Name : Maxwel mwala

Reg no : BSDS 2322

DATA WRANGLING AND EXPLORATORY ANALYSIS PROJECT

PROJECT TITTLE : **Analysis of online retail sales.**

**AIM AND DESCRIPTION ABOUT THIS PROJECT**

The aim of this project is to perform data wrangling and exploratory data analysis on the online retail dataset to answer key business questions and provide actionable insights. This includes identifying trends, understanding customer behavior, and segmenting customers based on their purchasing patterns. Online sales analysis involves examining historical sales data from an e-commerce platform to gain insights into customer behavior, product performance, and market trends. This analysis can help optimize pricing strategies, identify popular products, and improve marketing efforts. By leveraging techniques such as time series analysis and machine learning, the goal is to enhance the overall online shopping experience and drive higher sales and customer satisfaction.

**DATASET AND META DATA**

The name of data set used in this project is Amazon online sales data set. The source of this data set is kaggle . kaggle is well known platform where valid data sets are uploaded from different domain fields. In terms of timely this data set was uploaded a year ago. This data has 1466 records and 16 features. Some of features inclulde product id , product name , actual price , discount rate and category of the products ,review ID, user id, review id, review tittle and product link and many more . we have this kind of data set( Amazon) because Amazon is one of the big companies that sale products online.

**LIBRARIES USED**

**1 *pandas* :** data manipulation and analysis.

**2 *numpy* :** numerical operations and calculations.

**3 *matplotlib* :** for data visualization and plotting.

**4 *seaborn :*** Enhanced visualization and statistical plots.

**5 *Scipy* :** Scientific computing and advanced mathematical operations.

**STEPS INVOLVED TO PEFORM DATA WRANGLING AND EXPLORATORY ANALYSIS**

1. **Data cleaning.**

* In this step we perform cleaning processes like handling missing values , removing duplicates, replacing missing values and correcting data types .
* libraries involved are pandas as pd for stastical description using describe() method, dropnal() for removing missing values, fillnal() for filling missing values, head(10) and tail(10) for data set observation, identify areas of interest and over view in the data set, isnull.sum() for counting missing values, isnull.head() viewing rows where there is null values in the column, df.duplicates().any() to see if there is any duplicates values to insure reliability of data analysis and wrong skewness of and standard deviation, df.columns to show features only , fillna.median() to fill missing values with median e.g in rating account column, missing\_percentage.plot(kind='bar') to show percentage of missing values , df.isnull().sum().sort\_values(ascending = False) to finally check if there is any missing values and df.sum().sum() to check entire data set.

1. **Data transformation**

In transformation we change data types of variable from object to float data types. In this process it done two functions that is replace() and astype() to convert to float data types. Df.info() is used to show features data types. This will help to transform to correct data type for analysis. And we use StandardScaler() from sklearn to standardize our dataset.

* A feature called discounted price has been removed from the data set since it have same values or is redundance of discount percentage, user name, user id ,image link, item link, these features also have been removed because it does not add to our analysis.
* We have also change data types of categorical features such product name into numerical for easy computation. That feature engeenering.
* I have also remove certain characters in the some features like, euro symbol on price, % symbole on discount price this help easy mathematical operations.
* We have added new features like revenue and total revenue into a data set. this will help the company to predict their level of busness growth after specific period of time.

**3 . Exploratory data analysis.**

In the Explatory data l hve use different visualization like Scatter plots , histogram and correction analysis using matplotlib library and heatmap using seaborn library for correction analysis such as visualizing correlations if is negative or positive correlated, spearman correlation. I have also group and aggregate using groupby() function to calculate summary statistics for groups. For statistical test I have use ttest\_ind() to perform two independent two sample test. This is used to group means , hypothesis testing and assumption checking. df.infor() is used in the code to summary of dataset such as class type, index range, list of columns in the dataset. Contagious table to display frequency distribution of categorical variables , this makes it easier for comparison and analysis of different categories.

**KEY INSIGHTS FROM VISUALISATION**

* Total revenue after discount and rating have a weak positive correlation. This means that products with higher revenue after discount have lower ratings , but relation is not very strong.
* Distribution between total revenue after discount and actual price is positively skewed. That shows that their an opportunities due to increase discounted prices or discounted percentages to attract more customers.
* Visualization from the Scatter plot also shows that there is positive correlation between actual price and rating of each product. This shows that their more increase in demand in products with small prices hence higher rating.
* From our finds most popular products is Fire-boltt Ninja smart watch is the most popular followed by phoenix smart watch .
* Top rating products shows that is coming from technology related domains
* from box plot we can see that their is outliers on goods rating that their are more outliers this indicate that their are some products that dont buyed frequently.

**RECOMMENDATIONS.**

* When demand is low on certain products or time taken for a product to be purchased it better to put some discounts on those products hence higher ratings and improve overall sales within specific period of time.
* Most of top rating product is from tech domains like tablets , adapters, these products has to be priotised in meeting customer satisfaction.
* Products that don’t have good rating , its better to drop them for maxmum capital utilization.
* the output suggest that their may be oppurtunities to increase discount percentage to attract customers .

**FILES IN THE ZIP FOLDER**

Original data set

Cleaned data set

Transformated data set

Jupyter source file contain all steps