

ADS\_PHASE 3 PROJECT

IBM Naan Mudhalvan

**College Code:**7240 Dhaanish Ahmed Institute of Technology

**Project Name:** Future Sales Prediction.

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?

### 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:



|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ADS\_Phase3  **3.2 Begin**  import pand # Load CSV df = pd.read # Display th  print(df.hea Output:  Customer ID | Project  **building**  as as pd data  \_csv('sho e first few d(5))  Age | **the proj**  pping\_tren rows of th  Gender | **ect by loa**  ds.csv')  e DataFrame  Item Purchased | **d the dat**  Category | **aset:**  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  **3.3 Prepro** | 21  **cess Da** | Male  **taset:** | Sandals | Footwear | 90 | Rhode Island | M | Maroon | Spring |  |
|  | data.info()  Output: | | | | | | | | | |  |
| <class 'pandas.core.frame.DataFrame'> Index: 3900 entries, 1 to 3900  Data columns (total 18 columns):  # Column Non-Null Count Dtype   1. Age 3900 non-null int64 2. Gender 3900 non-null object 3. Item Purchased 3900 non-null object 4. Category 3900 non-null object 5. Purchase Amount (USD) 3900 non-null int64 6. Location 3900 non-null object 7. Size 3900 non-null object 8. Color 3900 non-null object 9. Season 3900 non-null object 10. Review Rating 3900 non-null float64 11. Subscription Status 3900 non-null object 12. Payment Method 3900 non-null object 13. Shipping Type 3900 non-null object 14. Discount Applied 3900 non-null object 15. Promo Code Used 3900 non-null object   1108-Jaya Engineering College | | | | | | | | | | | |

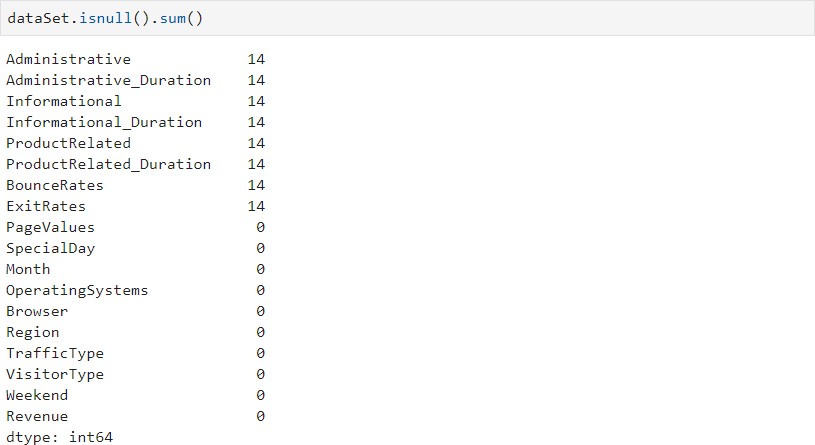
1. Previous Purchases 3900 non-null int64
2. Preferred Payment Method 3900 non-null object
3. Frequency of Purchases 3900 non-null object dtypes: float64(1), int64(3), object(14)

memory usage: 578.9+ KB

data.shape

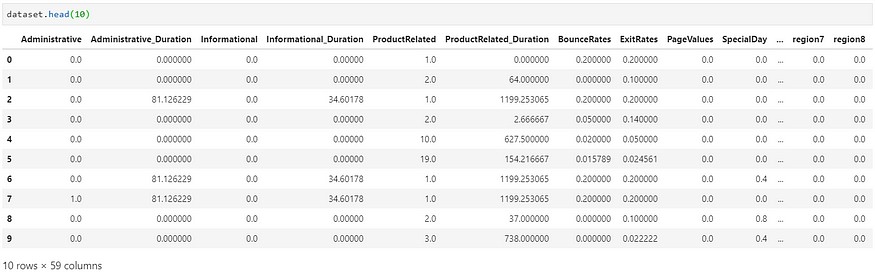
Output: (3900, 18)

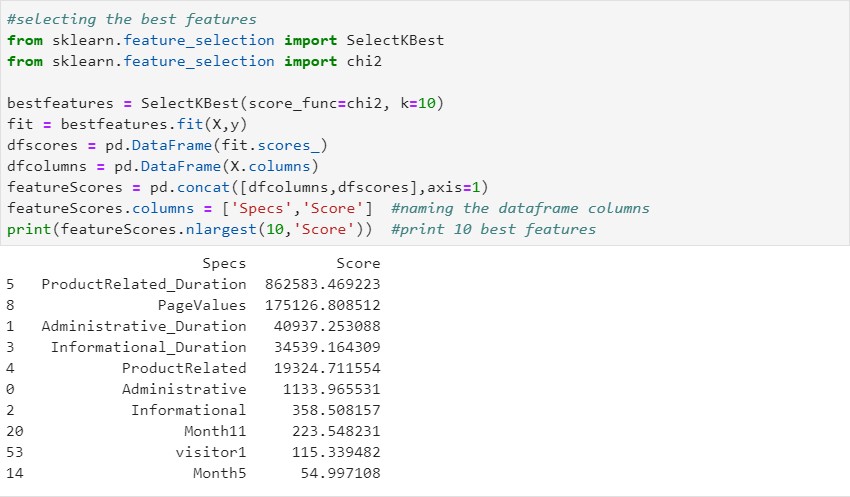
Handling missing data points:

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.

**Handling catagorical data:**

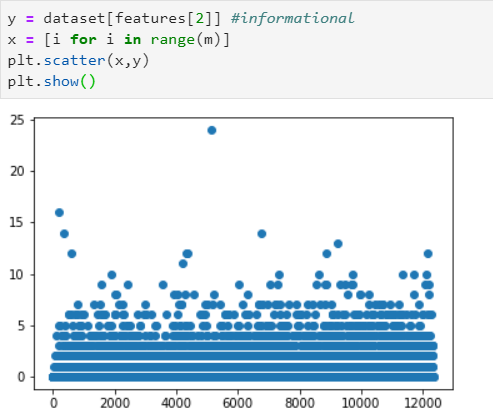
In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.



**Selecting the best features:**

## 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.



**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.