TP 5 Visualisation Avancée et Webscraping

TP Data Science M1IA

Ilyas Bambrik

2024

Table des matières

I - **Exercice : Visualisation et annotation** 3

II - **Exercice : Distributions** 6

III - **Exercice : WordCloud** 7

IV - **Exercice : Dashboards Interactives et Plotely** 8

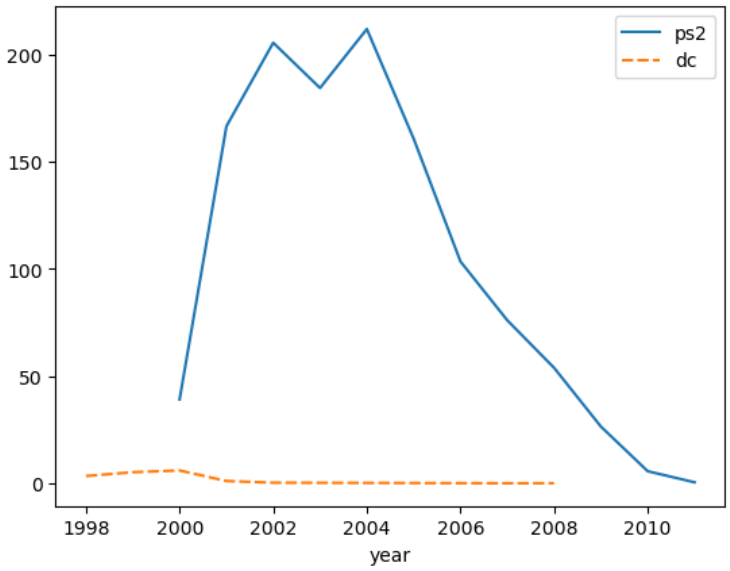
V - **Exercice : Webscraping** 11

Exercice : Visualisation et annotation

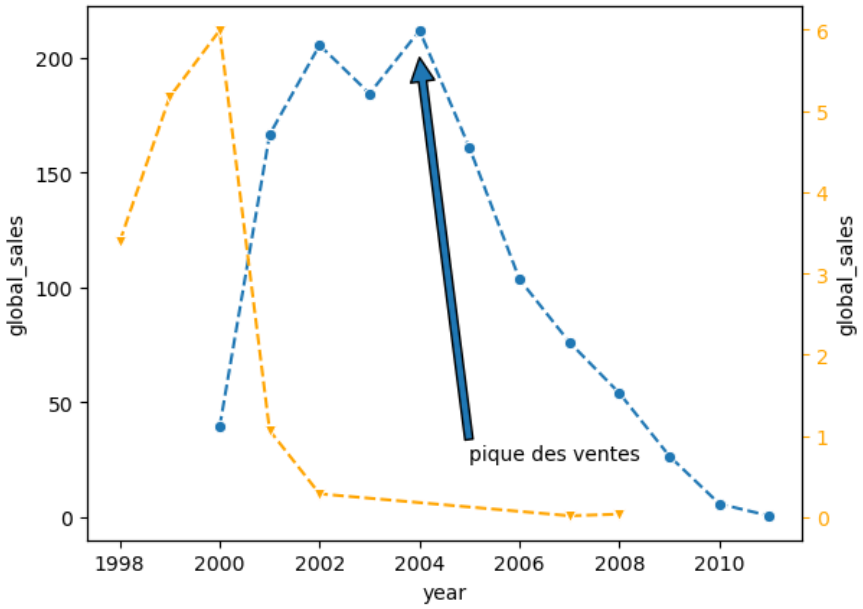
I

Question 1

- Lisez le DataSet *video_games_sales.csv*.
- Tracez un graphe *lineplot* pour les ventes accumulés sur *DC* (*Dreamcast*) et graphe *lineplot* pour les vente accumulés sur *PS3* dans la même axe en fonction de l'année.



- Refaite le même travail en utilisant *twinx* pour séparer l'axe des y des deux graphe et adapter l'échelle.
- Pour les deux graphes, utilisez deux couleurs et marqueurs (paramètres *marker* et *color*) distinctes.
- Utilisez un style de ligne discontinu (argument *linestyle="--"*).
- Utiliser la fonction *annotate* de l'objet axe pour annoter la valeur maximale du graphe des ventes PS3. Créez une flèche avec le paramètre *arrowprops*. Déplacez le texte de l'annotation en bas de la valeur annotée avec le paramètre *xytext*.



Question 2

- Tracez un barplot qui affiche pour chaque année les ventes PC, PS3 et Xbox360 (utilisez le paramètre *hue*).
- Tracez un box-plot des ventes au Japon pour les jeux GB (GameBoy) et NES (Super Nintendo).

Exercice : Distributions

II

Question

- Lisez *players.csv* avec pandas.
- Tracez la distribution des valeurs de la colonne *height_in_cm* avec *histplot*. Utilisez l'argument *kde* pour afficher la fonction de densité estimée.
- Tracez un *barplort* de la colonne *highest_market_value_in_eur* cumulée en fonction de la colonne *sub_position*. Élargissez la figure et faite une rotation 90° des étiquettes de l'axe des x afin que l'affichage soit claire.

Exercice : WordCloud

III

Question

- A partir du fichier *shakespare.txt,* créez un *wordcloud* simple. Les axes doivent être masqués de l'affichage avec un arrière plan blanc. Le *wordcloud* doit afficher seulement les 100 mots les plus importants du texte.
- Afficher le *wordcloud* à nouveau en ignorant les mots vides.
- Afficher le *wordcloud* à nouveau après avoir augmenté le nombre de termes affichés dans le *wordcloud* à 1000 et utilisez l'image *python-wordcloud.png* pour le masque et le schéma de coloriage.



Exercice : Dashboards Interactives et Plotely

IV

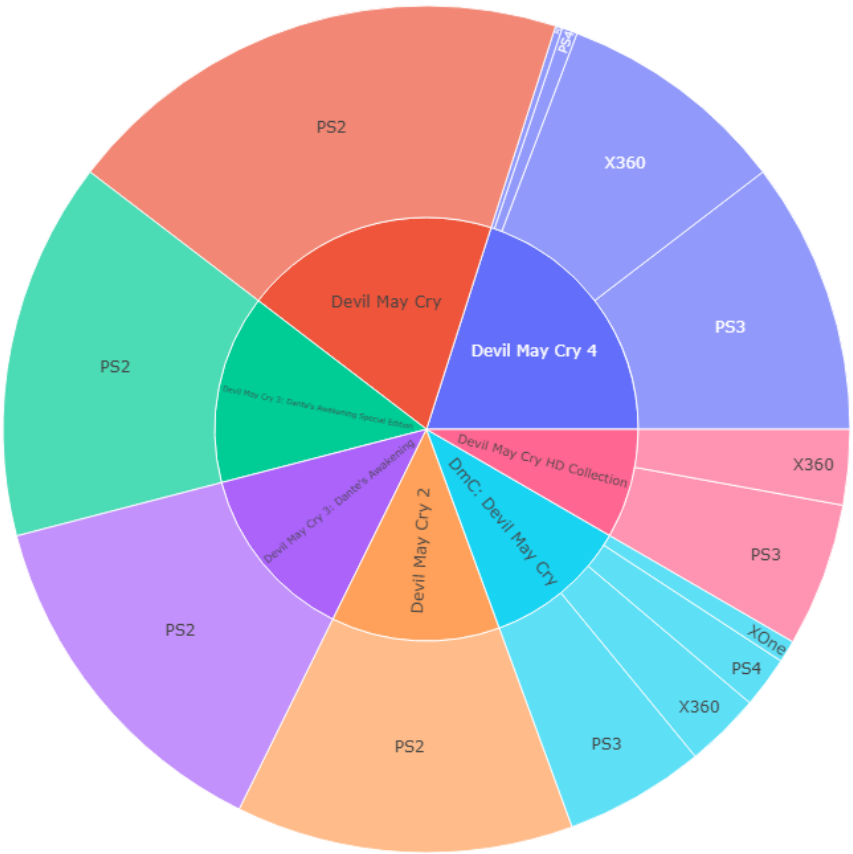
Lisez le DataSet *video_games_sales.csv*.

Question 1

Récupérez toutes les lignes où le nom du jeux contient la chaîne *Devil May Cry*.

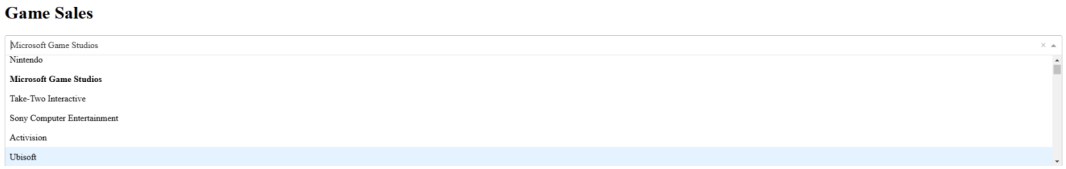
Question 2

A partir du résultat précédant, tracez un Sunburst plot divisant les ventes globales (*global_sales*) par le nom du jeux et puis par plate-forme.

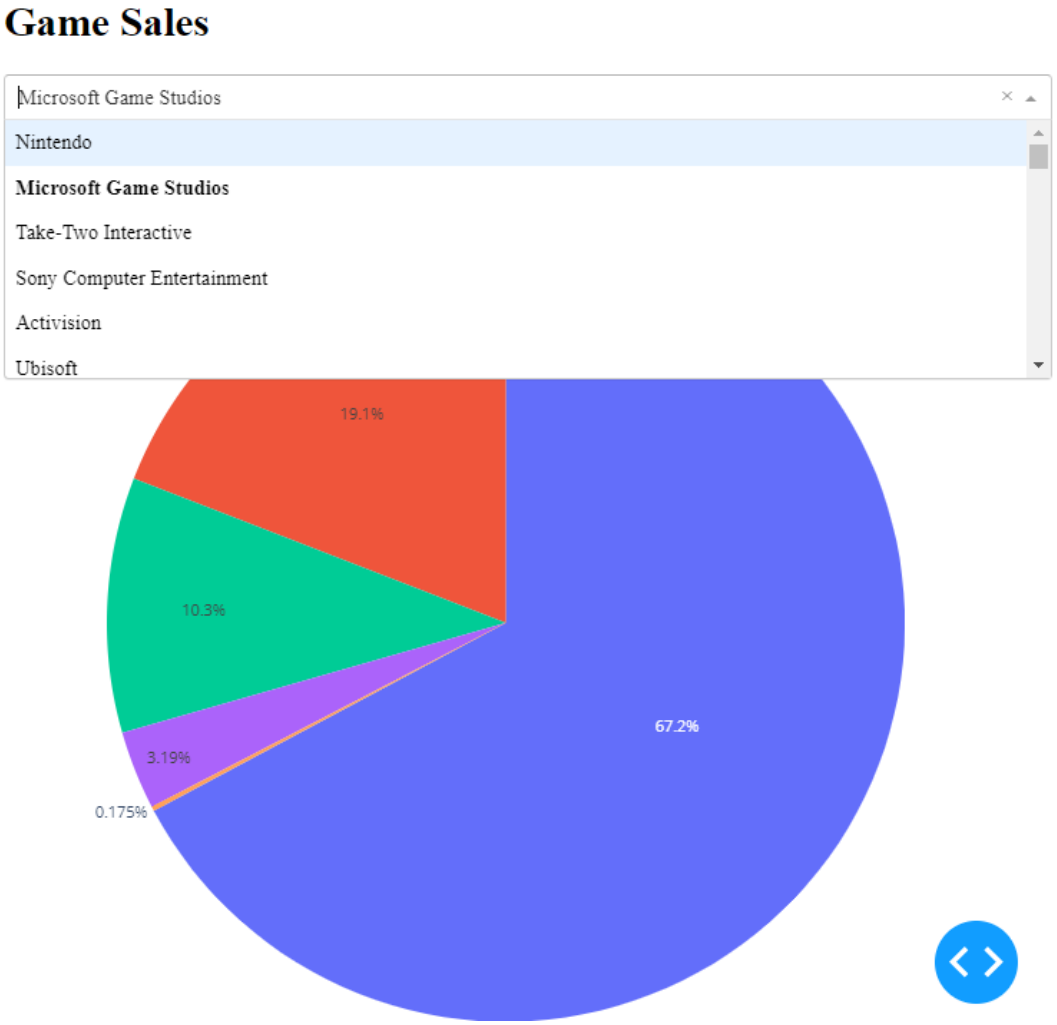


Question 3

Créez une application Dash qui contient un *Dropdown* contenant touts les noms des maisons de publication (*publisher*) présents dans *video_games_sales.csv*. Définissez l'identifiant de ce composant comme *game-selector*.

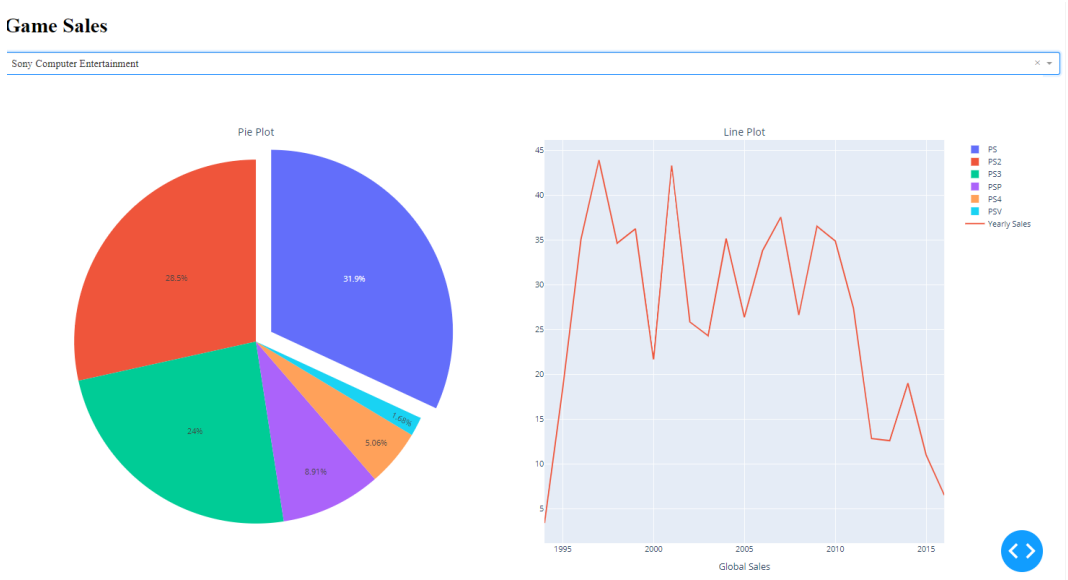


- Créer un composant *dcc.Graph* avec l'identifiant *sales-shares*.
- Définissez un *callback* dans votre application Dash qui filtre seulement les lignes associées *publisher* sélectionné *game-selector*.
- La fonction *callback* doit retourner un Pie plot représentant les proportions des ventes par plate-forme pour la maison de publication sélectionnée.
- Affich



Question 4

- Modifiez le code de l'application *Dash* afin d'afficher un deuxième graphique linéaire représentant l'évolution des ventes (*global_sales*)au fils des années (*year*) pour la maison de publication sélectionné (*publisher*). Vous devez utiliser le même composant identifié par *sales-shares* sans ajouter un autre *dcc. Graph* .
- Dans le premier Pie plot, la plus grande proportion (*platform*) doit être détachée du Pie plot par une distance de 0.1.



Exercice : Webscraping

V

Question

- Écrivez un programme avec le module *requests* pour récupérez la liste des université sous forme Json d'un pays donné. (par exemple http://universities.hipolabs.com/search?country=Algeria retourne la liste des université algériennes)
- Écrivez un programme avec les modules *BeautifulSoup* et *requets* afin de récupérer les prix et les noms des PC portables présents dans la page https://www.licbplus.com.dz/global-category/laptop-computers/3/225?sort_by=new