UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

DÉPARTEMENT D'INFORMATIQUE ET MATHÉMATIQUE

# TRAVAIL PRATIQUE 1

Travail présenté à M. Julien Maître par

Colin Bouchard - BOUC10049302

Frédérike Tremblay – TREF09519705

Comme exigence partielle du cours

Outils de programmation pour la science des données

8PRO408

27 octobre 2020

## 1. DESCRIPTION DE L'ENSEMBLE DES DONNÉES

- Dans quel objectif les données ont été collectés (s'il y a lieu)?

Tout d'abord, il est important de mentionner que dans le cadre de l'analyse des données du travail pratique 1, deux ensembles de données principaux seront analysés. Le premier ensemble de données correspond à des données d'ordre globale en libre accès sur les test de COVID-19 par une équipe de "Our World in Data" dans un but de générer des statistiques et des recherches (Roser *et al*. 2020). L'objectif principal, d'ailleurs mentionné par l'équipe, est de vérifier si le nombre de cas par pays est représentatif de la situation actuelle en fonction du niveau de test (Roser *et al*. 2020). L'équipe de recherche en charge de ces données tentent de vérifier quels sont les pays qui test adéquatement en ce moment : en effet, un pays ne procédant pas à un dépistage adéquat déclare ainsi actuellement un nombre de cas grandement sous-estimé (Roser *et al*. 2020).

Le deuxième ensemble de donnée concerne les données macroéconomique, et microéconomique respectivement. L'agrégat macroéconomique que nous avons sélectionnée est le S&P500; il s'agit d'un indice composite de 500 grandes entreprise américaine (Lou , 2018). Pour ce qui est des donnée micro-économique nous avons choisie trois actions du marché américain soit: google, zoom et costco. Les donnée financière ont été recueilli par yahoo finance dans le but d'offrir un service de données libre concernant les marchés financiers.

Nous avons comme objectif général d'établir les corrélations entre l'augmentation des cas de covid-19 et les marchés financier au cours de 2020; les données concerné ont donc été ciblé pour être comprise entre le mois de janvier et septembre 2020 inclusivement.

- Comment ont été collectées les données (s'il y a lieu) ?

Les données sur le niveau de dépistage de la COVID-19 sont accumulées de manière quotidienne à partir de l'agence European Center for Disease Prevention and Control (ECDC) (Roser et al. 2020; ECDC, 2020). L'ECDC rend disponible leurs données qu'ils collectent sur la base de rapports quotidiens des autorités sanitaires à l'échelle mondiale (ECDC, 2020). Celle-ci on été téléchargé en format .csv à partir du site web ourworldindata.org.

Pour ce qui est des données financière elle sont collectés par yahoo finance de façon journalière dans un but commercial, puisqu'il s'agit pour eux d'un avantage concurrentiel. Ses données on été télécharger en format .csv à partir du site de yahoo finance dans la section données historique. Le tout fut importer dans python en utilisant la librairie pandas avec sa fonction read_csv() résultant plusieurs DataFrame qui seront éventuellement combinée.

- Quelles sont les variables (à mettre sur forme de bullet point)?

Les données sur la COVID-19 regroupent les variables concernant le code ISO, le continent, la location (pays), la date, le nombre total de cas, le nombre de nouveau cas, le nombre de nouveaux cas adouci, le nombre total de décès, les nouveaux décès, les nouveaux décès adouci, le nombre total de cas par million, nombre de nouveau cas par million, le nombre de nouveaux cas adouci par million, le nombre total de mort par million, le nombre de nouveau mort par million, le nombre de nouveaux décès adouci par million, le nombre de nouveaux tests, le total de test, le total de test pas millier, le nombre de nouveaux tests par millier, le nombre de nouveaux tests adoucis, le nombre de nouveaux tests adoucis par millier, le nombre de test par cas, le taux positif, le nombre de tests par unité, l'index de rigueur, la population, la densité de la population, âge médian, le nombre de personne âgé de 65 ans et plus, le nombre de personnes âgé de 70 ans et plus, le produit intérieur brut par capita, pauvreté extrême, taux de décès cardiovasculaire, prévalence du diabète, nombre de femmes fumeuses, nombre d'hommes fumeurs, nombre d'installation de lavage de main, nombre de lits d'hôpital par millier, l'espérance de vie ainsi que l'indice de développement humain. Dans le contexte actuelle de nos recherche les valeur qui nous intéresse sont principalement la date, la localisation, et le nombre total de cas.

Pour ce qui est des données financières, autant pour les données microéconomique que les données macroéconomique; les variables sont les suivantes: la date, le prix d'ouverture, le plus haut prix, le plus bas prix, le prix de fermeture, l'adjacent de fermeture et le volume total pour chaque jours. On s'intéresse ici principalement au variable de date et de prix. Pour ce faire nous devons faire une moyenne entre les quatre valeur de prix afin d'obtenir les valeur d'une charte ohlc (Open+High+Low+Close/4). Cette valeur sera utilisé comme indice de prix journalier pour l'ensemble des données microéconomique et macroéconomique.

- Le nombre d'instance (se référer au document explicatif avec les notes)

L'ensemble de donnée concernant covid-19 contient 47 538 instances. Plusieurs de celle-ci ne nous intéresse pas puisqu'elle concerne d'autre pays que notre pays cible (les États Unis).

Pour ce qui est des donnée financière, elle on tous 189 instances; nous pouvons ici expliquer les données manquantes par la fermeture des marchés durant la fin de semaine.

- Le nombre de classe (s'il y a lieu)

Les données de covid-19 contiennent un nombre faramineux de classes qui ne nous intéresse pas tous. Trois principales classes nous intéressent particulièrement dans le cas présent soit: la date, la localisation et les nombre de cas totaux.

Pour ce qui est des données financières nous avons trois classes soit: la date, les prix et le volume. Seulement deux d'entre elle nous intéresse soit la date et le prix.

En sommes nous avons affaire avec quatre classes soit: la date, la localisation, le nombre de cas totaux et le prix.

- Les valeurs (intervalle min-max) que peuvent prendre chacune des variables?

Voici une description des valeurs qui a été fait par l'utilisation de la fonction pandas describe(). Tout d'abord pour les données covid:

| | total_cases |
|---|---|
| count | 4.692400e+04 |
| mean | 1.062962e+05 |
| std | 1.099990e+06 |
| min | 0.000000e+00 |
| 25% | 6.300000e+01 |
| 50% | 1.077000e+03 |
| 75% | 1.163425e+04 |
| max | 3.435072e+07 |

On peut voir ici que les valeurs sont trop grande pour être prise en compte par python il faudra donc reconvertir à partir de valeur scientifique.

On voit ici le tableau une fois convertie en valeur réel:

| | Cas totaux |
|---|---|
| Nombre total: | 47 538 |
| Moyenne: | 106296.2 |
| Déviation standard (Écart-type): | 1 099 990 |
| Minimum: | 0 |
| Q1 (25e percentile) | 63 |
| Q2 (médian) | 1077 |
| Q3 (75e percentile) | 11 634 |
| Maximum: | 34 350 720 |

Toutefois l'information qui nous intéresse concerne uniquement les États Unis et les données actuelle sont l'international; il faudra donc fractionner le dataset pour n'obtenir que l'information qui nous concerne.

Pour passer au donnée financière voici les résultat de la fonction describe de pandas pour chacun des datasets:

| | snp500_candle |
|---|---|
| count | 189.000000 |
| mean | 3103.004361 |
| std | 280.329424 |
| min | 2255.174988 |
| 25% | 2921.669983 |
| 50% | 3186.262512 |
| 75% | 3325.505005 |
| max | 3561.985046 |

| | google_candle |
|---|---|
| count | 189.000000 |
| mean | 1410.555167 |
| std | 134.213155 |
| min | 1050.698975 |
| 25% | 1356.640991 |
| 50% | 1435.475006 |
| 75% | 1498.243744 |
| max | 1700.391266 |

| | zoom_candle |
|---|---|
| count | 189.000000 |
| mean | 199.912758 |
| std | 107.172060 |
| min | 67.670000 |
| 25% | 112.145000 |
| 50% | 167.465000 |
| 75% | 257.672508 |
| max | 500.753243 |

| | costco_candle |
|---|---|
| count | 189.000000 |
| mean | 314.592844 |
| std | 18.075267 |
| min | 281.250000 |
| 25% | 302.162506 |
| 50% | 308.655006 |
| 75% | 326.864998 |
| max | 355.924995 |

On peut remarquer que tous les dataset financier on le même nombre de données soit 189. Le reste des données semble variée d'un dataset à l'autre. De plus, il est à noter que, tel que mentionné plus tôt, nous avons utilisé une valeur de charte ohlc obtenue à partir du calcul suivant:

```
SNP500['snp500_candle'] = (SNP500['Open'] + SNP500['High'] + SNP500['Low'] + SNP500['Close']) / 4
google['google_candle'] = (google['Open'] + google['High'] + google['Low'] + google['Close']) / 4
zoom['zoom_candle'] = (zoom['Open'] + zoom['High'] + zoom['Low'] + zoom['Close']) / 4
costco['costco_candle'] = (costco['Open'] + costco['High'] + costco['Low'] + costco['Close']) / 4
```

-ect…

Nous avons bien sûr isolé les valeur qui nous intéresse du dataset de la covid-19 par l'application de la ligne de code suivante:

```
covid_mod = covid[['Date', 'location', 'total_cases']].query("location == 'United States'").set_index('Date')
```

En résumé nous faisons la sélection des colonnes 'Date', 'location' et 'total_cases' lorsque la location est égale au État Unis; le tout avec la Date en index. L'utilisation de double bracket nous permet d'aller sélectionner les colonnes. Nous avons ensuite utilisé la fonction query() pour aller chercher la localisation précis et la fonction set_index pour placer la date en index.

Nous avons ensuite répété l'opération à double bracket sur les donnée financière pour sélectionner la 'Date' et la chandel de prix pour chacune des donnée financières et bien sûr nous avons placé la Date en index pour faciliter la concaténation à venir:

```
SNP500_mod = SNP500[['Date', 'snp500_candle']].set_index('Date')
google_mod = google[['Date', 'google_candle']].set_index('Date')
zoom_mod = zoom[['Date', 'zoom_candle']].set_index('Date')
costco_mod = costco[['Date', 'costco_candle']].set_index('Date')
```

Nous avons ensuite concaténé sur l'axis 1 (colonnes):

```
ensemble = pd.concat([covid_mod, SNP500_mod, google_mod, zoom_mod, costco_mod], axis=1)
```

## 2. VÉRIFICATION ET PRÉ-TRAITEMENT DES DONNÉES

- Combien de valeur manquante possède l'ensemble de données

Nous avons d'abord testé notre ensemble pour connaître ou se trouve nos données manquantes avec la fonction pandas isnull().any() qui nous créas un masque booléen:

```
print('Vérification de données manquantes pour l\'ensemble:\n', ensemble.isnull().any(), '\n\n')
```

```
Vérification de données manquantes pour l'ensemble:
 location         False
total_cases      False
snp500_candle     True
google_candle     True
zoom_candle       True
costco_candle     True
dtype: bool
```

Nous voyons donc que des données financières sont manquante; cela s'explique par la fermeture des marchés durant la fin de semaine. Nous devons donc calculer combien de valeur sont manquante. Nous utilison donc la fonction suivante:

```
print('Il y a ', ensemble['total_cases'].count() - ensemble['snp500_candle'].count(), ' donnees manquantes\n\n')
```

Résultat:

```
Il y a  88  donnees manquantes
```

- Quel(s) moyen(s) avez-vous utilisé pour gérer ses valeurs manquantes ?

Puisque les données manquantes sont des données financière qui sont vraisemblablement en provenance des fermeture de marché les fin de semaine, la meilleure option est de supprimer les donnée des ligne manquante et combler les graphique avec un 'backfill' puisque les donnée financière reste les même durant la fin de semaine (dû à la fermeture des marchés). Par contre,

pour les données de covid-19 il faudra faire un ‘frontfill’ puisqu’elle continue d’augmenter durant la fin de semaine.

Nous avons donc utilisé la fonction dropna() de pandas pour se débarrasser des valeur manquantes. Nous avons aussi utiliser la fonction drop() pour se débarrasser de la colonne ‘location’ dont nous n’avons plus besoin, il aurait été possible de s’en débarrasser plus tôt mais cela n'aurait pas affecté les opérations précédente:

```
e_formater = ensemble.dropna().drop(columns=['location'])
```

Il faut maintenant re-vérifier pour nos valeurs manquantes:

```
print('Vérification de données manquantes pour l\'ensemble:\n', e_formater.isnull().any(), '\n\n')
```

```
Vérification de données manquantes pour l'ensemble:
 total_cases      False
snp500_candle    False
google_candle    False
zoom_candle      False
costco_candle    False
dtype: bool
```

Excellent! Nous avons nos données tous ensemble dans le même DataFrame, avec aucune donnée manquantes.

- Un résumé du nombre d’instance par classes (s’il y a lieu)

Une fois les données formater, nous sommes passer de 277 instances (nombre d’instance des donnée covid après avoir sélectionné la localisation des état unis) à 189 instances; exactement le même nombres d’instance que l’ensemble des donnée financière.

- Quels sont les statistiques (moyenne,variance, écart-type) pour chaque variable

Grâce à la fonction pd.describe() utiliser plutot nous avons déjà extrait les valeur de moyenne et d’écart-type pour tout les variable de notre ensemble parcontre celle-ci était les description

précédent le pré-traitement. Nous allons maintenant appliquer les fonction pandas mean() [moyenne], var() [variance] et std() [écart-type] pour en extraire les valeurs statistiques.

```
# Moyenne
print('Moyenne total_cases:',e_formater[['total_cases']].mean(), '\n')  # valeur en affichage scientifique.
print('Moyenne snp500_candle:',e_formater[['snp500_candle']].mean(), '\n')
print('Moyenne google_candle:',e_formater[['google_candle']].mean(), '\n')
print('Moyenne zoom_candle:',e_formater[['zoom_candle']].mean(), '\n')
print('Moyenne costco_candle:',e_formater[['costco_candle']].mean(), '\n\n')

# Variance des donnee
print('Variance total_cases:',e_formater[['total_cases']].var(), '\n')  # valeur en affichage scientifique.
print('Variance snp500_candle:',e_formater[['snp500_candle']].var(), '\n')
print('Variance google_candle:',e_formater[['google_candle']].var(), '\n')
print('Variance zoom_candle:',e_formater[['zoom_candle']].var(), '\n')
print('Variance costco_candle:',e_formater[['costco_candle']].var(), '\n\n')

# Ecart-type
print('Ecart-type total_cases:',e_formater[['total_cases']].std(), '\n')  # valeur en affichage scientifique.
print('Ecart-type snp500_candle:',e_formater[['snp500_candle']].std(), '\n')
print('Ecart-type google_candle:',e_formater[['google_candle']].std(), '\n')
print('Ecart-type zoom_candle:',e_formater[['zoom_candle']].std(), '\n')
print('Ecart-type costco_candle:',e_formater[['costco_candle']].std(), '\n')
```

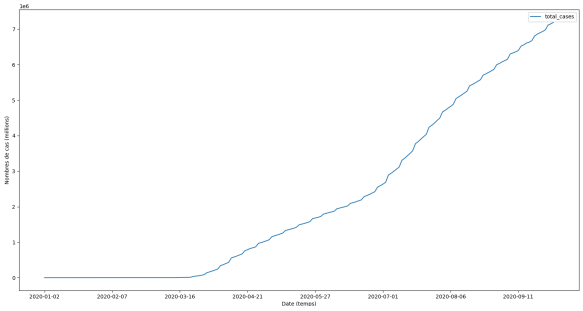
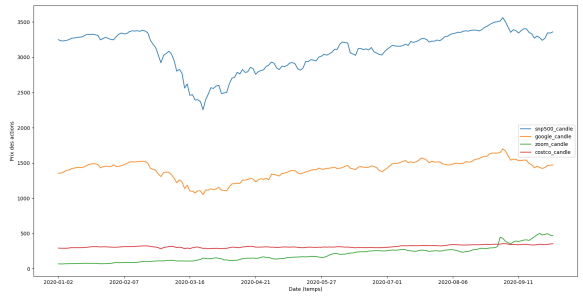
Voici donc les résultat pour chacune des variable:

| | total_cases | snp500 | google | zoom | costco |
|---|---|---|---|---|---|
| Moyenne | 2 213 757.00 | 3103 | 1 410.56 | 199.91 | 314.59 |
| Variance | 5 492 875 000 000.00 | 78 584.59 | 18 013.17 | 11 485.85 | 326.72 |
| Écart-type | 2 343 688.00 | 280.33 | 134.21 | 107.17 | 18.87 |

-Utilisation des outil de visualisation

Nous pouvons tout d'abord visualiser nos graphique simple en utilisant la librairie matplotlib.pyplot. Pour ce faire il faudra traiter les données Covid-19 séparément des données financière puisqu'il ne sont pas sur la même échelle, nous utiliserons donc les commande suivante:

```
plt.plot(e_formater[['total_cases']])
plt.xlabel('Date (temps)')
plt.ylabel('Nombres de cas')
plt.show()

finance = e_formater[['snp500_candle','google_candle','zoom_candle','costco_candle']]
finance.plot(kind='line', stacked=False)
plt.xlabel('Date (temps)')
plt.ylabel('Prix des actions')
plt.show()
```

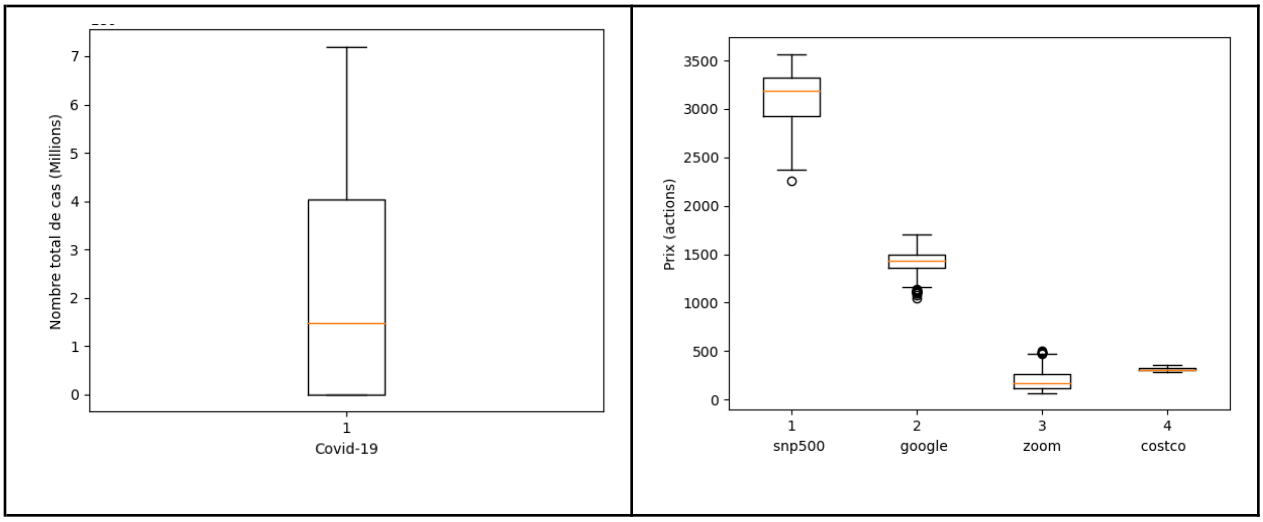
| | |
|---|---|
| Résultat:<br>  <br>  | En haut, nous avons la courbe du coronavirus au Etat-Unis; en bas, les courbe des actions.<br>On peut déjà commencer à observer des tendances. Il semblerait que l'appréhension du coronavirus est fait quelque peu baisser les marchés; mais au moment ou on voit une augmentation des cas au État-Unis, les marchés semble reprendre du mieux. Il s'agit peut-être du résultat du stimulus économique poussé par le gouvernement américain.<br><br>Toutefois il est trop tôt pour le savoir, il faudra observer les testes statistique pour en être sûre. |

Il est bien sûr intéressant d'observer les données selon plusieurs point de vue. Nous utiliserons donc les commande suivant pour visualiser nos données sous forme de boîte à moustache (boxplot):

```
np_total_cases = np.ravel(e_formater[['total_cases']].transpose().reset_index(drop=True).to_numpy())
np_snp500 = np.ravel(e_formater[['snp500_candle']].transpose().reset_index(drop=True).to_numpy())
np_google = np.ravel(e_formater[['google_candle']].transpose().reset_index(drop=True).to_numpy())
np_zoom = np.ravel(e_formater[['zoom_candle']].transpose().reset_index(drop=True).to_numpy())
np_costco = np.ravel(e_formater[['costco_candle']].transpose().reset_index(drop=True).to_numpy())

data_to_plot = [np_snp500, np_google, np_zoom, np_costco]

plt.boxplot(np_total_cases)
plt.ylabel('Nombre total de cas (Millions)')
plt.xlabel('Covid-19')
plt.show()

plt.boxplot(data_to_plot)
plt.ylabel('Prix (actions)')
plt.xlabel('snp500                google                zoom                costco')
plt.show()
```

La fonction transpose() de pandas nous permet de transformer nos column en ligne, nous supprimons ensuit nos index avec la fonction reset_index(drop=True), puis nous transformons le tout en tableau numpy avec la fonction to_numpy(). La fonction numpy ravel() nous sert ensuite à retirer la pairs de bracket excédentaire de notre tableau numpy.

Résultats:

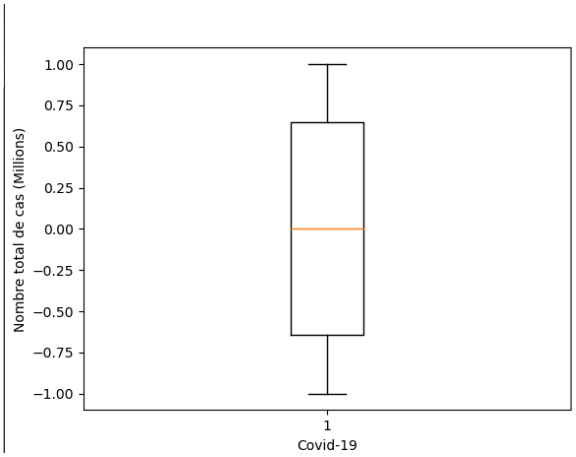


Encore une fois on peut observer les deux graphique pour voir des différence de distribution. Tout d'abord la distribution de Covid-19 ne semble pas être une distribution normale contrairement à l'action de Google ou Costco par exemple, toutefois les données du futurs reste manquantes ce qui pourrait expliquer cette tendance. De l'autre côté nous avons les action avec des tendances qui diverge les une par rapport aux autres. Tout d'abord le S&P500 avec sa moustache minimal plutôt basse pourrait indiquer la probabilité d'une corrélation négative par rapport au nombre de cas de Covid-19; toutefois il faudra valider le tout à l'aide d'un test statistique. D'un autre côté, Zoom avec sa moustache maximal plutôt élevé un peu comme celle des donnée sur le nombre de cas, nous laisse présager a une corrélation positive probable; encore une fois nous devrons valider avec un test statistique. Pour ce qui est de Google et Costco leur forme qui semble normal, nous laisse présager à une probable corrélation nul par rapport au données de Covid-19; un test statistique sera nécessaire.

Toutefois toute ses hypothèses sont invalider si l'on observe les données de Covid-19 transformer en sinus grâce a la fonction numpy sin(). Voici comment l'on s'y est pris:

```
plt.boxplot(np.sin(np_total_cases))
plt.ylabel('Nombre total de cas (Millions)')
plt.xlabel('Covid-19')
plt.show()
```

Résultat:

| | |
|---|---|
|  | Cette transformation en sinus permet de mieu observer la boite à moustache (boxplot) pour les données Covid. On peut même dire que les donnée on ete normaliser.<br><br>Lorsque l'on observe les valeurs transformer on peut voir une différence qui rendrais l'observation précédente invalide.<br><br>Il faudra donc bien tester les corrélation entre chaque variables pour déterminer si oui ou non la Covid-19 à influencer le marché américain. |

## 3. ÉTUDE STATISTIQUE DES DONNÉES ET INTERPRÉTATIONS

hypothèse null = L'augmentation des cas de covid-19 n'a pas eu d'effet sur les prix des actions à la bourse.
hypothèse alternative = L'augmentation des cas de covid-19 à eux un effet sur le prix des action à la bourse

- Étude statistique pour chaque variable par rapport par rapport aux différentes classes et /ou différentes valeurs
- des tests d'hypothèses
- corrélations possibles entre deux ou plusieurs variables
- des tests du Khi2
- etc
- **utliser des outils de visualisation des données

*inclure une autre méthode de traitement, ou statistique ou de visualisation, que nous n'avons pas vue en cours.

*démarche reposant sur des hypothèses et de la validation des hypothèses en fonction des résultats obtenus

*code = constituer d’un fichier main et de d’autres fichiers ou on va créer des fonctions (lire le document les fonctions en python)

*pr pwp : suivre le document conseils pour les rapports et les supports de présentation

Nous devons tout d‘abord établir notre hypothèse alternative. Nous prévoyons donc que l’augmentation des cas de covid-19 aurait eut un impact sur le prix des actions à la bourse. Notre hypothèse nul est : l’augmentation des cas de covid-19 n’a pas eux d’impact sur les prix des actions à la bourse.

Nous allons donc commencer par observer nos données en régression linéaire grâce à la librairie seaborn; nous pouvons du même coup etablir l'équation de la droite avec la librairie scikit-learn. Il nous faut donc importer seaborn et sklearn.

Nous entrons ensuite le code suivant:

```
# Regression linear simple
model = LinearRegression()

sns.regplot(x="total_cases", y="snp500_candle", data=e_formater)
plt.show()
model.fit(np_total_cases.reshape((-1, 1)), np_snp500)
print('Coefficient de determination (total_cases, snp500):', model.score(np_total_cases.reshape((-1, 1)), np_snp500))
print('b0:', model.intercept_)
print('b1:', model.coef_, '\n')

sns.regplot(x="total_cases", y="google_candle", data=e_formater)
plt.show()
model.fit(np_total_cases.reshape((-1, 1)), np_google)
print('Coefficient de determination (total_cases, google):', model.score(np_total_cases.reshape((-1, 1)), np_google))
print('b0:', model.intercept_)
print('b1:', model.coef_, '\n')

sns.regplot(x="total_cases", y="zoom_candle", data=e_formater)
plt.show()
model.fit(np_total_cases.reshape((-1, 1)), np_zoom)
print('Coefficient de determination (total_cases, zoom):', model.score(np_total_cases.reshape((-1, 1)), np_zoom))
print('b0:', model.intercept_)
print('b1:', model.coef_, '\n')

sns.regplot(x="total_cases", y="costco_candle", data=e_formater)
plt.show()
model.fit(np_total_cases.reshape((-1, 1)), np_costco)
print('Coefficient de determination (total_cases, costco):', model.score(np_total_cases.reshape((-1, 1)), np_costco))
print('b0:', model.intercept_)
print('b1:', model.coef_, '\n')
```

Tout d'abord on crée une variable model qui est une régression linéaire provenant de la classe linear_model de sklearn. Ensuite, on a la fonction regplot() de seaborn qui nous permet d'afficher de configurer notre graphique de régression linéaire simple; puis on utilise plt.show() pour afficher le graphique. Puis pour chaque graphique on utilise la fonction fit() du modèle de régression linéaire pour configurer les donnée pour obtenir ensuite notre coefficient de détermination, notre b0 et notre b1 avec les fonction respective score(), intercept_() et coef_().

Voici les resultats console:

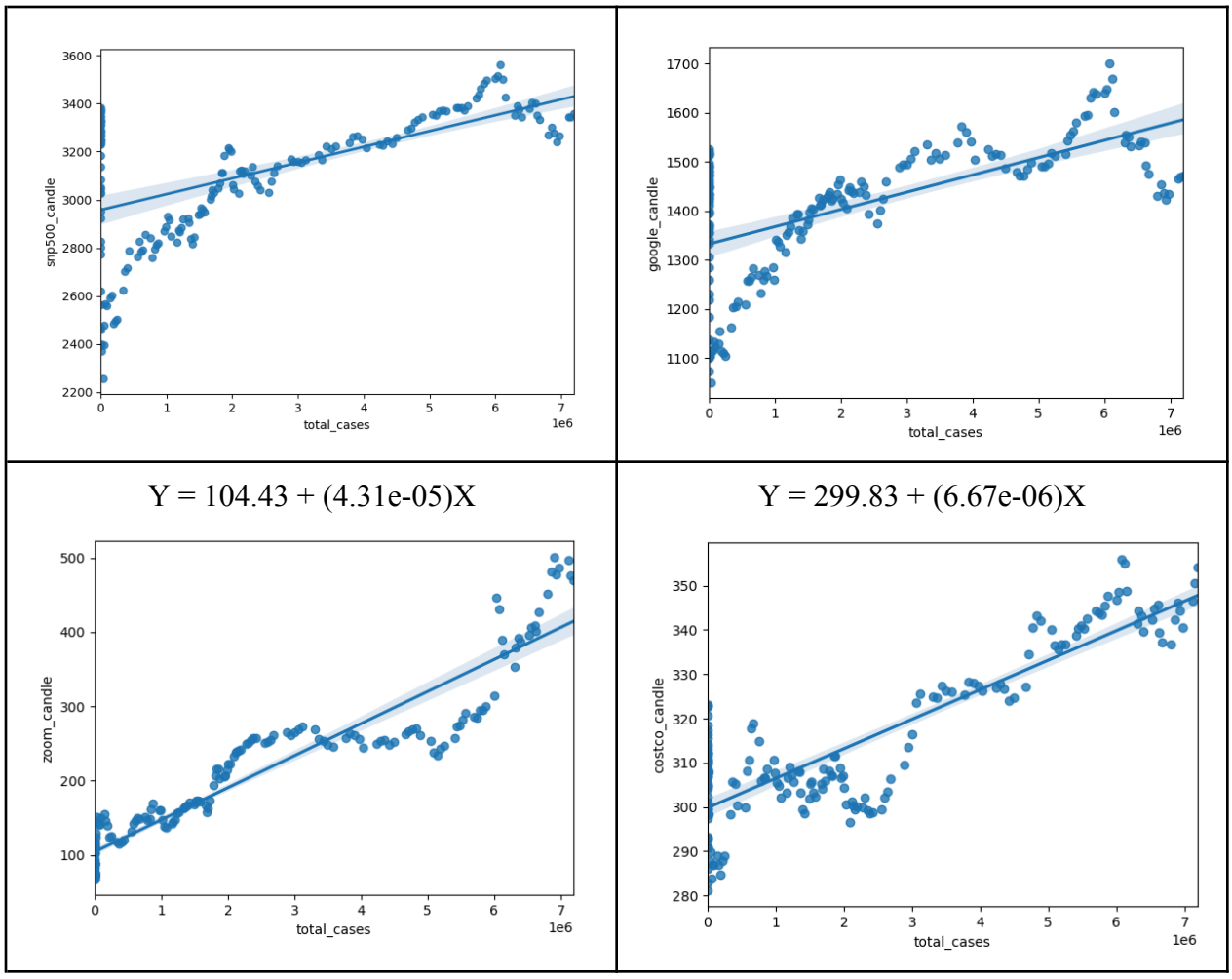
```
Coefficient de determination (total_cases, snp500): 0.3017166934525336
b0: 2957.559483477194
b1: [6.57004721e-05]

Coefficient de determination (total_cases, google): 0.3774545131301771
b0: 1332.66951249477
b1: [3.51825679e-05]

Coefficient de determination (total_cases, zoom): 0.8896503432820199
b0: 104.4307970423155
b1: [4.31311849e-05]

Coefficient de determination (total_cases, costco): 0.747232295496753
b0: 299.8343308128722
b1: [6.66672698e-06]
```

Nous avons donc tout le nécessaire pour  trouver les équation pour chaque tableau. De plus, grâce à notre coefficient de détermination r2 on peut observer si les regression sont significative.

On voit donc que seulement 30% des donnees du nombre total de cas, par rapport au prix du SNP500 signifie l'équation de la droite, tout comme google avec un faible 37%. D'un autre côté les droite de zoom et costco par rapport au nombre de cas est beaucoup plus significative avec des taux a 88% et 74% respectivement.

Nous voyons ici les résultat graphique des droite avec leur formule respective:

| | |
|---|---|
| Y = 2957.56 + (6.57e-05)X | Y = 1332.67 + (3.52e-05)X |




Nous voyons donc ici que la corrélation semble positive dumoin, d'un point de vue visuel. Il faudra quand même vérifier grâce à un test de corrélation. Toutefois la non-normalité des données de covid pourrait poser un problème en augmentant la possibilité d'erreur de type I (Bishara &al, 2012). C'est pourquoi nous testons les donnée normalisée grâce à la fonction sin() de numpy.

Nous pouvons donc tester la corrélation entre les variable avec les fonctions suivante:

```
# Correlations
print('Correlation entre nombre de cas de covid-19 et prix du S&P500:\n', stats.spearmanr(np_total_cases, np_snp500),
      '\n PearsonResult', stats.pearsonr(np_total_cases, np_snp500), '\n')

print('Correlation entre nombre de cas de covid-19 et prix de l\'action google:\n',
      stats.spearmanr(np_total_cases, np_google), '\n PearsonResult', stats.pearsonr(np_total_cases, np_google), '\n')

print('Correlation entre nombre de cas de covid-19 et prix de l\'action zoom:\n',
      stats.spearmanr(np_total_cases, np_zoom), '\n PearsonResult', stats.pearsonr(np_total_cases, np_zoom), '\n')

print('Correlation entre nombre de cas de covid-19 et prix de l\'action costco:\n',
      stats.spearmanr(np_total_cases, np_costco), '\n PearsonResult', stats.pearsonr(np_total_cases, np_costco), '\n')
```

Ce qui nous donne les résultat suivant:

```
Correlation entre nombre de cas de covid-19 et prix du S&P500:
 SpearmanrResult(correlation=0.35362666001999765, pvalue=5.989945143263039e-07)
 PearsonResult (0.5492874415572716, 2.7379977429385177e-16)

Correlation entre nombre de cas de covid-19 et prix de l'action google:
 SpearmanrResult(correlation=0.5187521665233179, pvalue=2.0651558872170662e-14)
 PearsonResult (0.6143732685673886, 5.349234790303906e-21)

Correlation entre nombre de cas de covid-19 et prix de l'action zoom:
 SpearmanrResult(correlation=0.9812989595601127, pvalue=9.074536466652109e-136)
 PearsonResult (0.9432127773106231, 1.9481915451404248e-91)

Correlation entre nombre de cas de covid-19 et prix de l'action costco:
 SpearmanrResult(correlation=0.6550055111965957, pvalue=1.5465780652547242e-24)
 PearsonResult (0.8644259919141445, 9.60295926565518e-58)
```

On peut donc confirmer ce qui avait été dit par rapport au données précédemment; C'est à dire que la corrélation entre le SNP500 et l'action de google par rapport au nombre de cas de covid-19 est très faible. Par contre, les action de zoom et costco sont eux beaucoup plus corrélé avec le nombre de cas de covid. On peut aussi voir que la corrélation semble positive et tout les test semble significatifs puisqu'il sont tous en dessous du taux de 0.0005. Nous pouvons maintenant passer à notre analyse de la variance (ANOVA).

Pour ce qui de nos ANOVA, nous utiliserons la fonction f_oneway() de Scipy comme suit:

```
# ANOVA
print('ANOVA unidirectionel(np_total_cases, np_snp500, np_google, np_zoom, np_costco): \n',
      stats.f_oneway(np_total_cases, np_snp500, np_google, np_zoom, np_costco), '\n')

print('ANOVA unidirectionel(np_total_cases, np_snp500): \n', stats.f_oneway(np_total_cases, np_snp500), '\n')

print('ANOVA unidirectionel(np_total_cases, np_google): \n', stats.f_oneway(np_total_cases, np_google), '\n')

print('ANOVA unidirectionel(np_total_cases, np_zoom): \n', stats.f_oneway(np_total_cases, np_zoom), '\n')

print('ANOVA unidirectionel(np_total_cases, np_costco): \n', stats.f_oneway(np_total_cases, np_costco), '\n')
```

Résultat:

```
ANOVA unidirectionel(np_total_cases, np_snp500, np_google, np_zoom, np_costco):
 F_onewayResult(statistic=168.4337840207612, pvalue=9.635932260464496e-109)

ANOVA unidirectionel(np_total_cases, np_snp500):
 F_onewayResult(statistic=168.15260499161053, pvalue=4.856847294251098e-32)

ANOVA unidirectionel(np_total_cases, np_google):
 F_onewayResult(statistic=168.41017649637487, pvalue=4.4410994390469873e-32)

ANOVA unidirectionel(np_total_cases, np_zoom):
 F_onewayResult(statistic=168.59454229743235, pvalue=4.1656594773120555e-32)

ANOVA unidirectionel(np_total_cases, np_costco):
 F_onewayResult(statistic=168.57707398160684, pvalue=4.1910031797243804e-32)
```

Nous pouvons donc observer que la variance des données est hautement significative avec des taux encore une fois bien en dessous de 0.0005

## 4. CONCLUSION DE L'ÉTUDE

Nous avons donc analysé nos données selon different aspect soit : des regression lineaire simple, des équations de droite, des test de corrélation (pearson et spearman) ainsi que des ANOVA unidirectionnel. Tout les teste effectuer porte à rejeter l'hypothèse nul qui était : l'augmentation des cas de covid-19 n'a pas eux d'impact sur les prix des actions à la bourse. Il faut par contre rester nuancer, car certain coefficient de détermination ainsi que certaine corrélation était plutôt faible. Par exemple: la courbe macro-économique représenté par S&P500 ainsi qu'une des trois action, celle de GOOGLE, avait des pente plus faible avec des coefficient de détermination plus bas, ainsi qu'une corrélation moin élevé que les autres actions. D'un autre côté les action de COSTCO et ZOOM sont sortie grande gagnante avec des coefficient de détermination beaucoup plus élevé ainsi qu'une corrélation qui s'élève au delà de 85%. Il est à ajouter une mention d'honneur à l'action de ZOOM qui avait la plus haute significativité dans tous les tests, ainsi que la pente la plus abrupte. On peut donc dire que certain marchés on profiter d'avantage de la pandémie que d'autre; mais que règle générale tous les marchés on monter malgré la chute abrupt des marchés en mars.

Puisque tous les marchés était déjà en hausse avant la pandémie, et que les donnée ne sont pas complètes (on ne connaît pas des donnée avenir) , il est difficile de dire si l'augmentation du prix des action est attribuable à la hausse de cas de covid-19. Par contre, une chose est certaine toute les variables semble alignées dans la même direction. De plus nos ANOVA ainsi que les test de corrélations, nous laisse croire que l'augmentation des cas de covid a eu un effet direct sur le prix des action ainsi que des agrégat économique comme le S&P500.

Vue la possibilité d'erreur de type I liée à la non-normalité des données (Bishara &al, 2012) il serait biaisée d'affirmer que nous rejetons l'hypothèse nul. Toutefois une observations méticuleuse des graphiques et des données tant à démontrer que l'appréhension du coronavirus et l'augmentation des cas de coronavirus à l'international aurait fait baisser le prix de certaine action ainsi que certain agrégat économique comme le S&P500 mais aussitôt l'augmentation des cas arriver au etat-unis on voit une augmentation des valeurs boursière (fort probablement dû au stimulus économique poussés par le gouvernement américain). Cette hypothèse expliquerait

pourquoi les variable semble parfois corrélé et parfois partiellement-corrélé. Ajoutons aussi que certain entreprise comme ZOOM ou COSTCO, on fort probablement profiter davantage du mode de vie imposer par la pandémie, ce qui expliquerait leurs forte corrélation avec les cas de covid-19.

Encore une fois il faut réitérer avec prudence que malgré des correlation parfois presque parfait, plusieurs facteurs sont en cause lorsque que l'on parle d'économie et que si on ne prend qu'une variable de l'équation économique alors il est certain que l'on pourrait faire dire ce que l'on veut au chiffre. Oui, les cas de covid et les prix des action vont dans la même direction; mais non il ne sont pas nécessairement directement liée ! Et cela même si les ANOVA unidirectionnel semble dire le contraire. Ce que l'on analyse dans ce travail n'est qu'une minuscule portion de l'economie, et une minuscule portion des variable qui pourrait l'affecter. Il faut donc prendre cette information avec un grain de sel.

Par contre l'information est probante ! Les droite de régression débute de valeur basse et finisse vers de valeurs haute, les équations sont positive, les corrélations sont positive, les ANOVA sont significative.

Tout point en direction d'une conclusion simple:
Plus les cas augmente au Etat-Unis et plus la bourse augmente au Etat-Unis.

Comme si le moteur économique n'avait pas vraiment arrêté, comme s'il était maintenue artificiellement en vie. Comme si l'économie avait la covid-19 mais était précieusement gardé sous un respirateur artificiel pour éviter que notre système ne se meurt.

Trêve de philosophie, on ne sait pas!

## 5. CONCLUSION GÉNÉRALE

- Résumer ce que vous avez apprécié, appris, moins apprécié par rapport aux données, aux outils utilisés ou encore le travail demandé dans ce TP1

## 6. RÉFÉRENCES

Anthony J Bishara et, James B Hittner, 2012, Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches, PMID: 22563845**,** DOI: 10.1037/a0028087

European Center for Disease Prevention and Control (ECDC). Mise à jour en 2020. Download the daily number of new reported cases of COVID-19 by country worldwide. Consulté le 2020-10-07,https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Roser M, Ritchie H, Ortiz-Ospina E et Hassel J. Mise à jour en 2020. Coronavirus Pandemic (COVID-19). Consulté le 2020-10-07, https://ourworldindata.org/coronavirus.

Lou C, 2018, Why Investors Love the S&P 500, consulté le 8 octobre 2020, https://money.usnews.com/investing/investing-101/articles/2018-10-02/why-investors-love-the-s-p-500

Documentation SciKit-Learn - Linear Regression, consulté le 10 octobre 2020, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Documentation SeaBorn - Linear Regression, consulte 10 octobre 2020, https://seaborn.pydata.org/tutorial/regression.html

Documentation NumPy - Ravel, consulté le 10 octobre 2020, https://numpy.org/doc/stable/reference/generated/numpy.ravel.html

Documentation SciPy - Statistics, consulté le 14 octobre 2020, https://docs.scipy.org/doc/scipy/reference/stats.html

## 7. SOURCES DES DONNÉES

Our World In Data, Covid-19, du 31 Décembre 2020 au 30 Septembre 2020, consulté le 8 octobre 2020 - https://covid.ourworldindata.org/data/owid-covid-data.csv

Yahoo finance, S&P500, du 1er Janvier 2020 au 30 Septembre 2020, consulté le 8 octobre 2020 - https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC

Yahoo finance, Zoom, du 1er Janvier 2020 au 30 Septembre 2020, consulté le 8 octobre 2020 - https://finance.yahoo.com/quote/ZM/history?p=ZM

Yahoo finance, Google, du 1er Janvier 2020 au 30 Septembre 2020, consulté le 8 octobre 2020 - https://finance.yahoo.com/quote/GOOG/history?p=GOOG

Yahoo finance, Costco, du 1er Janvier 2020 au 30 Septembre 2020, consulté le 8 octobre 2020 - https://finance.yahoo.com/quote/COST/history?p=COST

**Notes TP :**

- Notre question est pertinente ; est-ce qu'on va pouvoir répondre ?

- Est-ce que l'info compris dans le dataset de l'action sera suffisante pour expliquer la variation liée aux cas de COVID ; d'autres raisons pour laquelle une action varie ; chiffre d'affaire a augmenté de 20% , sorte de rattrapage, elle est sous évalué, en un jour elle prend 5%

- Influence du COVID sur la bourse, il pense qu'il serait plus intéressant de prendre plusieurs actions (+ micro ; Google, … pas utilisée un indice générale mais tu prends l'action d'une entreprise pour voir l'impact du covid sur la bourse)

- TSX60 … on peut partir en faisant notre analyse sur elle et confirmer en prenant quelques actions de bourse d'entreprise.

- Vérifier l'influence des tweet des Trump sur la bourse..

- Incertitude politique = fait le pont entre la covid et le marché boursier (définir une valeur de certitude ou de confiance dans la politique et de savoir si tu vas investir ajd ou pas et tester à prévoir les événements futurs). Si ça existe on pourrait faire ce pont là;

- Émettre des hypothèses et on analyse les données pour répondre à ces hypothèses (aussi des stat mais on doit les manipuler les données pour pouvoir répondre à cette hypothèse ou ces hypothèses).