EDA TP4

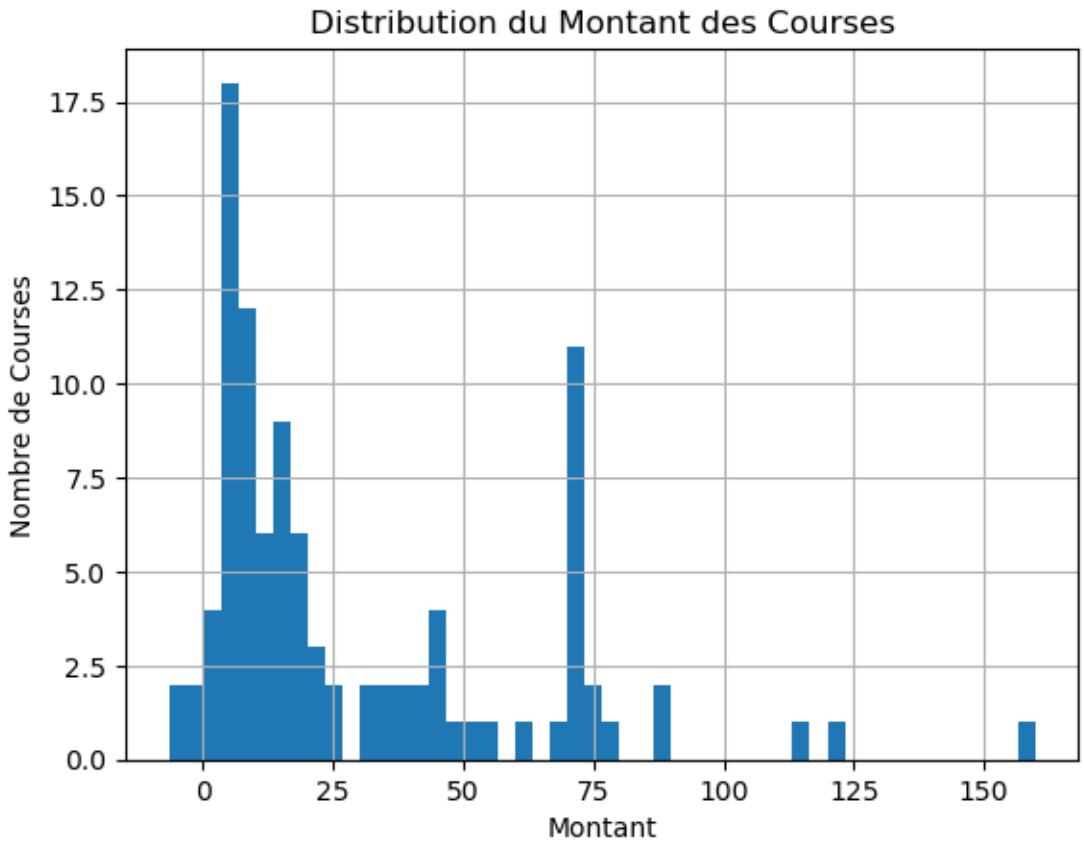
April 15, 2024

1 EDA TP4

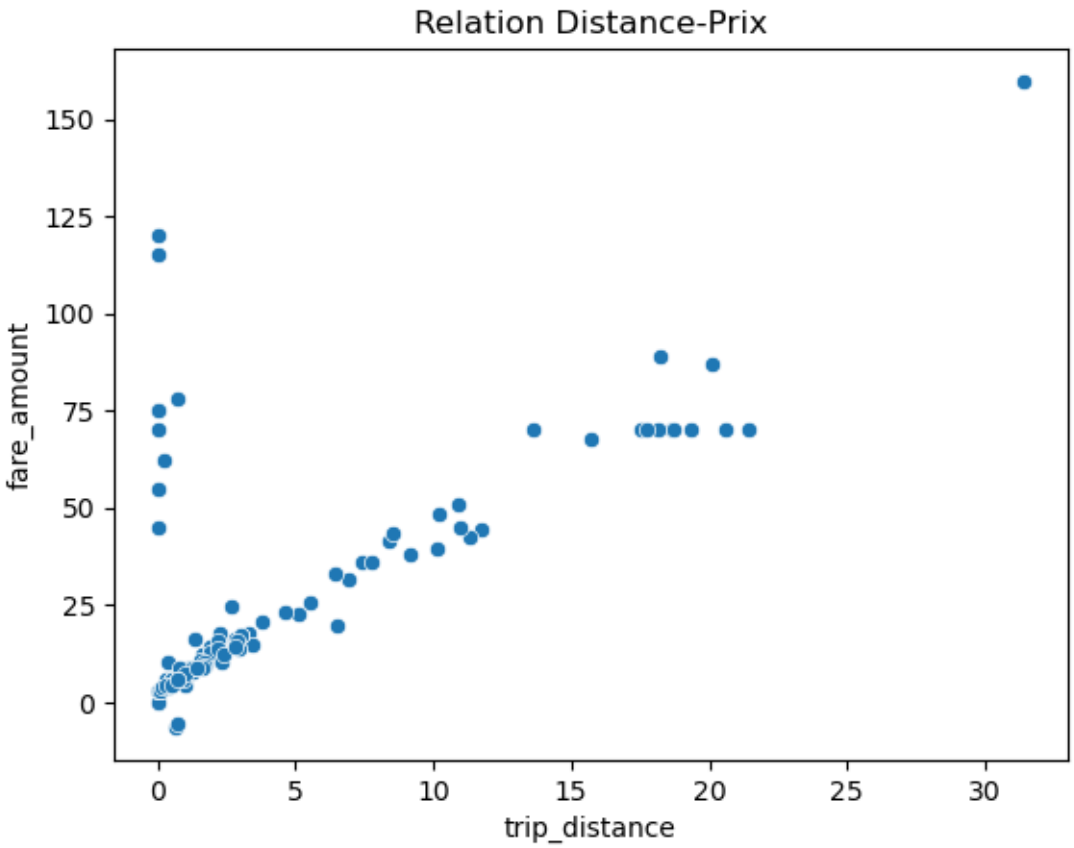
[8]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

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

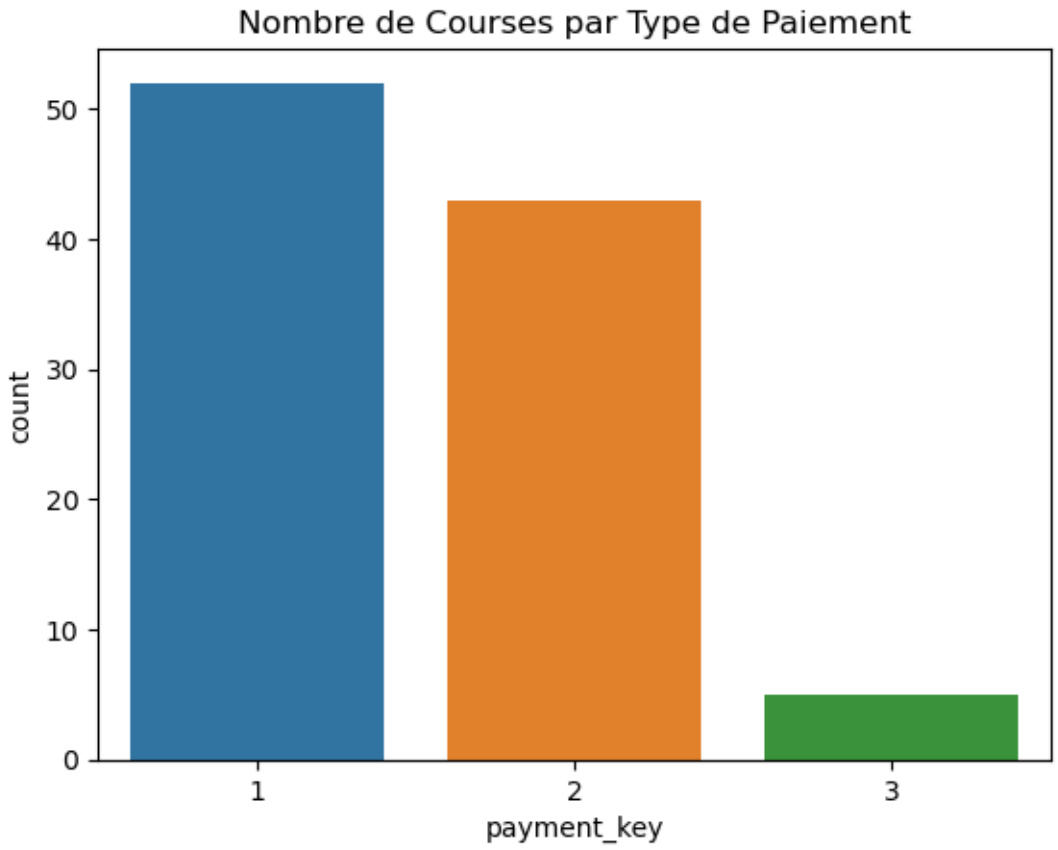
[3]:
```
df['fare_amount'].hist(bins=50)
plt.title('Distribution du Montant des Courses')
plt.xlabel('Montant')
plt.ylabel('Nombre de Courses')
plt.show()
```



[4]:
```
sns.scatterplot(x='trip_distance', y='fare_amount', data=df)
plt.title('Relation Distance-Prix')
plt.show()
```



[6]:
```
sns.countplot(x='payment_key', data=df)
plt.title('Nombre de Courses par Type de Paiement')
plt.show()
```

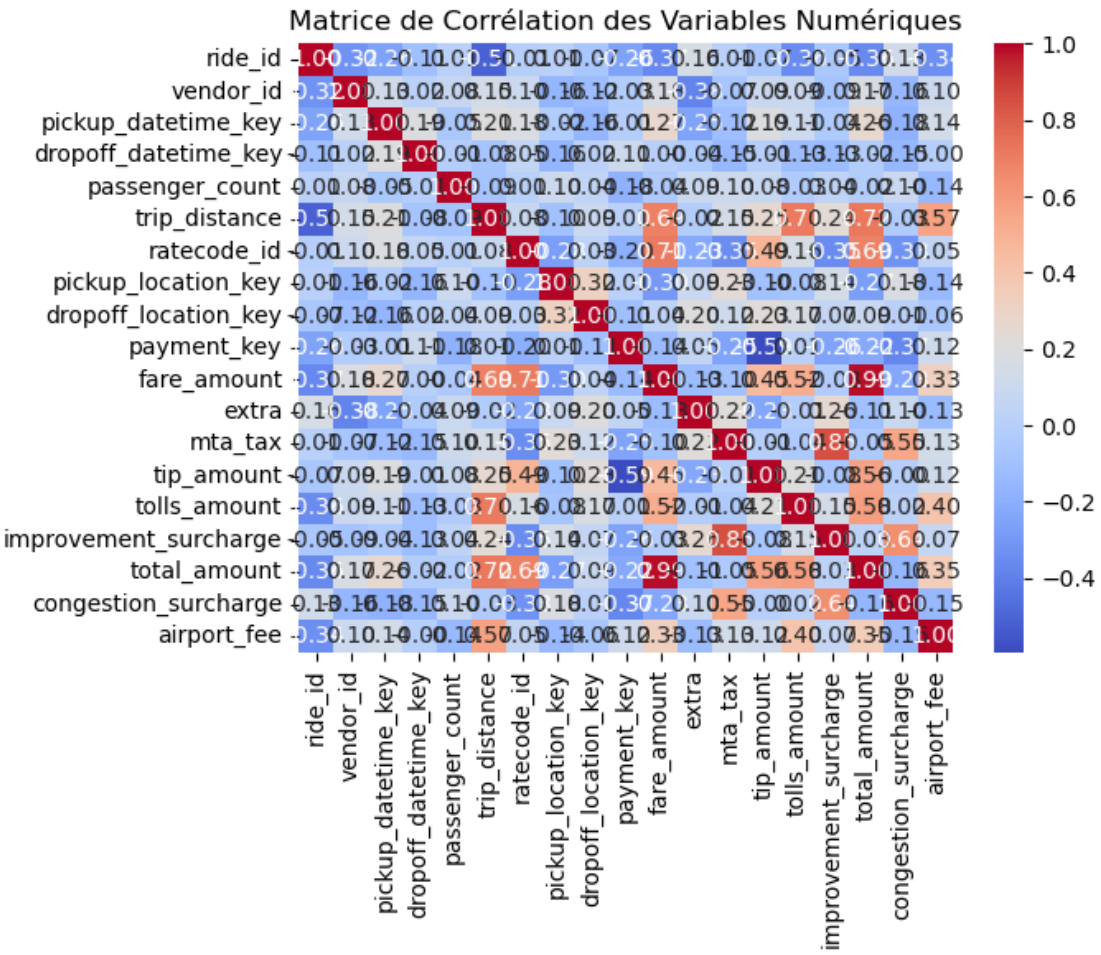


[13]:
```
df_numeric = df.select_dtypes(include=[np.number])

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()
```



[ ]:
```
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.
```