Flower Data Analysis

Mini Project

04/22/2024

IST652 - Scripting for Data Analysis

Shubham Gaikwad

216327540

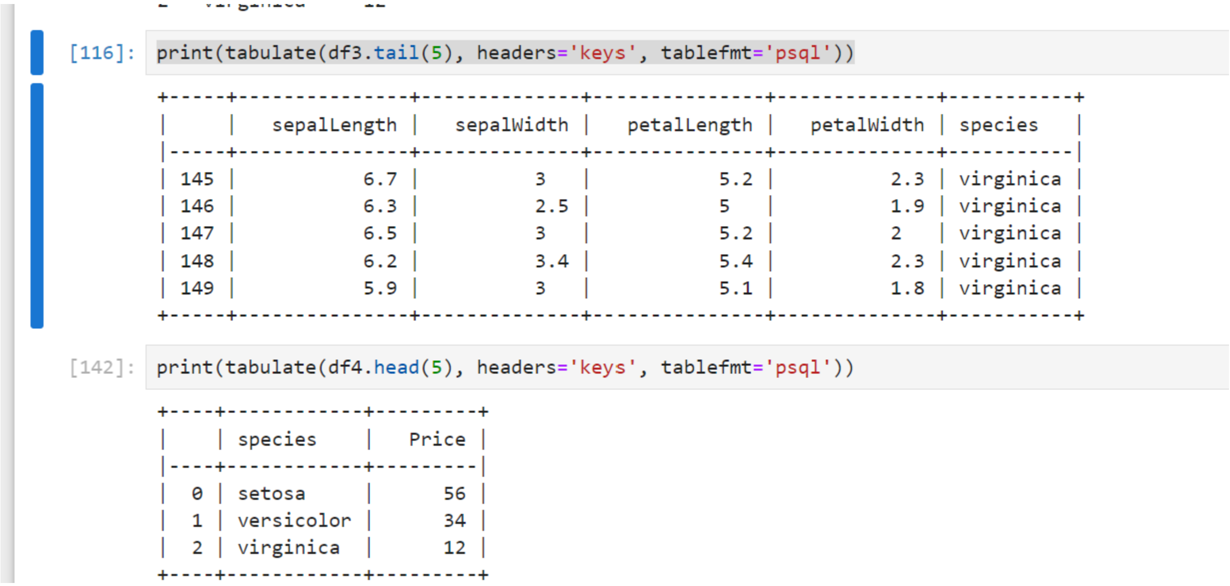
Project Objectives

- Evaluate the relationship between the physical characteristics of Iris flowers and their market prices.
- Identify patterns and trends in Iris flower pricing based on structural features across different species.
- Perform advanced visualization to understand features of data

Project Agenda

- Project Kickoff
- Data Collection
- Data Cleaning and Preparation
- Exploratory Data Analysis (EDA)
- Insights and Reporting
- Project Review and Closure
- Future Work and Recommendations

Data Summary



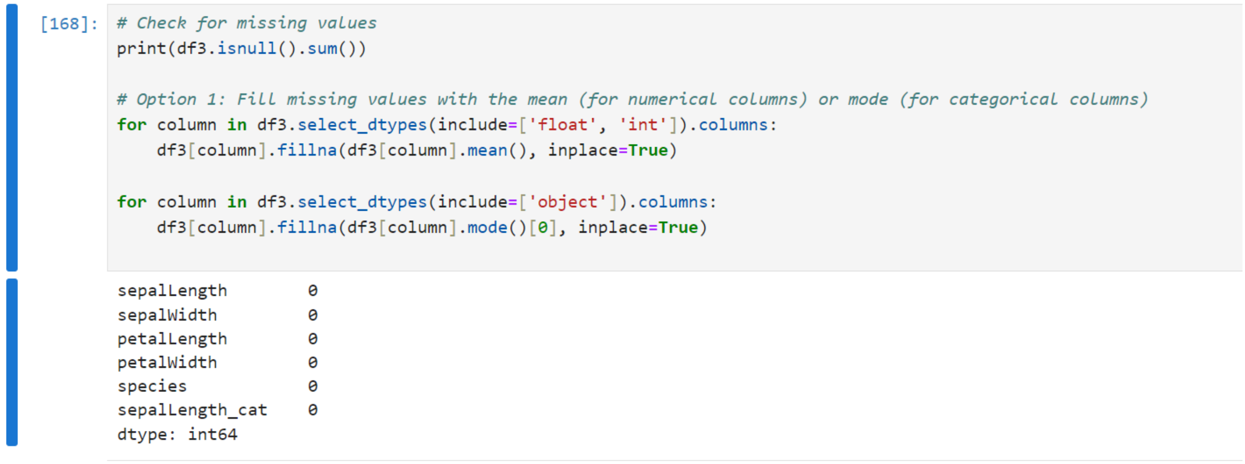
```
[116]: print(tabulate(df3.tail(5), headers='keys', tablefmt='psql'))
```

```
+-----+---------------+--------------+---------------+--------------+-----------+
|     |   sepalLength |   sepalWidth |   petalLength |   petalWidth | species   |
|-----+---------------+--------------+---------------+--------------+-----------|
| 145 |           6.7 |          3   |           5.2 |          2.3 | virginica |
| 146 |           6.3 |          2.5 |           5   |          1.9 | virginica |
| 147 |           6.5 |          3   |           5.2 |          2   | virginica |
| 148 |           6.2 |          3.4 |           5.4 |          2.3 | virginica |
| 149 |           5.9 |          3   |           5.1 |          1.8 | virginica |
+-----+---------------+--------------+---------------+--------------+-----------+
```

```
[142]: print(tabulate(df4.head(5), headers='keys', tablefmt='psql'))
```

```
+----+------------+---------+
|    | species    |   Price |
|----+------------+---------|
|  0 | setosa     |      56 |
|  1 | versicolor |      34 |
|  2 | virginica  |      12 |
+----+------------+---------+
```

Data Cleaning and Preparation



[168]:

```
# Check for missing values
print(df3.isnull().sum())

# Option 1: Fill missing values with the mean (for numerical columns) or mode (for categorical columns)
for column in df3.select_dtypes(include=['float', 'int']).columns:
    df3[column].fillna(df3[column].mean(), inplace=True)

for column in df3.select_dtypes(include=['object']).columns:
    df3[column].fillna(df3[column].mode()[0], inplace=True)
```

```
sepalLength        0
sepalWidth         0
petalLength        0
petalWidth         0
species            0
sepalLength_cat    0
dtype: int64
```

In this phase, I meticulously cleaned and prepared the datasets for analysis. The process involved several key steps:

Removing Duplicates: I eliminated any duplicate entries to ensure the integrity of the dataset, focusing on unique identifiers within the Iris dataset.

Handling Missing Values: I addressed missing values either by imputation or removal, depending on their impact on the overall dataset integrity and the subsequent analysis.

Data Type Conversion: I ensured that all data types were consistent with the expected format for analysis; for example, converting categorical data into a format suitable for modeling.

Merging Data: I merged the Iris structural dataset and the pricing data based on common attributes to facilitate a comprehensive analysis. This step was critical to align the physical characteristics with their respective prices.

Exploratory Of Data –

The EDA was aimed at uncovering patterns, anomalies, relationships, and trends in the data through the following methodologies:

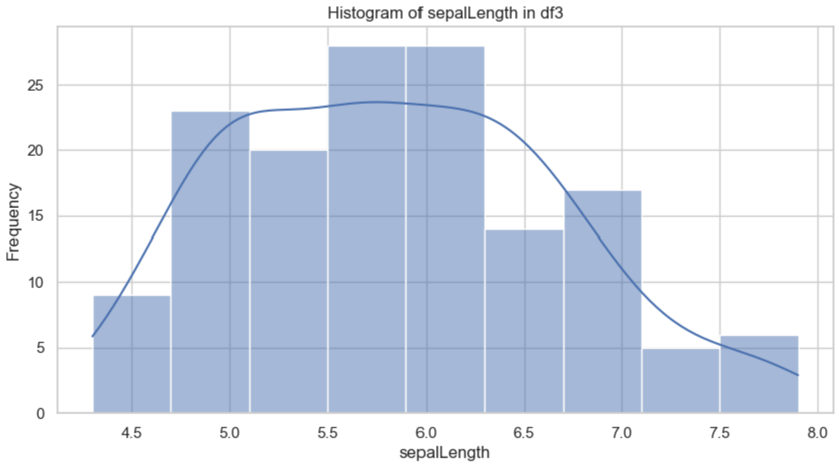
Statistical Summaries: I provided summaries that gave insights into the central tendency and variability of the data, helping to understand distributions of key variables.

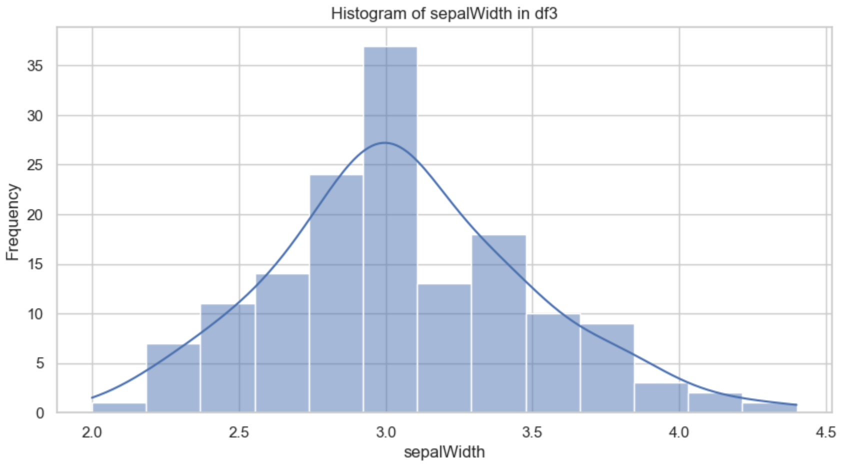
Visualization: I used various plots such as histograms, scatter plots, and box plots to visualize the data distributions and relationships between variables. For instance, histograms illustrated the distribution of flower prices and structural dimensions, while scatter plots helped to identify the correlation between flower size and its price.

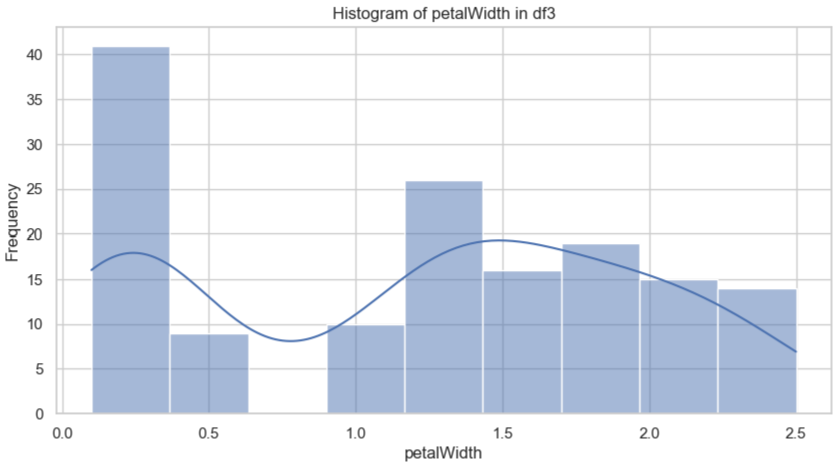
Correlation Analysis: I examined the correlations between different structural features of the flowers and their prices to identify the most influential factors driving the prices.

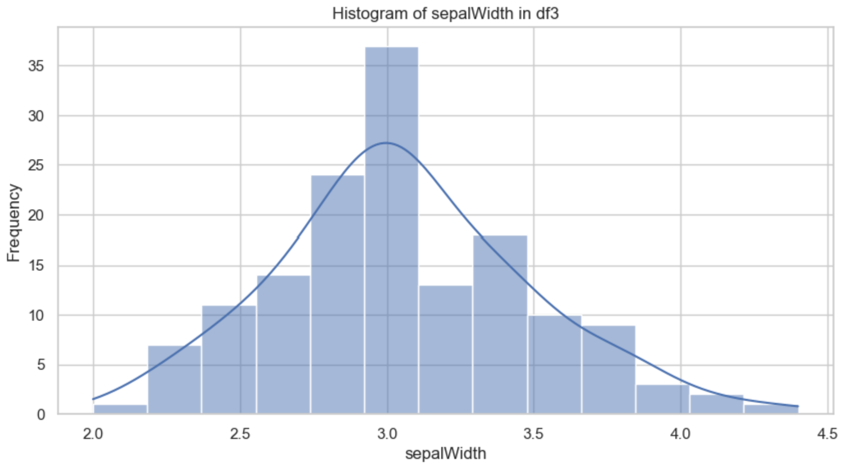
These explorations were instrumental in guiding the subsequent phases of modeling by providing a deep understanding of the data's characteristics and the relationships within.

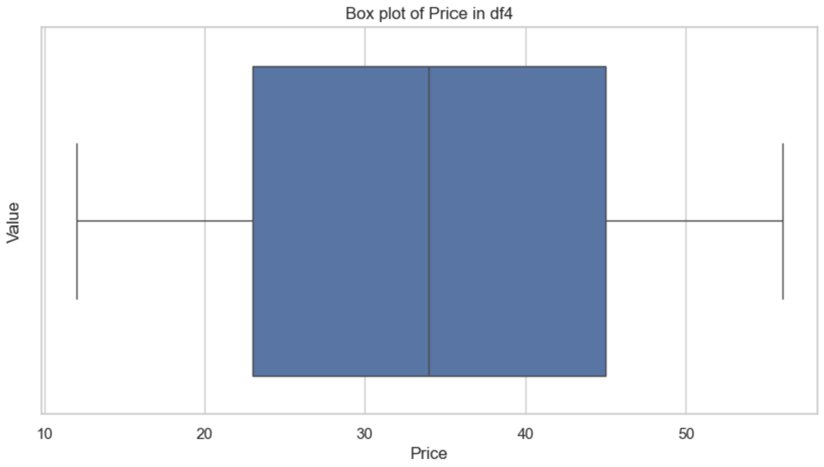
Histograms

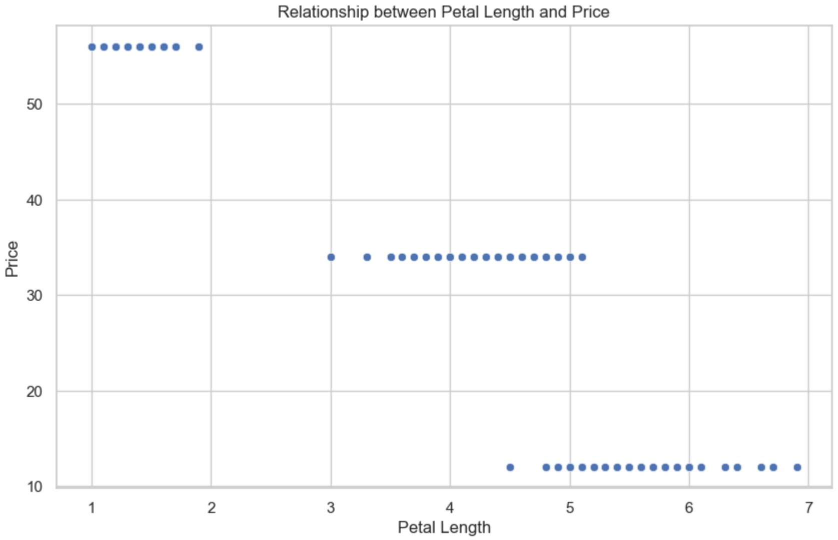


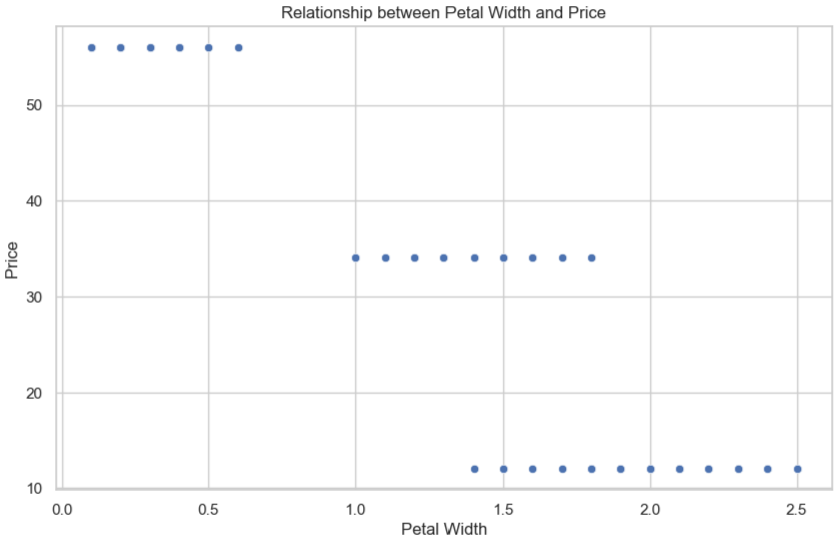

Histogram of sepalWidth in df3
Frequency
0
5
10
15
20
25
30
35
2.0
2.5
3.0
3.5
4.0
4.5
sepalWidth


Histogram of petalWidth in df3
Frequency
0
5
10
15
20
25
30
35
40
0.0
0.5
1.0
1.5
2.0
2.5
petalWidth

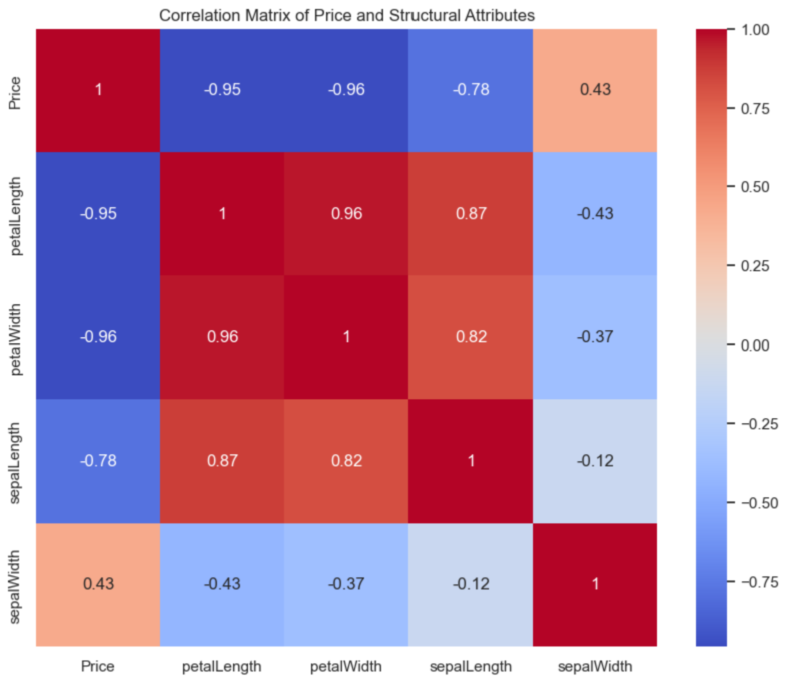

Histogram of sepalWidth in df3
Frequency
0
5
10
15
20
25
30
35
2.0
2.5
3.0
3.5
4.0
4.5
sepalWidth


Box plot of Price in df4
Value
10
20
30
40
50
Price

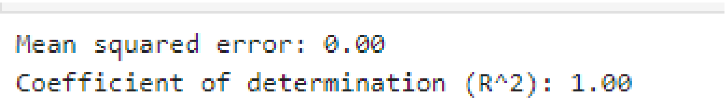

Relationship between Petal Length and Price
Price
10
20
30
40
50
1
2
3
4
5
6
7
Petal Length


Relationship between Petal Width and Price
Price
10
20
30
40
50
0.0
0.5
1.0
1.5
2.0
2.5
Petal Width

Understanding Correlation



Model Prediction [Linear Regression Model]



```
Mean squared error: 0.00
Coefficient of determination (R^2): 1.00
```

In the predictive modeling phase, I applied a Linear Regression model to predict Iris flower prices based on their structural attributes. Here’s the approach I took:

Model Selection: Linear Regression was chosen due to its efficacy in understanding the linear relationships between independent variables (flower structure) and the dependent variable (price).

Training the Model: I trained the model using a subset of the dataset, ensuring it could generalize well to new, unseen data.

Model Evaluation: I assessed the model's performance using standard metrics like R-squared and Mean Squared Error (MSE) on a validation set. These metrics provided insights into the accuracy and predictive power of the model.

Interpretation of Results: I interpreted the coefficients of the model to understand the influence of each structural attribute on the pricing. This analysis provided actionable insights that could be used for strategic pricing decisions in the floriculture market.

The predictive model not only highlighted the key predictors of price but also quantified their impact, offering valuable insights into how Iris flowers are valued in the market.

Each of these sections contributes to a thorough and insightful report that not only meets academic standards but also provides practical value to stakeholders interested in the floriculture industry.