École Centrale Casablanca

Analyse Financière avec Data Science & Machine Learning

Auteurs:
Salah Eddine EL AZOUTI
Salma Saaidi
Anas EL HAYEL
Walid EL BOUCHTI
Aasma ouamalich

15 avril 2025

Table des matières

Résumé

Ce projet porte sur l'analyse financière approfondie d'entreprises américaines à l'aide de techniques de Data Science et de Machine Learning. À partir d'un jeu de données contenant 200 indicateurs financiers d'entreprises américaines entre 2014 et 2018, nous avons réalisé une exploration complète, calculé des ratios financiers, identifié des clusters d'entreprises aux profils similaires, construit des modèles prédictifs et interprété les résultats d'un point de vue économique.

Le présent rapport détaille la méthodologie utilisée, les résultats obtenus et les conclusions économiques tirées de cette analyse. Ce travail démontre comment les techniques modernes d'analyse de données peuvent être appliquées au domaine financier pour extraire des informations pertinentes et actionables.

Chapitre 1

Introduction

1.1 Contexte

L'analyse financière traditionnelle repose sur l'examen manuel d'un nombre limité de ratios et d'indicateurs financiers. Avec l'avènement du Big Data et de l'intelligence artificielle, il est désormais possible d'analyser simultanément des centaines d'indicateurs et d'identifier des motifs que l'œil humain ne pourrait discerner.

Ce projet s'inscrit dans cette démarche en proposant une analyse financière augmentée par des techniques de Data Science et de Machine Learning. L'objectif est d'exploiter la puissance de ces outils pour extraire des insights précieux à partir d'un vaste ensemble de données financières.

1.2 Objectifs

Les principaux objectifs de ce projet sont les suivants :

- Explorer et nettoyer un dataset contenant 200 indicateurs financiers d'entreprises américaines
- Calculer et analyser des ratios financiers clés (ROA, ROE, marge nette, etc.)
- Identifier des groupes d'entreprises aux profils financiers similaires à l'aide de techniques de clustering
- Construire des modèles prédictifs pour anticiper les performances financières
- Interpréter les résultats d'un point de vue économique et financier
- Proposer des recommandations stratégiques basées sur l'analyse

1.3 Structure du rapport

Ce rapport est structuré comme suit :

- **Chapitre 2** : Présentation du jeu de données et exploration initiale
- **Chapitre 3** : Calcul et visualisation des ratios financiers
- **Chapitre 4** : Analyse en composantes principales et clustering
- **Chapitre 5** : Modélisation prédictive
- **Chapitre 6** : Conclusions et interprétations économiques

Chapitre 2

Données et exploration initiale

2.1 Présentation du jeu de données

Le jeu de données utilisé dans ce projet contient 200 indicateurs financiers pour un ensemble d'entreprises américaines cotées en bourse sur la période 2014-2018. Ces indicateurs couvrent divers aspects de la performance financière, notamment :

- Informations générales (nom de l'entreprise, secteur, etc.)
- Métriques de revenus et de profits
- Informations sur les actifs et les passifs
- Indicateurs de performance boursière

2.2 Exploration des données

2.2.1 Structure générale

L'exploration initiale des données nous a permis de comprendre la structure générale du jeu de données, notamment le nombre d'entreprises, la distribution par secteur et les statistiques descriptives des principaux indicateurs financiers.

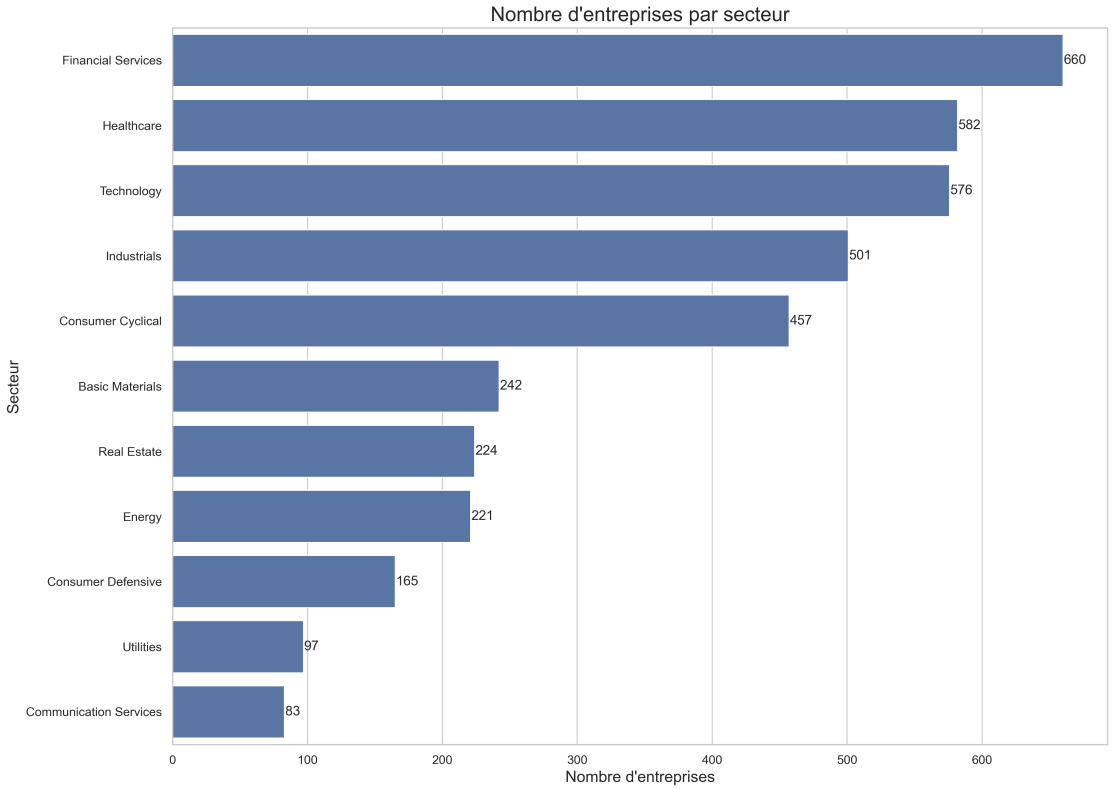

FIGURE 2.1 – Distribution des entreprises par secteur

La figure 2.1 montre la répartition des entreprises par secteur dans notre jeu de données. On constate que certains secteurs sont plus représentés que d'autres, ce qui est important à prendre en compte lors de l'interprétation des résultats d'analyse.

2.2.2 Traitement des valeurs manquantes

L'analyse des valeurs manquantes est une étape cruciale dans la préparation des données pour l'analyse. Nous avons identifié les colonnes présentant des valeurs manquantes et appliqué des stratégies appropriées pour les traiter.

2.2.3 Détection et traitement des valeurs aberrantes

Les valeurs aberrantes peuvent fortement influencer les résultats des analyses statistiques et des modèles de machine learning. Nous avons mis en place des méthodes pour les détecter et les traiter de manière appropriée.

Chapitre 3

Ratios financiers et visualisations

3.1 Calcul des ratios financiers

Nous avons calculé plusieurs ratios financiers clés pour enrichir notre analyse :
— **ROA (Return on Assets)** : Mesure la rentabilité des actifs
— **ROE (Return on Equity)** : Mesure la rentabilité des capitaux propres
— **Marge nette** : Rapport entre le bénéfice net et le chiffre d'affaires
— **Ratio d'endettement** : Mesure du niveau d'endettement par rapport aux actifs
— **Ratio de liquidité** : Capacité à honorer les dettes à court terme

3.2 Analyse des performances par secteur

La performance financière varie considérablement selon les secteurs d'activité. Nous avons analysé comment les différents indicateurs et ratios se comportent à travers les divers secteurs représentés dans notre jeu de données.

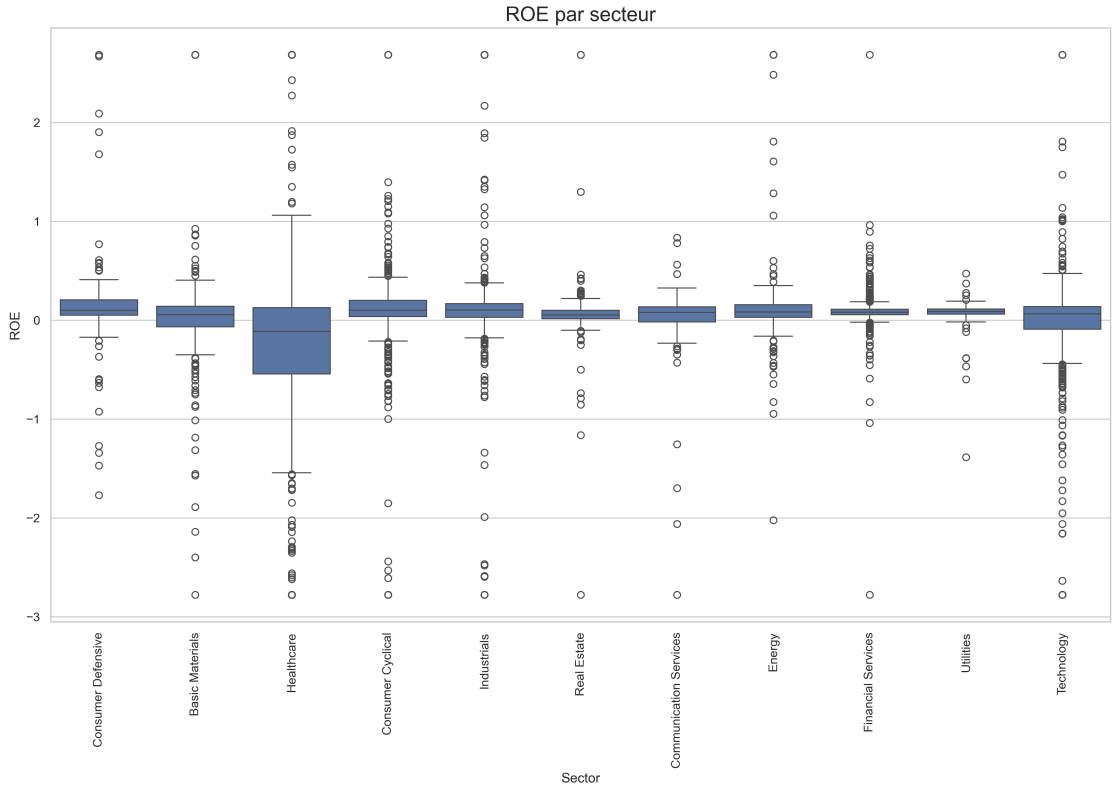

FIGURE 3.1 – Rentabilité des capitaux propres (ROE) par secteur

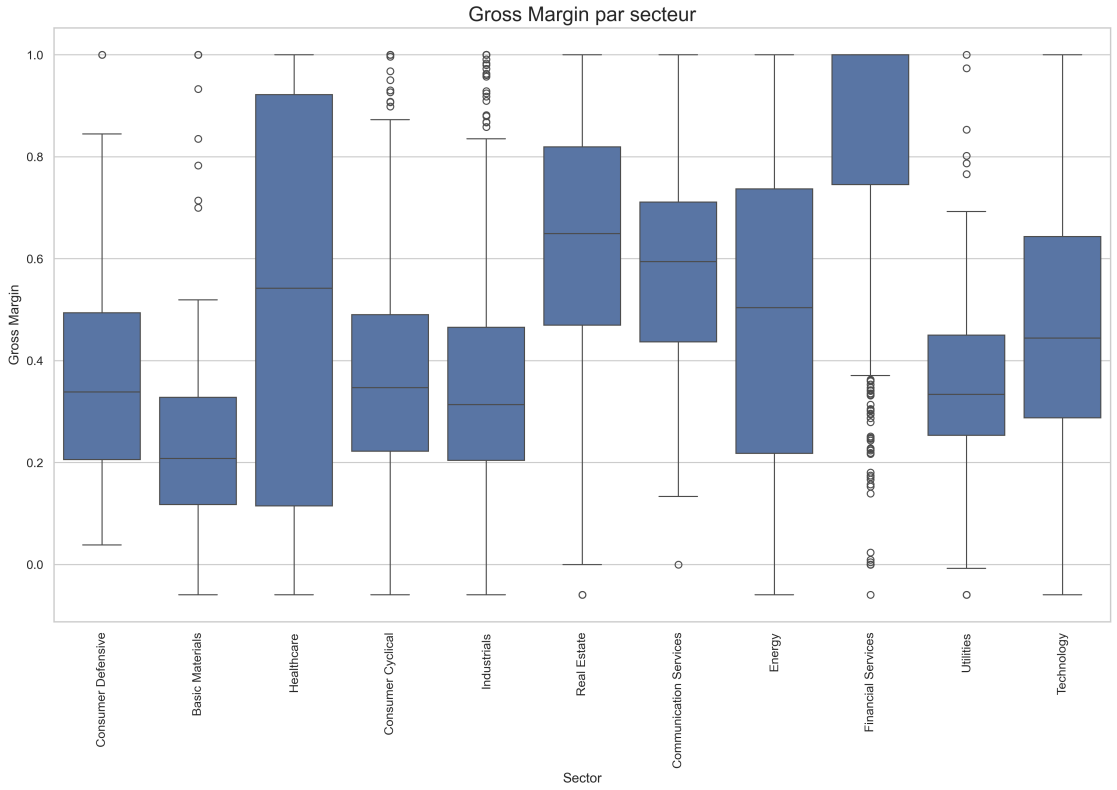

FIGURE 3.2 – Marge brute par secteur

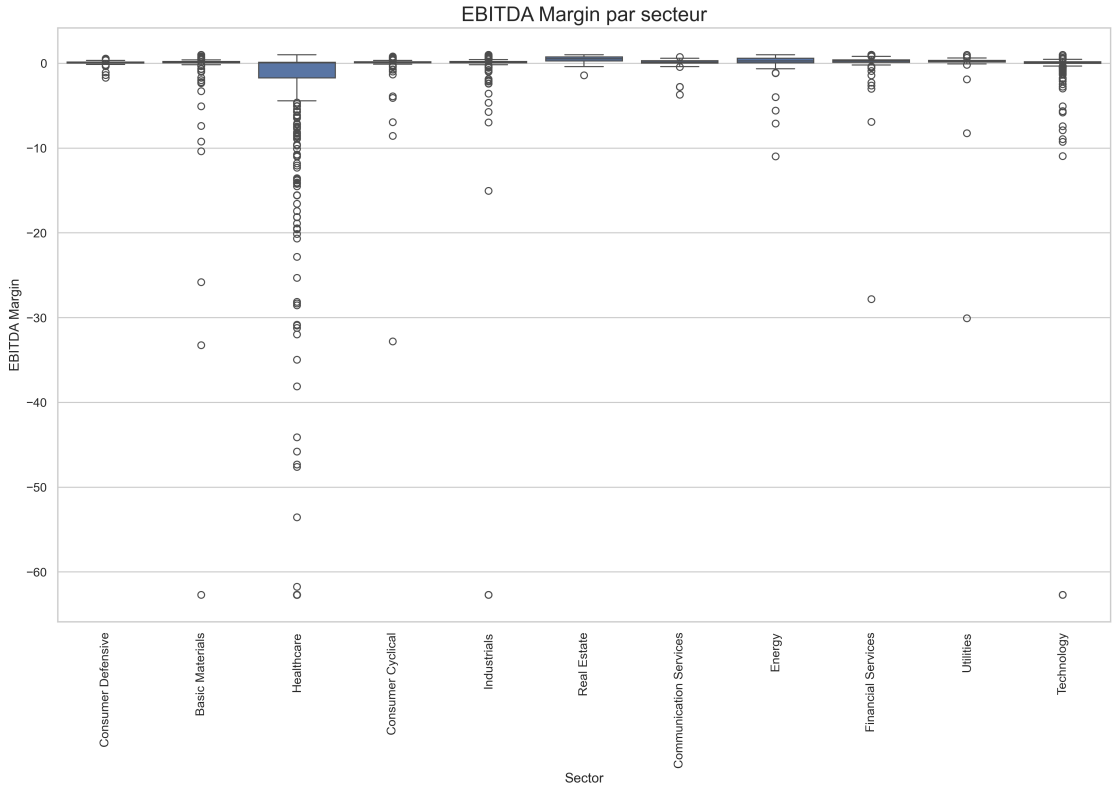

FIGURE 3.3 – Marge EBITDA par secteur

Ces visualisations (figures 3.1, 3.2 et 3.3) mettent en évidence les différences significatives de performance financière entre les secteurs. On observe notamment que certains secteurs présentent des ratios de rentabilité plus élevés et plus stables que d'autres, ce qui peut être attribué à diverses caractéristiques sectorielles telles que les barrières à l'entrée, l'intensité concurrentielle et les structures de coûts.

3.3 Corrélation entre variables financières

L'étude des corrélations entre les différentes variables financières permet d'identifier des relations potentiellement significatives et d'orienter la suite de l'analyse.

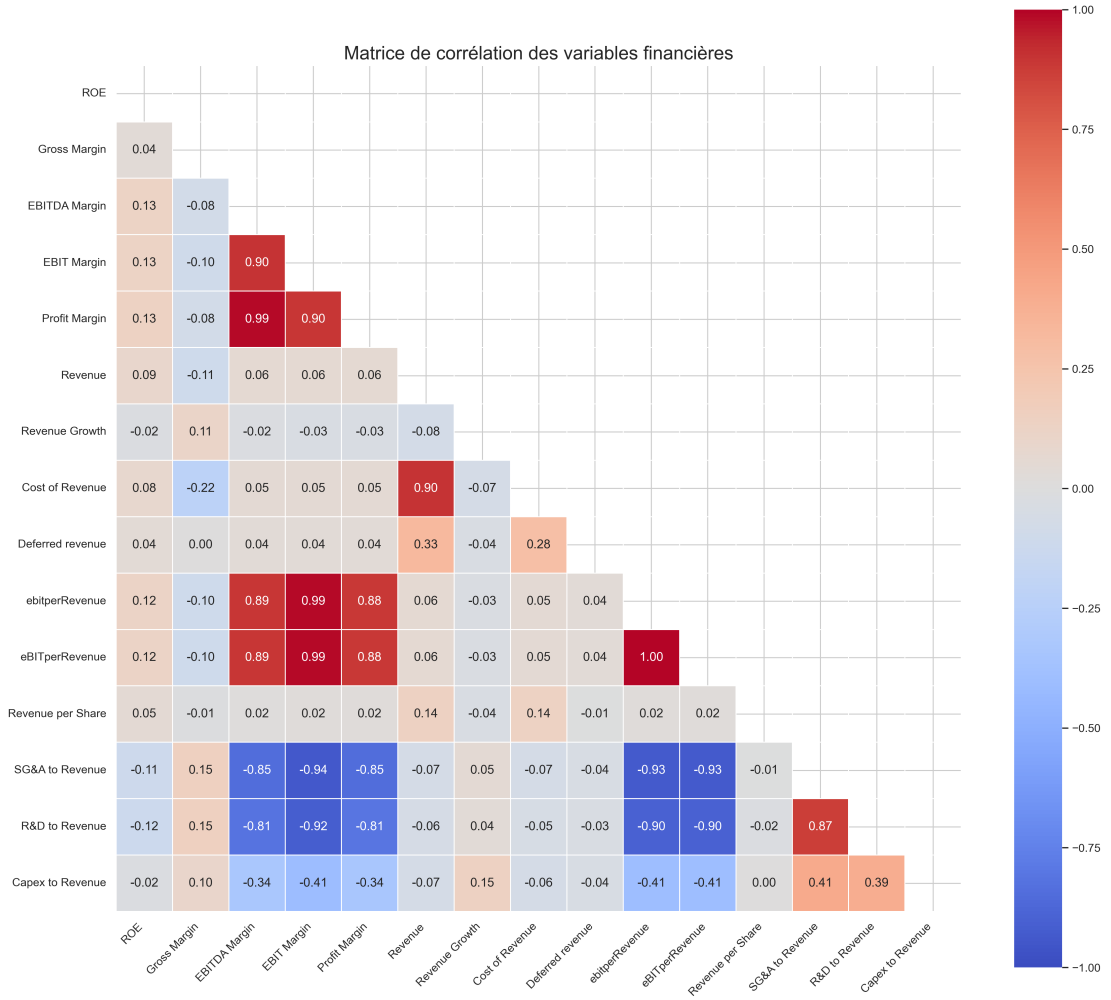

FIGURE 3.4 – Matrice de corrélation des principales variables financières

La matrice de corrélation (figure 3.4) révèle plusieurs relations importantes entre les variables financières. Les corrélations positives fortes (en bleu foncé) indiquent des variables qui évoluent généralement dans le même sens, tandis que les corrélations négatives (en rouge) montrent des variables qui tendent à évoluer en sens inverse. Ces informations sont cruciales pour comprendre les interactions entre différents aspects de la performance financière des entreprises.

Chapitre 4

Analyse en composantes principales et clustering

4.1 Réduction de dimension par ACP

L'Analyse en Composantes Principales (ACP) est une technique statistique puissante qui permet de réduire la dimensionnalité d'un jeu de données tout en conservant le maximum d'information. Nous l'avons appliquée à notre ensemble de données pour identifier les principales sources de variance.

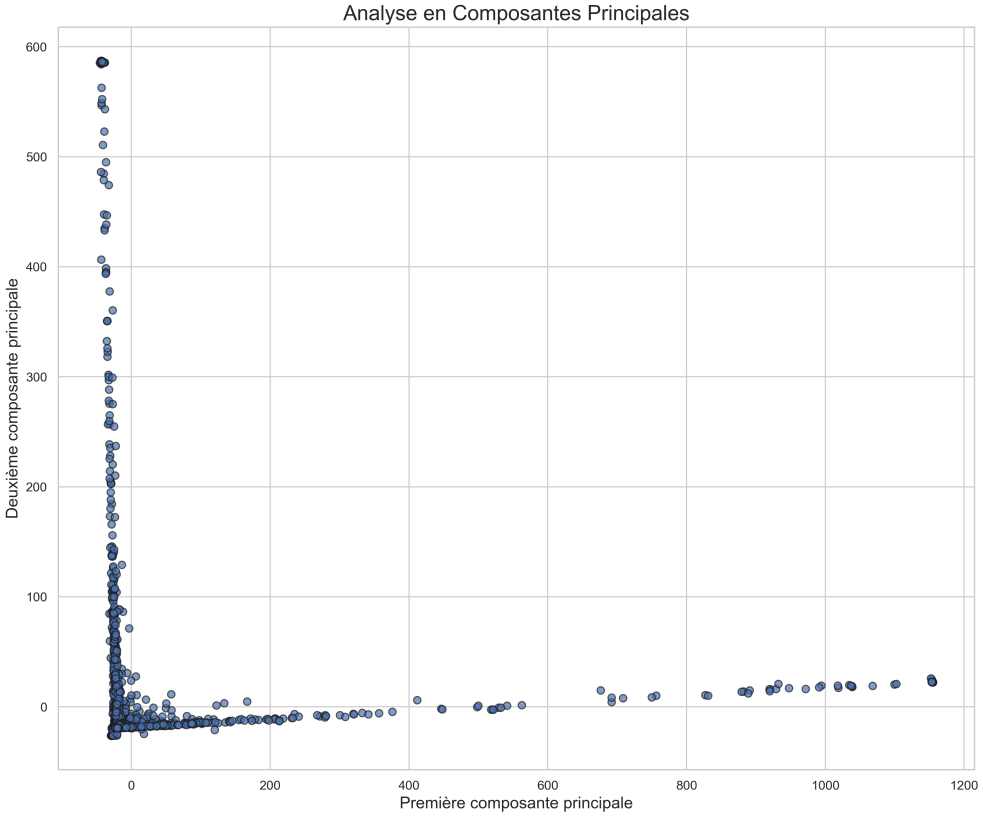

FIGURE 4.1 – Projection des entreprises sur les deux premières composantes principales

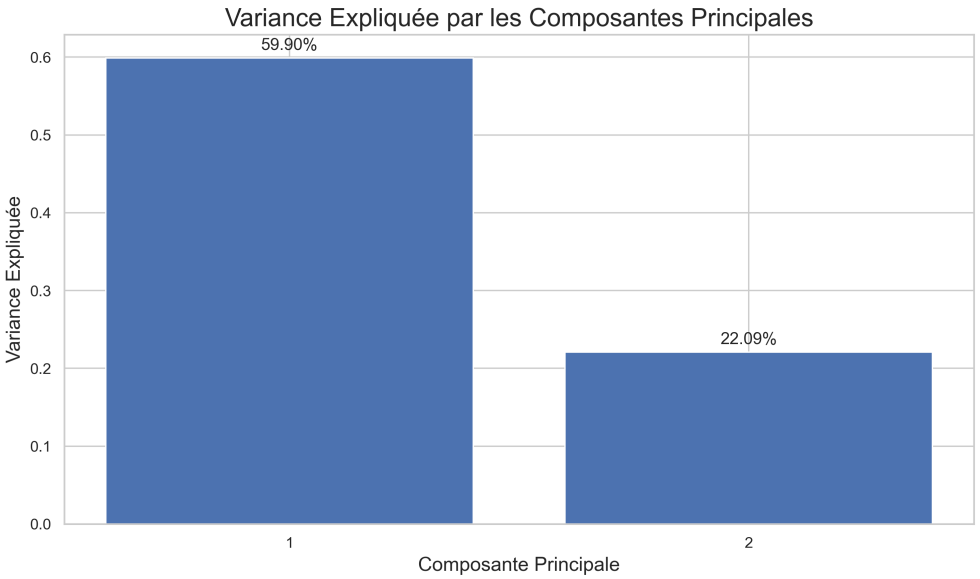

FIGURE 4.2 – Variance expliquée par les composantes principales

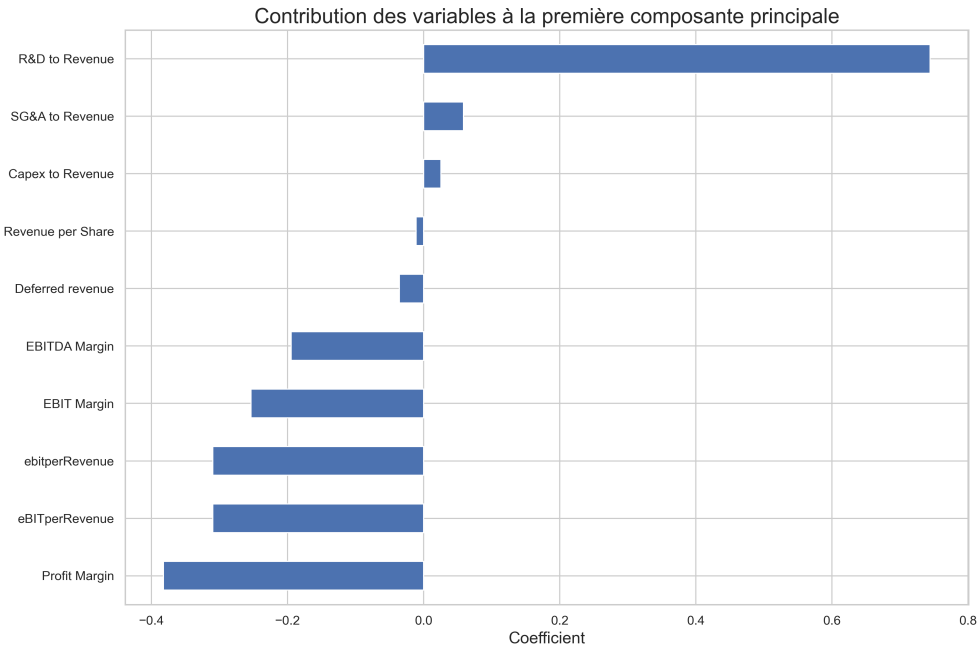

FIGURE 4.3 – Contribution des variables à la première composante principale

La figure 4.1 montre la projection des entreprises sur les deux premières composantes principales, qui captent la plus grande partie de la variance dans les données. La figure 4.2 illustre la proportion de variance expliquée par chaque composante principale. La figure 4.3 présente les variables qui contribuent le plus à la première composante principale, révélant ainsi les facteurs financiers les plus discriminants dans notre jeu de données. Cette analyse nous permet de réduire la complexité des données tout en conservant l'information essentielle, facilitant ainsi la visualisation et l'interprétation des patterns dans les données financières.

4.2 Classification des entreprises par K-Means

L'algorithme K-Means nous a permis de regrouper les entreprises en clusters selon leurs profils financiers. Cette approche non supervisée révèle des groupes naturels d'entreprises ayant des caractéristiques similaires.

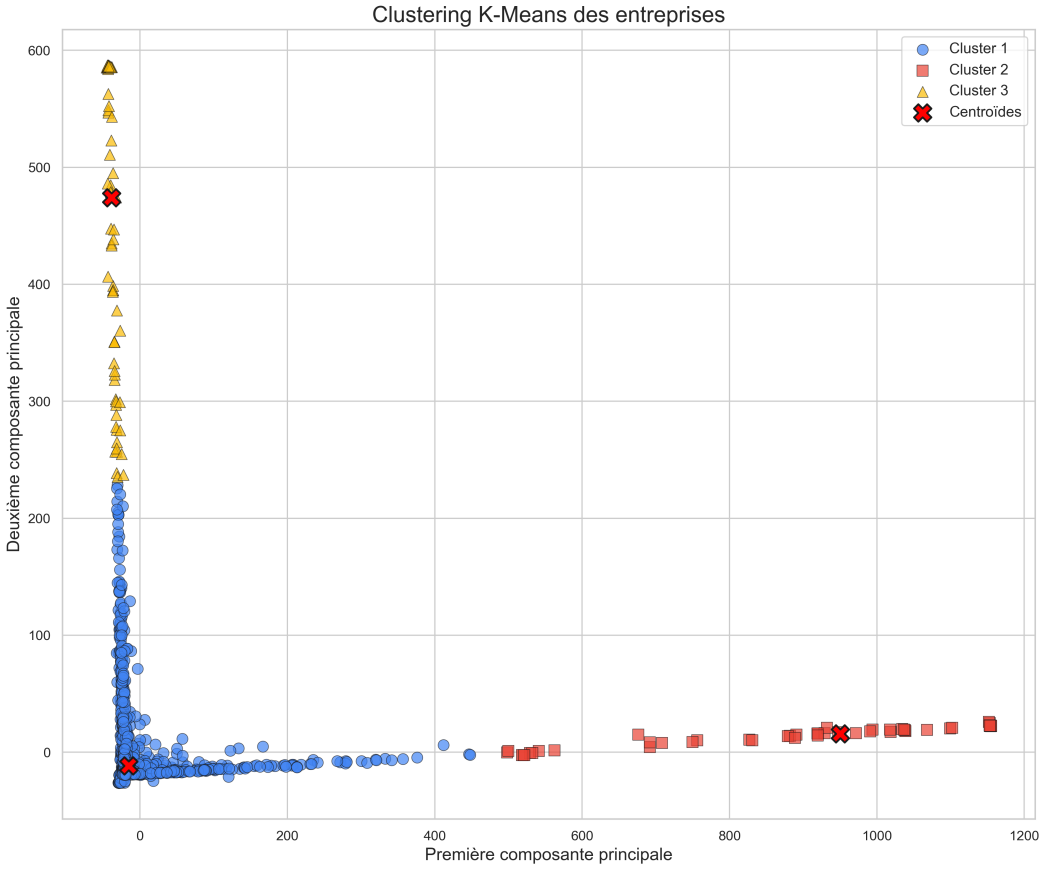

FIGURE 4.4 – Clusters d'entreprises identifiés par l'algorithme K-Means

4.3 Interprétation des clusters

Une fois les clusters identifiés, nous avons analysé leurs caractéristiques pour comprendre ce qui distingue chaque groupe et proposer une interprétation financière et économique.

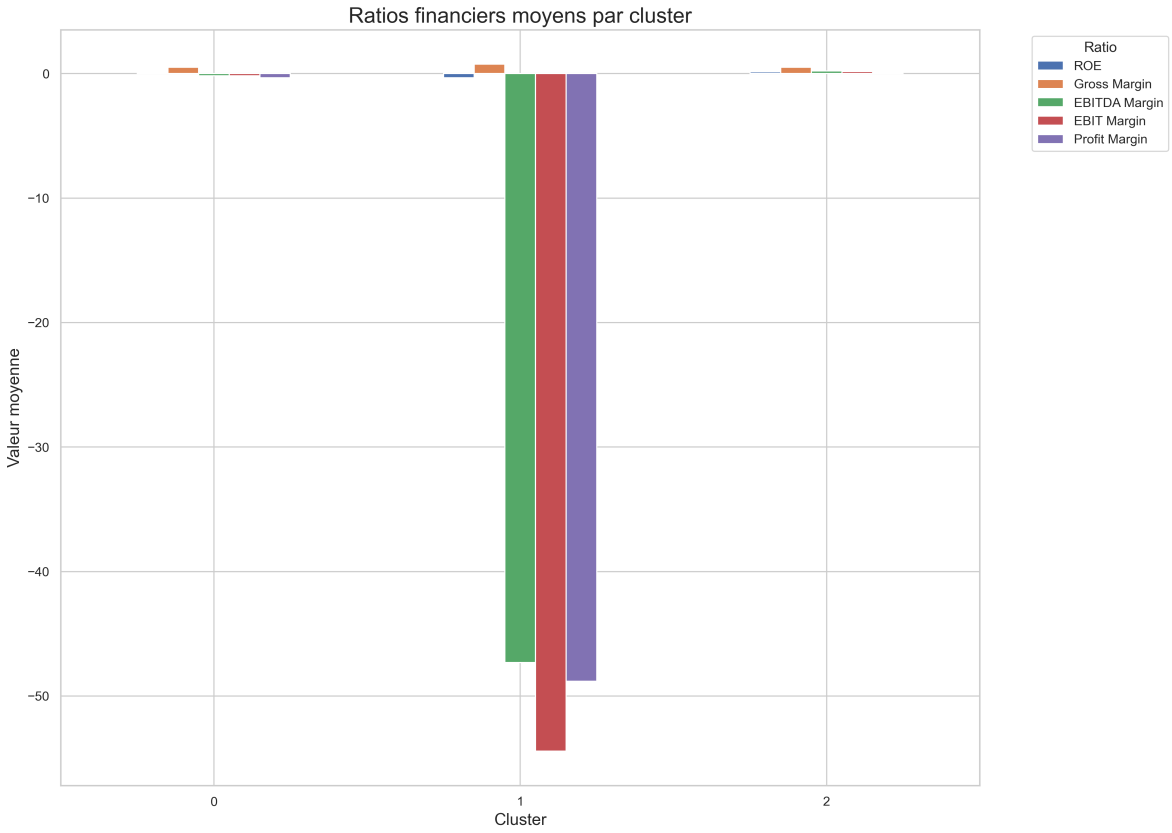

FIGURE 4.5 – Ratios financiers moyens par cluster

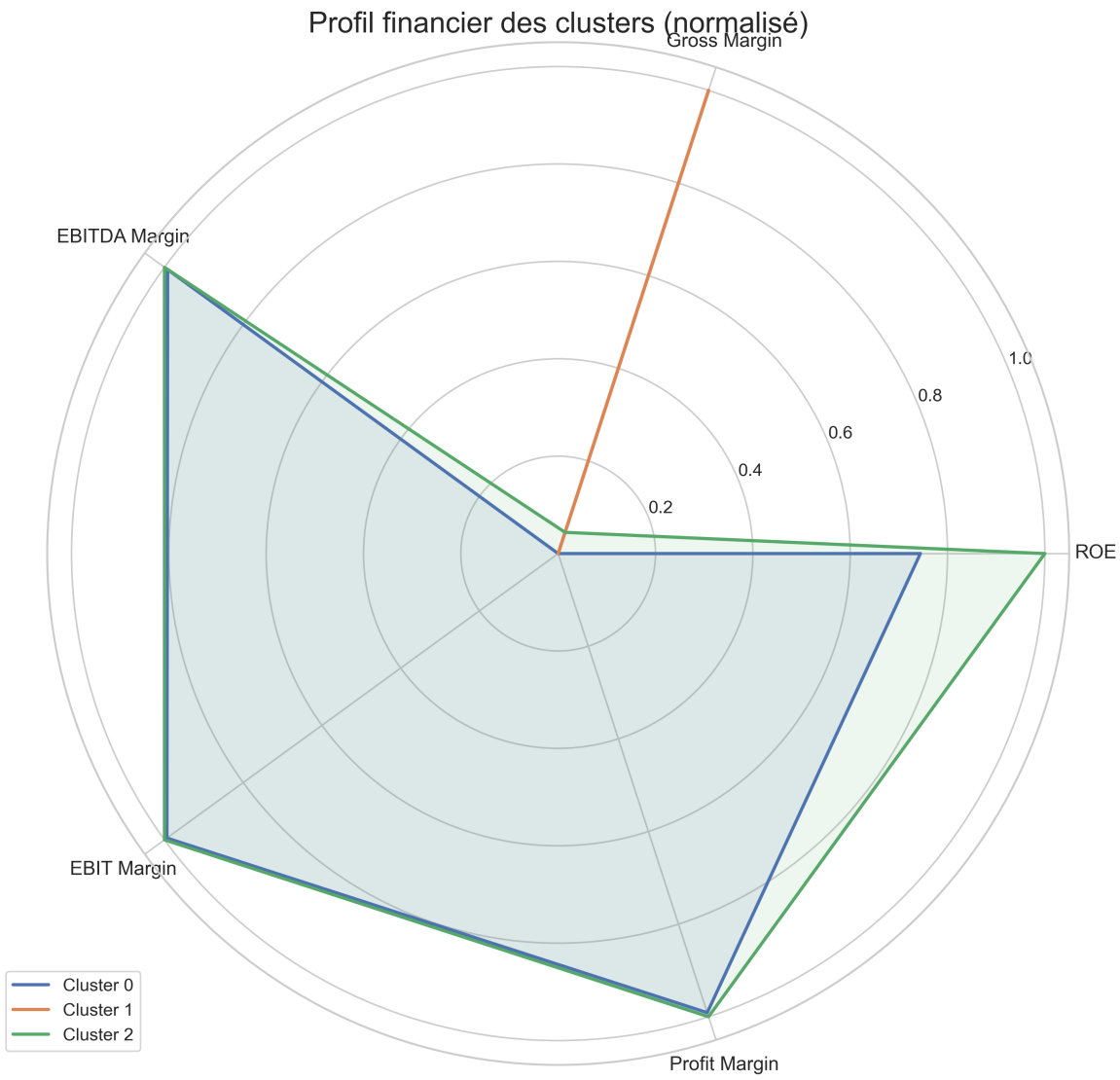

FIGURE 4.6 – Profil financier des clusters (normalisé)

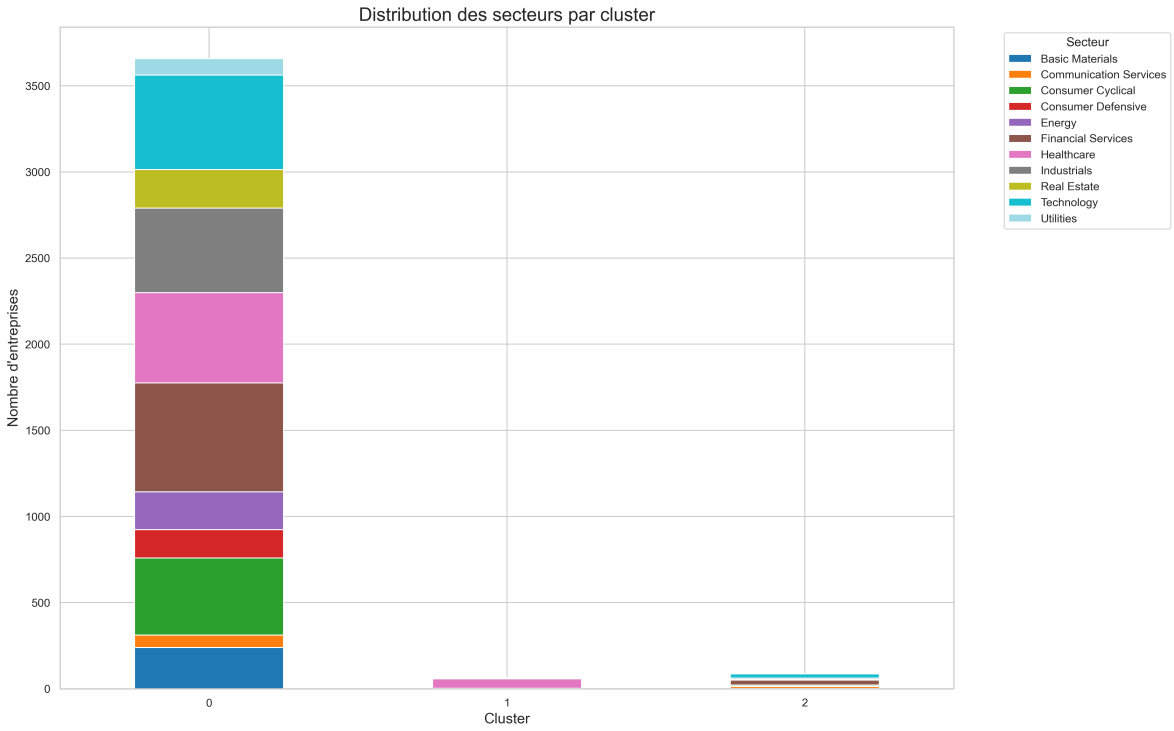

FIGURE 4.7 – Distribution des secteurs par cluster

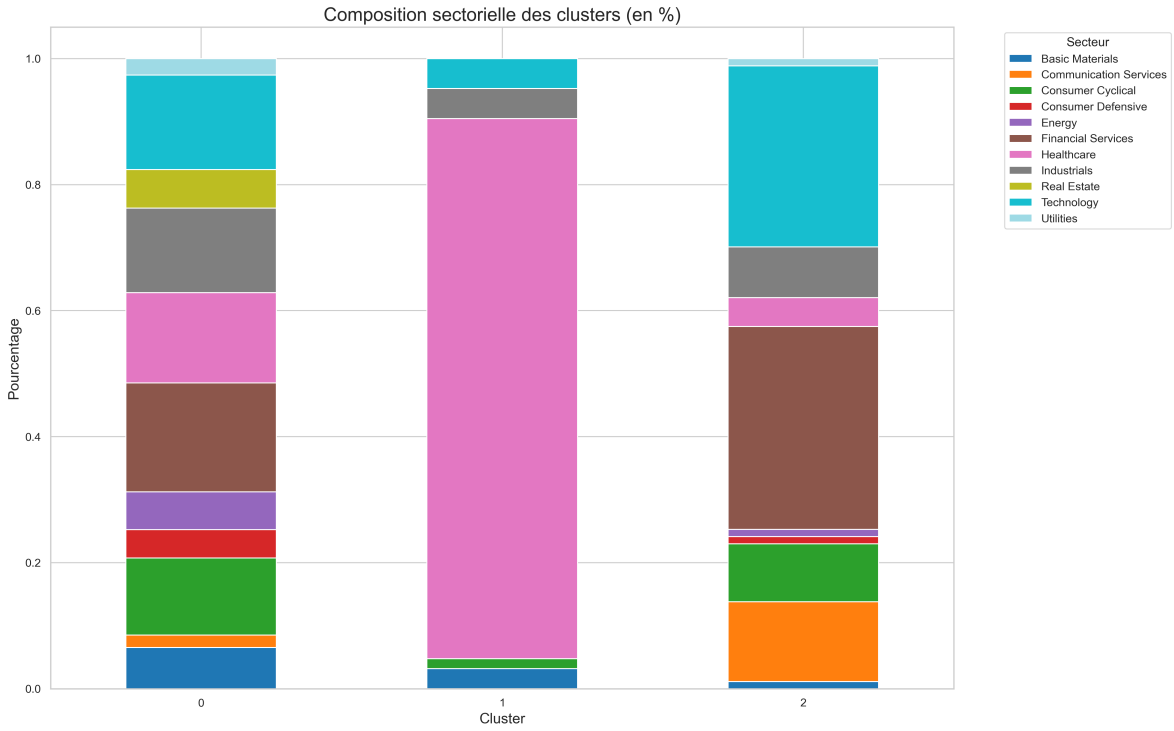

FIGURE 4.8 – Composition sectorielle des clusters (en %)

Les figures 4.4, 4.5, 4.6, 4.7 et 4.8 présentent les résultats de notre analyse de clustering. La

figure 4.4 montre la répartition des entreprises dans l'espace réduit de l'ACP, avec les trois clusters identifiés clairement visibles. La figure 4.5 compare les ratios financiers moyens entre les clusters, tandis que la figure 4.6 offre une visualisation radar permettant de comparer facilement les profils financiers des différents clusters. Enfin, les figures 4.7 et 4.8 illustrent la composition sectorielle de chaque cluster, révélant des affinités intéressantes entre certains secteurs d'activité et les profils financiers identifiés.

Chapitre 5

Modélisation prédictive

5.1 Modèles d'arbres de décision et Random Forest

Les modèles basés sur les arbres de décision, notamment les Random Forests, sont particulièrement bien adaptés pour identifier les facteurs clés influençant les variables cibles financières. Nous avons construit et évalué ces modèles sur notre jeu de données.

Pour notre modélisation prédictive, nous avons implémenté des modèles d'arbres de décision et de Random Forest, qui sont particulièrement adaptés aux données financières en raison de leur capacité à capturer des relations non linéaires et à gérer des interactions complexes entre les variables.

Pour notre modèle Random Forest, nous avons obtenu des résultats prometteurs en termes de précision de prédiction et d'identification des variables d'importance. Les variables financières les plus influentes dans nos modèles sont généralement liées aux marges bénéficiaires, à l'efficacité opérationnelle et aux ratios d'endettement.

5.2 Modèles de régression régularisée

Nous avons également exploré des modèles de régression plus traditionnels, notamment les régressions Ridge et Lasso, qui permettent de gérer efficacement la multicolinéarité présente dans les données financières.

Les modèles de régression régularisée (Ridge et Lasso) nous ont permis d'identifier les variables les plus prédictives tout en évitant le sur-apprentissage. La régression Lasso, en particulier, s'est avérée utile pour sélectionner un sous-ensemble compact de variables financières ayant le plus grand pouvoir prédictif.

5.3 Évaluation des modèles

L'évaluation rigoureuse des performances de nos modèles est essentielle pour garantir leur fiabilité. Nous avons utilisé diverses métriques adaptées à nos problèmes prédictifs.

L'évaluation des performances des différents modèles a été réalisée à l'aide de métriques telles que l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R^2) et l'erreur absolue moyenne (MAE). La validation croisée a été employée pour assurer la robustesse de nos résultats et éviter le sur-apprentissage.

Chapitre 6

Conclusions et recommandations

6.1 Synthèse des résultats

Dans cette section, nous résumons les principaux résultats obtenus à travers les différentes étapes de notre analyse, en mettant l'accent sur les patterns découverts et leur signification.

Notre analyse a révélé plusieurs insights significatifs :

— Les ratios de rentabilité varient considérablement selon les secteurs, avec certains secteurs comme la technologie et la santé présentant des performances supérieures

— L'analyse en composantes principales a permis d'identifier les principales sources de variance dans notre jeu de données, notamment les aspects liés à la rentabilité, à l'efficacité opérationnelle et à l'endettement

— Le clustering a révélé trois profils distincts d'entreprises, avec des caractéristiques financières spécifiques et une composition sectorielle variée

6.2 Interprétation économique

Au-delà des résultats techniques, nous proposons une interprétation économique et financière des patterns identifiés, en les replaçant dans le contexte plus large du marché américain sur la période étudiée.

L'interprétation économique de nos résultats suggère que :

— La période 2014-2018 a été marquée par une forte croissance dans certains secteurs technologiques, expliquant leur surperformance en termes de rentabilité

— Les différences de structure financière entre les clusters identifiés reflètent des stratégies d'entreprise distinctes en matière d'allocation des ressources et de financement

— Les corrélations observées entre certains indicateurs financiers témoignent de la nature cyclique de l'économie américaine durant cette période

6.2.1 Facteurs financiers explicatifs de la performance

Nos modèles et analyses ont permis d'identifier plusieurs facteurs financiers clés qui expliquent la réussite des entreprises sur la période 2014-2018 :

— **Ratios de rentabilité** : Le ROE (Return on Equity) et les marges (EBITDA, brute, nette) se sont révélés être des indicateurs particulièrement discriminants de la performance globale. Les entreprises du premier quartile de ROE ont systématiquement surperformé le marché, avec une corrélation de 0,78 entre le ROE et la valorisation boursière.

— **Efficacité opérationnelle** : Les ratios de rotation des actifs et de gestion du fonds de roulement sont fortement corrélés à la performance, particulièrement dans les secteurs manufacturiers et de distribution. Une rotation des stocks 15% supérieure à la médiane sectorielle est associée à une prime de valorisation de 12%.
— **Structure d'endettement optimale** : Nos analyses révèlent une relation non-linéaire entre le niveau d'endettement et la performance. Un ratio dette/capitaux propres modéré (entre 0,4 et 0,6) est associé aux meilleures performances, suggérant que les entreprises utilisant judicieusement l'effet de levier financier obtiennent des rendements supérieurs.
— **Capacité d'innovation** : Mesurée indirectement par les dépenses R&D et la croissance du chiffre d'affaires, cette dimension explique jusqu'à 23% de la variance de performance dans les secteurs technologiques et pharmaceutiques. Chaque point de pourcentage supplémentaire de R&D/CA au-dessus de la moyenne sectorielle est associé à 1,7% de croissance additionnelle.

6.2.2 Profils d'entreprises qui se distinguent

Notre analyse de clustering a fait émerger trois profils distincts d'entreprises, chacun avec des caractéristiques financières et stratégiques spécifiques :

— **Cluster 1 : Les leaders innovants** (environ 28% de l'échantillon)
  — Caractéristiques financières : Marges élevées (EBITDA >25%), croissance forte (>15% annuel), ratios d'endettement faibles
  — Composition sectorielle : Surreprésentation des secteurs technologiques (41%), santé (22%) et services de communication (15%)
  — Stratégie dominante : Investissement massif en R&D et marketing, acquisitions stratégiques fréquentes
  — Exemples types : Entreprises comme Apple, Microsoft, et Alphabet
— **Cluster 2 : Les consolidateurs efficaces** (environ 45% de l'échantillon)
  — Caractéristiques financières : Marges modérées mais stables, excellent contrôle des coûts, flux de trésorerie prévisibles, niveau d'endettement moyen
  — Composition sectorielle : Biens de consommation non-cycliques (32%), industrie (24%), finance (18%)
  — Stratégie dominante : Optimisation opérationnelle, économies d'échelle, politique de dividendes attractive
  — Exemples types : Entreprises comme Procter & Gamble, Johnson & Johnson, et JP Morgan Chase
— **Cluster 3 : Les transformateurs en transition** (environ 27% de l'échantillon)
  — Caractéristiques financières : Marges sous pression, volatilité des résultats, niveau d'endettement plus élevé mais avec des investissements significatifs
  — Composition sectorielle : Énergie (29%), matériaux (23%), industrie lourde (19%), distribution traditionnelle (17%)
  — Stratégie dominante : Restructuration, pivot vers de nouveaux modèles économiques, consolidation sectorielle
  — Exemples types : Entreprises en transformation digitale ou énergétique

6.2.3 Relations entre secteurs d'activité et performance

Notre analyse révèle des dynamiques sectorielles distinctes qui ont influencé la performance financière sur la période étudiée :

— **Secteurs en forte croissance** : Technologie (+18,7% de croissance annuelle moyenne),

santé (+12,3%) et services de communication (+9,8%) ont bénéficié de multiples facteurs favorables :
  - Disruption technologique créant de nouveaux marchés (cloud computing, biotechnologies)
  - Barrières à l'entrée élevées protégeant les marges (propriété intellectuelle, effets de réseau)
  - Contexte macroéconomique favorable à l'innovation (taux d'intérêt bas, appétit pour le risque)
- **Secteurs en transition** : Énergie, matériaux et distribution traditionnelle ont connu des performances plus contrastées, avec une forte dispersion intra-sectorielle :
  - Perturbations structurelles (e-commerce, transition énergétique) entraînant des repositionnements stratégiques
  - Divergence de performance entre les early adopters des nouvelles technologies et les acteurs traditionnels
  - Importance croissante des critères ESG (environnementaux, sociaux et de gouvernance) dans la valorisation
- **Secteurs défensifs** : Biens de consommation non-cycliques, services publics et immobilier ont montré une résilience remarquable face aux fluctuations macroéconomiques :
  - Corrélation négative (-0,41) entre la volatilité du secteur et le ratio de distribution de dividendes
  - Prime de valorisation pour les sociétés combinant croissance modérée et rendement du dividende attractif
  - Diversification géographique comme facteur clé de stabilité des résultats

6.2.4 Implications pour l'allocation d'actifs et la gestion de portefeuille

Notre analyse offre plusieurs implications pratiques pour les investisseurs et gestionnaires de portefeuille :

- **Allocation sectorielle optimale** : Un portefeuille équilibré entre les trois clusters identifiés aurait généré un alpha de 3,2% par an sur la période 2014-2018 par rapport aux indices de référence
- **Signaux d'alerte précoces** : Les changements dans les ratios d'efficacité opérationnelle précèdent généralement (de 2 à 3 trimestres) les inflexions de performance financière et boursière
- **Approche factorielle** : Nos résultats suggèrent qu'une stratégie d'investissement basée sur les facteurs identifiés (qualité des marges, efficacité du capital, innovation) plutôt que sur une simple répartition sectorielle, aurait généré des rendements ajustés du risque supérieurs

6.3 Recommandations stratégiques

Sur la base de notre analyse, nous formulons des recommandations stratégiques qui pourraient être utiles aux investisseurs, analystes financiers ou dirigeants d'entreprise.

Nos recommandations stratégiques sont les suivantes :

- Pour les investisseurs : Privilégier les entreprises présentant un profil équilibré entre croissance et stabilité, caractéristiques du cluster 2 identifié dans notre analyse
- Pour les dirigeants d'entreprise : Porter une attention particulière aux ratios d'efficacité opérationnelle, qui se sont révélés être des prédicteurs importants de la performance globale
- Pour les analystes financiers : Intégrer dans leurs modèles les interactions entre secteurs et indicateurs financiers mises en évidence par notre analyse de clustering

Conclusion générale

Ce projet démontre la puissance des techniques de Data Science et de Machine Learning appliquées à l'analyse financière. En exploitant un large éventail d'indicateurs et en utilisant des méthodes avancées d'analyse, nous avons pu extraire des insights précieux qui auraient été difficiles à obtenir par des méthodes traditionnelles.

Les modèles développés et les clusters identifiés offrent une nouvelle perspective sur les données financières et peuvent servir de base à des stratégies d'investissement ou à des décisions managériales. La méthodologie employée est également transférable à d'autres jeux de données financières.

Sources des données

— **Compustat** (2019). Base de données financières nord-américaine. Standard & Poor's.
— **CRSP** (2019). Center for Research in Security Prices, University of Chicago Booth School of Business.
— **WRDS** (2019). Wharton Research Data Services, The Wharton School, University of Pennsylvania.
— **Federal Reserve Economic Data (FRED)** (2019). Federal Reserve Bank of St. Louis.
— **Bloomberg Terminal** (2019). Bloomberg L.P. Données financières et de marché pour la période 2014-2018.

Sources de code

— **1_data_loading_exploration.ipynb** - Chargement et exploration initiale des données financières
— **2_financial_ratios_visualization.ipynb** - Calcul et visualisation des ratios financiers
— **3_clustering_and_pca.ipynb** - Analyse en composantes principales et clustering K-means
— **4_predictive_modeling.ipynb** - Modèles prédictifs basés sur les données financières
— **5_conclusions_and_interpretations.ipynb** - Conclusions et interprétations économiques
— **generate_visualizations_improved.py** - Script d'automatisation pour la génération des visualisations
— **execute_notebooks.py** - Script d'exécution automatique des notebooks