Iris Dataset Analytical Report

November 27th, 2023

Conducted By:
SapphireGaze

Table of Contents

1. Executive Summary

The purpose of this report is to analyze and depict the relationships within the Iris dataset. With that goal in mind, I was able to use various tools and techniques, such as PostgreSQL and pgAdmin, to analyze the Iris dataset and discover an underlying positive linear relationship between the variable pair petal length and petal width.

This analysis required quite a few stages to depict the nuanced relationship between the Iris statistics. First, it was necessary to integrate the Iris dataset into a PostgreSQL table in order to perform further analysis. Data preparation is then required to ensure the validity of the data provided, through the data preparation process, I was able to eliminate any existing rows with NULL values and converted the data types to be more convenient for future analysis.

The next phase in the analysis process is the preliminary analysis. I was able to use the summary statistics, as well as the Pearson Correlation Coefficient to determine the relationship between various data pairs. This proved to be an important stepping stone for future modeling as it narrowed down the analysis to a specific variable pair, petal length and petal width.

Modeling is the very last step of the data analysis process, with the focus now on a specific variable pair, I was able to employ various analytical techniques, such as linear regression/residual analysis, Coefficient of Determination, as well as the Root Mean Square Error, to come to the conclusion that there is an apparent positive linear correlation between the variable pair, petal length and petal width. Through this process, I was even able to discover interesting relations between the Iris statistics and species, which might warrant potential future analysis.

2. Introduction

The Iris dataset is a well-known and frequently used dataset in the field of machine learning and statistics. It was introduced by the British biologist and statistician Ronald A. Fisher in 1936 as an example of discriminant analysis. The dataset consists of measurements of sepal length, sepal width, petal length, and petal width for 150 iris flowers, representing three different species: Iris-setosa, Iris-versicolor, and Iris-virginica. Each species comprises 50 samples.

The Iris dataset has become a benchmark in the evaluation of various machine learning algorithms, particularly in the realm of classification. Its simplicity and clarity make it an ideal starting point for exploring and implementing data analysis and machine learning techniques. Researchers and practitioners often use the Iris dataset to demonstrate concepts and methods due to its balanced and easily interpretable nature.

For this analysis, I have chosen PostgreSQL as the database system and pgAdmin as the editor. PostgreSQL is a powerful open-source relational database management system known for its extensibility, reliability, and robust performance. It provides advanced features for handling complex data types and is widely used in various industries for managing large datasets.

By leveraging PostgreSQL with pgAdmin, this analysis aims to showcase the versatility and capabilities of open-source tools in handling and exploring the Iris dataset to discover the relations between various Iris statistics, with potential applications extending to more complex datasets in the future.

3. Data Preparation

In the initial phase of my analysis, I conducted a comprehensive analysis of the Iris dataset. This exploration revealed that the dataset consists of 150 rows and 5 columns, with four columns containing numerical features (sepal_length, sepal_width, petal_length, and petal_width) and one column for the text based categorical feature species.

One critical aspect of data preparation involves handling NULL values. Fortunately, the Iris dataset showed no instances of such NULL values, eliminating the need for removing invalid rows. Additionally, in order to better prepare for future analysis, I decided to convert the data types of the numerical columns from REAL to NUMERIC. I further delved into the influence of different species on the distribution of data. Skewness, a measure of distribution asymmetry, was calculated for sepal_length, sepal_width, petal_length, and petal_width, grouped by species. This analysis shed light on how species-specific characteristics impact the shape of the data distribution.

Moreover, to gain a nuanced understanding of the dataset, summary statistics were computed for each column, including mean, median, mode, minimum, maximum, range, standard deviation, variance, Q1, Q3, IQR, and skewness. This granular exploration allowed me to discern the unique characteristics of each species within the dataset.

Visual representations in the form of histograms were generated to complement my analysis throughout the data preparation phase, providing an intuitive view of data concentration and potential outliers. These histograms focused on sepal and petal length and width, unveiling distinct patterns for each species.

4. Preliminary Analysis

Early on in my investigation of the Iris dataset, I discovered important information that prepared the way for deeper analysis. Finding and eliminating significant outliers from each column was a crucial step in this procedure to make sure that extreme data points wouldn't excessively affect the results of the analysis. I decided to only remove the strong outliers as other outliers can easily be a natural variation within the Iris statistics, resulting in only one set of data points being removed, maintaining the dataset's integrity.

The analysis of the Pearson Correlation Coefficients (R values) between variable pairs proved to be a crucial discovery. Notably, I observed a low negative linear relationship between sepal length and sepal width, emphasizing their inverse proportionality. Moreover, a moderate negative linear relationship was indicated between sepal width and petal width through the analysis, shedding light on their relationships. In contrast, there was a strong positive linear relationship between petal length and petal width, as well as sepal length and petal length, aligning with my initial hypothesis regarding the linear interdependence of sepal and petal length statistics.

The exploration extended to the distribution characteristics of variable pairs within different Iris species. For instance, petal length demonstrated symmetry for iris-setosa, negative skewness for iris-versicolor, and positive skewness for iris-virginica. This understanding of the variation of distribution patterns provided valuable insights into the characteristics of each species.

The visual representations, including boxplots and heat maps, offered a comprehensive view of the dataset's characteristics. Boxplots illustrated the range of

data for different species, confirming my hypothesis that iris-setosa tends to have the lowest range, while iris-virginica exhibits the highest. Surprisingly, in sepal width statistics, iris-setosa demonstrated a higher range than both iris-versicolor and iris-virginica, introducing an interesting exception to the observed patterns, which proved to be a fascinating pattern that might warrant further investigation.

Through the preliminary analysis done, I was able to focus the efforts on the variable pair of particular interest, petal length and petal width. Their robust positive linear correlation indicated their potential significance in subsequent modeling and interpretation.

5. Modeling

Through the modeling process, the focus shifts from the entirety of the Iris dataset to individual Iris species, specifically Iris-setosa, Iris-versicolor, and Iris-virginica, to better understand the nuances in the linear relationship of the variable pair petal length and petal width. The following sections are different analytical methods employed in order to provide relevant insights into the relationships for each species.

Slope/Intercept and Linear Regression Analysis:

- For Iris-setosa, the regression analysis revealed a distinctive pattern. The slope and intercept, indicating a linear relationship between petal length and petal width, showed a unique distribution compared to the other two species.
- Moving to Iris-versicolor, the linear model exhibited a strong correlation between petal length and petal width. The slope and intercept values confirmed a consistent and positive linear relationship.
- Iris-virginica presented intriguing results. While maintaining a positive correlation, the slope was less steep, suggesting that a more significant increase in petal length is required for a substantial rise in petal width compared to the other species.

Residuals Analysis:

- The residual plots for each species visually depicted the variance from the regression line. Iris-setosa showcased a more scattered distribution, indicating higher variability in the relationship between petal length and petal width.
- Both Iris-versicolor and Iris-virginica displayed densely populated data points close to the regression line, underscoring a robust linear relationship. However,

subtle differences, such as the slope's steepness, appeared which further emphasized the species-specific characteristics.

Coefficient of Determination (R^2):

- Examining the R^2 values reinforced the overall trend observed in the previous portions of the analysis. For the entire Iris dataset, petal length and petal width exhibited the highest R^2 values, affirming the strength of the linear relationship.
- Surprisingly, when dissecting the species individually, distinct variable pairs emerged. Iris-setosa favored sepal length and sepal width over petal dimensions, Iris-versicolor upheld the dominance of petal length and petal width, and Iris-virginica leaned toward sepal length and petal length.

Root Mean Square Error (RMSE):

- The RMSE calculations offered a quantitative measure of the model's accuracy. Iris-setosa, with its dispersed residuals, displayed a slightly higher RMSE, aligning with the visual impression of higher variability.
- Iris-versicolor and Iris-virginica showcased lower RMSE values, signifying a more accurate fit of the linear regression model. The subtle distinctions in slope and intercept contributed to the variations in RMSE across species.

Visualization:

- The linear regression lines along with the residual plots histograms illustrate the regression lines and residuals for each species. The diverse patterns in scatterplots highlight the species-specific characteristics, with Iris-setosa deviating more significantly.
- The R^2 values between the variable pair petal length and petal width for each Iris species is depicted through heatmaps. While petal length and petal width

dominate for the overall dataset, the species-level analyses unearth unique preferences for variable pairs, underscoring the influence of species on the linear relationships.

In essence, this modeling section of the analysis not only validated the overall linear relationship between petal length and petal width but also uncovered species-specific variations. The examination of slopes, intercepts, residuals, and different evaluation metrics adds to our understanding, providing further insights of the Iris dataset.

6. Conclusions

From my analysis on the Iris dataset, I was able to come to the conclusion that there is a probable positive linear relationship between the variable pair petal length and petal width for the entire Iris dataset. This was depicted through several factors, such as the negative skewness presented through the summary statistics and the densely populated data points along the linear regression line (as seen in figure 1 below).

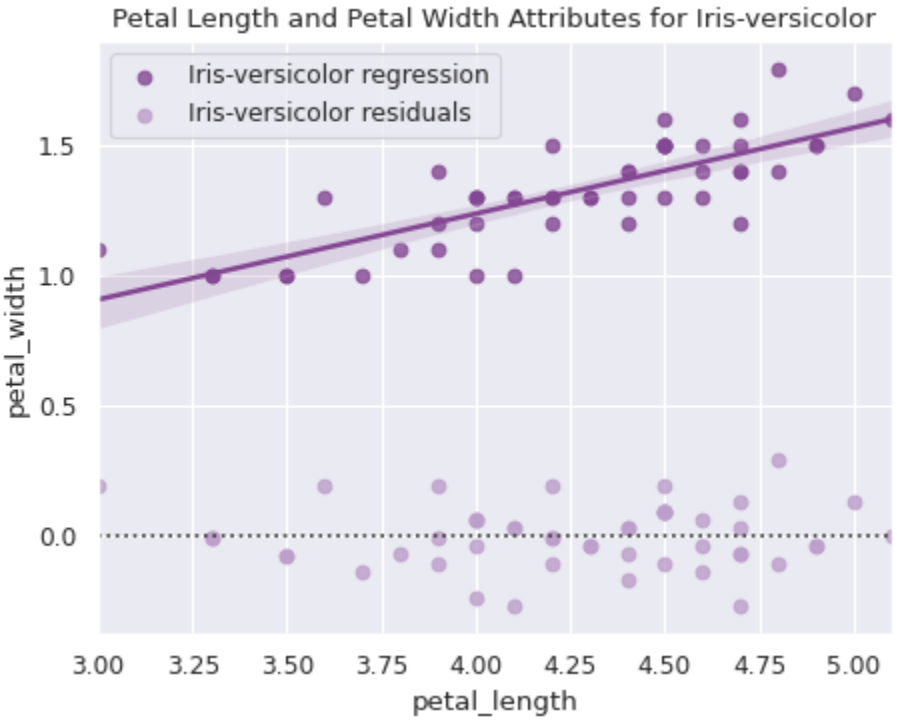

Figure 1. *Petal Length and Petal Width Attributes for Iris-versicolor*

Through the listed factors, it is likely that the said variable pair resembles a positive linear relationship as the increase in petal length leads to a direct increase in petal width. This relationship is further consolidated by the analysis of the Coefficient of Determination and the Root Mean Square Error (RMSE), as the Coefficient of Determination of the Iris dataset produced a value of 0.926, illustrating that the variations in the data points are a close fit for the linear regression model as the number approaches 1. The RMSE's ensued values are virtually 0 for each of the Iris species. The

analysis showed RMSE of 0.099 for Iris-setosa, 0.121 for Iris-versicolor, and 0.257 for Iris-virginica, demonstrating there are very few errors within the prediction of the model.

However, there are things that came to my attention during the analysis of the Iris dataset. Even though the variable pair petal length and petal width retains the most evident linear relationship for the entirety of the dataset, that is not the case if one were to analyze the data by breaking down the dataset into individual species. It appears that when the Iris dataset is broken down and analyzed by species, there are often other variable pairs that take precedence over the variable pair petal length and petal width in terms of linear relationships. Take the species Iris-setosa for example, the variable pair sepal length and sepal width has a significantly higher Coefficient of Determination when compared to the variable pair petal length and petal width (as seen below in figure 2), which is a possible indication that the sepal length/width variable pair taking precedence over petal length/width in regards to their linear relationship.

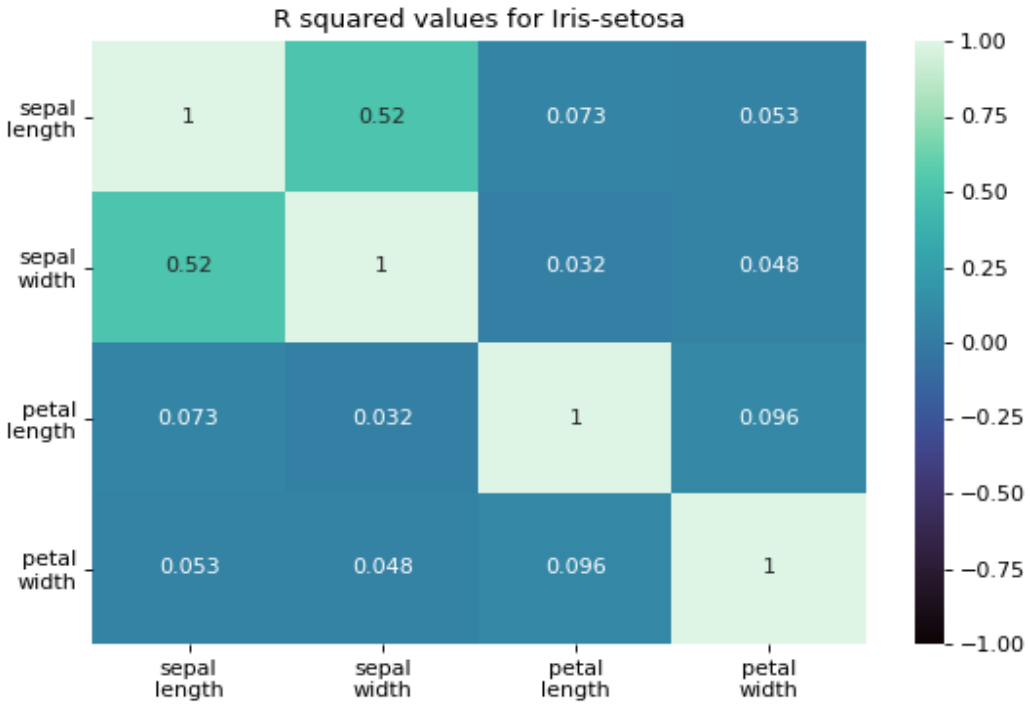

Figure 2. *R squared values for Iris-setosa*

Based on the information present from the analysis, the relationship of the variable pairs seems to be dependent on the species of Iris, which led me to believe that species plays an essential role in the development of Iris in terms of petal and sepal attributes. However, I cannot make a definitive conclusion regarding the impact of species in terms of petal/sepal relationships with the current data and analysis available, but this could be a possible next step for future research and analysis.

7. References

1. MathNerd. “Iris Flower Dataset.” *Kaggle*, 22 Mar. 2018, www.kaggle.com/datasets/arshid/iris-flower-dataset.

2. “Iris Flower Data Set.” *Wikipedia*, Wikimedia Foundation, 30 Nov. 2023, en.wikipedia.org/wiki/Iris_flower_data_set.

3. Bozkus, Emine. “Exploring the Iris Flower Dataset.” *Medium*, Medium, 19 Dec. 2022, eminebozkus.medium.com/exploring-the-iris-flower-dataset-4e000bcc266c.

8. Appendices

```
-- summary statistics query
WITH values AS (
      SELECT
            ROUND(AVG(sepal_length)::NUMERIC, 2) AS mean,
            ROUND(PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY
sepal_length)::NUMERIC, 2) AS median,
            ROUND(MODE() WITHIN GROUP (ORDER BY
sepal_length)::NUMERIC, 2) AS mode,
            ROUND(MIN(sepal_length)::NUMERIC, 2) AS minimum,
            ROUND(MAX(sepal_length)::NUMERIC, 2) AS maximum,
            ROUND(STDDEV(sepal_length)::NUMERIC, 2) AS
standard_deviation,
            ROUND(VARIANCE(sepal_length)::NUMERIC, 2) AS variance,
            ROUND(PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY
sepal_length)::NUMERIC, 2) AS q1,
            ROUND(PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY
sepal_length)::NUMERIC, 2) AS q3
      FROM public.iris
), combined_table AS (
      SELECT 1 AS sno, 'mean' AS statistic, mean AS value FROM values
      UNION ALL
      SELECT 2 AS sno, 'median' AS statistic, median AS value FROM
values
      UNION ALL
      SELECT 3 AS sno, 'mode' AS statistic, mode AS value FROM values
      UNION ALL
      SELECT 4 AS sno, 'minimum' AS statistic, minimum AS value FROM
values
      UNION ALL
      SELECT 5 AS sno, 'maximum' AS statistic, maximum AS value FROM
values
      UNION ALL
      SELECT 6 AS sno, 'range' AS statistic, maximum - minimum AS
value FROM values
      UNION ALL
      SELECT 7 AS sno, 'standard deviation' AS statistic,
```

```
standard_deviation AS value FROM values
      UNION ALL
      SELECT 8 AS sno, 'variance' AS statistic, variance AS value
FROM values
      UNION ALL
      SELECT 9 AS sno, 'Q1' AS statistic, q1 AS value FROM values
      UNION ALL
      SELECT 10 AS sno, 'Q3' AS statistic, q3 AS value FROM values
      UNION ALL
      SELECT 11 AS sno, 'IQR' AS statistic, q3 - q1 AS value FROM
values
      UNION ALL
      SELECT 12 AS sno, 'skewness' AS statistic, ROUND(3 * (mean -
median) / standard_deviation, 2) AS value FROM values
)
```

```
-- skewness analysis grouped by the species
SELECT
      species,
      ROUND(3 * (ROUND(AVG(sepal_length), 2) - ROUND(PERCENTILE_CONT(0.5)
WITHIN GROUP (ORDER BY sepal_length)::NUMERIC, 2)) /
ROUND(STDDEV(sepal_length), 2), 2) AS sepal_length_skewness,
      ROUND(3 * (ROUND(AVG(sepal_width), 2) - ROUND(PERCENTILE_CONT(0.5) WITHIN
GROUP (ORDER BY sepal_width)::NUMERIC, 2)) / ROUND(STDDEV(sepal_width), 2), 2)
AS sepal_width_skewness,
      ROUND(3 * (ROUND(AVG(petal_length), 2) - ROUND(PERCENTILE_CONT(0.5)
WITHIN GROUP (ORDER BY petal_length)::NUMERIC, 2)) /
ROUND(STDDEV(petal_length), 2), 2) AS petal_length_skewness,
      ROUND(3 * (ROUND(AVG(petal_width), 2) - ROUND(PERCENTILE_CONT(0.5) WITHIN
GROUP (ORDER BY petal_width)::NUMERIC, 2)) / ROUND(STDDEV(petal_width), 2), 2)
AS petal_width_skewness
FROM public.iris
GROUP BY species;
```

```
-- create Pearson Correlation Coefficient (R value) matrix based on
variable pairs
SELECT CORR(sepal_length, sepal_width) AS
sepal_length_vs_sepal_width,
      CORR(petal_length, petal_width) AS petal_length_vs_petal_width,
      CORR(sepal_length, petal_length) AS
```

```
sepal_length_vs_petal_length,
     CORR(sepal_width, petal_width) AS sepal_width_vs_petal_width
FROM public.iris;
```

```
-- query for calculate linear regression, residuals/residuals squared
using petal_length and petal_width as x and y
WITH setosa_values AS (
     SELECT REGR_SLOPE(petal_width, petal_length) AS slope,
          REGR_INTERCEPT(petal_width, petal_length) AS intercept
     FROM public.iris
     WHERE species = 'Iris-setosa'
)
SELECT
     petal_length,
     petal_width,
     setosa_values.slope * petal_length + setosa_values.intercept AS
regression,
     petal_width - (setosa_values.slope * petal_length +
setosa_values.intercept) AS residuals,
     POWER(petal_width - (setosa_values.slope * petal_length +
setosa_values.intercept), 2) AS squared_residuals
FROM public.iris, setosa_values
WHERE species = 'Iris-setosa';
```