|  |
| --- |
| Projet personnel à but formateur et exploratoire |
| Rapport d’Analyse de Données |
| *Exploration, traitement et visualisation de données* |

|  |
| --- |
| Sylvain Vasseur  08/04/2025 |

Table des matières

[**I.** **Introduction** 2](#_Toc195022679)

[A. Contexte 2](#_Toc195022680)

[B. Objectif 2](#_Toc195022681)

[C. Présentation des données 2](#_Toc195022682)

[**Identifiants** 2](#_Toc195022683)

[**Informations sur les commandes** 2](#_Toc195022684)

[**Informations sur le client** 2](#_Toc195022685)

[**Informations sur les produits** 3](#_Toc195022686)

[**II.** **Méthodologie** 3](#_Toc195022687)

[**III.** **Analyse** 3](#_Toc195022688)

[**IV.** **Discussion** 3](#_Toc195022689)

[**V.** **Conclusion** 4](#_Toc195022690)

1. **Introduction**

## Contexte

L’entreprise étudiée évolue dans un secteur très concurrentiel, où la demande des consommateurs ne cesse d'augmenter. Face à un environnement dynamique, cette entreprise cherche à mieux comprendre les facteurs clés influençant ses ventes et ses profits. Plus précisément, elle souhaite analyser quels produits, régions, catégories et segments de clients sont les plus rentables, afin de cibler ceux qui offrent le meilleur potentiel tout en évitant les segments moins performants. L’enjeu de cette analyse est de fournir des recommandations stratégiques permettant à l’entreprise de se différencier et d'optimiser ses résultats dans un marché en perpétuelle évolution.

## Objectif

L'objectif principal de cette analyse est d'identifier les facteurs clés influençant les performances de l’entreprise dans un marchédynamique. Pour ce faire, nous chercherons à analyser les relations entre différentes variables, qu'elles soient **qualitatives** (comme les catégories de produits) ou **quantitatives** (comme les ventes ou les profits). Cette approche complète vise à fournir des recommandations stratégiques pour aider l’entreprise à optimiser ses choix en termes de produits, régions, catégories et segments de clients à cibler, tout en évitant les éléments moins rentables.

## Présentation des données

Le dataset utilisé contient plusieurs variables qui permettent d'analyser les performances de l’entreprise. Voici un aperçu de ces variables :

### **Identifiants**

* **Row ID** (Numérique) : Identifiant unique pour chaque ligne
* **Order ID** (Numérique) : Identifiant unique de chaque commande
* **Customer ID** (Numérique) : Identifiant unique pour chaque client
* **Product ID** (Numérique) : Identifiant unique du produit

### **Informations sur les commandes**

* **Order Date** (Date) : Date de la commande du produit
* **Ship Date** (Date) : Date d'expédition du produit
* **Ship Mode** (Catégoriel) : Mode de livraison spécifié par le client
* **Quantity** (Numérique) : Quantité du produit commandé
* **Discount** (Numérique) : Remise appliquée sur le produit
* **Sales** (Numérique) : Montant des ventes du produit
* **Profit** (Numérique) : Profit ou perte réalisée sur la commande

### **Informations sur le client**

* **Customer Name** (Textuel) : Nom du client
* **Segment** (Catégoriel) : Segment auquel appartient le client
* **Country** (Textuel) : Pays de résidence du client
* **City** (Textuel) : Ville de résidence du client
* **State** (Textuel) : État de résidence du client
* **Postal Code** (Numérique) : Code postal de chaque client
* **Region** (Catégoriel) : Région où le client réside

### **Informations sur les produits**

* **Category** (Catégoriel) : Catégorie du produit commandé
* **Sub-Category** (Catégoriel) : Sous-catégorie du produit commandé
* **Product Name** (Textuel) : Nom du produit

1. **Méthodologie**
   1. **Préparation des données** : Nettoyage, traitement des valeurs manquantes, création de variables.

Avant de procéder à l’analyse, une étape de vérification de la qualité des données a été réalisée. Aucune valeur manquante n’a été détectée dans le jeu de données, ce qui a permis d’éviter les opérations de traitement ou d’imputation classiques. Le seul ajustement nécessaire a concerné le type des variables de dates (Order Date et Ship Date), qui ont été converties au format datetime afin de pouvoir effectuer d’éventuelles analyses temporelles. À ce stade, la création de nouvelles variables n’a pas été jugée nécessaire, les données disponibles étant déjà suffisamment structurées pour répondre aux objectifs de l’analyse.

* 1. **Démarche analytique** : Expliquer l’ordre logique des étapes choisies (ex. : pourquoi commencer par l’exploration, puis corrélation, puis segmentation, etc.).

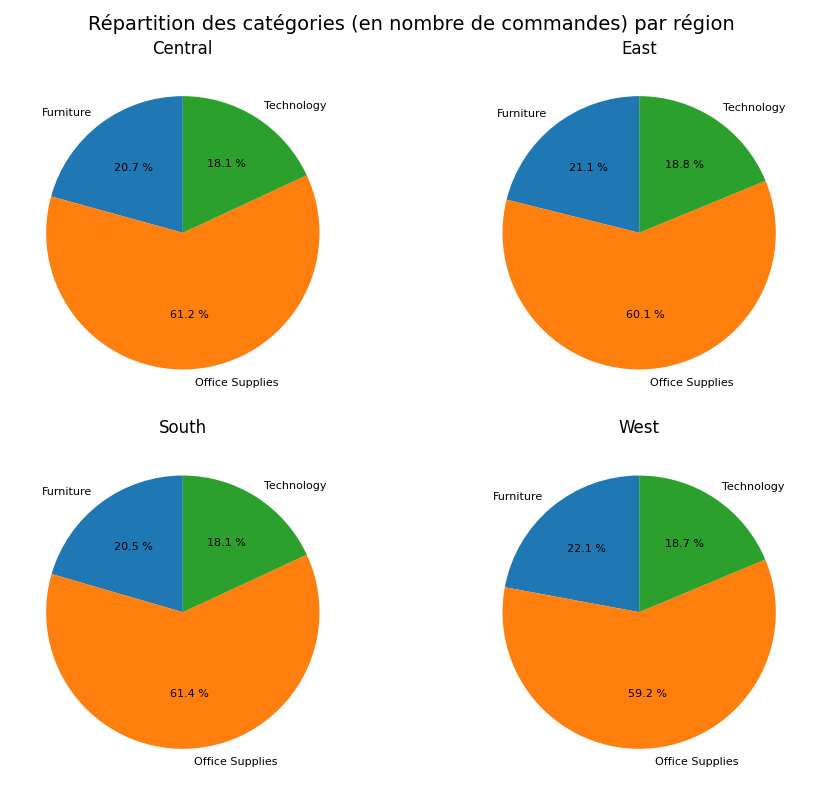
La première étape de cette analyse a consisté à explorer les données à l’aide de statistiques descriptives telles que la moyenne, l’écart-type, le minimum et le maximum. Cela a permis de se faire une idée globale de la distribution des variables quantitatives comme les ventes, les profits, les remises ou les quantités commandées. Le tableau ci-dessous synthétise ces indicateurs et a servi de base pour identifier des variables à analyser plus en détail dans les étapes suivantes (corrélations, segmentations, etc.).

Afin d’approfondir la compréhension des relations entre les variables quantitatives, une matrice de corrélation a été calculée. De manière générale, les corrélations observées restent faibles, la plus élevée étant de 0.48. On note toutefois quelques relations logiques :

* Une corrélation positive modérée entre **Profit** et **Sales** (*r = 0.48*) : ce qui peut s’expliquer par le fait qu’un volume de ventes plus élevé peut générer plus de bénéfices, bien que cela dépende aussi d'autres facteurs comme les remises.
* Une corrélation **négative** entre **Profit** et **Discount** (*r = -0.22*) : logique également, car offrir des remises importantes peut réduire les marges et donc impacter négativement le profit.
* Une corrélation faible entre **Sales** et **Quantity** (*r = 0.20*) : ce résultat suggère que le nombre d’articles vendus n’est pas strictement proportionnel au montant total des ventes. Cela peut être lié à la diversité des prix selon les produits.

Les autres relations sont proches de zéro, ce qui signifie qu’il n’existe pas de dépendances linéaires fortes entre ces variables dans l’ensemble du dataset.

Dans cette analyse, la segmentation ne consiste pas nécessairement à créer de nouveaux groupes à partir des données, mais plutôt à exploiter des variables déjà structurées en catégories distinctes. Par exemple, des variables comme le mode de livraison (Ship Mode) ou le segment client (Segment) comportent un nombre limité de modalités, ce qui permet de comparer directement les performances (ventes, profits, remises, etc.) entre ces groupes. L’objectif est alors d’identifier si certaines modalités se distinguent des autres par leur rentabilité ou leur comportement, afin d’éclairer les décisions stratégiques.

1. **Analyse**
   1. **Exploration des données** : Statistiques descriptives, visualisations simples.

Répartition des catégories

* 1. **Relations entre variables** : Analyse quali-quali, quanti-quanti, quali-quanti (on pourra y revenir ensemble plus en détail, t’inquiète).
  2. **Segmentation** : Regrouper ou découper les données selon des critères pertinents pour raffiner l’analyse.
  3. **Résultats & interprétations** : Ce qui se dégage concrètement des analyses, avec un lien vers l’objectif.

1. **Discussion**
   1. **Limites** : Ce qui a pu freiner ou restreindre l’analyse (données manquantes, périmètre, temps, etc.).
   2. **Recommandations** : Ce que l’analyse permet de proposer concrètement comme actions ou stratégies.
2. **Conclusion**
   1. **Résumé** : Retour sur les principaux résultats.
   2. **Pistes futures** : Suggestions pour aller plus loin, approfondir ou prolonger le travail.