Text

Description automatically generated

**MSc Information Technology** **and Data Analytics**

**COMP11122 – Data Minning & Business Intelligence.**

**Coursework 2nd Project.**

Table of Contents

[**Introduction** 3](#_Toc204348681)

[**Data Selection and Justification** 3](#_Toc204348682)

[**Data Cleaning and Preparation / Preprocessing** 4](#_Toc204348683)

[**Data Analysis and Visualization** 6](#_Toc204348684)

[**Regression Output Summary** 12](#_Toc204348685)

[**Interpretation of Regression Output** 12](#_Toc204348686)

[**Conclusion** 13](#_Toc204348687)

[**Clustering Analysis: KMeans on Quantity and Price** 14](#_Toc204348688)

[**Summary.** 15](#_Toc204348689)

**Introduction.**

Customers can now easily and conveniently buy goods and services online thanks to the growth of e-commerce in Pakistan. With a focus on consumer purchasing patterns, product preferences, payment options, and order distribution by region, this dataset-driven report provides an analytical overview of Pakistan's e-commerce transactions. This report finds valuable insights to support data-driven business decisions through extensive data cleaning, visualization, regression modelling, and clustering analysis. More than 1000 transaction records from various cities are included in the dataset, which includes variables like product category, payment method, quantity, price, and customer location. Using a variety of methods, such as statistical summaries, machine learning models, and data visualization, the aim of this analysis is to uncover hidden patterns in the data. The report's ultimate goal is to show how data science tools can be used practically to comprehend and improve e-commerce performance in a developing digital economy like Pakistan.

**Data Selection and Justification.**

This project's dataset reflects actual e-commerce transactions made in Pakistan. Important details like the invoice number, product category, subcategory, quantity, price, payment method, date, and customer city are all included. Because it provides a thorough understanding of consumer behaviour, product trends, and payment preferences in the Pakistani online retail market, this dataset was selected.

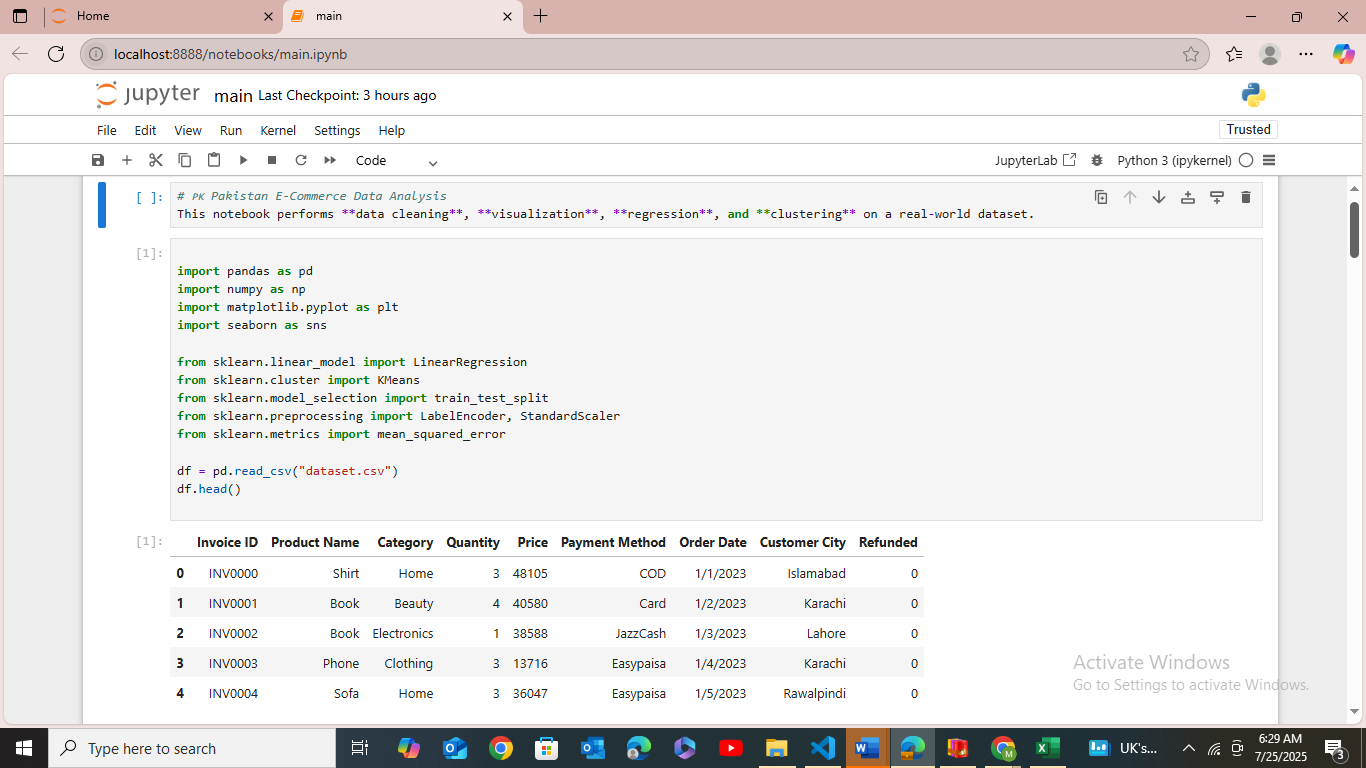
Every variable in the dataset has a distinct analytical function. For Example.

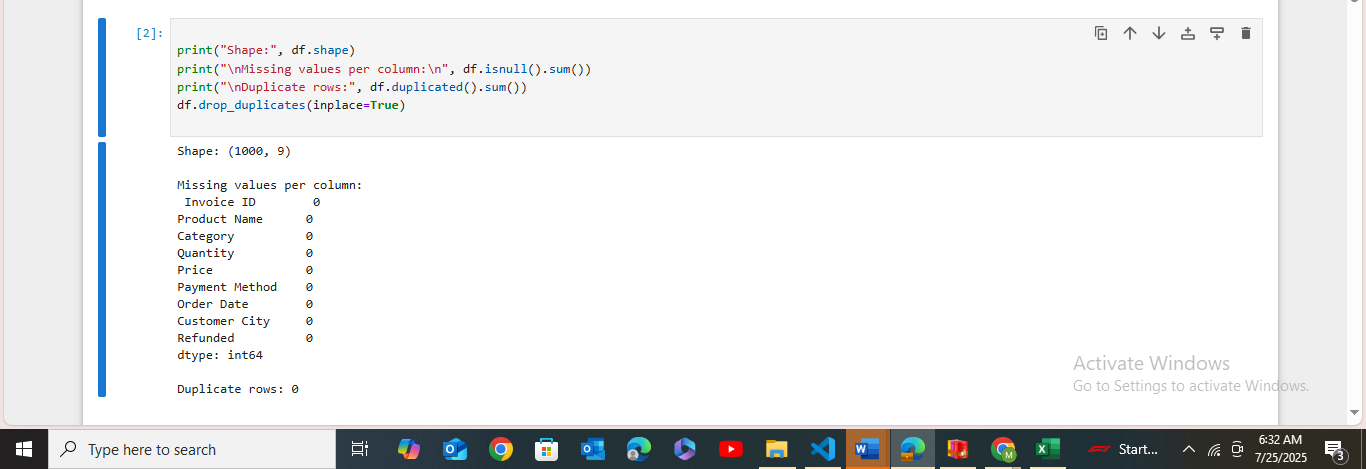
* **Product Category** and **Sub-Category** help in identifying top-selling product types.
* **Quantity** and **Price** are essential for sales trend analysis and regression modeling.
* **Payment Method** provides insights into customer payment behavior.
* **Customer City** enables geographical analysis and segmentation.

The dataset's moderate size (more than 1000 rows) made it suitable for clustering, regression balance between simplicity and richness made it a justified and practical choice for meeting the learning objectives modeling, and exploratory analysis without requiring a lot of processing power. Furthermore, the presence of categorical and numerical variables made it ideal for applying label encoding, visualizations (like count plots and histograms), and machine learning models (like linear regression and KMeans clustering). [C:\Users\Dell\Downloads\dataset.csv](file:///C:\Users\Dell\Downloads\dataset.csv) Click on the blue Link to View the Data Set.

**Data Cleaning and Preparation / Preprocessing.**

We carried out a number of preprocessing procedures after choosing the dataset to guarantee its quality and analytical suitability. Using df. isnull(). sum(), we first checked the dataset for missing values, but we couldn't find any. To preserve data integrity, we then looked for duplicate entries and eliminated them using df. drop duplicates(inplace=True).To prepare the data for machine learning models, we applied **Label Encoding** to categorical variables such as *Category*, *Payment Method*, *Customer City*, and *Product Name*. This transformation converted string values into numeric format required by regression and clustering algorithms. Additionally, summary statistics and data types were reviewed using df.describe() and df.info() to understand the structure and content of the dataset before proceeding with further analysis.



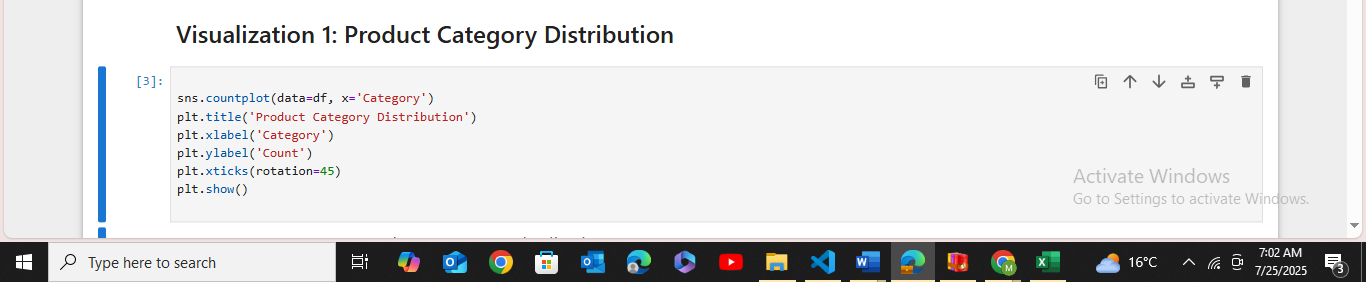


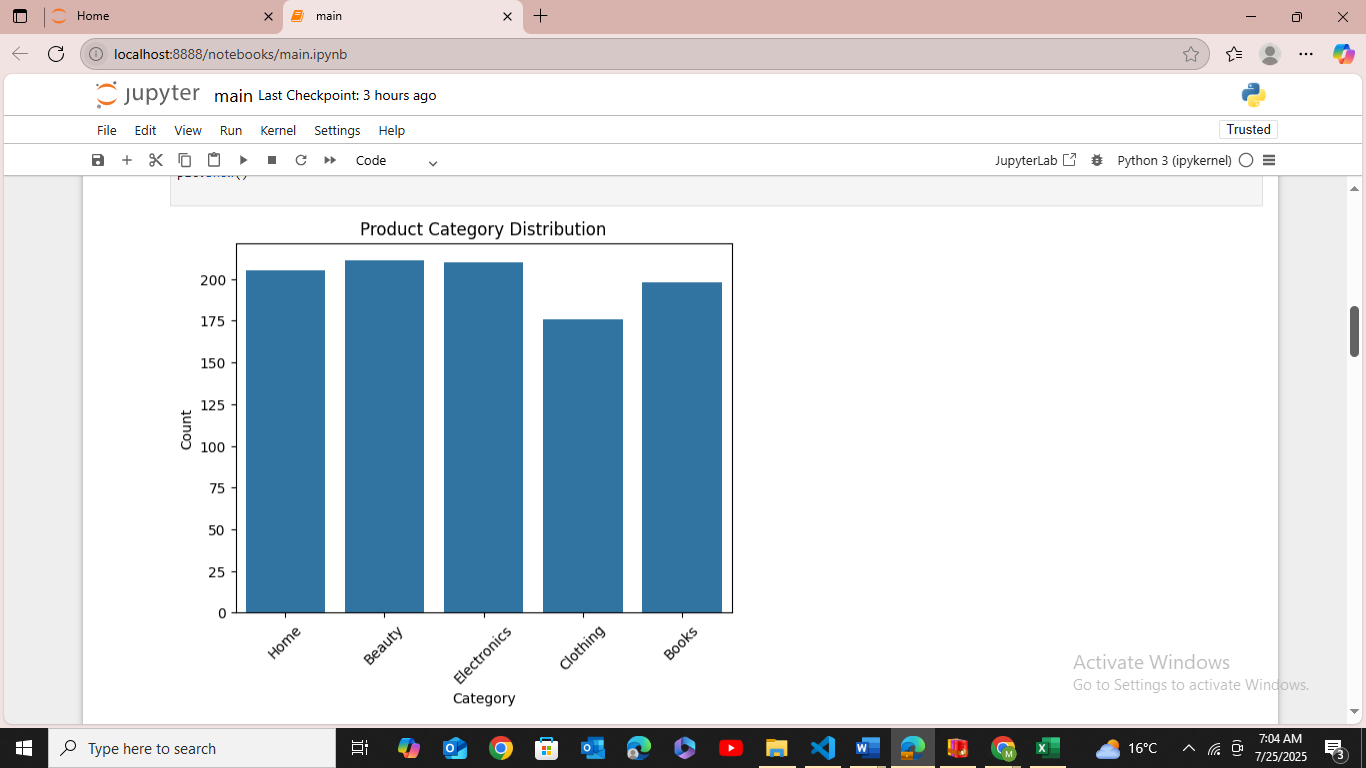
Explanation.

The dataset was imported into a Pandas Data Frame using the read\_csv() function to initiate the analysis process. To gain an initial understanding of the dataset's structure, the first few rows were displayed using the head() method. The dataset consisted of 1000 rows and 9 columns, including important attributes such as Invoice ID, Product Name, Category, Quantity, Price, Payment Method, Order Date, Customer City, and Refunded. A shape check confirmed the total size of the data as (1000, 9). To ensure the reliability of the dataset, a check for missing values was conducted using the isnull().sum() method, which revealed that there were no null entries across any of the columns. In addition, duplicate records were checked using the duplicated().sum() function, and the result showed zero duplicate rows. Although no duplicates were detected, the drop\_duplicates() method was applied as a precautionary measure. These initial cleaning steps were essential in preparing the dataset for the subsequent phases of visualization, regression, and clustering, ensuring that the data used for analysis was complete, accurate, and free from redundancy.

**Data Analysis and Visualization.**

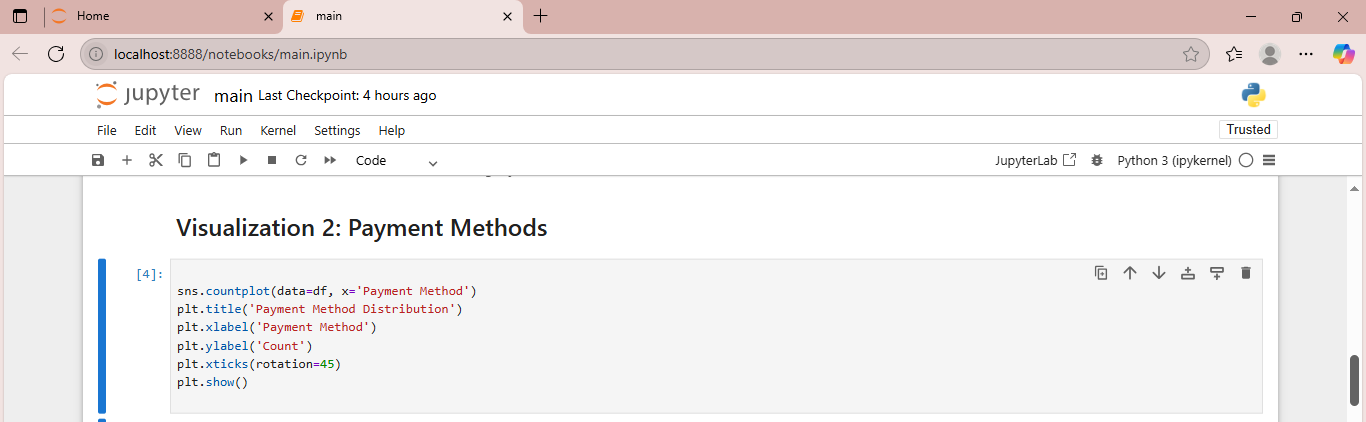
Several visualizations were made using Python libraries like Matplotlib and Seaborn in order to comprehend the distribution and relationships within the dataset. To display the distribution of products across various categories, a count plot was first created. The most and least common product types were identified with the aid of this visualization. To see the distribution and skewness of pricing in the dataset, a Price histogram was plotted, showing groups of items with low and high prices. A count plot for Payment Methods was made in order to further investigate categorical data, demonstrating that certain payment methods, such as JazzCash and Easypaisa, were utilized more frequently than others. In order to determine which cities had the most orders, a count plot was also used to visualize the customer distribution by city. To compare average product prices across categories, a bar plot was created for average price per product category. The foundation for the subsequent regression and clustering analysis was laid by these visualizations, which offered crucial insights into sales trends, consumer preferences, and product pricing behavior.

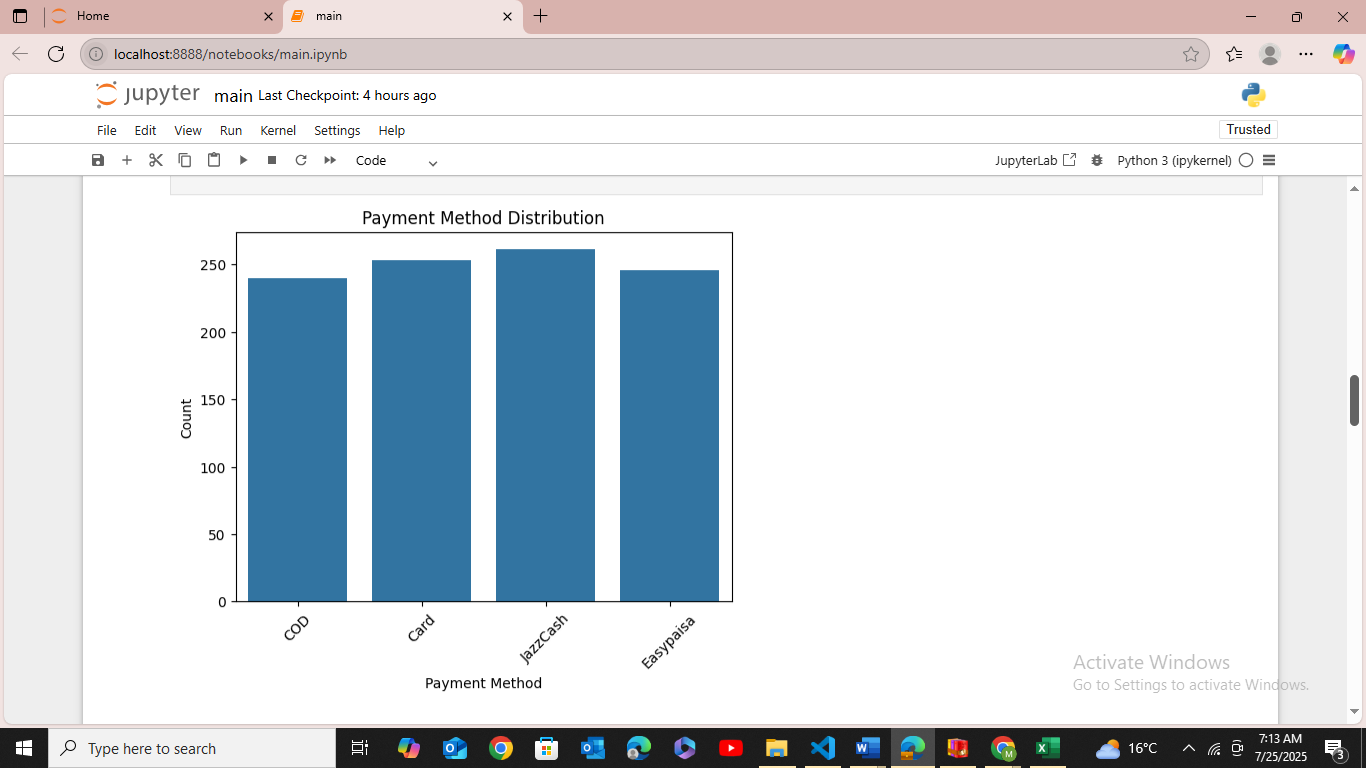




**Explanation**

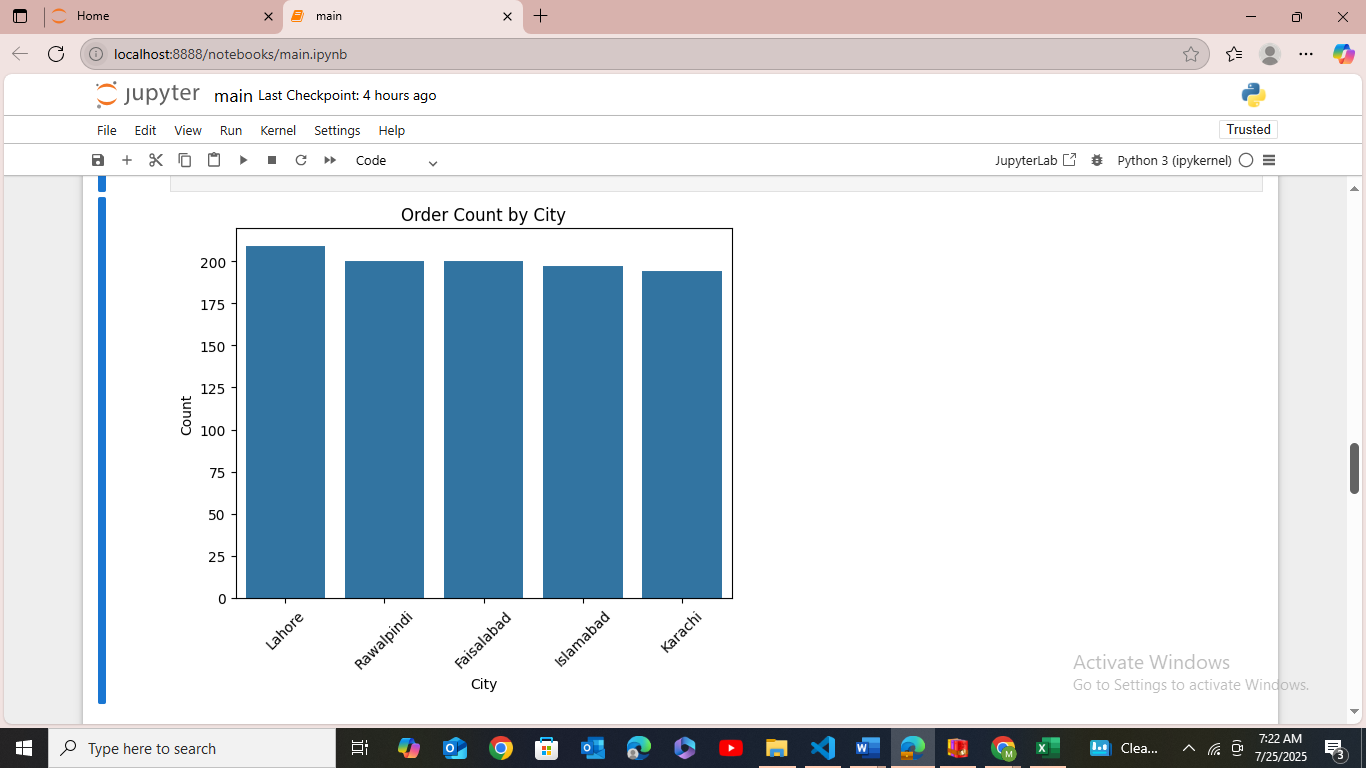
In the first visualization of our data analysis, we examined the distribution of products across different categories using a count plot from the Seaborn library. The x-axis represents various product categories such as *Home, Beauty, Electronics, Clothing,* and *Books*, while the y-axis shows the count of products sold in each category. This chart helps us identify which categories are more frequently appearing in the dataset. From the bar heights, it is evident that categories like Beauty and Electronics are among the most popular, whereas Clothing has the least count among all. This distribution insight can be helpful for understanding customer preferences and guiding inventory or marketing decisions. The graph also used rotated x-axis labels for better readability.





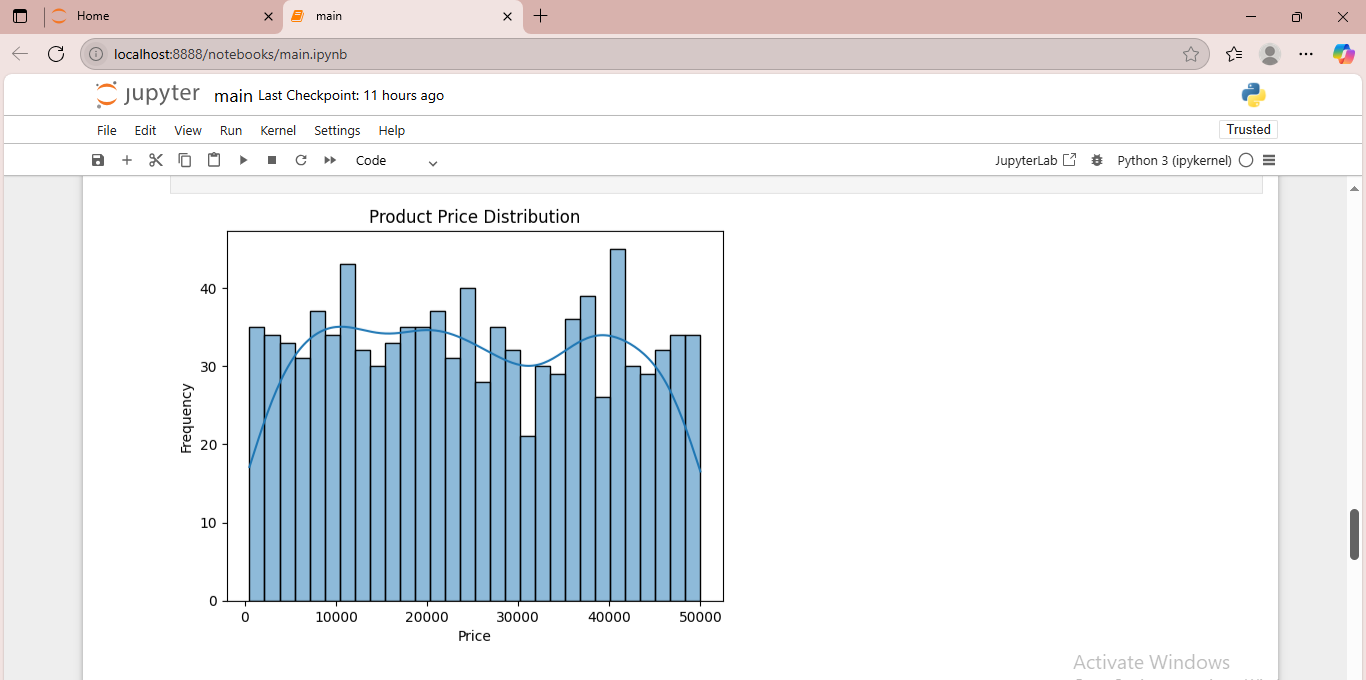
**Explanation.** In the second visualization, we explored the distribution of payment methods used by customers in the dataset. Using a count plot, we observed that JazzCash and Card payments were slightly more popular, followed by Easypaisa and Cash on Delivery (COD). This analysis helps us understand the customers’ payment preferences and can be useful for tailoring future e-commerce services or promotional strategies based on commonly used payment modes.





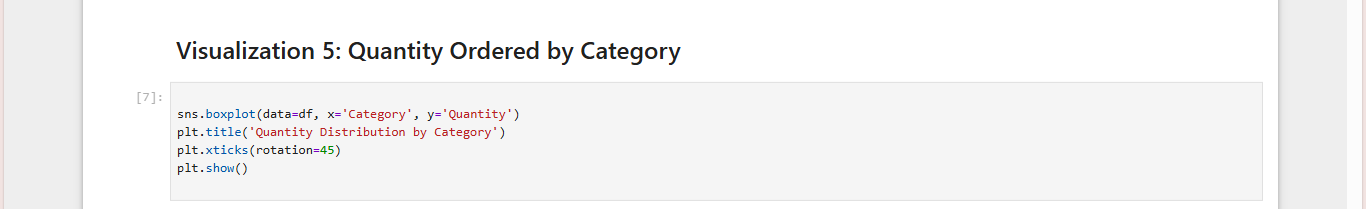
**Explanation.** In the third visualization, we analyzed the distribution of orders across different customer cities using a count plot. The X-axis represents various cities, including Lahore, Rawalpindi, Faisalabad, Islamabad, and Karachi, while the Y-axis indicates the number of orders from each location. According to the results, Lahore had the highest number of orders, followed closely by the other cities with relatively similar counts. This suggests that the customer base is spread fairly evenly across major cities in Pakistan, with Lahore being slightly more active. Such insights are useful for regional targeting and delivery optimization.

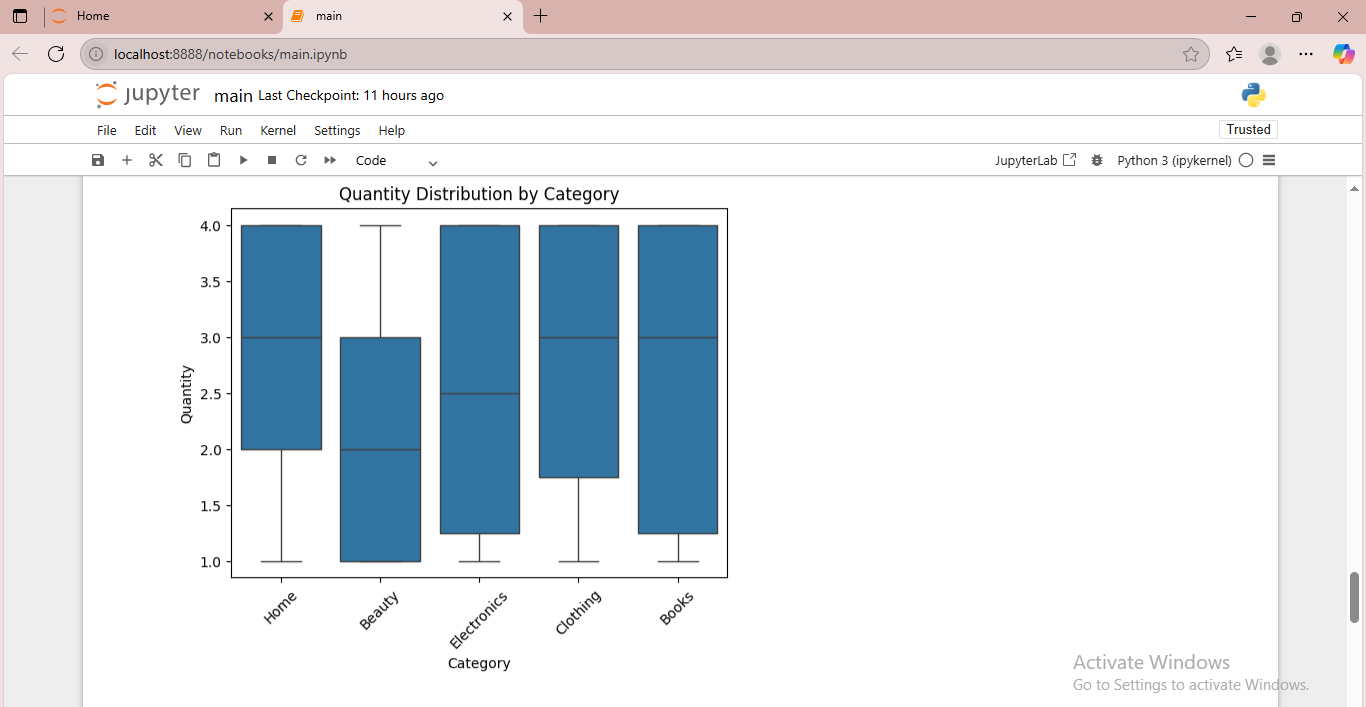




**Explanation.**

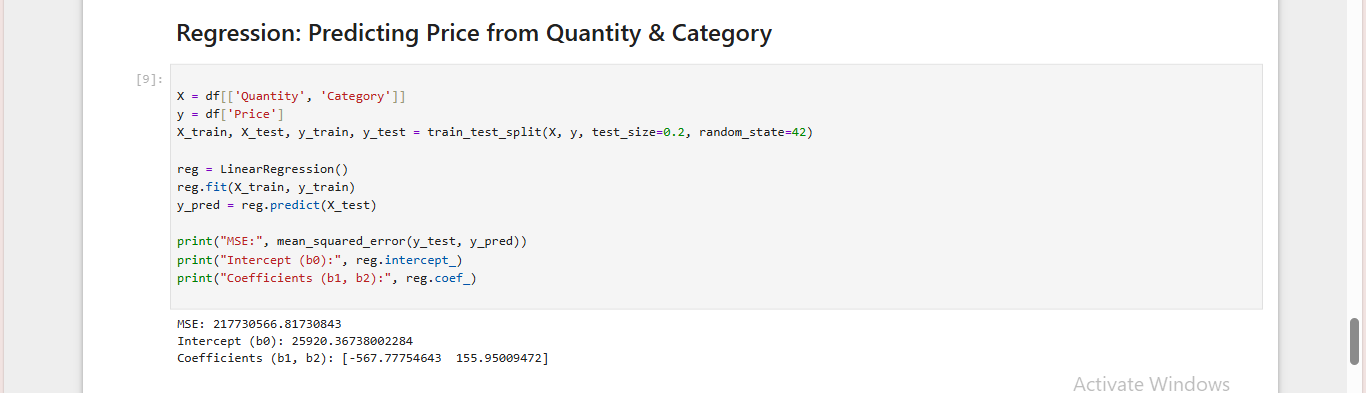
In this analysis, we used a histogram to visualize the distribution of product prices in the dataset. The histogram shows how frequently products fall into different price ranges, with the x-axis representing the price and the y-axis representing the frequency. We set the number of bins to 30, which divides the price data into 30 intervals, and enabled the kernel density estimate (KDE) line to show the overall trend. This visualization helps identify common pricing ranges and detect any skewness in the data. From the graph, we observed that prices are fairly evenly distributed across different ranges, indicating a balanced pricing strategy in the dataset.





**Explanation.**

This boxplot illustrates the distribution of product quantities ordered within each category. The categories shown include Home, Beauty, Electronics, Clothing, and Books. Each boxplot displays the spread of order quantities, highlighting the median, interquartile range, and potential outliers. From the graph, we can see that **Home, Clothing, and Books** categories have a higher median quantity ordered (around 3), while **Beauty and Electronics** have slightly lower medians (around 2 to 2.5). The vertical span of the boxes indicates that quantity values are more consistent for **Beauty** and **Electronics**, whereas categories like **Books** and **Clothing** have wider distributions, suggesting more variability in the number of items ordered per transaction.



# **Regression Output Summary**

|  |  |
| --- | --- |
| **Metric** | **Values** |
| Mean Squared Error (MSE) | 217735066.81730843 |
| Intercept (b₀) | 25920.36738002284 |
| Coefficient for Quantity | -567.77754643 |
| Coefficient for Category | 155.95009472 |

**Interpretation of Regression Output.**

In this linear regression model, the goal was to predict the Product Price based on two independent variables: Quantity and Category.

1. Mean Squared Error (MSE): 217,735,066.81

* This is a measure of the average squared difference between the actual prices and the prices predicted by the model.
* A high MSE indicates that the model’s predictions are not very close to the actual values.
* Since product prices can be high (in thousands), a large MSE value is expected, but this value still suggests that the model has limited prediction accuracy.

**Intercept (b₀): 25,920.36**

* The intercept represents the estimated price when both **Quantity** and **Category** are zero (baseline level).
* In real-world terms, this might not be meaningful (as you can't have zero quantity or undefined category), but mathematically, it acts as the starting point for price calculation.
* It means that the **base predicted price** starts at approximately **25,920 PKR**, and then adjustments are made based on quantity and category.

**Coefficient for Quantity: -567.78**

* This coefficient is negative, which means that as the quantity ordered **increases**, the predicted product price **decreases slightly**.
* Specifically, for each additional unit in quantity, the price decreases by approximately **568 PKR**.
* This could reflect **bulk discounting**, where ordering more units reduces the price per product.

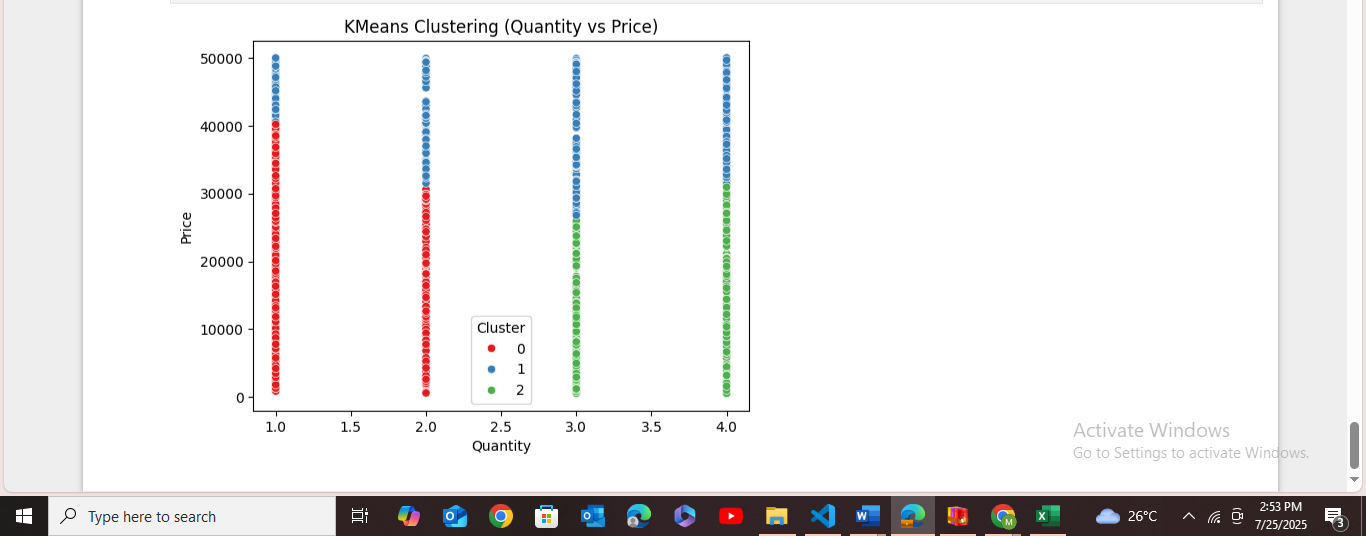
**Coefficient for Category: 155.95**

* This coefficient is positive, suggesting that **different product categories** have an increasing impact on price.
* Since "Category" is likely encoded numerically (e.g., 1 for Electronics, 2 for Clothing, etc.), each step up in category increases the predicted price by around **156 PKR**.
* This implies that **some categories are generally priced higher** than others.

**Conclusion**

* The regression analysis provides insight into how quantity and product category influence price... The **negative impact of quantity** suggests a possible discount trend, while the **positive category coefficient** indicates that more premium categories likely have higher prices.





# **Clustering Analysis: KMeans on Quantity and Price**

In this part of the analysis, the KMeans clustering algorithm was applied to identify patterns and groupings in the dataset based on **Quantity** and **Price**. Before clustering, the data was standardized using StandardScaler to ensure equal weightage of features during distance calculations.

We specified **3 clusters**, and the results were visualized using a scatterplot where:

* The **x-axis** represents the **Quantity** of items purchased.
* The **y-axis** shows the **Price** of the products.
* Each point is colored based on the assigned **Cluster** (0, 1, or 2).

From the plot:

* **Cluster 0 (Red)** primarily includes lower-priced items across all quantity levels.
* **Cluster 1 (Blue)** consists of mid-range to high-priced items, spanning various quantities.
* **Cluster 2 (Green)** is mostly made up of high-price products with higher quantity orders.

This clustering reveals that price and quantity have some meaningful segmentation in the dataset. The identified clusters can help in targeted pricing strategies, customer segmentation, and stock planning based on purchasing behavior.

# **Summary.**

This report analyzes Pakistan's e-commerce dataset containing over 1,000 real-world transactions to uncover customer behavior, pricing trends, and product patterns. The dataset includes features like Product Category, Quantity, Price, Payment Method, and Customer City. Initial steps involved data cleaning (handling missing values and duplicates) and transforming categorical data using label encoding for modeling purposes.

In the **data visualization** phase, several insights were uncovered using Seaborn and Matplotlib. Count plots highlighted that **Beauty and Electronics** were top-selling categories, while **Clothing** had fewer sales. A payment method analysis revealed **JazzCash and Card** were more popular, whereas **COD and Easypaisa** were less common. City-wise order distribution showed **Lahore** leading in customer activity. A histogram of product prices revealed a relatively balanced distribution, and boxplots showed quantity variation across categories, with **Books and Home** having higher average quantities per order.

A **linear regression model** was used to predict product prices using Quantity and Category as inputs. The model found that prices tend to **decrease slightly** with higher quantities (possibly due to discounts) and **increase with category type**, suggesting premium product categories cost more.

Finally, a **KMeans clustering algorithm** was applied on standardized Price and Quantity data. It grouped the dataset into **three clusters**: one with low-price items, another with mid- to high-price items, and a third combining high prices with higher quantities. These clusters highlight distinct consumer segments and can assist in **pricing strategies, customer targeting, and inventory management**.