In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
df=pd.read_csv('Diwali Sales Data.csv',encoding='latin1')
```

In [3]:
```
df.head()
```

Out[3]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	Stat
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtı
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Prades
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Prades
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnatak
4	1000588	Joni	P00057942	M	26-35	28	1	Gujar

In [4]:
```
df.shape
```

Out[4]:
```
(11251, 15)
```

In [5]:
```
!pip install -U ydata-profiling
```

```
Requirement already satisfied: ydata-profiling in c:\users\91901\documents\anacon
da1\lib\site-packages (4.6.1)
Requirement already satisfied: imagehash==4.3.1 in c:\users\91901\documents\anaco
nda1\lib\site-packages (from ydata-profiling) (4.3.1)
Requirement already satisfied: tqdm<5,>=4.48.2 in c:\users\91901\appdata\roaming
\python\python310\site-packages (from ydata-profiling) (4.64.1)
Requirement already satisfied: phik<0.13,>=0.11.1 in c:\users\91901\documents\ana
conda1\lib\site-packages (from ydata-profiling) (0.12.3)
Requirement already satisfied: wordcloud>=1.9.1 in c:\users\91901\documents\anaco
nda1\lib\site-packages (from ydata-profiling) (1.9.2)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in c:\users\91901\documents\ana
conda1\lib\site-packages (from ydata-profiling) (6.0)
Requirement already satisfied: visions[type_image_path]==0.7.5 in c:\users\91901
\documents\anaconda1\lib\site-packages (from ydata-profiling) (0.7.5)
Requirement already satisfied: statsmodels<1,>=0.13.2 in c:\users\91901\documents
\anaconda1\lib\site-packages (from ydata-profiling) (0.13.5)
Requirement already satisfied: dacite>=1.8 in c:\users\91901\documents\anaconda1
\lib\site-packages (from ydata-profiling) (1.8.1)
Requirement already satisfied: htmlmin==0.1.12 in c:\users\91901\documents\anacon
da1\lib\site-packages (from ydata-profiling) (0.1.12)
Requirement already satisfied: pydantic>=2 in c:\users\91901\documents\anaconda1
\lib\site-packages (from ydata-profiling) (2.4.2)
Requirement already satisfied: seaborn<0.13,>=0.10.1 in c:\users\91901\documents
\anaconda1\lib\site-packages (from ydata-profiling) (0.12.2)
Requirement already satisfied: pandas!=1.4.0,<2.1,>1.1 in c:\users\91901\appdata
\roaming\python\python310\site-packages (from ydata-profiling) (1.5.3)
Requirement already satisfied: numpy<1.26,>=1.16.0 in c:\users\91901\documents\an
aconda1\lib\site-packages (from ydata-profiling) (1.23.5)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in c:\users\91901\appdata\roam
ing\python\python310\site-packages (from ydata-profiling) (3.1.2)
Requirement already satisfied: typeguard<5,>=4.1.2 in c:\users\91901\documents\an
aconda1\lib\site-packages (from ydata-profiling) (4.1.5)
Requirement already satisfied: numba<0.59.0,>=0.56.0 in c:\users\91901\documents
\anaconda1\lib\site-packages (from ydata-profiling) (0.56.4)
Requirement already satisfied: matplotlib<=3.7.3,>=3.2 in c:\users\91901\appdata
\roaming\python\python310\site-packages (from ydata-profiling) (3.6.3)
Requirement already satisfied: multimethod<2,>=1.4 in c:\users\91901\documents\an
aconda1\lib\site-packages (from ydata-profiling) (1.10)
Requirement already satisfied: scipy<1.12,>=1.4.1 in c:\users\91901\appdata\roami
ng\python\python310\site-packages (from ydata-profiling) (1.10.0)
Requirement already satisfied: requests<3,>=2.24.0 in c:\users\91901\appdata\roam
ing\python\python310\site-packages (from ydata-profiling) (2.28.2)
Requirement already satisfied: PyWavelets in c:\users\91901\documents\anaconda1\l
ib\site-packages (from imagehash==4.3.1->ydata-profiling) (1.4.1)
Requirement already satisfied: pillow in c:\users\91901\appdata\roaming\python\py
thon310\site-packages (from imagehash==4.3.1->ydata-profiling) (9.4.0)
Requirement already satisfied: attrs>=19.3.0 in c:\users\91901\documents\anaconda
1\lib\site-packages (from visions[type_image_path]==0.7.5->ydata-profiling) (22.
1.0)
Requirement already satisfied: networkx>=2.4 in c:\users\91901\documents\anaconda
1\lib\site-packages (from visions[type_image_path]==0.7.5->ydata-profiling) (2.8.
4)
Requirement already satisfied: tangled-up-in-unicode>=0.0.4 in c:\users\91901\doc
uments\anaconda1\lib\site-packages (from visions[type_image_path]==0.7.5->ydata-p
rofiling) (0.2.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\91901\appdata\roaming
\python\python310\site-packages (from jinja2<3.2,>=2.11.1->ydata-profiling) (2.1.
2)
Requirement already satisfied: cycler>=0.10 in c:\users\91901\appdata\roaming\pyt
hon\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling) (0.1
```

```
1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\91901\appdata\roamin
g\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling)
(1.4.4)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\91901\appdata\roa
ming\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profilin
g) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\91901\appdata\roaming
\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling)
(3.0.9)
Requirement already satisfied: packaging>=20.0 in c:\users\91901\appdata\roaming
\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling)
(23.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\91901\appdata\roamin
g\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling)
(4.38.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\91901\appdata\roaming
\python\python310\site-packages (from matplotlib<=3.7.3,>=3.2->ydata-profiling)
(1.0.7)
Requirement already satisfied: llvmlite<0.40,>=0.39.0dev0 in c:\users\91901\docum
ents\anaconda1\lib\site-packages (from numba<0.59.0,>=0.56.0->ydata-profiling)
(0.39.1)
Requirement already satisfied: setuptools in c:\users\91901\documents\anaconda1\l
ib\site-packages (from numba<0.59.0,>=0.56.0->ydata-profiling) (65.6.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\91901\appdata\roaming\pyt
hon\python310\site-packages (from pandas!=1.4.0,<2.1,>1.1->ydata-profiling) (202
2.7.1)
Requirement already satisfied: joblib>=0.14.1 in c:\users\91901\appdata\roaming\p
ython\python310\site-packages (from phik<0.13,>=0.11.1->ydata-profiling) (1.2.0)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\91901\documents
\anaconda1\lib\site-packages (from pydantic>=2->ydata-profiling) (0.6.0)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\91901\documen
ts\anaconda1\lib\site-packages (from pydantic>=2->ydata-profiling) (4.8.0)
Requirement already satisfied: pydantic-core==2.10.1 in c:\users\91901\documents
\anaconda1\lib\site-packages (from pydantic>=2->ydata-profiling) (2.10.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\91901\appdata\ro
aming\python\python310\site-packages (from requests<3,>=2.24.0->ydata-profiling)
(1.26.14)
Requirement already satisfied: idna<4,>=2.5 in c:\users\91901\appdata\roaming\pyt
hon\python310\site-packages (from requests<3,>=2.24.0->ydata-profiling) (3.4)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\91901\appdata
\roaming\python\python310\site-packages (from requests<3,>=2.24.0->ydata-profilin
g) (3.0.1)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\91901\appdata\roami
ng\python\python310\site-packages (from requests<3,>=2.24.0->ydata-profiling) (20
22.12.7)
Requirement already satisfied: patsy>=0.5.2 in c:\users\91901\documents\anaconda1
\lib\site-packages (from statsmodels<1,>=0.13.2->ydata-profiling) (0.5.3)
Requirement already satisfied: colorama in c:\users\91901\appdata\roaming\python
\python310\site-packages (from tqdm<5,>=4.48.2->ydata-profiling) (0.4.6)
Requirement already satisfied: six in c:\users\91901\appdata\roaming\python\pytho
n310\site-packages (from patsy>=0.5.2->statsmodels<1,>=0.13.2->ydata-profiling)
(1.16.0)
```

In [6]:
```
from ydata_profiling import ProfileReport
```

In [7]:
```
profile = ProfileReport(df, title="DataSet Profile Report")
```

In [8]:
```
profile
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

'unnamed' is an unsupported type, check if it needs cleaning or further analysis — Unsupported

Reproduction

Analysis started	2023-11-07 13:36:37.889028
Analysis finished	2023-11-07 13:36:42.508342
Duration	4.62 seconds
Software version	ydata-profiling vv4.6.1 (https://github.com/ydataai/ydata-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22DataSet%20Profile%20Report%

Variables

Select Columns

User_ID

Real number (ℝ)

Distinct	3755
Distinct (%)	33.4%
Missing	0
Missing (%)	0.0%

Out[8]:

In [9]:
```
profile.to_widgets()
```

```
Render widgets:   0%|          | 0/1 [00:00<?, ?it/s]
VBox(children=(Tab(children=(Tab(children=(GridBox(children=(VBox(children=(Grids
pecLayout(children=(HTML(valu…
```

In [10]:
```
profile.to_notebook_iframe()
```

Overview

Dataset statistics

Number of variables	15
Number of observations	11251
Missing cells	22514
Missing cells (%)	13.3%
Duplicate rows	8
Duplicate rows (%)	0.1%
Total size in memory	1.3 MiB
Average record size in memory	120.0 B

Variable types

Numeric	3
Text	2
Categorical	8
Unsupported	2

Alerts

Dataset has 8 (0.1%) duplicate rows	**Duplicates**

In [11]:
```
orgin_df=df[['Gender','Age Group','Age','Marital_Status','State','Zone','Occupat
```

In [12]:
```
orgin_df.head()
```

Out[12]:

	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Orders
0	F	26-35	28	0	Maharashtra	Western	Healthcare	1
1	F	26-35	35	1	Andhra Pradesh	Southern	Govt	3
2	F	26-35	35	1	Uttar Pradesh	Central	Automobile	3
3	M	0-17	16	0	Karnataka	Southern	Construction	2
4	M	26-35	28	1	Gujarat	Western	Food Processing	2

In [13]:
```
orgin_df.shape
```

Out[13]:
```
(11251, 9)
```

In [14]:
```
orgin_df.isnull().sum()
```

Out[14]:
```
Gender             0
Age Group          0
Age                0
Marital_Status     0
State              0
Zone               0
Occupation         0
Orders             0
Amount            12
dtype: int64
```

In [15]:
```
orgin_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Gender          11251 non-null  object 
 1   Age Group       11251 non-null  object 
 2   Age             11251 non-null  int64  
 3   Marital_Status  11251 non-null  int64  
 4   State           11251 non-null  object 
 5   Zone            11251 non-null  object 
 6   Occupation      11251 non-null  object 
 7   Orders          11251 non-null  int64  
 8   Amount          11239 non-null  float64
dtypes: float64(1), int64(3), object(5)
memory usage: 791.2+ KB
```

In [16]:
```
amount_mean=orgin_df['Amount'].mean()
```

In [18]:
```
orgin_df['Amount']=orgin_df['Amount'].fillna(value=amount_mean)
```

```
C:\Users\91901\AppData\Local\Temp\ipykernel_16992\3564231867.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  orgin_df['Amount']=orgin_df['Amount'].fillna(value=amount_mean)
```

In [19]:
```
orgin_df['Amount'].isnull().sum()
```

Out[19]:
```
0
```

In [20]:
```
X=orgin_df.drop(columns=['Amount'])
y=orgin_df['Amount']
```

In [21]:
```
X.head()
```

Out[21]:

	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Orders
0	F	26-35	28	0	Maharashtra	Western	Healthcare	1
1	F	26-35	35	1	Andhra Pradesh	Southern	Govt	3
2	F	26-35	35	1	Uttar Pradesh	Central	Automobile	3
3	M	0-17	16	0	Karnataka	Southern	Construction	2
4	M	26-35	28	1	Gujarat	Western	Food Processing	2

In [22]:
```
from sklearn.model_selection import train_test_split
```

In [23]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=4
```

In [24]:
```
X_train.shape,X_test.shape,y_train.shape,y_test.shape
```

Out[24]:
```
((8438, 8), (2813, 8), (8438,), (2813,))
```

In [25]:
```
from sklearn.preprocessing import OneHotEncoder,StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
```

In [26]:
```
oht=OneHotEncoder(handle_unknown='ignore',sparse_output=False)
std=StandardScaler()
```

In [27]:
```
X_train
```

Out[27]:

	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Order
11070	F	26-35	30	0	Madhya Pradesh	Central	Banking	
7621	F	55+	73	1	Haryana	Northern	Healthcare	
10796	F	18-25	22	0	Gujarat	Western	Construction	
10239	M	26-35	29	0	Maharashtra	Western	Banking	
5652	M	46-50	46	0	Karnataka	Southern	Banking	
...	...	...	...	...	...	...	...	
5734	F	18-25	20	0	Bihar	Eastern	Banking	
5191	F	46-50	47	0	Uttarakhand	Central	Aviation	
5390	F	46-50	49	0	Kerala	Southern	Construction	
860	F	18-25	19	1	Bihar	Eastern	Aviation	
7270	F	36-45	39	1	Karnataka	Southern	Construction	

8438 rows × 8 columns

In [28]:
```
X_test
```

Out[28]:

	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Orde
5749	M	26-35	30	0	Haryana	Northern	Banking	
7335	F	26-35	31	0	Uttar Pradesh	Central	Retail	
5932	M	26-35	28	1	Uttarakhand	Central	Automobile	
3819	F	18-25	22	0	Karnataka	Southern	Chemical	
6395	F	18-25	22	0	Rajasthan	Northern	Food Processing	
...	...	...	...	...	...	...	...	
9112	M	18-25	22	1	Haryana	Northern	IT Sector	
2269	M	36-45	44	0	Andhra Pradesh	Southern	Banking	
9030	M	46-50	46	0	Bihar	Eastern	Hospitality	
9966	F	26-35	32	1	Karnataka	Southern	Healthcare	
7528	M	36-45	42	1	Maharashtra	Western	Banking	

2813 rows × 8 columns

In [29]:
```
X_train_scaled=oht.fit_transform(X_train)
X_test_scaled=oht.transform(X_test)
```

In [30]:
```
X_train_scaled
```

Out[30]:
```
array([[1., 0., 0., ..., 0., 1., 0.],
       [1., 0., 0., ..., 0., 0., 1.],
       [1., 0., 0., ..., 0., 0., 0.],
       ...,
       [1., 0., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 1., 0., 0.],
       [1., 0., 0., ..., 0., 1., 0.]])
```

In [31]:
```
X_test_scaled
```

Out[31]:
```
array([[0., 1., 0., ..., 0., 0., 1.],
       [1., 0., 0., ..., 1., 0., 0.],
       [0., 1., 0., ..., 0., 1., 0.],
       ...,
       [0., 1., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 0., 1., 0.],
       [0., 1., 0., ..., 0., 0., 0.]])
```

In [32]:
```
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
```

In [33]:
```
svr=SVR()
dtr=DecisionTreeRegressor()
knr=KNeighborsRegressor()
```

In [36]:
```
svr.fit(X_train_scaled,y_train)
```

Out[36]:
```
SVR()
```

In [37]:
```
dtr.fit(X_train_scaled,y_train)
```

Out[37]:
```
DecisionTreeRegressor()
```

In [38]:
```
knr.fit(X_train_scaled,y_train)
```

Out[38]:
```
KNeighborsRegressor()
```

In [39]:
```
y_pred_svr=svr.predict(X_test_scaled)
```

In [40]:
```
y_pred_dtr=dtr.predict(X_test_scaled)
```

In [41]:
```
y_pred_knr=knr.predict(X_test_scaled)
```

In [42]:
```
y_pred_svr
```

Out[42]:
```
array([8089.09506714, 8130.44530743, 8096.58316785, ..., 8099.84763544,
       8119.37959829, 8103.41641902])
```

In [43]:
```
y_pred_dtr
```

Out[43]:
```
array([ 7972., 11616.,  2002., ..., 19515.,  8535.,  1681.])
```

In [44]:
```
y_pred_knr
```

Out[44]:
```
array([ 7493.8, 11585.8,  8026. , ..., 10420.8, 11178.8,  7372. ])
```

In [45]:
```
from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
```

In [46]:
```
print("Mean absolute error:- ",mean_absolute_error(y_test,y_pred_svr))
print("Mean squared error:- ",mean_squared_error(y_test,y_pred_svr))
print("Root Mean squared error:- ",np.sqrt(mean_absolute_error(y_test,y_pred_svr
```

```
Mean absolute error:-  4193.6740966798
Mean squared error:-  29459301.56008552
Root Mean squared error:-  64.75858318925607
```

In [48]:
```
print("Mean absolute error:- ",mean_absolute_error(y_test,y_pred_dtr))
print("Mean squared error:- ",mean_squared_error(y_test,y_pred_dtr))
print("Root Mean squared error:- ",np.sqrt(mean_absolute_error(y_test,y_pred_dtr
```

```
Mean absolute error:-  5654.507289388122
Mean squared error:-  51284745.25594983
Root Mean squared error:-  75.19645795772645
```

In [47]:
```
print("Mean absolute error:- ",mean_absolute_error(y_test,y_pred_knr))
print("Mean squared error:- ",mean_squared_error(y_test,y_pred_knr))
print("Root Mean squared error:- ",np.sqrt(mean_absolute_error(y_test,y_pred_knr
```

```
Mean absolute error:-  4519.1313660053975
Mean squared error:-  31316789.272688925
Root Mean squared error:-  67.224484869766
```

In [49]:
```
print("Score :- ",r2_score(y_test,y_pred_svr))
print("Score :- ",r2_score(y_test,y_pred_dtr))
print("Score :- ",r2_score(y_test,y_pred_knr))
```

```
Score :-  -0.06779897129072521
Score :-  -0.8588966922897798
Score :-  -0.13512655082134173
```

In [50]:
```
test=pd.DataFrame([['M','46-50',50,0,'Uttar Pradesh','Central','Healthcare',3]],
```

In [51]:
```
test
```

Out[51]:

	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Orders
0	M	46-50	50	0	Uttar Pradesh	Central	Healthcare	3

In [52]:
```
test_scaled=oht.transform(test)
```

In [53]:
```
svr.predict(test_scaled)
```

Out[53]:
```
array([8115.10030671])
```

In [54]:
```
dtr.predict(test_scaled)
```

Out[54]:
```
array([4536.])
```

In [55]:
```
knr.predict(test_scaled)
```

Out[55]:
```
array([6332.2])
```

In [ ]: