|  |  |
| --- | --- |
|  | **Namal University Mianwali**  **Department of Computer Science** |

**DAV Project Phase 3**

**Data Cleaning (Missing Values, Duplicates, Outliers)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Course** | **Phase Lead** | **Group Members** | **Department** |
| Data Analysis and Visualization | Usaira Shahbaz (NUM-BSCS-2022-27) | Malik Zaheer Abbas (NUM-BSCS-2022-50)  Ahmad Mustafa (NUM-BSCS-2022-37) | Computer Science, Namal University, Mianwali |

### **Step 1: Import and Initial Cleaning**

**Objective:** Load the dataset and remove irrelevant or redundant features.

*import pandas as pd*

*df = pd.read\_csv("/content/FINAL.csv")*

*df.drop(columns=['review\_comment\_title'], inplace=True)*

Removed the review\_comment\_title column as it added no analytical value and had excessive missing values.

### **Step 2: Identify Missing Values**

We checked for missing data across all columns:

missing\_values = df.isnull().sum()

print(missing\_values[missing\_values > 0])

**Summary of Missing Values:**

* Missing values per column:

seller\_city 833

seller\_state 833

product\_category\_name\_english 2567

price 833

payment\_value 3

payment\_type 3

order\_delivered\_customer\_date 3421

review\_score 997

review\_comment\_message 69062

review\_creation\_date 997

dtype: int64

Total missing values in the dataset: 79549

### **Step 3: Handling Missing Values**

We followed a stepwise approach to handle null values logically:

#### ➤ **Step 3.1:** **Remove rows with missing review\_comment\_message**

**Reason:** It’s a key column for sentiment analysis; rows with no comment add little value. (We can’t analyze **sentiment** from an empty comment.)

*df.dropna(subset=['review\_comment\_message'], inplace=True)*

#### ➤ **Step 3.2: Drop rows missing key transactional fields**

**Columns:** seller\_city, seller\_state, product\_category\_name\_english, payment\_value

*df.dropna(subset=['seller\_city', 'seller\_state', 'product\_category\_name\_english', 'payment\_value'], inplace=True)*

#### ➤ **Step 3.3: Fill missing delivery dates using the **median** date**

**Reason:** Delivery dates are continuous time-based data; the median is robust against skewed values.

*df['order\_delivered\_customer\_date'] = pd.to\_datetime(df['order\_delivered\_customer\_date'])*

*median\_date = df['order\_delivered\_customer\_date'].median()*

*df['order\_delivered\_customer\_date'] = df['order\_delivered\_customer\_date'].fillna(median\_date)*

**Final Check:** All missing values were successfully handled.

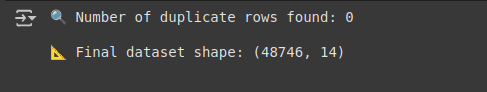
#### 

### **Step 4: Detecting of Duplicates Values:**

To ensure data integrity, we checked for duplicate records:

duplicates = df.duplicated().sum()

**Result:** As in our case no duplicate value found.



### **Step 5: Outlier Detection**

We used the **IQR (Interquartile Range)** method to detect outliers in numerical columns.

*for col in ['price', 'payment\_value']:*

*Q1 = df[col].quantile(0.25)*

*Q3 = df[col].quantile(0.75)*

*IQR = Q3 - Q1*

*lower = Q1 - 1.5 \* IQR*

*upper = Q3 + 1.5 \* IQR*

*df = df[(df[col] >= lower) & (df[col] <= upper)]*

#### **Outlier Detection**

Outliers were found in the following columns:

| Column | Number of Outliers |
| --- | --- |
| price | 402 |
| payment\_value | 543 |

These numbers were obtained by comparing each data point to the 1.5 \* IQR threshold for each numerical column.

### **Step 5: Outlier Handling (Impute method)**

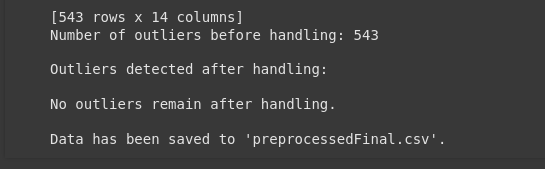
All identified outliers were removed from the dataset using conditional filtering based on IQR bounds.

*# Impute outliers with the median value*

*median\_value = df[column].median()*

*df\_cleaned = df.copy()*

*df\_cleaned[column] = df[column].apply(lambda x: median\_value if x < lower or x > upper else x)*

****

**Rationale:**  
Removing outliers helps in reducing skewness and enhancing model accuracy.

### **Step 6: Saving (CSV File):**

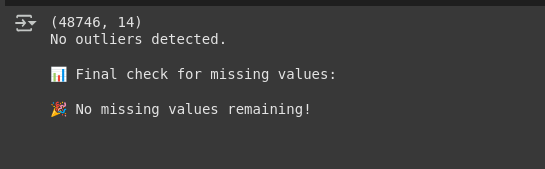
*# Save the cleaned DataFrame to a new CSV file*

*df\_cleaned.to\_csv('preprocessedFinal.csv', index=False)*

*print("\nData has been saved to 'preprocessedFinal.csv'.")*

**

### **Step 7: Final Confirmation (New Saved CSV File):**

****

### **Why These Steps?**

* **Removing rows vs filling values**: Removed when the missing data was excessive or crucial (like review\_comment\_message).
* **Filling with median**: Prevented bias that could occur with mean in skewed distributions.
* **IQR for outliers**: Robust method suitable for continuous features.