Komal

Exploratory Data Analysis (EDA) Report for Online Retail Dataset

Contents

[Exploratory Data Analysis (EDA) Report for Online Retail Dataset 1](#_Toc194181724)

[1. Introduction 1](#_Toc194181725)

[2. Data Loading and Overview 1](#_Toc194181726)

[2.1 Importing Libraries 1](#_Toc194181727)

[2.2 Reading the Dataset 1](#_Toc194181728)

[2.3 Summary Statistics 1](#_Toc194181729)

[3. Data Cleaning 2](#_Toc194181730)

[3.1 Handling Missing Values 2](#_Toc194181731)

[3.2 Removing Outliers 2](#_Toc194181732)

[4. Data Exploration 2](#_Toc194181733)

[4.1 Most Sold Products 2](#_Toc194181734)

[4.2 Top Countries by Sales 2](#_Toc194181735)

[5. Visualizations & Insights 3](#_Toc194181736)

[5.1 Boxplot of Unit Prices 3](#_Toc194181737)

[5.2 Histogram of Unit Prices 3](#_Toc194181738)

[5.3 Scatter Plot of High-Price Outliers 3](#_Toc194181739)

[5.4 Top 10 Most Expensive Transactions 3](#_Toc194181740)

[6. Key Findings 3](#_Toc194181741)

[7. Conclusion 3](#_Toc194181742)

# Exploratory Data Analysis (EDA) Report for Online Retail Dataset

# 1. Introduction

* This document presents an exploratory data analysis (EDA) of the Online Retail dataset. The dataset contains 541,909 rows and 8 columns, providing transactional records of an e-commerce company based in the UK.
* Dataset Source: [Online Retail Dataset](https://archive.ics.uci.edu/dataset/352/online+retail)
* Code Link: <https://github.com/komal7-alpha/PYTHON/blob/48f91110e1c869b56b4ecd61d169466d99392600/Exploratory%20Data%20Analysis/Online_Retail/OR_Komal.ipynb>

# 2. Data Loading and Overview

## 2.1 Importing Libraries

* The required Python libraries were imported:
  + pandas for data manipulation
  + seaborn and matplotlib.pyplot for visualization

## 2.2 Reading the Dataset

* The dataset was read using pd.read\_csv('OnlineRetail.csv') and stored in a DataFrame.

A screenshot of a computer

AI-generated content may be incorrect.

## 2.3 Checking Data Structure

* The df.info() function was used to inspect the dataset.
* The dataset consists of object, integer, and float data types.
* Missing values were observed in Description and CustomerID columns.

A screenshot of a computer program

AI-generated content may be incorrect.

## 2.4 Summary Statistics

* The dataset's numerical columns were analyzed using df.describe(). The table below presents key summary statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Quantity** | **UnitPrice** | **CustomerID** |
| Count | 541909 | 541909 | 406829 |
| Mean | 9.55 | 4.61 | 15287.69 |
| Std Dev | 218.08 | 96.76 | 1713.60 |
| Min | -80995 | -11062.06 | 12346 |
| 25% | 1.00 | 1.25 | 13953 |
| 50% | 3.00 | 2.08 | 15152 |
| 75% | 10.00 | 4.13 | 16791 |
| Max | 80995 | 38970.00 | 18287 |

# 3. Data Cleaning

## 3.1 Handling Missing Values

* The CustomerID column contains missing values, which may require imputation or removal based on the analysis purpose.
* The Description column also has missing values, which may affect item-level insights.

A screenshot of a computer

AI-generated content may be incorrect.

## 3.2 Removing Outliers

* Negative values in Quantity and UnitPrice indicate returns or incorrect data.
* Entries where Quantity <= 0 or UnitPrice <= 0 were removed to ensure data accuracy.

A graph of a number of objects

AI-generated content may be incorrect.

# 4. Data Exploration

## 4.1 Most Sold Products

* The table below shows the top 10 most sold products based on the total quantity sold:

|  |  |
| --- | --- |
| **Product Description** | **Quantity Sold** |
| WORLD WAR 2 GLIDERS ASSTD DESIGNS | 53847 |
| JUMBO BAG RED RETROSPOT | 47363 |
| ASSORTED COLOUR BIRD ORNAMENT | 36381 |
| POPCORN HOLDER | 36334 |
| PACK OF 72 RETROSPOT CAKE CASES | 36039 |
| WHITE HANGING HEART T-LIGHT HOLDER | 35317 |
| RABBIT NIGHT LIGHT | 30680 |
| MINI PAINT SET VINTAGE | 26437 |
| PACK OF 12 LONDON TISSUES | 26315 |
| PACK OF 60 PINK PAISLEY CAKE CASES | 24753 |

## 4.2 Top Countries by Sales

* The dataset includes transactions from multiple countries. The table below lists the top 5 countries by transaction count:

|  |  |
| --- | --- |
| **Country** | **Transaction Count** |
| United Kingdom | 495478 |
| Germany | 9495 |
| France | 8557 |
| EIRE | 8196 |
| Spain | 2533 |

# 5. Visualizations & Insights

## 5.1 Boxplot of Unit Prices

* A boxplot was created to visualize outliers in the UnitPrice column. The plot showed that most unit prices are concentrated between 0 and 5, while outliers extend up to thousands. These outliers were marked for further investigation.

A screen shot of a graph

AI-generated content may be incorrect.

## 5.2 Histogram of Unit Prices

* A histogram was used to understand the distribution of UnitPrice. The majority of unit prices fall below 50, while a few extreme values exist. A red dashed line was used to mark the lower limit, and a green dashed line marked the upper limit.

A graph with a line

AI-generated content may be incorrect.

## 5.3 Scatter Plot of High-Price Outliers

* A scatter plot was created to highlight high-price outliers. Most points lie within the normal range, while extreme values are concentrated above the green dashed line, indicating unusual prices.

A screen shot of a graph

AI-generated content may be incorrect.

## 5.4 Top 10 Most Expensive Transactions

* A bar plot was created showing the top 10 highest-priced transactions, with StockCode on the y-axis and UnitPrice on the x-axis.
* The green dashed line represents the upper limit for typical prices, helping to visually separate high-price outliers.

A graph showing the top 10 high price outliers

AI-generated content may be incorrect.

# 6. Key Findings

* The dataset contains transactions from 541,909 records, mostly from the United Kingdom.
* There are missing values in CustomerID and Description.
* The most sold product is "WORLD WAR 2 GLIDERS ASSTD DESIGNS" with 53,847 sales.
* The highest priced transaction in the dataset is £38,970.00.
* A significant number of transactions had negative or zero values in Quantity and UnitPrice, which were removed during cleaning.

A graph of stock codes

AI-generated content may be incorrect.

# 7. Conclusion

* The EDA process provided key insights into the dataset, including:
  + The presence of missing values and outliers.
  + The dominance of UK-based transactions.
  + The necessity of data cleaning to remove negative values.
  + Identification of best-selling products and most expensive transactions.
* Further analysis can be done for customer segmentation and revenue trends to improve business insights.