Greetings

Today, we are delving into an exciting topic: Python Exploratory Data Analysis (EDA), focusing on analyzing Diwali sales data. By the end of this session, you will have a solid understanding of how to use Python to explore, clean, visualize, and draw insights from datasets.

The first step in any data analysis project is importing the Python libraries we will need for working with data and visualizing the analysis.

**Pandas** - For loading, manipulating and analyzing dataset. It allows us to store data in DataFrames similar to Excel tables and has many built-in methods.

**NumPy** - Provides support for mathematical and numerical operations on arrays and matrices which is useful for statistical analysis.

**Matplotlib and Seaborn** - Primary Python data visualization libraries that allow us to create plots, charts, histograms to visualize trends and patterns in data.

**Data Transformation Libraries** - Libraries like scikit-learn can be used for data transformations at later stages of analysis.

We will begin by introducing the Diwali sales dataset. In this dataset, we have columns such as userID, customer\_name, product\_ID, gender, Age\_group, age, marital status (representing married or unmarried), state, zone, occupation, product\_category, orders, amount, status, and an unnamed column. These are the columns available in our dataset.

Further, we looked into the shape of the dataset, which shows the number of columns and rows we have: 11251 rows and 15 columns, respectively.

With this much data available, let's inspect the top and bottom rows of the dataset. To display the topmost rows, we can use the code **data.head()**, and similarly, **data.tail()** shows the bottommost rows from the dataset.

Furthermore, we looked into the information of the dataset using **data.info()**. This method displays the number of columns, number of rows, null values, data types of the columns, and the memory location it occupies.

Now that we have picked all the vegetables, our next step is to clean them.  
Data Cleaning:

In this step, we understand the presence of null or missing values and address duplicate rows. We then proceed to handle missing values, typically through one of three methods: Acceptance, Deletion, or Imputation. Imputation involves replacing null values with predetermined values based on the category of the data.

If we're not utilizing a machine learning model, we may simply opt to delete rows with null or missing values, as well as duplicate rows.

We observed that two columns have 100% null values, and the amount column has 12 rows with missing values.

As part of the data reduction step, we remove unwanted columns from the dataset. Certain columns or variables can be dropped if they do not contribute value to our analysis.

Removing duplicates from the data is essential because it can distort results.

Up to this point, we have cleaned our vegetables; next, we proceed to food preparation.

Finally, we enter into EDA, which stands for Exploratory Data Analysis.

Using **data.describe().T**, we gain insight into the statistical description of the data.

The need to find mean, standard deviation, max, min, and count (measures of central tendency) serves several purposes:

* Summarizing data to get a quick overview of its distribution and characteristics.
* Handling missing values by understanding their impact on the overall dataset.
* Identifying extreme values that deviate significantly from the rest of the data. These outliers can indicate data quality issues, such as data entry errors or measurement inconsistencies.

After describing the data, we proceed to separate numerical and categorical columns based on their data types.

Moving on to analysis, we start with univariate analysis, which focuses on describing or summarizing one variable at a time.

Specifically, we begin by analyzing numerical variables. All numerical columns are iterated through using a for loop to find the skewness of each column.

Following this, we plot a histogram or distplot and a boxplot. The histogram displays the distribution of the data, while the boxplot defines the outliers in the data.

When handling outliers, there are three steps: retaining them for machine learning models, deleting them, or transforming them.

One common transformation method is the logarithmic transformation, which compresses large ranges of values into a smaller range.

Skewness impacts various statistical measures such as measures of central tendency, variance, and standard deviation. It can also affect the validity of assumptions underlying certain statistical tests.

Next, we analyze categorical variables using countplot or barplot, which display the count of each unique value.

Further, we move on to bivariate analysis, which involves describing or summarizing two variables and exploring the relationship between them.

Here, we plot a heatmap or pairplot to visualize the relationship between variables. It's better to use a heatmap because it clearly denotes the values of correlation.

Correlation measures the statistical relationship between two variables, indicating the extent to which changes in one variable are associated with changes in another variable. It only signifies the strength and direction of the relationship, ranging from -1 to +1.

The correlation coefficient (often denoted by r) quantifies the strength and direction of the linear relationship between two variables. The coefficient of determination, R^2, represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s), ranging from 0 to 1. It indicates the percentage of variation in the dependent variable explained by the independent variable(s).

From this plot, we select the dependent and feature engineering columns to make decisions. In the Diwali sales data, we select "orders" and "price" as dependent variables. This helps us forecast sales and determine the number of orders from different geographical locations.

Taking "order" as the dependent variable, we use the groupby method to find the number of orders from different states, zones, product categories, and product IDs. Then, we plot pie charts for each category.

Similarly, taking "price" as the dependent variable, we use the groupby method to find the net sales from different categories. Then, we plot a barplot for the same.

Finally, we conclude by summarizing the overall analysis and the insights gained.

In conclusion, after conducting a thorough exploratory data analysis (EDA) on the Diwali sales data, we have successfully prepared a comprehensive understanding of the dataset. Just as a chef prepares a meal to serve, we have carefully examined, cleaned, and analyzed the data to uncover valuable insights. Now, we're ready to present our findings, allowing stakeholders to make informed decisions based on our analysis. Just like serving a delicious dish, our analysis aims to satisfy the hunger for actionable insights, contributing to the success of future endeavors.