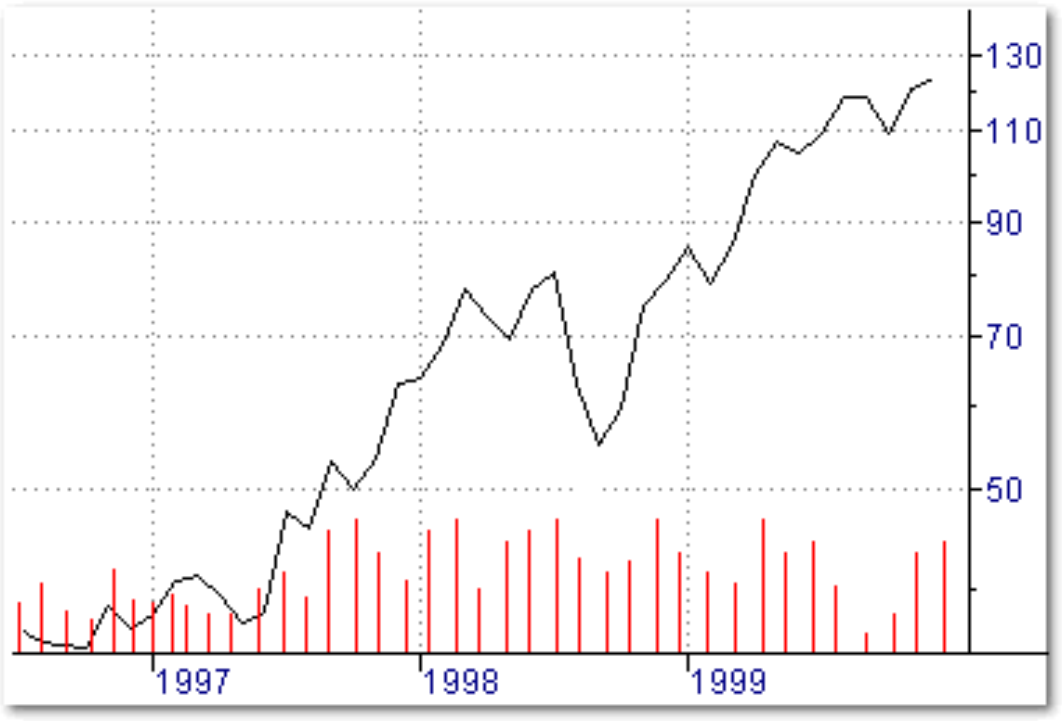
PROJET DE STATISTIQUES

Comprendre un monde financier complexe par l'intermédiaire des statistiques



Un projet mené par :
TARLET Quentin
BARDIN-ENDERLIN Malo
BAUMBERGER Max
SERENE Théo

Motivation & Présentation du sujet **3**
I) Étude univariée des variables utilisées **4**
Variables quantitatives 4
Variables qualitatives 7
II) Problématiques **8**
Problématique n°1 : Comment peut-on prédire l'évolution de la valeur de l'indice du S&P500 en fonction de plusieurs variables explicatives ? 8
Problématique n°2 : Comment évolue la volatilité et les rendements du S&P500 en période de crise 13
Problématique n°3 : Comment le secteur d'activité d'une entreprise, sa performance financière et son nombre d'employés sont-ils interconnectés ? 17
III) Conclusion **20**
IV) Vocabulaire & Annexe **21**

Motivation & Présentation du sujet

Nous avons choisi ce projet car nous sommes tous passionnés de finance des marchés et essayer de comprendre comment fonctionnent les rouages du système financier mondial est quelque chose qui nous passionne. Ainsi, par l'intermédiaire de ce projet, nous souhaitons améliorer nos connaissances en la matière tout en exploitant la puissance des statistiques afin d'utiliser les données du passé pour prédire le futur.

Afin de nous aider dans cette quête, nous allons récupérer l'ensemble de nos données du site Yahoo Finance avec la bibliothèque yfinance dont vous pouvez trouver la documentation juste ici : https://pypi.org/project/yfinance/.

Nos données commencent dès le 1er janvier 2023 jusqu'à maintenant, ce qui veut dire que vous pouvez compiler le code en 2025, les données les plus récentes datent de la veille. Nous avons choisi d'éviter de prendre des données avant 2023 car la période Covid a fortement impacté les marchés financiers, rendant nos modèles de prévisions bien moins efficaces (pour le moment, prédire une crise est quasiment impossible). Cependant, il était important pour nous d'avoir un code fonctionnel même avec des données récentes afin de pouvoir travailler avec un jeu de données en perpétuelle évolution. Nous mettons un point d'honneur à avoir des données précises et à jour, comme ce que nous fournit yfinance.
Seule la problématique n°2 utilise des données antérieures, car elle porte justement sur les périodes de crise.

Pour ce projet, nous allons essayer de répondre à 3 grandes problématiques qui constituent aujourd'hui le paysage des marchés financiers. Nous chercherons tout d'abord à prévoir le prix futur d'un actif. En effet, comme tout bon investisseur, nous sommes à la recherche du profit et un modèle permettant de prévoir cette évolution nous permettra d'engranger des gains plus qu'intéressants.

Nous commencerons par définir l'ensemble des variables utilisées (quantitatives et qualitatives). Nous continuerons ensuite par développer nos trois problématiques pour enfin arriver à une conclusion générale du sujet. Vous trouverez à la fin de ce document une annexe (pour le code) mais également une page vocabulaire pour certains termes techniques.

I) Étude univariée des variables utilisées

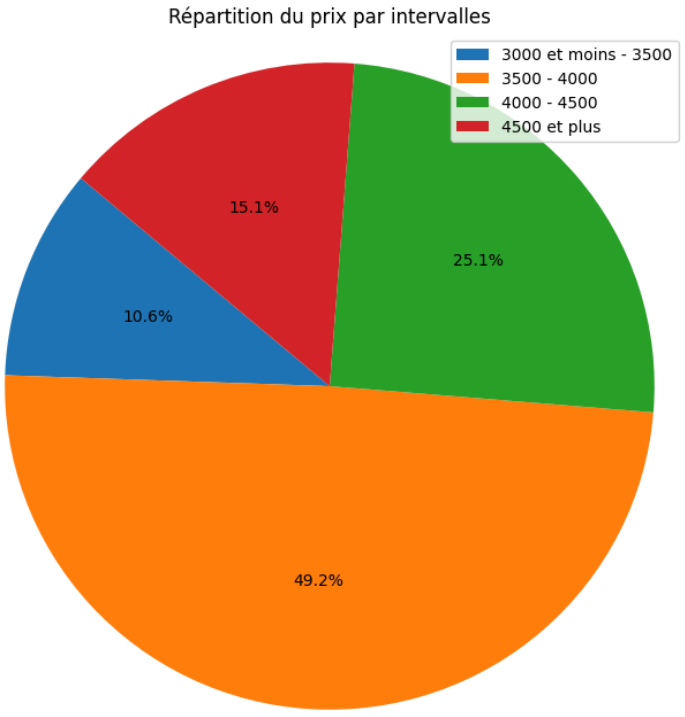
Variables quantitatives

Indice :
Dans le contexte de notre projet, le prix est une donnée cruciale qui reflète la valeur monétaire d'un actif ou d'une monnaie. Cette dernière, accompagnée du temps, permet de tracer l'évolution d'un actif en fonction des jours, des semaines et même des années.
Il faut savoir que la notion de prix est variable, en effet une action Tesla ne vaut pas la même chose en euro, en dollars ou en pesos argentin, il est donc important, quand on indique le prix, d'y ajouter sa monnaie.

Le prix d'un actif est facilement récupérable par l'intermédiaire de notre bibliothèque.
Nous prenons ici la valeur de l'indice du S&P 500, l'indice de référence pour la finance mondiale.

```
La moyenne du prix est : 4464.653596169279
La médiane du prix est : 4405.7099609375
Minimum :  3808.10009765625  Maximum :  5254.35009765625
La variance des prix est de : 147313.05644058232  et l'écart-type :  383.8138304446341
```

Nous pouvons également étudier la répartition des différentes valeurs en 4 grandes intervalles : entre 3500$ et 4000$, 4000$ et 4500$, 4500$ et 5000$ enfin 5000$ et plus.



Nous pouvons remarquer que, la majeure partie du temps, l'indice a évolué entre 3500 et 4000$. Ce modèle n'est qu'une représentation univariée qui n'est pas très précise, l'évolution du prix étant généralement lié à une date.

Volume :

Le volume est une métrique très importante des marchés financiers. Elle permet entre autres d'indiquer la liquidité d'un actif (le nombre de d'acheteurs et de vendeurs). Plus un actif est liquide, moins il est volatile et plus vite s'effectuent les transactions.

```
La moyenne du prix est : 4011340664.652568
La médiane du prix est : 4405.7099609375
Minimum :  1639500000  Maximum :  9354280000
La variance du volume est de : 5.539258462795569e+17  et l'écart-type :  744261947.3542612
```

Comme pour le prix, le volume a besoin d'une deuxième variable pour prendre tout son sens. Il faut donc le coupler généralement à un prix et/ou une date afin de pouvoir l'exploiter.

Date :

Je ne vais pas expliquer ce qu'est une date. Cependant, dans notre modèle (et dans les marchés financiers en général), toute variable est étudiée en fonction du temps (par seconde, minutes, jours, mois …). Dans le cadre de notre projet, afin d'éviter une surcharge des données pouvant ralentir l'exécution du code, nous allons utiliser un format journalier. Il est cependant intéressant de noter que le format de la date n'a pas de réelle incidence sur les moyennes ou médianes de nos variables puisque nous récoltons nos données sur 15 mois (plus de 300 valeurs).

Il est important de noter que les marchés financiers ne sont pas ouverts 24h/24 ni 7j/7. En effet, le NASDAQ n'est ouvert que de 15h30 à 22h (heure française) et les jours de la semaine. Il est important de noter que la bourse est également fermée les jours fériés et le week-end. Cependant, les ordres boursiers peuvent être mis en attente les jours de fermeture rendant l'ouverture généralement mouvementée.

VIX :

VIX est le diminutif de 'Chicago Board Options Exchange Volatility Index'. Il s'agit d'un instrument de mesure utilisé pour suivre la volatilité de l'indice S&P 500. C'est l'indice de volatilité le plus connu sur le marché.

2008 :

```
En 2008, la moyenne du VIX était :  32.663611086588055
En 2008, la médiane du VIX était :  25.0600004196167
Minimum :  16.299999237060547  Maximum :  80.86000061035156
La variance du VIX était de : 269.12427700030935  et l'écart-type :  16.405007680592817
```

2015 :

```
En 2015, la moyenne du VIX était :  16.668007998827445
En 2015, la médiane du VIX était :  15.289999961853027
Minimum :  11.949999809265137  Maximum :  40.7400016784668
La variance du VIX était de : 18.858530005780917  et l'écart-type :  4.3426409022369
```

2020 :

```
En 2020, la moyenne du VIX était :  29.277103212144638
En 2020, la médiane du VIX était :  26.770000457763672
Minimum :  12.100000381469727  Maximum :  82.69000244140625
La variance du VIX était de : 152.61164400541256  et l'écart-type :  12.353608541855799
```

2023 :

```
En 2023, la moyenne du VIX était :  16.870040000915527
En 2023, la médiane du VIX était :  16.9350004196167
Minimum :  12.069999694824219  Maximum :  26.520000457763672
La variance du VIX était de : 9.854435087615354  et l'écart-type :  3.1391774539862114
```

Rendement :

Le rendement d'un actif financier, tel que le S&P500, mesure le gain ou la perte réalisé(e) sur cet actif sur une période donnée. On utilise ici la formule du rendement logarithmique pour déterminer le rendement quotidien.

$$r = ln\left(\frac{V_t}{V_{t-1}}\right)$$

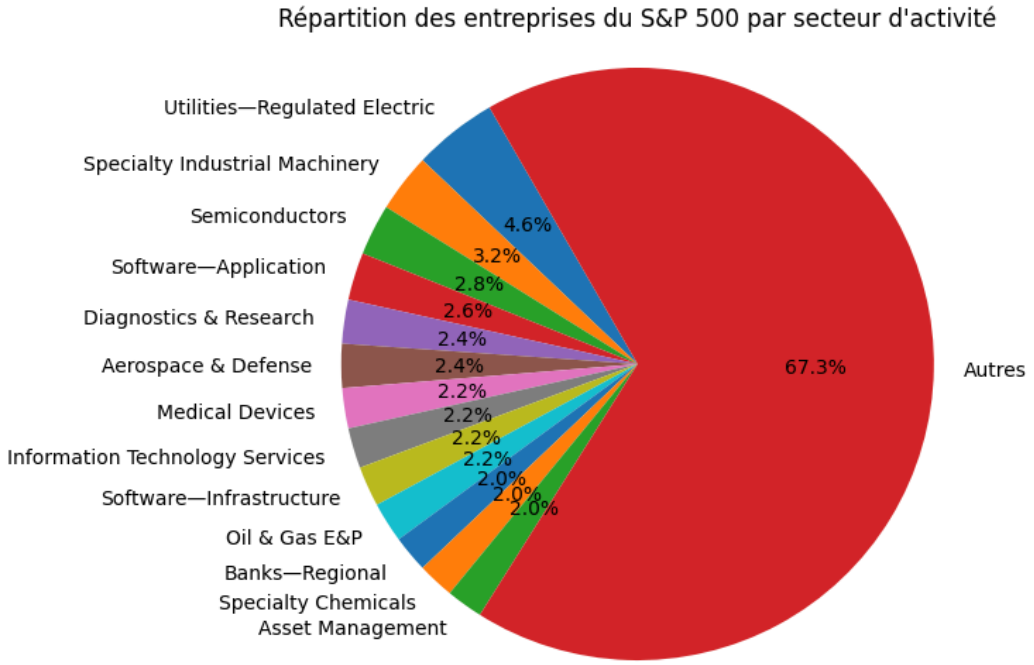
```
Depuis 2023, la moyenne du VIX était :  0.0008793731712155409
Depuis 2023, la médiane du VIX était :  0.0007506937530339595
Minimum :  -0.020244783973471698  Maximum :  0.02258383558903089
La variance du VIX était de : 6.406849320917879e-05  et l'écart-type :  0.008004279680844417
```

On retrouve logiquement une moyenne et une médiane proche de 0.

Variables qualitatives

Secteur d'activité :

Dans notre data frame, à l'aide de yfinance, nous avons à notre disposition les différents secteurs d'activité présents sur le nasdaq. Il existe plus de 45 secteurs d'activité différents mais nous avons choisi de nous intéresser aux 14 majeurs, ceux qui sont mis en évidence sur le graphique.



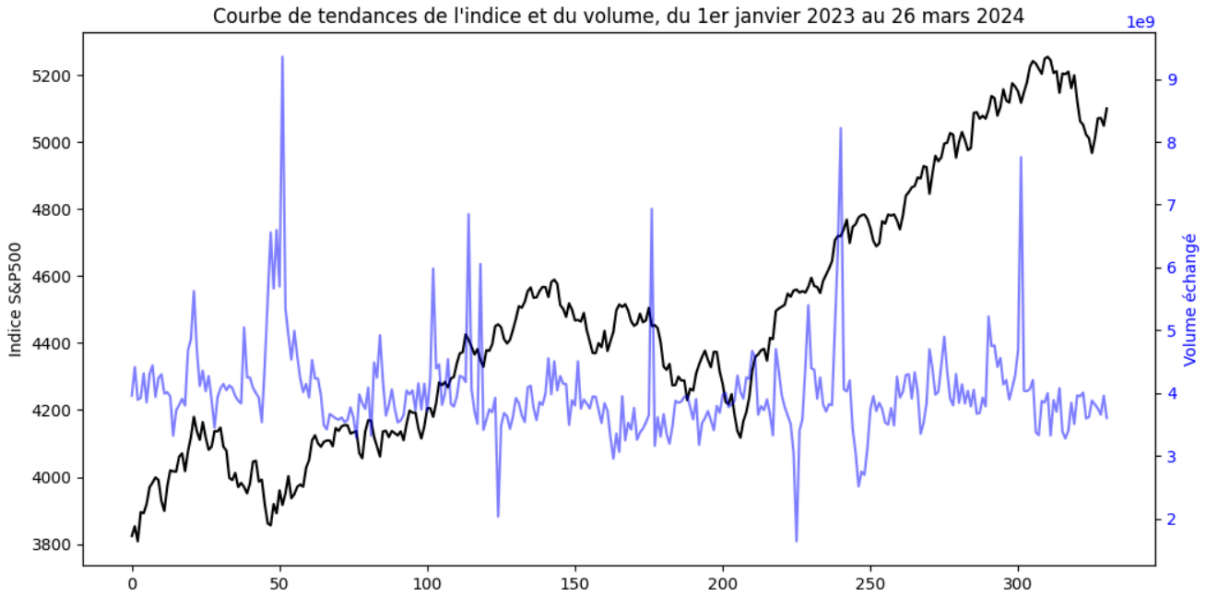
Nous avons choisi majoritairement des variables quantitatives car ce sont elles qui constituent la quasi-totalité des données disponibles sur un marché financier.

II) Problématiques

Problématique n°1 : Comment peut-on prédire l'évolution de la valeur de l'indice du S&P500 en fonction de plusieurs variables explicatives ?

Visualisation 2D des différentes variables :

Dans un premier temps nous allons chercher à effectuer une régression linéaire tri-variée, avec le temps et le volume comme variables explicatives du prix de fermeture de chaque séance.

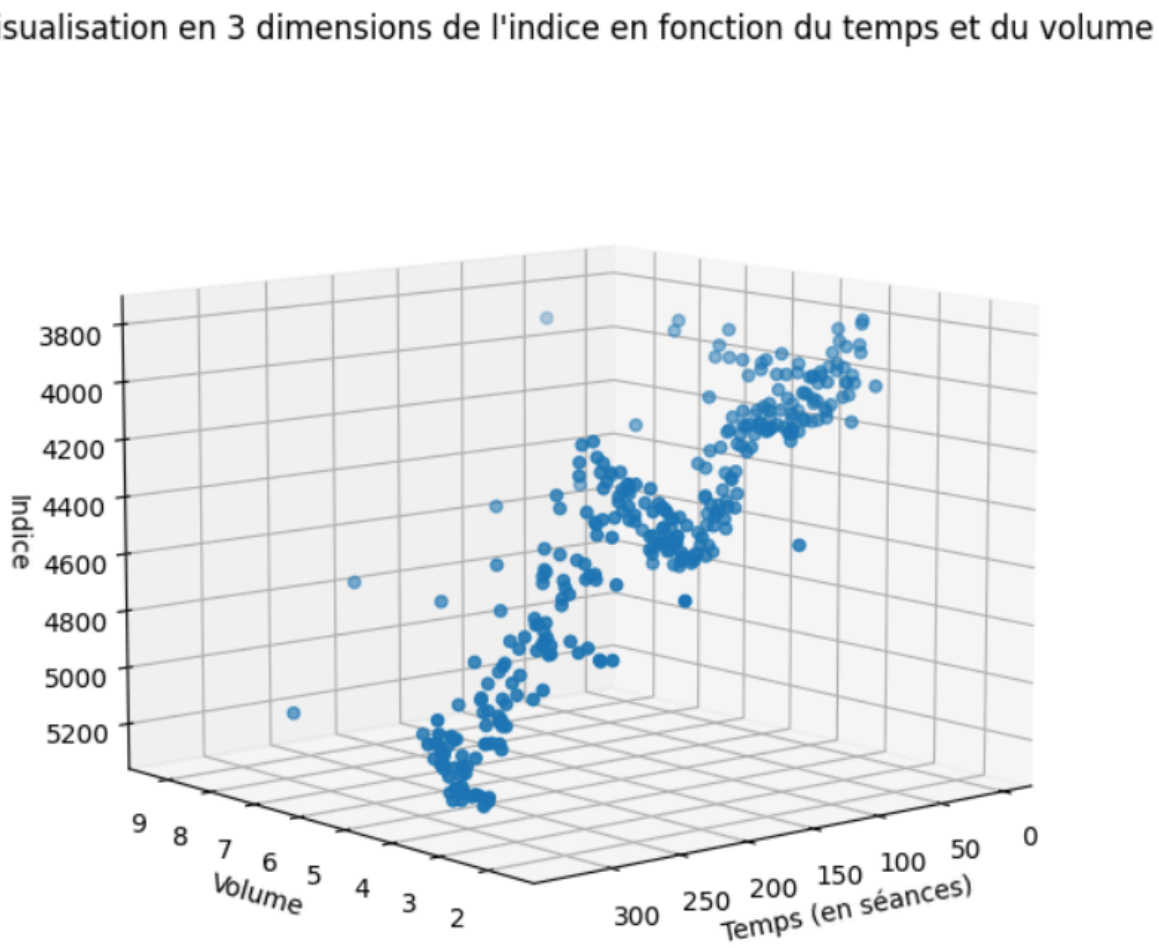


Sur ce premier graphique, on observe que le volume présente de courtes phases de pics ou de chutes, entrecoupées de périodes plus stables et plus longues. Dans le cas d'un indice financier comme le S&P500, les pics de volumes coïncident souvent avec des annonces sur les marchés financiers. À l'inverse, le volume à tendance à chuter dans des phases plus stables.

De plus, on observe aussi qu'avec le temps (ici décompté en séances), le S&P500 suit une tendance haussière.

Il est donc pertinent de commencer notre analyse avec ces 3 variables.

Visualisation 3D :



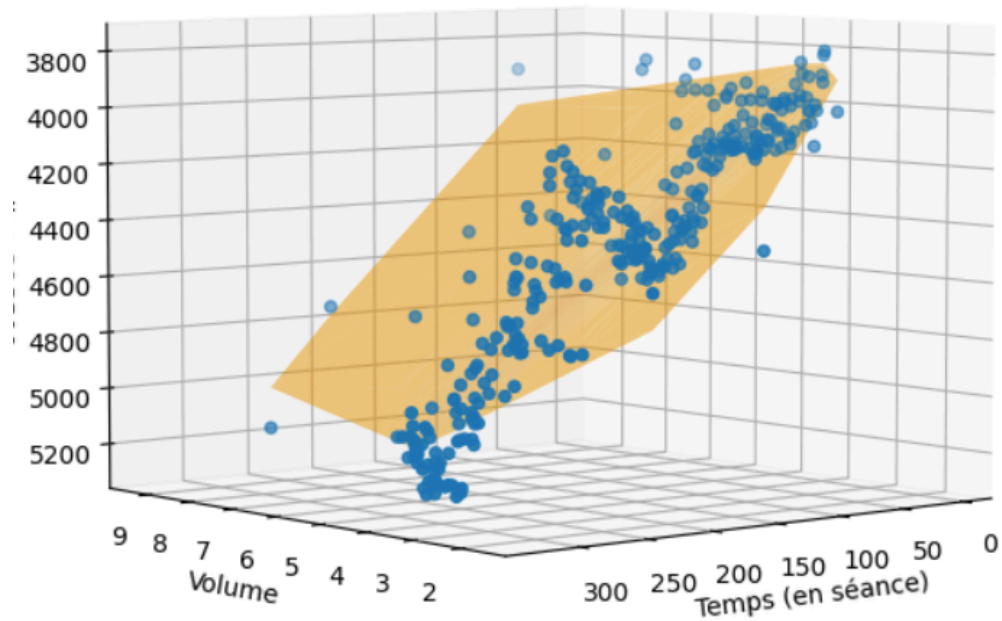
Ici nous obtenons une visualisation en 3 dimensions de nos variables, qui va servir de base à la suite d'une analyse statistique.

À l'aide du module scipy et de la fonction linalg(), nous établissons des valeurs □0,□1, □2.

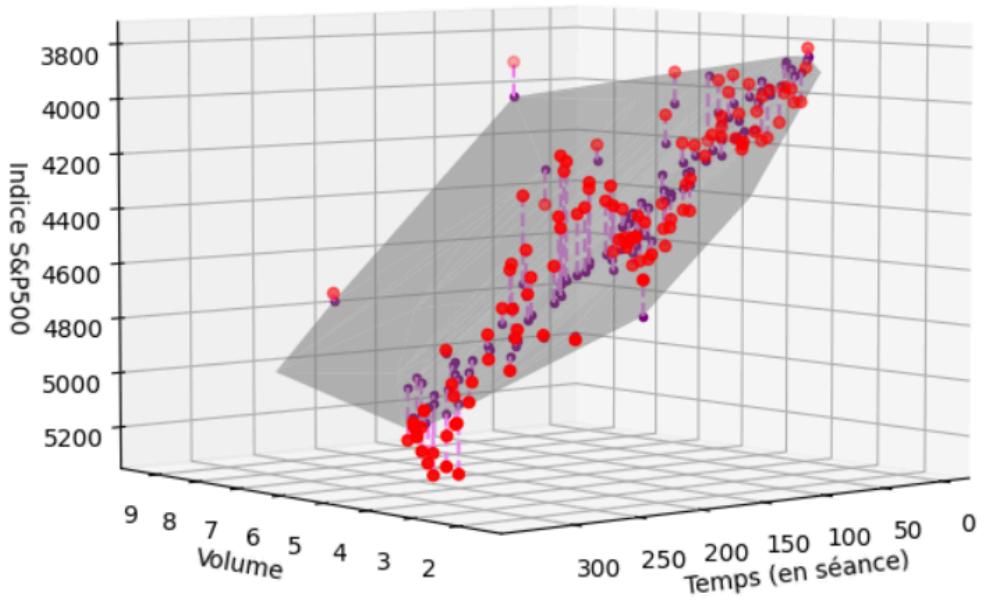
```
Beta0: 3843.6218658804482
Beta1: 3.705061449936089
Beta2: 1.4936476605468682e-09
```

Ce qui nous permet de tracer leplanaffineoptimal(ausensdesmoindrescarrés) traversant notre nuage de point, d'après l'équation □0 + □1*X0 + □*X1, avec X0 le temps et X1 le volume.

Plan affine optimal du nuage de points



Plan affine optimal du nuage de points, avec visualisation des moindres carrés



Le 2ème graphique proposé n'affiche qu'une valeur sur trois afin de rendre le tout plus lisible.

Coefficient de détermination R^2 :

Maintenant qu'on a établi un premier modèle de régression, on calcule le coefficient de détermination afin d'en déterminer la qualité.

```
Le coefficient de détermination est :  0.8526678739701683
```

La valeur de 0,85 n'est pas mauvaise, mais on peut essayer de l'améliorer. Et pour ce faire, nous allons introduire une 4ème variable : le cours du dollar par rapport à l'euro.
En effet, le S&P500 regroupe les 500 plus grosses entreprises américaines, et le dollar est la monnaie des États Unis, souvent utilisé comme étalon. Le cours du dollar face à l'euro est en général un bon indicateur de la santé des marchés américains face aux marchés européens,et semble donc pertinent à mettre en lien dans l'analyse de l'évolution de l'indice du S&P500.

Régression linéaire quadri-variée :

On ajoute donc à notre régression la variable X2 qui correspond à la valeur du dollar face à l'euro.
On obtient alors de nouveaux coefficients □0,□1, □2, et un nouveau coefficient □3, ainsi qu'un nouveau coefficient de détermination.

```
Beta0: 9199.456898586172
Beta1: 9.880438968265023e-09
Beta2: -5829.1701644192535
Beta3: 3.7005990775037887
Coefficient de détermination : 0.8909181412148661
```

Notre coefficient est passé à 0.89, ce qui montre une meilleure qualité de cette régression.
On ne peut plus visualiser le plan étant donné que l'on fonctionne désormais dans 4 dimensions.

Prédiction :

On peut désormais mettre en place un modèle de prédiction de la valeur de l'indice du S&P500, en donnant des valeurs à nos variables explicatives. Notre variable "Temps" est déjà donnée (pas de surprise dans la numérotation des séances). Nos variables "Volume" et "Cours du dollar" doivent-elles être prédites au préalable. Néanmoins, même sans connaître ces variables précisément, on peut se servir du modèle pour prédire l'évolution de l'indice en fonction des tendances du marché (volume inhabituellement élevé suite à des publications de résultats, dollar en grande forme,...).

Exemples :

En reprenant les valeurs de volume des 5 séances précédentes et en paramétrant une valeur du dollar stable :

	Volume	Cours Dollars	Séance
0	3820250000	0.9365	332
1	3751400000	0.9369	333
2	3656740000	0.9368	334
3	3958050000	0.9381	335
4	3604140000	0.9383	336

Le modèle nous donne :

```
Erreur quadratique moyenne : 126.0342852085301
[4997.88874145 4998.69736553 5002.33209592 5000.21401171 5000.31957857]
```

Les valeurs données sont cohérentes avec la valeur actuelle de l'indice.

En simulant un effet d'annonce impactant fortement le volume échangé le lundi, avant de revenir progressivement à des valeurs classiques, et une valeur du dollar identique à celle des 5 séances précédentes :

	Volume	Cours Dollars	Séance
0	6950250000	0.93824	332
1	4451400000	0.93861	333
2	4256740000	0.93420	334
3	3958050000	0.93459	335
4	3804140000	0.93470	336

Le modèle nous donne :

```
[5008.34308347 4993.11456995 5021.92774364 5021.26245652 5023.2424628 ]
```

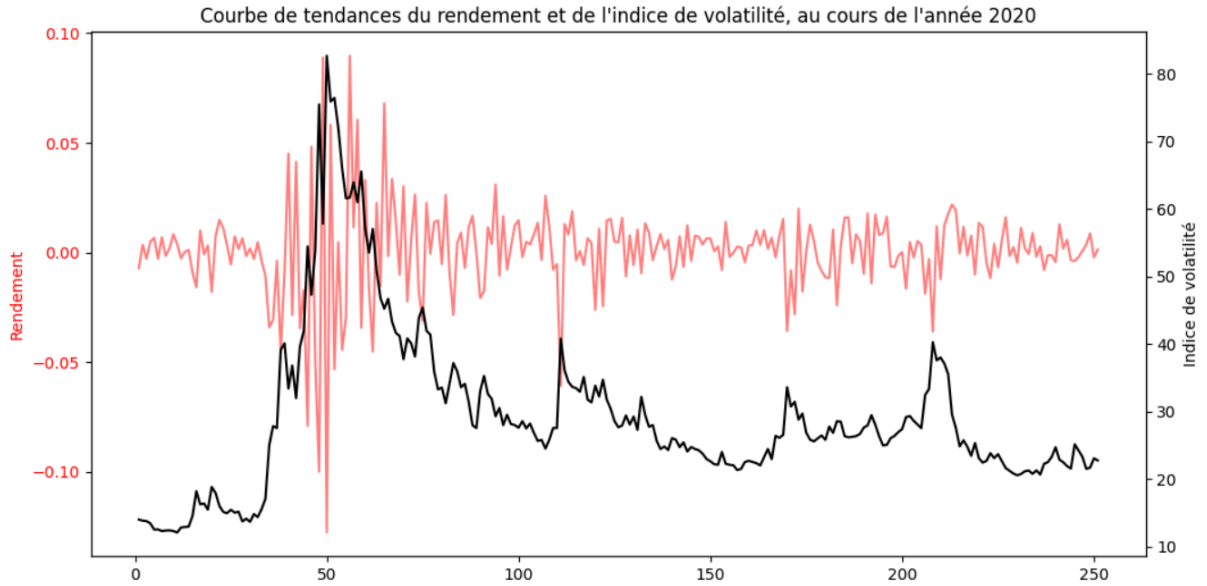
Les valeurs données sont cohérentes avec la valeur actuelle de l'indice (un "bond" de 15 points n'a rien d'extrême en période de mouvement sur le marché américain).

Limites du modèle :

De manière assez évidente, notre modèle présente de grosses limites : l'évolution de la valeur d'un indice ou du prix d'un actif en finance dépend de bien plus de 3 variables. Particulièrement dans le cas d'un indice composé de 500 entreprises, dans des secteurs très variés. De plus, pour pouvoir réellement étudier le cours du dollar il faudrait un double étalon or et euro, pour s'assurer qu'une apparente bonne performance ne soit pas basée uniquement sur une mauvaise performance du marché européen.
Néanmoins l'utilisation du volume est un indicateur intéressant car il permet de simuler des effets d'annonces, composantes très importantes des marchés financiers.
Enfin, on pourrait aussi utiliser l'évolution du cours de certaines valeurs refuges comme l'or ou le bitcoin, qui, comme leur nom l'indique, ont tendance à attirer les investisseurs lorsque le marché va mal, ou encore un indice de volatilité comme le VIX.

Problématique n°2 : Comment évolue la volatilité et les rendements du S&P500 en période de crise

Analyse préliminaire :



Ce premier graphique représente les courbes de tendance du rendement du S&P500 et du VIX en 2020. On observe une corrélation assez nette : une augmentation du VIX coïncide une plus grande variation du rendement.
Il y a en fait un lien de causalité : si la volatilité de l'indice est plus élevée, alors les risques pris sur les positions sont plus importants, ce qui résultent en des gains (ou des pertes) plus élevés, et donc une valeur absolue de rendement plus élevée.
Il est donc pertinent de mettre en perspective ces 2 variables.

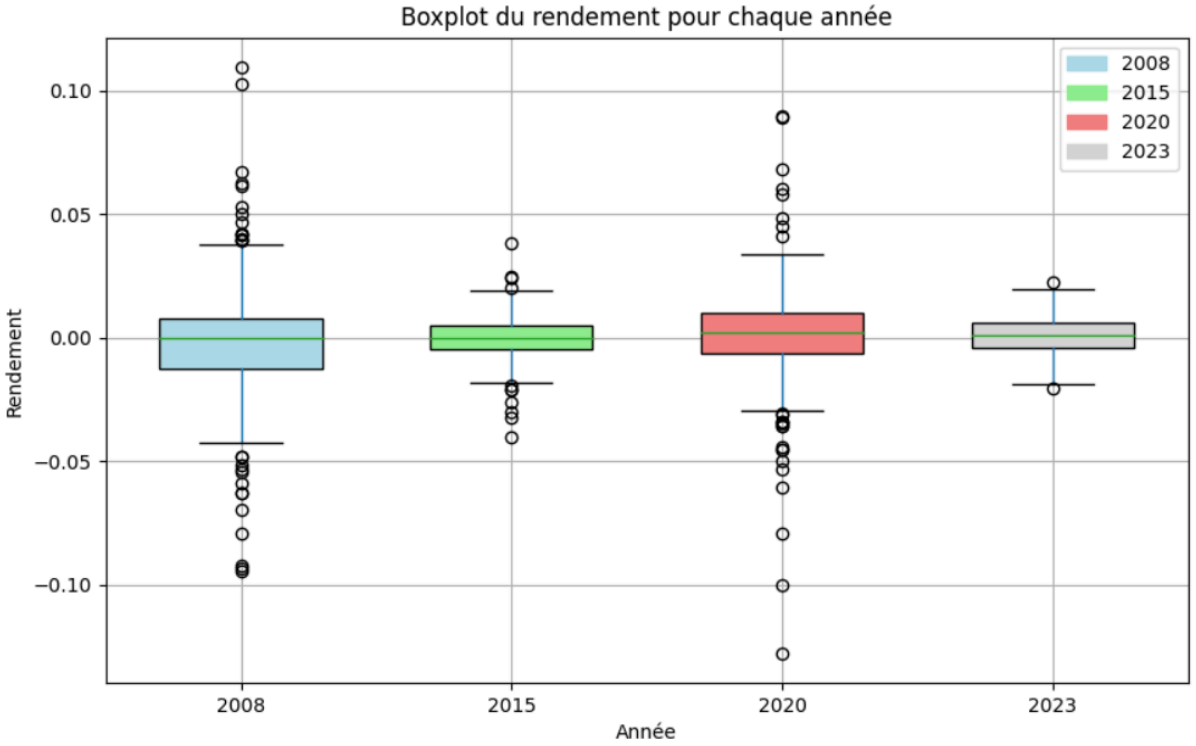
De plus, le pic du VIX sur ce graphe correspond au début de la crise COVID (confinements, augmentation des contaminations), ce qui permet d'émettre l'hypothèse qu'une période de crise provoque une forte volatilité.

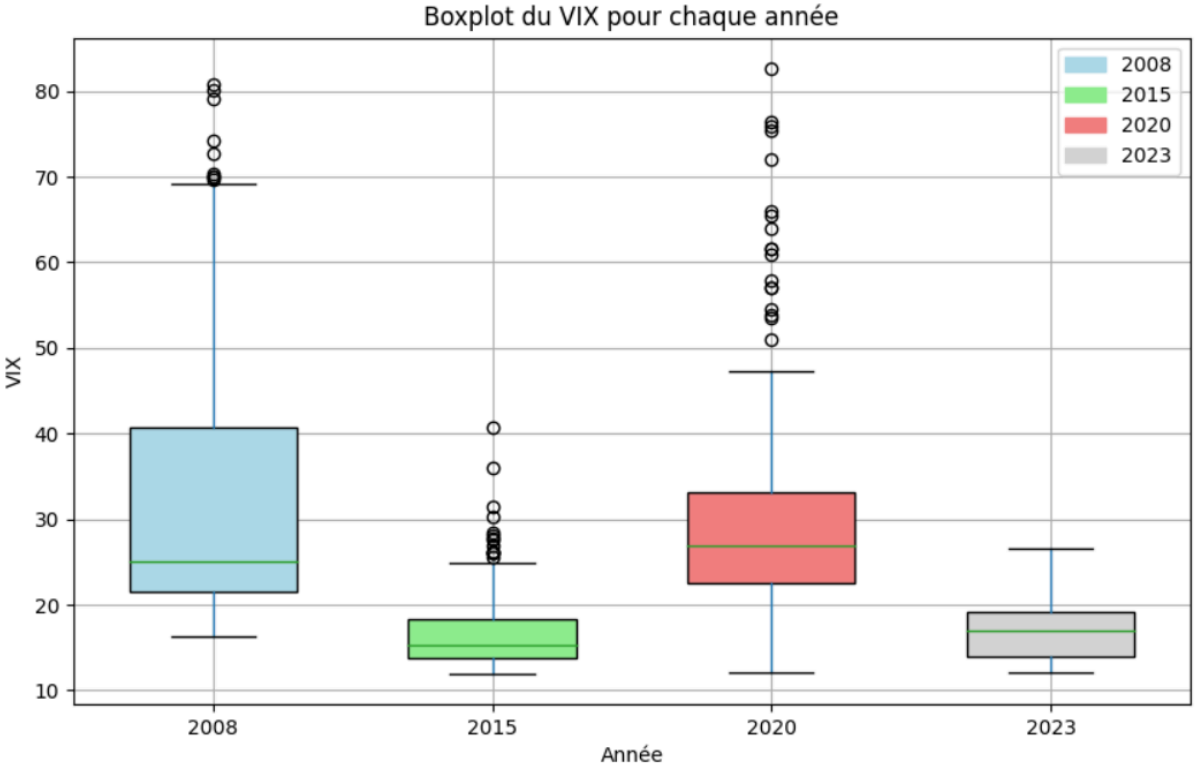
Pour plus d'indication sur le fonctionnement du VIX nous vous prions de regarder l'annexe.

Comparaison des périodes de crise et des périodes stables :

Nous allons donc répondre à notre problématique en mettant 4 années en parallèle : 2 années stables, 2015 et 2023, et 2 années de crise, 2008 (crise de subprimes) et 2020 (crise COVID). La comparaison de 2020 à 2023 est particulièrement intéressante car du point de vue des marchés l'année 2023 est considérée comme l'année de "retour à la normale" à la suite de la crise COVID.

Dans un premier temps nous allons donc tracer les diagrammes “boxplot” des 4 années, pour le rendement, et pour le VIX.

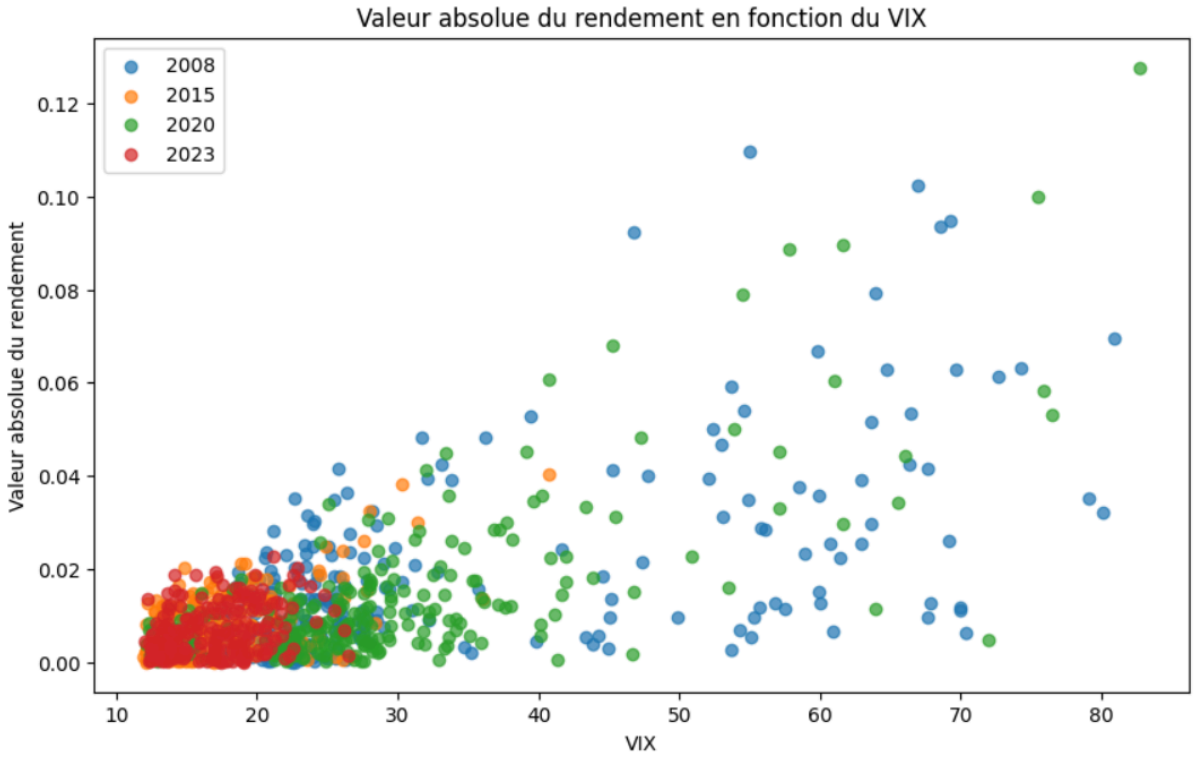




Ces diagrammes viennent clairement appuyer notre hypothèse.

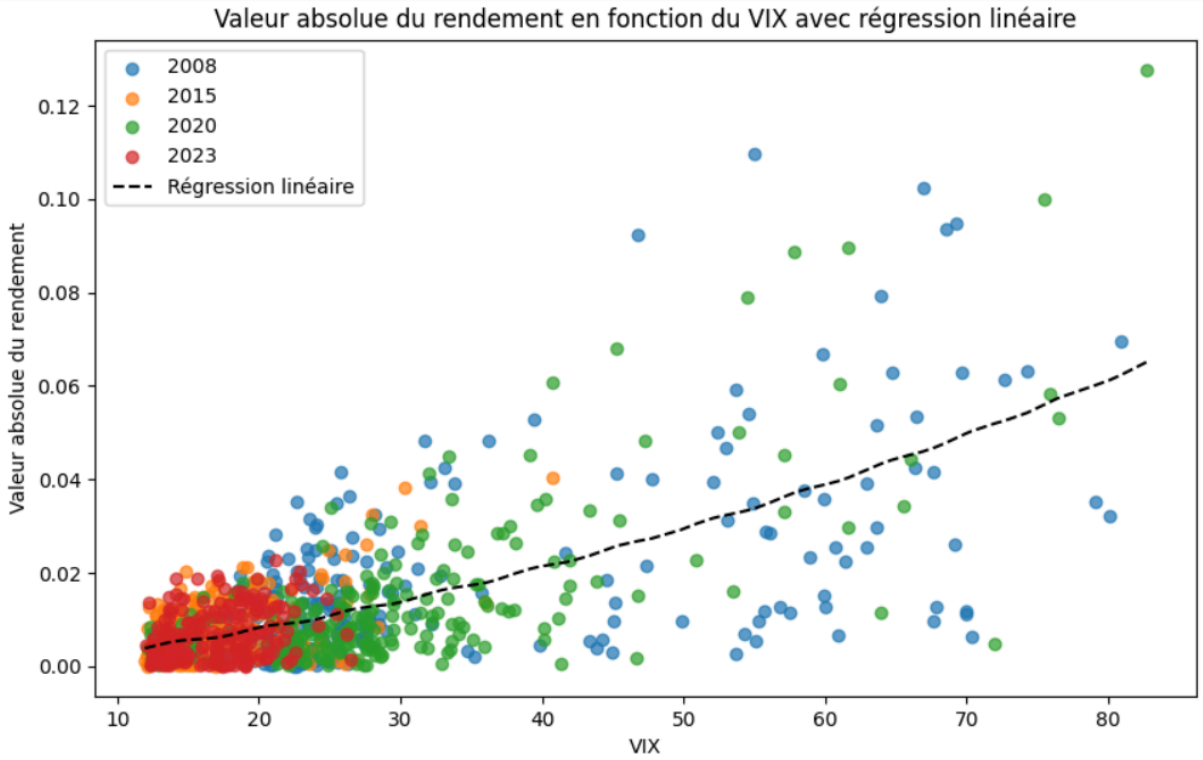
Pour le rendement la médiane oscille par définition autour de 0, ce sont les quartiles, l'écart interquartile et les “outliers” qui sont intéressants. On observe en période de crises des rendements nettement plus élevés (valeurs atypiques autour des 0,10 tandis qu’en années

stables elles sont inférieures à 0,05, et bien moins nombreuses). Les écart-interquartiles sont eux aussi plus importants.
Pour le VIX, on observe un une médiane plus élevée et un écart interquartile plus grand les années de crises. Ont observé que lors des années crises il y a de nombreux “outliers”, qui représentent des données atypiques, généralement atteintes aux moments clés de la crise. En 2020 plus d’un quart des séances (+ de 60 séances) présentaient un VIX > 30, et en 2008 un quart des séances présentaient un VIX > 40. À titre indicatif, un VIX > 30 représente une période de crise. De plus, les valeurs > 80 atteintes en 2008 puis 2020 sont des records historiques. À l’inverse, en 2015 seules 4 séances présentent un VIX > 30, et en 2023 aucune.



Enfin, on trace sur un graphe le nuage de points des valeurs absolues du rendement par rapport au VIX. On prend ici les valeurs absolues car ce qui nous intéresse est la variation absolue des rendements en période de crise (comme vu précédemment, au final ces rendements ont une médiane proche de 0 quelque soit la période). Ce graphique explicite bien le fait que des valeurs extrêmes de rendement ne peuvent être obtenues qu’en période de forte volatilité, et que ces périodes de forte volatilité coïncident aux périodes de crise. On observe par exemple que jamais un rendement absolu ne dépasse 0.10 avec un VIX inférieur à 50.
Néanmoins dans ce graphe les valeurs minimums ne sont pas réellement pertinentes car le VIX étant en réalité une valeur qui peut passer rapidement d’un extrême positif à un extrême négatif, il est courant d’observer une séance à rendement faible entre 2 séances aux rendements atypiques opposés, bien que sur l’ensemble le VI reste élevé.

Et c’est celà qui fait qu’il n’est pas vraiment possible d’établir un modèle de régression très pertinent, comme on le voit dans le prochain graphe.



En effet même une formule poussée pour notre régression ne nous donne pas un modèle très fiable :

y=☐0+☐1*x+☐2*sin(x)+☐3*x^2

```
R^2 =  0.4574163860113857
β0, β1, β2 =  [0.00000000e+00 3.09556714e-04 3.14121344e-04 5.80595234e-06]
```

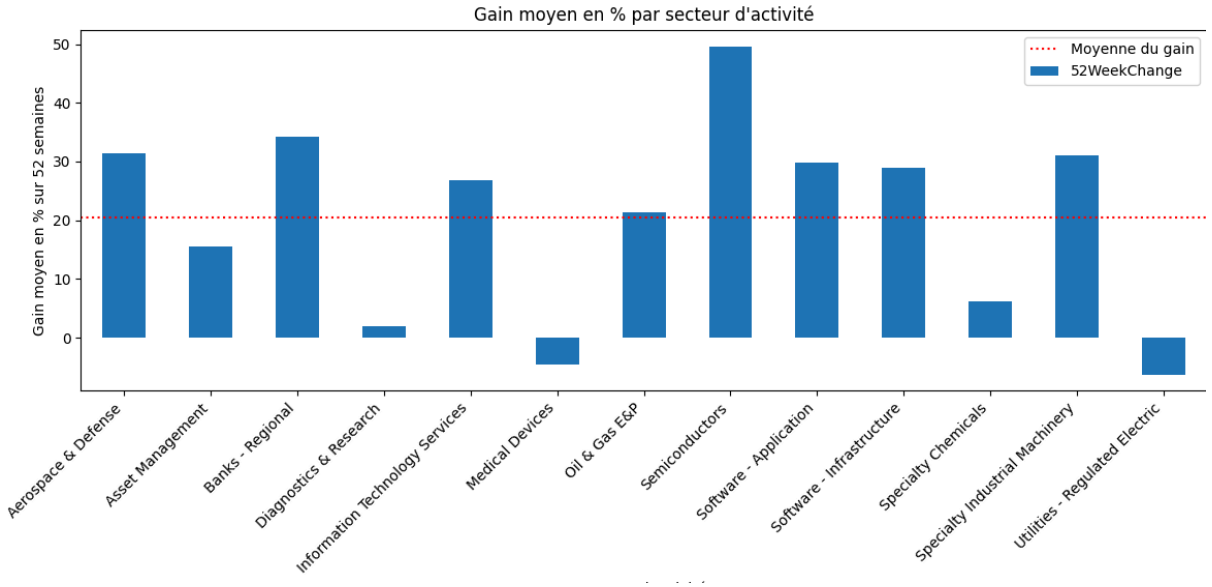
Cela n'est pas surprenant, et les différents graphes suffisent à montrer qu'un haut rendement nécessite une haute volatilité du marché, et que les périodes de crises provoquent une haute volatilité.

On peut donc conclure que qu'une période de crise entraîne bel et bien un période de forte volatilité, et donc des valeurs de rendement positifs ou négatifs plus extrêmes.

Problématique n°3 : Comment le secteur d’activité d'une entreprise, sa performance financière et son nombre d'employés sont-ils interconnectés ?

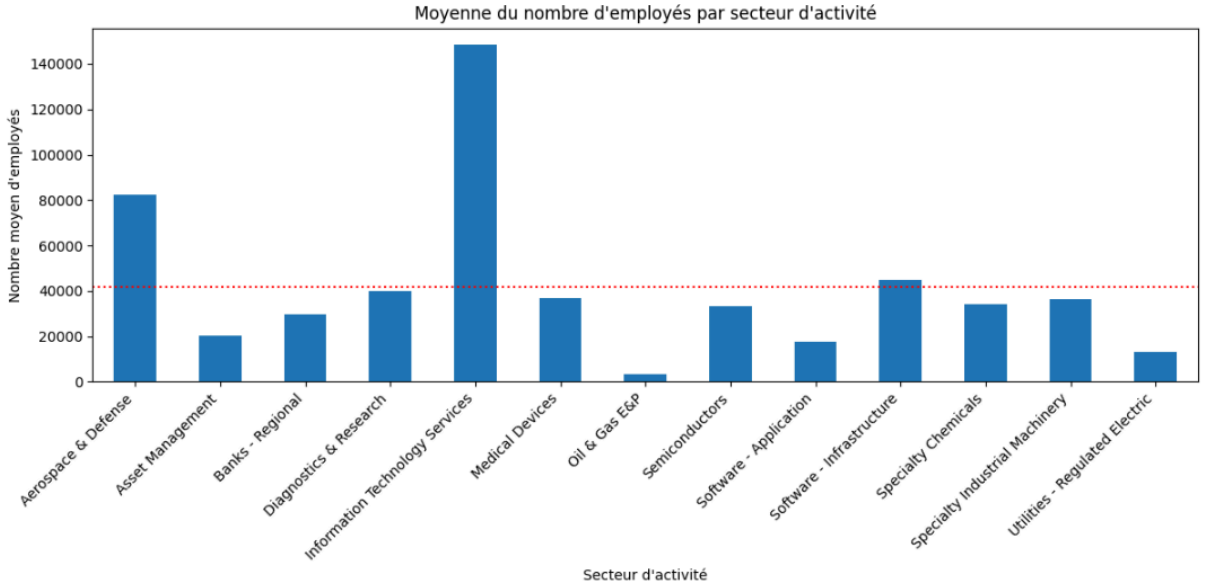
Tout d’abord, il est important de noter que toutes ces informations sont belles et bien disponibles dans le dataframe de yfinance, même le nombre d’employés.

Nous prenons ainsi l’ensemble des 13 secteurs d’activités en enlevant la catégorie “Autres” comme indiqué juste ici. En traçant le gain de toutes ces entreprises sur 52 semaines nous observons bien une tendance générale :

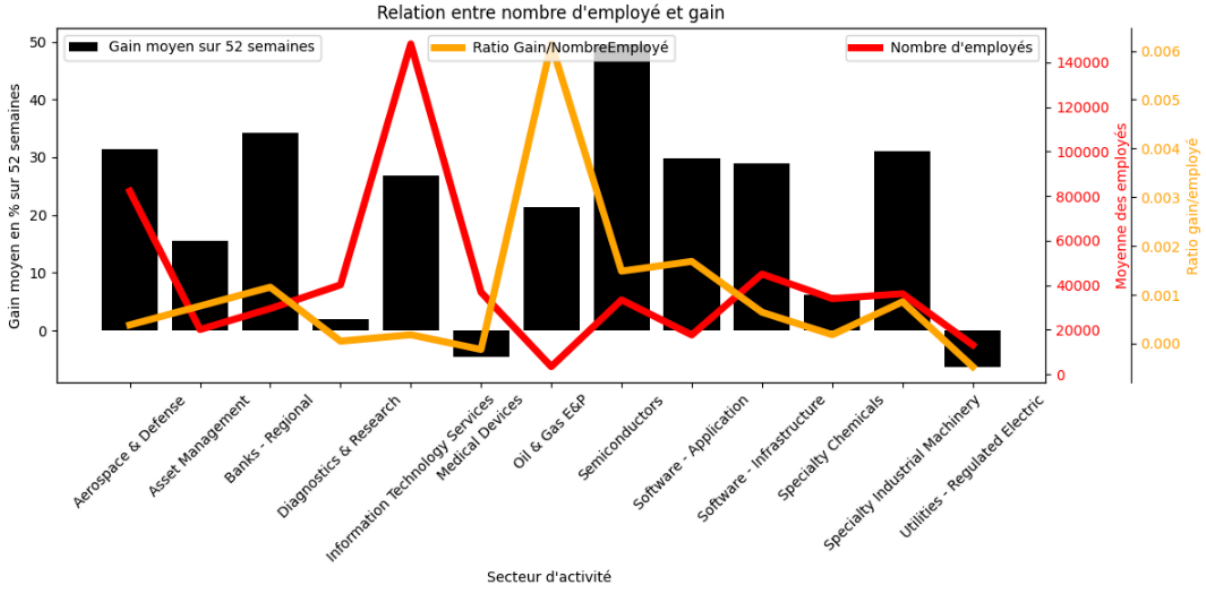


La grande majeure partie des entreprises en profite et annonce des résultats positifs. Le secteur qui profite le plus est celui des semiconducteurs, qui est porté par NVIDIA et par AVGO qui réalisent +220% et +118% respectivement sur une année. Cette forte augmentation de ces deux géants perturbe la moyenne du secteur. En effet, NVIDIA et d’autres constructeurs de carte graphique ont vu leur capitalisation boursière exploser ces derniers mois au vu de l’avènement de l’IA, nous vous renvoyons sur l’excellente analyse de zonebourse quelques mois avant la flambée de l’action : https://shorturl.at/acCYZ . De ce fait, la médiane d’augmentation du secteur n’est que de 26%, loin des 50% de moyenne.

Nous pouvons également mettre en avant le nombre d'employé dans chaque secteur



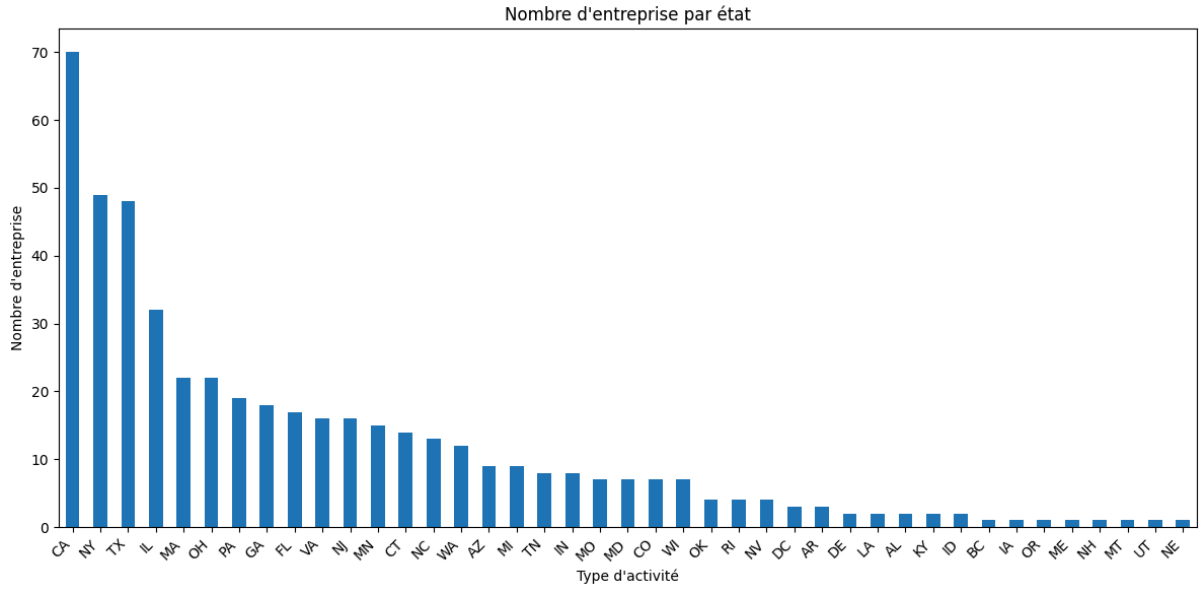
Et nous pouvons combiner ces deux graphiques afin de produire un graphique sur deux axes.



Nous avons également choisi d'ajouter une donnée représentant le gain par employé. Nous pouvons donc observer que les secteurs avec un gain négatif ou faible avec beaucoup d'employés possèdent un ratio très faible. Nous remarquons une anti-symétrie avec le domaine des "Diagnostics & Research" et celui des "Oils & Gas E&P" prouvant qu'il n'y pas de lien entre le gain d'une entreprise et son nombre d'employés.

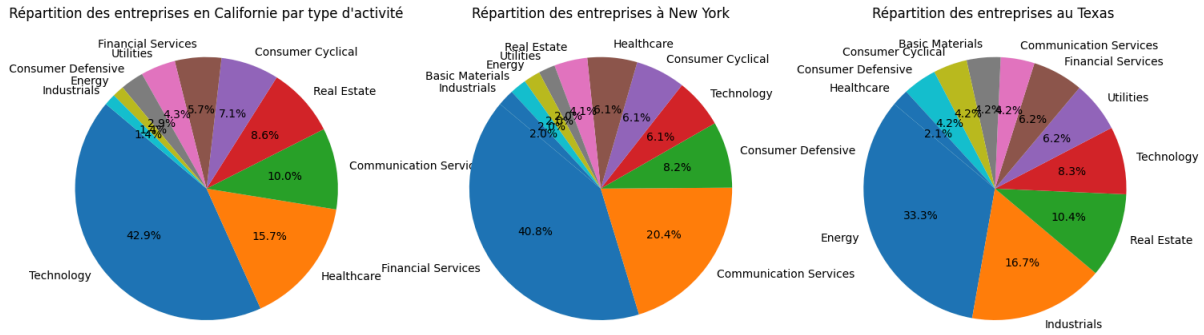
Nous souhaitions aussi comparer les gains des entreprises avec son secteur géographique. Pour cela, nous avons ce graphique de répartition des entreprises dans les différents états. Nous avons choisi ici de nous intéresser non pas au secteur d'activité mais plus au type d'activité dans un sens plus global. La différence peut sembler abstraite donc voici un exemple. Deux entreprises peuvent avoir un secteur d'activité différent comme Asset

Management et Banks - Régional mais cependant, le type d'activité est de la finance. Nous avons fait cela afin de pouvoir regrouper plus facilement nos données.



Nous avons décidé également de creuser dans les 3 états les plus représentés à savoir la Californie, le Texas et l'État de New-York.

En Californie, la technologie est en tête avec la Silicon Valley, New-York la finance est en première place (première place boursière du monde) et le Texas lui possède une majorité d'entreprise lié à l'énergie (ce qui est logique au vu de la topographie du terrain).



Ainsi, nous venons de prouver que le nombre d'employés ainsi que les gains des entreprises ne sont pas corrélés. A l'inverse, le secteur d'activité est lié à la performance d'un actif. Néanmoins, il est important de préciser que lorsqu'une nouvelle technologie est découverte, c'est toutes les entreprises du secteur qui augmentent. De plus, ces augmentations sont également à l'actualité (on pourrait parler du secteur de la défense en plein boom ces dernières semaines suite à la montée des tensions).

III) Conclusion

Au cours de cette étude, nous avons pu répondre à 3 problématiques que nous nous posions à propos du fonctionnement des marchés financiers. Chaque problématique apportait différentes variables tout en poussant un questionnement et une méthode d'approche du problème différente à chaque fois.
Si nous devions apporter un point de conclusion global sur l'ensemble de nos problématiques, il serait quelque peu critiquable. En effet, il est difficile de prévoir un marché et il est encore plus difficile d'en tirer des bénéfices, la concurrence est rude et les pertes peuvent être très importantes. Il est donc important de prendre du recul et de combiner bon sens, expérience et statistiques afin de maximiser ses gains mais surtout de limiter ses pertes.

Nous voulions absolument apporter des connaissances poussées sur le sujet et c'est pour cela qu'il existe une annexe mise à votre disposition pour tous les termes techniques.
A terme, nous aimerions automatiser cette appréciation des données afin de pouvoir la lier avec un robot trader qui pourrait, à notre place, placer des ordres d'achat ou de vente. Cette étude statistique est donc une belle entrée en matière pour notre projet personnel.

Ce fut un travail très intéressant et surtout très enrichissant dont nous sommes tous très fiers. A des fins pédagogiques, nous vous fournissons un lien Google Colab vers l'ensemble du code au début de notre annexe.

Merci pour votre lecture,

IV) Vocabulaire & Annexe

Extrait de code (vous pouvez retrouver l'ensemble du code sur le lien suivant) : https://colab.research.google.com/drive/1_HYo9LvUr_hkIv9VecdlFe_ehJO2Si4F?usp=sharing)

```
X = sp500_np[:, [8, 5]]
y = sp500_np[:, 4]

X_with_intercept = np.column_stack([np.ones(X.shape[0]), X])

X_with_intercept = X_with_intercept.astype(float)
y = y.astype(float)

coefficients = np.linalg.inv(X_with_intercept.T @ X_with_intercept) @ X_with_intercept.T @ y

Beta0, Beta1, Beta2 = coefficients[0], coefficients[1], coefficients[2]
print("Beta0:", Beta0)
print("Beta1:", Beta1)
print("Beta2:", Beta2)
```

Calcul des coefficients beta

```
fig_3d = plt.figure(figsize=(12, 7))
ax_3d = fig_3d.add_subplot(projection="3d", elev=-175, azim=130)
ax_3d.scatter(sp500_np[::3, 8], sp500_np[::3, 5], sp500_np[::3, 4], label='Données réelles', color='red')

X0 = sp500_historical['Temps (1d)']
X1 = sp500_historical['Volume']
y_pred = Beta0 + Beta1*X0 + Beta2*X1

ax_3d.plot_trisurf(X0, X1, y_pred, alpha=0.5, color='grey', label='Plan affine de régression')

ax_3d.set_xlabel("Temps (en séance)")
ax_3d.set_ylabel("Volume")
ax_3d.set_zlabel("Indice S&P500")
ax_3d.legend("")
plt.title("Plan affine optimal du nuage de points, avec visualisation  des moindres carrés")
n = sp500_np.shape[0]
for i in range(0, n, 3):
    ax_3d.plot(sp500_np[i, 8], sp500_np[i, 5], (Beta0 + Beta1*sp500_np[i, 8] + Beta2*sp500_np[i, 5]), marker=".", color="purple")
    ax_3d.plot([sp500_np[i, 8], sp500_np[i, 8]], [sp500_np[i, 5], sp500_np[i, 5]], [(Beta0 + Beta1*sp500_np[i, 8] + Beta2*sp500_np[i, 5]), sp500_np[i, 4]], color="violet",

plt.show()
```

Affichage du plan avec les moindres carrés

```python
X = merged_np[:, [2, 3, 4]]
y = merged_np[:, 1]


X_with_intercept = np.column_stack([np.ones(X.shape[0]), X])

X_with_intercept = X_with_intercept.astype(float)
y = y.astype(float)

coefficients = np.linalg.inv(X_with_intercept.T @ X_with_intercept) @ X_with_intercept.T @ y
Beta0, Beta1, Beta2, Beta3 = coefficients

X0 = merged_df['Séance']
X1 = merged_df['Volume']
X2 = merged_df['Cours Dollars']
y_pred = Beta0 + Beta1*X0 + Beta2*X1 + Beta3*X2

reg_mod = LinearRegression().fit(X, y)
r_squared = reg_mod.score(X, y)


print("Beta0:", Beta0)
print("Beta1:", Beta1)
print("Beta2:", Beta2)
print("Beta3:", Beta3)
print("Coefficient de détermination :", r_squared)
```

Régression améliorée avec le cours du dollar

```python
future_dates = pd.date_range(start='2024-04-29', periods=1, freq='D')
future_volume = [6950250000,4451400000,4256740000,3958050000,3804140000] #REMPLIR AVEC DES DONNEES COHERENTES
future_dollar_price = [0.93824,0.93861,0.93420,0.93459,0.93470]
future_seance = [332,333,334,335,336]

future_data = pd.DataFrame({
    'Volume': future_volume,
    'Cours Dollars': future_dollar_price,
    'Séance': future_seance
})

print(future_data)
```

```
       Volume  Cours Dollars  Séance
0  6950250000        0.93824     332
1  4451400000        0.93861     333
2  4256740000        0.93420     334
3  3958050000        0.93459     335
4  3804140000        0.93470     336
```

Prédiction de la valeur de l'indice

```
X = merged_df[['Volume', 'Cours Dollars', 'Séance']]
y = merged_df['Indice']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

predictions = model.predict(X_test)

mse = mean_squared_error(y_test, predictions)
print("Erreur quadratique moyenne :", mse ** 0.5)

future_predictions = model.predict(future_data)

print(future_predictions)
```

Prédiction de la valeur de l'indice

```
plt.figure(figsize=(10, 6))
plt.scatter(data_2008['VIX'], abs(data_2008['Return']), label='2008', alpha=0.7)
plt.scatter(data_2015['VIX'], abs(data_2015['Return']), label='2015', alpha=0.7)
plt.scatter(data_2020['VIX'], abs(data_2020['Return']), label='2020', alpha=0.7)
plt.scatter(data_2023['VIX'], abs(data_2023['Return']), label='2023', alpha=0.7)

x_vect = all_data["VIX"]
y_vect = abs(all_data["Return"])
n = len(x_vect)
scores = []
scores_ajustes = []
p = 3
x_mod_tab = np.ones((n, p+1))
x_mod_tab[:, 1] = x_vect
x_mod_tab[:, 2] = np.sin(x_vect)
x_mod_tab[:, 3] = x_vect**2


reg_mod = LinearRegression().fit(x_mod_tab, y_vect)
score = reg_mod.score(x_mod_tab, y_vect)
scores += [score]
scores_ajustes += [1-(n-1) * (1-score) / (n-(p + 1))]
print("R^2 = ", score)
print("β0, β1, β2 = ", reg_mod.coef_)
print(reg_mod.intercept_)
y_hat = reg_mod.intercept_ * x_mod_tab[:, 0]
for k in range(1, p+1):
  y_hat += reg_mod.coef_[k] * x_mod_tab[:, k]
plt.plot(x_vect, y_hat, color='black', linestyle='--', label=f'Régression linéaire')
```

Tentative de régression pour la problématique n°2

```
fig, ax1 = plt.subplots(figsize=(12, 6))
ax1.bar(evolution_prix_moyen_sans_autre.index, evolution_prix_moyen_sans_autre.values, color='black')
ax1.set_xlabel("Secteur d'activité ")
ax1.set_ylabel("Gain moyen en % sur 52 semaines")
ax1.tick_params(axis='y')

ax2 = ax1.twinx()
ax2.plot(moyenne_employes_par_secteur_filtre.index, moyenne_employes_par_secteur_filtre.values, color='red',linewidth=5)
ax2.set_ylabel("Moyenne des employés", color='red')
ax2.tick_params(axis='y', labelcolor='red')

ax3 = ax1.twinx()
ax3.spines['right'].set_position(('outward', 60))
ax3.plot(ratio_employe_gain.index, ratio_employe_gain.values, color='orange',linewidth=5)
ax3.set_ylabel('Ratio gain/employé', color='orange')
ax3.tick_params(axis='y', labelcolor='orange')

ax1.tick_params(axis='x', rotation=45)
ax2.tick_params(axis='x', rotation=45)
ax3.tick_params(axis='x', rotation=45)

ax1.legend(["Gain moyen sur 52 semaines"], loc="upper left")
ax2.legend(["Nombre d'employés"], loc="upper right")
ax3.legend(["Ratio Gain/NombreEmployé"], loc="upper center")

plt.title("Relation entre nombre d'employé et gain")

plt.tight_layout()
plt.show()
```

Graphique du ratio gain/employé, employé et gain par secteur pour la problématique n°3

Vocabulaire :

Actif financier : Un actif financier est un titre ou un contrat, généralement négociable sur un marché financier. Il y en a de très nombreuses sortes, des plus simples : actions, obligations, aux plus complexes : options, swaps, dérivés de crédit…Dans cette étude nous étudierons uniquement les actions.

Volatilité : La volatilité est un indicateur qui mesure l'amplitude de ces hausses et de ces baisses.

VIX : L'indicateur VIX est souvent appelé l'indice de la peur des marchés financiers.Son calcul repose sur la volatilité des options d'échéance 30 jours sur le S&P 500. Cette volatilité implicite est elle-même obtenue à partir de la médiane entre le prix d'achat et le prix de vente de toutes les options. Un VIX inférieur à 30 indique une confiance certaine et un VIX supérieur à 70 indique une peur extrême.