Rendu : Python pour la data

1. Choix du dataset :

Pour ce rendu de projet, j'ai choisi le dataset "Titanic" provenant de Kaggle. Il contient les informations sur 891 passagers du Titanic avec à la fois des valeurs numériques (age et tarif du billet) mais aussi catégorielles (sexe, classe, survie et port d'embarquement).

Ce dataset est idéal pour un rendu de projet d'analyse de données car il permet de réaliser :

- Des analyses statistiques descriptives.
- Des visualisations variées.
- Des analyses par catégorie.

2. Étape d'analyse :

2.1 Chargement et exploration :

Le dataset a été chargé avec "Pandas". J'ai commencé par observer les colonnes, les types de données, les valeurs manquantes et la présence de doublons.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
720	721	1	2	Harper, Miss. Annie Jessie "Nina"	female	6.0	0	1	248727	33.0	NaN	S

2.2 Nettoyage :

J'ai commencé par supprimer les doublons (drop_duplicates). Ensuite, j'ai remplacé les valeurs manquantes par :

- Âge : Remplacer par la médiane afin de garder une valeur réaliste sans supprimer de lignes.
- Embarked : Remplacer par la valeur la plus fréquente.
- Cabin : Remplacer par "Unknow" car il y avait trop de valeurs manquantes.

J'ai aussi converti les colonnes "Sex", "Pclass", "Embarked" et "Survived" en type "category" et supprimé les colonnes "PassengerId", "Name" et "Ticket" afin de simplifier l'analyse.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
526	1	2	female	50.0	0	0	10.5	Unknown	S

2.3 Analyses statistiques :

J’ai commencé par faire un descriptif (médiane, moyenne, écart type, valeur plus petite et valeur plus grande) des colonnes avec le type number c'est-à-dire les colonnes “Age”, “Fare”, “SibSp”, “Parch”.

	Age	SibSp	Parch	Fare
mean	29.361582	0.523008	0.381594	32.204208
median	28.000000	0.000000	0.000000	14.454200
std	13.019697	1.102743	0.806057	49.693429
min	0.420000	0.000000	0.000000	0.000000
max	80.000000	8.000000	6.000000	512.329200

Ensuite, j’ai calculé les corrélations entre les différentes colonnes afin d’identifier les relations entre les variables numériques.

	Age	SibSp	Parch	Fare
Age	1.000000	-0.233296	-0.172482	0.096688
SibSp	-0.233296	1.000000	0.414838	0.159651
Parch	-0.172482	0.414838	1.000000	0.216225
Fare	0.096688	0.159651	0.216225	1.000000

Et enfin, j’ai calculé la moyenne d'âge par sexe, on peut noter que la moyenne d’age des femmes est de 28 ans tandis que celle des hommes est de 30 ans. Mais aussi, j’ai calculé la moyenne du prix de classe, on peut donc noter que la moyenne de la première classe est de 80$, la deuxième classe est 20$ et la troisième classe 13$.

3. Résultats principaux :

3.1 Statistiques descriptives :

On observe que l'âge moyen des passagers est de 30 ans tandis que la médiane est de 28 ans et l’écart type de 14 ans.

On note que pour les tarifs des billets, le tarif moyen est de 32$ tandis que la valeur la plus élevée dépasse les 500$.

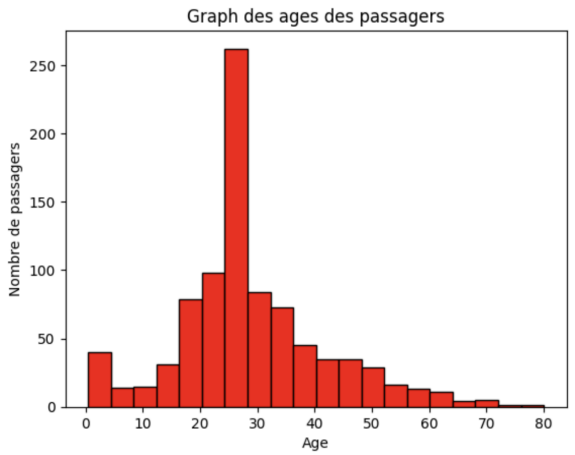
3.1 Corrélations :

“Fare” est fortement corrélé avec “Pclass” : les passagers de la première classe payaient beaucoup plus cher.

“SibSp” et “Parch” sont légèrement corrélés, ce qui reflète la présence de famille à bord.

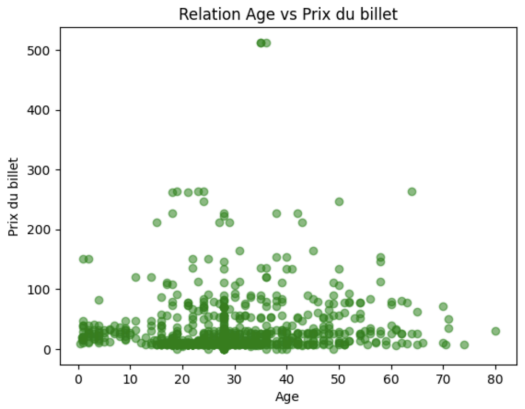
4. Visualisations :

1. Histogramme :



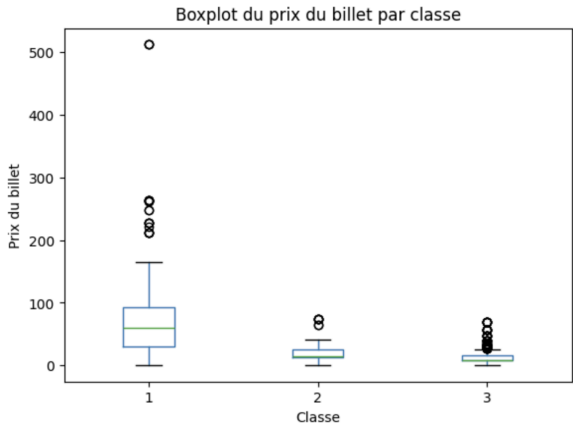
On observe que l'histogramme des âges démontre que la majorité des passagers avaient entre 20 et 40 ans.

2. Boxplot :



Ce graphique montre clairement que la première classe avait des billets plus chers et de valeurs extrêmes.

3. Scatter plot :



La majorité des passagers avaient pris la troisième classe mais aussi que le nombre de passagers de la première classe est supérieur à celui de la deuxième classe.

5. Conclusion :

Ce projet m’a permis de redécouvrir le processus d’analyse de données : nettoyage, exploration, statistiques et visualisation. Le dataset “Titanic“ offre un exemple correct pour comprendre comment les variables numériques et catégorielles interagissent et comment tirer des conclusions à partir de données réelles.