

Tutoriel avec exercices : Analyse des données Titanic en Python

Votre Nom

December 19, 2024

Ce tutoriel vous guide pas à pas pour analyser la célèbre base de données **Titanic** en Python, en utilisant la bibliothèque **pandas** pour manipuler les données et **matplotlib** pour les visualiser.

Étapes du tutoriel

1. Chargement des données

Utilisez le fichier `data1.csv` contenant les informations des passagers du Titanic.

```
1 import pandas as pd
2 from matplotlib import pyplot as plt
3
4 # Charger les données
5 df = pd.read_csv('data1.csv')
6
7 # Afficher un aperçu du tableau
8 print(df.to_string())
```

Question 1 : Combien de colonnes contiennent des données numériques ? Utilisez la méthode `df.info()` pour répondre.

2. Gestion des données manquantes

Pour traiter les colonnes ayant des valeurs manquantes :

```
1 # Supprimer les colonnes contenant des valeurs manquantes
2 df_cleaned = df.dropna(axis=1)
3 print(df_cleaned.info())
```

```

4
5 # Afficher un échantillon aléatoire de 10 lignes
6 print(df_cleaned.sample(10).to_string())

```

Question 2 : Quelle méthode pouvez-vous utiliser pour supprimer uniquement les lignes contenant des valeurs manquantes ? Affichez le tableau résultant.

3. Analyse exploratoire des données

1. Pour sélectionner les colonnes `age` et `survived` pour une analyse spécifique :

```

1 df_subset = df[['age', 'survived']]
2 print(df_subset.head())

```

Question 3 : Sélectionnez les colonnes `age`, `survived` et `fare` puis affichez les statistiques descriptives (`mean`, `min`, `max`) pour ces colonnes. Quelle est la valeur moyenne de l'âge des passagers ?

2. Visualisez la relation entre l'âge et la classe en colorant par survie :

```

1 plt.scatter(df['age'], df['pclass'], c=df['
    survived'])
2 plt.xlabel('Âge')
3 plt.ylabel('Classe')
4 plt.title('Âge vs Classe coloré par survie')
5 plt.show()

```

Question 4 : Filtrez les passagers ayant plus de 30 ans et appartenant à la première classe, puis affichez leurs informations.

4. Comparaison entre hommes et femmes

```

1 # Hommes
2 df_male = df[df['sex'] == 'male']
3 plt.subplot(1, 2, 1)
4 plt.scatter(df_male['age'], df_male['pclass'], c=df_male[
    'survived'], s=4)
5 plt.title('Hommes')
6
7 # Femmes
8 df_female = df[df['sex'] == 'female']

```

```

9 plt.subplot(1, 2, 2)
10 plt.scatter(df_female['age'], df_female['pclass'], c=df_
    female['survived'], s=4)
11 plt.title('Femmes')
12
13 plt.show()

```

Question 5 : Calculez le nombre total de passagers pour chaque sexe et chaque classe. Affichez les résultats dans un tableau.

5. Moyenne des taux de survie par catégorie

Regroupez les données par sexe et classe pour calculer la survie moyenne :

```

1 print(df.groupby(['sex', 'pclass'])['survived'].mean())

```

Question 6 : Parmi les femmes de la première classe, quel est le pourcentage de survie ?

6. Bonus : Analyse des tarifs

1. Sélectionnez les passagers ayant payé un tarif inférieur à la moyenne :

```

1 mean_fare = df['fare'].mean()
2 df_low_fare = df[df['fare'] < mean_fare]

```

2. Visualisez l'âge en fonction de la classe pour ces passagers :

Question 7 : Représentez par un nuage de points l'âge en fonction de la classe pour ces passagers.

3. Affichez un histogramme des tarifs :

```

1 plt.hist(df['fare'], bins=10)
2 plt.xlabel('Tarifs')
3 plt.ylabel('Fréquence')
4 plt.title('Distribution des tarifs')
5 plt.show()

```

Question 8 : Affichez la répartition des passagers ayant payé moins de 50 unités monétaires en fonction de leur classe. Quelle classe a la majorité ?