png

│ ├── sepal\_width\_histogram.png

│ ├── petal\_length\_histogram.png

│ ├── petal\_width\_histogram.png

│ ├── scatter\_plots.png

├── README.md

└── requirements.txt

**iris\_EDA\_G00305450.py:** Main script that performs the EDA.

**output\_files1/:** Directory where the output files (summary statistics, histograms, scatter plots) are saved.

**README.md:** This readme file, the file you are reading.

**requirements.txt:** List of required Python packages

# Results

**Summary Statistics:** Stored in output\_files1/summary\_statistics.txt.

**Histograms:** Stored as PNG files in output\_files1/.

**Scatter Plots:** Stored as scatter\_plots.png in output\_files1/.

**Correlation Heatmap:** Stored as correlation\_matrix.png in output\_files1/.

# References

[1] UC Irvine Machine Learning Repository. Iris data set. [Iris - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/53/iris)

[2] Pandas Documentation. <https://pandas.pydata.org/docs/>

[3] Seaborn heatmap Documentation <https://seaborn.pydata.org/generated/seaborn.heatmap.html>

[4] Mathplotlib histrogram Documentation <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html>

[5] Exploratory Data Analysis on Iris Dataset <https://www.geeksforgeeks.org/exploratory-data-analysis-on-iris-dataset/>