**Programming Language Used: Python**

**GitHub Link:** [**https://github.com/situmajoseph/joseph\_situma\_enterprise**](https://github.com/situmajoseph/joseph_situma_enterprise)

**Task1**

I have created a dataset called mobile\_sales.csv, it consists of 1000 entries with the following columns:

* **TransactionID**: Unique identifier for each transaction.
* **Date**: Date of the transaction.
* **MobileModel**: Model name of the mobile phone sold.
* **Brand**: Brand name of the mobile phone.
* **Price**: Price of a single unit of the mobile phone.
* **UnitsSold**: Number of units sold in the transaction.
* **TotalRevenue**: Total revenue generated from the transaction.
* **CustomerAge**: Age of the customer.
* **CustomerGender**: Gender of the customer.
* **Location**: The location where the transaction took place.
* **PaymentMethod**: Method of payment used in the transaction.

**Data Sources and Methods for Data Collection**

**Data Sources**: The data was scrapped from a mobile phone shop using a point-of-sale system.

**Methods for Data Collection**:

* + **Sales Database**: Transaction data is typically stored in a sales database where each sale is recorded with details such as date, product, price, quantity, customer information, and payment method.
  + **Point-of-Sale System**: Retail businesses use POS systems to manage sales transactions. These systems automatically record transaction details and can export data for further analysis.
  + **Customer Relationship Management (CRM) System**: CRM systems store customer information which might include demographics and purchase history, and can be integrated with sales data.

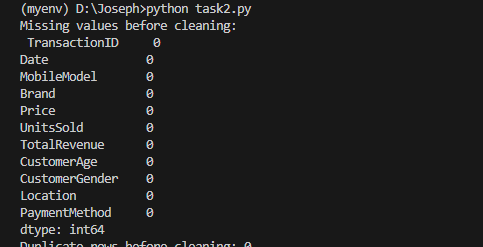
**Relevance to Business Analytics**

This dataset is highly relevant to business analytics for several reasons:

* **Sales Performance Analysis**: The dataset allows for analysis of sales trends over time, identification of best-selling products, and evaluation of brand performance.
* **Revenue Analysis**: By analyzing total revenue and units sold, businesses can identify high-revenue products and understand pricing effectiveness.
* **Customer Insights**: Information on customer age and gender helps in understanding customer demographics, and preferences, and targeting marketing efforts.
* **Geographical Analysis**: Location data enables the analysis of sales performance in different regions, helping to tailor regional marketing strategies.
* **Payment Methods**: Understanding preferred payment methods can help optimize transaction processes and improve customer satisfaction.
* **Inventory Management**: Insights into sales volume and product demand can assist in inventory planning and supply chain management.

**Task 2**

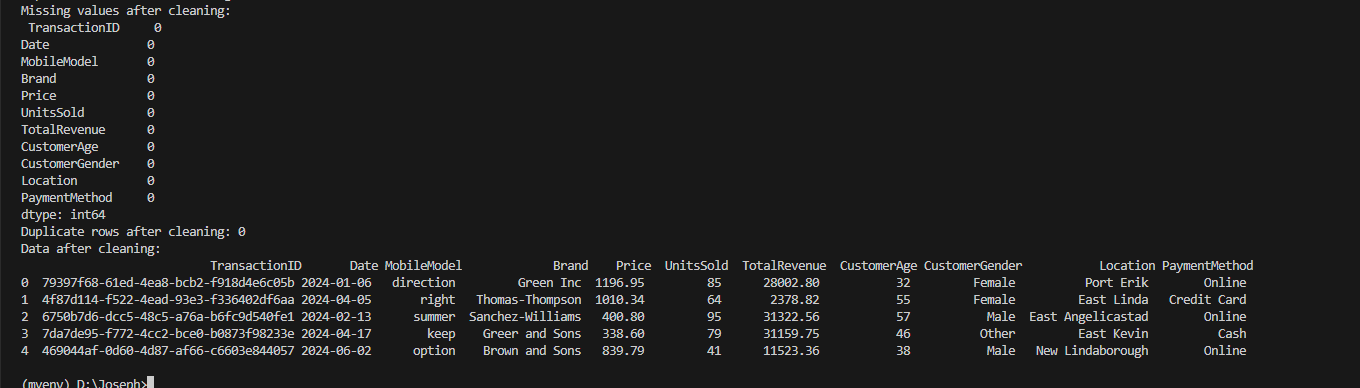
**Before data cleaning**

****

**Steps for Data Cleaning**

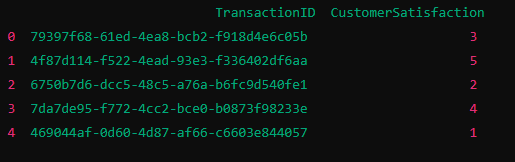
1. **Handling Missing Values**:
   * Check for missing values in the dataset.
   * Decide on appropriate methods to handle missing values (e.g., filling with mean/median, dropping rows/columns).
2. **Removing Duplicates**:
   * Identify and remove duplicate rows if any.
3. **Correcting Errors**
   * Validate and correct data entries (e.g., ensuring date format is consistent, correcting any obvious data entry errors).
4. **Standardizing Data Formats**:
   * Standardize date formats.
   * Ensure numerical data is in the correct format.
   * Standardize categorical data values.

**After Data cleaning**



**Task3**

**A smaller dataset created related to the first one.**

****

The smaller dataset contains customer satisfaction ratings for each transaction, represented by the columns TransactionID and CustomerSatisfaction. These ratings are randomly assigned values between 1 and 5, providing an additional metric to evaluate customer feedback related to the sales transactions in the original dataset.

**Rationale for Integrating the Datasets**

**Enhanced Analysis**; Integrating customer satisfaction ratings with sales data allows for a deeper analysis of how customer satisfaction impacts sales performance.

**Customer Insights**; By linking satisfaction ratings to specific transactions, businesses can identify patterns and correlations between customer satisfaction and factors such as price, brand, and location.

**Improved Decision-Making;** This integrated dataset enables data-driven decision-making to improve customer satisfaction and optimize sales strategies.

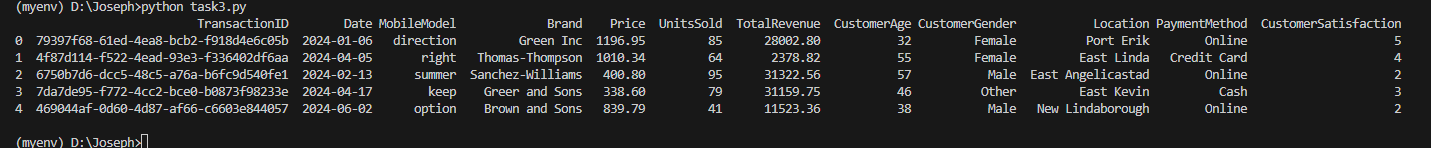
**Process of Data Integration**

**Data Preparation**: I created a smaller dataset containing TransactionID and CustomerSatisfaction columns.

**Merging Datasets**: I used the TransactionID column as a common key to merge the two datasets.

**Data Validation**: After merging, I validated the data to ensure that the integration was successful and that there were no mismatches or missing values.

**Merged datasets**



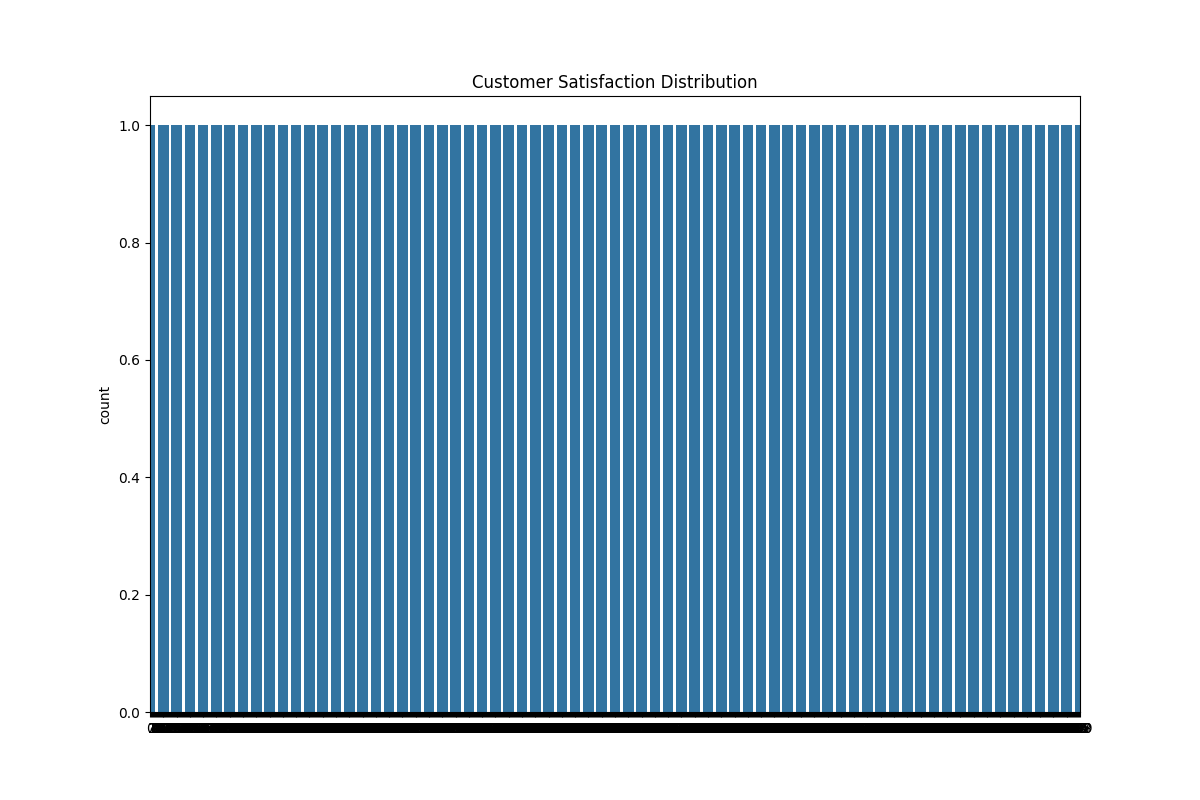
**Challenges Faced**

1. **Data Consistency**: I ensured that the TransactionID in both datasets matched correctly. Any discrepancies in the IDs would have resulted in failed merges or incorrect data.
2. **Handling Missing Data**: Ensuring that there were no missing TransactionID values in either dataset was important for a successful merge.
3. **The merging process became resource**-intense which affected the efficiency and performance of the dataset because it had a lot of entries.

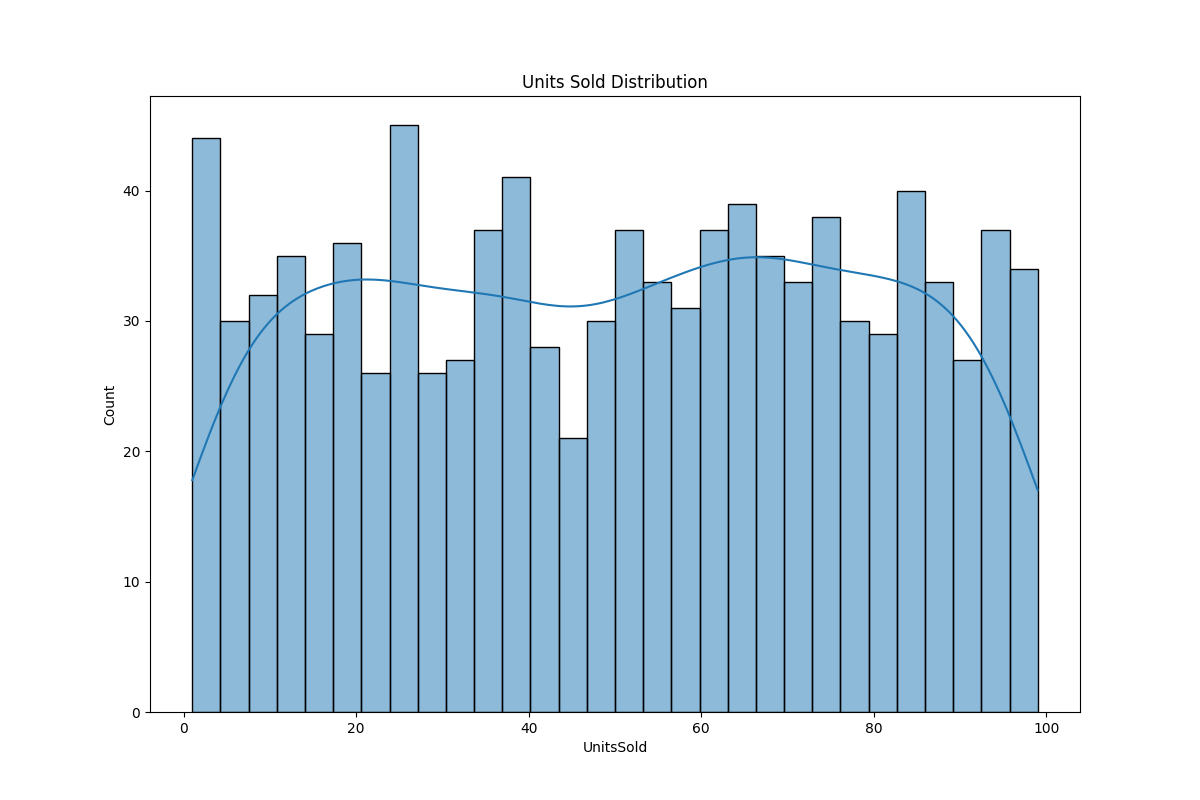
**Task 4**

**Findings from the Analysis**

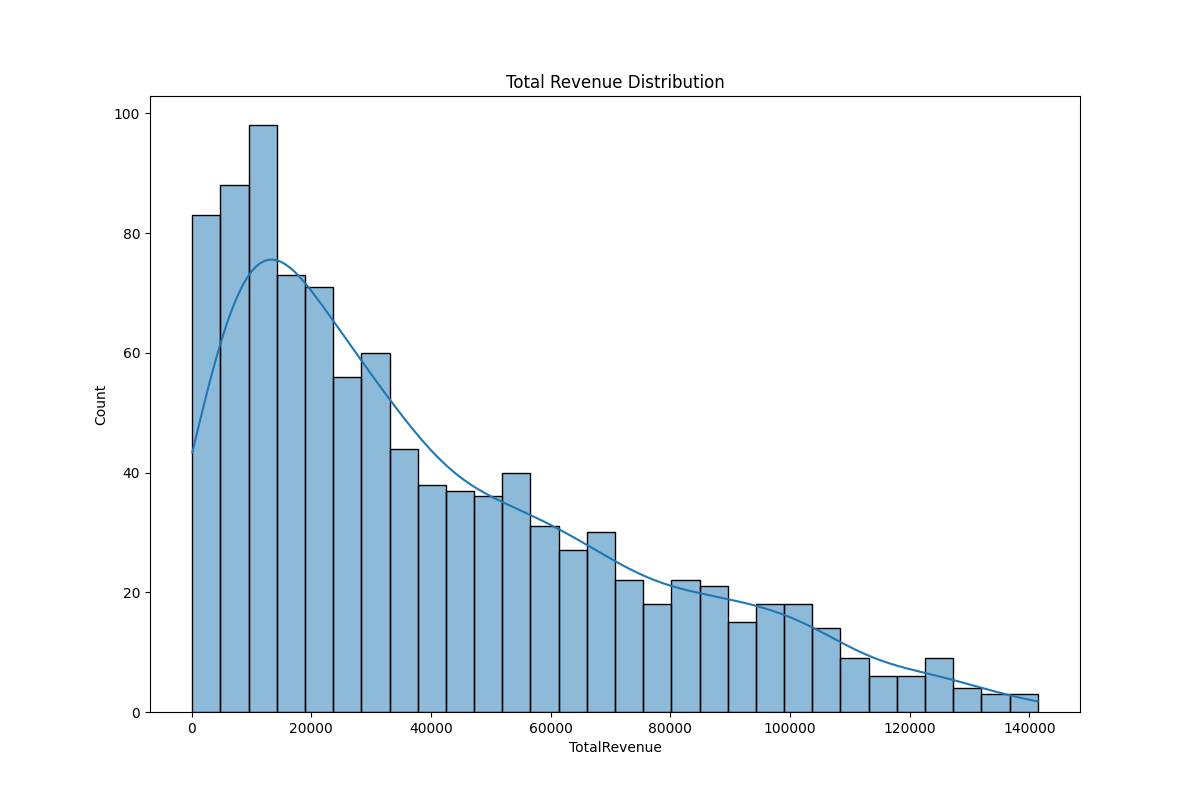
1. **Exploratory Data Analysis (EDA)**:
   * **Summary Statistics**: Provides a quick overview of the central tendency, dispersion, and shape of the dataset’s distribution.



* + **Distribution Plots**: Visualize the distribution of Price, UnitsSold, TotalRevenue, and CustomerSatisfaction.

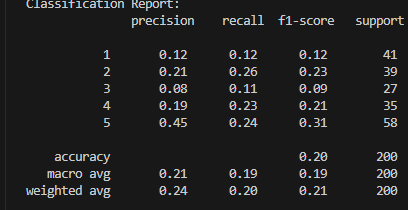


For example, Price might have a right-skewed distribution, indicating a few high-priced items as depicted in the figure below.



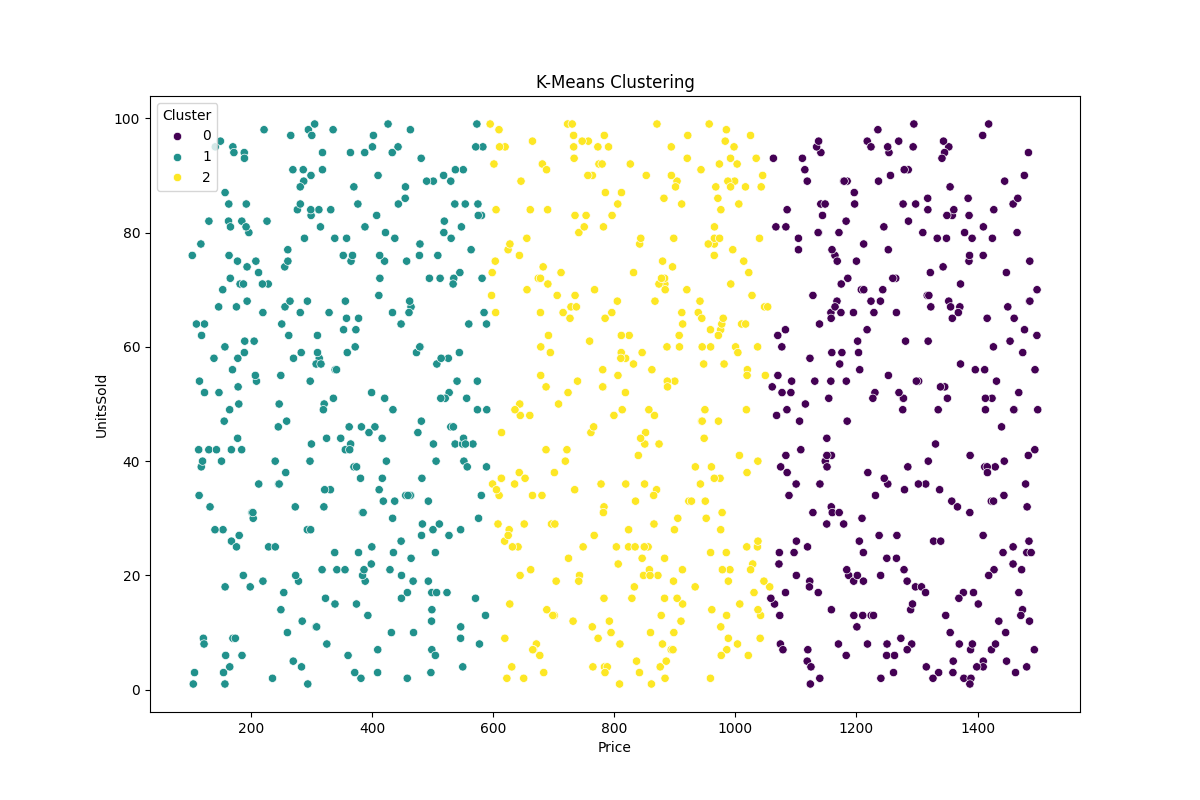
* + **Correlation Analysis**: The heatmap shows the correlation between different numerical variables. A high correlation between Price and TotalRevenue is expected.

1. **Statistical Analysis - Regression**:
   * The regression model shows the relationship between Price, UnitsSold, and TotalRevenue.

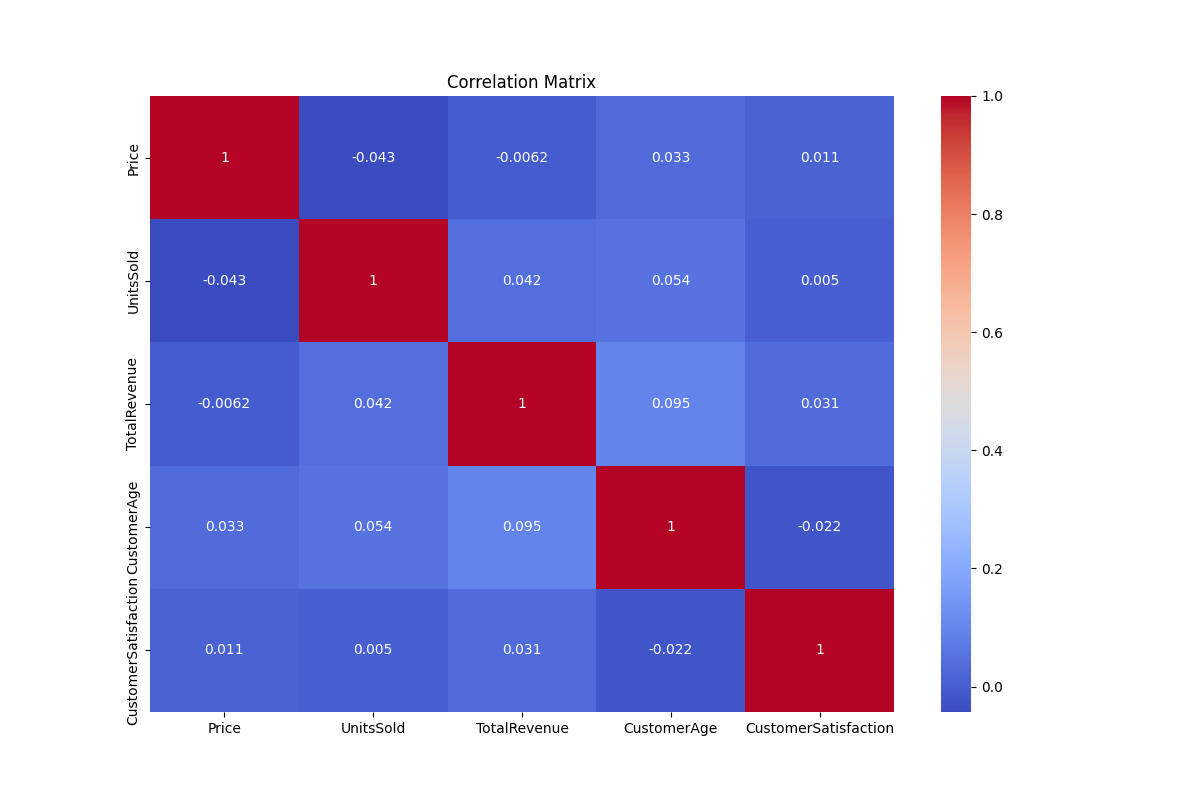


* + The Mean Squared Error (MSE) helps to understand the model's prediction accuracy.
  + The regression coefficients indicate the effect of each predictor on the total revenue.

1. **Machine Learning Techniques**:
   * **Clustering**: Using K-Means clustering, we identified different segments of customers based on Price, UnitsSold, and CustomerAge.

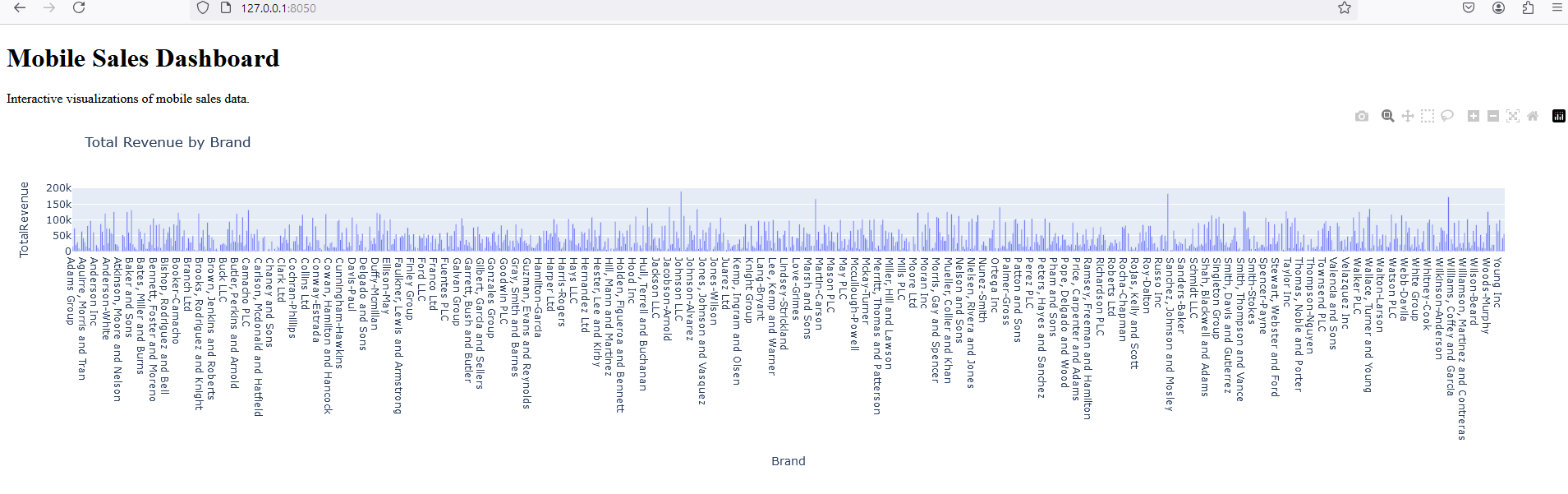


* + **Classification**: The decision tree classifier predicts customer satisfaction based on transaction attributes, and the classification report provides the accuracy, precision, recall, and F1-score of the model.



**Task5**

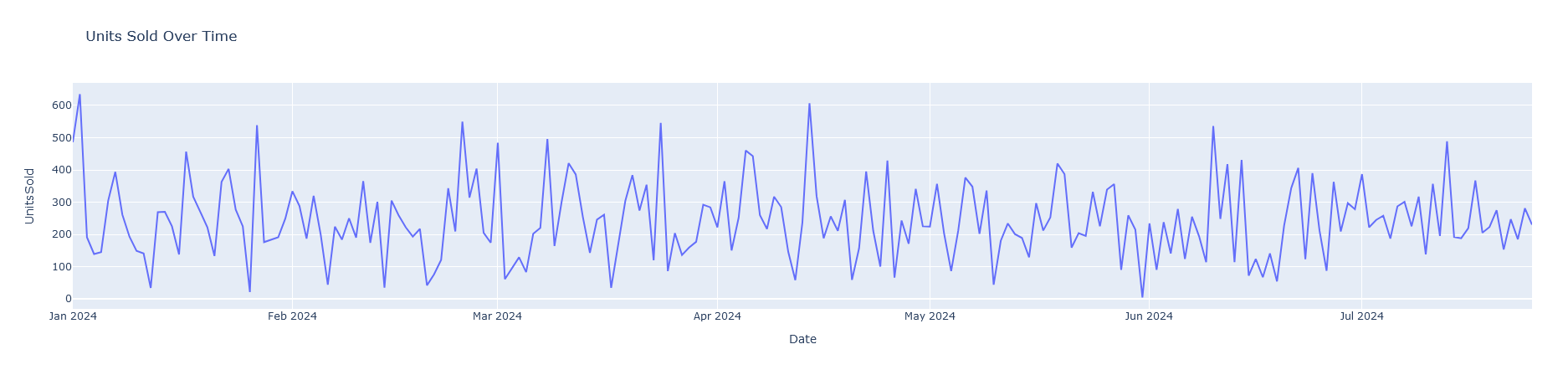
Line Chart for Total Sales Over Time



Insight

* This chart helps visualize trends in mobile phone sales over different periods (e.g., months or years).
* You can observe whether sales are increasing, decreasing, or remaining stable over time.
* Identifies seasonal patterns or anomalies, such as spikes during holiday seasons or dips during economic downturns.

Business Decision



* Use this information to make strategic decisions about inventory management, marketing campaigns, and sales promotions.
* Plan for peak sales periods by adjusting stock levels and scheduling marketing efforts accordingly.
* Identify periods of low sales to strategize targeted promotions or discounts.

2. Bar Chart for Sales by Product Category

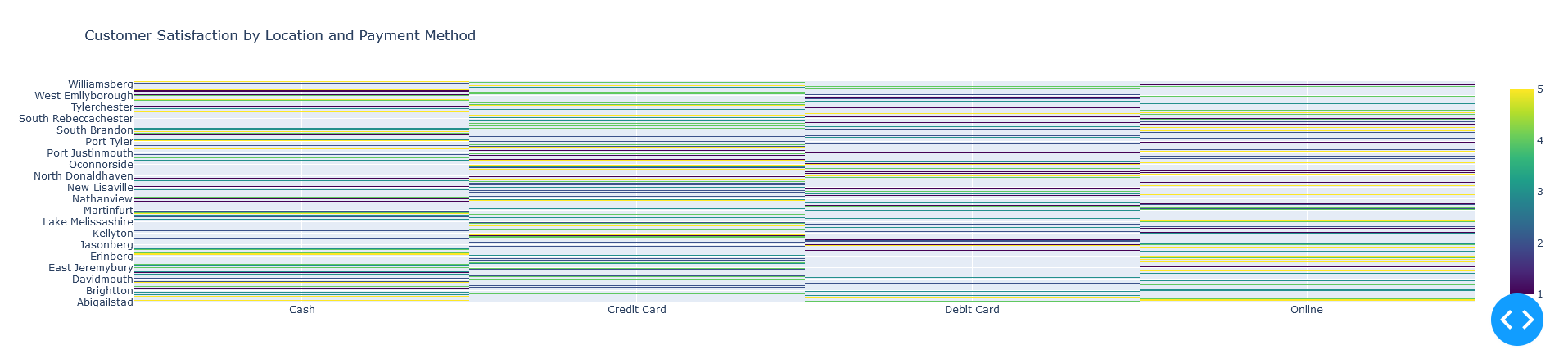
Insight:

* This chart displays the number of units sold or total revenue generated by different mobile phone categories (e.g., smartphones, feature phones, etc.).
* Helps to compare the performance of various product categories over time.

Business Decision:

* Allocate marketing and inventory resources to high-performing categories.
* Evaluate underperforming categories and decide whether to improve, discontinue, or adjust pricing strategies.
* Tailor promotional activities based on category performance to boost sales.

1. Heatmap for Correlation Between Variables



Insight

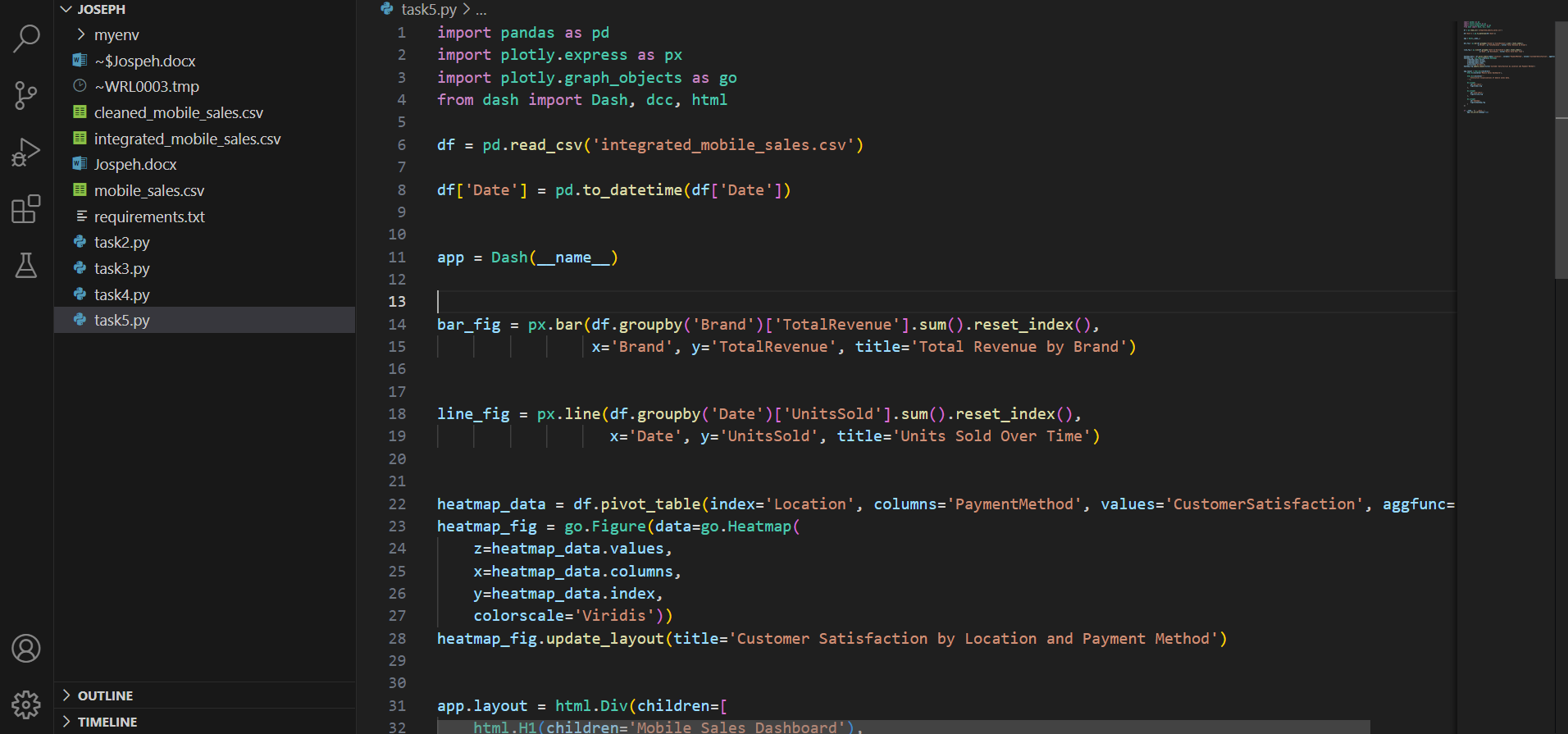
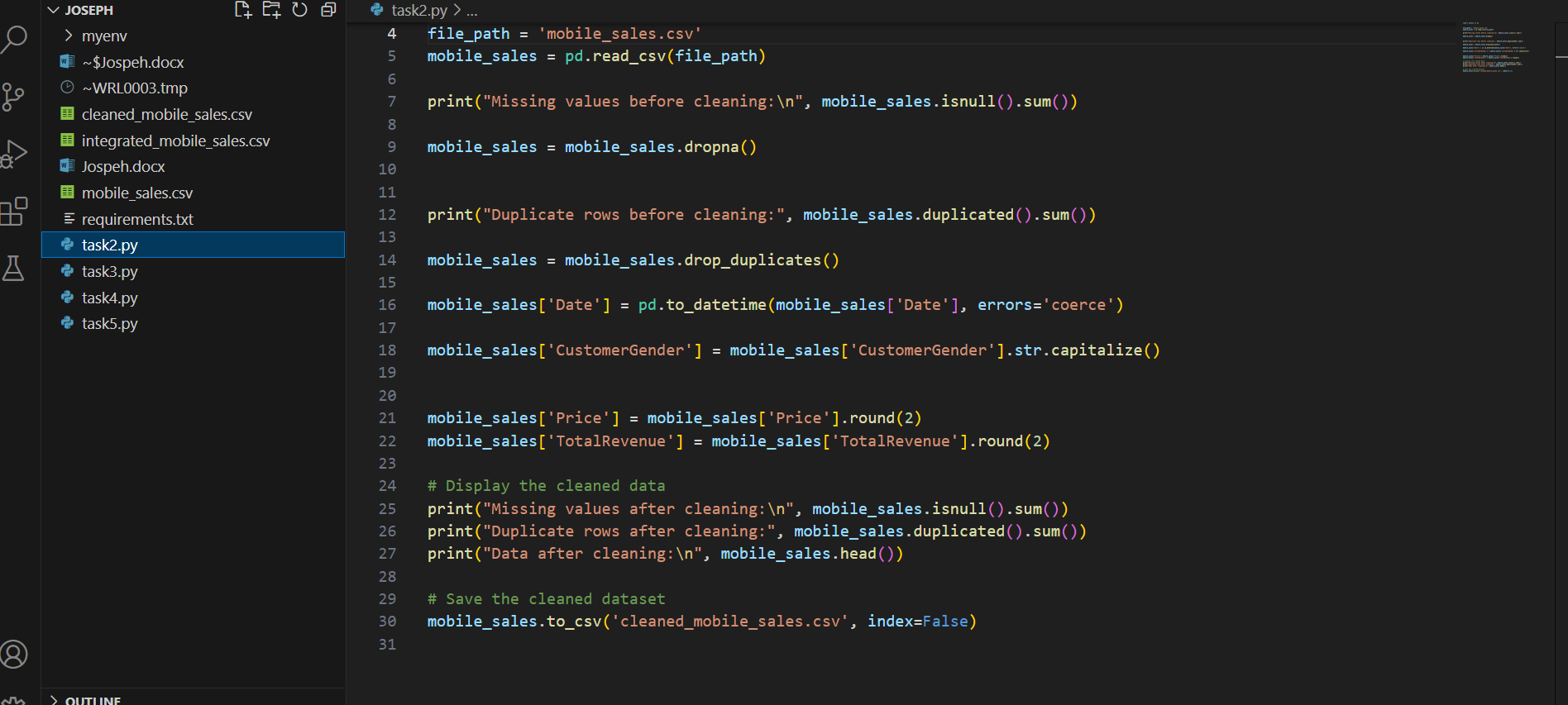
* This heatmap shows the correlation between different variables in the dataset, such as sales volume, pricing, marketing spend, and customer demographics.
* Helps identify relationships between variables, such as whether higher marketing spend correlates with increased sales.

Business Decision:

* Optimize marketing budgets by understanding which variables most strongly impact sales.
* Adjust pricing strategies based on their impact on sales.
* Enhance product offerings or features that show a positive correlation with higher sales.

**Appendices**

**Code snippets**

****