Consignes projet statistique - Baptiste Gautier

Voici les consignes attendues pour le projet de statistique de mon côté. Marc-Arthur vous joindra les consignes et les données pour sa partie du projet. La soutenance finale sera une présentation des deux projets. A vous de voir si vous voulez traiter les deux séparément ou faire des liens entre les deux sujets.

Lien des données

Vous trouverez le lien des données ici:
https://explore.data.gouv.fr/fr/datasets/586dae65a3a7290df6f4be90/?Code%20de%20la%20discipline__exact=disc06&Ann%C3%A9e__exact=2020#/resources/c7c9642b-9fa1-40a0-83d5-1615c15b4178

Il s'agit de données concernant le taux d'insertion des étudiants diplômés du supérieur.

Vous pouvez les télécharger en csv en cliquant sur le lien en haut à droite. Ensuite vous pouvez les charger sur python avec pandas comme ceci:

Python

```python
import pandas
from pandas import DataFrame


def load_data(path: str = "../../data/database.csv") -> DataFrame:
    """
    Load data from a CSV file with pandas.

    Parameters
    ----------
    path : str
        The path to the CSV file.

    Returns
    -------
    DataFrame
        The data loaded from the CSV file.
    """
    return pandas.read_csv(path, delimiter=";")
```

Ou avec polars comme cela:

```python
import polars as pl

def load_data(path: str = "../../data/database.csv") -> pl.DataFrame:
    """
    Load data from a CSV file using Polars.

    Parameters
    ----------
    path : str
        The path to the CSV file.

    Returns
    -------
    pl.DataFrame
        The data loaded from the CSV file.
    """
    return pl.read_csv(path, sep=";")
```

Ensuite je vous laisse opérer de votre magie !

Attendus

L'objectif du projet est d'expliquer la colonne du **Taux d'insertion**. Vous pouvez pour cela poser un modèle explicatif, faire des tests statistiques, …

Vous constituerez pour ce travail des groupes de **4 maximum** (même groupe pour les deux projets).

Rendu

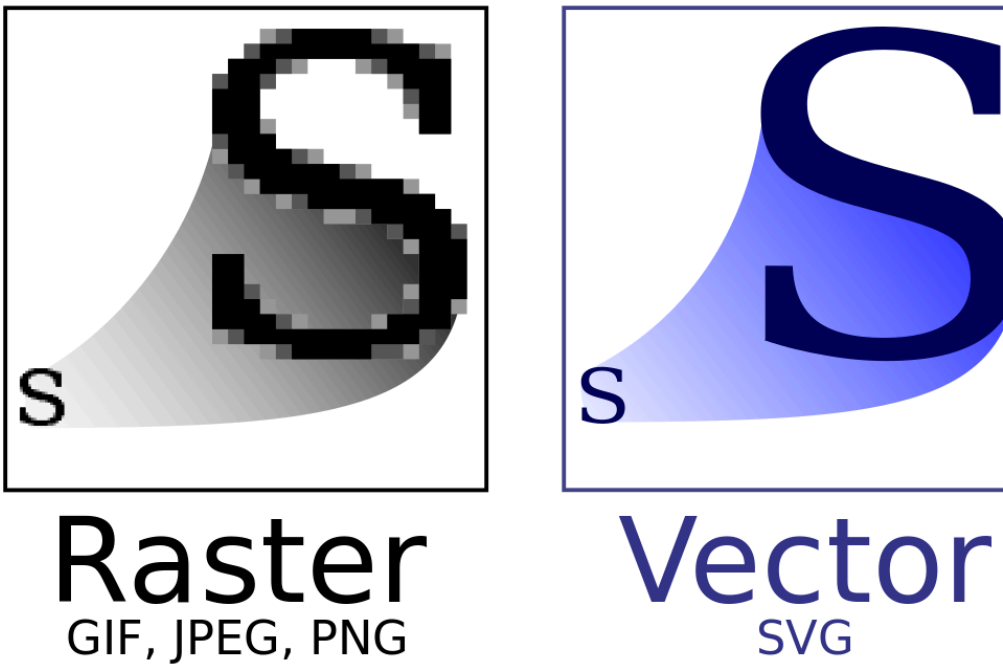
Le rendu se scinde en deux parties:

- **Le code** (notebook ou repo git) à déposer le **12 janvier** sur la plateforme**.**

- **La soutenance.** Elle durera **9 minutes** (et 5 minutes de questions/réponses) où vous présenterez vos résultats. Notez bien que les **9 minutes sont pour la présentation des deux parties**: celles de Marc-Arthur et la mienne. Vous n'êtes évidemment pas obligé de présenter les données en détail et pourrez vous concentrer directement sur la présentation de vos analyses. **Les soutenances auront lieu le 13 et 14 janvier**.

Tips pour préparer vos slides:

Pour extraire des images de votre code python sur un powerpoint, vous pouvez les sauvegarder en format svg.



Étirer un png donnera un rendu pixelisé alors qu'un svg supporte naturellement les variations de taille. Attention toutefois, les svg ne sont pas lus par tous les supports. En python:

```python
def ma_fonction_graphique(path_to_save: str = "ma_figure.svg") -> None:
      ... # mon plot ici
      plt.savefig(path_to_save, bbox_inches="tight")
```

Plus d'infos:
https://chatgpt.com/share/6754006e-0708-800c-a718-56d226238164

Grille d’évaluation

La grille d’évaluation pour ma partie du projet est la suivante (la note obtenue sera moyennée avec la note de la partie de Marc-Arthur pour obtenir votre note finale):

Description du critère	Nombre de points
Vous avez produit un modèle explicatif du taux d’insertion. Ma base est un R2 de 0.65, en utilisant les variables simples à traiter du tableau. Evidemment, si vous avez une moins bonne R2 mais une explication correcte de vos résultats et des analyses pertinentes, vous aurez tous les points.	10
Vous avez rendu votre présentation intéressante. Il s’agit d’évaluer ici votre capacité à transmettre des informations techniques en les rendant intelligibles et captivantes.	5
Travail complémentaire / Ouverture. Il s’agit de proposer une analyse un peu plus poussée que ce que l’on a vu ensemble. Cela peut être une méthode d’analyse, une technique de prévision, un GLM, …, ou simplement l’exploitation des données “one-hot” (les colonnes à droite) un peu plus compliquées à exploiter que les données à gauche des données.	5

Annexe

Voici le graphique des résultats de ma baseline (régression linéaire en prenant en compte les données les plus faciles à exploiter):