**Introduction**

**Problem statement:**

To understand customer purchasing behavior and identify trends and patterns in order data from a retail store that sells Men's fashion and Mobile & Tablets.

**Data description**

* The dataset is taken from Kaggle.
* The provided data is a transactional dataset capturing online customer purchases. It contains information on order ID, order date, item ID, SKU, quantity ordered, price, payment method, and customer information like name, gender, age, email, etc. Additionally, the data includes information on the order's status, such as received, complete, canceled, and order refunded, which can help to identify potential issues in the purchasing process.
* Moreover, the data includes details on the item categories purchased, allowing for the identification of popular items and trends among customers. The dataset can be used for various purposes, such as analyzing customer purchase behavior, identifying the most popular products, and understanding payment preferences.
* By analyzing this data, businesses can gain insights into the shopping preferences of their customers, optimize their sales strategy, and improve customer satisfaction. For example, businesses can identify which products sell better during certain times of the year, which payment methods are preferred, and which marketing strategies are most effective.

**Methodology:**

There are several data-cleaning tasks that need to be performed on this dataset. Here are some of them:

Convert date fields to a consistent format: The "order\_date" field is in the format "dd-mm-yyyy", but the "month" field is in the format "MMM-yy". We should convert all date fields to a consistent format, such as "yyyy-mm-dd".

Remove unnecessary columns: Some columns such as "bi\_st", "ref\_num", and "Discount\_Percent" seem to be unnecessary, so we can drop them from the dataset.

Convert categorical variables to consistent format: The "category" and "payment\_method" columns should be converted to a consistent format, such as lowercase.

Deal with missing data: We should check for missing data in the dataset and decide how to handle them. If there are missing values in critical fields such as "order\_id", we may need to remove those rows altogether.

Fix data types: Some fields such as "qty\_ordered" and "age" are numeric but are stored as strings. We should convert them to the appropriate data type.

Check for duplicates: We should check for and remove any duplicate rows in the dataset.

Preprocessing the data:

Text

Description automatically generated

Text

Description automatically generated

To effectively perform data cleaning on that dataset, the following steps could be taken:

• Removed any unnecessary columns that didn't provide meaningful information: Went through the data set and identified any columns that were irrelevant or redundant and removed them to streamline the data set.

• Checked for missing or null values and handled them accordingly: Scanned through the data set to identify any missing or null values. If any were found, either removed the entire row or replaced the missing value with an appropriate value (e.g., mean, median, or mode).

• Ensured that all column headers were formatted uniformly and were easy to read: Reviewed column headers to ensure they were standardized and easy to understand. This step improved the readability of the data set and reduced the risk of confusion when analyzing it.

• Checked for any duplicate rows and removed them if necessary: Identified any rows with duplicate data and removed them from the data set to prevent bias and inaccuracies in the data.

• Standardized the format of data in each column to ensure consistency: Reviewed each column to ensure that the data was formatted consistently, for example, ensuring that dates were formatted in the same way or that units of measurement were standardized throughout the dataset.

• Checked for outliers and errors in the data and handled them accordingly: Identified any data points that appeared to be outliers or errors and handled them accordingly. For example, removing or correcting these values could help prevent biases and inaccuracies in the data.

By following these steps, we ensured that the data set was clean, consistent, and ready for analysis.

**Design Process**

**Insights:**

To remove unnecessary columns and produce illuminating visuals, a Python analysis was first conducted. Next, the data was transferred to Tableau to design an interactive dashboard that includes seven visual representations, which offer a comprehensive and valuable perspective on the data. Our analysis highlights that Texas ranks as the highest revenue-generating state in the US, with California closely following suit. Additionally, our findings indicate that the month of December experiences the highest sales figures, primarily attributed to Christmas. Moreover, the age demographic of 30-40 showcases the highest revenue numbers. We have also generated revenue breakdowns per category, revealing that females generate higher revenue than males when purchasing mobiles and tablets. Additionally, our analysis indicates that the southern region boasts the highest revenue figures, accounting for 38.37% of total revenue. These are the trends that we have analyzed through our data exploration and visualization efforts.

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface, chart

Description automatically generated