

# Revision

January 12, 2024

## 0.1

### *Rvision*

Dans ce fichier, nous allons effectuer une analyse complète des données.

## 0.2 1) Identification du problème

La première étape de toute analyse consiste à identifier le problème que nous allons étudier. Dans ce fichier, notre problème est de déterminer quels sont les facteurs qui influencent la satisfaction des individus. Pour aborder ce problème, nous formulons plusieurs hypothèses. Par exemple, nous pourrions supposer que le revenu, l'éducation, l'expérience professionnelle et le secteur d'activité ont une influence sur la satisfaction. Pour répondre à ces questions, nous allons effectuer les analyses suivantes :

- Régression linéaire simple : nous allons essayer de prédire la satisfaction en nous basant sur une variable qui présente une relation linéaire avec la satisfaction.
- Régression linéaire multiple : nous allons essayer de prédire la satisfaction en nous basant sur plusieurs variables qui présentent une relation linéaire avec la satisfaction.
- ANOVA (Analyse de la variance) : nous allons utiliser l'ANOVA pour déterminer si les moyennes de satisfaction diffèrent significativement entre différents groupes ou catégories de variables.
- ACP (Analyse en composantes principales) : nous allons utiliser l'ACP pour réduire la dimensionnalité de nos données et identifier les variables qui contribuent le plus à la variation de la satisfaction.
- ACF (Analyse des corrélations fonctionnelles) : nous allons utiliser l'ACF pour étudier les relations fonctionnelles entre les variables et la satisfaction.

## 0.3 2) Collection des données

La deuxième étape est généralement la phase de collection des données. Pour cela, nous devons suivre plusieurs étapes :

1. Identification de l'échantillon : Il est essentiel de choisir un échantillon représentatif qui correspond à notre problème d'étude. Cela garantit que les résultats obtenus sont généralisables à la population cible.
2. Collecte des données : La collecte des données peut être réalisée à l'aide de questionnaires, d'entretiens, d'observations, etc. Il est important de collecter un maximum de données pertinentes pour notre analyse.

Dans notre cas, nous utiliserons des données générées par l'IA, juste pour la simplicité.

## 0.4 3) Prétraitement des données

La troisième étape consiste au prétraitement des données. Cela implique la suppression des variables perturbantes, le filtrage et la suppression des colonnes indésirables, ainsi que la transformation des données en un format avec lequel nous pouvons travailler.

Pour effectuer le prétraitement des données, nous pouvons utiliser des techniques telles que :

- Nettoyage des données : Suppression des valeurs manquantes, des valeurs aberrantes et des doublons.
- Transformation des données : Normalisation, discrétisation, codage des variables catégorielles, etc.
- Sélection des variables : Identification des variables les plus pertinentes pour notre analyse.

Une fois que nous avons prétraité nos données, nous sommes prêts à passer à l'étape suivante de notre analyse.

Maintenant, on commence notre analyse, d'abord on faire la lecture des données.

•

### 0.4.1 Lecture des données

```
[ ]: data <- read.csv("data/data.csv")  
head(data)
```

		Revenu <int>	Experience <int>	Education <chr>	Satisfaction <dbl>	Secteur <chr>	Age <int>
A data.frame: 6 × 6	1	60000	5	Bachelor	8.5	IT	32
	2	75000	8	Master	7.2	Finance	45
	3	50000	3	Secondaire	6.0	Ventes	28
	4	90000	9	Ph.D.	9.3	Recherche	38
	5	120000	11	Bachelor	8.8	Ingénierie	50
	6	95000	6	Master	7.5	IT	34

La description des données : 1. **Revenu** : Le revenu en dollars. 2. **Expérience** : Le nombre d'années d'expérience dans le domaine de travail. 3. **Éducation** : Le niveau d'éducation, catégorisé en Secondaire, Bachelor (Licence), Master et Ph.D. 4. **Satisfaction** : Le niveau de satisfaction, mesuré sur une échelle donnée. 5. **Secteur d'activité** : Le secteur d'activité dans lequel la personne travaille. 6. **Âge** : L'âge de la personne.

Dans notre cas, nous avons aucune valeur nulle.

## 0.5 4) Analyse Statistique

Maintenant on commence notre Analyse Statistique, et la première chose est de déterminer les types des variables à étudier.

### 0.5.1 Types des variables

- **Variables quantitatives :**

1. Revenu - Quantitative continue (mesurée en dollars).

2. Expérience - Quantitative discrète (mesurée en années).
3. Satisfaction - Quantitative continue (mesurée sur une échelle).

- **Variables qualitatives :**

1. Éducation - Qualitative ordinale (niveau d'éducation catégorisé en Secondaire, Bachelor, Master, Ph.D.).
2. Secteur - Qualitative nominale (catégorie du secteur d'activité).
3. Âge - Qualitative ordinale (groupe d'âge).

### 0.5.2 Analyse univariée

L'analyse univariée est une étape initiale dans l'exploration des données qui vise à comprendre la distribution et les caractéristiques d'une seule variable à la fois. Elle permet d'obtenir des informations statistiques et des tendances centrales pour une variable donnée, qu'elle soit quantitative ou qualitative.

**Pour les variables quantitatives : - Description générale :** La description d'une variable quantitative implique l'analyse de la tendance centrale, qui est souvent représentée par la moyenne, la médiane et le mode. On examine également la dispersion des données, généralement mesurée par l'écart type. - **Visualisation :** Pour visualiser une variable quantitative, on utilise généralement des histogrammes, des boîtes à moustaches (box plots) et des diagrammes de densité.

**Pour les variables qualitatives : - Description générale :** Pour les variables qualitatives, l'analyse se concentre sur la fréquence de chaque catégorie. On examine également les proportions et les pourcentages pour avoir une idée de la répartition des catégories. - **Visualisation :** Les diagrammes en barres (bar plots) et les diagrammes circulaires (pie charts) sont couramment utilisés pour visualiser les variables qualitatives.

Ces analyses univariées fournissent une compréhension approfondie des caractéristiques individuelles des variables, ce qui peut orienter la sélection des méthodes d'analyse statistique plus avancées par la suite.