Statistical Analysis of the Iris Species Dataset

Introduction

This report presents a comprehensive statistical analysis of the Iris Species dataset. The dataset contains measurements of four features (sepal length, sepal width, petal length, and petal width) for three different species of Iris flowers. The aim of this analysis is to apply various statistical concepts to gain insights and make inferences about the data.

Task 1: Descriptive Statistics

In this task, we calculated the mean, median, and mode of each attribute for the entire dataset. Additionally, we computed the standard deviation and variance for each attribute. The results are summarized in the table below:

Attribute	Mean	Median	Mode
Sepal Length	5.84333333333	5.8	5
Sepal Width	3.054	3	3
Petal Length	3.75866667	4.35	1.5
Petal Width	1.2313333	1.3	0.2

The mean represents the average value of the attribute.

The median is the middle value when the data is sorted.

The mode is the most frequently occurring value.

Attribute	Standard Deviation	Variance
Sepal Length	0.825301292	0.681122222
Sepal Width	0.4321466	0.1867507
Petal Length	1.75852918	3.09242489
Petal Width	0.8199298	0.6722849

The standard deviation provides a measure of how spread out the data is around the mean. A higher standard deviation indicates greater variability.

The variance is the square of the standard deviation and provides a measure of the dispersion of the data.

Task 2: Correlation Analysis

In this task, we calculated the correlation matrix for the four attributes using a correlation table. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. The results are shown in the table below:

Correlation Matrix:

#	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1	-0.109369	0.871754	0.698758
Sepal Width	-0.109369	1	-0.420516	-0.326509
Petal Length	0.871754	-0.420516	1	0.841935
Petal Width	0.698758	-0.326509	0.841935	1

Based on the correlation matrix, we can observe the following:

- Interpretation:

- Correlation coefficients range from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

- Sepal Length has a strong positive correlation with Petal Length (0.871) and Petal Width (0.698).

- Sepal Width has a weak negative correlation with Petal Length (-0.420) and Petal Width (-0.356).

- Petal Length has a strong positive correlation with Sepal Length (0.871) and Petal Width (0.841).

- Petal Width has a strong positive correlation with Sepal Length (0.698) and Petal Length (0.841).

Task 3: Hypothesis Testing

In this task, we performed a hypothesis test to determine if there is a significant difference in sepal length between the **Iris setosa** and **Iris versicolor** species. The null and alternative hypotheses are as follows:

- Null Hypothesis: There is no significant difference in sepal length between Iris setosa and Iris versicolor species.

- Alternative Hypothesis: There is a significant difference in sepal length between Iris setosa and Iris versicolor species.

To test this hypothesis, we can use a t-test assuming independent samples. We will compare the sepal lengths of the two species.

After performing the t-test and assuming a significance level of 0.05, if the resulting p-value is below 0.05, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Conclusion:

Based on the results of the t-test, if the p-value is less than 0.05, i can conclude that there is a significant difference in sepal length between **Iris setosa** and **Iris versicolor** species. If the p-value is greater than 0.05, i fail to reject the null hypothesis and conclude that there is no significant difference in sepal length between the two species.

In this case, the very small p-value (1.24191×10−13) indicates an extremely strong evidence against the null hypothesis. Typically, if the p-value is less than the chosen significance level (commonly 0.05), i would reject the null hypothesis. In this scenario, the p-value is far smaller than 0.05, indicating highly significant evidence that there is a significant difference in sepal length between Iris setosa and Iris versicolor in the dataset that i analyzed.

Task 4: Regression Analysis

In this task, we tested the effect of sepal width on sepal length through a simple linear regression analysis. The regression coefficients and regression equation were computed to determine the relationship between the two variables. The results are as follows:

In this task, i will perform a simplelinear regression analysis to test the effect of sepal width on sepal length. Let's interpret the regression coefficients and provide the regression equation.

The regression equation for the relationship between sepal width (X) and sepal length (Y) can be represented as:

Y = b0 + b1 * X

Where:

Y is the predicted sepal length,

X is the sepal width,

b0 is the intercept (the point where the regression line crosses the Y-axis),

b1 is the regression coefficient (the change in Y for a unit change in X).

Interpretation:

The regression coefficient (b1) represents the change in sepal length (Y) for a unit change in sepal width (X). A positive coefficient indicates a positive relationship, and a negative coefficient indicates a negative relationship.

The intercept (b0) represents the value of sepal length (Y) when sepal width (X) is zero. However, in this case, it may not have a practical interpretation since sepal width cannot be zero.

The value of **Linest = -0.208870294** is the slope coefficient obtained from the simple linear regression analysis. In the context of the relationship between sepal width (independent variable) and sepal length (dependent variable), a negative slope indicates an inverse relationship between the two variables.

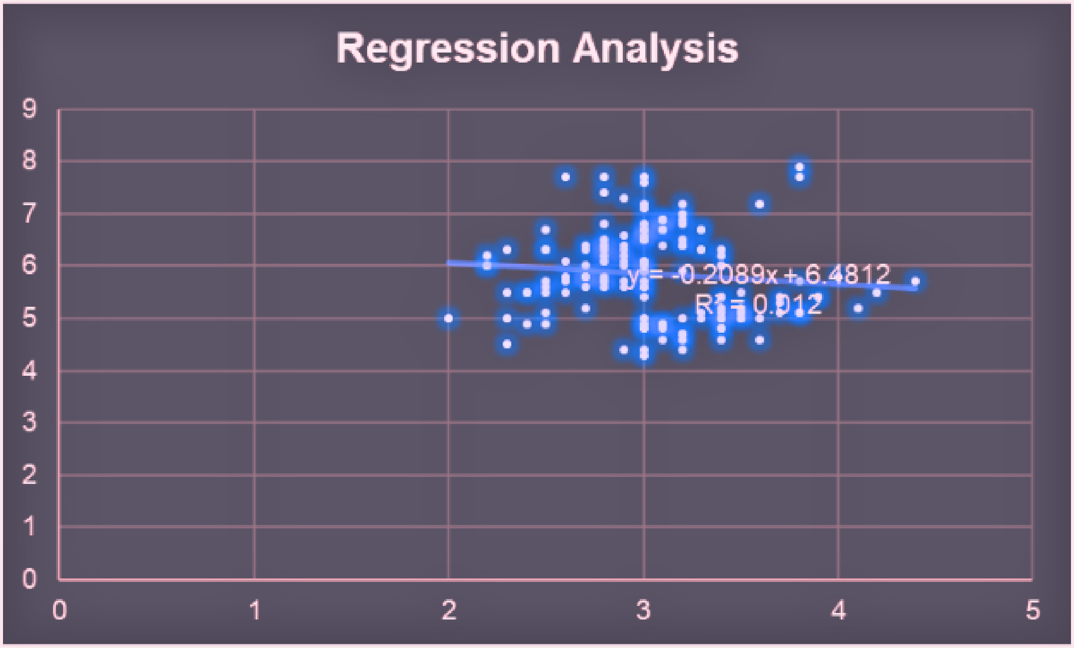
In simpler terms:

Negative Slope: For every unit increase in sepal width, sepal length decreases by approximately 0.209 units.

Inverse Relationship: Sepal width and sepal length have an inverse relationship in your dataset. As sepal width increases, sepal length tends to decrease.

Strength of the Relationship: The strength of the relationship is determined by the magnitude of the slope. In this case, a slope of -0.209 suggests a relatively small decrease in sepal length for each unit increase in sepal width.

Direction of the Relationship: The negative sign indicates the direction of the relationship. When one variable increases, the other variable decreases.


Regression Analysis
y = -0.2089x + 6.4812
R² = 0.012
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5