Projet d'Analyse de Données

Visualisation des Données du Dataset Sample - Superstore et Évaluations des Performances Commerciales

Note de Synthèse

Réalisé par :
Lucas ALIOME

EURECOM - Sophia Antipolis
Année académique 2025

Contexte et Problématique

Dans un contexte où la donnée constitue un atout stratégique majeur, la capacité à analyser et interpréter les informations commerciales est devenue incontournable pour guider les choix d'une entreprise. Ce projet s'inscrit dans cette logique en exploitant le fichier "**Sample – Superstore.xls**", qui recense les ventes d'une enseigne fictive sur l'ensemble du territoire américain.

L'objectif est de transformer ces données brutes en visualisations claires et en indicateurs pertinents, afin de mieux évaluer les performances commerciales, repérer les zones de rentabilité ou de faiblesse, et proposer des recommandations concrètes pour améliorer les résultats.

L'analyse couvre la **période 2014–2018** et se concentre exclusivement sur les données internes, sans intégrer les facteurs macroéconomiques extérieurs. Elle vise à fournir des éléments d'aide à la décision pour les équipes marketing et commerciales, à travers une lecture multi-échelle des ventes et de leurs déterminants internes.

Méthodologie et Exploitation Technique

L'exploitation du fichier "*Sample - Superstore.xls*" a été effectuée à l'aide du langage de programmation Python, dans un environnement Jupyter Notebook. Cet environnement a permis d'utiliser les différentes bibliothèques Python nécessaires à l'extraction des données:

- **pandas** pour la manipulation et le nettoyage des données
- **matplotlib** et seaborn pour la création de visualisation de graphique
- **numpy** pour les calculs numériques

Cette méthodologie a permis d'explorer efficacement les tendances, d'identifier des corrélations significatives, et de produire des graphiques clairs et explicatifs illustrant les résultats de l'analyse. Dans les sections suivantes, les fonctions Python ne seront pas présentées, celles-ci étant disponibles dans l'intégralité du projet accessible sur GitHub: https://github.com/Lucas-Aliome/Sales-Performance-Analytics

Analyse du DataFrame

Le fichier a été chargé en DataFrame avec pandas, facilitant ainsi l'exploration et la manipulation des données. Avant l'analyse, il est crucial de comprendre le contenu :

- Le DataFrame compte **9 994 lignes** (correspondant à 9 994 commandes passées entre 2014 et 2018) et **21 colonnes.**
- Il y a en réalité **793 clients distincts**.
- L'examen des données inclut la vérification des valeurs manquantes et la suppression des colonnes inutiles pour améliorer la qualité du jeu de données.

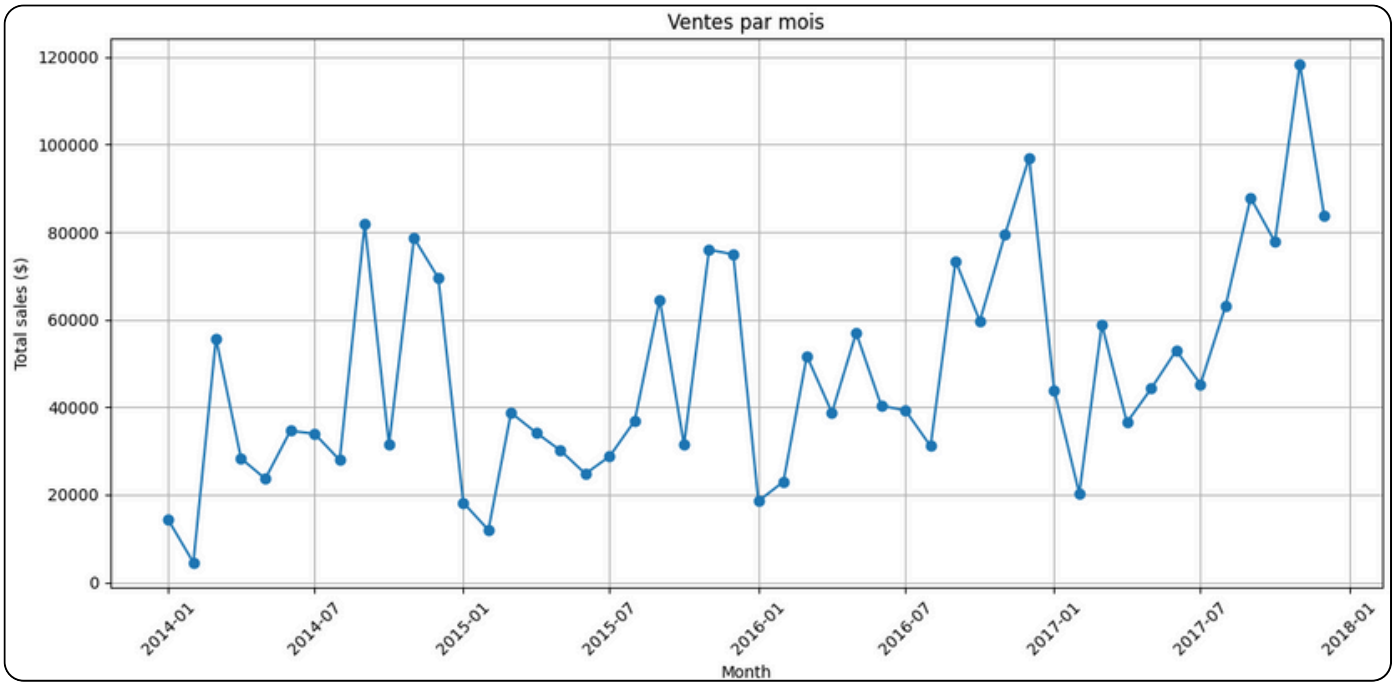
Les colonnes importantes retenues pour l'analyse sont notamment :

- **Order ID** (identifiant unique de commande)
- **Order Date** et **Ship Date** (dates pour étudier délais et saisonnalité)
- **Segment** (informations clients)
- **Category et Sub Category**(classification des produits)
- **Sales**, **Quantity**, **Profit** (indicateurs clés de performance commerciale)
- **City**, **State**, **Region** (données géographiques)

Les autres colonnes, moins pertinentes ou redondantes, seront exclues pour garantir la clarté des résultats.

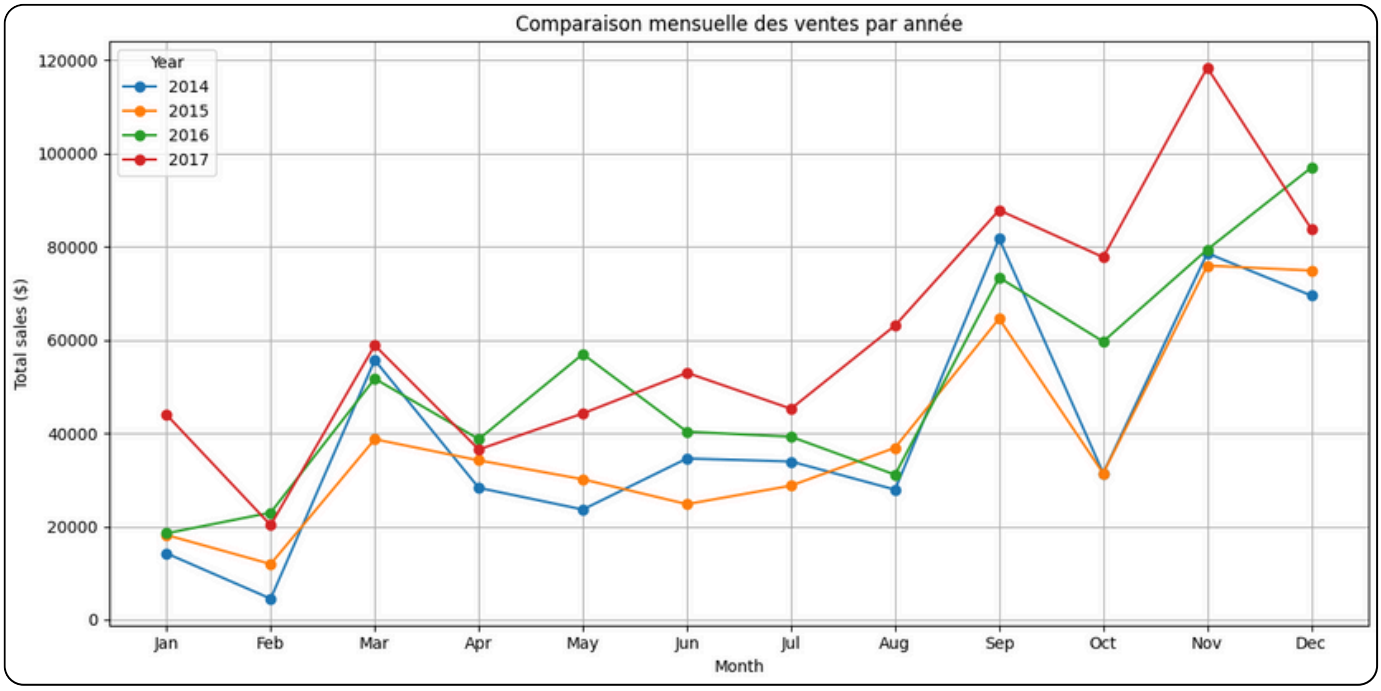
Analyse des Ventes et Profits

L'étude des ventes constitue un volet central de l'analyse commerciale. Elle permet d'évaluer la performance financière globale et de dégager des tendances clés, tant au niveau des segments clients que des régions ou des modes de livraison, et délais de livraison. Les indicateurs tels que le chiffre d'affaires moyen, la répartition des ventes, ainsi que l'impact des variables explicatives sur les revenus seront examinés pour mieux comprendre les leviers de croissance et les opportunités d'optimisation.



Les ventes enregistrées entre 2014 et 2018 suivent une tendance globalement croissante: en dessous de 20 000$ en Janvier 2014 et juste au dessus de 80 000$ en Décembre 2017. Cependant, cette progression n'est ni linéaire, ni polynomiale : elle se caractérise par une forte volatilité. L'analyse chronologique fait apparaître des fluctuations marquées, avec des pics de ventes récurrents durant la période estivale et des baisses significatives en hiver.

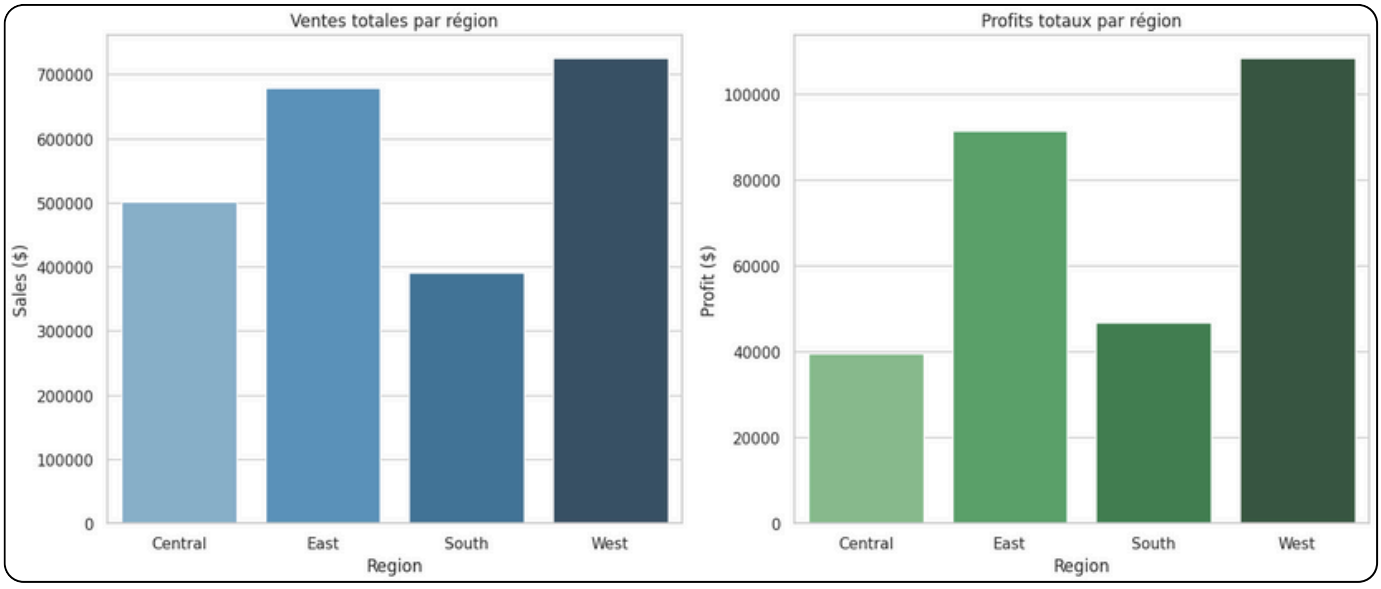
Ces variations saisonnières suggèrent une **influence notable de la temporalité sur les comportements d'achat**, ce qui pourrait orienter les décisions commerciales.



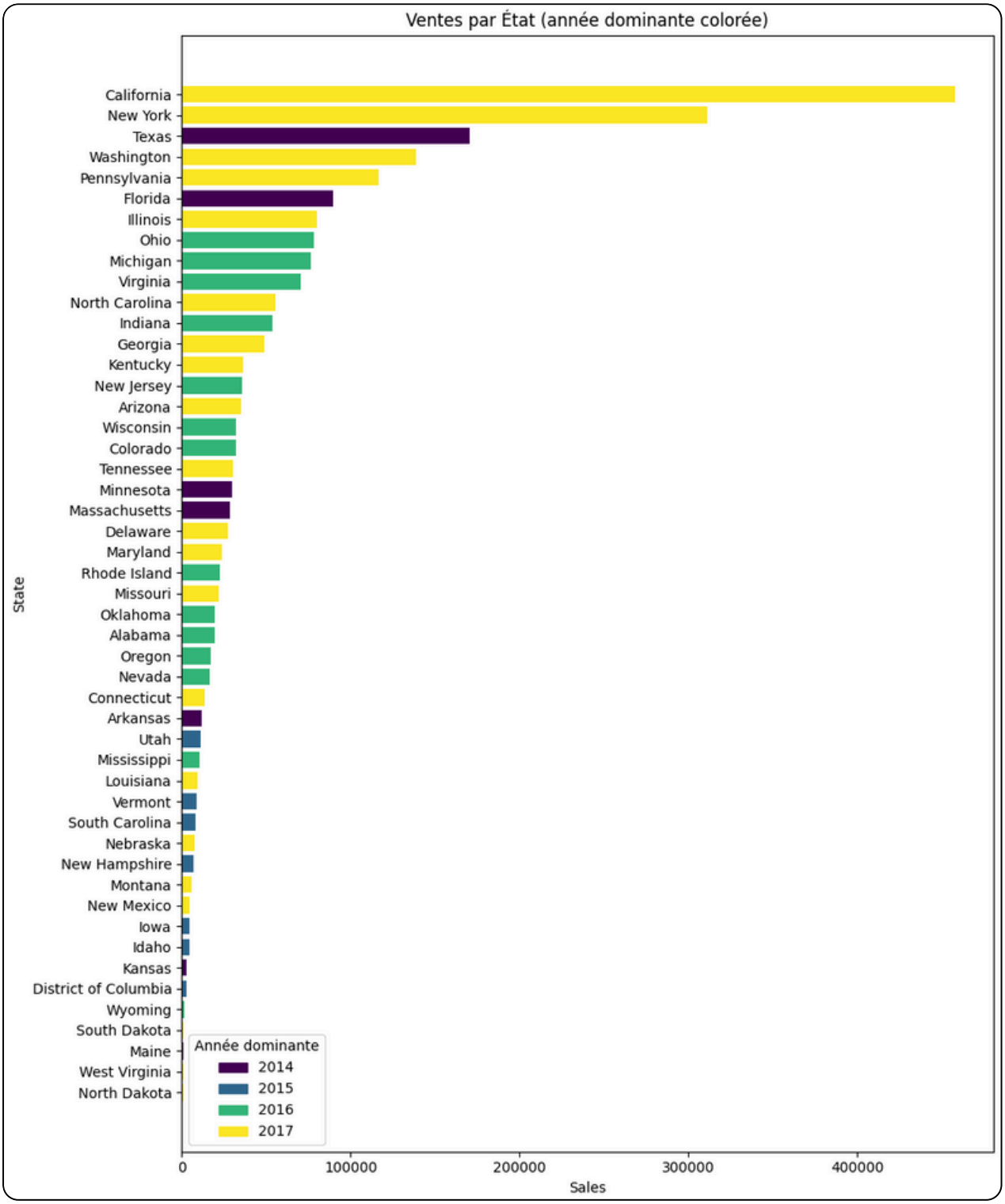
Ce graphique illustre l'évolution mensuelle des ventes sur quatre années d'activité. On observe une croissance progressive du chiffre d'affaires, avec une tendance globale à la hausse au fil du temps.
L'année 2017 se distingue par des **performances particulièrement solides**, dépassant les autres années la plupart des mois, sauf en mai et décembre.

L'analyse révèle un **comportement saisonnier régulier** et **une certaine cyclicité** :

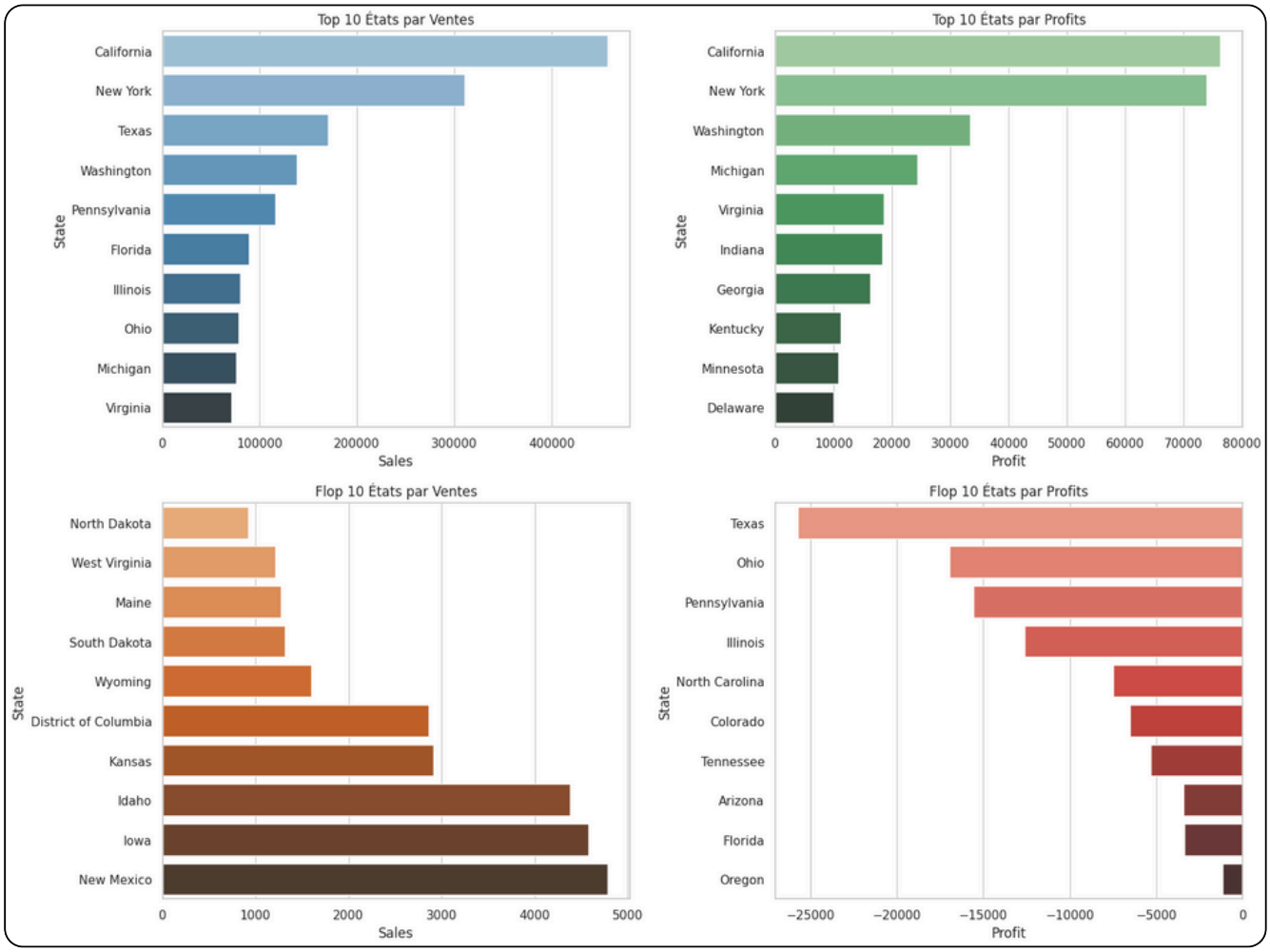
- une hausse récurrente des ventes en février,
- une baisse marquée en mars,
- une stabilité relative en juin,
- une forte augmentation entre août et septembre,
- suivie d'un recul en octobre, puis d'une reprise en fin d'année.

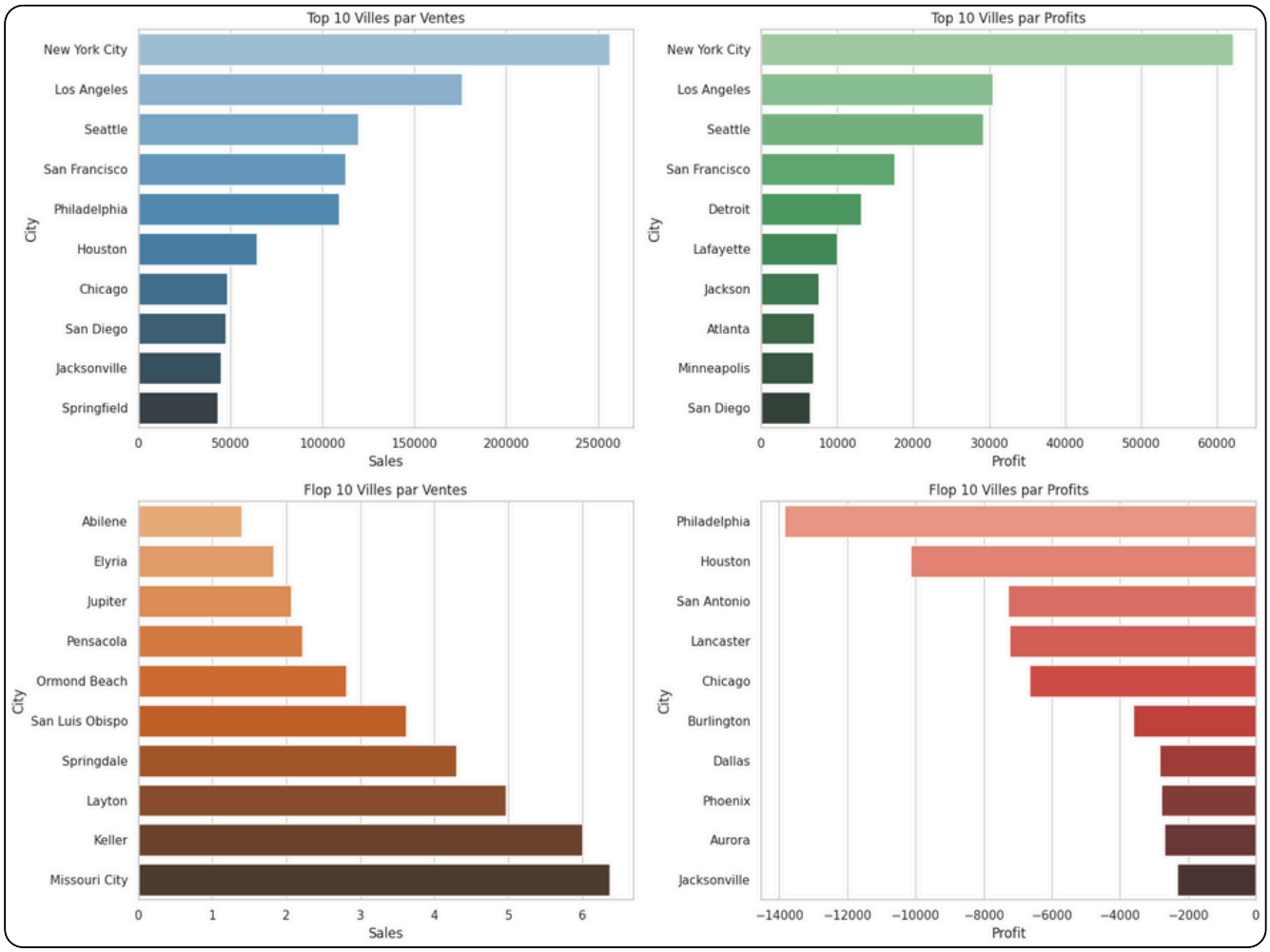


Les régions West et East dominent en ventes et en profits. La région Central génère peu de profit malgré un bon volume de ventes, ce qui soulève des questions sur la rentabilité. La région South, en retard sur tous les plans, nécessite une analyse pour identifier les causes de ses faibles performances.



Les ventes varient fortement selon les États : la Californie, le Texas et New York concentrent plus de 40 % du chiffre d'affaires, surtout en 2017. À l'inverse, des États comme l'Arkansas ou l'Iowa restent peu contributeurs, avec des ventes limitées, surtout en début de période. Ces écarts s'expliquent par la densité de population, la maturité du marché ou encore la présence d'actions marketing ciblées.


Top 10 États par Ventes
California
New York
Texas
Washington
Pennsylvania
Florida
Illinois
Ohio
Michigan
Virginia
State
0
100000
200000
300000
400000
Sales
Top 10 États par Profits
California
New York
Washington
Michigan
Virginia
Indiana
Georgia
Kentucky
Minnesota
Delaware
State
0
10000
20000
30000
40000
50000
60000
70000
80000
Profit
Flop 10 États par Ventes
North Dakota
West Virginia
Maine
South Dakota
Wyoming
District of Columbia
Kansas
Idaho
Iowa
New Mexico
State
0
1000
2000
3000
4000
5000
Sales
Flop 10 États par Profits
Texas
Ohio
Pennsylvania
Illinois
North Carolina
Colorado
Tennessee
Arizona
Florida
Oregon
State
−25000
−20000
−15000
−10000
−5000
0
Profit

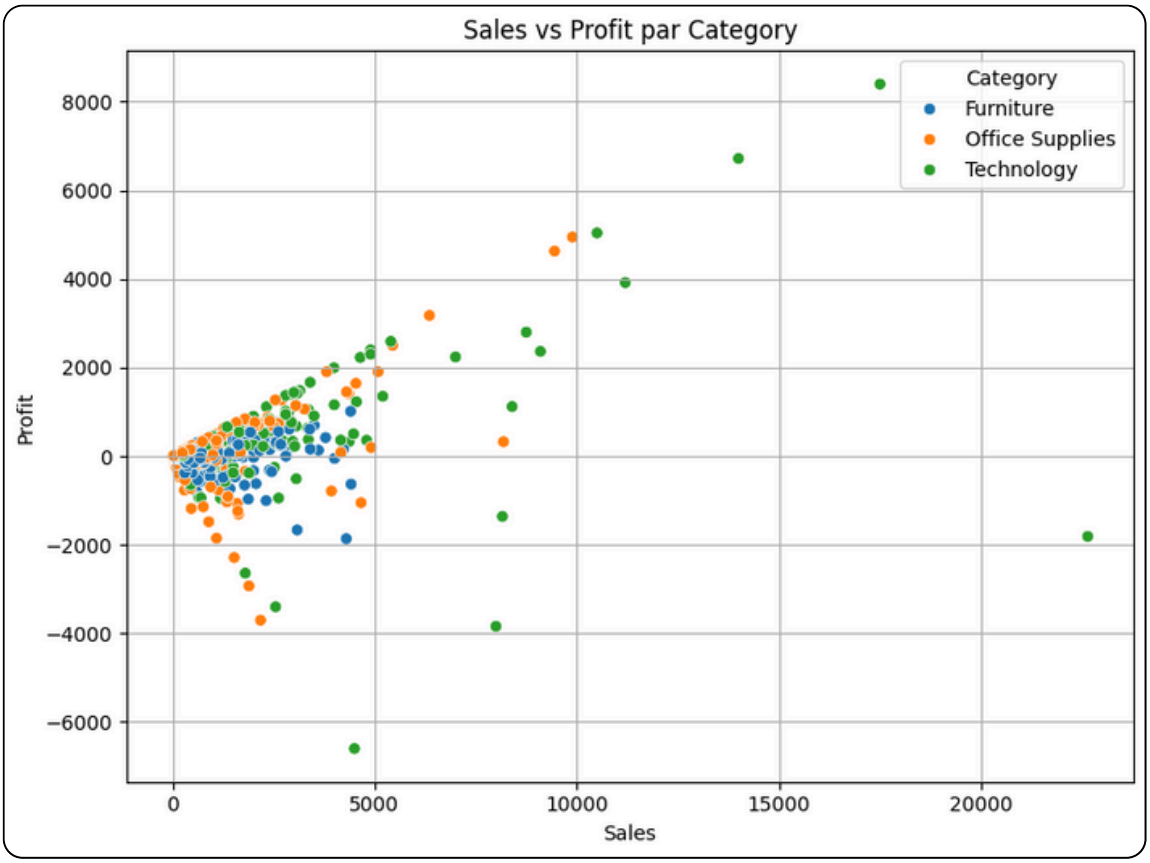

Top 10 Villes par Ventes
New York City
Los Angeles
Seattle
San Francisco
Philadelphia
Houston
Chicago
San Diego
Jacksonville
Springfield
City
0
50000
100000
150000
200000
250000
Sales
Top 10 Villes par Profits
New York City
Los Angeles
Seattle
San Francisco
Detroit
Lafayette
Jackson
Atlanta
Minneapolis
San Diego
City
0
10000
20000
30000
40000
50000
60000
Profit
Flop 10 Villes par Ventes
Abilene
Elyria
Jupiter
Pensacola
Ormond Beach
San Luis Obispo
Springdale
Layton
Keller
Missouri City
City
0
1
2
3
4
5
6
Sales
Flop 10 Villes par Profits
Philadelphia
Houston
San Antonio
Lancaster
Chicago
Burlington
Dallas
Phoenix
Aurora
Jacksonville
City
−14000
−12000
−10000
−8000
−6000
−4000
−2000
0
Profit

Utilisation des Key Perfomance Indicators (KPI)

Après l'analyse temporelle et géographique, l'étude des indicateurs clés de performance (KPI) permet de mieux comprendre les dynamiques commerciales:

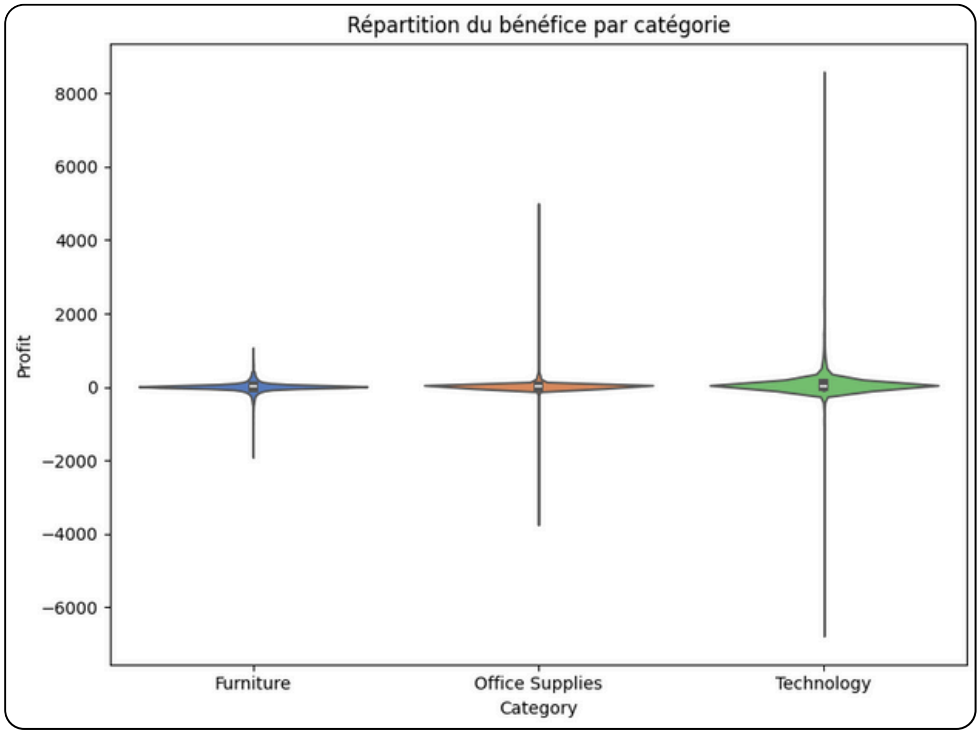
- Vente moyenne par commande : **229,86 $**
- Chiffre d'affaires total : **2,3 M$**
- Profit total : **286 397 $**
- Commande médiane : **54,49 $** → forte asymétrie
- Quantité moyenne par vente : **3,79 articles** (souvent entre 1 et 5 articles), confirmant un modèle B2C
- Articles moyens par client : **7,56**
- Délai moyen de livraison : **3,96 jours** (variant selon l'État)
- Segments clients : **5 191 Consumer**, **3 020 Corporate**, **1 783 Home Office**

Corrélations et Facteurs Explicatifs des Profits

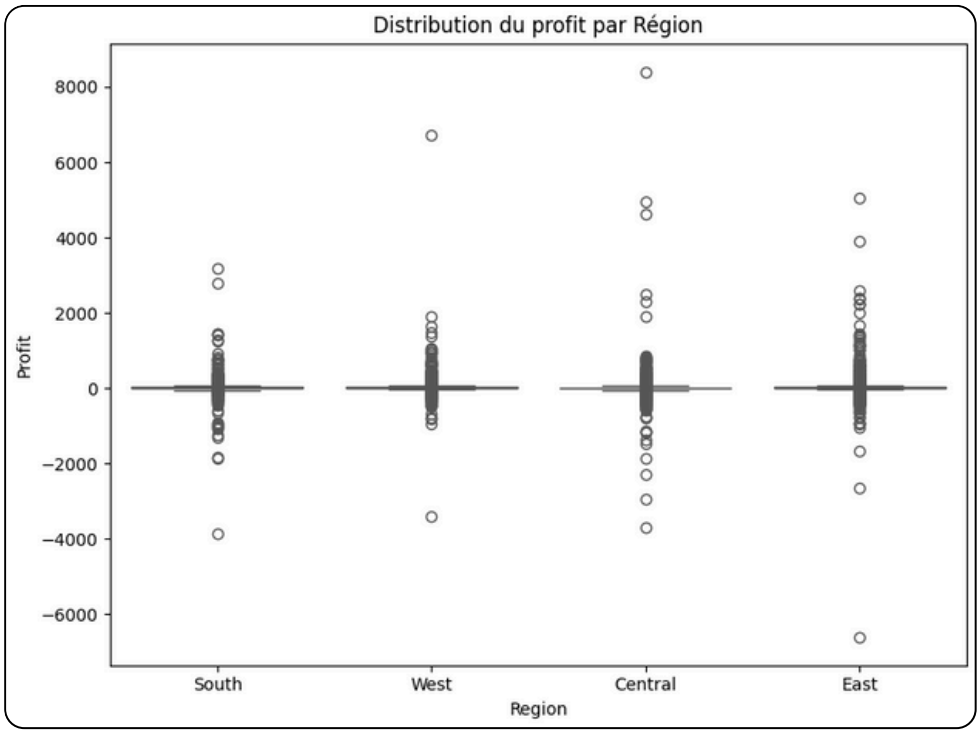


Le graphique montre une corrélation générale entre ventes élevées et profit, mais de nombreuses commandes affichent un profit négatif malgré un chiffre d'affaires important.

Ces pertes pourraient s'expliquer par **des remises trop importantes ou des coûts mal maîtrisés**, notamment dans la catégorie Office Supplies, souvent déficitaire.

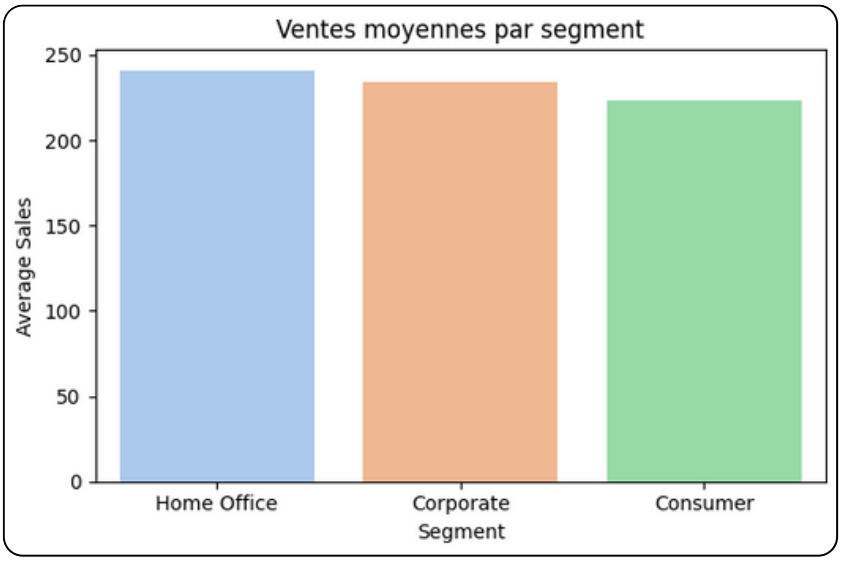


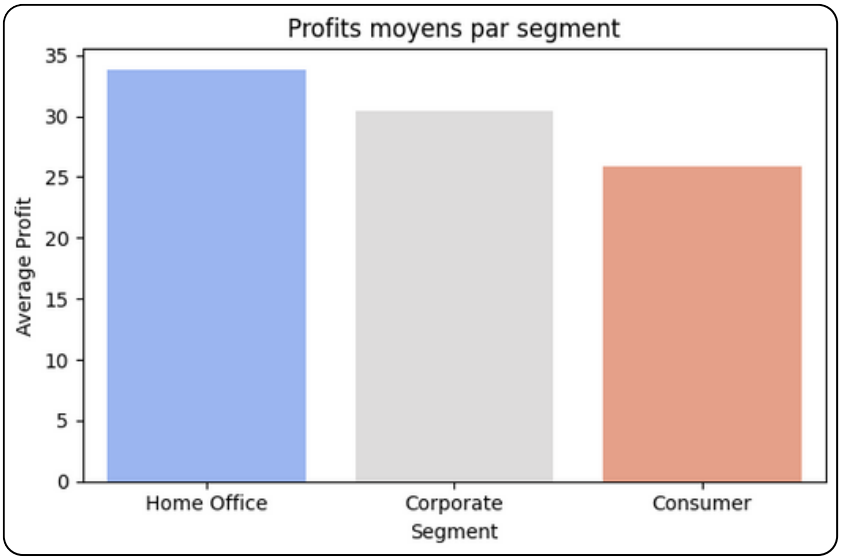
Le graphique met en évidence des profils de rentabilité contrastés selon les catégories de produits. Furniture affiche des profits plus faibles mais réguliers, traduisant une activité stable et prévisible. Office Supplies et surtout **Technology montrent une forte variabilité**, avec des ventes très rentables mais aussi de lourdes pertes. Cela suggère un potentiel de marge élevé, mais aussi un risque accru, nécessitant un **pilotage attentif des remises et des coûts**.

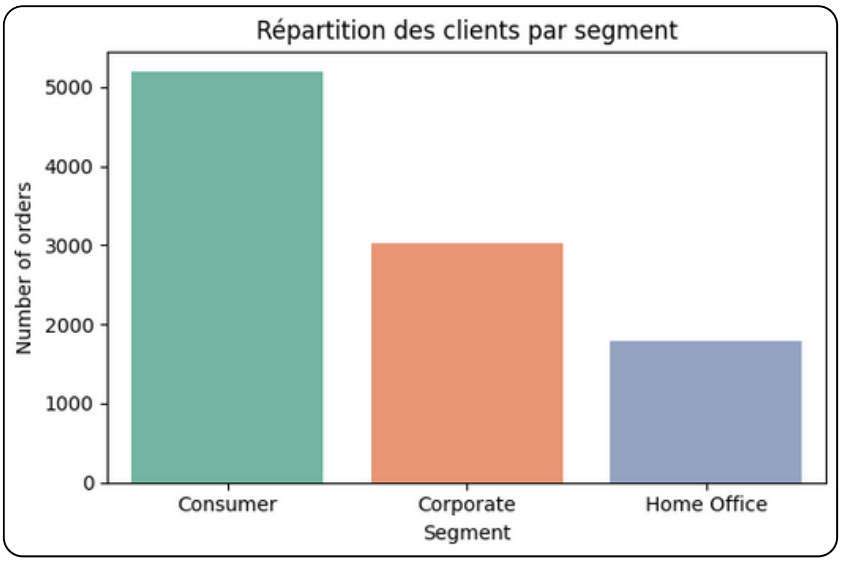


Le graphique montre que toutes les régions sont globalement rentables, mais avec de fortes disparités. West et East se démarquent par des profits élevés et stables, tandis que Central et South affichent des bénéfices plus faibles et plus irréguliers. Une attention particulière doit être portée à ces dernières, pour identifier et corriger les sources de pertes et **améliorer leur rentabilité**.

Visualisation des Performances par Segment

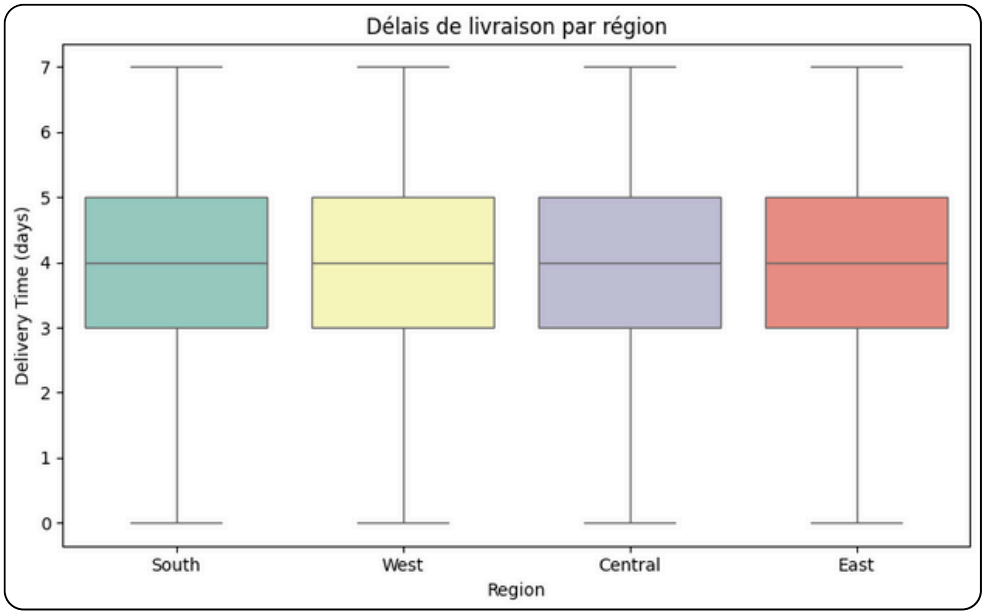


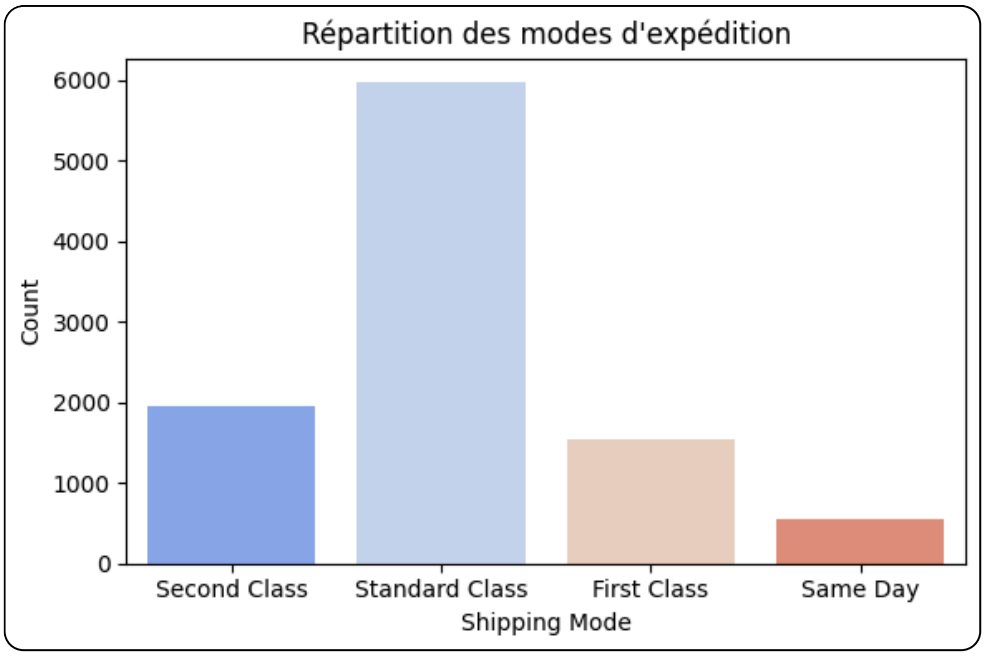




L'analyse croisée des trois graphiques révèle plusieurs enseignements clés. Le profit moyen représente environ **14,4 % du chiffre d'affaires**, avec une rentabilité relative plus élevée dans le segment Home Office. Le segment Consumer génère plus de 5 000 commandes, contre moins de 2 000 pour Home Office, mais son profit reste proportionnellement plus faible. Cela suggère que les marges y sont réduites, sans doute à cause de **remises fréquentes ou de faibles montants par commande**. À l'inverse, les segments Corporate et Home Office affichent une meilleure rentabilité, bien qu'ils enregistrent moins de ventes. Cela en fait des cibles prioritaires pour une **stratégie d'expansion**, tandis que le segment Consumer nécessiterait une **optimisation des marges par une politique commerciale mieux adaptée**.

Influence des Modes et Délais de Livraison



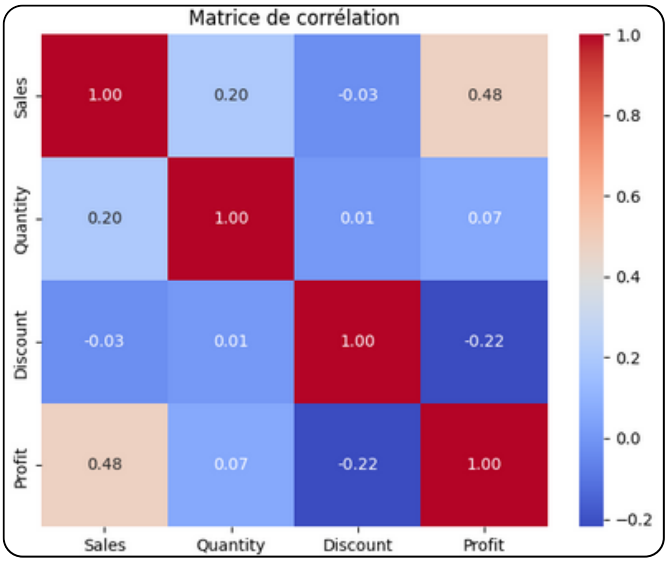


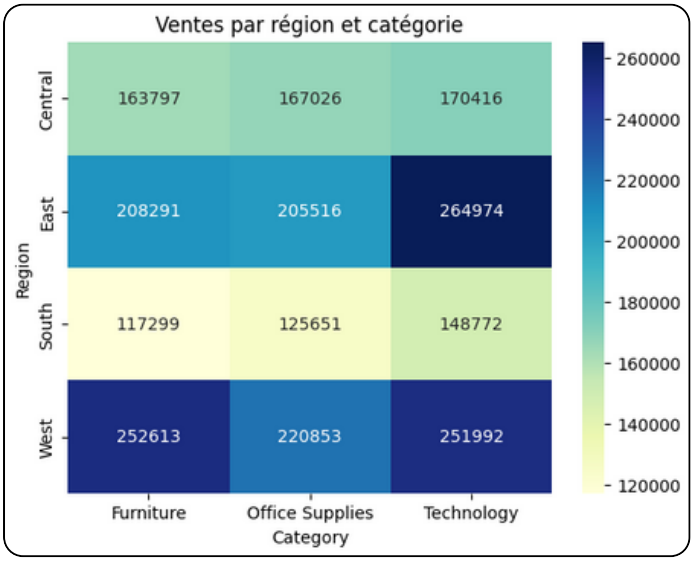
Les délais de livraison apparaissent globalement homogènes selon les régions, traduisant une logistique efficace à l'échelle nationale. Le mode Standard Class domine largement avec près de 6 000 commandes, tandis que Same Day reste marginal.

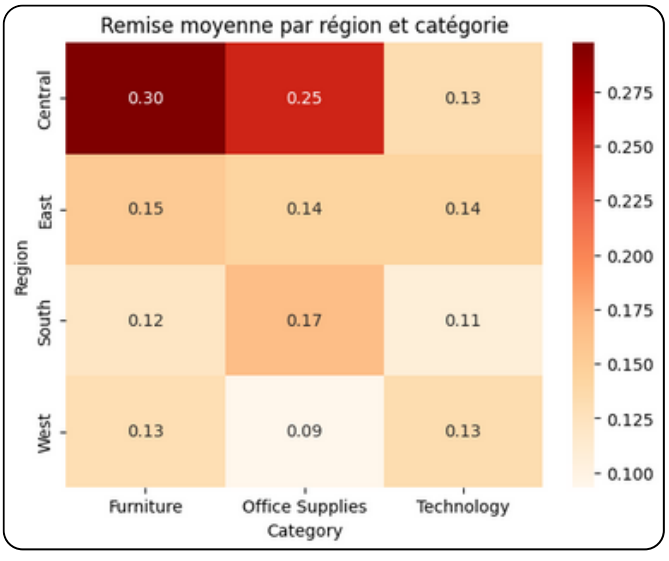
Pour optimiser la rentabilité, il serait judicieux de mieux positionner la "First Class", en la rendant plus attractive auprès des clients à forte valeur, via des **actions ciblées ou des avantages perçus renforcés**.

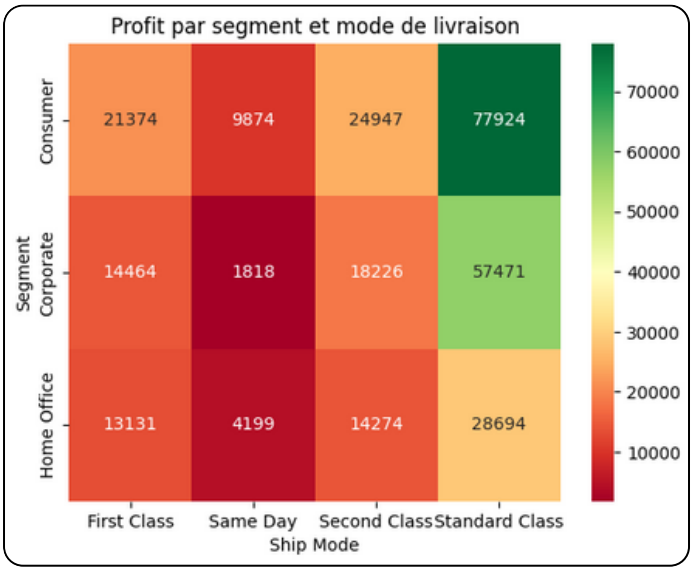
Caractéristiques des Matrices de Corrélation (Heatmap)

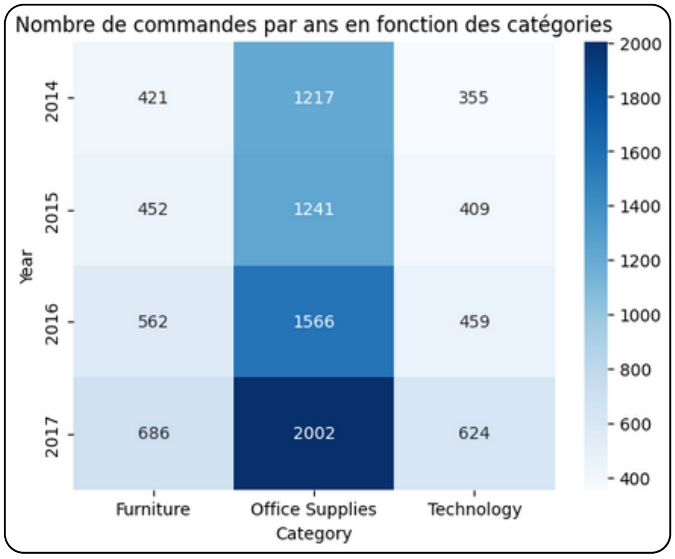
Pour mieux comprendre les relations entre les variables quantitatives, des matrices de corrélation ont été générées sous forme de heatmaps. Ces visualisations permettent d'identifier rapidement les dépendances linéaires entre les indicateurs clés. Ces analyses sont essentielles pour visualiser les axes d'amélioration.











L'analyse des cinq heatmaps révèle plusieurs points clés :

- Les remises sont trop fréquentes, nuisant surtout à la rentabilité, et nécessitent une **politique plus ciblée**.
- Le ratio ventes/profit reste globalement positif malgré quelques anomalies.
- **Augmenter la quantité moyenne par commande** permettrait d'optimiser coûts et marges.

Géographiquement :

- La **région Sud est en retard**, nécessitant des **actions marketing ou logistiques**.
- La **région Centrale applique trop de remises**, surtout sur les fournitures, ce qui réduit la rentabilité.
- La technologie subit peu de remises, cohérent avec sa forte volatilité de profit.

Par catégorie :

- Office Supplies croît rapidement, dépassant Technology et Furniture en volume.
- Furniture progresse lentement mais reste stable et moins risqué.

Enfin, par segment :

- Consumer et Corporate génèrent le plus de profits, surtout via Standard Class.
- Il serait utile de rendre la First Class plus attractive ou de **créer une offre premium plus rentable**.

Synthèse de l’Analyse des Performances Commerciales

L’étude croisée des ventes, des bénéfices et des indicateurs logistiques met en évidence plusieurs leviers d’amélioration pour optimiser la performance commerciale de l’entreprise. Si les ventes progressent globalement d’année en année, elles restent sensibles à la saisonnalité, aux politiques de remises et aux disparités régionales.

Les segments les plus rentables ne sont pas nécessairement ceux qui génèrent le plus de volume, ce qui souligne l’intérêt de cibler des stratégies différenciées selon le profil client. De même, la performance des catégories de produits appelle à privilégier la stabilité du secteur Furniture et à mieux encadrer la volatilité du secteur Technology.

Enfin, une meilleure gestion des remises, un rééquilibrage régional, et une optimisation des modes de livraison pourraient permettre à l’entreprise d’améliorer durablement son chiffre d’affaires et sa marge.

Conclusion

Cette étude a permis d’analyser les performances commerciales de l’entreprise à partir du fichier de données “Sample - Superstore.xls”, en fournissant une représentation détaillée des ventes et des marges selon les régions géographiques et les périodes temporelles. L’analyse a mis en évidence les principaux facteurs influençant les variations du chiffre d’affaires, en lien avec les catégories de produits et les segments de clientèle. Ces résultats contribuent à répondre à la problématique générale : comprendre les interactions complexes entre produits, performances régionales et marges afin d’optimiser la stratégie commerciale de l’entreprise.

Pour prolonger cette analyse, il serait pertinent de réaliser des projections quantitatives pour l’année 2019 à l’aide de modèles statistiques ou de méthodes avancées d’apprentissage automatique, telles que les réseaux de neurones récurrents (RNN) ou les réseaux à mémoire à long terme (LSTM), afin de modéliser et prédire les dynamiques des ventes et des profits. Néanmoins, il convient de souligner que ces approches ne tiennent pas compte des événements exogènes majeurs, notamment la pandémie de COVID-19 apparue en 2019, ce qui pourrait entraîner une dégradation significative de la qualité prédictive des modèles dans ce contexte.

Sources et Références Bibliographiques

[1] Dataset : Sample - Store.xls (source : Kaggle)

[2] Provost, F., & Fawcett, T. (2013): *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*

[3] McKinney, W. (2022): *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*

[4] VanderPlas, J. (2016): *Python Data Science Handbook: Essential Tools for Working with Data*