```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv('F:/DS Project/Sales_data.csv',encoding = 'latin1')

#average spending of people from different age group
data = df.groupby('age')['Total_amount'].mean()
data  = pd.DataFrame({'age':data.index, 'Average_purchase':data.values})
plt.figure(figsize = (16,4))
plt.plot('age','Average_purchase','ys-',data = data);
plt.grid();
plt.xlabel('age group');
plt.ylabel('Average_Purchase amount in rs');
plt.title('Age group vs average amount spent');

#age group and gender have high visiting rate
data_Age = df.groupby('age')['age'].count()
data_Sex = df.groupby('gender')['gender'].count()
data_Age = pd.DataFrame({'age':data_Age.index, 'Count':data_Age.values})
data_Sex = pd.DataFrame({'gender':data_Sex.index, 'Count':data_Sex.values})
plt.figure(figsize = (16,16))
plt.subplot(121)
plt.pie(data_Age['Count'],labels = data_Age['age'],autopct='%1.1f%%',shadow=True);
plt.title('Age split in data');
plt.subplot(122)
plt.pie(data_Sex['Count'],labels = data_Sex['gender'],autopct='%1.1f%%',shadow=True);
plt.title('Gender Split in data');
```

```
#top 10 selling product

data7 = pd.read_csv('F:/DS Project/mustcart/products1.csv',encoding = 'latin1')

data1 = df.groupby('product_id').agg({'Total_amount':'sum'}).reset_index()

data2 = df['product_id'].value_counts()

data2 = pd.DataFrame({'product_id':data2.index, 'Count':data2.values})

data8 = pd.merge(data1,data2,left_on='product_id',right_on='product_id',how = 'left');

data9 = pd.merge(data8,data7,left_on='product_id',right_on='product_id',how = 'left');

data10 = data9.sort_values(['Total_amount'],ascending=False)[0:10];

data10


fig, ax = plt.subplots(figsize=(16,10), dpi= 80)

sns.stripplot(data10.product_id, data10.Total_amount, jitter=0.05, size=8, ax=ax, linewidth=.5)

plt.title('Top 10 selling items', fontsize=22)

plt.ylabel('Total amount it was purchased in thousands of rs')

plt.xlabel('Product ids')

plt.show()


plt.figure(figsize=(16,6));

plt.grid();

plt.plot(data10['product_id'],data10['Total_amount'],'o-');

plt.xlabel('product IDs');

plt.ylabel('Total amount in thousands of rs');

plt.title('Top 10 Products with highest sales and Count displayed');

for a,b,c in zip(data10['product_id'], data10['Total_amount'], data10['Count']):

   plt.text(a, b+1000, str(c))

plt.show();
```

```
#high purchase rate based on marital status and gender
data = df.groupby(['gender','maritial status'])['gender'].count();
plt.figure(figsize=(16,13));
plt.subplot(211)
plt.pie(data.values,labels = data.index,autopct='%1.1f%%',shadow=True);
plt.title('Plot of split of gender and marital status in the data');
data = df.groupby(['gender','maritial status'])['Total_amount'].mean()
data.unstack(level=1).plot(kind='bar');

#popular product for each age group
data6 = pd.read_csv('F:/DS Project/mustcart/products.csv',encoding = 'latin1')
data4 = df.groupby('age')['product_id'].apply(lambda x: x.value_counts().index[0]).reset_index()
data5 = pd.merge(data4,data6,left_on='product_id',right_on='product_id',how = 'left');
data5

#purchase percent for each age group and for Gender Group
data = df.groupby('age')['Total_amount'].sum()
data_Sex = df.groupby('gender')['Total_amount'].sum()
plt.figure(figsize=(16,16));
plt.subplot(121)
plt.pie(data.values,labels = data.index,autopct='%1.1f%%',shadow=True);
plt.title('Percent amount spent per age group');
plt.subplot(122)
plt.pie(data_Sex.values,labels = data_Sex.index,autopct='%1.1f%%',shadow=True);
plt.title('Percent amount spent per gender');
```

```
#predicting purchase amount using linear regression
data.describe()
data['User_ID'].nunique()
data.info()
data = data.drop(['UserId','product_id'], axis=1)
data.info()
df_Gender = pd.get_dummies(data['gender'])
df_Age = pd.get_dummies(data['age'])
data_final = pd.concat([data, df_Gender, df_Age], axis=1)
data_final.head(100)
X = data_final[['F', 'M', '0-17', '18-25', '26-35', '36-45', '46-50', '51-55', '55+']]
y = data_final['Total_amount']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4)
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train, y_train)
print(lm.fit(X_train, y_train))
print('Intercept parameter:', lm.intercept_)
coeff_df = pd.DataFrame(lm.coef_, X.columns, columns=['Coefficient'])
print(coeff_df)
predictions = lm.predict(X_test)
print("Predicted purchases (in rupees) for new costumers:", predictions)

from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
```