Projet: Analyse et Prédiction de Marchés Financiers

Joris Rabilloud - Escen Tech N5

Github

https://github.com/Joriiss/projet-traitement-donnees

▼ 🔎 1. Exploration et compréhension

1. Importation et inspection

cac40.csv

En ouvrant le fichier cac40.csv dans Sheets, je vois qu'il comporte 7 colonnes. Chaque ligne correspond à une date entre le 02/01/2023 et le 01/01/2025. Pour chaque ligne on a le prix d'ouverture du CAC 40, le prix de fermeture, le max et le min sur la journée ainsi que le volume d'échanges sur la journée.

Les données ne sont pas triés chronologiquement par défaut.

Je remarque que certaines valeurs sont manquantes. Par example pour le 08/02/2023, il manque le volume.

Date	Index	Open	High	Low	Close	Volume
2023-02-08	CAC40	995.42	1001.84	992.33	994.87	

Pour le 20/12/2023, le volume à une valeur négative, ce qui ne devrait pas être possible

Date	Index	Open	High	Low	Close	Volume
2023-12-20	CAC40	998.03	999.37	997.65	996.87	**-299298.0**

Il y a aussi des dates manquantes, par exemple, on passe du 2023-01-20 au 2023-01-23.

Certaines valeurs ont un symbole "€" en plus du nombre.

Il faudra donc nettoyer les données pour supprimer les lignes manquant des valeurs ou avec des valeurs impossibles.

Aperçu des 5 premières lignes - cac40.csv

Date	Index	Open	High	Low	Close	Volume
2023-01-02	CAC40	1001.64	1004.9	998.99	1001.74	440686.0
2023-01-03	CAC40	998.83	1006.71	994.86	998.08	416463.0
2023-01-04	CAC40	997.4	1004.78	991.12	1003.08	345144.0
2023-01-05	CAC40	999.27	1006.17	997.81	1002.38	978499.0
2023-01-06	CAC40	1002.39	1005.52	995.61	1003.64	776999.0

sto.parquet

Ce fichier contient les données pour l'action Euro Stoxx 50. Il contient 700 lignes et 6 colonnes (Date, Close, High, Low, Open et Volume).

La première ligne contient uniquement la mention “^STOXX50E” donc elle devra être supprimée.

Aperçu des 5 premières lignes - Euro Stoxx 50 (^STOXX50E)

Date	Close	High	Low	Open	Volume
	^STOXX50E	^STOXX50E	^STOXX50E	^STOXX50E	^STOXX50E
19360.0	3882.2900390625	3921.3701171875	3852.070068359375	3852.070068359375	33158800
19361	3973.969970703125	3975.5	3891.0400390625	3891.0400390625	39907300
19362	3959.47998046875	3974.610107421875	3950.419921875	3967.010009765625	26926500
19363	4017.830078125	4017.85009765625	3951.929931640625	3959.9599609375	28231800

2. Nettoyage et mise en forme

1. Cohérence temporelle

Les valeurs couvrent une période de 731 jours dont 523 jours de trading (les marchés sont fermés les week-ends).

Aucune date ne manque au dataset. En revanche, 10 dates sont en double (ex : le 2023-04-18).

En dehors des doublons, les dates sont cohérentes, ils ne manquent pas de jour dans la période donnée.

1. **Valeurs vides**
   - Date : 0
   - Index : 0
   - Open : 6
   - High : 7
   - Low : 3
   - Close : 4
   - Volume : 10

   30 valeurs manquantes en tout → Les lignes contenant une valeur vide seront supprimés car nous n’avons pas le moyen d’estimer la valeur manquante.
3. **Doublons**
   - 10 lignes sont totalement identiques
   - 0 lignes avec la même date des valeurs différentes

   10 doublons en tout → les lignes seront supprimés
4. **Valeurs négative**
   - Date : 0
   - Index : 0
   - Open : 0
   - High : 0
   - Low : 0
   - Close : 1
   - Volume : 5

   6 valeurs négatives en tout → les lignes seront supprimés car impossible de connaître la valeur réel.
5. **Optimisations**
   - Conversion du champ “Date” en Datetime
   - Conversion des champs numériques en float32

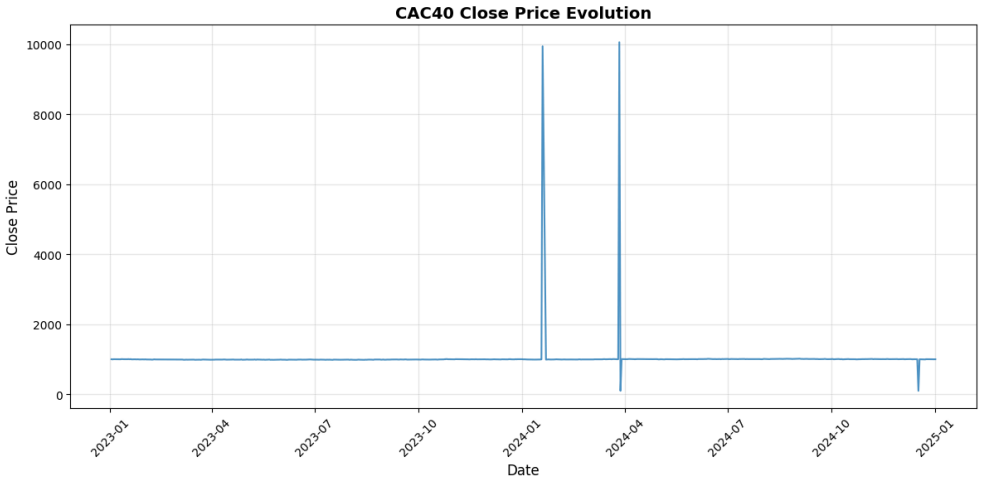
- Suppression du champ “Index” car la valeur est toujours “CAC40”
- Tri par date croissante
- Suppression des doublons, valeurs vides et négatives

Avant optimisations, le dataset était de 0.09 MB pour 533 lignes et 7 colonnes.

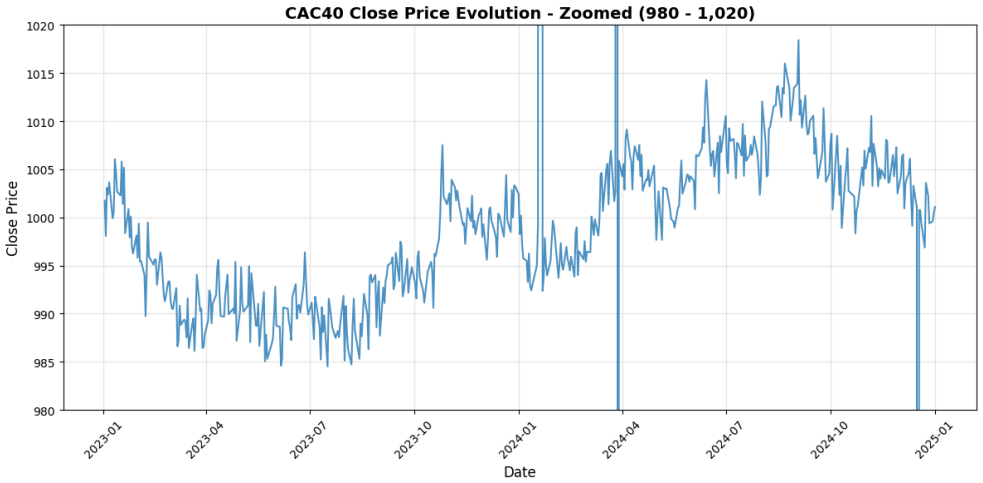
Après optimisations, le dataset est de 0.01 MB (-84.9%) avec 488 lignes et 6 colonnes. 45 lignes et 1 colonne ont été supprimés.

3. Analyse visuelle de base via graphiques (Matplotlib)

Évolution du prix de close du CAC40 sur la période

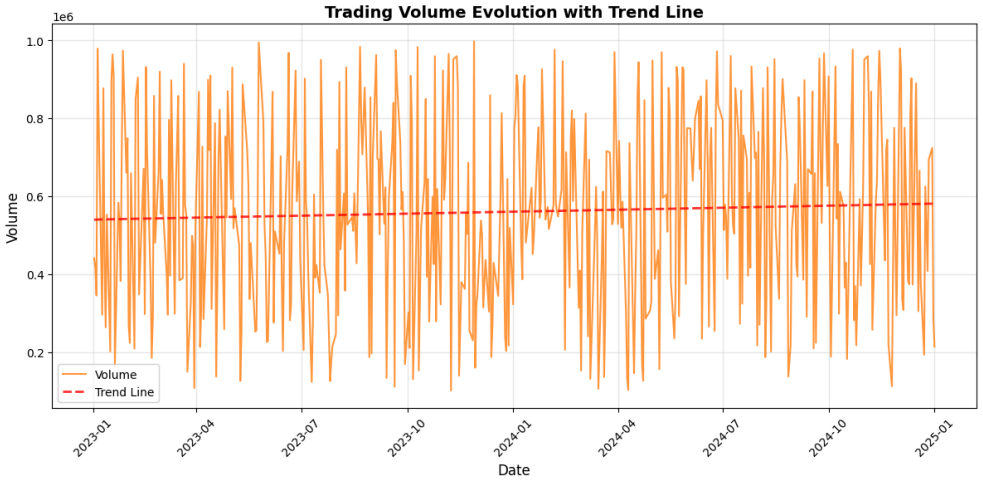


Remarque : On remarque 4 valeurs qui semblent aberrantes



Avec un zoom entre 980 et 1020, on peut mieux voir la réelle tendance.

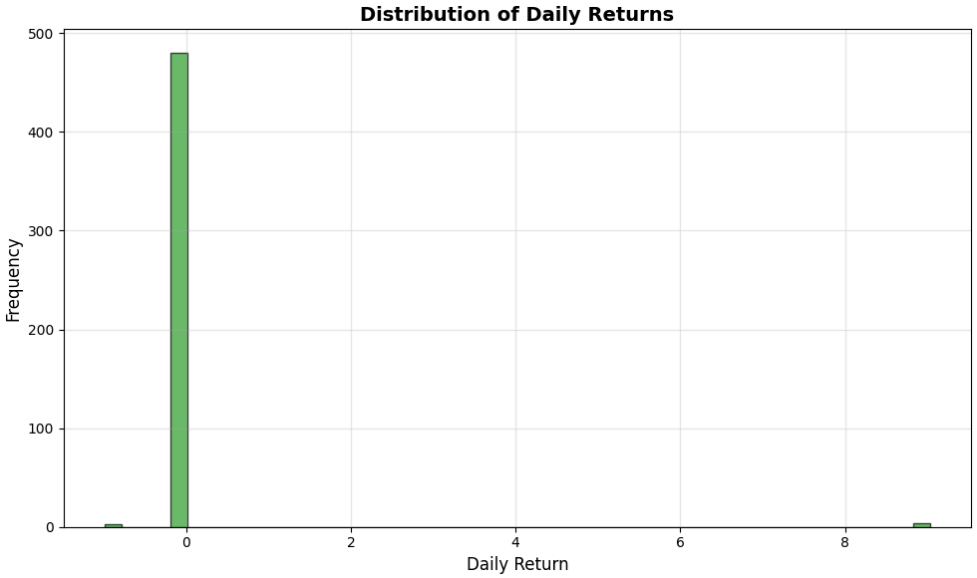
Évolution du volume d’échanges



Histogramme de distribution des rendements

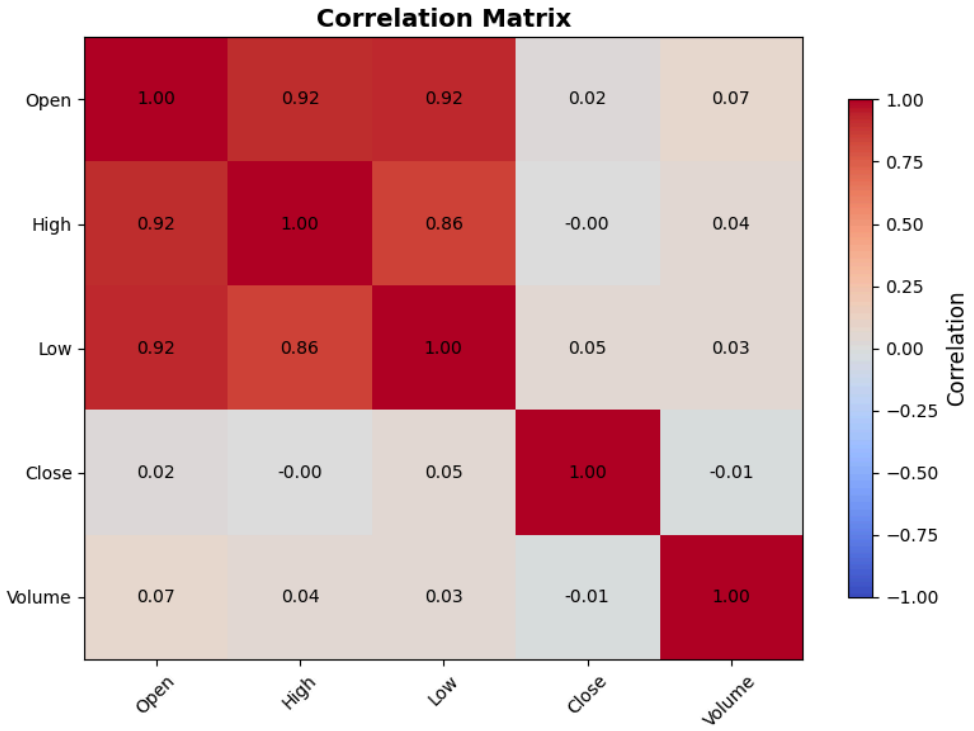
Le rendement est calculé avec le pourcentage de changement du prix de close d'un jour à l'autre

```
daily_returns = cac40_df['Close'].pct_change().dropna()
```



On remarque là aussi des valeurs aberrantes, une inférieure au reste et une bien supérieure.

Matrice de corrélation



La matrice révèle une forte corrélation entre Open/High (0.92), Open/Low (0.92) et High/Low (0.86).

Ces corrélations sont plutôt logiques.

Le volume d'échanges n'est pas corrélé avec les prix de la journée, ce qui semble logique. Le volume d'échanges ne devrait pas être influencés par le prix en soit mais plutôt par les variations des prix au sein de la journée et des évènements attendus.

Par contre, le prix de Close n'est pas corrélé avec Open, High ou Low, ce qui est étrange. Cela s'explique sans doute par le fait que Close contient des valeurs extrêmes (les 4 valeurs sur le graphique d'évolution du prix de close du CAC40 sur la période).

La prochaine étape sera de détecter et supprimer ces valeurs aberrantes.

▼ 🕵️ 2. Détection d'anomalies

Choix des méthodes

1. **Z-Score**

J'ai décidé de commencer avec la méthode Z-Score sur les rendements journaliers pour détecter des variations de prix inhabituelles. J'utilise un seuil classique de $|z\text{-score}| > 3$.

2. **Écart inter quartile**

Ensuite, j'utilise IQR (écart inter quartile) pour détecter des valeurs extrêmes sur Open, Close, High, Low et Volume. Le seuil est de 1.5.

3. Écart avec la moyenne mobile

Calcul de l'écart avec une moyenne mobile sur 1 mois de trading pour détecter des pics de valeurs. Le seuil sera de 2 écarts-types avec la moyenne.

4. **Comparaison avec EURO STOXX 50**

Finalement, on compare avec l'euro stoxx 50 pour voir si des anomalies similaires existent et donc peuvent être expliqué par des événements extérieurs.

L'utilisation de plusieurs méthodes permet plus de robustesse pour réduires les faux positifs.

Implémentation

1. **Z-score**

   La méthode Z-Score avec |Z| > 3 permet de détecter :

   - 4 anomalies dans les rendements journaliers :

Date	Close	Daily_Return	ZScore (>3?)
2024-1-19	9942.50	8.948768	10.875587
2024-3-27	100057.50	9.030417	10.975579
2024-3-29	1005.90	8.995627	10.932974
2024-12-18	1000.77	9.008902	10.949230

   - 2 anomalies sur Close

Date	Close	ZScore (>3?)
2024-01-19	9942.5	15.394662
2024-03-27	10057.5	15.593349

   - Aucune anomalie sur le volume, low, high et open

2. **Écart interquartile**

   En utilisant Q1 - 1.5 x écart interquartile comme seuil bas et Q3 + 1.5 x écart interquartile en seuil haut, on obtient :

   - 9 anomalies sur les rendements journaliers

     Seuil bas (Q1 - 1,5×IQR) : -0,0082
     Seuil haut : (Q3 + 1,5×IQR) : 0,0082
     IQR: 0.0165

Date	Close	Daily_Return	Seuil dépassé
2023-02-09	999.469971	0.009821	> 0,0082
20204-28	987.169983	-0.008238	< -0,0082
2024-01-19	9942.500000	8.948768	> 0,0082
2024-01-22	992.369995	-0.900189	< -0,0082
...			

     Les 4 anomalies détectées avec le ZScore apparaissent dans cette liste

   - 4 anomalies sur le prix de fermeture (Close)

Date	Close
2024-01-19	9942.5
2024-03-27	10057.5
2024-03-28	100.635002
2024-12-17	99.987999

     Les 2 anomalies détectées avec le ZScore apparaissent dans cette liste

   - 0 anomalies pour le volume, le prix max, min et d'ouverture

3. **Écart avec la moyenne mobile**

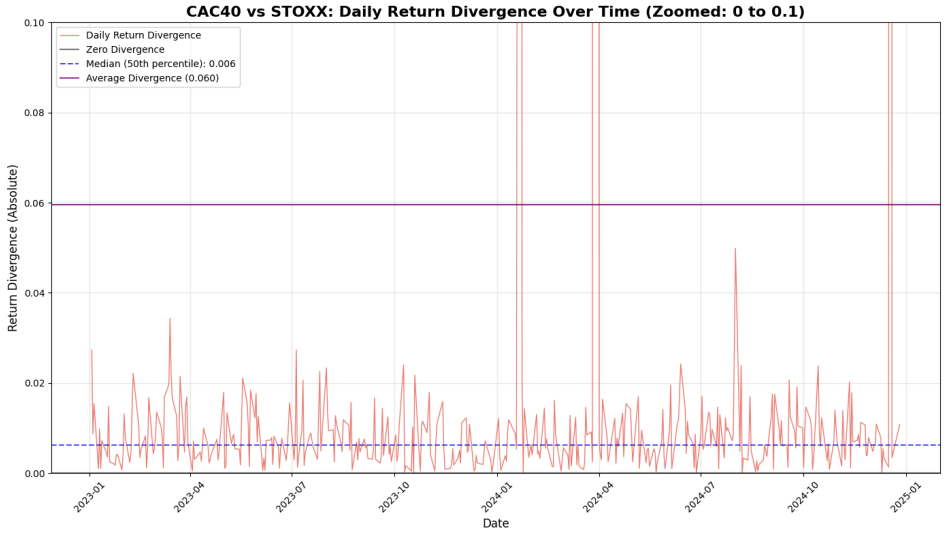
   En calculant l'écart avec la moyenne sur 20 jours (correspondant à un mois de trade) avec un seuil de 3σ, on obtient les anomalies suivantes :

   - 1 anomalie pour Low
   - 3 pour Close

- 4 pour le rendement journalier
- 0 pour Open, High et Volume

4. Comparaison avec Euro Stoxx 50

En comparant les valeurs du CAC40 et de Euro Stoxx 50, on obtient une divergence moyenne de **0.0596.** Cela signifie que la différence de retours moyens entre les deux stocks est de 5.96%.
En revanche en regardant le graphique, on peut voir que la moyenne est élevée dû à des valeurs aberrantes dans les valeurs du CSV. La divergence médiane est autour de 0.006 soit une différence de 0.6% dans les retours des 2 stocks. → C'est une divergence basse, les marchés varient en concordance.



Avec cette méthode on peut détecter des anomalies dans les valeurs en regardant les valeurs supérieures aux 99è centile. Cela nous renvoie 4 valeurs anormales :

Date	Close
2024-01-22	992.4
2024-03-27	10057.5
2024-12-17	99.9
2024-12-18	1000.77

Conclusion

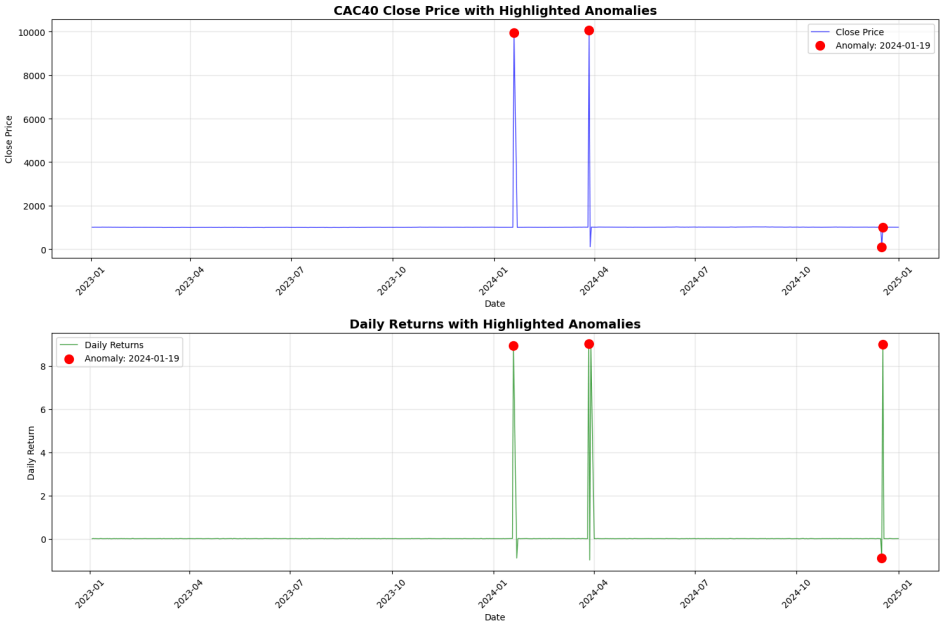
- Synthèses des anomalies détectées

En total 9 lignes ont été détectées comme "anomalie" par au moins une des méthodes :

Date	Close	Méthode de détection
2023-02-09	999.47	IQR
2023-04-28	987.17	IQR
2024-01-19	9942.5	Z-Score, IQR, Rolling
2024-01-22	992.37	IQR, High-Divergence
2024-03-27	10057.5	Z-Score, IQR, Rolling, High-Divergence
2024-03-28	100.6	IQR
2024-03-29	1005.91	Z-Score, IQR
2024-12-17	99.99	IQR, Rolling, High-Divergence
2024-12-18	1000.77	Z-Score, IQR, Rolling, High-Divergence

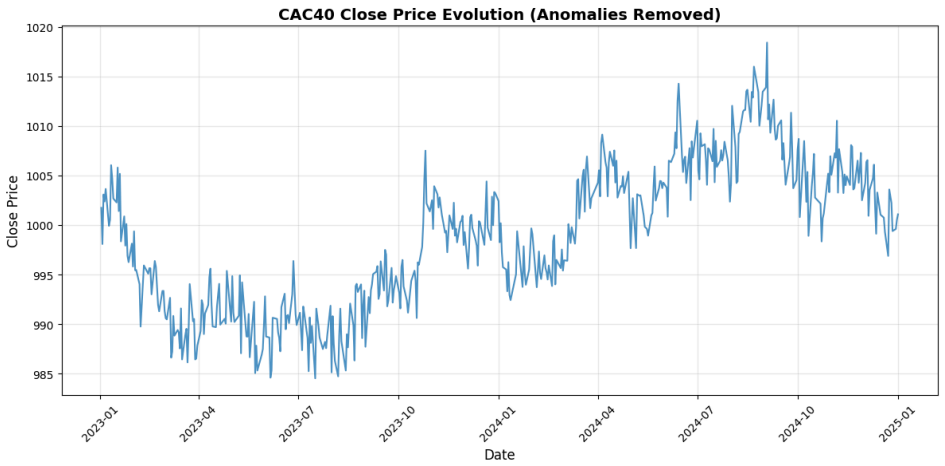
J’ai décidé de retenir les anomalies détectés par au moins 3. Si une valeur semble anormale mais que la divergence avec Euro Stoxx est faible, dans ce cas on peut expliquer ce changement par un événement qui a affecté les marchés et affectés le cours de ces actions.

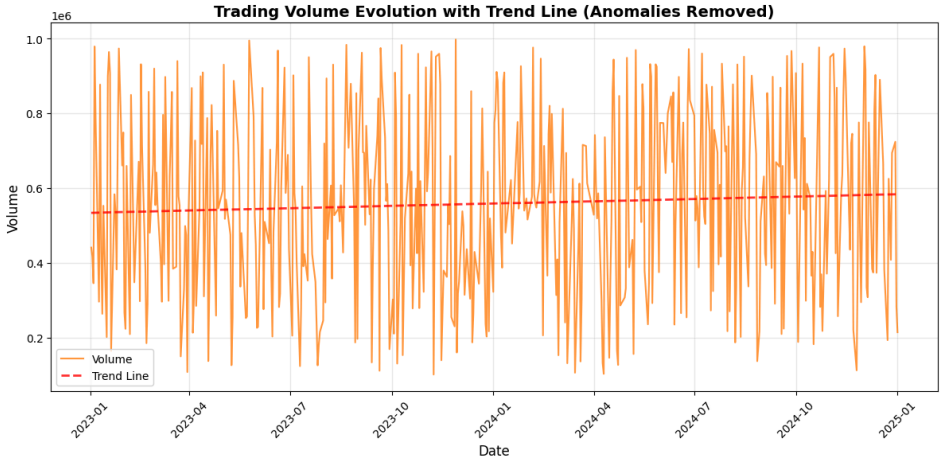
On supprime donc les valeurs pour : **2024-01-19, 2024-03-27, 2024-12-17 et 2024-12-18.**

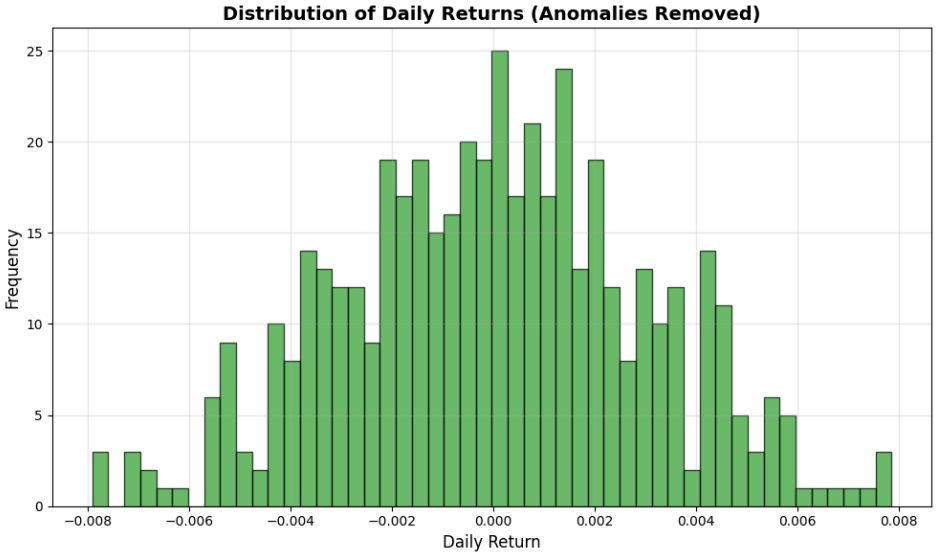


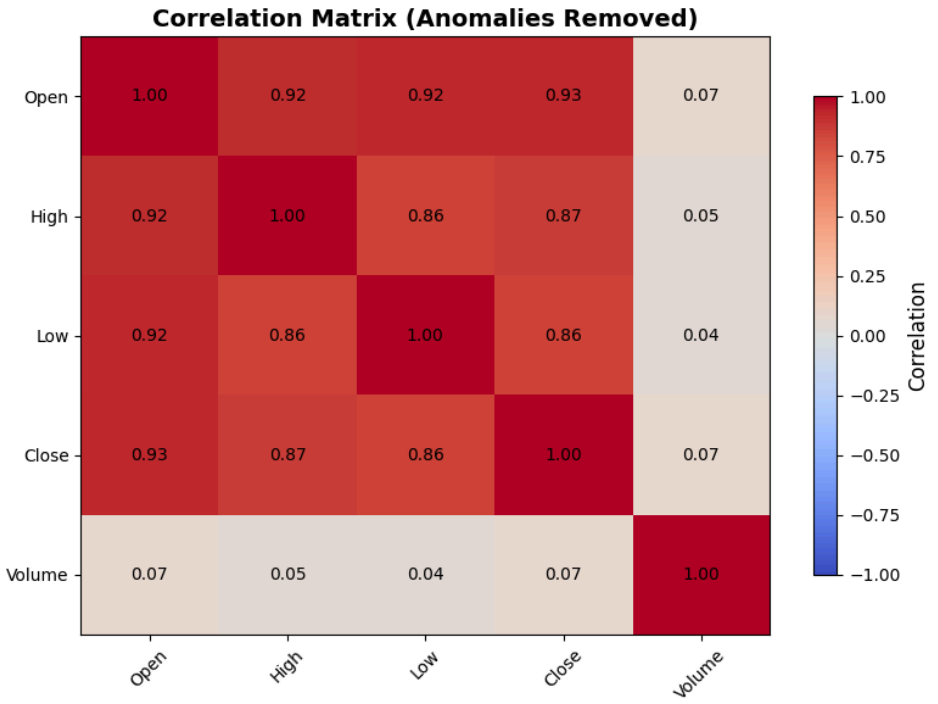
→ On voit une valeur très faible le 28/03 qui semble ne pas avoir été détecté par ma méthode.

Analyse visuelle de base via graphiques avec les 9 anomalies supprimées




Trading Volume Evolution with Trend Line (Anomalies Removed)
1e6
1.0
0.8
0.6
0.4
0.2
Volume
Volume
Trend Line
2023-01
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
Date


Distribution of Daily Returns (Anomalies Removed)
25
20
15
10
5
0
Frequency
−0.008
−0.006
−0.004
−0.002
0.000
0.002
0.004
0.006
0.008
Daily Return

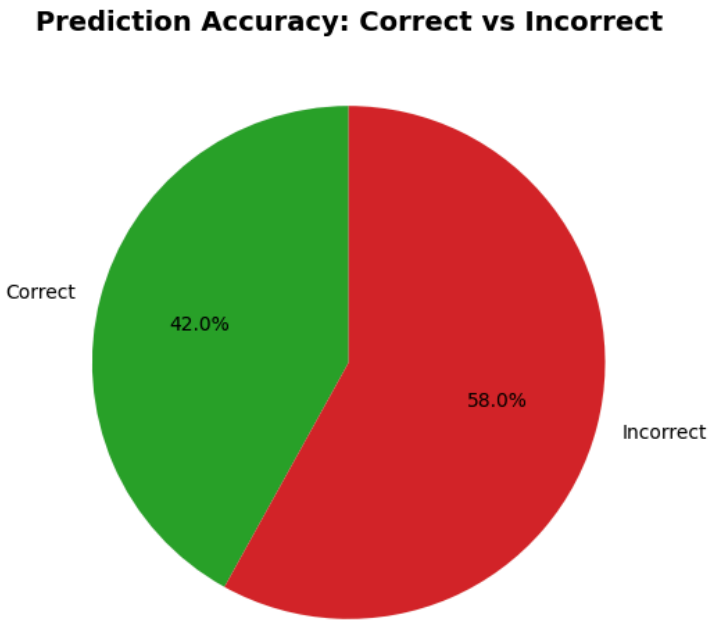

Correlation Matrix (Anomalies Removed)
Open
High
Low
Close
Volume
1.00 0.92 0.92 0.93 0.07
0.92 1.00 0.86 0.87 0.05
0.92 0.86 1.00 0.86 0.04
0.93 0.87 0.86 1.00 0.07
0.07 0.05 0.04 0.07 1.00
Open
High
Low
Close
Volume
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation

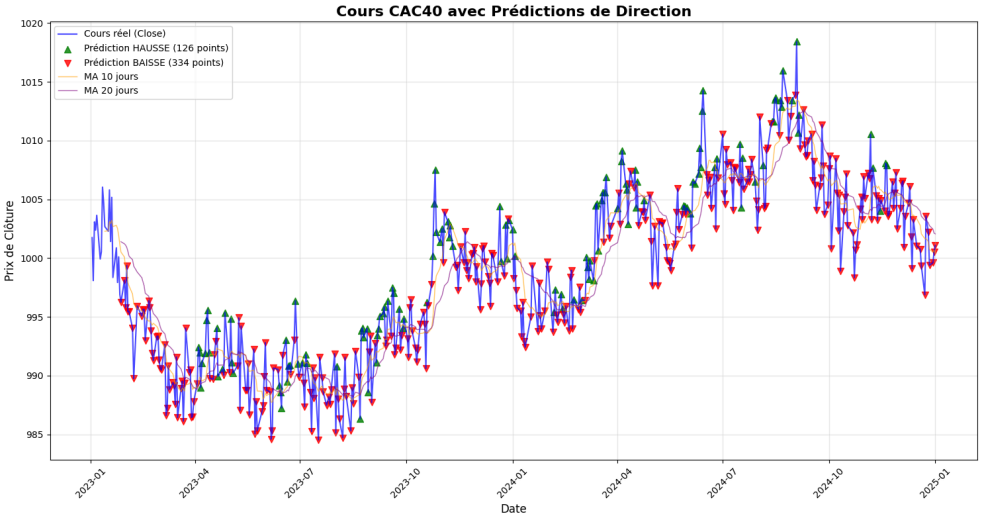
Une fois les anomalies supprimées les graphiques réalisés à l'étape 1 semble plus cohérent. On peut maintenant déceler une corrélation entre Close et Open / High / Low.

▼ 📈 3. Phase d'analyse statistique

Le modèle nous permet de faire 460 prédictions. Il y a 479 dates dans le dataset mais il faut calculer la moyenne mobile sur 20 jours pour pouvoir appliquer notre règle de prédiction.

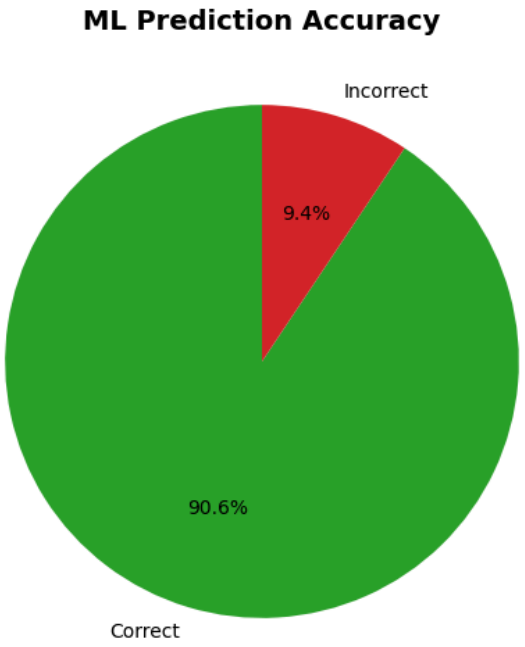
Sur les 460 prédictions (126 de hausse, 334 de baisse), 193 sont corrects. Soit une précision de 41,96%.

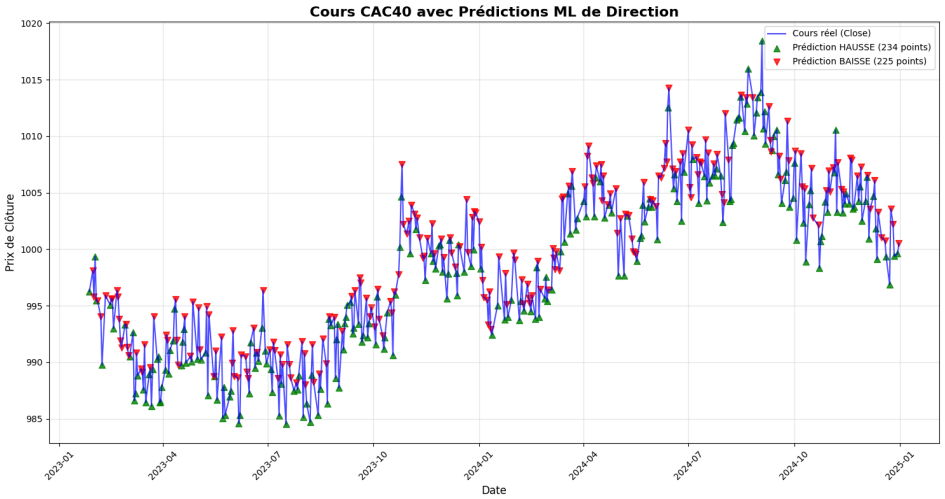




▼ 🧠 4. Prédiction avec du Machine Learning

Le modèle de prédiction avec ML génère 459 prédictions (234 de hausse et 225 de baisse) dont 416 corrects soit 90,63% de précison.





▼ ✍️ 5. Conclusion

Principaux Résultats

Qualité des données

Le nettoyage a permis d'éliminer 45 lignes problématiques (30 valeurs manquantes, 10 doublons, 6 valeurs négatives) sur 533 lignes initiales, résultant en un dataset propre de 488 observations.

Détection d'anomalies

L'approche à permis d'identifier des valeurs aberrantes cohérentes entre les différentes techniques. La comparaison avec l'Euro Stoxx 50 a confirmé que certaines anomalies n'étaient pas dues à des événements de marché mais à des erreurs de données (divergence médiane de 0,6% entre les deux indices).

Performance des modèles

Les résultats obtenus montrent une différence significative entre les deux approches :

- **Modèle statistique simple** (moyennes mobiles + momentum) : 41,96% de précision sur 460 prédictions (126 hausses, 334 baisses, 193 correctes)

- **Modèle Machine Learning** (RandomForest) : 90,63% de précision sur 459 prédictions (234 hausses, 225 baisses, 416 correctes)

Le modèle de Machine Learning à permis de faire de meilleures prédiction, avec une distribution plus équilibrée entre prédictions de hausse et de baisse, suggérant une meilleure capture des patterns complexes du marché.

Interprétation et Limites

Forces du projet

- **Rigueur méthodologique** : L'utilisation de plusieurs méthodes de détection d'anomalies a permis de limiter les faux positifs et d'assurer la qualité des données.
- **Performance du modèle ML** : Une précision de 90,63% est un résultat encourageant pour la prédiction de mouvements de marché.
- **Optimisation technique** : La réduction de 84,9% de la taille des données facilite le traitement et améliore les performances.

Limites identifiées

- **Échantillon temporel limité** : Deux ans de données historiques peuvent ne pas capturer tous les cycles de marché et événements exceptionnels (crises, bulles).
- **Simplicité du modèle statistique** : La règle basée sur les moyennes mobiles (41,96% de précision) s'avère insuffisante, confirmant la nécessité d'approches plus sophistiquées.
- **Prédiction binaire** : La classification hausse/baisse ne fournit pas d'information sur l'amplitude des mouvements, limitant son utilité pour des stratégies de trading réelles.

Perspectives d'amélioration

Pour aller plus loin, il serait pertinent de :

- Intégrer des données sur une période plus longue (10-15 ans) incluant différents cycles économiques
- Ajouter des variables externes (taux d'intérêt, inflation, indicateurs économiques, sentiment de marché)