

Projet Data Visualisation

CAPEL Alexandre

2025-10-20

Consignes

Dans ce projet, vous devez effectuer une analyse statistique plus ou moins complète sur un jeu de données (proposé ou de votre choix; dans le second cas choisissez un jeu de données avec au moins cinq six variables).

Le projet est séparé en deux parties :

- la première partie consiste en la rédaction d'un rapport détaillé avec .Rmd contenant une description de votre jeu de données, les démarches statistiques, les graphiques et les **potentielles** (j'insiste sur le potentiel, c'est pas grave si il n'y en a pas) conclusions que vous pouvez en tirer.
- la seconde partie correspond au développement d'une application à l'aide de R-shiny. Elle devrait me permettre de naviguer de manière interactive à travers l'ensemble de vos graphes et devra proposer une interface soignée. Soyez créatifs en développant cette application, ça ne sera que valorisé !

Barême

Rapport statistique - 10 points

En ce qui concerne la rédaction du rapport vous serez notés sur :

- la structure du rapport
- les graphiques (purement esthétique mais je serez sévère dans le cas d'oubli de titre etc...)
- la pertinence des descriptions statistiques pour répondre au problème posé.
- la beauté finale du rapport (titres, utilisation des outils markdown pour rendre la lecture agréable).

L'application - 10 points

Pour ce qui est de l'application, une application non fonctionnelle sera un 0 (malheureusement) instantané pour cette partie. Veillez vraiment à ce que l'application fonctionne avant de me rendre vos projets.

Ici vous serez noté sur :

- l'interface de l'application.
- les bugs potentiels.
- la qualité des graphiques (si il y en a en plus comparé à votre rapport).
- les codes utilisés.

Rendu

Le projet se fera par groupe de 3 ou 4 donc trouvez la bonne équipe afin de bien travailler ensemble ;).

Il faudra me rendre votre rapport dans fichier `.html` (ou `.pdf` au choix). Dans le rapport je ne dois voir aucun code sauf dans un cas exceptionnel où vous créez des fonctions pour faire parler les données. Dans ce cas là, je vous prierai de mettre ces codes là dans une section *Annexe*.

Enfin, pour l'application vous avez le choix entre :

- m'envoyer un seul fichier `.R` qui permet une compilation autonome de l'application (contenant les ui et server).
- m'envoyer le lien vers un dépôt distant (GitHub, je n'ai pas accès au GitLab Polytech...) que je clonerai et essaierai de faire fonctionner dans mon dépôt local (dans ce cas là, n'oubliez pas de rédiger un tutoriel dans le `README` afin que je puisse faire fonctionner l'application comme vous le voulez).

Le projet sera à rendre avant le 20 novembre 2025 en envoyant un mail à l'adresse suivante : `alexandre.capel@umontpellier.fr` (sans le **etu** !!!). Tout retard entraînera le retrait d'un point par jour.

Jeux de données proposés

Tous les jeux de données sont disponibles sur le Moodle du cours (ou sur la page GitHub). S'il vous plaît, un seul jeu de données par groupe.

1 - Les voyages en taxi à New York

Téléchargez le fichier `taxi.csv`. Il s'agit d'un jeu de données avec 19 variables et décrivant un million de courses pendant le mois de janvier 2015. Le jeu de données comporte des données sur les courses de taxi dans la ville de New York. Elle décrit :

- la distance parcouru.
- les moyens de paiement.
- l'heure de prise en charge et l'heure de dépôt.
- nombre de passager etc...

Table 1: Première lignes du jeu de données `taxi.csv` (sous ensemble des variables)

VendorID	Date	Pickup_time	passenger_count	trip_distance	fare_amount
2	2015-01-21	09:04:52	1	0.88	5.0
1	2015-01-05	17:46:19	1	0.60	4.5
1	2015-01-10	22:22:05	2	9.20	30.5
2	2015-01-05	13:03:55	1	2.31	9.5
1	2015-01-21	01:17:20	1	1.60	6.5
1	2015-01-17	22:03:26	1	0.60	4.0

Piste de réflexion. On pourrait se demander ce qui influe sur les prix des courses (localisations, temps, heure de pointe) par exemple.

Librairie utile : `lubridate` pour les données temporelles.

2 - Les AirBnb à Seattle

Téléchargez le fichier `seattle.csv`. Il s'agit d'un jeu de variables possédant 20 variables et quelques milliers d'individus. Ce jeu de données décrit les visites de plusieurs logement Airbnb situé à Seattle (et ses alentours) avec quelques descriptions techniques :

- nombre de chambre.
- salle de bain.
- prix de l'appartement.
- localisation dans la ville etc...

Table 2: Première lignes du jeu de données `seattle.csv` (sous ensemble des variables)

room_id	room_type	address	price	bathrooms
2318	Entire home/apt	Seattle, WA, United States	250	2.5
3335	Entire home/apt	Seattle, WA, United States	100	1.0
4291	Private room	Seattle, WA, United States	82	1.0
5682	Entire home/apt	Seattle, WA, United States	49	1.0
6606	Entire home/apt	Seattle, WA, United States	90	1.0
9419	Private room	Seattle, WA, United States	65	3.0

Piste de réflexion. Quels facteurs influencent le prix d'une location Airbnb ?

3 - Données sur les passagers du Titanic

Téléchargez le fichier `titanic.csv`. Il s'agit d'un jeu de données avec 13 variables et presque un millier d'individus. Il répertorie les passagers du Titanic avec toutes les informations utiles :

- si ils ont survécu ou non.
- le prix du ticket.
- le sexe.
- l'âge etc...

Table 3: Première lignes du jeu de données `titanic.csv` (sous ensemble des variables)

Name	Sex	Pclass	Survived	Fare
Braund, Mr. Owen Harris	male	3	0	7.2500
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	1	1	71.2833
Heikkinen, Miss. Laina	female	3	1	7.9250
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	1	1	53.1000
Allen, Mr. William Henry	male	3	0	8.0500
Moran, Mr. James	male	3	0	8.4583

Piste de réflexion. Comprendre ce qui favoriserait le fait qu'un passager survive ou non.

4 - Les logements dakarois

Téléchargez le fichier `dakar.csv`. Il s'agit d'un jeu de données avec 29 variables et une quantaine d'individus correspondant à des logements dans la capitale du Sénégal. Pour chaque individu, on a une liste de variable décrivant le logement :

- loyer.
- surface du terrain.
- distance avec le bord de mer (sous forme qualitative).
- nombre de pièce etc...

Table 4: Première lignes du jeu de données `dakar.csv` (sous ensemble des variables)

id	Loyer	SurfTerrain	Type	Egout	Quartier
1	850	598	Villa	Non	Mamelles
2	590	95	Appart	Oui	Plateau
3	200	90	Appart	Oui	Mermoz
4	970	396	Villa	Non	Mamelles
5	600	589	Villa	Oui	PointE
6	90	70	Appart	Oui	Medina

Piste de réflexion. On aimerait comprendre les facteurs qui rentrent en compte dans le prix du loyer.

5 - Étude sur les maladie cardiaque

Téléchargez le fichier `heart.csv`. Il s'agit d'un jeu de données contenant 19 variables pour plusieurs centaines de milliers d'individus. Chaque individu est un patient atteint (ou non) d'une maladie cardiaque avec plusieurs variables le décrivant :

- son âge.
- si il est diabétique ou non.
- la temps de sommeil.
- si il est atteit d'un cancer de la peau etc...

Table 5: Première lignes du jeu de données heart.csv (sous ensemble des variables)

HeartDisease	Smoking	Sex	AgeCategory	Asthma
No	Yes	Female	55-59	Yes
No	No	Female	80 or older	No
No	Yes	Male	65-69	Yes
No	No	Female	75-79	No
No	No	Female	40-44	No
Yes	Yes	Female	75-79	No

Piste de réflexion. L'objectif pourrait être de comprendre quels facteurs pourraient emmener à contracter une maladie cardiaque.

6 - Le votre

Si les jeux de données ne sont pas à votre goûts (ce que je peux comprendre) libre à vous d'en trouver un et de me le proposer. Dans ce cas là, prévenez moi rapidement pour que je puisse le regarder et le valider.