Projet

Pour ce projet on récupère un dataset regroupant les jeux vidéos vendus sur PS4. On commence ainsi par importer les librairies ainsi que le dataset :

Entrée [426]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from scipy.stats import zscore
from sklearn.linear_model import LinearRegression
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

Entrée [408]:
```
#charger les données
#https://www.kaggle.com/sidtwr/videogames-sales-dataset
dataset = 'PS4_GamesSales.csv'
df = pd.read_csv(dataset, encoding = 'latin-1')
df.head()
```

Out[408]:

	Game	Year	Genre	Publisher	North America	Europe	Japan	Rest of World	Global
0	Grand Theft Auto V	2014.0	Action	Rockstar Games	6.06	9.71	0.60	3.02	19.39
1	Call of Duty: Black Ops 3	2015.0	Shooter	Activision	6.18	6.05	0.41	2.44	15.09
2	Red Dead Redemption 2	2018.0	Action-Adventure	Rockstar Games	5.26	6.21	0.21	2.26	13.94
3	Call of Duty: WWII	2017.0	Shooter	Activision	4.67	6.21	0.40	2.12	13.40
4	FIFA 18	2017.0	Sports	EA Sports	1.27	8.64	0.15	1.73	11.80

Une fois la librairie et le dataset importés, on commence le nettoyage du dataset :

Entrée [410]:
```
df.dtypes
```

Out[410]:
```
Game             object
Year            float64
Genre            object
Publisher        object
North America   float64
Europe          float64
Japan           float64
Rest of World   float64
Global          float64
dtype: object
```

Entrée [411]:
```
df.tail()
```

Out[411]:

	Game	Year	Genre	Publisher	North America	Europe	Japan	Rest of World	Global
1029	Fallen Legion: Flames of Rebellion	NaN	Role-Playing	NaN	0.0	0.0	0.0	0.0	0.0
1030	Radial G Racing Revolved	2017.0	Racing	Tammeka Games	0.0	0.0	0.0	0.0	0.0
1031	The Mummy Demastered	NaN	Action	NaN	0.0	0.0	0.0	0.0	0.0
1032	Project Nimbus: Code Mirai	NaN	Action	NaN	0.0	0.0	0.0	0.0	0.0
1033	Battle Chef Brigade	NaN	Action	NaN	0.0	0.0	0.0	0.0	0.0

On constate qu'il y a des jeux sans année et sans éditeur, on supprime donc les jeux du dataset où ces valeurs manquent

Entrée [412]:
```
df.isna().sum()
```

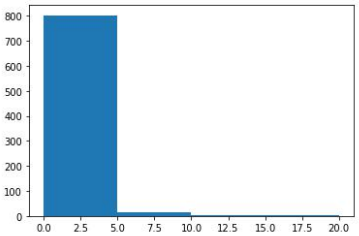
Out[412]:
```
Game              0
Year            209
Genre             0
Publisher       209
North America     0
Europe            0
```

En recoupant les différents éléments et en ayant clean le dataset on peut en découler cette **problématique** : Existe-t-il une région plus propice à la vente d'un jeu ?

Pour répondre à cette problématique on réalise différents graphiques pour observer le comportement de nos colonnes :

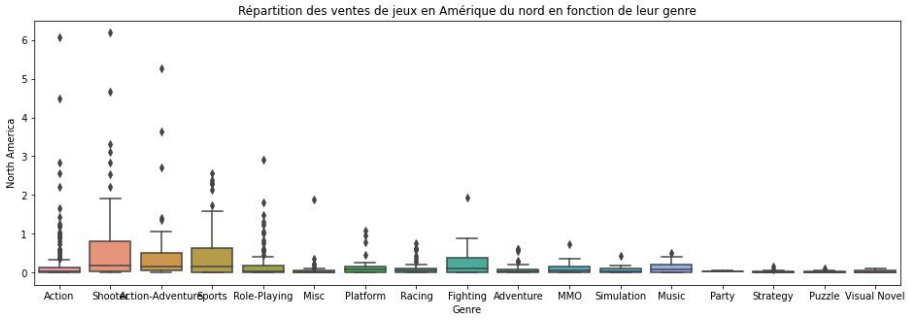
On constate via ce graphique que la majorité des jeux ont peu de ventes totales. Nous allons donc chercher à trouver et nous rapprocher de la catégorie de jeu qui se vendent le plus :

```
Entrée [362]: plt.hist(serie,
                       bins=bins);
```



On va donc chercher à trouver quels sont les caractéristiques des jeux qui fonctionnent dans les différents pays. Pour cela nous utilisons les boîtes à moustaches :

```
Entrée [449]: plt.figure(figsize=(16,5))
sns.boxplot(x='Genre',
            y='North America',
            data=df)
plt.title("Répartition des ventes de jeux en Amérique du nord en fonction de leur genre")
plt.show()
```

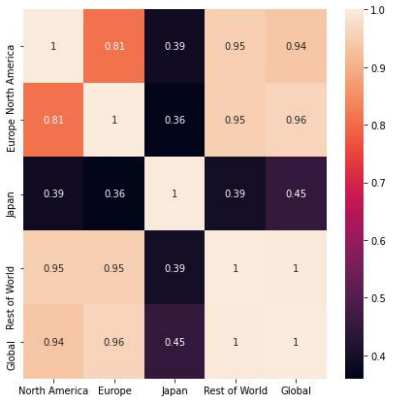


Cela nous permet de constater que certains genre de jeux ont plus de popularité selon certains pays.

Une fois que nous avons déterminé le genre nous pouvons analyser une heatmap pour voir quel corrélation existerait entre les ventes totales d'un jeu et les différentes région de distribution :

```
Entrée [454]: plt.figure(figsize=(7,7))
sns.heatmap(df.corr(), annot=True)
```

```
Out[454]: <AxesSubplot:>
```



Cette Heatmap nous permet d'émettre l'hypothèse qu'il existe une corrélation forte entre les ventes totales d'un jeu et le nombre de ventes en Europe et Amérique du Nord.

Pour déterminer si cette hypothèse est vrai nous allons pouvoir tester nos variables dans un algorithme qui nous servira à faire des prédictions et démontrer des corrélations.

```
Entrée [471]: X = df[['North America','Europe', 'Japan','Rest of World']]
              Y = df['Global']
```

Nous allons pouvoir prendre commes critères nos régions, qui nous serviront à savoir si celles ci ont une corrélation forte avec les ventes totales.

Comme attendu, d'après nos hypothèses, il existe une corrélation forte entre les ventes en Amérique du Nord et en Europe par rapport aux ventes totales, et celle-ci est plus faible pour le Japon et le reste du monde :

```
Entrée [479]: print('Coef : ',lin.coef_)
              print('Intercept : ',lin.intercept_)

              Coef :  [1.12821592 1.09334429 0.9975151  0.44392399]
              Intercept :  0.00020291092247370912
```