Présentation du projet :

L’objectif de ce projet est d’analyser systématiquement l'impact de différentes variables économiques sur le S&P 500, en utilisant deux méthodes **statistiques** et **machine learning.** Pour la méthode statistique nous avons utilisé :

* Le score R² & l'erreur quadratique moyenne (MSE),

ainsi que des métriques supplémentaires comme :

* L'Akaike Information Criterion (AIC) & le Bayesian Information Criterion (BIC)

pour évaluer les combinaisons de variables. Cela permet d’obtenir une combinaison d’indicateurs macro et micro permettant de traquer la variance du S&P. Cependant, cette méthode n’informe pas sur la pondération de chaque indicateur. Elle se concentre sur les principaux contributeurs de l’explication de la variance du S&P.

Pour la méthode en machine learning, nous avons appliqué le **PCA (Principal Component Analysis)** pour réduire la dimensionnalité du jeu de données, en nous concentrant sur les caractéristiques les plus influentes.

Pour aller plus loin, nous avons introduit des décalages (lags de 5 à 60 jours) dans les variables indépendantes pour créer des indicateurs avancés pour le S&P 500 et mesurer ce dernier pour les 2 méthodes.

Lexique :

**Matrice de covariance** : Mesure la variance partagée entre variables ; indique comment elles évoluent ensemble.

**Matrice de corrélation** : Mesure normalisée (-1 à 1) de la relation linéaire entre variables ; 1 = corrélation parfaite, 0 = aucune.

**R²** : Indique la part de variance expliquée par le modèle ; 1 = ajustement parfait, 0 = aucun.

**MSE** : Moyenne des erreurs au carré entre valeurs prédites et réelles ; plus c’est bas, meilleur est l’ajustement.

**AIC**: Mesure la qualité d’ajustement en pénalisant la complexité ; plus c’est bas, meilleur est le modèle.

**BIC** : Similaire à l’AIC, mais avec une pénalité plus forte pour la complexité ; favorise souvent des modèles plus simples.

Conclusions de l’étude :

* Le composite issu de l’étude statistique révèle un R² = 97.56% et un MSE de 0.14% pour un décalage de 53 Jours
* Le composite issu du PCA révèle un R² = 93.44% et un MSE de 0.37% pour un décalage de 5 Jours

Nous constatons que le composite avancé issu des résultats du PCA n’est pas aussi performant que celui déterminé par l’étude statistique

* **A faire : Backtest sur quelques années et détermination (en %) de la véracité du modèle par année + Explications si divergences conséquentes**

### Étapes clés du processus :

### 1. Base de Données

Nous travaillons avec un ensemble de données comprenant plusieurs variables économiques, avec pour start\_date le 01/01/2004, notamment :

Données Macro

* **Taux d'intérêt** : Market Yield on U.S. Treasury Securities at 10-Year
* **M1**
* **M2**
* **Global GDP**
* **Mortgage Rate 30y**
* **Indices de confiance des consommateurs** : Mesures de l'optimisme économique.
* **Chômage** : Taux de chômage US.
* **Inflation** : CPI US
* **Données historiques du S&P 500**

Autre

* **VIX**

Commo

* **Cuivre**
* **Or**
* **Brent**
* **Natural Gas**

Attention : Les données doivent avoir la même plage historique

### 2. Hypothèses

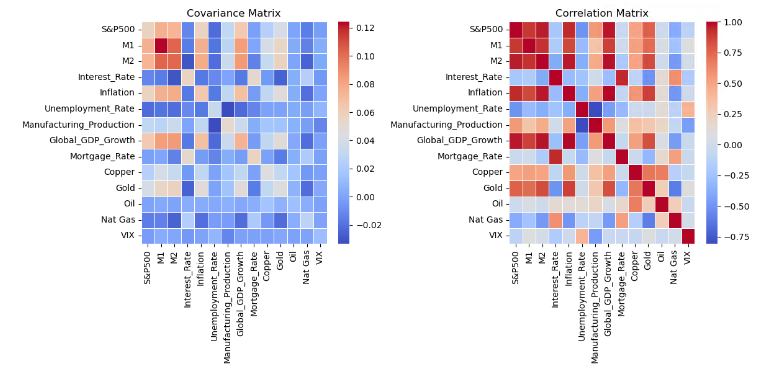
Avant d'analyser les données, nous établissons des hypothèses de travail, telles que :

* Les taux d'intérêt ont un impact inverse sur le S&P 500 ; des taux plus élevés peuvent réduire les investissements des entreprises et donc faire baisser les actions.
* Une augmentation de l'indice de confiance des consommateurs est positivement corrélée avec le S&P 500, car une confiance accrue pourrait stimuler les dépenses des consommateurs et, par conséquent, les bénéfices des entreprises.
* Un taux de chômage en baisse est généralement considéré comme un signe de croissance économique, ce qui peut avoir un effet positif sur le S&P 500.

### 3. Processus et Méthodologie

#### a. Préparation des Données

* **Collecte des Données** : Importation des données économiques et du S&P 500 à partir de sources fiables (<https://fred.stlouisfed.org/series> & <https://finance.yahoo.com/markets/world-indices/>).
* **Nettoyage des Données** : Traitement des valeurs manquantes et élimination des anomalies pour assurer l'intégrité des données.
* **Matrices de Covariances et Corrélation** - figures 1 & 2 :
  + Corrélation : Calcul de la matrice de corrélation entre les variables économiques et le S&P 500. Cela permet d'identifier la force et la direction des relations linéaires entre les variables.
  + Covariance : Calcul de la matrice de covariance pour analyser la variance conjointe entre les variables économiques. Cela permet d'évaluer comment les variations de deux variables interagissent.



* **Transformation des Données** : Normalisation ou standardisation des variables pour faciliter la comparaison (0 à 1 avec 1 = max sur la période) – figure 3.

Une image contenant texte, ligne, Parallèle, Tracé

Description générée automatiquement

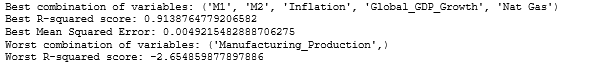
#### b. Analyse des Combinaisons de Variables – Partie uniquement statistique

* **Génération de Combinaisons** : Création de toutes les combinaisons possibles des variables économiques.
* **Évaluation des Modèles** : Pour chaque combinaison, utilisation d’une régression linéaire pour évaluer le modèle :
  + Calcul du **score R²** pour mesurer la proportion de variance expliquée.
  + Calcul de l'**Erreur Quadratique Moyenne (MSE)** pour évaluer l'ajustement du modèle.

Une image contenant texte, reçu, Police, capture d’écran

Description générée automatiquement

* **Best et Worst combinaison** : récupérer la meilleure et la pire combinaison des variables économiques avec comme critère de comparaison le R² et le MSE



Ce qui donne en comparaison – figure 4 :

Une image contenant Tracé, ligne, diagramme, capture d’écran

Description générée automatiquement

Le composit de la figure 4 affichant le R² le plus élevé et l’erreur quadratique la plus faible serait : M1 / M2 / Inflation / Global GDP / Natural Gas

Worst combinaison – figure 5 :

Une image contenant ligne, Tracé, texte, diagramme

Description générée automatiquement

Logique pour la figure 5… étant donné que le worst n’est composé que de la production manuf

#### c. Incorporation des Critères d'Évaluation

* **Critères AIC et BIC** : Ajout de l'AIC (Akaike Information Criterion) et du BIC (Bayesian Information Criterion) pour évaluer la qualité des modèles en prenant en compte la complexité :
  + Un AIC ou BIC plus bas indique un meilleur modèle tout en pénalisant les modèles plus complexes.

AIC : Mesure qui cherche à évaluer la qualité relative d'un modèle statistique en prenant en compte la qualité de l'ajustement (via la somme des résidus #erreur) et le nombre de paramètres du modèle. Il favorise les modèles bien ajustés, mais pénalise ceux qui sont trop complexes (avec trop de paramètres). Plus l'AIC est faible, meilleur est le modèle.

BIC : Similaire à l'AIC, mais il pénalise plus fortement les modèles complexes. Le BIC favorise donc généralement des modèles plus simples par rapport à l'AIC, surtout lorsque le nombre d'observations est élevé. Là encore, un BIC plus faible est préférable.

**Différence entre AIC et BIC :**

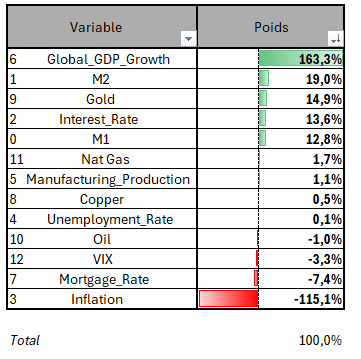
* L'AIC est plus axé sur la recherche d'un modèle qui minimise la perte d'information sans trop se soucier du nombre d'observations.
* Le BIC, en revanche, prend en compte la taille de l'échantillon et pénalise davantage les modèles complexes, surtout lorsque le nombre d'observations est élevé.

En prenant le top 5 des combinaisons on obtient :

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

La première combinaison donne des résultats intéressants (R² élevé et faible marge d’erreur) en plus d’avoir les plus bas AIC/BIC. Ci-dessous la constitution du composite – figure 6 :



#### d. Création d'Indicateurs Avancés – Partie Machine Learning

* **Offset des Variables** : Nous avons appliqué des décalages (lags) de 5 jours à 60 jours sur les variables indépendantes pour créer des indicateurs avancés. Cela permet de prédire un leading indicateur assez proche des variations du S&P 500 en utilisant les valeurs passées des variables.
* **Optimisation des offsets** : Test de différents délais pour déterminer ceux qui maximisent le score R² du modèle.

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

Figure 7

Une image contenant texte, Tracé, ligne, capture d’écran

Description générée automatiquement

Avec la combinaison sélectionnée, nous arrivons à obtenir le meilleur couple R²/MSE pour un offset de 53 jours. Déterminons le poids des éléments contribuant principalement à la variance du S&P.

#### e. Analyse en Composantes Principales (PCA)

* **Application du PCA** : Le PCA réduit la dimensionnalité des données, permettant ainsi de conserver les variables les plus pertinentes tout en éliminant celles qui ajoutent peu d'informations. Nous cherchons à interpréter au moins 90% de la variation de la variance du S&P 500. Le nombre de PC ressort dans notre étude à 5 et nous arrivons à 93.8% - figure 8 :

Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquementUne image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

**Interprétation des Résultats** :

* **PC1 – Tendances Macroéconomiques Globales**
  + Dans PC1, les variables les plus influentes sont *l'inflation*, la *masse monétaire (M2)*, la *croissance du PIB* et *l'or* 🡪 PC1 pourrait donc représenter une "**tendance macroéconomique globale**".
* **PC2 – Facteurs de Taux d'Intérêt et d'Incertitude**
  + PC2 est dominé par les *taux hypothécaires*, les *taux d'intérêt*, le *chômage* et le *VIX* 🡪 PC2 pourrait indiquer des "**conditions de crédit et d'incertitude**". Ce PC semble donc déconnecté des tendances de croissance captées par PC1.
* **PC3 – Dynamique des Matières Premières et de la Production**
  + PC3 montre une influence marquée du *pétrole*, du *cuivre* et de la *production manufacturière* 🡪 PC3 pourrait capter la "**dynamique des matières premières**" et ses effets sur la production industrielle, avec des liens entre les prix des ressources et le secteur manufacturier, une tendance indépendante des précédentes.
* **PC4 – Volatilité des Marchés et Crédit**
  + PC4 est fortement influencé par le *VIX*, les *taux hypothécaires* 🡪 PC4 pourrait capturer un "**indice de volatilité et de crédit**", indépendant des aspects de croissance macroéconomique et des matières premières, montrant comment les changements de liquidité et la volatilité affectent les marchés.
* **PC5 – Effets des Énergies et du Travail**
  + PC5 est dominé par le *gaz naturel*, le *VIX*, les *taux d'intérêt*, et le *chômage*. Il semble capter des effets liés aux prix de l'énergie et à l’emploi, ainsi que la volatilité 🡪 PC5 pourrait donc représenter une "**composante énergie et emploi**", montrant comment les coûts énergétiques et la volatilité interagissent avec le marché du travail.

Figure 9

Une image contenant texte, reçu, Police, blanc

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquement

C//C – Pondération :

L’étude statistique avait relevé que la combinaison optimale sans tenir compte des pondérations et sans avoir le pourcentage exact de l’explication de la variance du S&P était :

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Tandis que notre modèle PCA nous indique que la combinaison optimale serait :

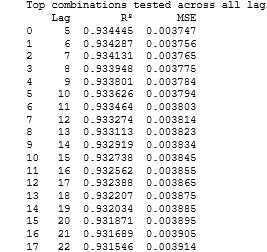
Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

La pondération du composite basé sur le PCA représente une sorte de moyenne pondérée des contributions des variables principales, issues des composantes significatives (celles capturant 95% de la variance). Ces pondérations sont calculées en intégrant les contributions des variables sur chaque composante principale, ajustées pour l'importance relative de chaque composante dans la variance totale du dataset.

* **M1, Global\_GDP\_Growth et inflation** ont des pondérations relativement élevées (15.4%, 13.3%, et 13.3%), probablement parce qu’elles sont significatives dans les composantes principales dominantes (PC1) et représentent des variables macroéconomiques globales, reflétant la tendance de croissance et de liquidité.
* **Copper, Gold, et Manufacturing Production** ont des pondérations intermédiaires (11.4%, 10.2% et 7.4%), indiquant une importance modérée, en ligne avec les variations liées aux matières premières et à la production.
* **Interest Rate et Mortgage Rate** ont des pondérations faibles (2.9% et 6.5%), malgré leur importance dans PC2, car la variance expliquée par PC2 peut être plus faible, diminuant leur impact dans le composite.

Teston la combinaison issue du ML et retrouvons le lag correspondant :



Les résultats sont moins précis que la combinaison issue de l’étude statistique (R² = 97.56% / MSE = 0.14%). Le lag serait de 5 jours pour la combinaison issue du PCA (contre 53 pour la combinaison statistique) – figure 10 :

Une image contenant texte, Tracé, ligne, capture d’écran

Description générée automatiquement

Avec un zoom sur 4 ans – figure 11 & 12 :

Une image contenant texte, ligne, Tracé, capture d’écran

Description générée automatiquement

Une image contenant texte, ligne, Tracé, Police

Description générée automatiquement

Sur l’année 2022, notre composite issu du PCA a été totalement décorrélé du S&P. Logique vu la composition…

En ce qui concerne le composite issu de l’étude statistique, la précision a été plus élevé que celle du PCA – figure 13 & 14

Une image contenant texte, Tracé, ligne, diagramme

Description générée automatiquement

Une image contenant texte, Tracé, ligne, diagramme

Description générée automatiquement

### 4. Algorithme

Schéma simplifié de l'algorithme :

1. **Importer les bibliothèques nécessaires** : Pandas, NumPy, scikit-learn, statsmodels, etc.
2. **Charger et nettoyer les données** (Fred & Yahoo Finance – Pas de données bloomberg)
3. **Statistique**
   * **Pour chaque combinaison de variables** :
     1. Ajuster un modèle de régression linéaire.
     2. Calculer R², MSE, AIC, et BIC.
     3. Enregistrer les résultats.
   * **Pour chaque variable avec décalage** :
     1. Ajuster le modèle et évaluer.
4. **Machine Learning : Application du PCA** :
   * Identifier les composantes principales (au moins 90% de la variance).
   * Calculer R², MSE, AIC, et BIC.
   * Enregistrer les résultats.
   * Pour chaque variable avec décalage :
     1. Ajuster le modèle et évaluer.
5. **Visualiser les résultats** : Tracer les relations entre les meilleures combinaisons de variables et le S&P 500.

### 5. Résultats

En quelques mots, nous avons été en mesure de :

* Identifier les combinaisons de variables qui ont le plus d'impact sur le S&P 500 en construisant deux composites issues d’une approche statistique et d’une approche en machine learning.
* Mis en lumière du composite de variables économiques permettant d’évaluer l’impact de ces dernières sur les variations de la variance du S&P.
* **A faire : Backtest sur quelques années et détermination (en %) de la véracité du modèle par année + Explications si divergences conséquentes**
* **Réaliser l’étude par GICS**
* **Ajouter davantage de données macro & micro (pour l’instant 14)**
* **Peut-être étendre sur l’Europe ?**