***Exploratory Data Analysis (EDA) Project Report***

***1. Introduction***

***Project Title****: Exploratory Data Analysis of Amazon Sales Dataset*

***Objective****:-*

* *The primary objective of analyzing the Amazon Sales Dataset is delve into product categories, prices, ratings, and sales patterns to identify characteristics that resonate with consumers and propel them to purchase.*
* *Delve into product categories, prices, ratings, and sales patterns to identify characteristics that resonate with consumers and propel them to purchase.*
* *Translate insights into actionable recommendations that optimize product development, inform marketing strategies, and boost your competitive edge.*
* *Equip businesses with the knowledge to create products that cater to evolving consumer needs and desires.*
* *Craft communication strategies that resonate with specific demographics and maximize engagement.*
* *Facilitate a marketplace where products find their perfect match in the hearts of consumers*

***Dataset****: The dataset used for this analysis includes the following columns:*

***Source:****This dataset is scraped from the official website of Amazon*

* *product\_id - Product ID*
* *product\_name - Name of the Product*
* *category - Category of the Product*
* *discounted\_price - Discounted Price of the Product*
* *actual\_price - Actual Price of the Product*
* *discount\_percentage - Percentage of Discount for the Product*
* *rating - Rating of the Product*
* *rating\_count - Number of people who voted for the Amazon rating*
* *about\_product - Description about the Product*
* *user\_id - ID of the user who wrote review for the Product*
* *user\_name - Name of the user who wrote review for the Product*
* *review\_id - ID of the user review*
* *review\_title - Short review*
* *review\_content - Long review*
* *img\_link - Image Link of the Product*
* *product\_link - Official Website Link of the Product*

***2. Methodology***

***2.1 Data Cleaning***

* ***Missing Values****: Identified and handled missing values by either filling them with appropriate values or removing rows/columns with excessive missing data.*
* ***Data Types****: Converted columns to appropriate data types (e.g., date and time columns).*

***2.2 Data Exploration***

* ***Descriptive Statistics****: Calculated summary statistics to understand the central tendency, dispersion, and distribution of the data.*
* ***Correlation Analysis****: Explored relationships between numerical variables to identify correlations.*

***2.3 Visualization***

* ***Histograms****: Used histograms to visualize the distribution of continuous variables like actual price,frequency.*
* ***Scatter Plots****: Created scatter plots to examine relationships between variables such as actual\_price vs. rating.*
* ***Heatmaps****: Attempted to visualize correlations between numerical variables (if applicable).*

***2.4 Correlation Analysis***

* ***Correlation Between Variables****: Heatmap analysis reveals that actual\_price has a strong/weak correlation with quantity.*

***3. Find Duplications and Analyse them***

*Removing duplicates is one of the most important part of the data wrangling process, we must remove the duplicates in order to get the correct insights from the data.*

*If you do not remove duplicates from a dataset, it can lead to incorrect insights and analysis.*

*Duplicates can skew statistical measures such as mean, median, and standard deviation, and can also lead to over-representation of certain data points.*

*It is important to remove duplicates to ensure the accuracy and reliability of your data analysis.*

***4. Visualizations***

***4.1 Scatter Plot of Actual vs. Predicted Values***

* ***Description****: This scatter plot compares actual vs. predicted values for rating and actual\_price.*

***4.2 Histogram of Total Price***

* ***Description****: This histogram displays the distribution of actual\_price with frequency.*

***4.3 Correlation Heatmap***

* ***Description****: This heatmap shows the correlation between numerical variables.*

***5. Grouping & Aggregation***

* *Calculate mean sales by product category*
* *Print mean sales by product category*
* *Calculate summary statistics for groups*
* *Create pivot tables*
* *Statistical testing*

***6. Conclusion***

* *The primary goal of this project is to analyze the Amazon Sales dataset and identify insights based on the data. The Amazon Sales dataset is a valuable resource for businesses and researchers alike. It provides a wealth of information about customer behavior, product trends, and market conditions. By conducting exploratory data analysis (EDA) on this dataset, businesses can gain valuable insights that can help them make better decisions about their products, marketing, and operations.*
* *Recommendations: Based on the analysis, it is recommended to focus on improving offerings in popular pizza categories and consider pricing strategies to increase sales.*

***6. Future Work***

* ***Further Analysis****: Investigate other features such as customer demographics or time-based trends.*
* ***Advanced Modeling****: Apply predictive modeling techniques to forecast future sales and customer behavior.*

***7. References***

* ***Tools Used****: Python, Pandas, NumPy, Matplotlib, Seaborn*