MA5755 – Data Analysis & Visualization in R/Python/SQL
Assignment 1

Anurag Dhiman - MA25M004

TASK 1

ANSWER TASK 1:

How many samples and features are there?

- From the output, there are 150 samples and 4 features.

What are the feature names?

- The features in dataset named as sepal length (cm), sepal width (cm), petal length (cm) petal width (cm).

What is the target variable?

- The target variable is named as species in dataframe.

TASK 2

ANSWER TASK 2:

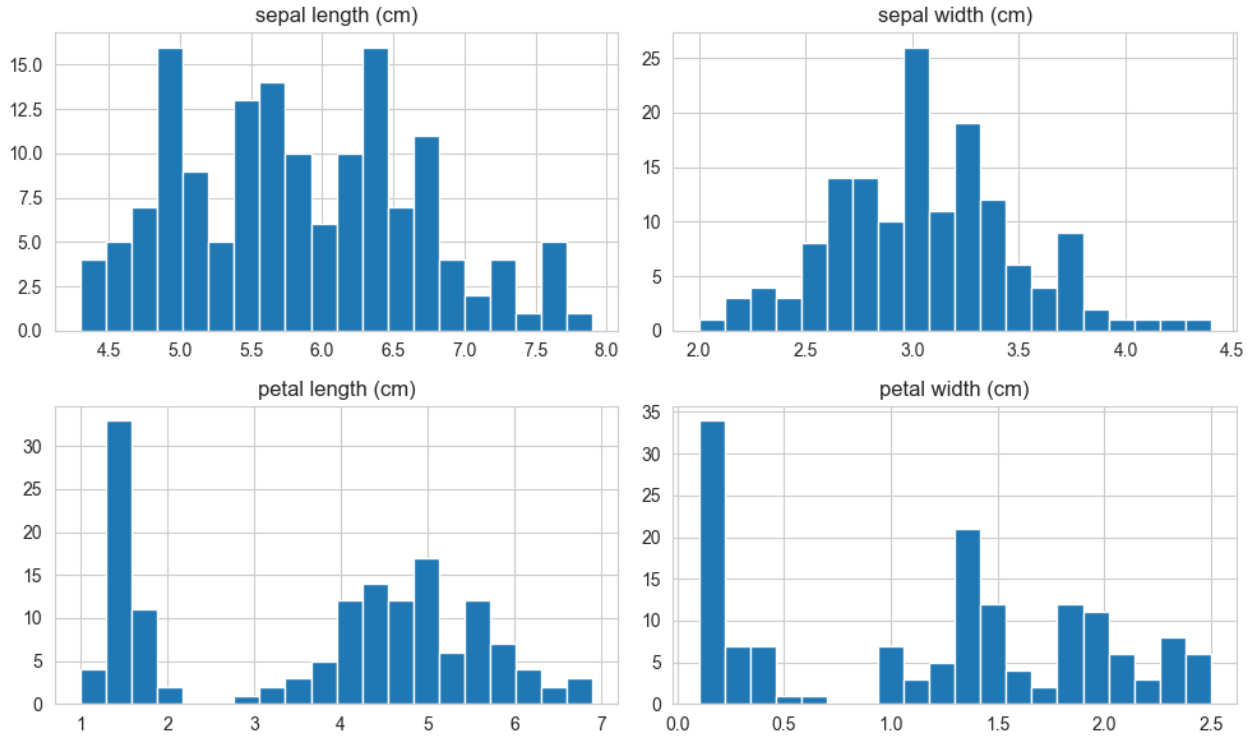
Which features show the largest variability?

- The petal length feature shows the largest variability because range (max - min) is approximately 5.9

Are the feature scales comparable?

- No, because features have different ranges. SO that makes feature scales not comparable.

TASK 3



ANSWER TASK 3:

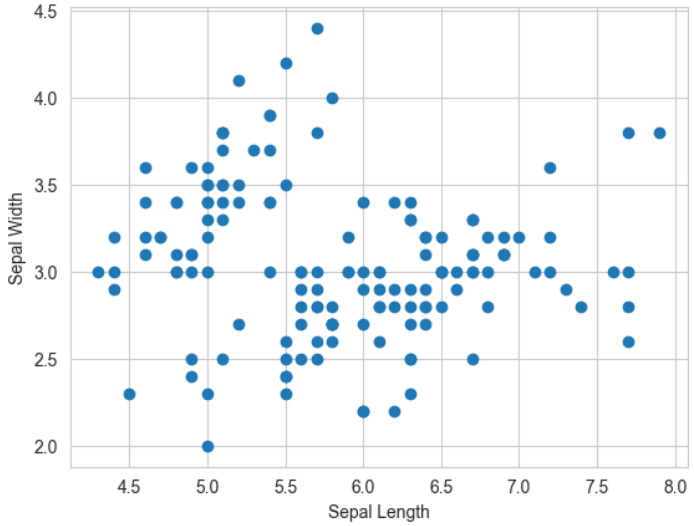
Are the distributions symmetric or skewed?

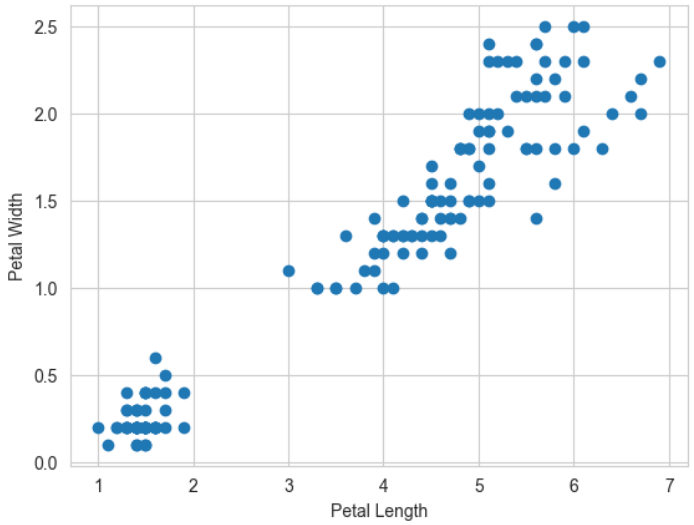
- The distribution are as:
  sepal length is roughly symmetrix, slightly bell shaped.
  Sepal width is roughly symmetric, with a slighly right skew.
  Petal length is showing bimodal distribution (two peaks). It is clearly skewed, not symmetric.
  Petal width is showing bimodal distribution (two peaks). It is heavily skewed.

Do you observe any potential outliers?

- The outlier on sepal width is visible on both sides like (around 2.0 cm and 4.0-4.4 cm). In petal length and petal width, it is hard to identify outliers due to bimodal nature.

TASK 4





ANSWER TASK 4:

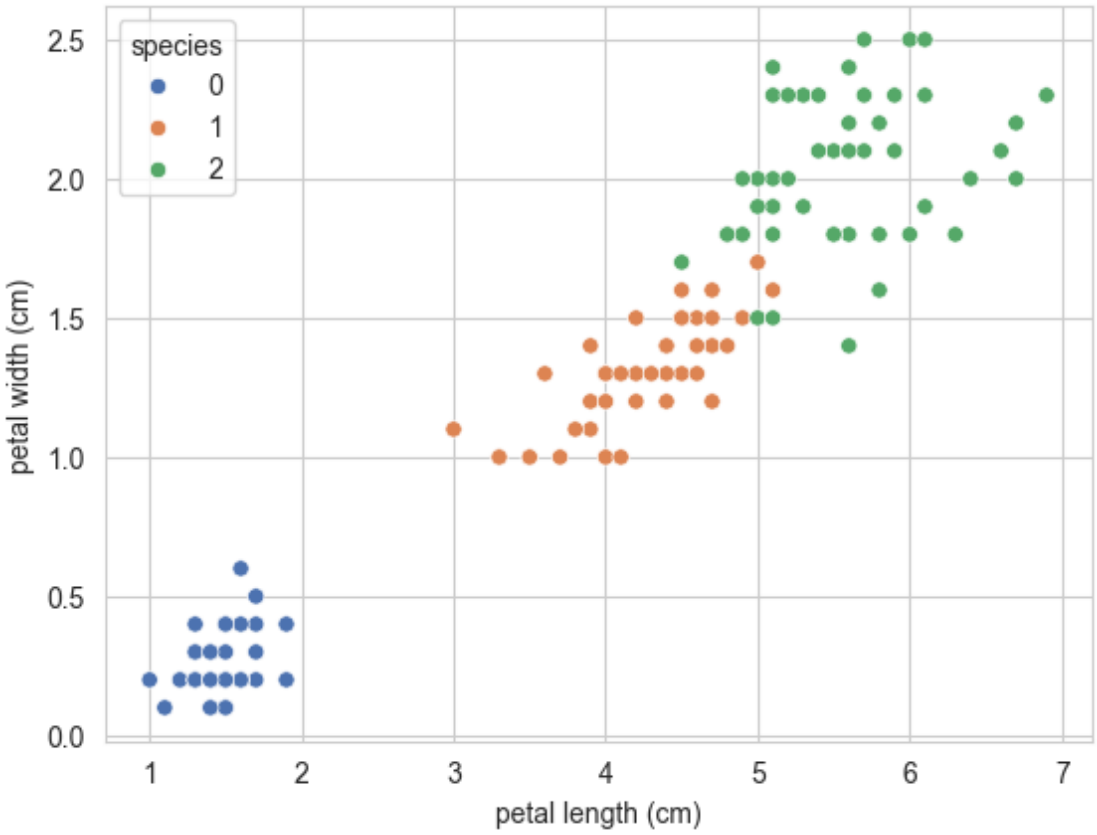
Do you observe any clustering or separation?

- Sepal length vs Sepal width: Shows some overlap with no clear clustering - the data points are more scattered and mixed together, making it difficult to distinguish distinct groups.
- Petal length vs Petal width: Shows very clear clustering and separation into distinct groups.

Which features appear most informative visually

- The plot between petal width and petal length is most informative visually because of clear separation of data points into distinct cluster.

TASK 5



ANSWER TASK 5:

Does labeling change your interpretation?

- Yes, labeling clearly helped to see three different clusters are related the three iris species. Labeling significantly enhanced the interpretation.

Which classes appear easiest or hardest to separate visually?

- The class 0 is easily separable because of clear separate cluster. But class 1 and 2 is hardest to separate because they show some overlap.

TASK 6

ANSWER TASK 6:

Why is exploratory data analysis important before modeling?

- Exloratory data analysis helps us understand the data's structure, distribution, and patterns before jumping into modeling. It reveals important characteristics like feature scales, missing values, outliers, and relationships between variables. This understanding guides our choice of preprocessing steps (like scaling or handling outliers) and helps us select appropriate models. Without exploratory data analysis, we'd be building models blindly without knowing if our data even supports our modeling goals.

What could go wrong if we skip visualization?

- We might miss critical insights like class imbalances, outliers, or non-linear relationships that affect model performance. For example, without visualizing the Iris dataset, we wouldn't know that the features have different scales (requiring normalization) or that setosa is completely separable while the other species overlap. We could waste time using complex models when simple ones would work, or use inappropriate algorithms that assume data properties our dataset doesn't have. Visualizations also help catch data quality issues early.

Which feature(s) would you expect to be useful for classification, and why?

- Petal length and petal width would be the most useful features for classification. From the scatter plots, these features show clear separation between species with minimal overlap - setosa forms a completely distinct cluster, while versicolor and virginica are also distinguishable. The histograms showed bimodal distributions for these features, indicating they capture species-specific differences. Sepal measurements, on the other hand, showed more overlap and would be less discriminative on their own, though they could still provide supplementary information.