|  |  |
| --- | --- |
| **Electronic Sales** | Final Project |

**Description & Source**

*This dataset contains sales transaction records for an electronics company over a one-year period, spanning from September 2023 to September 2024. It includes detailed information about customer demographics, product types, and purchase behaviors we brought this data from Kaggle.com.*

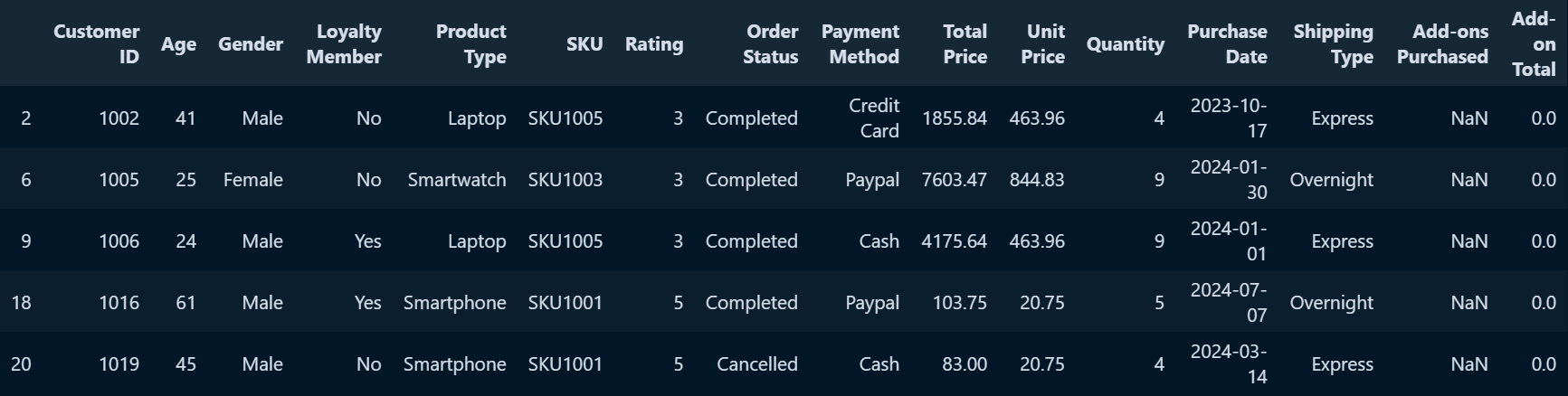
**Objectives**

1. ***What is the most sold product according to the gender?***
2. ***What is the most high-rated product?***
3. ***How does loyalty membership impact customer spending and add-on purchases?***
4. ***What are the trends in purchasing behavior based on shipping type and purchase dates?***
5. ***Which product types generate the most revenue, and how do add-ons contribute to overall sales?***

**Methodology**

# **Cleaning**

# *We’ve checked if there is any null or duplicated data. So, we found no duplicated rows but found many null data in two columns:* ***Gender -> 1*** *&* ***Add-ons Purchased -> 4868***

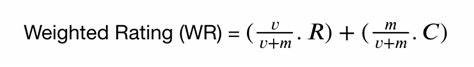
*  
First 5 rows that indicate NaN values in “Add-ons Purchased” column*

*From this picture we realized that the “Add-ons Purchased” is NaN when “Add-ons Total” = zero. So, there is no value in “Add-ons Purchased” and we can replace it with N/A value. Then, NaN values of Add-ons Purchased been solved.*

*The NaN value in “Gender” can be replaced with any Female/Male value, so we replaced it with Male. It wouldn’t effect distribution at all.*

1. **Top Rated Product**

*Each system need a function that calculates the top rated products to indicate the demographic filtering to recommend a list of Weighted Rating (WR) for each product. The formula to give a product rating is:*

*  
IMDB formula*

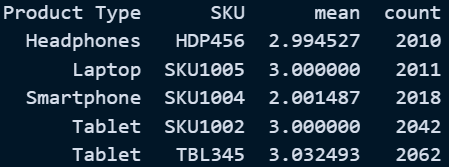
* + *v: is the number of votes for the product.*
  + *m: is the minimum votes required to be listed in the chart.*
  + *R: is the average rating of the product.*
  + *C: is the mean vote across the whole report.*

*First, we group ['Product Type', 'SKU'] by ['Rating'] to calculate the mean and the average for each product. This is what we gain:*

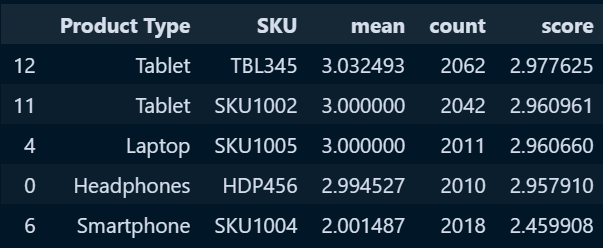
*Then, we calculate the C from all the products we preserved and the value of the average of all products is: 2.921*

*To calculate m (quantile) value by the 65% of the data, we founded that the minimum quantile to be listed in the rating of products is: 2005.4*

*From what we gained, we calculate the last value we need and it’s the q value, the qualified products in the list. We calculate these products by looking if the product’s count greater than the minimum quantile needed. This is the qualified products:*

*  
Qualified Products*

*Finally we can calculate the* ***Weighted Rating (WR)*** *scores for each of the qualified products we’ve gain by the previous formula of IMDB and sorting the values discerningly. So, this is the final result:*

*  
Top Rated Products*

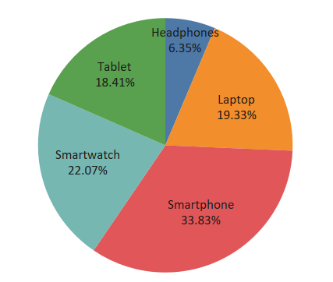
# **Analysis and RFM**

# *In this analysis, our team processed a dataset by applying various techniques to uncover key insights and trends. Below is a detailed overview of the steps and findings:*

1. *Data Sorting:*
   * *We sorted the dataset based on the purchase date, organizing the data from the most recent to the oldest transactions. This enabled a clearer understanding of purchasing patterns over time.*
2. *Boxplot Visualization:*
   * *Using Seaborn, we created individual boxplots for each column to visually assess the distribution of values, detect any outliers, and understand the spread of data for each feature. Each figure was plotted separately for clarity and detailed analysis.*
3. *Correlation Analysis:*
   * *A correlation matrix was computed to measure the relationships between all the columns in the dataset. Among the variables, the strongest correlation was observed between the Unit Price and Total Price columns, with a correlation coefficient of 0.674. This indicates a relatively strong positive relationship between the unit price of items and the total transaction value.*
4. *Conclusion:*
   * *These steps provided significant insights, particularly highlighting how changes in unit price influence the total price, which can help guide further business decisions or optimizations.*

# **Visualization and Insights**

# *In this project, our team utilized a variety of visual tools to uncover meaningful insights from the Customer Behavior dataset. Below are the key visualizations we implemented:*

1. *Total Sales and Year-on-Year Growth:*
   * *We created a calculated field for Total Sales of 2024 and a corresponding Year-on-Year Growth field. This allowed us to track performance trends and spot fluctuations over time.*
   * *Similarly, we calculated the growth in Total Products Sold, helping us understand product demand dynamics.*
2. *Product Type and Total Sales Pie Chart:*
   * *We designed a pie chart representing the distribution of total sales by product type, with percentages clearly indicating the contribution of each category. This gave us a snapshot of the product categories driving the most revenue.*
3. *Shipping Type and Purchase Date by Quarter:*
   * *A bar chart was created to analyze the relationship between shipping types and sales data based on the purchase date, segmented by quarter. This visualization helped us recognize sales peaks by quarter and the preferred shipping methods.*

*A graph of different colored bars

Description automatically generated*

1. *Purchase Date by Total Sales Line Chart:*
   * *To provide a continuous view of sales trends over time, we implemented a line chart illustrating total sales by purchase date. This line chart allowed us to identify periods of growth, dips, and patterns in consumer buying behavior*

A graph with lines and numbers

Description automatically generated

IIV. Future considerations

1. Expand loyalty programs to boost customer retention.
2. Further analysis of customer demographics could reveal deeper insights into buying behavior.
3. Implement targeted marketing campaigns based on customer segmentation