UNIVERSITÉ SORBONNE PARIS NORD
M1 BIDABI

INTRODUCTION A PYTHON

Travaux portants sur l'analyse d'actifs financiers sous python

❖ Participant:

- **KAMELA K. Alesterd Percys** N° Etudiant : 12214385

❖ Sous la Supervision de :

- **Mme JABRI RIM**

Structure du rapport.

INTRODUCTION ET PROBLEMATIQUE

I. COLLECTE ET PREPARATION DES DONNEES

- Source et Collecte de donnée
- Nettoyage des données
- Transformation des données

II. L'ANALYSE ET LE TRAITEMENT

- Exploration et Analyse des Données
- Modélisation et Prédiction

III. VISUALISATION DES DONNEES.

- Présentation des outils et librairies de visualisation graphique
- Présentation des résultats de l'analyse et de la modélisation
- Fournir des recommandations d'investissement sur la base des analyses

CONCLUSION

INTRODUCTION ET PROBLEMATIQUE

Pendant les années 1950, beaucoups d'acteurs économiques investissaient sur la base des recommandations de d'autres personnes. Les décisions d'investissement n'étaient pas issues d'un processus de décisions pertinent et développé. Ce qui a conduit à plusieurs scandales financiers qui n'ont cessé de resurgir au cours du temps sous plusieurs formes à savoir, la crise des subprimes aux USA, la covid à l'international, Evrgrande en chine, et plus récemment le cas des banques SVB.

Tous ces scandales financiers ont conduit à une forte règlementations et lois vis-à-vis de tous les acteurs des marchés financier, ainsi qu'une intégration total des technologies de l'information et de la communication pour facilité et mieux gérer les différents acteurs et intervenants des marchés ainsi qu'une augmentation du nombre d'actifs financier.
Notre préoccupation est celle de savoir comment utiliser l'historique des prix pour étudier l'évolution des actifs financier, avec pour objectif de déterminer les risques et recommander un actif correspondant au profil et aux préférences d'un investisseur.

I. COLLECTE ET PREPARATION DES DONNEES

1. Les ressources utilisées pour ce travail

- L'environnement de travail python et ses librairies (pandas, numpy, matplotlib) ainsi que certains sites internet tel que stackoverflow, google trend, Yahoo finance, google colab

- Les données financières utilisées dans le cadre de cette étude sont des données structurée, issues de la cotation journalière des actifs GAFAM allant de 2007 à nos jours. Soit une durée minimum de **5840 jours**. Ces actifs cotés à la bourse américaine du NYSE et ces choix se justifie par la gratuité des ressources.

```
assets = ["AMZN","META","MSFT","GOOG","TSLA","NVDA"]
start = dt.datetime.today() - dt.timedelta(5840)
end = dt.datetime.today()
```

- **Source et Collecte de donnée**

La collecte de données s'est faite via l'API (Application Programing Interface) de Yahoo finance qui est la branche de Yahoo qui s'occupe de la finance et des marchés financier et qui met gratuitement à disposition des données de cottions d'actif sous différentes formats (API, CSV, JSON). Augmenter a cela l'accès est rapide et peut facilement être importer dans python avec les paramètres OHLCV (Open, High, Low, Close, Volume) relatif à l'actif souhaiter. Nous avons travaillé avec la modalité associée au prix de clôture des actifs

```
for price in assets:
    p_close[price] = yf.download(price,start,end)["Close"]
```

2. Transformation des données

Nous avons procédé à l'étape de transformation avant le nettoyage paracerque toutes les données issues de l'API n'étaient pas utile. L'étape de transformation consistait essentiellement a créé des data-Frames pour stocker les données de prix et volume essentielles pour la suite de nos travaux.

```
# Creation de la Dataframe
p_close = pd.DataFrame()

# Creation de la Dataframe
p_volume = pd.DataFrame()
```

3. Nettoyage des données (Data Frames)

L'étape essentiel avant le nettoyage était la sélection des données parmi celles issues de la source de données. Apres inspection des partie entête (Head) et fin (Tail) de nos données, nous avons procédés à un **remplacement par zéro** des valeurs manquante (Nan) sur la cotation de **Meta** (Facebk) et **TSLA qui n'étaient pas encore coté en 2007**

- Données initiales avec **présence de NAN**

META	MSFT	GOOG	TSLA
NaN	28.850000	11.775861	NaN
NaN	28.600000	11.855812	NaN

- Données Final **Nettoyé**

META	MSFT	GOOG	TSLA
0.000000	28.850000	11.775861	0.000000
0.000000	28.600000	11.855812	0.000000

- Nous avons recensé trois méthodes de suppression des NAN, tel que le remplacement par la moyenne des données non NAN, suppressions des cellules contenant les NAN, remplacement des NAN
- Nous avons opté pour le remplacement des valeurs manquantes (NAN) par zéro afin de ne pas modifier la structure de nos donnée et erroné nos résultats et analyses.

```
close_p = p_close.fillna(0)
```

II. L'ANALYSE ET LE TRAITEMENT DES DONNEES

1. Exploration et Analyse des Données

```
# observation du head et tail
close_p.head(100)
```

Date	AMZN	META	MSFT	GOOG
2007-04-17	2.2535	0.0	28.850000	11.775861
2007-04-18	2.2495	0.0	28.600000	11.855812
2007-04-19	2.2320	0.0	28.690001	11.747219
2007-04-20	2.2475	0.0	29.020000	12.016957
2007-04-23	2.2385	0.0	28.780001	11.932275

L'observation du **Head** et **Tail** de nos données nous révèle que nos données sont des **booléens** de types **série financière continue**, car plusieurs facteurs (prix, volumes…) sont étudiés et l'intervalle n'est pas fini, mais continue et axé sur l'évolution journalière.

- Les valeurs quand a elles nous montrent que le prix de certains actifs évolue plus vite que d'autres.
- Certains actifs clôturent à la hausse (prix du jour n supérieur a celui du jour n -1) contrairement à d'autre clôturent à la baisse (prix du jour n inférieur à celui du jour n -1)

2. Modélisation et Prédiction

Les outils statistiques utilisé pour notre analyse sont :

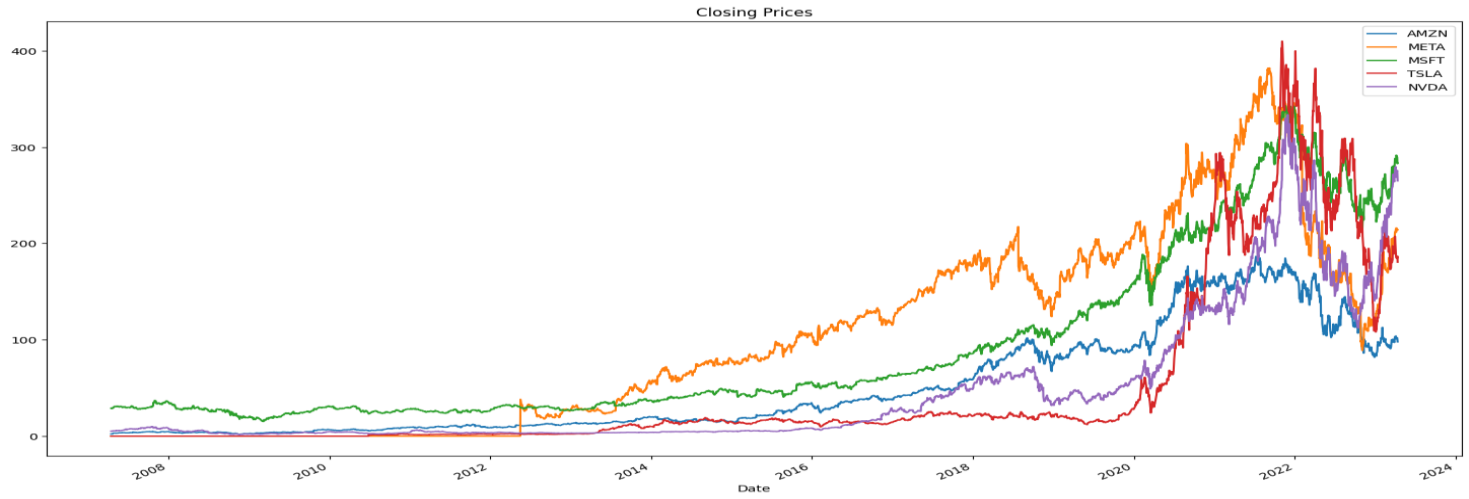
- **Etude des tendances** pour observer l'évolution des prix suivant des phases de hausse et de baisse
- **Etude de la volatilité** permet d'étudier **la volatilité** (vitesse et fréquence) de variation du prix de l'actif étudier, permettant ainsi d'identifier le risque associé aux variations du prix de l'actif
- **Etude des moyennes mobiles**, il est complémentaire à l'étude de tendance dans le but d'identifier les patterns de l'actif étudier
- **Nuage de points** pour visualiser les relations bivariées entre toutes nos variables avec chaque mini-diagramme qui représente la relation entre une paires de variable
- **Histogramme et Box-plot** aident à identifier et ressortir des propriétés propres à la structure de l'actifs

III. VISUALISATION DES DONNEES

1. Présentation des résultats et statistiques et Analyse

Afin d'améliorer notre compréhension des chiffres décris dans l'analyse précèdent, nous avons procédé à une étude graphique afin de voir le visuel qui se cacha dans nos données.

- **Etude de tendance (**observer l'évolution des prix en phases de hausse et de baisse)



Une Interprétation du graphique ci-dessus est la suivante :

- Depuis 2012, tous ces actifs connaissent une croissance avec des prix qui évoluent significativement à la hausse.
- De 2020 à nos jours, les prix se chevauchent entre eux et atteint de nouveau points hauts ou points bas
- L'actif avec la plus grande variation et prix le plus haut est Tesla

Ajouter à ces points, nous pouvons associer d'autres facteurs tel que :

- Les outils utiliser pour coter les actifs et réguler les marchés financiers ne cesse d'évoluer
- Les cycles économiques dépendent majoritairement de la politique économique appliqué dans la zone économique concerné et qui est relatif à la durée du mandat présidentiel (4 ans aux USA et 5 en France)

Tous ces facteurs nous permettent de réduire notre analyse de la volumétrie sur une durée de 8 ans qui correspond à 2 mandats présidentiel, soit 2 cycles économiques. Tout en tenant compte du fait que nos actifs sont cotés sur la bourse Américaine du NYSE et sont soumis aux lois Américaines.

- **Etude de la volatilité Visualisation des données (Volumétrie)**

- Création de la data frame **basé sur le volume**

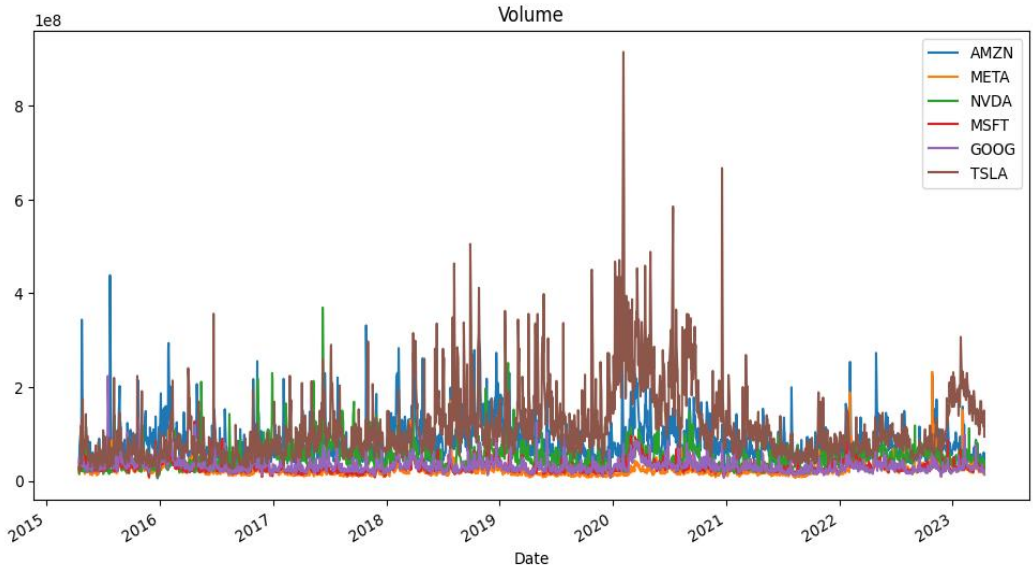
```
for price in assets:
    p_volume[price] = yf.download(price,start,end)["Volume"]
```

- **Remplacement par zéro des NAN** (si elles existe dans notre Data Frame)

Le Plot de la volumétrie se présente comme suite :

Sur la base de cette image, **nous constatons que :**

- Tesla possède la plus forte et grande volumétrie par rapport aux autres actifs
- Pour la suite, nous avons refait un deuxième plot de volumétrie **en retirant l'actif Tesla**



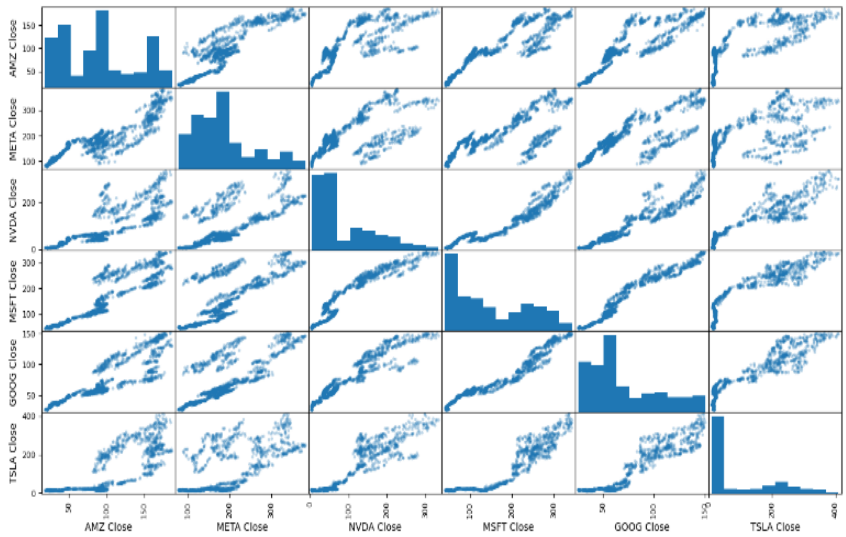
Nous avons progressivement répété le procédé. Ce qui nous a permis de classer nos actifs en deux catégories (ceux à forte volatilité, ceux à moyenne et basse volatilité) Modélisation et Prédiction

- **Actifs à forte volatilité :** Nous remarquons une **dominance** de TESLA (en bleu), suivi de AMAZON (orange) et enfin NVIDIA (vert). Nous pouvons encore dire que Tesla est l'actif dont le prix possède la plus grande variation le plus, suivi par Amazon et enfin NVIDIA

- **Actifs à moyenne volatilité** Nous avons une **dominance** de **META** (en bleu), suivi de **GOOGLE** (vert) et enfin **MICOSOFT** (orange). Nous pouvons encore dire que Meta est l'actif dont le prix possède la plus grande variation, suivi par Google et enfin Microsoft.

- **Présentation des Nuages de points** (relations bivariées entre toutes nos variables)

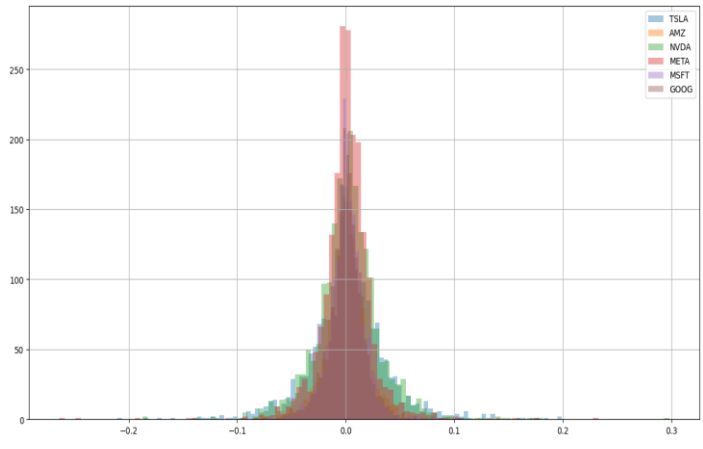
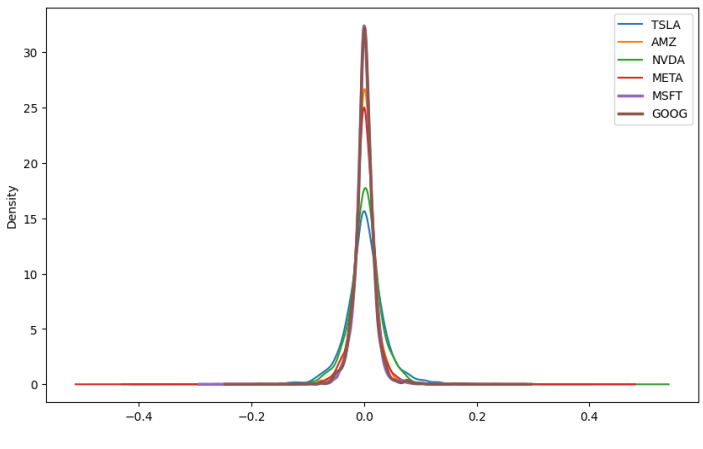
Une interprétation des relations entre nos variables peut être faite comme suite

GRAPHIQUE ET VISUEL DU NUAGE DES POINTS	VALEURS NUMERIQUE DES CORRELATIONS
	

ACTIFS AVEC NIVEAUX DE CORRELATION		
FORTE	**MOYENNE**	**FAIBLE VOIR ABSCENTE**
Microsoft-Google, Nvidia-Microsoft Microsoft et Nvidia, Google et Nvidia Google-Tesla, Microsoft et Amzone	**Google et Meta**, Tesla et Nvidia **Tesla et Amazone**	Nvidia et Meta, **Tesla et Meta**
Forte synchronisation de l'évolution du prix des deux actifs au cours du temps	Synchronisation Moyenne sur l'évolution du prix des deux actifs cours du temps	Absence de Synchronisation sur l'évolution du prix des deux actifs au cours du temps

- **Présentation de l'Histogramme, Densité et des Box-plots (Boites à Moustaches)**

L'histogramme nous a permis d'étudier l'évolution du rendement moyen journalier. Ce rendement journalier se calcul grâce à la formule $Rt = (\frac{Pt}{Pt-1}) - 1$

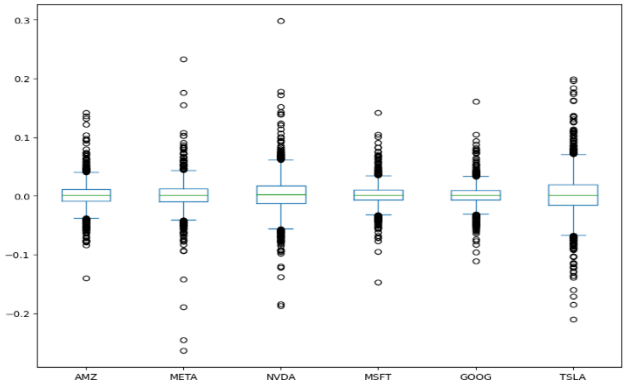
Histogramme	KDE (Kernel Density Estimation)
	

Interpretation se fera sur 2 dimensions (en ordonnées : le niveau atteint et en Abscisse : Frequence) de Volatilité

- l'histogramme a gauche nous montre les variations des rendements moyennes journalier des actifs
  - les rendement journaliers sont fortement centré autour de 00 et varient majoritairement entre -0.1 et +0.1
  - il montre la moyenne des rendements **positif** (coté droit) ou negatif (coté gauche), Ce rendement **est positif** lorque le prix de l'actif coture à la hausse entre 2 jours successif ou négatif s'il cloture à la baisse
  - nous observons que **META** est l'actif avec le rendement moyen (positif et negatif) **le plus élevé et stable**
- Le graphique d'estimation de Denssité (KDE) qui nous permet de visualisé la distribution graphic de nos actif conformement aux interpretations ci-dessus, il ressort egalement le niveau atteint par chaques actif tel que google qui se superpose avec Microsoft, suivi par Amazon, Meta(facebook), Nvidia et enfin tesla.
- La Base des deux graphiques (Kde+histo) montrent que **Tesla et Nvidia** ont les plus grandes Fréquences, mais sont déséquilibré sur le niveau qui peut etre plus positif que négatif et inversement
- **Google et MSFT** se superposent et dominent tous les autres actifs en matière de Fréquence et Niveau(positif, negatif) du rendement Moyen de nos actifs

- **Histogramme et Box-plot**

Le box-plot ci-dessous est associé au rendement moyen journalier calculé en amont. Il nous montre une distribution de nos données et de valeurs extrême. Ainsi, nous observons que :



- La plage des données est compris entre -0.2 et +0.3

- Tesla possède le box **le plus grand** contrairement à google et Microsoft qui possèdent les plus petit et resserré. Cela traduit la faible variations des prix de clôture de Microsoft et Google, contrairement à de fortes variations du prix de clôture de Tesla ainsi que sa forte fréquence de volatilité.

- La médiane est unique (0) et le niveaux des quartiles (1^{er} ,2^{e},3^{e},4^{e}) est relatif au comportement de l'actif.

Les valeurs aberrantes sont représentées par les petits cercles (°°°) en dehors de la plage du nuage de points et peuvent être interpréter comme suite :

- Tesla possède de fortes et large variation du rendement, contrairement à Microsoft Où c'est très resserré. Ajouter à cela, Meta possède des niveaux de rendement plus élevé que Tesla, et les espaces révèlent que les rendements de Tesla sont plus fréquents que celles de Meta
- AMZ et Nvidia possède un nombre de rentabilités positif qui dépasse les rentabilités négatives

CONCLUSION

Parvenue à la fin de notre étude qui consistait à étudier le comportement et l'évolution des actifs financier dans le but de recommander à un potentiel investisseur, il en ressort que la recommandation d'un actif peut se faire sur la base des points développées tout au long de notre étude qui sont (la tendance, la volatilité, la rentabilité moyenne sans oublier l'interaction et la corrélation entre les actifs) a ces éléments seront combinées d'autres facteur relatif au profil de l'investisseur tel que son niveau et profil de risque (risquophobe, risquophile), la durée de l'investissement (court, long, moyen terme), le nombre d'actifs et le capital allouer a chacun des actifs.

La suite de nos travaux consistera a créé un algorithme de trading qui sur la base d'élément présenter dans notre projet associé à d'autres conditions de marchés, pourront (sous l'accord du client) placer des ordres d'achat ou de vente de l'actif souhaiter par le client.