EDA TP4

April 22, 2024

1 EDA TP4

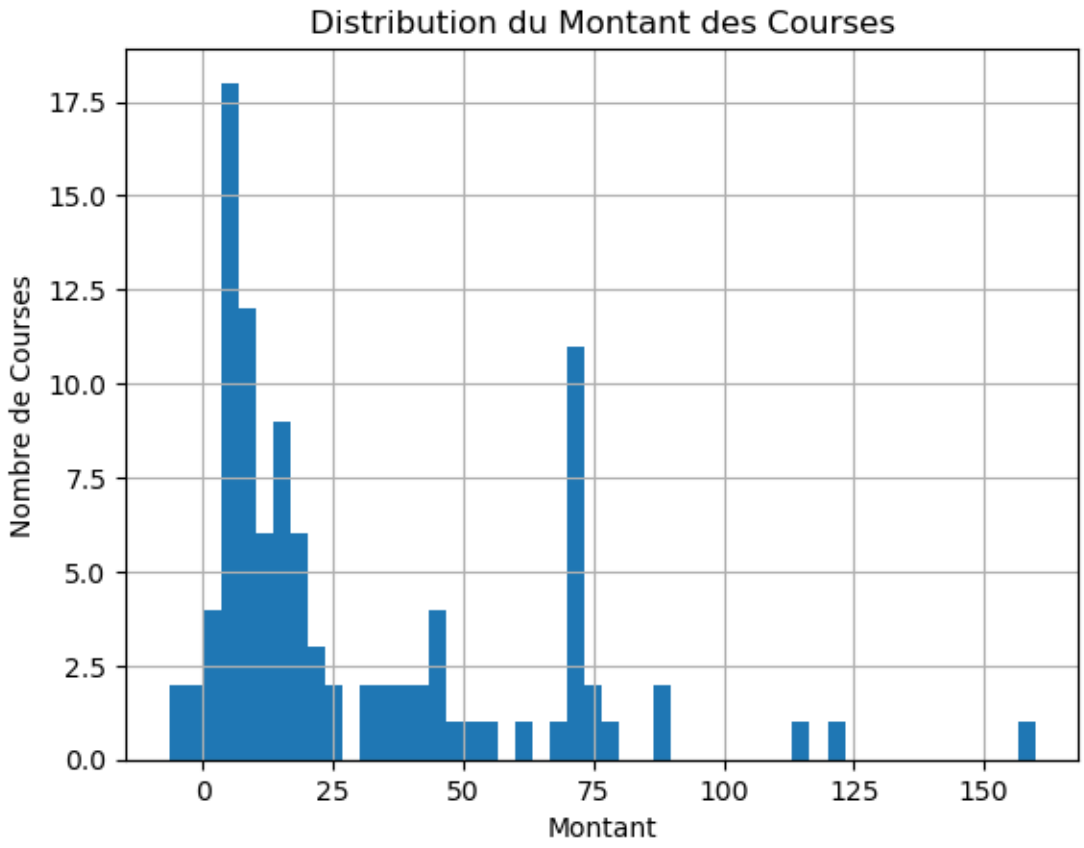
LAPUJADE Quentin, ROUARD Ulysse, NOVINC Alexis

[15]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data-1713136906221.csv')
```

1.1 1. Histogramme du Montant des Courses :

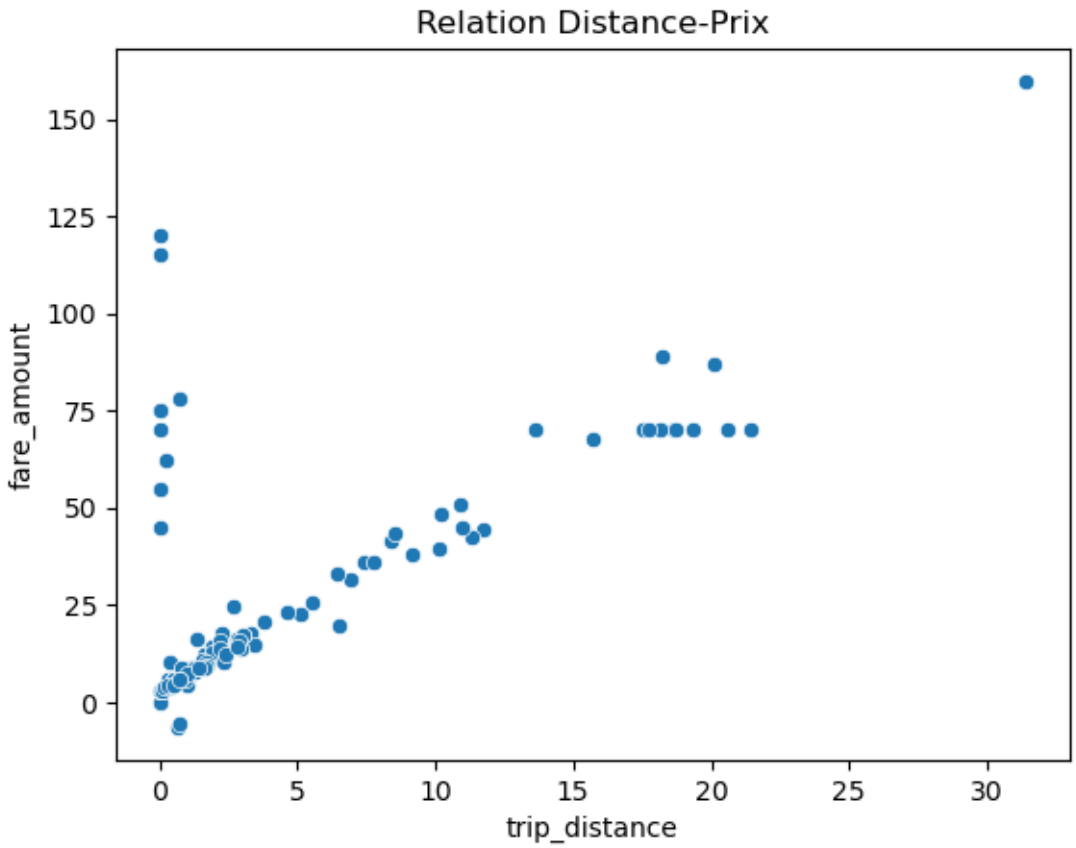
[16]:
```
df['fare_amount'].hist(bins=50)
plt.title('Distribution du Montant des Courses')
plt.xlabel('Montant')
plt.ylabel('Nombre de Courses')
plt.show()
```



Ce graphique présente la distribution des montants des courses. Un histogramme est utilisé avec 50 intervalles (bins), ce qui permet de voir la fréquence des courses pour différents montants. L'axe des abscisses montre les montants de la course, et l'axe des ordonnées montre le nombre de courses correspondantes. Ce graphique est utile pour identifier la gamme de tarifs les plus courants ainsi que pour détecter des anomalies telles que des montants exceptionnellement élevés ou bas.

1.2 2. Graphique de Dispersion Distance-Prix :

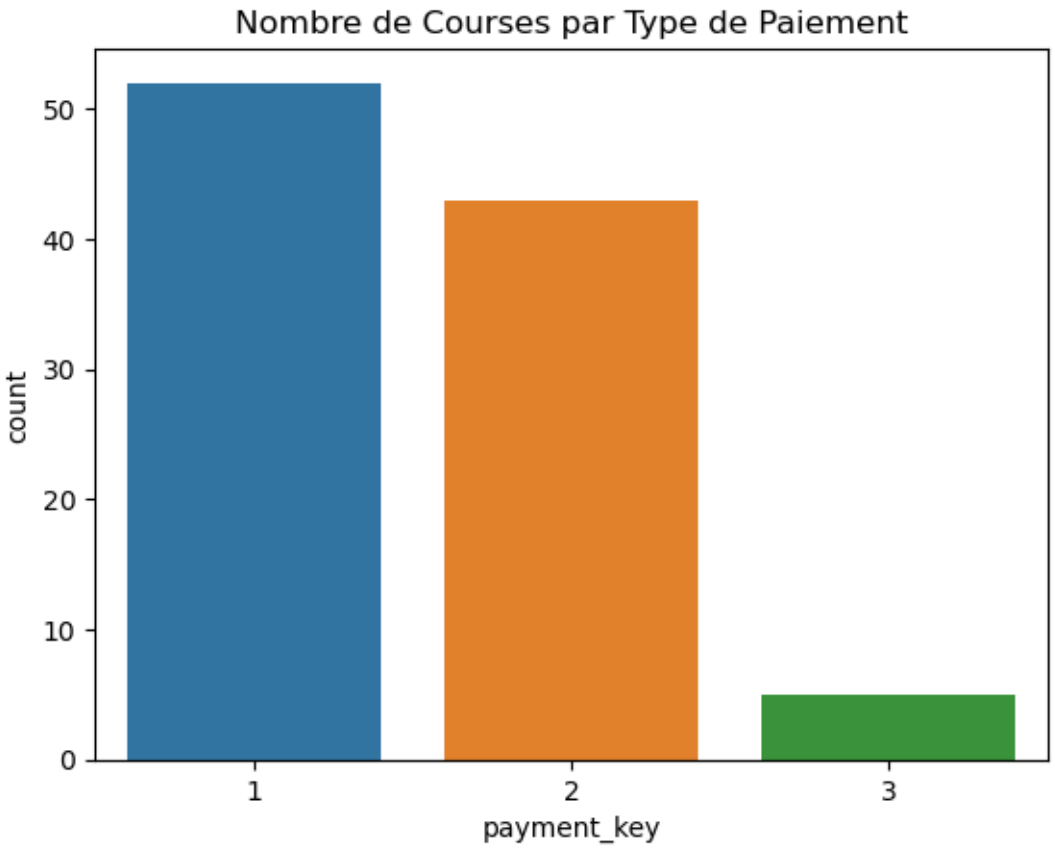
[17]:
```
sns.scatterplot(x='trip_distance', y='fare_amount', data=df)
plt.title('Relation Distance-Prix')
plt.show()
```



Ce graphique en nuage de points montre la relation entre la distance parcourue et le montant payé pour la course. L'axe des abscisses représente la distance de la course et l'axe des ordonnées représente le montant de la course. La dispersion des points peut indiquer s'il existe une corrélation entre la distance parcourue et le montant facturé, ce qui peut aider à comprendre la tarification des courses.

1.3 3. Graphique à Barres du Nombre de Courses par Type de Paiement :

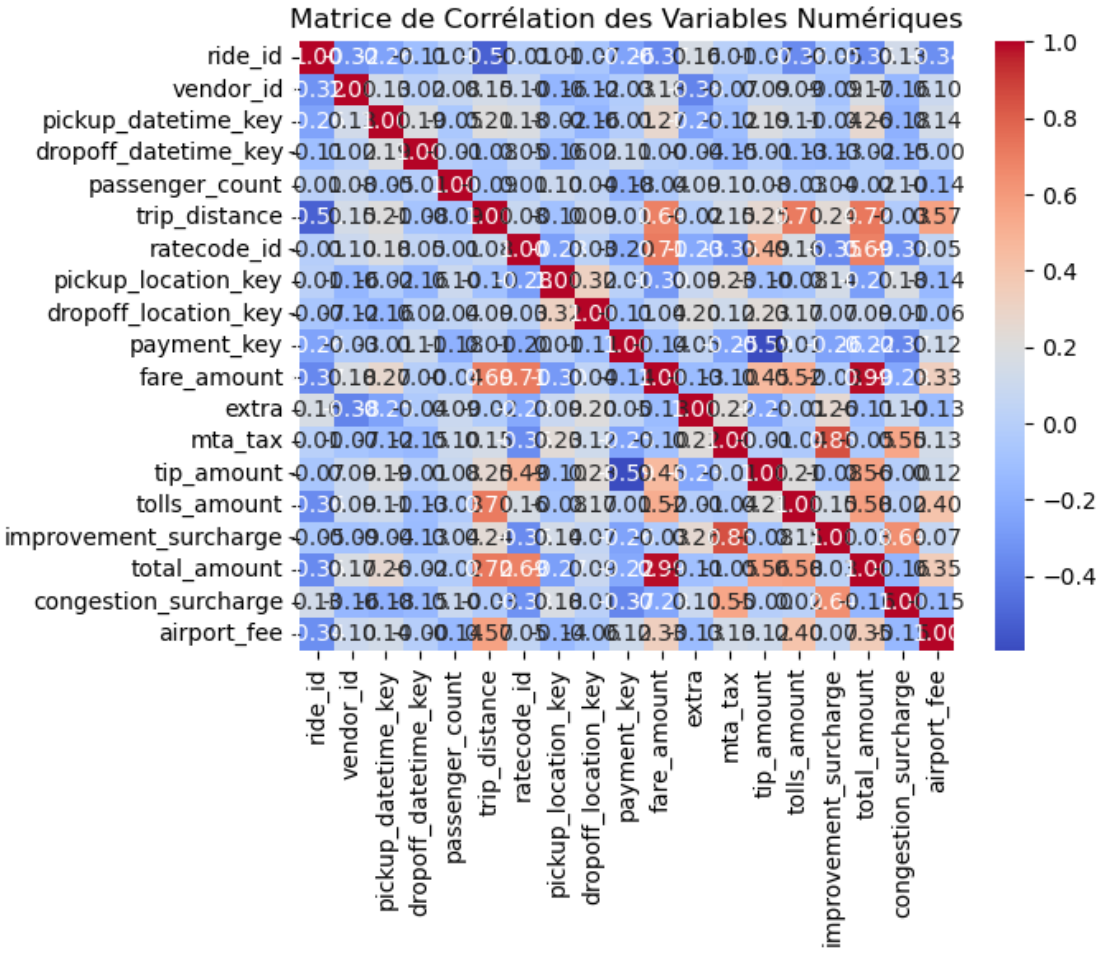
[18]:
```
sns.countplot(x='payment_key', data=df)
plt.title('Nombre de Courses par Type de Paiement')
plt.show()
```



Ce graphique à barres montre le nombre de courses en fonction du type de paiement, représenté par payment_key. Chaque barre représente un type de paiement différent, et la hauteur de la barre indique le nombre de courses payées par ce moyen. Ce type de visualisation aide à comprendre les préférences de paiement des clients et à évaluer l'utilisation relative des différentes méthodes de paiement.

1.4 4. Matrice de Corrélation des Variables Numériques :

[19]:
```
df_numeric = df.select_dtypes(include=[np.number])
corr_matrix = df_numeric.corr()
import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Matrice de Corrélation des Variables Numériques')
plt.show()
```



La matrice de corrélation est représentée sous forme de heatmap et montre comment les différentes variables numériques sont corrélées entre elles. Les couleurs varient selon l'intensité de la corrélation, allant de valeurs négatives (bleu) à valeurs positives (rouge). Des annotations indiquent les coefficients de corrélation. Cette visualisation est cruciale pour détecter des relations linéaires potentielles entre les variables et peut aider à identifier les variables qui ont une influence significative les unes sur les autres.

2 Définir les KPIs

Définir les KPIs Revenu Total : Description : Le revenu total généré par toutes les courses. Calcul : Somme de fare_amount. Tarif Moyen : Description : Le tarif moyen par course. Calcul : Moyenne de fare_amount. Nombre de Courses par Heure de la Journée : Description : Distribution du nombre de courses sur différentes heures de la journée pour identifier les pics de demande. Calcul : Compter le nombre de courses par heure, en extrayant l'heure de pickup_datetime ou dropoff_datetime. Nombre de Passagers : Description : Total de passagers transportés. Calcul : Somme de passenger_count. Distance Moyenne des Courses : Description : Distance moyenne parcourue par course. Calcul : Moyenne de trip_distance. Répartition des Types de Paiement

: Description : Analyse de la répartition des différents types de paiement pour comprendre les préférences des clients. Calcul : Compter le nombre de courses par payment_key.