# Projet exploitation massive des données

## *Etude du prix de l'action Airbus du 01/01/13 au 01/01/23 et prévision sur 5 jours.*

**SOMRANI Ines**

**VERGOZ Margaux**

**DIALLO Thierno Mamadou**

**Avant-propos**

Au cours du projet, nous avons examiné la performance historique de l'action Airbus en utilisant les données de clôture de ses prix. En utilisant ces données, nous avons entraîné un modèle de régression linéaire qui nous permet de faire des prévisions sur les prix futurs de cette action.

L'étude comporte 3 étapes clés : la récupération des données, le traitement des données et le modèle de prévision.

## Récupération des données

*Lien des données utilisées : https://finance.yahoo.com/quote/AIR.PA/*

Pour récupérer les données on utilise la bibliothèque "Yahoo Finance" pour récupérer l'historique de l'action Airbus en utilisant le "ticker" AIR.PA qui est un identifiant unique pour cette action.

La méthode yf.Ticker est utilisée pour créer un objet pour l'action Airbus en utilisant son ticker. La méthode history est ensuite appelée sur cet objet pour récupérer les données historiques de l'action sur une période de 10 ans, allant du 1er janvier 2013 au 1er janvier 2023.

Les données récupérées sont ensuite stockées dans un data frame appelé df. Le data frame peut être affiché en utilisant le dernier appel df. Les données affichées incluent les informations sur les prix de clôture de l'action Airbus pour chaque jour de la période spécifiée.

On récupère donc un tableau de l'historique des données comportant diverses informations pour chaque jour : (Annexe 1)

- Le prix d'ouverture (Open)
- Le prix le plus élevé atteint par l'action pendant la journée (High)
- Le prix le plus bas atteint par l'action pendant la journée (Low)
- Le prix de clôture (Close)
- Le nombre de transaction dans la journée (Volume)
- Les dividendes si il y en a (Dividende)
- Les fractionnements d'actions si il y en a (Stock splits)

## Traitement des données

La phase de traitement des données est indispensables pour les préparer à l'analyse ou à la modélisation. En général, le traitement des données vise à fournir une base solide pour l'analyse et la modélisation des données afin de générer des résultats fiables et précis.

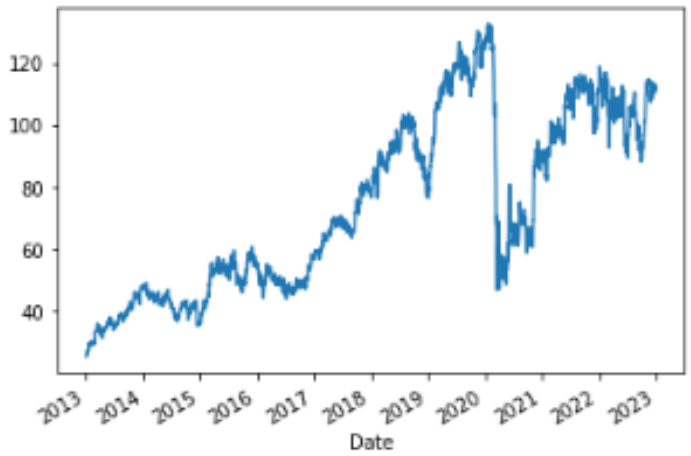
On a d'abord effectuer un nettoyage des données vérifiant et en supprimant la présence de doublons et de valeurs manquante dans le data frame.

Puis nous avons générer, à titre informatif, les statistiques élémentaires des variables étudiés (Annexe 2). La méthode df.describe() est utilisée pour afficher un résumé statistique des

données contenues dans le data frame df. Cette méthode calcule et retourne plusieurs statistiques élémentaires pour chaque variable du data frame, notamment la moyenne, la déviation standard, le minimum, le quartile 1 (Q1), le median (Q2), le quartile 3 (Q3) et le maximum.

Enfin nous avons générer un graphique des données du prix de clôture car c'est la variable que l'on a choisir d'étudier et de prédire (Figure 1) . La méthode de tracé de pandas crée automatiquement un graphique représentant les données de la colonne 'Close' en fonction de la date. La méthode de tracé permet de visualiser rapidement les tendances et les fluctuations du prix de clôture au fil du temps.

Figure 1 : Prix de clôture de l'action Airbus




**Modèle de prévision**

Pour effectuer la prévision des données on a choisi une régression linéaire. La prévision par régression est une technique d'analyse statistique qui permet de prédire une variable dépendante en fonction d'une ou plusieurs variables indépendantes. Cette technique est souvent utilisée pour les prévisions à court terme, comme dans ce cas où nous souhaitons prédire les prix futurs de l'action Airbus.

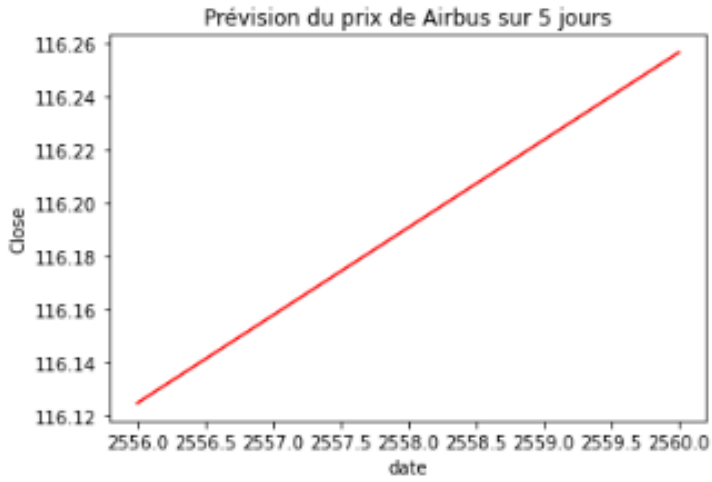
Tout d'abord, le code ajoute une colonne date au data frame df car cette colonne n'était pas présente dans le data frame d'origine. La colonne date est créée en utilisant la méthode range avec comme argument la longueur du data frame df.

Ensuite, nous avons défini les variables indépendantes (X) et dépendantes (y) en sélectionnant respectivement la colonne date et le prix de clôture dans le dataframe. Nous avons converti ces données en tableaux numpy pour pouvoir les utiliser dans le modèle de régression linéaire.

Le modèle de régression linéaire est entraîné en utilisant les données d'entraînement (toutes les observations sauf les 5 dernières) et est ensuite utilisé pour faire des prévisions sur les 5 derniers jours. Le code affiche les prévisions en formatant la date et le prix prévu pour chaque jour. (Annexe 3)

Enfin, le code affiche un graphique représentant les prévisions pour les 5 derniers jours en utilisant la fonction plt.plot(). (Figure 2)

Figure 2 :


Prévision du prix de Airbus sur 5 jours
116.26
116.24
116.22
116.20
116.18
116.16
116.14
116.12
Close
2556.0
2556.5
2557.0
2557.5
2558.0
2558.5
2559.0
2559.5
2560.0
date

# Annexes

Annexe 1 : Historique des données

| Date | Open | High | Low | Close | Volume | Dividends | Stock Splits |
|---|---|---|---|---|---|---|---|
| 2013-01-02 00:00:00+01:00 | 25.370847 | 25.649926 | 25.269363 | 25.425817 | 1526764 | 0.0 | 0.0 |
| 2013-01-03 00:00:00+01:00 | 25.370852 | 25.561133 | 25.286282 | 25.539991 | 802652 | 0.0 | 0.0 |
| 2013-01-04 00:00:00+01:00 | 25.586500 | 26.089689 | 25.455418 | 26.047403 | 1547823 | 0.0 | 0.0 |
| 2013-01-07 00:00:00+01:00 | 26.089689 | 26.225000 | 25.992434 | 26.005119 | 1370626 | 0.0 | 0.0 |
| 2013-01-08 00:00:00+01:00 | 26.140429 | 26.372995 | 26.043175 | 26.153114 | 1911342 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 2022-12-23 00:00:00+01:00 | 111.599998 | 112.260002 | 110.940002 | 111.739998 | 467239 | 0.0 | 0.0 |
| 2022-12-27 00:00:00+01:00 | 112.940002 | 113.680000 | 112.300003 | 112.820000 | 487884 | 0.0 | 0.0 |
| 2022-12-28 00:00:00+01:00 | 113.000000 | 113.099998 | 111.459999 | 111.620003 | 533122 | 0.0 | 0.0 |
| 2022-12-29 00:00:00+01:00 | 111.300003 | 112.019997 | 110.580002 | 112.019997 | 484868 | 0.0 | 0.0 |
| 2022-12-30 00:00:00+01:00 | 111.680000 | 111.760002 | 110.820000 | 111.019997 | 544003 | 0.0 | 0.0 |

Annexe 2 : Statistique élémentaire

| | Open | High | Low | Close | Volume | Dividends | Stock Splits |
|---|---|---|---|---|---|---|---|
| count | 2561.000000 | 2561.000000 | 2561.000000 | 2561.000000 | 2.561000e+03 | 2561.000000 | 2561.0 |
| mean | 74.229588 | 75.156173 | 73.292251 | 74.234305 | 2.081904e+06 | 0.004549 | 0.0 |
| std | 28.721905 | 29.014111 | 28.444007 | 28.726560 | 1.303556e+06 | 0.079757 | 0.0 |
| min | 25.370847 | 25.561133 | 25.269363 | 25.425817 | 0.000000e+00 | 0.000000 | 0.0 |
| 25% | 48.352899 | 48.889336 | 47.598460 | 48.280724 | 1.303346e+06 | 0.000000 | 0.0 |
| 50% | 67.827516 | 68.609932 | 66.869223 | 67.850685 | 1.800277e+06 | 0.000000 | 0.0 |
| 75% | 101.031455 | 101.918727 | 99.556314 | 100.952583 | 2.482765e+06 | 0.000000 | 0.0 |
| max | 132.196170 | 133.074425 | 130.916977 | 132.692581 | 2.689999e+07 | 1.800000 | 0.0 |

Annexe 3 : Prévision des 5 prochains jours

```
5 prix prévus :
Date : 2023-01-02 00:00:00 Prix prévu : 116.13
Date : 2023-01-03 00:00:00 Prix prévu : 116.16
Date : 2023-01-04 00:00:00 Prix prévu : 116.19
Date : 2023-01-05 00:00:00 Prix prévu : 116.22
Date : 2023-01-06 00:00:00 Prix prévu : 116.26
```