Botanica iris

Derrière chaque pétale, une vérité statistique.

Introduction

Vous êtes chercheur en botanique, vous appliquez des méthodes d'analyse exploratoire au dataset Iris… Et peu à peu, les données révèlent ce que l'œil ne voit pas : des motifs subtils entre la longueur des sépales et l'espèce, des regroupements inattendus, des indices presque invisibles.

Ce qui n'était qu'un simple tableau devient une carte : celle d'un écosystème codé, où chaque valeur chuchote une vérité.

À travers les graphes, vous voyagez — de jardin en jardin, de Setosa en Virginica, jusqu'à ce que les fleurs ne soient plus seulement observées, mais comprises.

Et dans ce monde numérique, une idée s'impose :

🌸 La nature aussi parle en statistiques. Il suffit d'écouter.

Iris Versicolor | Iris Setosa | Iris Virginica

Pourquoi R et l'Analyse Exploratoire des Données ?

L'analyse exploratoire est une **étape fondamentale en Data Science**. Avant de développer des modèles avancés, il est essentiel de **comprendre les données, détecter les tendances, les anomalies et préparer les analyses futures**.

R est l'un des **langages les plus puissants et populaires pour l'analyse de données** grâce à :

- ➔ Sa richesse en fonctions statistiques
- ➔ Ses packages spécialisés comme ggplot2, dplyr et tidyr
- ➔ Sa flexibilité pour manipuler et visualiser des datasets complexes

🎯 **Ce projet a pour objectif de vous apprendre à :**

- ➔ Explorer un dataset réel **(IRIS)**
- ➔ Appliquer des techniques **statistiques descriptives avancées**
- ➔ Créer et interpréter des **visualisations graphiques efficaces**

Description du sujet

Vous êtes **data analyst junior** et devez analyser le **dataset IRIS**, qui contient des mesures de fleurs appartenant à trois espèces différentes.

Votre mission est de :

- Explorer et comprendre la structure des données
- Effectuer des **analyses statistiques complètes** pour extraire des insights
- Produire des **visualisations impactantes** qui mettent en lumière les tendances
- Présenter et expliquer vos résultats de manière claire et synthétique

Problématiques à explorer :

1. Quelles sont les relations entre les différentes variables ?
2. Peut-on prédire l'espèce d'une fleur en fonction de ses mesures ?
3. Comment identifier les valeurs aberrantes dans le dataset ?

Ce qu'on attend de vous

- → **Manipulation de données sous R** : Importation, exploration et transformation
- → **Statistiques descriptives avancées** : Moyennes, médianes, quartiles, dispersion, corrélation
- → **Visualisation des données avec ggplot2** : Histogrammes, boxplots, scatter plots, heatmaps
- → **Communication des résultats** : Structuration et présentation des analyses

Ce projet est divisé en **4 grandes étapes**, chacune avec des objectifs clairs et une mise en pratique sous R.

Étape 1 : Exploration et Préparation des Données

Explorer et préparer les données afin de comprendre leur structure et identifier les premières tendances.

Objectif : Comprendre la structure et identifier les premières tendances.

- Charger et observer le dataset IRIS
- Analyser les types de variables et leurs distributions
- Détecter des valeurs manquantes

Étape 2 : Statistiques Descriptives et Analyse des Relations

Objectif : Décrire les caractéristiques des données et identifier des tendances.

- Mesurer la tendance centrale et de dispersion

Analyse de la corrélation entre les variables

Interprétation attendue :

➔ Quelles variables sont fortement corrélées ?

➔ Comment ces mesures influencent-elles l'espèce d'une fleur ?

- Représenter le visuelle de la corrélation avec un heatmap

Étape 3 : Visualisation et Identification des Tendances

Objectif : Communiquer efficacement les résultats via des graphiques pertinents.

Créer des :

- Histogrammes pour voir la répartition des valeurs
- Boxplots pour détecter les outliers

Analyse attendue :

➔ Existe-t-il des différences significatives entre les espèces ?

➔ Quels sont les outliers identifiables sur les boxplots ?

- **Scatter Plot pour analyser les relations entre les variables**

Questions à répondre :

Vous devrez répondre à la problématique

➔ Les espèces de fleurs sont-elles bien séparées sur le scatter plot ?

➔ Quels patterns peuvent être exploités pour classifier les espèces ?

Étape 4 : Synthèse et Communication des Résultats

Objectif : Présenter une analyse claire et exploitable.

- Rédaction d'un rapport avec les principaux insights
- Structuration d'une présentation sur google slides
- Explication des choix méthodologiques et des visualisations

Rendu

L'évaluation du projet se fera sur deux livrables :

Présentation explicative sous forme de diapositives

- ➔ Présentation du dataset et objectifs
- ➔ Statistiques descriptives et analyse des corrélations
- ➔ Visualisations et interprétations
- ➔ Synthèse et conclusions

Repository GitHub public `Project_R1_Name` contenant :

- ➔ Code R propre et commenté
- ➔ Graphiques générés avec ggplot2
- ➔ Fichier README expliquant la démarche
- ➔ Rapport synthétique en PDF

Base de connaissances

- Data Import in R
- Base R Statistics
- ggplot2 Reference