



## ETUDE SUR LES ACTIONS

QUELS FACTEURS PEUVENT INFLUENCER LE PRIX  
D'UNE ACTION BOURSIERE ?

---

Jordan NAMOR – Développeur Data 2021

## TABLE DES MATIERES

<b>1</b>	<b>INTRODUCTION.....</b>	<b>3</b>
1.1.	PRESENTATION DU PROJET.....	3
1.1.1.	DEMOCRATISATION DE LA BOURSE .....	3
1.1.2.	LES ACTIONS.....	3
1.1.3.	LES FACTEURS LIES A L'EVOLUTION DE LA SOCIETE COTEE .....	4
1.1.4.	LES FACTEURS LIES AU SECTEUR DE LA SOCIETE.....	4
1.1.5.	LES FACTEURS LIES A L'ENVIRONNEMENT ECONOMIQUE.....	4
1.2.	CAHIER DES CHARGES .....	5
1.2.1.	LE CONTEXTE .....	5
1.2.2.	LES OBJECTIFS .....	5
1.2.3.	LANGAGES ET OUTILS UTILISES.....	5
1.2.4.	GESTION DU PROJET .....	7
<b>2</b>	<b>CONCEPTION DE LA BASE DE DONNEES.....</b>	<b>8</b>
2.1.	LES ETAPES DE LA PREPARATION DE LA DONNEE.....	8
2.1.1.	COLLECTE DE LA DONNEE .....	8
2.1.2.	DESCRIPTION DES DONNEES.....	12
2.1.3.	NETTOYER ET VALIDER LES DONNEES .....	15
2.2.	MODELES DE DONNEES .....	16
2.2.1.	CREER LA BASE DE DONNEES .....	17
2.2.2.	STOCKER LES DONNEES DANS MYSQL.....	18
2.2.3.	SAUVEGARDE JOURNALIERE DES DONNEES.....	19
2.2.4.	SECURISATION DE LA BASE DE DONNEES .....	20
<b>3</b>	<b>AXES D'ANALYSES.....</b>	<b>20</b>
3.1.	REQUETES.....	20
3.2.	DEVELOPPEMENT DU SITE WEB .....	23
<b>4</b>	<b>BILAN DE PROJET.....</b>	<b>25</b>
4.1.	OBSERVATIONS .....	25
4.2.	CONCLUSIONS .....	25
4.3.	LES DIFFICULTES .....	26
4.4.	LES AXES D'AMELIORATIONS.....	26
4.5.	REMERCIEMENTS.....	27

## 1. INTRODUCTION

### 1.1.1. PRESENTATION DU PROJET

Le mot « bourse » désigne le lieu public où s'assemblent, à certaines heures, les négociants, les banquiers, les agents de change, les courtiers, pour traiter d'affaires. Son origine vient du nom du lieu où les échanges de créances et de titres divers entre banquiers s'effectuaient au XIV<sup>ème</sup> siècle.

Les plus grandes places boursières à travers le monde concentrent la majorité des flux financiers et des transactions internationales. Certains marchés dictent le rythme de la finance mondiale, en plus d'être les phares économiques des pays où ils siègent.

### 1.1.2. DEMOCRATISATION DE LA BOURSE

Traditionnellement, les bourses étaient des bâtiments physiques dans chaque pays où elles étaient en activité, mais depuis le passage à la négociation électronique, beaucoup ont dû fermer leurs salles de marché pour passer aux plateformes en ligne. Cependant, les institutions elles-mêmes existent toujours. Elles servent à évaluer la santé économique d'un pays. En somme, elles sont les indicateurs de la puissance économique mondiale.

### 1.1.3. LES ACTIONS

En bourse, plus de 6 milliards d'actions s'échangent chaque jour, ce qui correspond à plus de 5000 milliards de dollars en moyenne.

Le cours d'une action est principalement influencé par l'appréhension que les investisseurs ont de la société qui l'émet. Parmi les principaux facteurs qui vont influencer le cours d'une action en bourse, on retrouve différents facteurs :

- Les facteurs propres à l'évolution de l'entreprise
- Les facteurs liés au secteur de la société
- Les facteurs liés à l'environnement dans lequel évolue l'entreprise

Pour nous permettre de comprendre ces influences et de prévenir un possible investissement boursier, le projet consistera à étudier ces principaux facteurs et d'établir des comparaisons entre elles pour peser le poids de leur impact sur les marchés.

#### 1.1.4. LES FACTEURS LIES A L'EVOLUTION DE LA SOCIETE COTEE

Les attentes sur les bénéfices de l'entreprise représentent un critère très important pour les investisseurs. En effet, ce critère peut être l'objet d'anticipation par les analystes sur les futurs bénéfices (ou pertes) de la société.

Les actifs détenus par l'entreprise pourront également la valoriser sur les marchés. On peut citer ici, à titre d'exemples, la détention d'un brevet, d'un patrimoine immobilier ou d'une exclusivité de vente. L'innovation est également un des facteurs clés de la réussite d'une entreprise.

#### 1.1.5. LES FACTEURS LIES AU SECTEUR DE LA SOCIETE

Le secteur dans lequel évolue une entreprise est aussi un facteur qui pourra modifier son cours. Une entreprise saine dans un secteur en difficulté subira l'impact négatif de l'image que son secteur a dans le marché. On pense ici aux valeurs financières qui ont perdu beaucoup de crédibilité sur les marchés après la crise de 2008.

Un des facteurs pouvant aussi valoriser une action en bourse est un changement radical qui affecte le secteur d'activité. L'apparition d'un nouveau produit ou de nouvelles habitudes de consommation sont des facteurs qui peuvent déterminer l'évolution de la valeur d'une action.

Une autre opportunité est la consolidation d'un secteur. Le rachat de sociétés du même secteur permet à l'entreprise de se constituer comme leader d'un marché et d'y occuper une situation dominante.

#### 1.1.6. LES FACTEURS LIES A L'ENVIRONNEMENT ECONOMIQUE

Les grandes tendances économiques et politiques peuvent aussi susciter des craintes ou des apaisements sur les marchés boursiers entraînant tout un marché à la hausse ou à la baisse en raison de fondamentaux macroéconomiques favorables ou inquiétants. Les mesures fiscales, budgétaires et surtout monétaires d'un pays peuvent tantôt rassurer ou inquiéter les investisseurs en bourse.

Les habitudes de consommation peuvent aussi être un élément à analyser dans l'évolution potentielle du cours d'une action. Certaines opportunités sont cycliques. Certains secteurs qui ont dû effectuer des restructurations peuvent sortir d'un cycle de décroissance en étant davantage consolidés. D'autres opportunités sont plus structurelles. C'est le cas, par exemple, des sociétés actives dans l'éducation. Alors que les pays émergents passent d'une civilisation essentiellement agraire à une transition vers une économie industrielle et de services, il existe

un besoin structurel d'amélioration du système éducatif. Ces sociétés peuvent alors bénéficier d'une demande en bourse plus importante.

## 1.2. CAHIER DES CHARGES

### 1.2.1. LE CONTEXTE

- Enjeux : avoir une meilleure compréhension sur l'influence du cours de l'action d'une entreprise
- Problématique : quels sont les facteurs qui peuvent influencer le prix d'une action ?
- Stratégie :
  - Déterminer s'il y a des corrélations entre les différents facteurs financiers
  - Analyser si le secteur d'activité a une influence sur une action boursière
  - Observer les différents résultats financiers
- Domaine d'application : les actions américaines, française et allemandes

### 1.2.2. LES OBJECTIFS

Observer les différents facteurs disponibles qui pourraient exercer une influence sur le prix d'une action en bourse :

- Résultats financiers : chiffre d'affaires, capitalisation boursière, ...
- Secteurs d'activités : technologie, santé, industrie, ...
- Prix d'une action : comparaison entre les différentes entreprises

### 1.2.3. LANGAGES ET OUTILS UTILISES

Outils	Utilisation
Python	Langage orienté objet pour traiter la Data et développer des applications

SQL	Langage informatique servant à exploiter des bases de données relationnelles
HTML	Langage de balisage conçu pour représenter les pages web
CSS	Langage informatique qui décrit la présentation des documents HTML
Jupyter Notebook	Application web pour exécuter du code : sert aux chargement, exploration, traitement et nettoyage des données
Visual Studio Code	IDE pour la création des scripts Python
MySQL	SGBDR : Système de Gestion de Bases de Données Relationnelles, qui stocke des données de façon organisée et cohérente
MySQL Workbench 8.0	IDE SQL pour la conception de la base de donnée, la visualisation et l'administration de la base de données
Amazon RDS	Service Amazon qui permet de créer et de gérer une base de données
Plotly	Visualisation des données (Data Visualisation)
Dash	Framework open-source de développement web en Python, permet de visualiser les données via une application web
Trello	Gestion de projet, permet de visualiser les tâches selon la méthode Agile

#### 1.2.4. GESTION DU PROJET

Un projet peut se diviser en plusieurs étapes selon une hiérarchie qui semble bien précise, et c'est durant le développement de celui-ci que le développeur souhaiterait revenir à une étape précédente pour modifier ou ajouter une fonctionnalité. Cette action pourrait lui engendrer une perte de temps considérable si celle-ci est répétée.

Dans le cadre de notre formation, nous avons passé la certification aux méthodes agiles. Celle-ci nous permet de travailler en adoptant une organisation qui vise l'efficacité maximale. L'idée avec les méthodes apprises, c'est de pouvoir travailler tout en gagnant un maximum de temps, de prendre en compte l'erreur humaine en nous adaptant si le travail est retardé.

Mon organisation personnelle s'inspire de la méthode appelée Kanban, qui permet de prendre chacune des tâches et de la faire naviguer d'une colonne à une autre selon sa progression.

Dans l'image ci-dessous, je prends chacune des tâches nécessaires à l'obtention de la certification puis la divise en sous-tâche pour me donner une idée du travail à effectuer. Une fois que la tâche est faite, celle-ci devient verte :



## 2. CONCEPTION DE LA BASE DE DONNEES

### 2.1. LES ETAPES DE LA PREPARATION DE LA DONNEE

#### 2.1.1. COLLECTE DE LA DONNEE

Les sources des données pour constituer la base de données provient de Wikipédia et Google Finance



#### WIKIPEDIA – DOW JONES / CAC40 / DAX

- Format : HTML
- Contenu : Le Dow Jones, le CAC40 ainsi que le DAX étant des indices économiques qui regroupent les entreprises dans leur pays d'origine, Wikipédia nous met à disposition plusieurs informations tels que le chiffre d'affaires, le revenu net après impôt, sa capitalisation boursière ou le nombre d'employé par entreprise. Les résultats les plus récents correspondent à l'année de 2020, l'année de 2021 n'étant pas encore terminée.
- Détail des colonnes : 11 colonnes
- Lien :
  - [https://en.wikipedia.org/wiki/Dow\\_Jones\\_Industrial\\_Average](https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average)
  - [https://en.wikipedia.org/wiki/CAC\\_40](https://en.wikipedia.org/wiki/CAC_40)
  - <https://fr.wikipedia.org/wiki/DAX>

Afin de débiter le Web scraping, la récolte des données s'effectuera grâce à Requests, un module Python permettant de récupérer les données d'une page web. Pour donner un exemple, les données de base se présentent de cette façon :

DJIA component companies, showing trading exchange, <a href="#">ticker symbols</a> and industry						
Company	Exchange	Symbol	Industry	Date added	Notes	Index weighting
<a href="#">3M</a>	NYSE	<a href="#">MMM</a>	<a href="#">Conglomerate</a>	1976-08-09	As Minnesota Mining and Manufacturing	3.73%
<a href="#">American Express</a>	NYSE	<a href="#">AXP</a>	<a href="#">Financial services</a>	1982-08-30		3.20%
<a href="#">Amgen</a>	NASDAQ	<a href="#">AMGN</a>	<a href="#">Pharmaceutical industry</a>	2020-08-31		4.52%
<a href="#">Apple Inc.</a>	NASDAQ	<a href="#">AAPL</a>	<a href="#">Information technology</a>	2015-03-19		2.75%
<a href="#">Boeing</a>	NYSE	<a href="#">BA</a>	<a href="#">Aerospace and defense</a>	1987-03-12		4.26%



Celles-ci sont ensuite récupérées sur Jupyter Notebook :

```
import requests
from bs4 import BeautifulSoup
import time
import pandas as pd
import numpy as np
import re
pd.set_option('display.html.use_mathjax', False)
pd.options.display.float_format = '{:.2f}'.format

url = 'https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average'
response = requests.get(url)
soup = BeautifulSoup(response.content)
liens = soup.find('table', attrs = {'id': 'constituents'})

liste = ['Company', 'Symbol', 'Industry']
df = pd.concat(pd.read_html(str(liens)))[liste].head(5)
df
```

	Company	Symbol	Industry
0	3M	MMM	Conglomerate
1	American Express	AXP	Financial services
2	Amgen	AMGN	Pharmaceutical industry
3	Apple Inc.	AAPL	Information technology
4	Boeing	BA	Aerospace and defense

Comme on le voit ci-dessus, nous pouvons aussi récupérer les liens qui mènent aux entreprises pour avoir des informations encore plus détaillées :


```
liens = soup.find_all('tr')

dict1 = {}
for tr in liens:
    try:
        a = tr.find('a')
        link = a['href']
        cond = link.split('/')[2].replace('_', ' ').replace('%26', '&').replace('%27', "'")
        if cond in list(df['Company']):
            dict1[cond] = 'https://en.wikipedia.org'+link
    except:
        pass
dict1

{'3M': 'https://en.wikipedia.org/wiki/3M',
 'American Express': 'https://en.wikipedia.org/wiki/American_Express',
 'Amgen': 'https://en.wikipedia.org/wiki/Amgen',
 'Apple Inc.': 'https://en.wikipedia.org/wiki/Apple_Inc.',
 'Boeing': 'https://en.wikipedia.org/wiki/Boeing'}
```

La transformation des informations détaillées se fait ainsi dans un dataframe, puis le processus de collecte sera répété pour les entreprises françaises et allemandes :

**3M Company**



3M headquarters in Maplewood, Minnesota

**Formerly** Minnesota Mining and Manufacturing Company (1902–2002)

**Type** Public

**Traded as** NYSE: MMM  DJIA Component S&P 500 Component

**Industry** Conglomerate

**Founded** June 13, 1902; 119 years ago (as Minnesota Mining and Manufacturing Company) Two Harbors, Minnesota, U.S. <sup>[1]</sup>

**Founders** Dr. J. Danley Budd Henry S. Bryan William A. McGonagle John Dwan Hermon W. Cable <sup>[2]</sup>

**Headquarters** Maplewood, Minnesota, U.S.

**Area served** Worldwide

**Key people** Mike Roman (Chairman, President, & CEO) <sup>[3]</sup>

**Revenue**  US\$32.18 billion (2020) <sup>[4]</sup>

**Operating income**  US\$7.161 billion (2020) <sup>[4]</sup>

**Net income**  US\$5.38 billion (2020) <sup>[4]</sup>

**Total assets**  US\$47.3 billion (2020) <sup>[4]</sup>

**Total equity**  US\$12.931 billion (2020) <sup>[4]</sup>

**Number of employees**  94,987 (2020) <sup>[4]</sup>

**Website** [www.3m.com](http://www.3m.com) 

```
df2 = pd.DataFrame(dict1.items(), columns=['Company', 'Link'])

for i in df2['Link']:
    links = []
    response = requests.get(i)

    soup = BeautifulSoup(response.content)
    description = soup.find('table', attrs = {'class': 'infobox vcard'})
    df_data = pd.concat(pd.read_html(str(description)))

    name = i.split('/')[0].replace('_', ' ').replace('%26', '&').replace('%27', '')

    list_of_numbers = ["Revenue", "Operating income", "Net income", "Total assets", "Total equity", "Number of employees"]
    df_data = df_data.pivot_table(columns=0, aggfunc='first')[list_of_numbers]
    df_data['Company'] = name

    if 'Revenue' in df2.columns:
        df_data = df_data.set_index('Company')
        df2.update(df_data)
    else:
        df2 = df2.merge(df_data, how='left', on='Company').set_index('Company')

df2 = df2.reset_index()

df_merge = df.merge(df2, how='inner', on='Company')
df_merge
```

	Company	Symbol	Industry	Link	Revenue	Operating income	Net income
0	3M	MMM	Conglomerate	<a href="https://en.wikipedia.org/wiki/3M">https://en.wikipedia.org/wiki/3M</a>	US\$32.18 billion (2020) <sup>[4]</sup>	US\$7.161 billion (2020) <sup>[4]</sup>	US\$5.38 billion (2020) <sup>[4]</sup>
1	American Express	AXP	Financial services	<a href="https://en.wikipedia.org/wiki/American_Express">https://en.wikipedia.org/wiki/American_Express</a>	US\$36.09 billion (2020) <sup>[1]</sup>	US\$4.3 billion (2020) <sup>[1]</sup>	US\$3.14 billion (2020) <sup>[1]</sup>
2	Amgen	AMGN	Pharmaceutical industry	<a href="https://en.wikipedia.org/wiki/Amgen">https://en.wikipedia.org/wiki/Amgen</a>	US\$25.424 billion (2020)	US\$9.674 billion (2019)	US\$7.842 billion (2019)
3	Apple Inc.	AAPL	Information technology	<a href="https://en.wikipedia.org/wiki/Apple_Inc.">https://en.wikipedia.org/wiki/Apple_Inc.</a>	US\$274.515 billion <sup>[4]</sup> (2020)	US\$66.288 billion <sup>[4]</sup> (2020)	US\$57.411 billion <sup>[4]</sup> (2020)
4	Boeing	BA	Aerospace and defense	<a href="https://en.wikipedia.org/wiki/Boeing">https://en.wikipedia.org/wiki/Boeing</a>	US\$ 58.16 billion <sup>[4]</sup> (2020)	US\$ -12.76 billion <sup>[4]</sup> (2020)	US\$ -11.94 billion <sup>[4]</sup> (2020)

## GOOGLE FINANCE

- Format : CSV
- Contenu : Google nous met à disposition des formules qui nous permettent d'étudier le cours de l'actif souhaité. On peut ainsi récupérer les données historiques jusqu'à aujourd'hui.
- Détail des colonnes : 7 colonnes

La récupération des données historiques n'est possible qu'avec un fichier Sheets, et nous avons besoin de Python pour automatiser la récolte. Nous utilisons donc l'API que Google nous met à disposition pour nous permettre d'interagir avec celui-ci :

```
import pandas as pd
import time
import gspread
from google.oauth2.service_account import Credentials
from gspread_pandas import Spread, Client

scope = ['https://spreadsheets.google.com/feeds',
         'https://www.googleapis.com/auth/drive']

# Authentification aux API Google pour utiliser Les fichiers Sheets
credentials = Credentials.from_service_account_file('stocks_sheet.json', scopes=scope)
client = Client(scope=scope, creds=credentials)
spread_2 = Spread("stocks_2", client=client)
spread_miss = Spread("stocks_analysis_miss", client=client)
```

Nous demandons ensuite au fichier de récupérer les données de tous les actifs que nous avons récoltés sur Wikipédia. Si l'origine d'une entreprise est différente, le code s'adapte :



The screenshot shows a Google Sheets spreadsheet titled 'stocks\_2' with columns A, B, and C. The data in the spreadsheet is as follows:

	A	B	C
1	Date	Open	High
2	02/01/2018 16:00	235.78	237.07
3	03/01/2018 16:00	235.07	235.73
4	04/01/2018 16:00	237	239.44
5	05/01/2018 16:00	238.65	240.9
6	06/01/2018 16:00	239.38	240.94
7	07/01/2018 16:00	239.6	241.78
8	08/01/2018 16:00	241	242.57
9	09/01/2018 16:00	240.74	242.34
10	10/01/2018 16:00	243.07	246
11	11/01/2018 16:00	245.3	247.19
12	12/01/2018 16:00	246.85	248.53
13	13/01/2018 16:00	248.13	249
14	14/01/2018 16:00	246.63	248.5
15	15/01/2018 16:00	247.13	247.89
16	16/01/2018 16:00	246.91	247.26
17	17/01/2018 16:00	247.99	248.54
18	18/01/2018 16:00	250	254.8
19	19/01/2018 16:00	253.43	256.77
20	20/01/2018 16:00	258.51	259.34
21	21/01/2018 16:00	255.6	255.69
22	22/01/2018 16:00	251.49	253.13
23	23/01/2018 16:00	247.44	250.09
24	24/01/2018 16:00	246.43	248.75
25	25/01/2018 16:00	243.5	244.4
26	26/01/2018 16:00	227.51	234.57
27	27/01/2018 16:00	233.19	237.52
28	28/01/2018 16:00	233.16	233.31
29	29/01/2018 16:00	224.61	226.96
30	30/01/2018 16:00	227.49	230.38

The Python script on the right interacts with the spreadsheet using the gspread library. It defines a DataFrame for the stock data and uses the gspread\_pandas library to interact with the spreadsheet. The script includes a loop to append data to the DataFrame and update the spreadsheet, and a try-except block to handle errors.

## INVESTISSEURS

- Format : CSV
- Contenu : Les investisseurs jouent un rôle important concernant le prix d'une action. Ils prennent en compte différents facteurs concernant des analyses qui ont pour but d'être investies dans les marchés financiers, ils sont donc à l'origine de la demande. Étant impossible de récupérer les données de tous les investisseurs au monde, nous avons simulé les données à l'aide de Python pour nous permettre de mieux appréhender leur fonctionnement.
- Détail des colonnes : 7 colonnes

Nous créons deux listes qui servent à simuler l'investissement sur plusieurs actifs financiers entre le 1<sup>er</sup> Janvier 2018 et le 23 Juillet 2021. Le marché étant fermé le week-end, le code prendra ce critère en compte :

```
liste = []
for i in range(4000):
    liste.append(random.choice(df_company.id_company))
len(liste)

4000

start_date = datetime.date(2018, 1, 1)
end_date = datetime.date(2021, 7, 23)

time_between_dates = end_date - start_date
days_between_dates = time_between_dates.days

random.seed(a=None)

liste_1 = []
weekdays = [5,6]
while len(liste_1) != 4000:
    random_number_of_days = random.randrange(days_between_dates)
    random_date = start_date + datetime.timedelta(days=random_number_of_days)
    if random_date.weekday() not in weekdays:
        liste_1.append(random_date)

liste_1.sort(reverse=True)
len(liste_1)

4000
```

Puis le tout est regroupé de sorte à pouvoir créer la table d'investissement :

```
df_order = pd.DataFrame()
df_order['id_order'] = np.random.randint(78252128, 168266028, 4000)
df_order['id_type'] = np.random.randint(1, 4, 4000)
df_order['id_investor'] = np.random.randint(1, 10, 4000)
df_order['id_company'] = liste
df_order['times'] = liste_1
df_order['amount'] = np.random.randint(-1500, 1500, 4000).astype(np.int32)
df_order
```

	id_order	id_type	id_investor	id_company	times	amount
0	80605340	3	5	FRE	2021-07-22	1362
1	134328631	3	8	ENGI	2021-07-22	927
2	81278896	1	1	EN	2021-07-22	-113
3	94695675	2	3	V	2021-07-22	-253
4	106698236	1	5	HER	2021-07-21	-42

## 2.1.2. DESCRIPTION DES DONNEES

Tableau 1 - Company

Nom de la colonne	Datatype / Format	Description
id_company	object	Identifiant de l'entreprise

company	object	Nom de l'entreprise
revenue	object	Chiffre d'affaires (2020)
operating_income	object	Bénéfice avant intérêts et impôts
net_income	object	Bénéfice après intérêts et impôts
total_assets	object	Toutes les choses que l'entreprise possède et qui peuvent être converties en espèces
total_equity	object	Montant total investi dans une entreprise par les investisseurs
employees	object	Nombre d'employé
market_cap	object	Capitalisation boursière
sector	object	Secteur d'activité
area	object	Pays d'origine

Tableau 2 - Orders

Nom de la colonne	Datatype / Format	Description
id_orders	object	Identifiant de l'ordre
times	Date	Date de l'ordre effectué

id_investor	object	Identifiant de l'investisseur
last_name	object	Nom
first_name	object	Prénom
title_courtesy	object	Titre de courtoisie
type	object	Type d'investissement (dépôt, retrait, Profit/Perte)
amount	INT	Profit/Perte
company	object	Nom de l'entreprise investie
area	object	Pays d'origine de l'investisseur

La création des tables se base sur des critères de redondance et de dépendance. Toutes valeurs qui se répètent dans un jeu de données doit trouver sa propre table pour des questions d'optimisation et d'organisation. Dans cet exemple, le tableau Company nous dégage 3 classes d'entités qui doivent figurer dans la base de données :

- **La notion géographique** : le pays est identifiable par son nom, ce qui nous permet de remplacer sa valeur par un code unique, et de prendre moins de place dans la base de données
- **Les valeurs financières** : les résultats de l'entreprise sont détaillés
- **Le secteur d'activité** : le secteur étant de type object et indépendant, celui-ci est remplacé par un code numérique et identifiable

Le tableau Orders nous indique également 5 classes d'entités dont deux ayant déjà été identifiées (géographique et financière) :

- **Les ordres** : les informations relatives aux positions effectuées sont uniquement gardées, comme par exemple la somme d'argent que l'investisseur a gagnée

- **Le type d'investissement** : plusieurs types d'ordres sont identifiables ce qui nous pousse à créer une table supplémentaire
- **L'investisseur** : les informations personnelles de l'investisseur sont réunies dans une seule table et seul le code investisseur nous permet d'identifier qui a été à l'origine d'une position

Il ne reste plus qu'une dernière notion non précisée dans les tableaux précédents :

- **La notion de bourse** : regroupe les données historiques concernant les actifs boursiers liés aux entreprises

### 2.1.3. NETTOYER ET VALIDER LES DONNEES

L'étape de nettoyage s'effectue avant l'insertion des données. Cette opération est très importante car elle nous permettra d'importer de la donnée propre, utilisable et exploitable.

Les différentes étapes de nettoyage qui ont été effectuées :

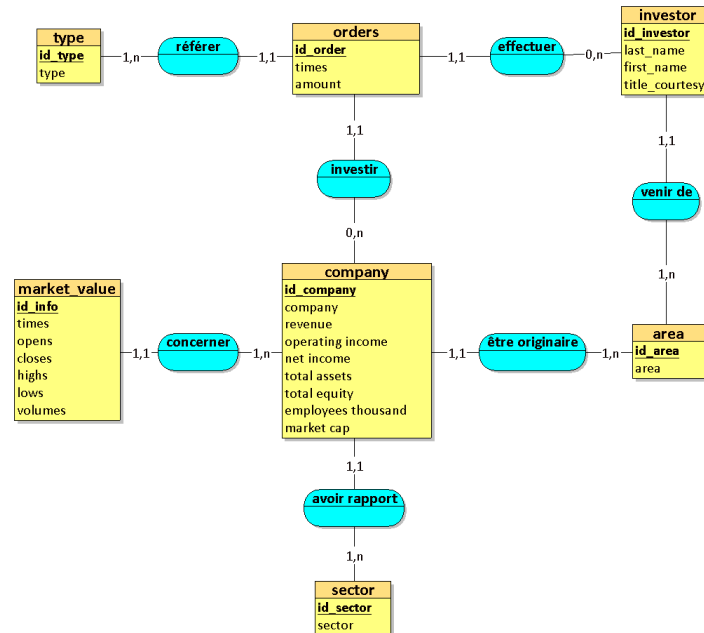
- Supprimer les données redondantes et superflues (NaN, #N/A)
- Ajouter les valeurs manquantes : il peut arriver que des valeurs n'aient pas été récupérées
  - Identifier celles-ci
  - Répéter le processus de récolte
- Adapter les données à une structure standard :
  - Harmoniser la nature des valeurs (milliard, million)
  - Normalisation les types de colonne
  - Exemple avec \$1,234.56 billion et \$1,234.56 million. La fonction replace() permet de prendre en compte ce critère en retirant le point, le symbole du dollars et en multipliant la valeur pour l'adapter en milliard.
- Réorganiser les colonnes
- Masquer les données privées ou sensibles
- Sauvegarder une copie du travail effectué (checkpoint) puis exporter les données en CSV

Il peut arriver qu'une fois les données nettoyées, des erreurs apparaissent soit durant l'insertion de celles-ci, soit durant les analyses quand on s'aperçoit que les données sont aberrantes et non réalistes. Il devient alors nécessaire de les corriger.

On obtient à la fin 7 fichiers de données prêts à être injectés dans les tables.

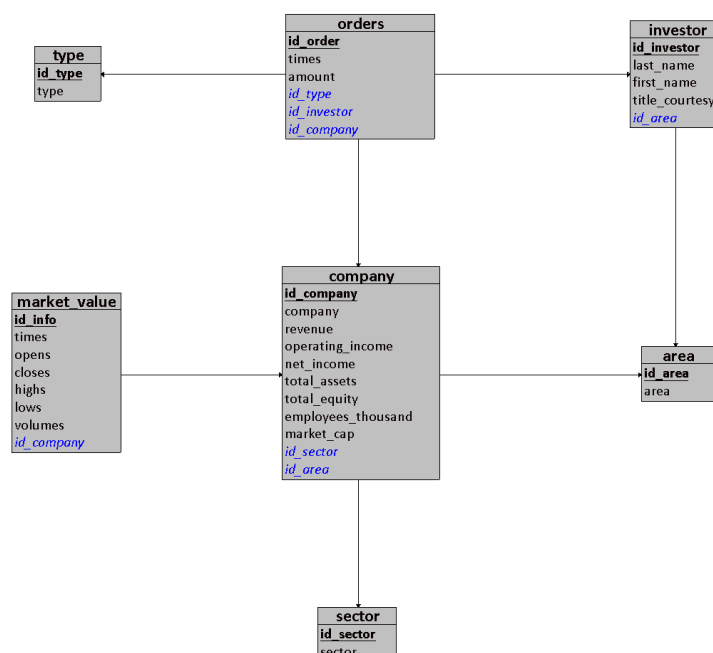
## 2.2. MODELES DE DONNEES

À partir des différents fichiers et de l'objectif général du projet, le modèle conceptuel suivant est établi :



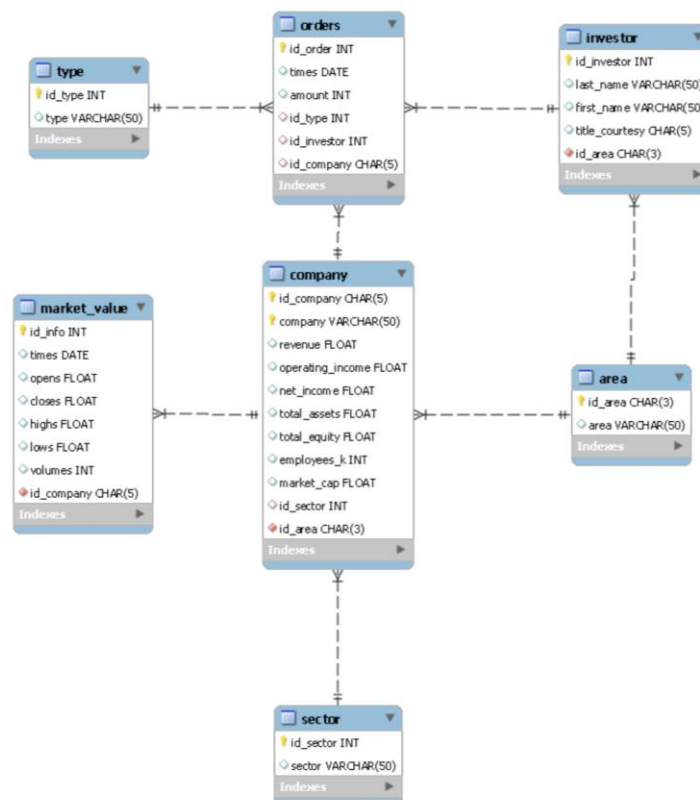
• **Figure 1** : modèle conceptuel de données

Suite à cela, le modèle relationnel de données ci-dessous peut être établi. Comme nous avons uniquement des relations (one to many), il n'y pas eu de création de nouvelles tables :



• **Figure 2** : modèle relationnel de données





• **Figure 3** : modèle physique de données

## CHOIX DU MODÈLE RELATIONNEL

- L'intégrité des données est primordial
- Les données à stocker sont des données structurées
- Le traitement des données est ordonné et non redondant
- Les informations relatives aux tables sont organisées

### 2.2.1. CREER LA BASE DE DONNEES

Lorsque le nettoyage et la préparation des données sont terminées, celles-ci peuvent être stockées vers une application tierce. Pour cela, il est nécessaire de créer une base de données en utilisant MySQL et RDS.

MySQL est système qui peut stocker et gérer les données. Celui-ci opte pour une approche appelée base de données relationnelle. Les données sont ainsi divisées en plusieurs zones de stockage séparées – appelées tables – plutôt que de tout regrouper dans une seule grande unité de stockage.

RDS quant à lui est un service Amazon qui nous permet également de créer et de gérer notre base de données, mais sa particularité repose aussi sur l'automatisation des tâches qui pourraient être chronophages. Les services Amazon sont très utiles pour nous permettre de :

- Optimiser la mémoire et les performances de notre base de données
- Automatiser la sauvegarde journalière des données
- Accroître la sécurité en chiffrant nos données
- Avoir le choix entre plusieurs moteurs de base de données comme PostgreSQL ou encore MySQL

MySQL sera donc inclus dans RDS pour la création et la gestion de notre base de données, puis nous nous connecterons à celle-ci pour y insérer nos fichiers CSV :

```
# Connection à l'instance de la database
def connect_to_instance_db():
    try:
        conn = pymysql.connect(
            host='db-company.cwm601wtiqo3.eu-west-3.rds.amazonaws.com',
            user='admin',
            password=ACCESS_DB,
            charset='utf8mb4')
    except pymysql.err.OperationalError as e:
        raise e
    else:
        print("Connection Successful!")
    return conn

connection = connect_to_instance_db()
cursor = connection.cursor()

# Création de la database
def create_db(cursor):
    queries = (
        "DROP DATABASE IF EXISTS db_company",
        "CREATE DATABASE db_company",
        "USE db_company"
    )

    for query in queries:
        cursor.execute(query)
        connection.commit()
```

## 2.2.2. STOCKER LES DONNEES DANS MYSQL

Deux choix sont possibles :

- Importation des données depuis Jupiter Notebook (sous le format de dataframe) à MySQL grâce à la librairie pymysql
- Importation des données directement depuis MySQL en insérant les fichiers CSV nettoyés au préalable

```

# Importation des données
df_type = pd.read_csv('data/type.csv')
df_sector = pd.read_csv('data/sector.csv')
df_area = pd.read_csv('data/area.csv')
df_investor = pd.read_csv('data/investor.csv')
df_company = pd.read_csv('data/company.csv')
df_company['employees_k'] = (df_company['employees_k'] * 0.001)
df_market_value_short_2 = pd.read_csv('data/market_value_short_2.csv')
df_orders = pd.read_csv('data/orders.csv')

# Insertion du DataFrame ligne par ligne
def insert_into_table(data, table):
    # Création d'une liste de colonnes pour l'insertion
    cols = ','.join(str(x) for x in data.columns)

    for i, row in data.iterrows():
        query = f"INSERT IGNORE INTO {table}({cols}) VALUES(" + "%s,"*(len(row)-1) + "%s)"
        cursor.execute(query, tuple(row))
        print(i)

    connection.commit()

insert_into_table(df_type, 'type')
insert_into_table(df_sector, 'sector')
insert_into_table(df_area, 'area')
insert_into_table(df_investor, 'investor')
insert_into_table(df_company, 'company')
insert_into_table(df_orders, 'orders')
insert_into_table(df_market_value_short_2, 'market_value')

```

### 2.2.3. SAUVEGARDE JOURNALIERE DES DONNEES

Il peut arriver que la suppression accidentelle de nos données représente une perte de temps significative. Comme dit précédemment, Amazon automatise la sauvegarde de nos données de façon journalière, ce qui nous permet de travailler avec plus de sûreté.

Name	DB source	Creation time	Status	Progress	VPC
rdscdb-company-2021-08-11-10-02	db-company	Wed Aug 11 2021 11:02:31 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-12-10-02	db-company	Thu Aug 12 2021 11:02:55 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-13-10-02	db-company	Fri Aug 13 2021 11:02:42 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-14-10-02	db-company	Sat Aug 14 2021 11:02:31 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-15-10-02	db-company	Sun Aug 15 2021 11:02:32 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-16-10-02	db-company	Mon Aug 16 2021 11:02:47 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-17-10-02	db-company	Tue Aug 17 2021 11:02:19 GMT+0100	available	Completed	vpc-d6d31dbe
rdscdb-company-2021-08-18-10-02	db-company	Wed Aug 18 2021 11:02:53 GMT+0100	available	Completed	vpc-d6d31dbe

## 2.2.4. SECURISATION DE LA BASE DE DONNEES

La donnée représente une source d'information très importante, surtout si elle est privée et réservée uniquement aux membres d'une entreprise. Amazon nous permet de renforcer la sécurité de nos données en les chiffrant, et en limitant l'intrusion de notre base données uniquement à celles et ceux qui possèderaient la même adresse IP. Pour ma part, toutes sont acceptées, mais les utilisateurs auront une utilisation limitée uniquement au traitement de la donnée.

```
Anacoda Prompt (anaconda3) - mysql -h db-company.cwm601wtiqo3.eu-west-3.rds.amazonaws.com -u admin -p

(base) C:\Users\namor>mysql -h db-company.cwm601wtiqo3.eu-west-3.rds.amazonaws.com -u admin -p
Enter password: *****
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 3281
Server version: 8.0.23 Source distribution

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> CREATE USER 'jordan_new'@'%' IDENTIFIED BY 'welcome123' ;
Query OK, 0 rows affected (0.21 sec)

mysql> GRANT SELECT ON *.* TO 'jordan_new'@'%' WITH GRANT OPTION ;
Query OK, 0 rows affected (0.31 sec)

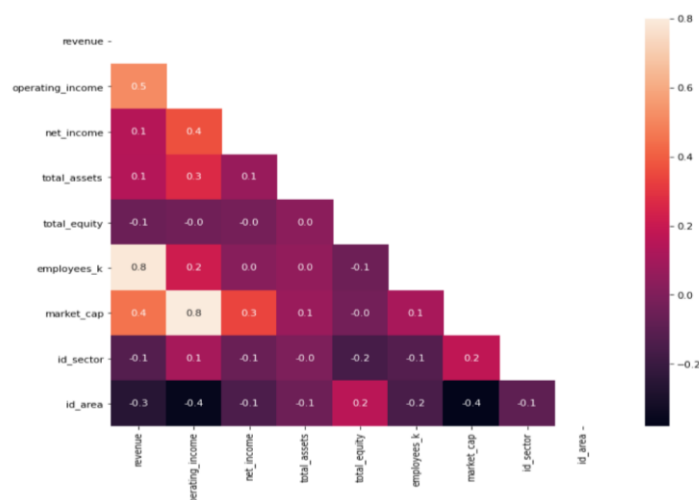
mysql> SHOW GRANTS FOR 'jordan_new' ;
+-----+
| Grants for jordan_new@% |
+-----+
| GRANT SELECT ON *.* TO 'jordan_new'@'%' WITH GRANT OPTION |
+-----+
1 row in set (0.20 sec)

mysql>
```

## 3. AXES D'ANALYSES

### 3.1. REQUETES

Voici une visualisation qui n'apparaît dans l'application mais qui peut orienter nos axes d'analyses en nous aidant à savoir quel facteur exerce une influence sur une autre.



Plus les couleurs sont claires, plus la corrélation est élevée. Par exemple, nous observons que le chiffre d'affaires est fortement corrélé au nombre d'employé. Notre travail est de vérifier ces exemples d'analyses directement avec les facteurs qui nous sont disponibles.

### REQUÊTE : EXEMPLE 1 – QUELS SONT LES MEILLEURES ENTREPRISES ?

Le but de cette requête est de récupérer les dix meilleures entreprises ayant la capitalisation boursière la plus grande. Cet indicateur peut nous révéler des informations importantes sur ce que vaut l'entreprise sur le marché boursier et à quel point celle-ci est bien cotée au vu des investisseurs :

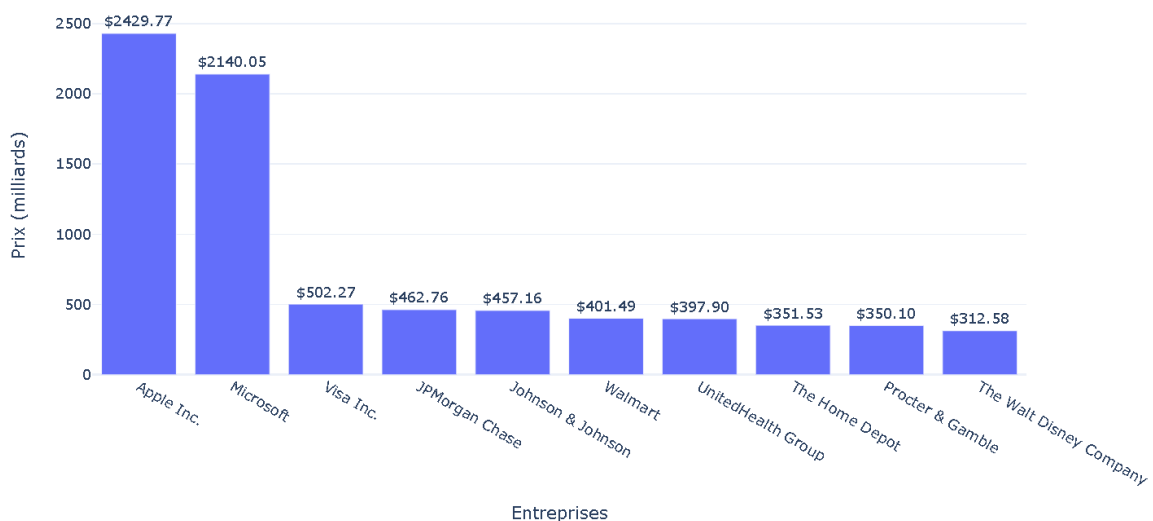
```
SELECT company, market_cap
FROM company
ORDER BY market_cap DESC
LIMIT 10;
```



	company	market_cap
►	Apple Inc.	2429.77
	Microsoft	2140.05
	Visa Inc.	502.269
	JPMorgan Chase	462.757
	Johnson & Johnson	457.157
	Walmart	401.491
	UnitedHealth Group	397.903
	The Home Depot	351.534
	Procter & Gamble	350.097
	The Walt Disney Company	312.585

La requête ci-dessus nous montre qu'Apple est l'entreprise ayant la capitalisation boursière la plus élevée avec une valeur de 2429.77 milliards de dollars. Cette information constitue l'introduction de l'objectif de ce projet, à savoir si les facteurs que nous possédons actuellement ont représenté une influence sur cette valeur.

Top 10 des entreprises ayant les capitalisations boursières les plus élevées



Le graphique regroupe toute entreprise qu'elle soit américaine, française ou allemande. Pourtant, celui-ci nous montre une domination clairement visible. Apple et Microsoft sont des entreprises du secteur technologique, peut-on penser qu'une entreprise qui est dans ce secteur a plus de chance d'être cotée sur les marchés financiers ? Notre application cherche à le savoir.

## REQUÊTE : EXEMPLE 2 – QUELS SONT LES MEILLEURS SECTEURS ?

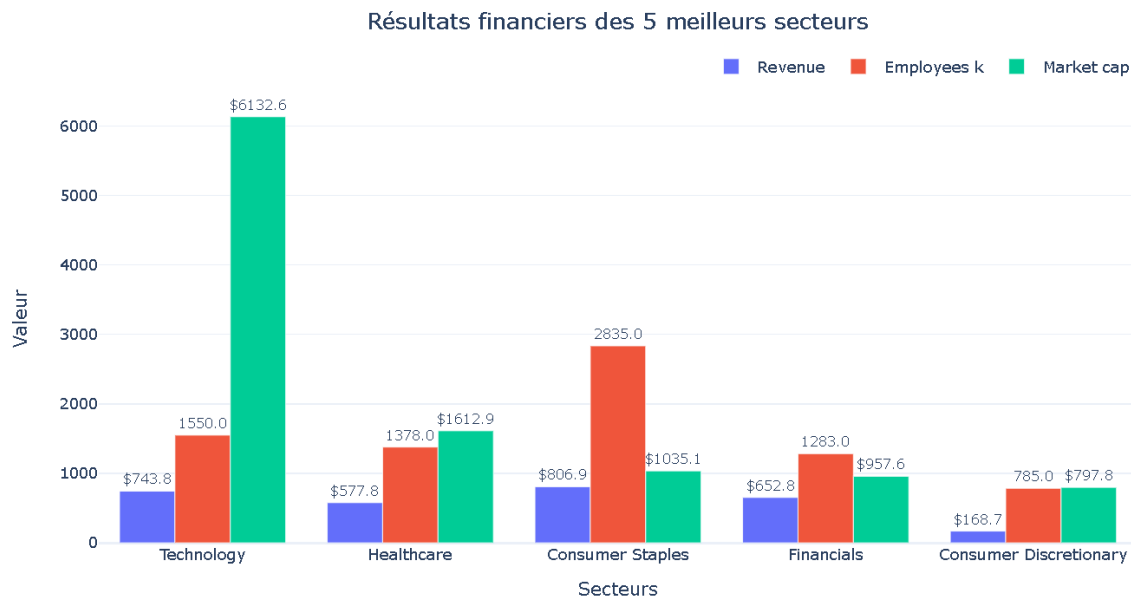
Pour répondre à cette question, nous allons pencher notre axe d'analyse sur ce sujet, à savoir les secteurs, et observer les différents critères financiers puis les comparer entre elles.

```
SELECT SUM(revenue) as revenue, SUM(employees_k) as employees_k, SUM(market_cap) as market_cap, C.id_sector, S.sector
FROM company C
LEFT JOIN sector S ON C.id_sector = S.id_sector
GROUP BY id_sector
ORDER BY market_cap DESC
LIMIT 5;
```



	revenue	employees_k	market_cap	id_sector	sector
►	743.7720165252686	1550	6132.575043005403	17	Technology
	577.7749853134155	1378	1612.9180251802318	13	Healthcare
	806.9200096130371	2835	1035.137981414795	8	Consumer Staples
	652.7910034656525	1283	957.6419948041439	11	Financials
	168.67599868774414	785	797.7669982910156	6	Consumer Discretionary

Il semblerait que le résultat soit à la hauteur de ce que nous pensions. Le secteur technologique est premier au classement et regroupe les capitalisations boursières les plus élevées. Comparé aux autres, la différence est nettement visible.



### 3.2. DEVELOPPEMENT DU SITE WEB

La visualisation utilisée pour réaliser ces graphiques est Plotly. En plus de donner des visualisations esthétiques, elles sont également interactives, ce qui nous permet d'observer les informations avec plus de détails qu'une librairie classique.

Une fois connecté à la base de données, nous pouvons réaliser différentes requêtes qui nous permettront de modéliser les visualisations. Dans cet exemple, la fonction `get_bar()` récupère les informations du dataframe puis retourne un diagramme à barres horizontales.

```
def get_bar(df, x, y):
    fig = go.Figure()

    fig.add_trace(go.Bar(x=df[y], y=df[x], text=df[y].map('${:.2f}'.format), orientation='h'))

    fig.update_layout(margin=dict(l=10, r=10, t=10, b=10),
                      template="plotly_white",
                      showlegend=False,
                      barmode='group',
                      yaxis={'categoryorder': 'total ascending'},
                      uniformtext_minsize=8,
                      uniformtext_mode='hide',
                      xaxis_title=None, #y.capitalize(),
                      yaxis_title=None #x.capitalize(),
                      )

    fig.update_traces(textposition='outside')
    return fig
```

Ce qui nous permet d'obtenir la visualisation montrée précédemment sur les dix meilleures entreprises en matière de capitalisation boursière.

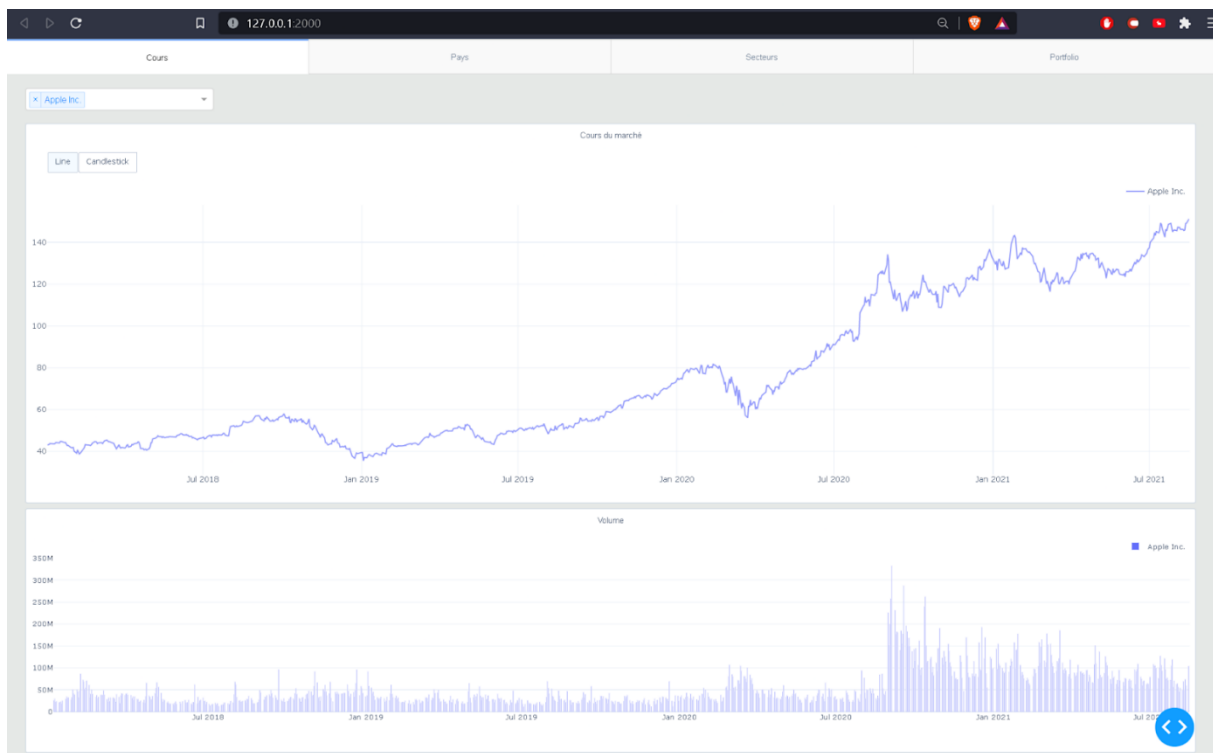
```

html.Div(
    className="four columns volume_div",
    children=[
        html.P("Entreprises comptant la capitalisation boursière la plus élevée"),
        dcc.Graph(
            id="live-update-graph",
            style={"height": "91%", "width": "100%"},
            config=dict(displayModeBar=False, scrollZoom=False),
            figure=get_bar(
                get_data_from_sql(
                    ''' SELECT id_company, market_cap FROM company ORDER BY market_cap DESC LIMIT 10; '''
                ), 'id_company', 'market_cap'
            )
        )
    ],
),

```

Toutes les visualisations sont disponibles sur l'application web par le biais de Dash, développé également par Plotly. Le but est d'observer des axes d'analyses qui permettraient de comprendre les influences exercées sur le prix d'une action. Pour cela, les facteurs financiers, liés au nombre d'employé et les secteurs seront pris en compte.

Les informations relatives aux investisseurs nous permettent d'observer un échantillon de leur comportement sur les marchés financiers, mais ne constitue aucunement une observation réelle.





## 4. BILAN DE PROJET

### 4.1. OBSERVATIONS

Pour rappel, le but était d'apporter une vision plus claire sur ce qui pourrait influencer le prix d'une action, basé sur les différents facteurs qui nous sont à disposition.

Pour cela, les points abordés sont :

- Les résultats financiers (chiffre d'affaires, capitalisation boursière, bénéfice avant/après impôts, ...)
- Le secteur d'activité
- Le pays d'origine
- Les investisseurs

Suivant les différentes observations effectuées, nous pouvons remarquer les faits suivants :

- 67% des quinze entreprises ayant le prix d'une action le plus élevé sont américaines, 20% sont françaises et 13% sont allemandes.
- 100% des dix entreprises ayant la capitalisation boursière la plus élevée sont américaines
- 70% d'entre-elles ont le chiffre d'affaires le plus élevé, et c'est à 50% le même cas concernant le nombre d'employé

De plus, il semblerait que la capitalisation boursière représente un critère assez rassurant pour les investisseurs. En prenant en compte le fait que ce même facteur représente 10'000 milliards de dollars pour les entreprises américaines, 40% de cette valeur totale est attribuée à Apple et Microsoft réunie, appartenant justement au secteur technologique.

Le secteur étant un facteur très important dans l'évaluation des cours, les observations montrent que le secteur technologique domine complètement les marchés financiers, en plus d'appartenir en grande majorité au top des secteurs présentant les résultats les plus élevés.

### 4.2. CONCLUSIONS

Le cours d'une action sera donc fonction de facteurs endogènes : sa structure financière, sa rentabilité, les anticipations sur ses bénéfices, sa part du chiffre d'affaires, sa place sur le marché par rapport à ses concurrents, ses actifs. Mais cette évolution sera aussi fonction de facteurs exogènes : le pays d'origine, l'appréhension du secteur dans lequel elle évolue.

L'ensemble de ces facteurs doit donc être correctement ciblé et analysé par l'investisseur.

### 4.3. LES DIFFICULTES

Le projet a suscité un nombre d'heure conséquent pour apporter un résultat potable et présentable. Les différentes attentes vis-à-vis de celui-ci ont rendu la tâche complexe en sous-estimant la charge de travail nécessaire à l'établissement d'une fonctionnalité que nous aimerions observer dans le projet.

Ainsi, parmi toutes les difficultés qui ont été observées durant ce travail, nous pouvons énumérer celles-ci :

- Créer un tableau de bord en ligne et le dynamiser
- Collecter et nettoyer la donnée
- Apprendre des technologies nouvelles (HTML, CSS, Dash, Plotly)

Plotly est une librairie très intéressante et surtout fonctionnelle, ce qui en fait également son plus gros défaut. Celle-ci est tellement modifiable qu'elle demande un apprentissage profond pour s'en familiariser. Le chemin pour obtenir la visualisation souhaitée peut être très fastidieux et complexe, ce qui explique que toutes les améliorations possibles n'ont pas pu être effectuées pour une question chronologique.

La création d'une application web représente également un vrai défi. Cependant, l'apprentissage du HTML et du CSS est très intéressant, car il vient à récolter et développer le fruit de notre créativité.

### 4.4. LES AXES D'AMELIORATIONS

Des axes d'amélioration sont à apporter pour parfaire le projet :

- Améliorer l'interface web
- Actualiser les informations (si possible selon la disponibilité de la data – après 2016) - Créer une veille automatique sur le sujet
- Créer un calendrier avec les grandes dates/événements sur le sujet
- Croiser ces données avec ceux du réchauffement climatique par exemple
- Utiliser des méthodes en Machine Learning et en intelligence artificielle pour analyser les différents actifs financiers
- Récupérer les différents articles qui parlent de sujets financiers et effectuer de la NLP

#### 4.5. REMERCIEMENTS

Je souhaite remercier les 4 formateurs qui ont accompagné la promotion 2021 de Développeur Data de Nanterre, sans qui ce projet n'aurait jamais pu aboutir :

- MME Manel BOUMAIZA,
- M. Josselin TOBELEM,
- M. David AZRIA,
- M. Nicolas ZANFORLINI

Et enfin, je tiens à remercier les 20 apprenants de notre promotion, avec qui j'ai eu l'honneur de partager des moments qui resteront inoubliables.