



IRIS FLOWER DATASET

An overview



TABLE OF CONTENTS

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



PROJECT OVERVIEW

▶	Project overview.....	03
▶	Dataset overview.....	08
▶	Analysis	12
▶	Data cleaning.....	12
▶	Descriptive analysis.....	15
▶	Statistical analysis.....	21
▶	Machine learning.....	24
▶	Summary	26
▶	References.....	32



PROJECT AIMS

The aim of this project is to use the [Iris Flower Dataset](#) to showcase an [example data analysis project](#) for beginner analysts.

To be used in the blog series:

[\[Data Analysis Project\] Iris Flower Dataset \(9 steps\)](#)

01. To find and [import](#) a dataset from Kaggle.

02. To [clean](#) the data by checking for: duplicates, nulls and outliers.

03. To [visualise](#) the data to aid in data analysis & data exploration.

04. To [statistically analyse](#) the data to see if our findings are statistically significant.

05. To use [machine learning](#) to predict the species of flower given unseen input data.



PROJECT PROCESS

Here is a visual representation of the project
in 5 steps

01.

kaggle

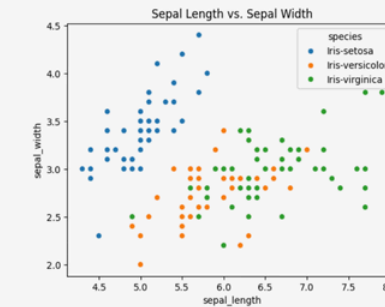
Import

02.

```
# Count duplicate values  
df.duplicated().sum()  
| ✓ 0.0s  
3
```

Clean

03.



Visualise

04.

(p-value: 0.0022585277836218586)

Statistically analyse

05.

```
▼ RandomForestClassifier  
RandomForestClassifier()
```

Machine learning



PROJECT QUESTIONS

To help structure the analysis, here are 9 questions we will investigate.

01. What is the distribution of each feature in the dataset?
02. What are the characteristics of each iris species?
03. How are the features correlated with each other?
04. Are there any outliers?
05. Are there any missing values?



PROJECT QUESTIONS

To help structure the analysis, here are 9 questions we will investigate.

06. Are the features on the same scale?

07. Which classification algorithm is most suitable for this task?

08. How will we assess the accuracy of our model?

09. How accurate is our model?



DATASET OVERVIEW

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



DATASET OVERVIEW

The [Iris dataset](#) is a well-known dataset in the field of machine learning and consists of measurements of [four features](#) of [three species](#) of iris flowers:

- Setosa
- Versicolor
- Virginica

IRIS_FLOWER_DATAFRAME

sepal_length	FLOAT
sepal_width	FLOAT
petal_length	FLOAT
petal_width	FLOAT
species	STRING

DATASET FEATURES

The dataset includes four features, which represent various dimensions of the iris flowers:

- Sepal Length (in centimeters)
- Sepal Width (in centimeters)
- Petal Length (in centimeters)
- Petal Width (in centimeters)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   sepal_length    150 non-null   float64
1   sepal_width     150 non-null   float64
2   petal_length    150 non-null   float64
3   petal_width     150 non-null   float64
4   species         150 non-null   object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

DATASET OBSERVATIONS

There are a total of 150 observations, with 50 samples from each species. There is no missing data.

Each observation includes the measurements of the four features and is labeled with the corresponding species.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   sepal_length    150 non-null    float64
1   sepal_width     150 non-null    float64
2   petal_length    150 non-null    float64
3   petal_width     150 non-null    float64
4   species         150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

ANALYSIS

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



Data cleaning

There are 5 columns of data:

- 4 numeric
- 1 categorical

There are 150 rows of data, with no null values in any of the columns

The min, max and standard deviation tell us that there are not likely to be many, if any, outliers.

- As the **{min + standard deviation}** & **{max – standard deviation}** values are close to the mean value

SUMMARY STATISTICS

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Data cleaning

1. Handling missing values

- There are no missing values in the dataset

2. Handling duplicates

- There are 3 duplicate rows
- The decision was taken not to remove them as (i) there aren't many of them and (ii) it's feasible for flowers to have the same characteristics — akin to multiple people having the same height and weight.

3. Converting data types

- The data types are already correct: **float** for the numeric values & **string** for the species column

4. Handling outliers

- There are no outliers in the dataset (Z-score ± 3)

01.

```
sepal_length    0
sepal_width     0
petal_length     0
petal_width     0
species         0
dtype: int64
```

Missing values

03.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   sepal_length 150 non-null    float64
1   sepal_width  150 non-null    float64
2   petal_length 150 non-null    float64
3   petal_width  150 non-null    float64
4   species      150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

Convert data types

02.

```
# Count duplicate values
df.duplicated().sum()
```

✓ 0.0s

3

Duplicates

04.

```
sepal_length  sepal_width  petal_length  petal_width  species  sum_lengths
```

Rows where the z-score is over 3 — in our case, none.

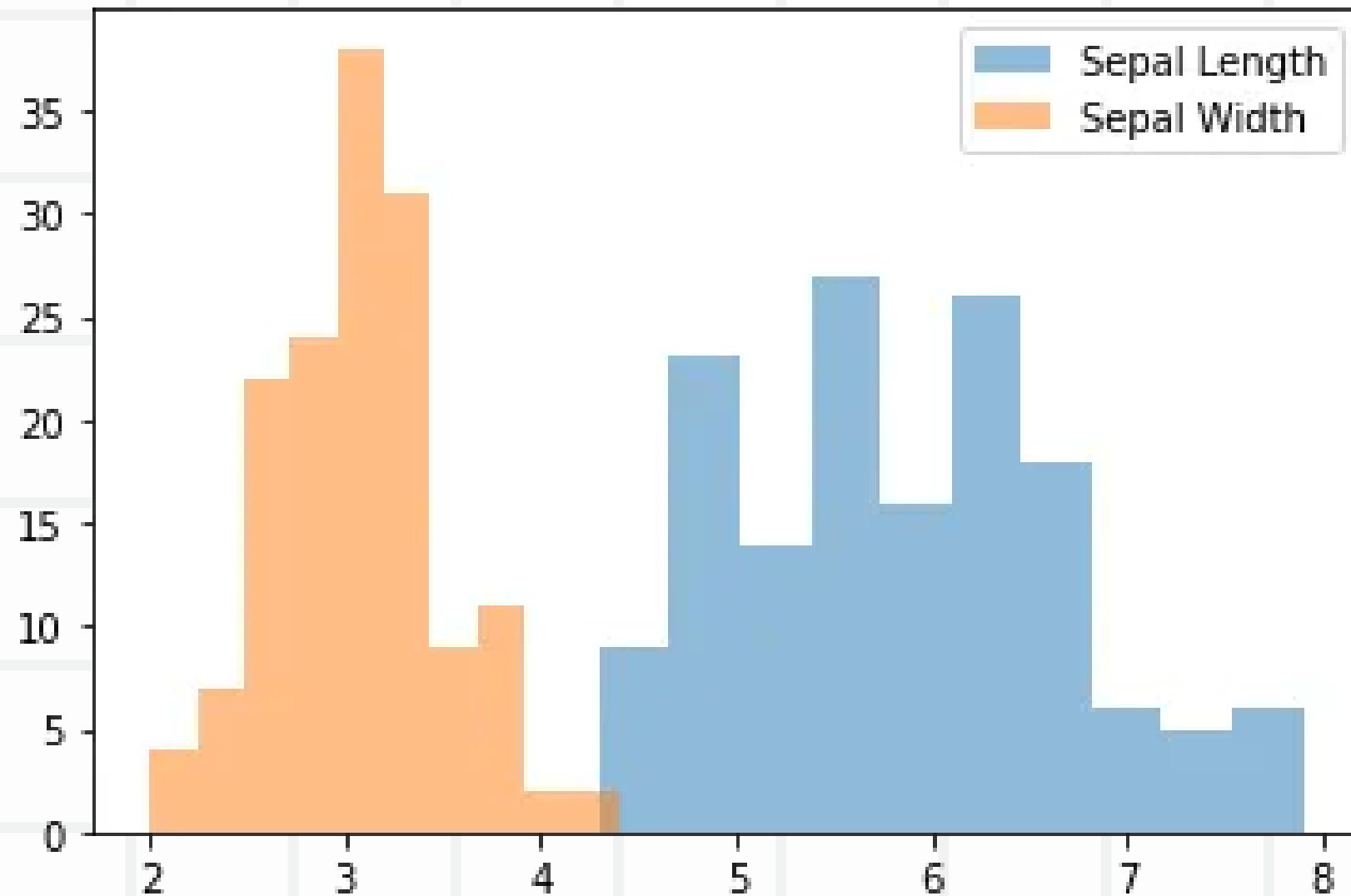
Outliers

ANALYSIS

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



SEPAL LENGTH vs. SEPAL WIDTH



Analysis: histogram

Histograms are useful for understanding the distribution of a variable. They display the frequency or density of data points within specified bins or intervals.

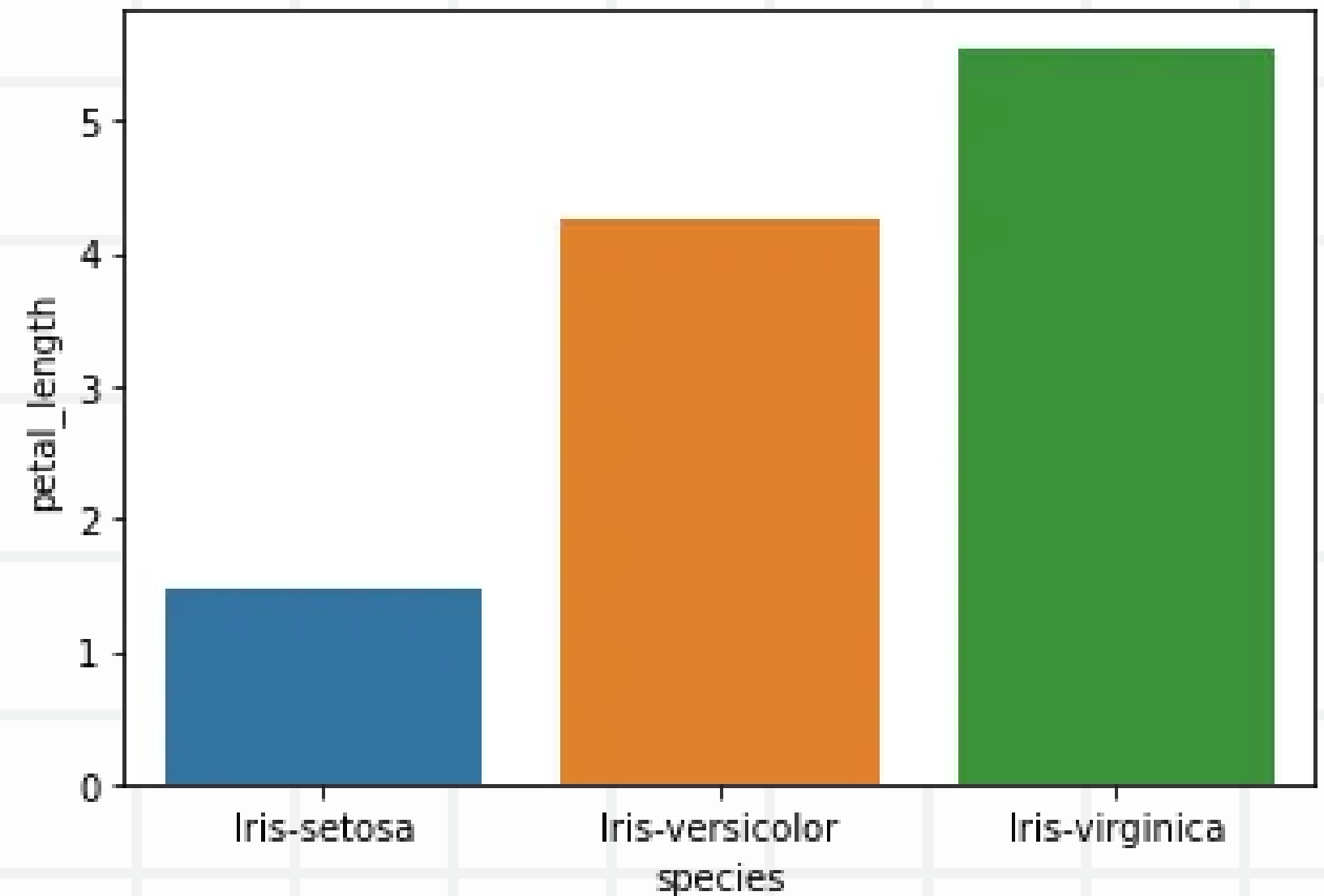
- Immediately, we can see that there is likely to be **distinct categories** or **groupings** to the data (i.e. different species in our case), as there are multiple peaks in the data (multi-modal distribution), particularly for the **sepal length**.
- **Sepal length** is almost always greater than **sepal width**.
- **Sepal length** has a larger range than **sepal width**.

Analysis: bar chart

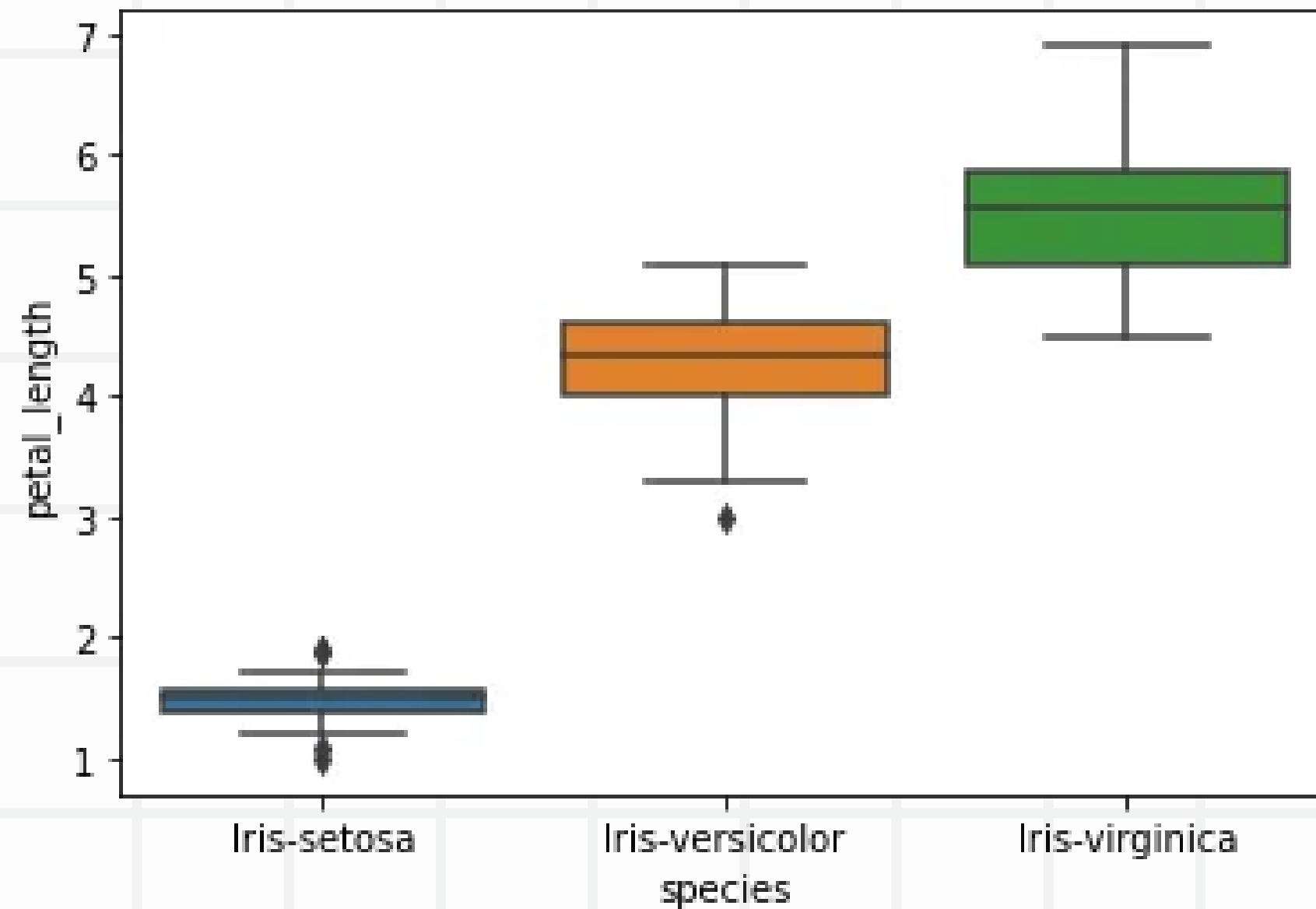
Bar charts are effective for comparing the **means** or **counts** of a numerical variable across different categories.

- **Setosa** has the lowest average petal length than the other species (~1.5cm).
- **Versicolor** has a lower average petal length than Virginica but a higher petal length mean than Virginica (~4.2cm).
- **Virginica** has the highest average petal length than the other species (~5.3cm).

AVERAGE PETAL LENGTH vs. SPECIES



PETAL LENGTH vs. SPECIES



Analysis: box plot

Box plots are ideal for visualizing the distribution of a numerical variable across different categories or groups. They display the **median**, **quartiles**, and **potential outliers**.

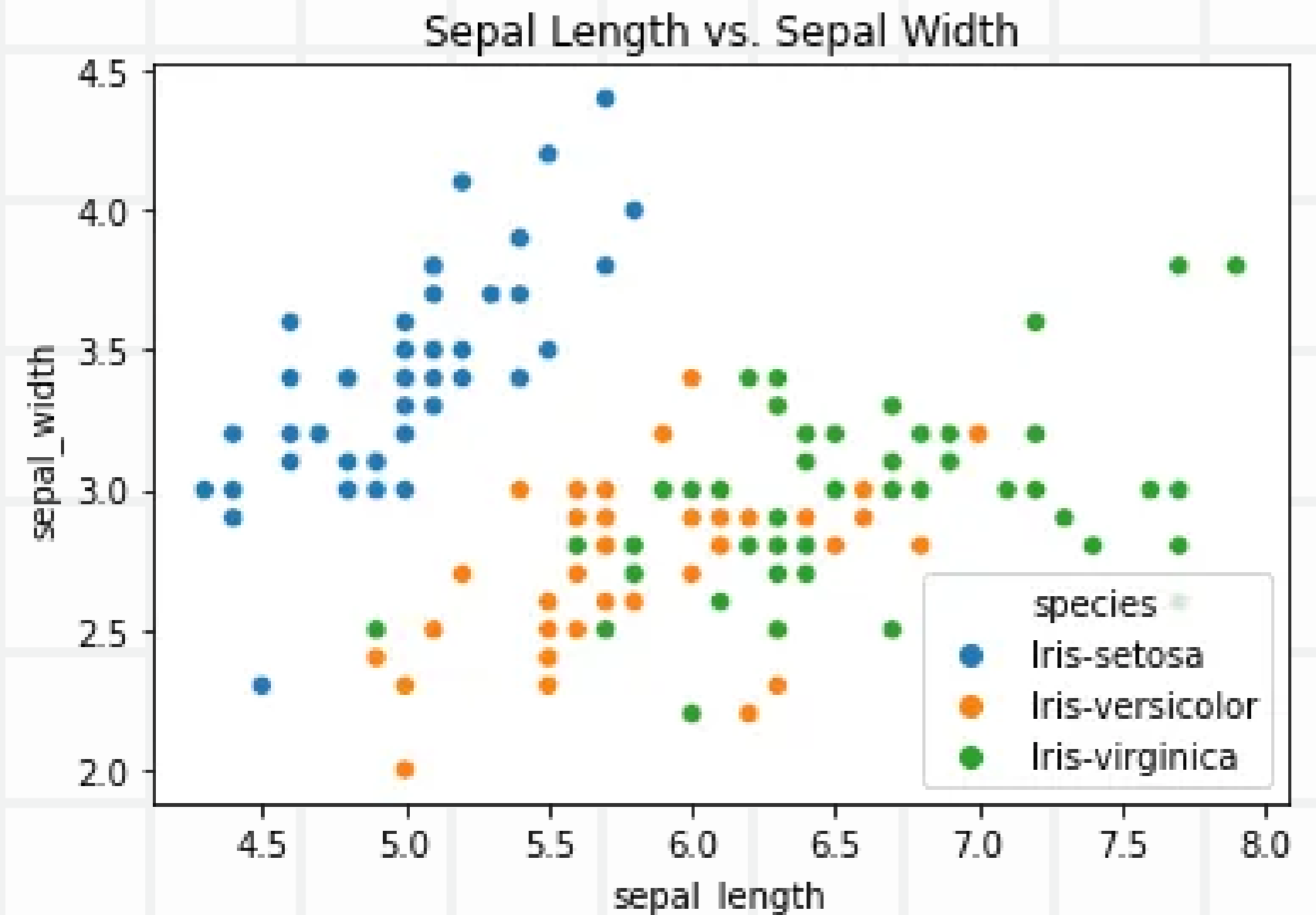
- There seems to be **very few outliers** for each species, with Setosa having the most.
- **Setosa** has the **smallest petal length** and a very small range.
- **Versicolor** and **Virginica's** petal length overlaps more.
 - However, Virginica is on average larger than Versicolor.

Analysis: scatter plot

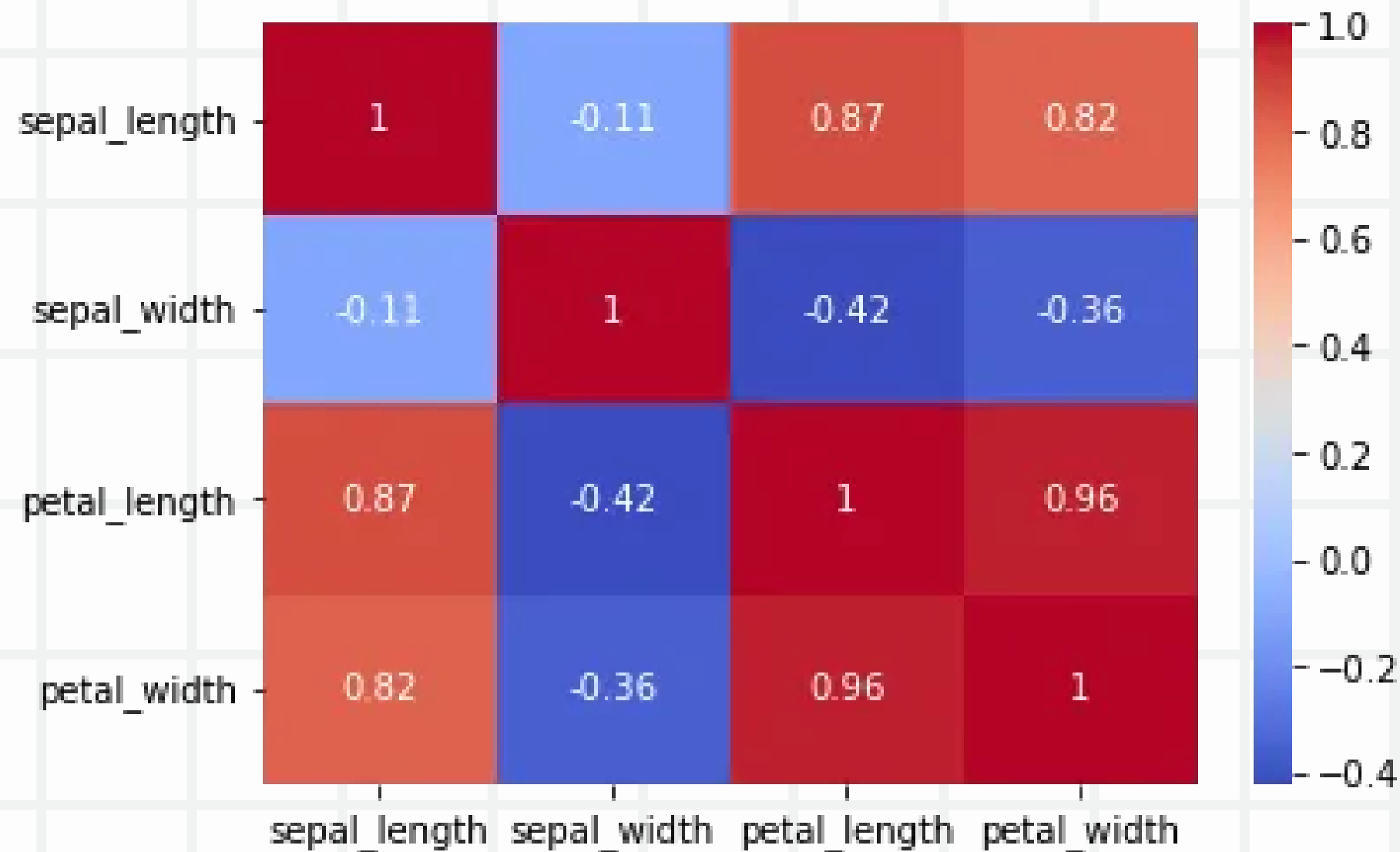
Scatter plots are excellent for visualizing the relationship between two continuous variables. They help you identify patterns, clusters, outliers, and correlations in your data.

- **Setosa** has the shortest sepal length and widest sepal width. The two variables have a strong positive correlation.
 - That is, an increase in sepal length results in an increase in sepal width.
- **Versicolor & Virginica** have a weaker positive correlation between sepal width and sepal length.
- **Versicolor & Virginica's** sepal length and width overlap with each other.
 - However, Versicolor's sepal tends to be slightly shorter on average in comparison to Virginica's.

SEPAL LENGTH vs. SEPAL WIDTH by SPECIES



CORRELATION ANALYSIS



Analysis: heatmap

Correlation matrices (heatmaps) are used to understand the **strength** and **direction of relationships between multiple variables**. They are crucial for identifying which variables are positively or negatively correlated.

- We can see that {sepal length and petal length} & {petal width and petal length} are very strongly correlated
 - That is, when sepal length increases, so does petal length.
 - That is, when petal width increases, so does petal length.
- We can also see a negative correlation between {petal length and sepal width}
 - That is, when petal length increases, sepal width decreases.

ANALYSIS

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



Analysis: statistical


The aim of this statistical analysis is to test whether there is a **significant difference in sepal length between the Iris species**.

Hypothesis:

- [H0] Null hypothesis: There's no difference in sepal length between Iris species.
- [H1] Alternative hypothesis: Sepal length differs between Iris species.

Result:

We can see here that our p-value is below 0.05, so we can reject our null hypothesis and accept our alternative hypothesis:

-  There is a difference in the sepal length between Iris species.

ONE-WAY ANOVA

```
import pingouin as pg

# Welch's ANOVA test
result = pg.welch_anova(data=df, dv='sepal_length', between='species')

# Access the p-value from the result
p_value = result['p-unc'].values[0]

# Format the p-value to display with all decimals
formatted_p_value = "{:.35f}".format(p_value)

# Return a message depending on the result of the p-value
if p_value < 0.05:
    print(f"The sepal length differs significantly between Iris species. \np-value: {formatted_p_value}")
else:
    print(f"There's no significant difference in sepal length between Iris species. \np-value: {formatted_p_value}")
```

[illegible]

The results of our one-way ANOVA test

TUKEY'S HSD POST-HOC TEST

Tukey's HSD Post-Hoc Test:

Multiple Comparison of Means – Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Iris-setosa	Iris-versicolor	0.93	0.001	0.6862	1.1738	True
Iris-setosa	Iris-virginica	1.582	0.001	1.3382	1.8258	True
Iris-versicolor	Iris-virginica	0.652	0.001	0.4082	0.8958	True

Tukey's HSD Post-Hoc Test performed on the Iris dataset

Analysis: statistical

The ANOVA test reveals significant differences in sepal length among Iris species, but it doesn't specify exactly which species differ significantly.

For that we need to use **Tukey's post-hoc test**.

Result:

Iris-setosa vs. Iris-versicolor: Mean difference = ~0.93 & p-value < 0.001 (very significant).

- ✓ There is a significant difference

Iris-setosa vs. Iris-virginica: Mean difference = ~1.582, & p-value < 0.001 (very significant).

- ✓ There is a significant difference

Iris-versicolor vs. Iris-virginica: Mean difference = ~0.652 & p-value < 0.001 (very significant).

- ✓ There is a significant difference

ANALYSIS

▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32



Machine learning

Taking our analysis one step further, we can use the data that we have to try and **predict the category of an iris flower**, using its **petal & sepal values**.

As we have the {species} column, our data is **labelled**, allowing us to use a **supervised learning model**: the random forest algorithm to make the predictions.

Results:

- 🚩 100% of our test data was correctly identified by our model's predictions
 - 10 of our predictions correctly guessed Iris-setosa, where it should have guessed Iris-setosa.
 - 9 of our predictions correctly guessed Iris-versicolor, where it should have guessed Iris-versicolor.
 - 11 of our predictions correctly guessed Iris-virginica, where it should have guessed Iris-verginica.

RANDOM FOREST: EVALUATION METRICS

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	10
Iris-versicolor	1.00	1.00	1.00	9
Iris-virginica	1.00	1.00	1.00	11
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

A classification report of our Random Forest model on our Iris dataset

SUMMARY

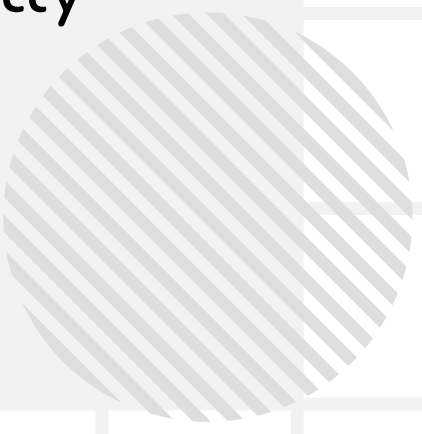
▶▶▶	Project overview.....	03
▶▶▶	Dataset overview.....	08
▶▶▶	Analysis	12
▶▶▶	Data cleaning.....	12
▶▶▶	Descriptive analysis.....	15
▶▶▶	Statistical analysis.....	21
▶▶▶	Machine learning.....	24
▶▶▶	Summary	26
▶▶▶	References.....	32





SUMMARY


1. Iris Setosa:

- **Sepal Characteristics:** Iris setosa has the shortest sepal length and widest sepal width among the three species.
 - **Petal Characteristics:** It has the shortest petal length and petal width.
 - **Distinct Features:** Iris setosa is the most distinguishable species, with significantly different sepal and petal dimensions compared to the other two species.
- 



SUMMARY


2. Iris Versicolor:

- **Sepal Characteristics:** Iris versicolor has intermediate values for sepal length and sepal width.
 - **Petal Characteristics:** It has moderate petal length and petal width, falling between setosa and virginica.
 - **Overlap with Other Species:** Versicolor's characteristics overlap with both setosa and virginica, making it less distinct.
- 





SUMMARY

3. Iris Virginica:

- **Sepal Characteristics:** Iris virginica typically has the longest sepal length among the three species.
 - **Petal Characteristics:** Iris virginica has the longest petal length and wider petal width.
 - **Overlap with Versicolor:** While virginica has some overlap with versicolor, it generally has larger petal dimensions, making it distinguishable.
- 


SUMMARY

4. Statistical analysis:

- **Iris-setosa vs. Iris-versicolor:** The mean difference is approximately 0.93, and the p-adj value is less than 0.001 (very significant).
 -  Therefore, there is a significant difference in sepal length between Iris-setosa and Iris-versicolor.
- **Iris-setosa vs. Iris-virginica:** The mean difference is approximately 1.582, and the p-adj value is less than 0.001 (very significant).
 -  This indicates a significant difference in sepal length between Iris-setosa and Iris-virginica.

SUMMARY

4. Statistical analysis:

- **Iris-versicolor vs. Iris-virginica:** The mean difference is approximately 0.652, and the p-adj value is less than 0.001 (very significant).
 -  This shows a significant difference in sepal length between Iris-versicolor and Iris-virginica.


REFERENCES

- ▶▶▶ Project overview..... 03
- ▶▶▶ Dataset overview.....08
- ▶▶▶ Analysis 12
 - ▶▶▶ Data cleaning..... 12
 - ▶▶▶ Descriptive analysis..... 15
 - ▶▶▶ Statistical analysis.....21
 - ▶▶▶ Machine learning..... 24
- ▶▶▶ Summary 26
- ▶▶▶ References.....32





REFERENCES

- 🔍 **Géron, A. (2023).** *'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems'*. O'Reilly Media.
 - 🔍 **Glen, S. (2023).** *'Z-Score: Definition, Formula and Calculation.'* StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/probability-and-statistics/z-score/>. Last accessed: 02 August 2024.
 - 🔍 **Ektamaini. (2020).** *'Z score for Outlier Detection — Python'*. GeeksforGeeks. <https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/>. Last accessed: 02 August 2024.
- 

REFERENCES

- 🔍 **Stephanie, G (n.d.).** *‘Welch’s ANOVA: Definition, Assumptions’* StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/what-is-statistical-significance/>. Accessed: October 25, 2023.
- 🔍 **Ott, R. L., & Longnecker, M. (2015).** *‘An Introduction to Statistical Methods and Data Analysis (7th ed.)’*. Cengage Learning.
- 🔍 **Stephanie, G (n.d.).** *‘Statistical Significance: Definition, Examples’* StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/what-is-statistical-significance/>. Accessed: October 25, 2023.

REFERENCES

- 🔍 **Lund Research Ltd. (2018).** *'One-way ANOVA in SPSS Statistics'*. Lund Research Ltd. <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php#:~:text=Typically%2C%20a%20one%2Dway%20ANOVA,commonly%20used%20for%20two%20groups>). Accessed: October 25, 2023.
- 🔍 **Stangor, C. (2011).** *'Research methods for the behavioral sciences (4th ed.)'*. Mountain View, CA: Cengage.
- 🔍 **Surbhi, S. (2017).** *'Difference Between Null and Alternative Hypothesis.'* Key Differences. <https://keydifferences.com/difference-between-null-and-alternative-hypothesis.html>. Accessed: October 25, 2023.

REFERENCES

- 🔍 **Walters, S. (2020).** *'Psychology — 1st Canadian Edition'*. Thompson Rivers University. <https://psychology.pressbooks.tru.ca/chapter/3-2-psychologists-use-descriptive-correlational-and-experimental-research-designs-to-understand-behaviour/#:~:text=Descriptive%20research%20is%20designed%20to,to%20assess%20cause%20and%20effect>. Accessed: October 25, 2023.
- 🔍 **Powers, D. M. (2011).** *'Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation'*. Journal of Machine Learning Technologies, 2(1), 37–63.