Botanica Iris

A story of petals and sepals

Introduction :

We're in possession of a single data set that contains information about 3 different species of the flower iris.

The main goal of this project is to learn how to properly use R in terms of data management and visualization.

In order to do so, we were given a starter sheet, which had most of the techniques we would be using for the following tasks.

1) Why R ? :

a) Richness of statistical functions :

Statistical functions are native in R, same for correlations.

b) Specialized packages :

Everything is available in order to visualize data like ggplot2 or corrplot.
Same for modifying and restructuring data with dplyr or tidyr.

c) Flexibility needed for complex data-set :

Able to transfer data in and out universally with file format ranging from CSV to JSON and going by Excel or XML.
Data transformation, Data analysis, Data visualization *pipeline.*

2) Data refinement :

a) Step 1 : Load & Explore :

```
data(iris)
```

```
str(iris)
summary(iris)
```

b) Step 2 : Filter & Prepare :

```
setosa <- iris %>%
    filter(Species == "setosa") %>%
    select(-Species)
```

c) Transform for visualization :

```
setosa_long <- setosa %>%
  pivot_longer(everything(),
               names_to = "Measurement",
               values_to = "Value")
```

d) Enrich with statistics :

```
stats_summary <- setosa_long %>%
  group_by(Measurement) %>%
  summarise(mean = mean(Value),
            median = median(Value))
```

3) Data analysis with statistical techniques :

“

Only work on *setosa* will be shown, for full insights on the whole dataset, please refer to the key insights.

“

a) Central Tendencies :
Mean : Sepal = (5.01 * 3.43 cm), Petal = (1.46 * 0.25 cm).
Median ~~ Mean → Symmetrical distribution.

b) Dispersion measures :
Maximum coefficient of variation (CV) on petal width (44%) → High variability.

```
cv_petal_width = sd(Petal.Width)/mean(Petal.Width)*100
```

Min CV on sepal length (7%) → Stable measure.

c) Correlation Analysis :
Significance tests showing correlation : r = 0.74.

4) Data analysis with visualization :

a) Density Histograms:
Distribution of each measure.
Statistical annotations (mean/median).
Overlapping of density curbs.

```
hist(iris$Petal.Length, breaks = 10, col = "lightblue")
```

b) Correlation Matrixes :
4 by 4 matrix showing correlating sepal measures unique to Setosa.
Colored Heatmap.
Hierarchical clustering.
Visual significance.

```
cor_matrix <- cor(iris[, 1:4])
corrplot::corrplot(cor_matrix, method = "color")
```

c) Comparative Boxplots :
Outliers detection.
Inter-measures comparison.

```
ggplot(iris, aes(x = Species, y = Petal.Length, fill = Species)) +
  geom_boxplot() +
  theme_classic()
```

5) Key insights :

a) Setosa's aka The Miniature :
Setosa is unique among the 3 species for its signature ***miniature*** size and petals, and seeing a unique positive correlation among the sepals in the dataset.

b) Setosa's correlation coefficient between sepals measures :
Setosa's **sepals** are strongly correlated between their length and width.

c) Setosa's steady measurements :
Setosa measurements have little to **no variation**.

d) Versicolor aka The Medium :
   Veriscolor is recognised by its ***average*** size

e) Versicolor's correlation coefficient between petals measures :
   Versicolor's **petals** are strongly correlated between their length and width.

f) Virginica aka The Giant :
   Virginica is the ***largest*** specie in all dimensions

g) Virginica's High Variability :
   We found the highest **variability** in the virginica's measurements.