Projet d'Analyse de Données

Visualisation des Données du DataFrame Sample - Store et Optimisation des Performances Commerciales

Note de Synthèse

Réalisé par :
Lucas ALIOME

EURECOM - Sophia Antipolis
Année académique 2025

Contexte et Problématique

Dans un contexte économique où les données constituent un levier stratégique majeur, la capacité à analyser et interpréter les informations commerciales devient essentielle pour orienter les décisions d'une entreprise. Ce projet s'inscrit dans cette démarche, en exploitant le fichier de données "***Sample – Superstore.xls***", qui regroupe les ventes réalisées par une enseigne fictive répartie sur l'ensemble des États-Unis. L'objectif principal est de transformer ces données brutes en visualisations claires et en indicateurs pertinents, afin de mieux comprendre les performances commerciales de l'entreprise, d'identifier les zones de rentabilité ou de faiblesse, et de formuler des recommandations concrètes pour optimiser ses résultats.

Les données étudiées couvrent la période de 2014 à 2018. Dans le cadre de cette analyse, le contexte géopolitique américain et mondial, qui pourrait influencer l'économie, ne sera pas pris en compte. L'objectif reste de livrer une analyse pertinente des ventes à différentes échelles, d'identifier les facteurs internes pouvant impacter les résultats, et de formuler des conseils éclairés à destination de l'équipe marketing et commerciale de l'entreprise.

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.

Analyse du DataFrame

Le fichier a été chargé sous la forme d'un DataFrame à l'aide de la bibliothèque pandas, offrant une structure tabulaire facilitant l'exploration et la manipulation des données.

Dans un premier temps, il est essentiel de bien comprendre les données avant de procéder à leur analyse. Cela implique d'examiner la taille du DataFrame, le nombre d'index, ainsi que la pertinence des différentes colonnes pour l'analyse. Une attention particulière est portée à la détection des valeurs manquantes, et, le cas échéant, à la suppression des colonnes jugées inutiles afin d'optimiser la qualité du jeu de données.

Le DataFrame contient **9 994 lignes** et **21 colonnes**. Cela ne veut pas dire qu'il y 9994 clients différents, mais simplement que 9994 commandes ont été passées entre 2014 et 2018, il y a en réalité **793 clients**.

Concernant les index, les colonnes jugées pertinentes pour l'analyse incluent notamment :

- **Order ID** : identifiant unique de chaque commande ;
- **Order Date et Ship Date** : dates permettant d'étudier les délais et la saisonnalité ;
- **Customer ID et Segment** : informations clientèles ;
- **Category** : classification des produits ;
- **Sales, Quantity et Profit** : variables quantitatives clés pour l'évaluation de la performance commerciale ;
- **City, State, Region** : variables donnant des informations géographiques.

D'autres colonnes moins informatives ou redondantes seront exclues du traitement afin de garantir la pertinence et la clarté des résultats.

Analyse Exploratoire des Colonnes

Une analyse individuelle des colonnes a été menée afin d'identifier les distributions, les relations et les éventuelles anomalies présentes dans les données. Chaque variable a été examinée selon sa nature, sa contribution potentielle à l'analyse, ainsi que ses corrélations avec d'autres indicateurs clés tels que le chiffre d'affaires ou la marge bénéficiaire.

Certaines colonnes n'ont pas été interprétées dans cette étude, notamment :
Row ID, *Order ID*, *Customer ID*, *Customer Name*, *Country*, *Postal Code*, *Product ID* et *Product Name*.

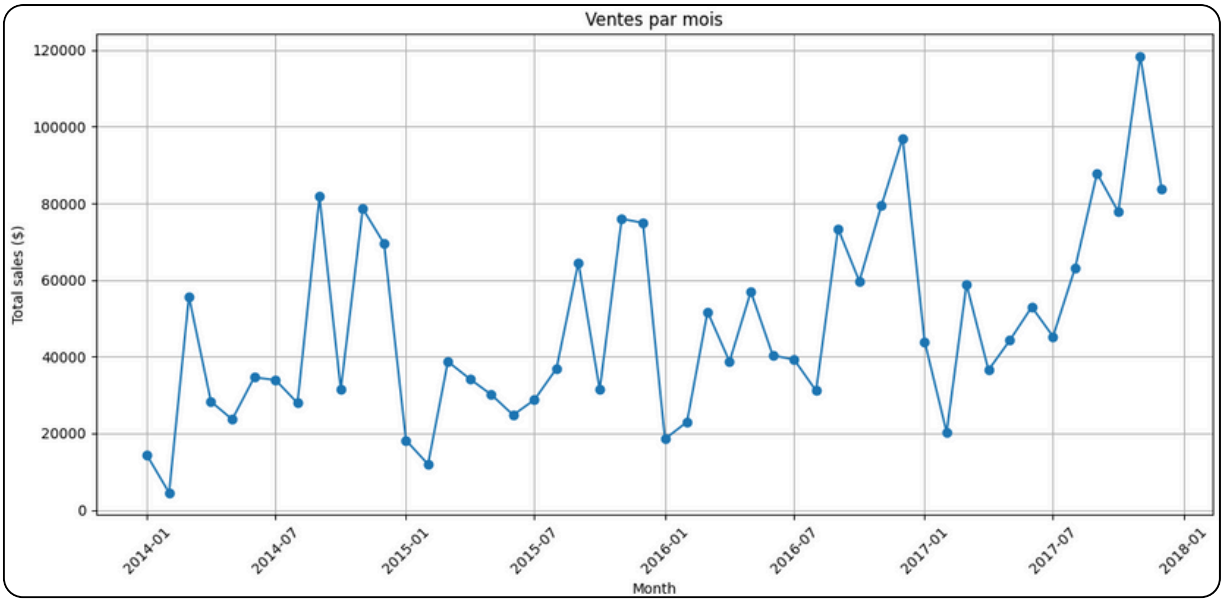
En revanche, les colonnes analysées présentent un intérêt stratégique majeur pour la prise de décision commerciale au sein de l'entreprise. Parmi celles-ci figurent notamment :

- **Segment** : catégorise les clients en Consumer, Corporate Segment et Home Office ;
- **Region** : indique la région géographique des commandes (South, West, East, Central) ;
- **Ship Mode** : précise le mode de livraison (Standard Class, Second Class, First Class).

La colonne Segment renseigne sur le type de clientèle auquel appartient chaque commande. La colonne Region permet de situer géographiquement les commandes, une analyse sera également réalisée au niveau des villes américaines. Enfin, la colonne Ship Mode décrit les différents modes de livraison utilisés.

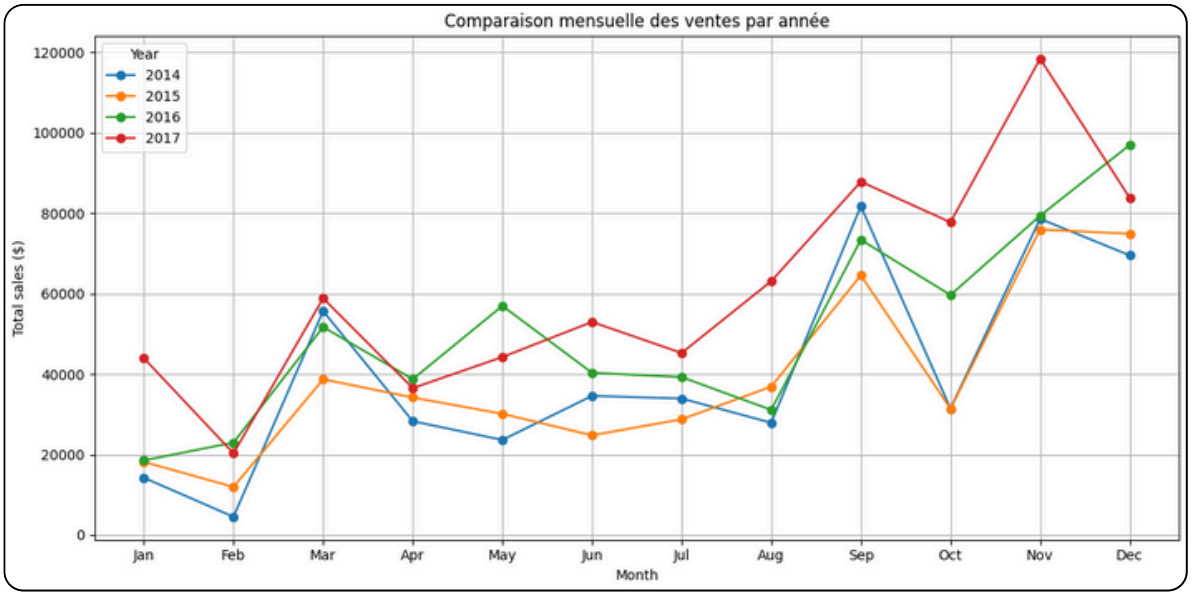
Analyse des Ventes

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.



De manière générale, 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.

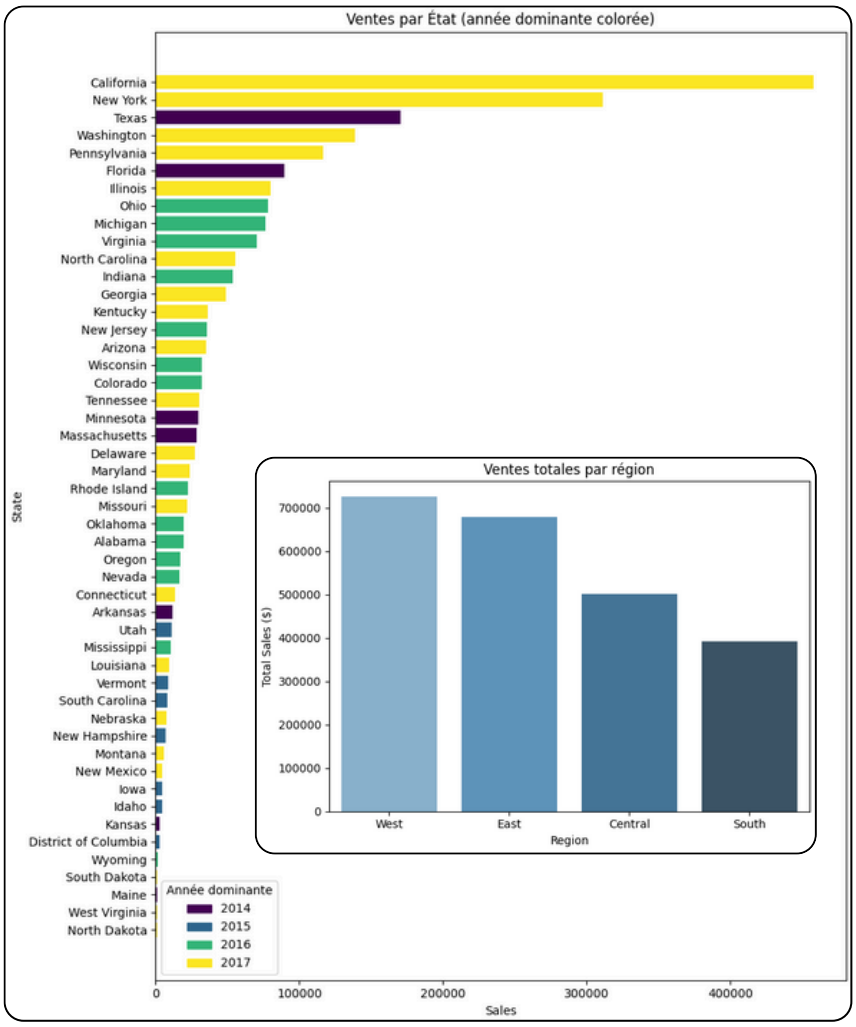
Ces variations saisonnières suggèrent une **influence notable de la temporalité sur les comportements d'achat**, ce qui pourrait orienter les décisions commerciales.



Ce graphique permet d'observer l'évolution des ventes mensuelles sur les quatre années d'activité. On remarque une croissance progressive du chiffre d'affaires au fil des années, avec des courbes qui suivent globalement une tendance ascendante. L'année 2017 se distingue par des performances particulièrement solides, surpassant les autres années sur la majorité des mois, à l'exception notable de mai et décembre.

L'analyse met en évidence un **comportement saisonnier régulier et une certaine cyclicité** :

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



Après avoir mis en évidence l'influence de la saisonnalité sur les ventes, il est tout aussi essentiel d'analyser l'impact de la dimension géographique. Les disparités entre les États sont significatives : les plus développés économiquement — tels que la Californie, le Texas ou New York — concentrent à eux seuls plus de **40 % du chiffre d'affaires total**.

Dans ces États, l'année 2017 se démarque comme la plus performante, confirmant une dynamique de croissance soutenue. À l'inverse, dans des États moins contributeurs comme l'Arkansas ou l'Iowa, les ventes restent faibles et semblent se concentrer davantage sur les premières années de la période (notamment 2014), traduisant un développement commercial plus limité.

Ces écarts peuvent s'expliquer par des facteurs exogènes tels que la densité de population, la région (Sud très en retard), la maturité du marché, ou encore l'intensité des actions marketing et publicitaires, davantage présentes dans les zones urbaines à fort potentiel

Utilisation des Key Perfomance Indicators (KPI)

Après l'analyse temporelle et géographique des ventes, il est pertinent d'examiner les Key Performances Indicators, afin de synthétiser les dynamiques commerciales et d'identifier les variables les plus influentes.

Concernant les ventes, les résultats suivants ont été observés :

- vente moyenne par commande : **229,86 $**,
- chiffre d'affaires total : **2 297 200,86 $**,
- vente la plus élevée : **638,48 $**,
- vente la plus faible : **44 $**,
- médiane des ventes : **54,49 $**, (ce qui révèle une forte asymétrie dans la distribution).

L'analyse de la variable Quantity (quantité d'articles par commande) donne également des informations importantes :

- moyenne par vente : **3,79 articles**,
- total vente : **37 873 articles**,
- vente maximale : **14 articles**,
- vente minimale : **1 article**,
- médiane : 3 **articles**.

Ces données suggèrent que la majorité des commandes sont de taille modérée, comprises entre 1 et 5 articles. Cela confirme une dynamique commerciale orientée vers les particuliers (**modèle B2C**), plutôt que vers des achats en gros (B2B).

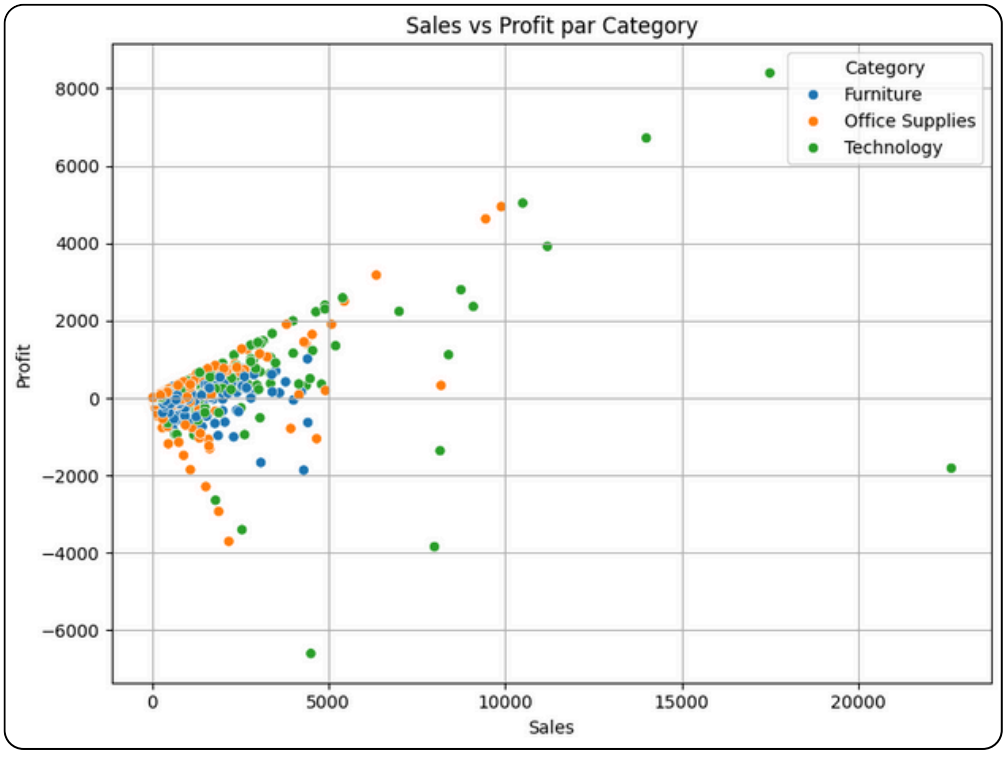
D'autres indicateurs complètent cette analyse :

- Le nombre moyen d'articles par client est estimé à **7,56** (et pas par vente ou commande comme ci-dessus),
- Le délai moyen de livraison est de **3,96 jours**, avec des disparités régionales : par exemple, 2,86 jours en moyenne pour le North Dakota contre 5,70 jours dans le District of Columbia.

Enfin, la répartition des clients par segment est la suivante : 5 191 clients dans le segment Consumer, 3 020 dans Corporate, et 1 783 dans Home Office.

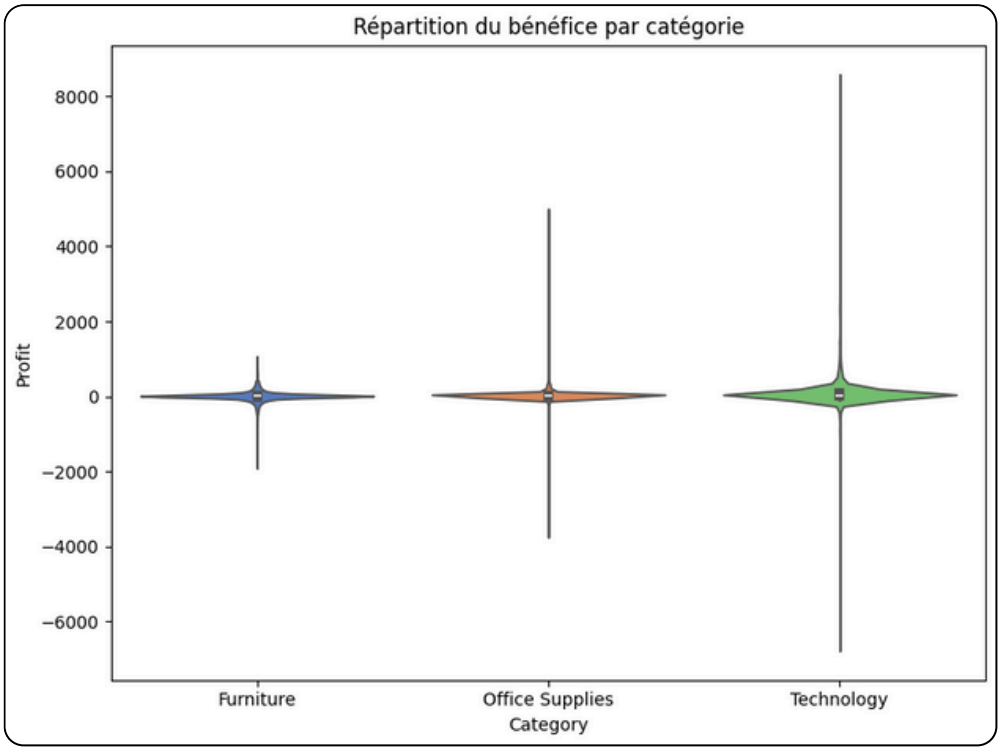
Le profit total généré s'élève à **286 397,02 $**.

Corrélations et Facteurs Explicatifs des Ventes



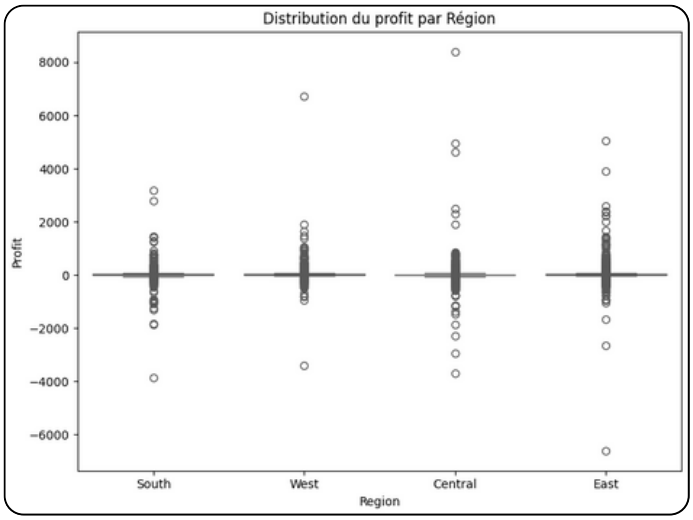
Ce graphique met en évidence la relation entre le chiffre d'affaires (Sales) et le bénéfice (Profit) pour chaque commande. Une tendance générale se dessine, suggérant qu'un volume de vente élevé est souvent corrélé à une rentabilité accrue. Cependant, un grand nombre de points apparaissent dans la zone des pertes, indiquant des ventes substantielles qui génèrent malgré tout un profit négatif.

Ces anomalies peuvent résulter de politiques de remises excessives ou d'un déséquilibre entre les coûts d'acquisition et les prix de vente. Il est également préoccupant de constater qu'un nombre significatif de produits engendrent systématiquement des marges négatives, en particulier dans la catégorie Office Supplies.



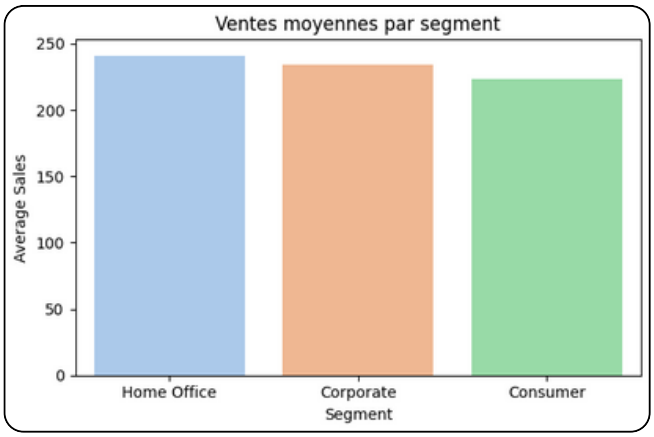
Ce graphique illustre la répartition du bénéfice (Profit) selon les trois principales catégories de produits. La catégorie Furniture se distingue par une plus grande stabilité : bien que les profits soient globalement plus faibles, la dispersion est limitée, suggérant une activité moins risquée et plus prévisible.

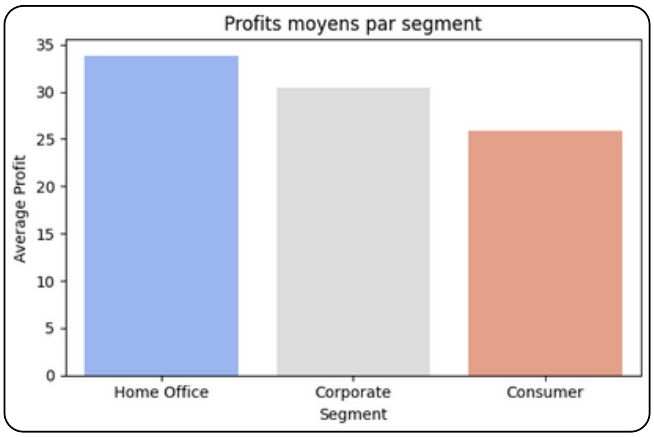
En revanche, les catégories Office Supplies et Technology présentent une forte variabilité. La catégorie Technology, en particulier, montre des profits élevés sur certaines ventes, mais aussi des pertes marquées, traduisant un comportement plus aléatoire et potentiellement risqué. Ces fluctuations importantes impliquent que, bien qu'elle soit capable de générer d'excellentes marges, cette catégorie nécessite un suivi renforcé afin de maîtriser les risques liés aux remises ou aux coûts de revient.

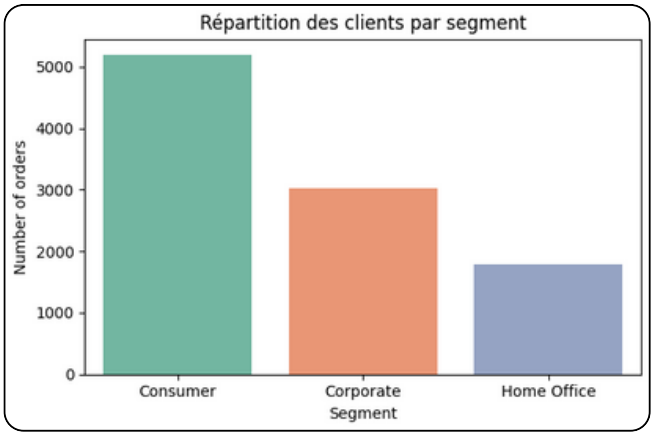


Ce graphique présente la répartition du bénéfice par région des États-Unis. L'ensemble affiche des résultats positifs, ce qui est encourageant, mais des disparités notables apparaissent (graphique page 3). Les régions West et East enregistrent les meilleures performances, avec des profits élevés et relativement homogènes, particulièrement dans la région West, où les ventes sont majoritairement rentables. À l'inverse, les régions Central et South génèrent des bénéfices plus faibles et présentent une plus grande dispersion. Ces pertes, bien que ponctuelles, doivent être analysées en détail pour en identifier les causes (remises excessives, types de produits...) et éviter qu'elles ne se reproduisent.

Il est donc essentiel pour ces régions d'optimiser leur stratégie commerciale, en ciblant les produits ou segments générateurs de marges négatives, afin de renforcer la rentabilité globale.



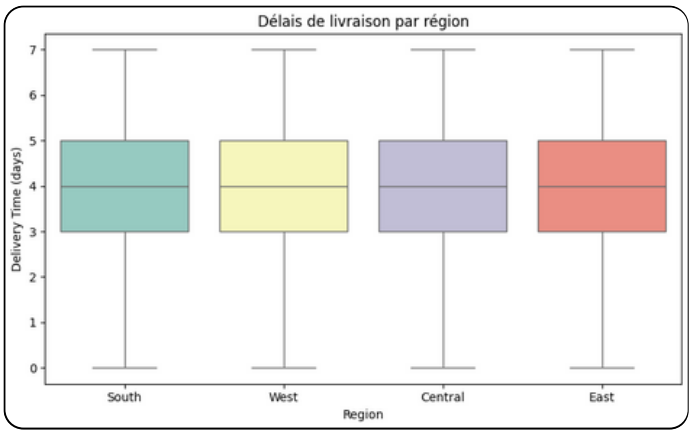


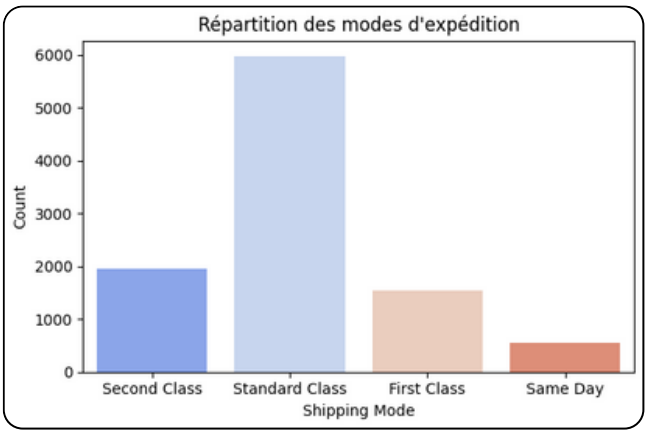


L'analyse croisée des trois graphiques permet de dégager plusieurs enseignements clés concernant la rentabilité par segment de clientèle. Dans l'ensemble, le profit moyen représente environ 14,4 % du chiffre d'affaires, avec une contribution plus significative du segment Home Office en termes de rentabilité relative.

Le segment Consumer génère à lui seul plus de 5 000 commandes, contre moins de 2 000 pour le segment Home Office. Cependant, malgré ce volume important, le profit généré par les consommateurs reste proportionnellement plus faible. Cela suggère que les marges dans ce segment sont réduites, probablement en raison de remises plus fréquentes ou de commandes de plus faible valeur.

À l'inverse, les segments Corporate et Home Office présentent un meilleur ratio de rentabilité, bien qu'ils représentent un volume de commandes inférieur. Un levier stratégique pertinent consisterait à augmenter le volume de ventes dans ces deux segments, qui démontrent déjà un bon potentiel de marge, tout en optimisant la politique commerciale appliquée au segment Consumer afin d'en améliorer la rentabilité.



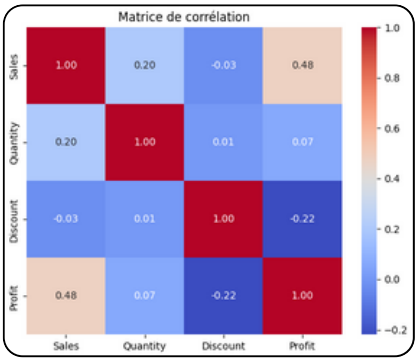


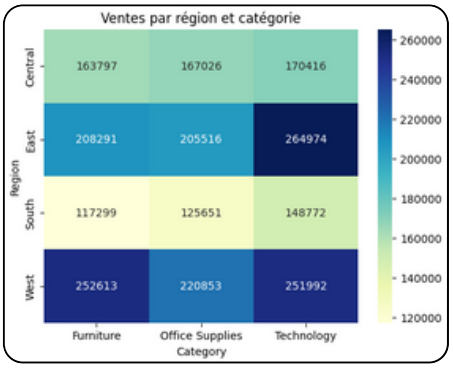
Le premier graphique montre une homogénéité des délais de livraison selon les régions, suggérant une logistique bien répartie à l'échelle nationale. Le second indique que le mode "Standard Class" domine largement, avec près de 6 000 commandes, tandis que les livraisons Same Day restent marginales.

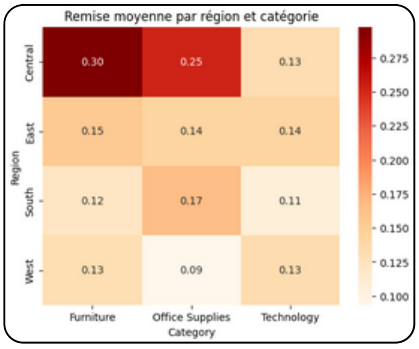
Il serait pertinent de renforcer l'attractivité de la "First Class", en ajustant son positionnement ou en ciblant les clients à forte valeur pour améliorer sa rentabilité.

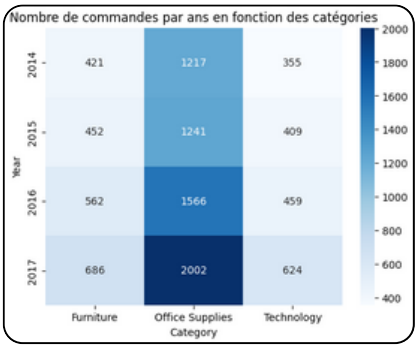
Analyse des Ventes par Matrices de Corrélation (Heatmap)

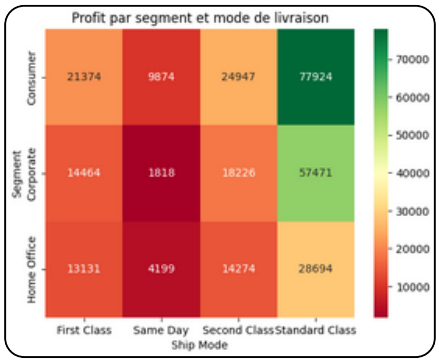
Pour mieux comprendre les relations entre les variables quantitatives, des matrices de corrélation ont été générées sous forme de heatmaps. Ces visualisations permettent d'identifier rapidement les dépendances linéaires entre les indicateurs clés. Ces analyses sont essentielles pour visualiser les axes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











L'examen des cinq heatmaps permet de dégager plusieurs enseignements clés :

- Les remises (Discount) sont trop fréquemment appliquées, ce qui impacte négativement à la fois les ventes et surtout la rentabilité. Une politique de réduction plus ciblée semble nécessaire pour éviter les effets délétères sur le long terme.
- Le ratio ventes/profit reste globalement satisfaisant, ce qui montre que l'activité génère des marges positives malgré certaines anomalies ponctuelles.
- Il serait pertinent de favoriser une augmentation de la quantité moyenne par commande, afin d'optimiser les coûts logistiques et les bénéfices unitaires.

Sur le plan géographique :

- La région Sud reste en retrait par rapport aux autres en termes de performance commerciale ; des efforts marketing ou logistiques pourraient y être envisagés pour stimuler les ventes.
- La région Centrale applique un niveau de remise trop élevé, en particulier sur les fournitures, ce qui compromet la rentabilité locale.
- À l'inverse, la technologie bénéficie de très peu de remises, ce qui est cohérent compte tenu de sa forte volatilité en termes de profit.

Concernant les catégories de produits :

- Les Office Supplies connaissent une croissance notable, en particulier en 2017, et surpassent même la Technology et le Furniture en volume de ventes.
- Le secteur Furniture progresse lentement mais reste le plus stable, ce qui le rend moins risqué malgré un potentiel de croissance limité.

Enfin, du côté des segments :

- Consumer et Corporate Segment génèrent le plus de profit, en grande majorité via le mode de livraison Standard Class.
- Il pourrait être intéressant de diversifier les offres de livraison, soit en rendant la First Class plus attractive, soit en développant une nouvelle offre premium mieux calibrée pour la rentabilité.

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 jeu de données Sample - Store.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.