

Spécifications fonctionnelles détaillées


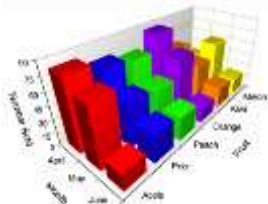
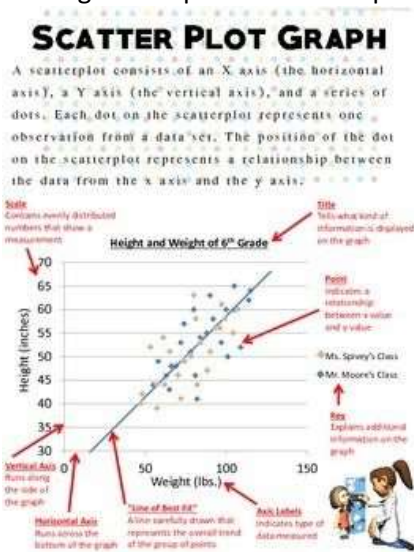
PLAN :

1. Scrapping et constitution du jeu de données
2. Classification

SPECIFICATIONS :

1. Scrapping et constitution du jeu de données

| Date MAJ | Spécifications |
|----------|---|
| 12/02 | <p>Script de scrapping :</p> <p>Le scrapping des charts se base sur un script qui récupère les résultats de google image pour une requête. Ce script est disponible sur le GitHub au nom de google-scrapper_2.0.py.</p> <p>Nous l'avons obtenu en faisant un tour d'horizon de l'existant, avec comme critère de trouver un script qui permette de scrapper à la fois une image et un json avec des descripteurs. Nous avons trouvé un premier script mais qui avait le défaut de n'extraire qu'un nombre limité d'image. En cherchant plus avant, nous avons trouvé un nouveau script qui a permis de surmonter ce bridage. Nous l'avons adapté et modifié selon nos besoins, notamment en termes de nommage des fichiers sortants et de facilitation du procédé de lancement de requête afin de permettre une meilleure parallélisation des tâches au sein du groupe.</p> <p>Le script prend en entrée le terme de la recherche google image à effectuer et renvoie tous les résultats disponibles sur la page jusqu'à ce que plus aucun résultat ne s'affiche. Par résultat, on entend une image et un fichier json dont le nommage répond à la règle suivante : requête_numérotation.format <u>Exemple :</u> line_chart_30.jpg bar_chart_5.json</p> <p>Le script utilise une méthode de scrolling avec un driver adapté pour chrome et des paramètres de gestion du temps (afin de ne pas être repéré comme un bot.)</p> <p>Nous avons dans un premier temps lancé trois requêtes distinctes correspondant aux trois types de graphiques sur lesquelles nous allons nous concentrer :</p> <ul style="list-style-type: none">- Line Chart (requête : line_chart)- Bar Chart (requête : bar_chart)- Scatter Plot (requête : scatter plot) |

| | |
|-------|---|
| | <p>Les résultats se trouvent dans un dossier situé dans le répertoire du scrapper, dont le nom est « dataset ». Chaque sous-dossier de « dataset » correspond au résultat d'une requête.</p> |
| 12/02 | <p>Nettoyage du jeu de donnée :</p> <p>Une fois les trois requêtes lancées et les dossiers de données brutes constituées, nous avons effectué une première phase de tri manuel avec comme objectif d'obtenir 230 images propres et utilisables.</p> <p>Voici les critères que nous avons utilisé pour exclure les images non conformes. Ils sont issus de l'expérience. Cette liste n'est pas exhaustive mais correspond au cas qui reviennent le plus souvent.</p> <p>Pour les trois catégories (line chart, bar chart, scatter plot) :</p> <ul style="list-style-type: none"> - Fond hétérogène ne permettant pas un contraste clair  <ul style="list-style-type: none"> - Graphique en 3 dimensions :  <ul style="list-style-type: none"> - Légende représentant une part trop importante de l'image  |

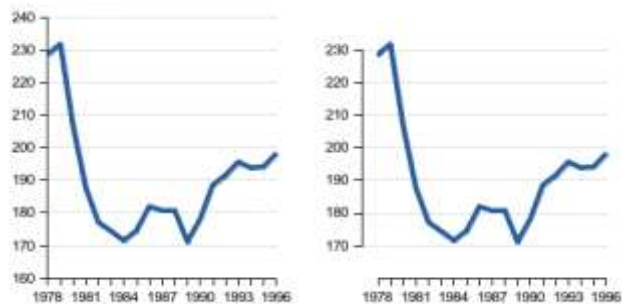
- Graphique regroupant 2 catégories (ex : à la fois bar et line chart)



- Graphique trop simpliste (type icône ou vecteur)



- Image comportant plusieurs graphiques :

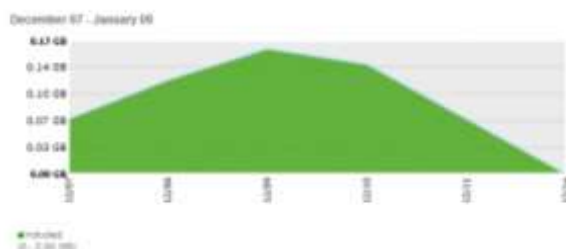


- Graphique tracé manuellement :

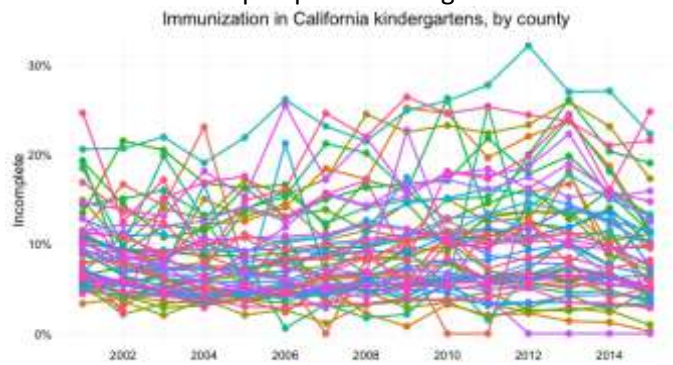


Critères spécifiques aux line Charts :

- Zone pleine colorée en dessous de la ligne :



- Nombre trop important de lignes :



Critères spécifiques aux Scatter Plot :

- Fond de cartes :



Critères spécifiques aux Bar Chart :

2. Classification