**Analysis of the Iris Dataset**

Exploratory Data Analysis and Visualization

**Date:** 2nd August 2024

**Name**: Ishita Singh

This project involves performing Exploratory Data Analysis (EDA) on the Iris dataset to understand its key characteristics, visualize the data, and explore correlations between different features.

The Iris dataset is a well-known dataset in the field of machine learning and statistics. It consists of data collected from three species of iris flowers: Setosa, Versicolor, and Virginica. Here's a brief overview of its features and target variable:

* **Features**:

1. **Sepal** **Length**: The length of the sepal in centimeters.

2. **Sepal** **Width**: The width of the sepal in centimeters.

3. **Petal** **Length**: The length of the petal in centimeters.

4. **Petal** **Width**: The width of the petal in centimeters.

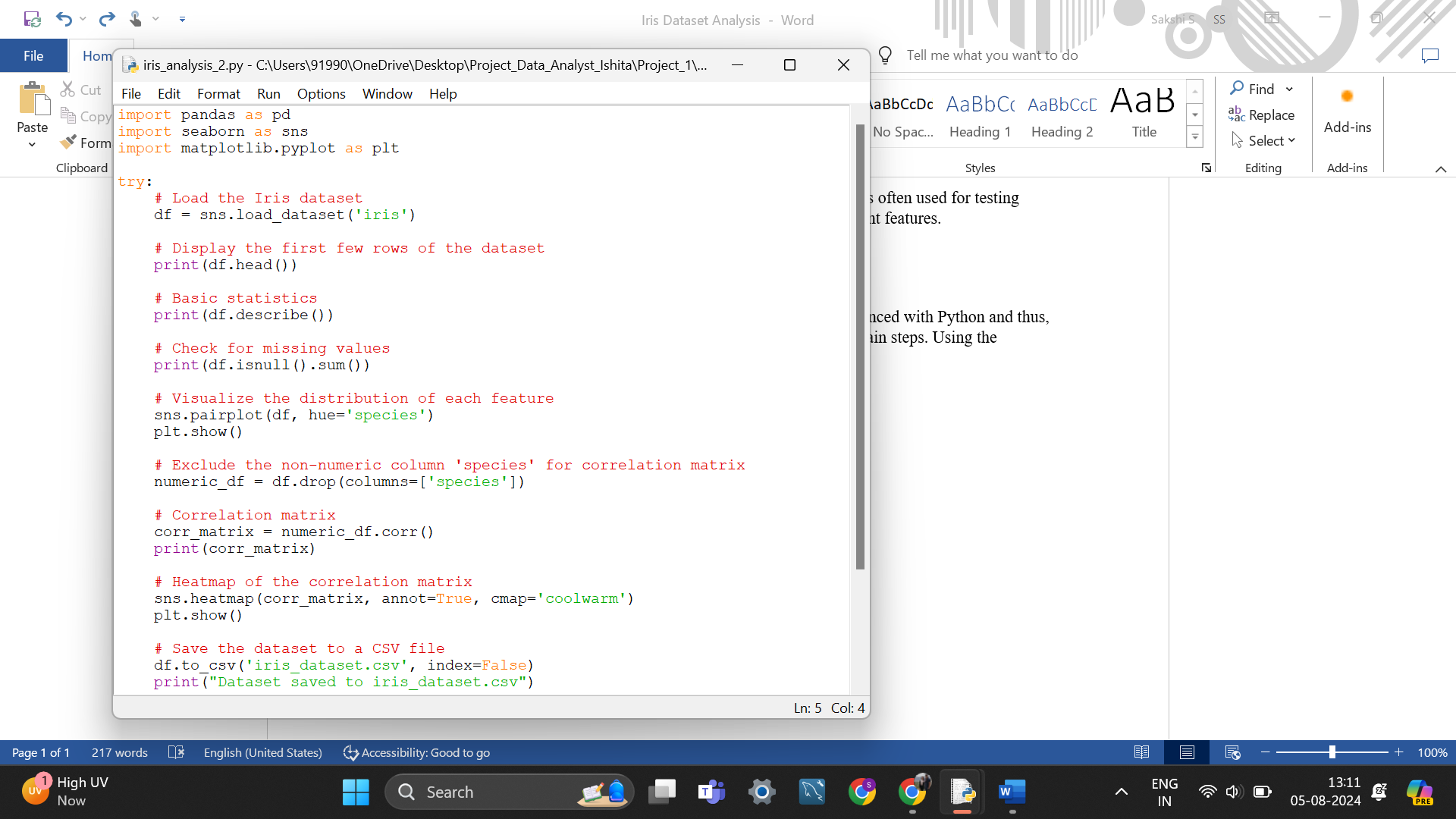
* **Target Variable:**

**Species**: The species of the iris flower. It is a categorical variable with three classes: Setosa, Versicolor, and Virginica.

The dataset contains 150 samples, with 50 samples from each species. It is often used for testing classification algorithms and visualizing the relationships between different features.

**Working on Python:**

I downloaded Python and installed the required libraries. I am not experienced with Python and thus, used ChatGPT to help learn. I went onto YouTube as well to work on certain steps. Using the following code, I started the python script:



After this, to be informed if any errors were being generated I used the “try-except code”. Then, after loading the data, I asked python to generate a few rows of the data to see if it has been successfully loaded. This was followed by generating basic statistics like: count, max, min, mean, 25th, 50th and 75th percentiles. Following are the description of the statistics and their interpretation in the Iris Dataset:

1. **Count**: The total number of observations or data points in the dataset.
2. **Max**: The maximum value in a dataset. It represents the largest observation.
3. **Min**: The minimum value in a dataset. It represents the smallest observation.
4. **25th Percentile (Q1)**: The value below which 25% of the observations fall. It is also known as the first quartile.
5. **Median (50th Percentile or Q2)**: The middle value of the dataset when it is ordered. It divides the data into two equal halves.
6. **75th Percentile (Q3)**: The value below which 75% of the observations fall. It is also known as the third quartile.

**What Each Statistic Tells About the Data in the Iris Dataset:**

1. **Count**: This tells you how many iris flowers were measured for each feature. In the Iris dataset, the count for each feature should be 150, as there are 150 observations.
2. **Max**: This shows the maximum recorded value for each feature. For example, the maximum petal length might indicate the largest petal size observed in the dataset.
3. **Min**: This shows the minimum recorded value for each feature. For example, the minimum sepal width could reveal the smallest sepal width observed.
4. **25th Percentile (Q1)**: This provides insight into the lower end of the distribution. For instance, if the 25th percentile of petal width is 1.2 cm, it means that 25% of the iris flowers have a petal width of 1.2 cm or less.
5. **Median**: This is a measure of central tendency that indicates the middle value of the feature when all observations are sorted. It gives a central point of the data distribution. For example, the median sepal length tells you the central value of sepal length.
6. **75th Percentile (Q3)**: This provides insight into the upper end of the distribution. For example, if the 75th percentile of sepal length is 6.0 cm, it means that 75% of the iris flowers have a sepal length of 6.0 cm or less.

Together, these statistics offer a comprehensive summary of the distribution of each feature in the Iris dataset, helping you understand the spread, central tendency, and range of the data.

**Visualizations Overview**

**Pair Plot**

The pair plot provides scatter plots for each pair of features and kernel density estimates (KDE) for the diagonal elements, showing the distribution of each feature. This visualization helps in identifying relationships and distributions among the features for different species.

**Correlation Matrix**

The correlation matrix uses a heatmap to display the Pearson correlation coefficients between all pairs of features. This helps in understanding the linear relationships between the features.

**Interpretation of Visualizations**

**Pair Plot Analysis**

The pair plot includes scatter plots and KDEs for sepal length, sepal width, petal length, and petal width. Key observations from the pair plot are:

1. **Sepal Length vs. Sepal Width:**

* Setosa (blue) has distinct, smaller sepal lengths and widths compared to Versicolor (orange) and Virginica (green).
* There is no strong linear relationship between sepal length and sepal width for any species.

2. **Sepal Length vs. Petal Length:**

* There is a clear linear separation between Setosa and the other two species.
* Versicolor and Virginica overlap but can still be distinguished to some extent.

3. **Sepal Length vs. Petal Width:**

* Setosa is well-separated from Versicolor and Virginica.
* There is a positive correlation between sepal length and petal width for Versicolor and Virginica.

4. **Sepal Width vs. Petal Length:**

* Setosa shows a distinct grouping with smaller petal lengths.
* Versicolor and Virginica show some overlap.

5. **Sepal Width vs. Petal Width:**

* Setosa is again distinct with smaller petal widths.
* Versicolor and Virginica overlap but show a positive correlation.

6. **Petal Length vs. Petal Width:**

* This plot shows the most clear separation between the three species.
* Setosa has the smallest petal length and width.
* Versicolor and Virginica are linearly separable to some extent.

**Correlation Matrix Analysis**

The correlation matrix provides Pearson correlation coefficients between features. The key observations from the heatmap are:

1. **Sepal Length:**

* Highly positively correlated with petal length (0.87) and petal width (0.82).
* Negatively correlated with sepal width (-0.12).

2. **Sepal Width:**

* Negatively correlated with petal length (-0.43) and petal width (-0.37).
* Weak negative correlation with sepal length (-0.12).

3. **Petal Length:**

* Highly positively correlated with petal width (0.96).
* Negatively correlated with sepal width (-0.43).

4**. Petal Width:**

* Highly positively correlated with petal length (0.96).
* Negatively correlated with sepal width (-0.37).

**Interpretation of Correlation Coefficients**

**High Correlation (close to 1 or -1):**

* Indicates a strong linear relationship between two features.
* For example, petal length and petal width have a high positive correlation (0.96), indicating that as one increases, the other tends to increase as well.

**Low Correlation (close to 0):**

* Indicates a weak or no linear relationship between two features.
* For example, sepal length and sepal width have a low negative correlation (-0.12), indicating little to no linear relationship.

**Conclusion**

The exploratory data analysis of the Iris dataset reveals significant insights:

**1. Species Separation:**

* Setosa is distinctly separable from Versicolor and Virginica based on all features.
* Versicolor and Virginica show some overlap but can be distinguished using combinations of petal length and petal width.

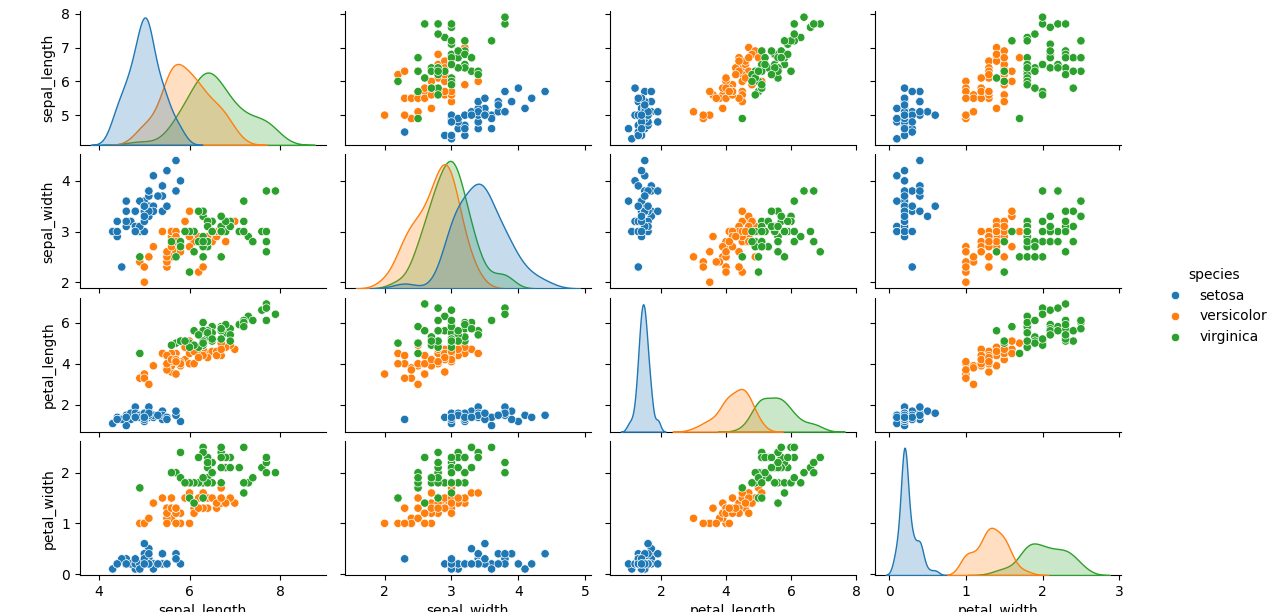
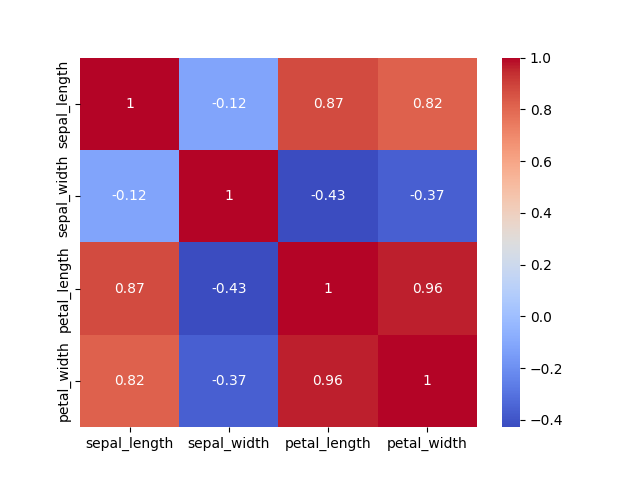
**2. Feature Relationships:**

* Petal length and petal width have the highest positive correlation, suggesting these features are strongly related.
* Sepal width has weak correlations with other features, indicating it may not be as useful for distinguishing between species.

These insights can guide further analysis and model building for classification tasks using the Iris dataset.

**Figures**

* **Figure 1:** Pair plot of sepal length, sepal width, petal length, and petal width for Setosa, Versicolor, and Virginica species.
* **Figure 2:** Correlation matrix heatmap displaying Pearson correlation coefficients between features.

By understanding the relationships and distributions of features within the Iris dataset, we can make informed decisions for data preprocessing, feature selection, and model development for tasks such as classification and clustering.

**Figure 1:**

**Figure 2:**