ADS_PHASE 3 PROJECT

IBM Naan Mudhalvan

College Code:1108(Jaya Engineering College).

Project Name: Future Sales Prediction.

Team Members:

1. **Thirukarthikeyan (B.Tech[Information Technology])**
2. **Veluprasath (B.Tech[Information Technology])**
3. **Praveen kumar (B.Tech[Information Technology])**
4. **Vignesh (B.Tech[Information Technology])**
5. **Abinesh(B.Tech[Information Technology])**

PHASE 3 PROJECT PROBLEM STATEMENT

Project name: Future sales prediction

Synopsis:

Aim

- **3.1 Dataset and its detail explanation.**
- **3.2 Begin building the project by load the dataset.**
- **3.3 Preprocess Dataset.**
- **3.4 Performing Different Analysis needed.**
- **Conclusion.**

Aim:

Clearly define the objectives of your sales prediction project. What specific sales metrics or time frames are you trying to predict? What decisions or actions will these predictions inform?

3.1 Dataset and its detail explanation:

About Dataset:

Context:

- ✓ The Customer Shopping Preferences Dataset offers valuable insights into consumer behavior and purchasing patterns.
- ✓ Understanding customer preferences and trends is critical for businesses to tailor their products, marketing strategies, and overall customer experience.
- ✓ This dataset captures a wide range of customer attributes including age, gender, purchase history, preferred payment methods, frequency of purchases, and more.
- ✓ Analyzing this data can help businesses make informed decisions, optimize product offerings, and enhance customer satisfaction.
- ✓ The dataset stands as a valuable resource for businesses aiming to align their strategies with customer needs and preferences.
- ✓ It's important to note that this dataset is a Synthetic Dataset Created for Beginners to learn more about Data Analysis and Machine Learning.

Content:

- ✓ This dataset encompasses various features related to customer shopping preferences, gathering essential information for businesses seeking to enhance their understanding of their customer base.

- ✓ The features include customer age, gender, purchase amount, preferred payment methods, frequency of purchases, and feedback ratings.

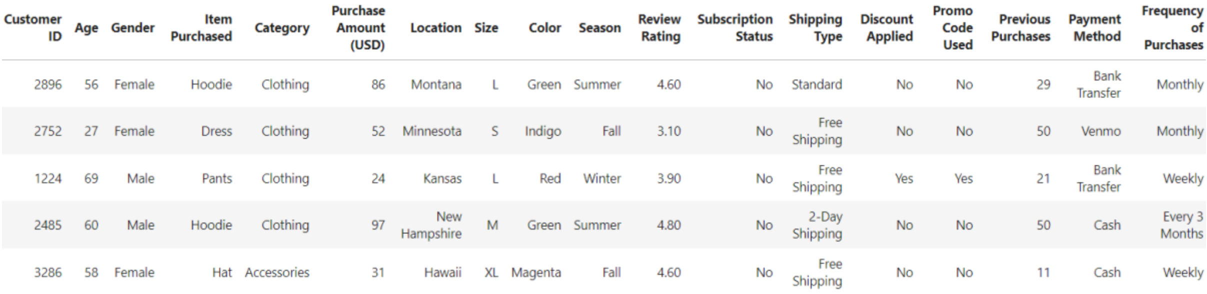
- ✓ Additionally, data on the type of items purchased, shopping frequency, preferred shopping seasons, and interactions with promotional offers is included.

- ✓ With a collection of 3900 records, this dataset serves as a foundation for businesses looking to apply data-driven insights for better decision-making and customer-centric strategies.

Dataset Glossary (Column-wise):

- **Customer ID -** Unique identifier for each customer.
- **Age -** Age of the customer.
- **Gender -** Gender of the customer (Male/Female).
- **Item Purchased -** The item purchased by the customer.
- **Category -** Category of the item purchased.
- **Purchase Amount (USD) -** The amount of the purchase in USD.
- **Location -** Location where the purchase was made.
- **Size -** Size of the purchased item.
- **Color -** Color of the purchased item.
- **Season -** Season during which the purchase was made.
- **Review Rating -** Rating given by the customer for the purchased item.
- **Subscription Status -** Indicates if the customer has a subscription (Yes/No).
- **Shipping Type -** Type of shipping chosen by the customer.
- **Discount Applied -** Indicates if a discount was applied to the purchase (Yes/No).
- **Promo Code Used -** Indicates if a promo code was used for the purchase (Yes/No).
- **Previous Purchases -** Number of previous purchases made by the customer.
- **Payment Method -** Customer's most preferred payment method.
- **Frequency of Purchases -** Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly).

Structure of the Dataset:



Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
2896	56	Female	Hoodie	Clothing	86	Montana	L	Green	Summer	4.60	No	Standard	No	No	29	Bank Transfer	Monthly
2752	27	Female	Dress	Clothing	52	Minnesota	S	Indigo	Fall	3.10	No	Free Shipping	No	No	50	Venmo	Monthly
1224	69	Male	Pants	Clothing	24	Kansas	L	Red	Winter	3.90	No	Free Shipping	Yes	Yes	21	Bank Transfer	Weekly
2485	60	Male	Hoodie	Clothing	97	New Hampshire	M	Green	Summer	4.80	No	2-Day Shipping	No	No	50	Cash	Every 3 Months
3286	58	Female	Hat	Accessories	31	Hawaii	XL	Magenta	Fall	4.60	No	Free Shipping	No	No	11	Cash	Weekly

3.2 Begin building the project by load the dataset:

import pandas as pd

# Load CSV data

df = pd.read_csv('shopping_trends.csv')

# Display the first few rows of the DataFrame

print(df.head(5))

Output:

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Revi Rati
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	

3.3 Preprocess Dataset:

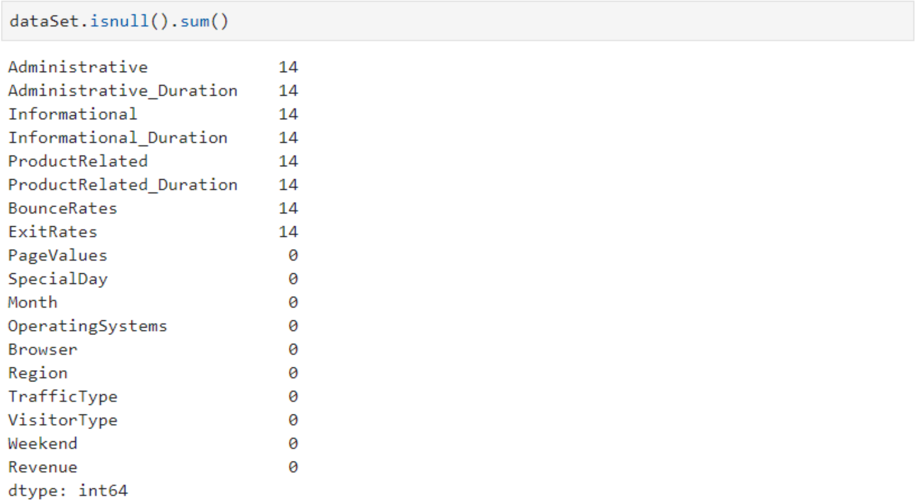
```
  data.info()

Output:
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 3900 entries, 1 to 3900
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Age                     3900 non-null   int64  
 1   Gender                  3900 non-null   object 
 2   Item Purchased          3900 non-null   object 
 3   Category                3900 non-null   object 
 4   Purchase Amount (USD)   3900 non-null   int64  
 5   Location                3900 non-null   object 
 6   Size                    3900 non-null   object 
 7   Color                   3900 non-null   object 
 8   Season                  3900 non-null   object 
 9   Review Rating           3900 non-null   float64
 10  Subscription Status     3900 non-null   object 
 11  Payment Method          3900 non-null   object 
 12  Shipping Type           3900 non-null   object 
 13  Discount Applied        3900 non-null   object 
 14  Promo Code Used         3900 non-null   object 
 15  Previous Purchases        3900 non-null   int64
 16  Preferred Payment Method  3900 non-null   object
 17  Frequency of Purchases    3900 non-null   object
dtypes: float64(1), int64(3), object(14)
memory usage: 578.9+ KB
```

```
data.shape
                Output:     (3900, 18)
```

Handling missing data points:

There can be random missing data points in the dataset, which if not handled properly may raise errors later, or may lead to inaccurate inferences. First, we found out if there are any missing values. The value next to each feature name shows the number of missing data points per each column.

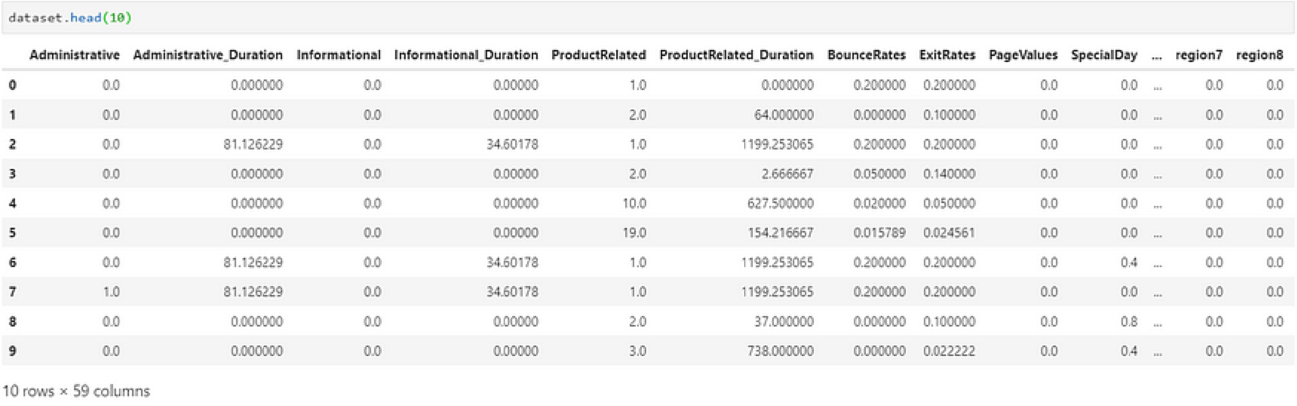


```
dataSet.isnull().sum()
```

```
Administrative             14
Administrative_Duration    14
Informational              14
Informational_Duration     14
ProductRelated             14
ProductRelated_Duration    14
BounceRates                14
ExitRates                  14
PageValues                  0
SpecialDay                  0
Month                       0
OperatingSystems            0
Browser                     0
Region                      0
TrafficType                 0
VisitorType                 0
Weekend                     0
Revenue                     0
dtype: int64
```

Handling catagorical data:

In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.

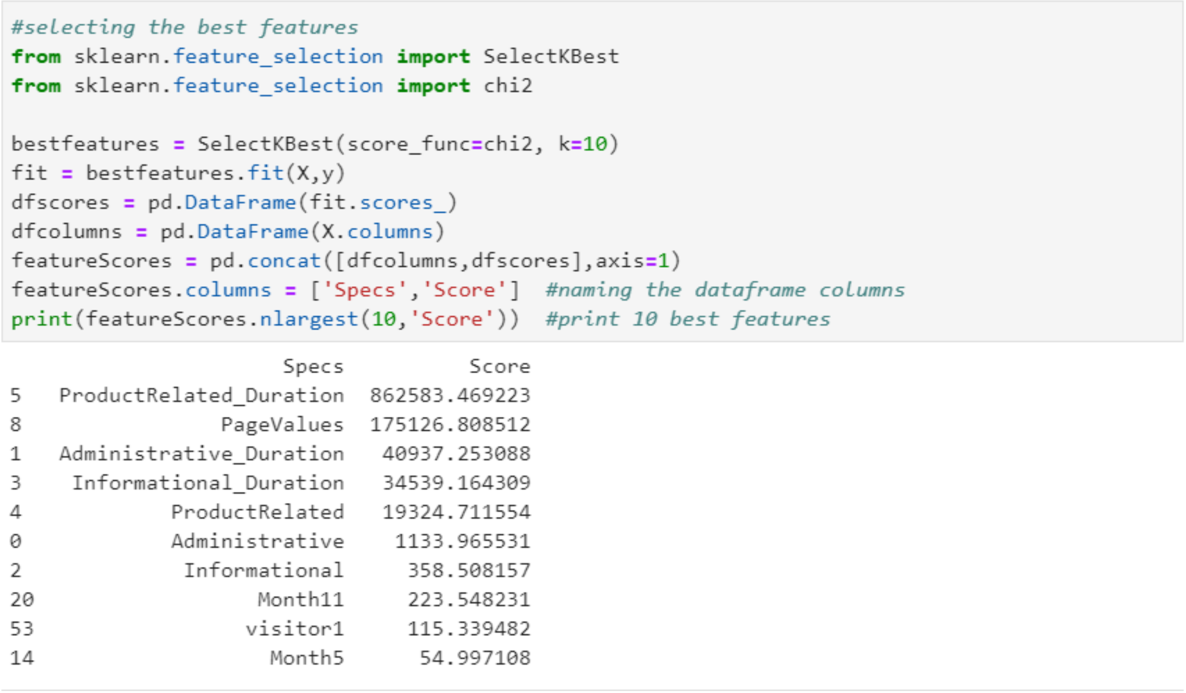


```python
dataset.head(10)
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	...	region7	region8
0	0.0	0.000000	0.0	0.00000	1.0	0.000000	0.200000	0.200000	0.0	0.0	...	0.0	0.0
1	0.0	0.000000	0.0	0.00000	2.0	64.000000	0.000000	0.100000	0.0	0.0	...	0.0	0.0
2	0.0	81.126229	0.0	34.60178	1.0	1199.253065	0.200000	0.200000	0.0	0.0	...	0.0	0.0
3	0.0	0.000000	0.0	0.00000	2.0	2.666667	0.050000	0.140000	0.0	0.0	...	0.0	0.0
4	0.0	0.000000	0.0	0.00000	10.0	627.500000	0.020000	0.050000	0.0	0.0	...	0.0	0.0
5	0.0	0.000000	0.0	0.00000	19.0	154.216667	0.015789	0.024561	0.0	0.0	...	0.0	0.0
6	0.0	81.126229	0.0	34.60178	1.0	1199.253065	0.200000	0.200000	0.0	0.4	...	0.0	0.0
7	1.0	81.126229	0.0	34.60178	1.0	1199.253065	0.200000	0.200000	0.0	0.0	...	0.0	0.0
8	0.0	0.000000	0.0	0.00000	2.0	37.000000	0.000000	0.100000	0.0	0.8	...	0.0	0.0
9	0.0	0.000000	0.0	0.00000	3.0	738.000000	0.000000	0.022222	0.0	0.4	...	0.0	0.0

10 rows × 59 columns

Selecting the best features:

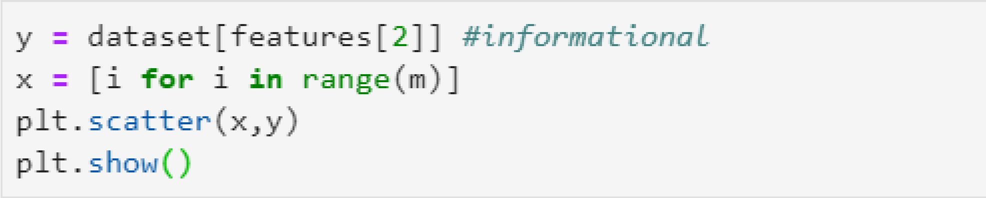


```python
#selecting the best features
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

bestfeatures = SelectKBest(score_func=chi2, k=10)
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns
print(featureScores.nlargest(10,'Score'))  #print 10 best features
```

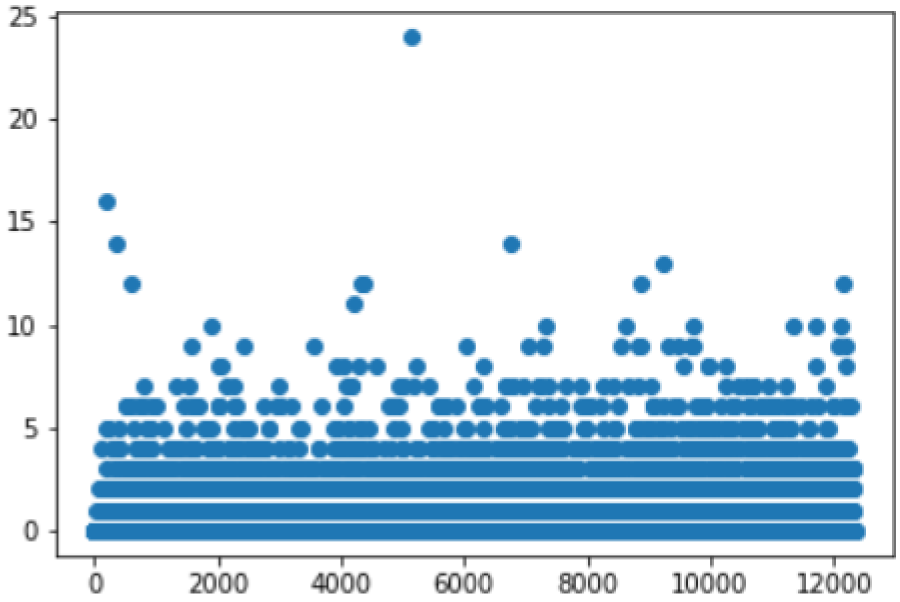
```
                      Specs          Score
5   ProductRelated_Duration  862583.469223
8                PageValues  175126.808512
1   Administrative_Duration   40937.253088
3    Informational_Duration   34539.164309
4            ProductRelated   19324.711554
0            Administrative    1133.965531
2             Informational     358.508157
20                  Month11     223.548231
53                 visitor1     115.339482
14                   Month5      54.997108
```

Outliers:

In statistics, an **outlier** is a **data** point that differs significantly from other observations. An **outlier** may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the **data** set.



```
y = dataset[features[2]] #informational
x = [i for i in range(m)]
plt.scatter(x,y)
plt.show()
```



Conclusion:

Start by clearly defining the objectives of our sales prediction project. Knowing what specific sales metrics or time frames you aim to predict is essential.