Analysis of the Campaign Marketing: Data Cleaning, Transformation, and Visualization

Maha Hanif
Oumaima Mouimi

March 7, 2024

Contents

1 Introduction

The primary objective of this project is to thoroughly analyze the marketing campaign data, aiming to evaluate the campaign’s effectiveness and customer segmentation. The data set encompasses various columns covering campaign details, consumer demographics, sales figures, engagement metrics, and more.

2 Data Cleaning and Transformation

Data cleaning and transformation play a crucial role in ensuring the reliability and accuracy of our analysis. By cleaning the data, we remove inconsistencies, errors, and missing values that could otherwise introduce bias and lead to incorrect conclusions. Additionally, transforming the data into a more suitable format allows us to extract meaningful insights and patterns more effectively.

Through meticulous data cleaning, we ensure that our dataset is free from outliers, typographical errors, and inconsistencies. This not only enhances the quality of our analysis but also improves the robustness of our findings. Moreover, transforming the data into a standardized format enables us to perform more accurate comparisons and calculations across different variables and time periods.

2.1 Date Formatting

The date columns were converted to actual date formats using the following code:

```
for col in data.columns:
    if data[col].dtype == ’object’:
        try:
            data[col] = pd.to_datetime(data[col])
        except ValueError:
            pass
```

3 Handling Missing Data with the MICE Algorithm

The full name of MICE is Multivariate Imputation by Chained Equations. It is a statistical method used to handle missing data in a dataset.

3.1 Why is MICE Important?

Here are some reasons why MICE is considered important:

- **Preserves relationships:** MICE preserves the relationships between variables in the original data, which is important for producing accurate results in machine learning models.

- **Reduces bias:** By imputing missing values multiple times and combining the results, MICE reduces the amount of bias that is introduced into the data.

- **Flexibility:** MICE is a flexible technique that can handle different types of missing data, including both missing at-random and missing not-at-random data.

- **Handles large amounts of missing data:** MICE is particularly useful for datasets with large amounts of missing data, where other imputation techniques may not be appropriate.

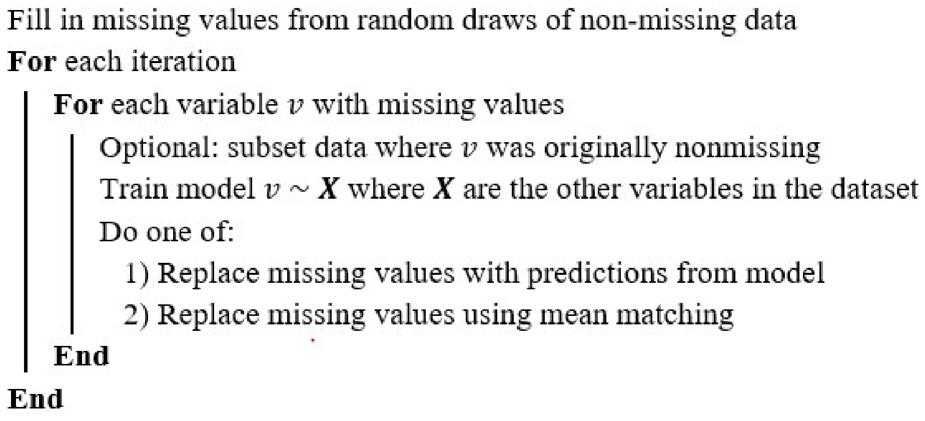
- **Comprehensive:** MICE provides a comprehensive way to handle missing data by taking into account the uncertainty associated with imputing missing values.

3.2 How Does the MICE Algorithm Work?

To understand the functionality of the MICE algorithm, here is a quick intuitive explanation (not the exact algorithm):

1. Start with the variable that contains missing values as a response variable Y and use other variables as predictors X.
2. Build a model with the observations where Y is not missing.
3. Predict the missing observations in Y using the model.

This process is repeated multiple times by doing random draws of the data and taking the mean of the predictions to create multiple imputations.



```
Fill in missing values from random draws of non-missing data
For each iteration
    For each variable $v$ with missing values
        Optional: subset data where $v$ was originally nonmissing
        Train model $v \sim \boldsymbol{X}$ where $\boldsymbol{X}$ are the other variables in the dataset
        Do one of:
          1) Replace missing values with predictions from model
          2) Replace missing values using mean matching
    End
End
```

Figure 1: Pseudo-code for the MICE algorithm.

The pseudo-code provided above offers a simplified view of the complex mechanism behind MICE imputation.

3.3 Why We Chose the MICE Algorithm

Among various imputation methods, we selected the MICE algorithm due to its robustness and ability to maintain the inherent relationships within the data. Unlike simple imputation methods, MICE uses the information from the entire dataset, preserving the statistical properties and minimizing biases that could affect subsequent analysis. This is particularly important for our dataset, where the pattern of missingness might be complex and not completely at random.

4 Outlier Detection & Typographical Errors

Outliers were detected using the Interquartile Range (IQR) method:

```
def detect_outliers(df):
    outlier_indices = []

    for column in df.columns:
        if df[column].dtype in ['int64', 'float64']:
            Q1 = df[column].quantile(0.25)
            Q3 = df[column].quantile(0.75)
            IQR = Q3 - Q1
            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR
            outliers = df[(df[column] < lower_bound) |
            (df[column] > upper_bound)].index
            outlier_indices.extend(outliers)

    outlier_indices = list(set(outlier_indices))
    return df.iloc[outlier_indices]

# Detecting outliers
outliers_df = detect_outliers(data)
print("Outliers in the dataset:")
print(outliers_df)
```

After seeing the unique values in each column, we noticed that data does not include typos.

5 Aggregated Data Insights

The aggregation of the marketing campaign data using various grouping criteria has led to valuable insights. By grouping the data by campaign type and region, we computed aggregate sales figures, units sold, client engagement metrics, and satisfaction rates. Our key findings are as follows:

- The *Direct Advertising* campaign in the *North* region achieved the highest total sales with a considerable average, although the *Online Advertising* strategy in the *East* excelled in terms of average sales.
- When it comes to units sold, *Online Advertising* in the *North* demonstrated superior performance in total units sold.
- Engagement with clients was significantly higher for *Online Advertising* campaigns, especially in the *South* and *North* regions.
- The *East* region maintained a higher satisfaction rate across different campaign types, indicating a positive reception of marketing efforts.
- Interestingly, the *Social Networks* campaign in the *West* saw the highest median sales, suggesting targeted strategies in this area may be particularly effective.

Key performance indicators were calculated for each campaign type within each region, leading to the following observations:

- **Sales Performance:** The 'Direct Advertising' campaign exhibited notable average sales in the East region, whereas 'Online Advertising' achieved the highest mean in the South region.
- **Units Sold:** The North region saw the highest number of units sold for 'Email' campaigns, indicating a successful penetration in this demographic.
- **Client Engagement:** 'Online Advertising' in the South had significantly higher client engagement, showing the effectiveness of digital campaigns in this area.
- **Conversion and Satisfaction:** Across campaign types, the 'Social Networks' campaign in the West reported an impressive median for both conversion rate and client satisfaction percentage.

These insights can guide future marketing strategies by identifying the most effective campaigns and the regions where the customers are most engaged and satisfied.

6 Data visualization

In this section, we present various visualizations to explore different aspects of the marketing campaign data.

6.1 Line Plot: Sales Over Years

The following line plot illustrates the trend of sales over the years:

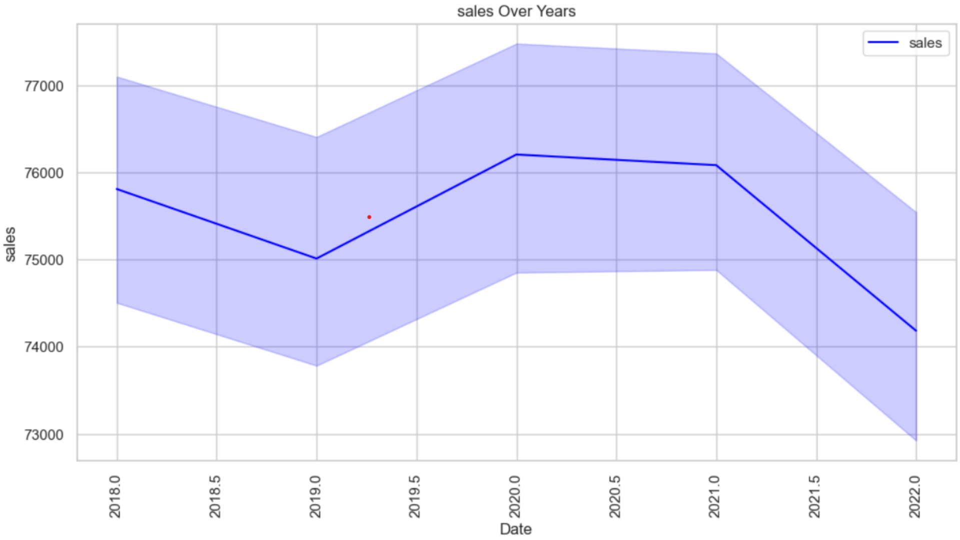


Figure 2: Sales Over Years

6.2 Line Plot: Sales Over Time

The following line plot illustrates the trend of sales over time:

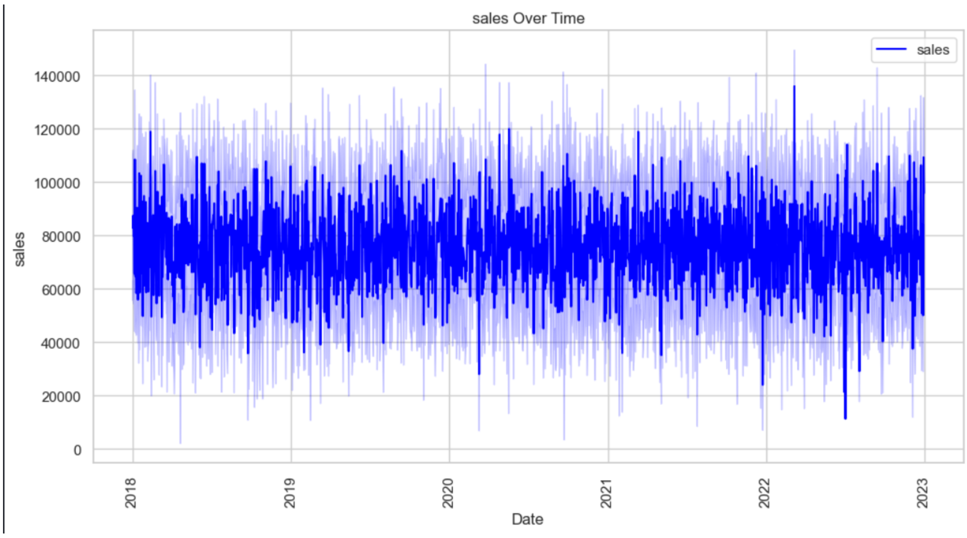


Figure 3: Sales Over Time

This line plot shows the variation in sales over the entire time period covered by the dataset. The x-axis represents the date, and the y-axis represents the sales amount. From the plot, we can observe the overall trend and any seasonal patterns or anomalies in the sales data.

6.3 Line Plot: Clients Engagement Over Years

The following line plot illustrates the trend of client engagement over the years:

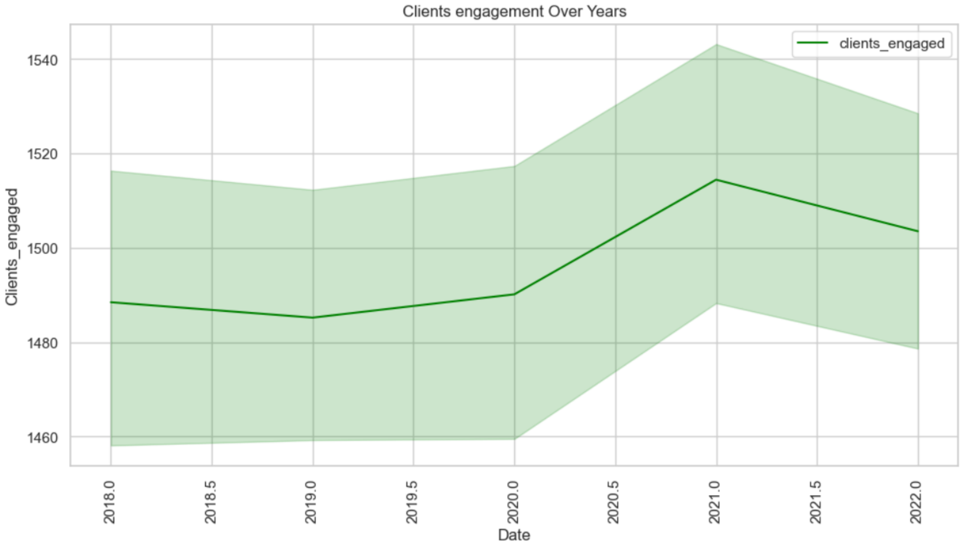


Figure 4: Clients Engagement Over Years

This line plot shows the variation in client engagement over the years. The x-axis represents the year, and the y-axis represents the number of clients engaged. From the plot, we can observe any trends or patterns in client engagement over time.

6.4 Line Plot: Clients Engagement Over Time

The following line plot illustrates the trend of client engagement over time:

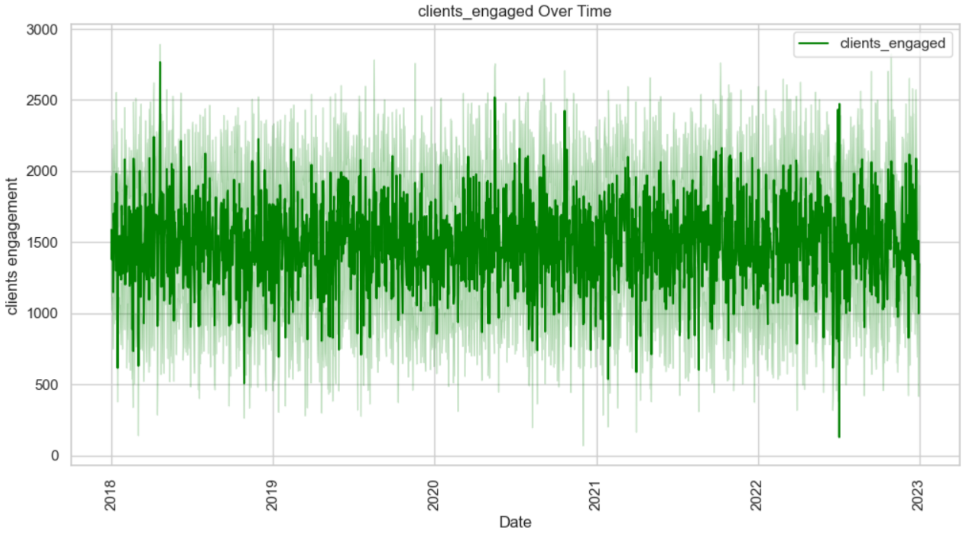


Figure 5: Clients Engagement Over Time

This line plot shows the variation in client engagement over the entire time period covered by the dataset. The x-axis represents the date, and the y-axis represents the number of clients engaged. From the plot, we can observe any trends or patterns in client engagement over time.

6.5 Histogram Plot: Distribution of Sales by Campaign Type

The following histogram plot illustrates the distribution of sales by campaign type:

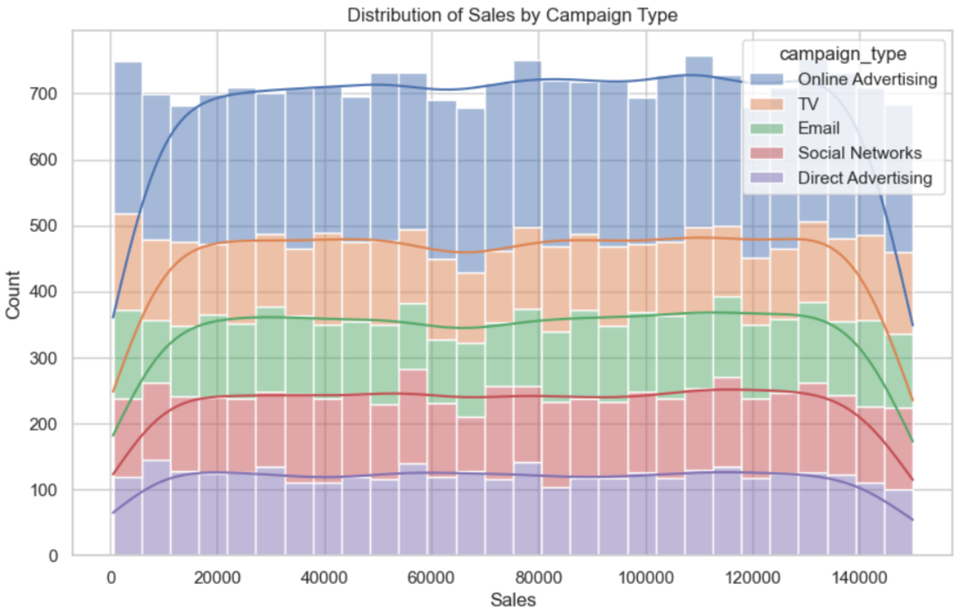


Figure 6: Distribution of Sales by Campaign Type

This histogram plot shows the distribution of sales across different campaign types. Each campaign type is represented by a different color in the plot. The x-axis represents the sales amount, and the y-axis represents the count of observations. From the plot, we can observe the distribution of sales for each campaign type and any differences or similarities between them.

6.6 Box Plot: Distribution of Sales by Region

The following box plot illustrates the distribution of sales by region:

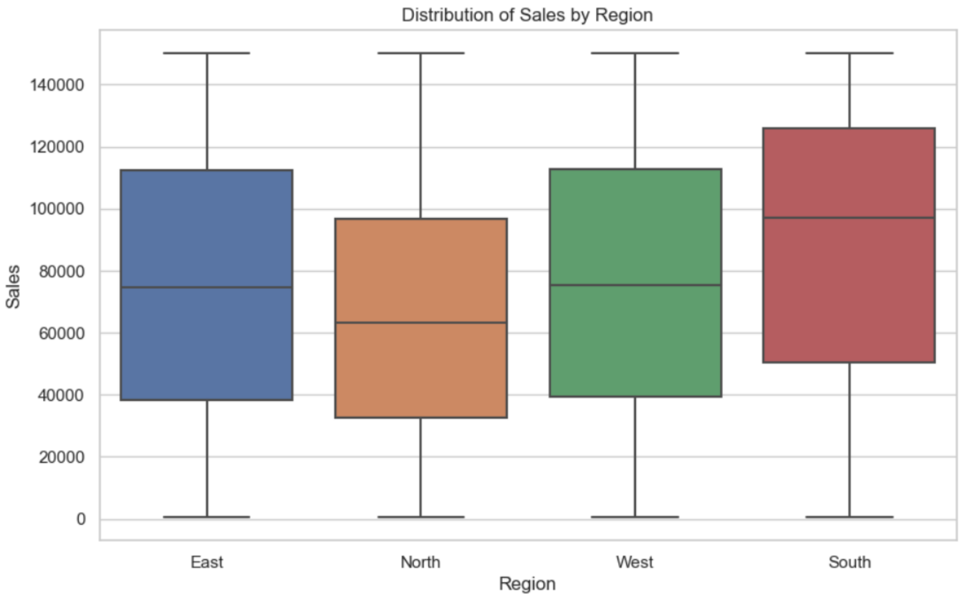


Figure 7: Distribution of Sales by Region

This box plot shows the distribution of sales across different regions. Each box represents the interquartile range (IQR) of sales for a specific region, with the median value indicated by the line inside the box. The whiskers extend to show the range of the data, excluding outliers. From the plot, we can observe any variations in sales distribution among different regions and identify potential outliers.

6.7 Bar Plot: Average Conversion Rate by Customer Age Range

The following bar plot illustrates the average conversion rate by customer age range:

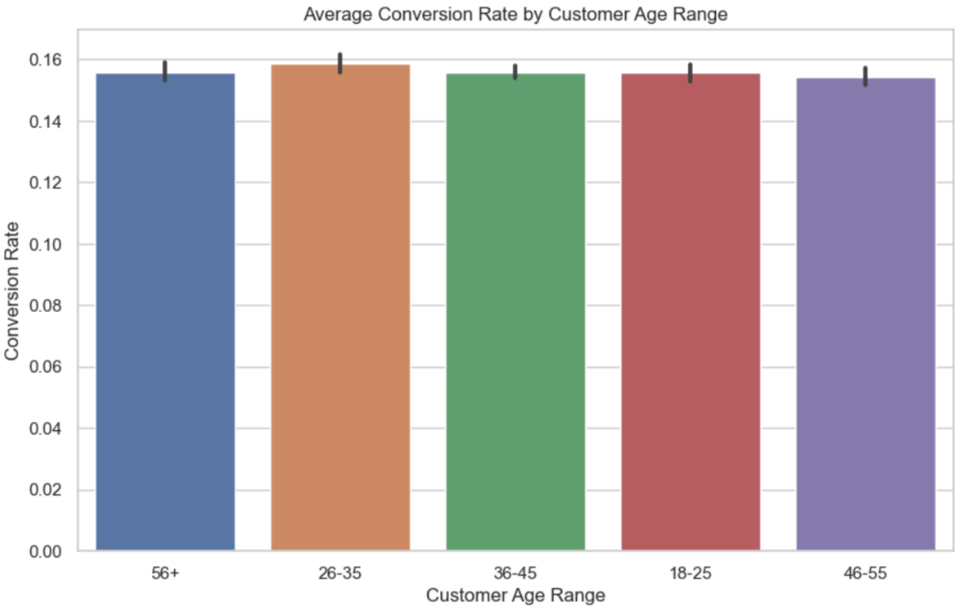


Figure 8: Average Conversion Rate by Customer Age Range

This bar plot shows the average conversion rate for different customer age ranges. Each bar represents the average conversion rate for a specific age range. From the plot, we can observe any variations in conversion rates across different age groups and identify potential trends or patterns.

6.8 Scatter Plot: Click Rate vs Conversion Rate

The following scatter plot illustrates the relationship between click rate and conversion rate:

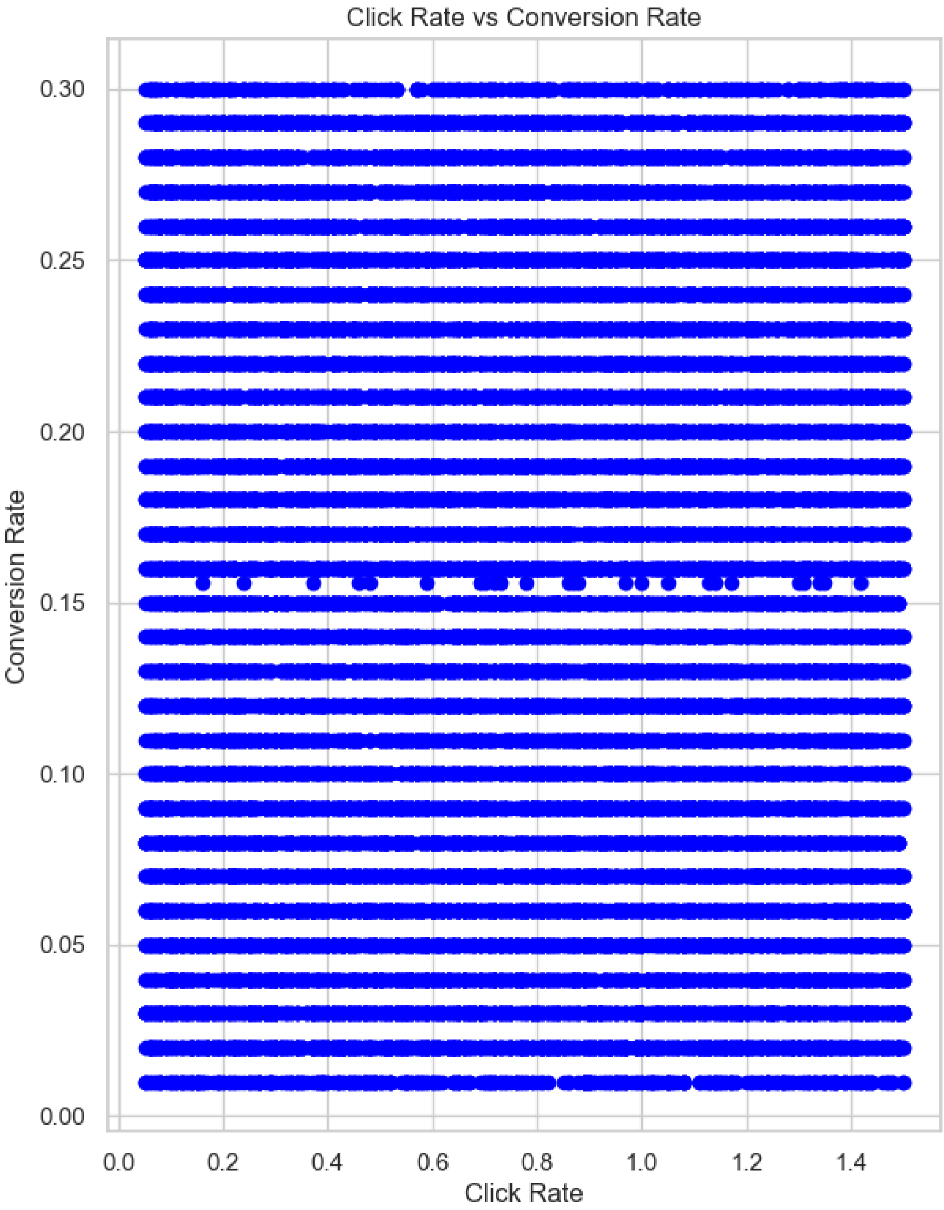


Figure 9: Click Rate vs Conversion Rate

This scatter plot shows the relationship between click rate and conversion rate. Each point represents a data point, where the x-coordinate represents the click rate and the y-coordinate represents the conversion rate. From the plot, we can observe any patterns or correlations between these two variables.

6.9 Scatter Matrix Plot

The following scatter matrix plot illustrates the relationships between different variables:

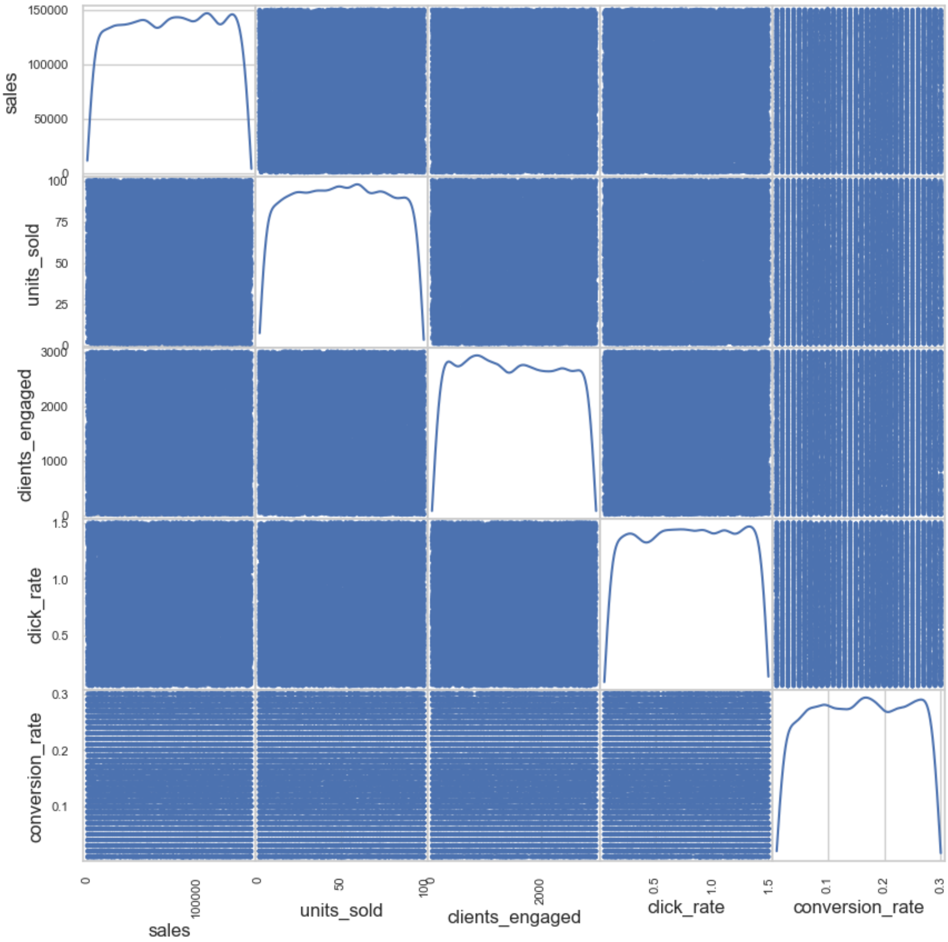


Figure 10: Scatter Matrix Plot

This scatter matrix plot provides a visual overview of the relationships between different variables in the dataset. Each scatter plot in the matrix represents the relationship between two variables, and the diagonal plots show the distributions of individual variables.

6.10 Violin Plot: Gender vs. Sales

The following violin plot illustrates the relationship between gender and sales:

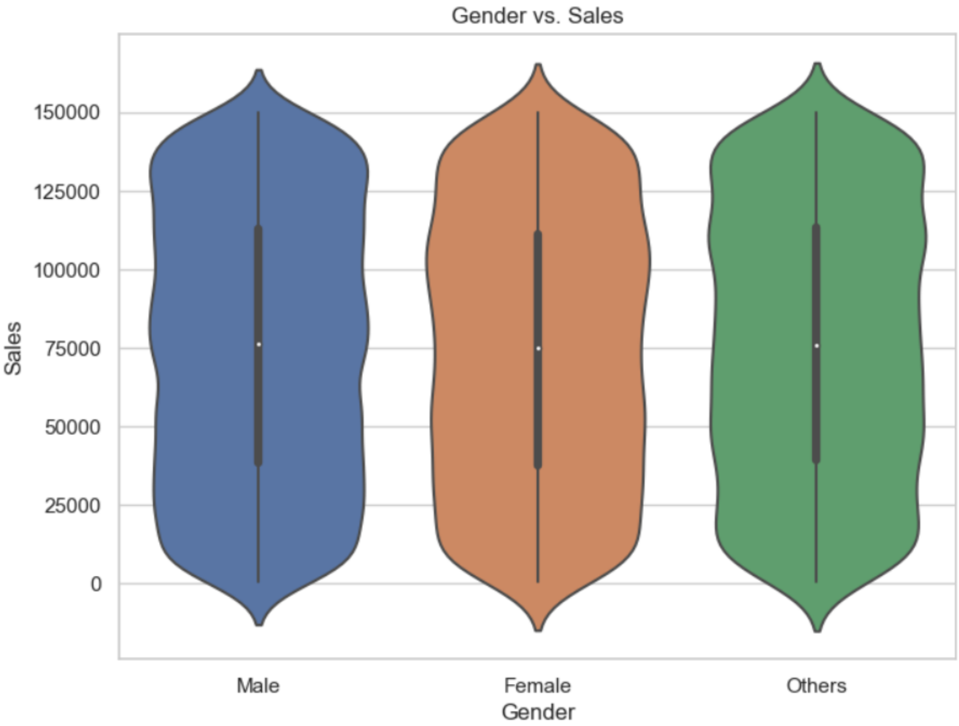


Figure 11: Violin Plot: Gender vs. Sales

This violin plot shows the distribution of sales across different genders. Each violin plot represents the distribution of sales for a specific gender. From the plot, we can observe any variations in sales distribution between different genders.

6.11 Correlation Heatmap

The following heatmap illustrates the correlation between different numerical variables:

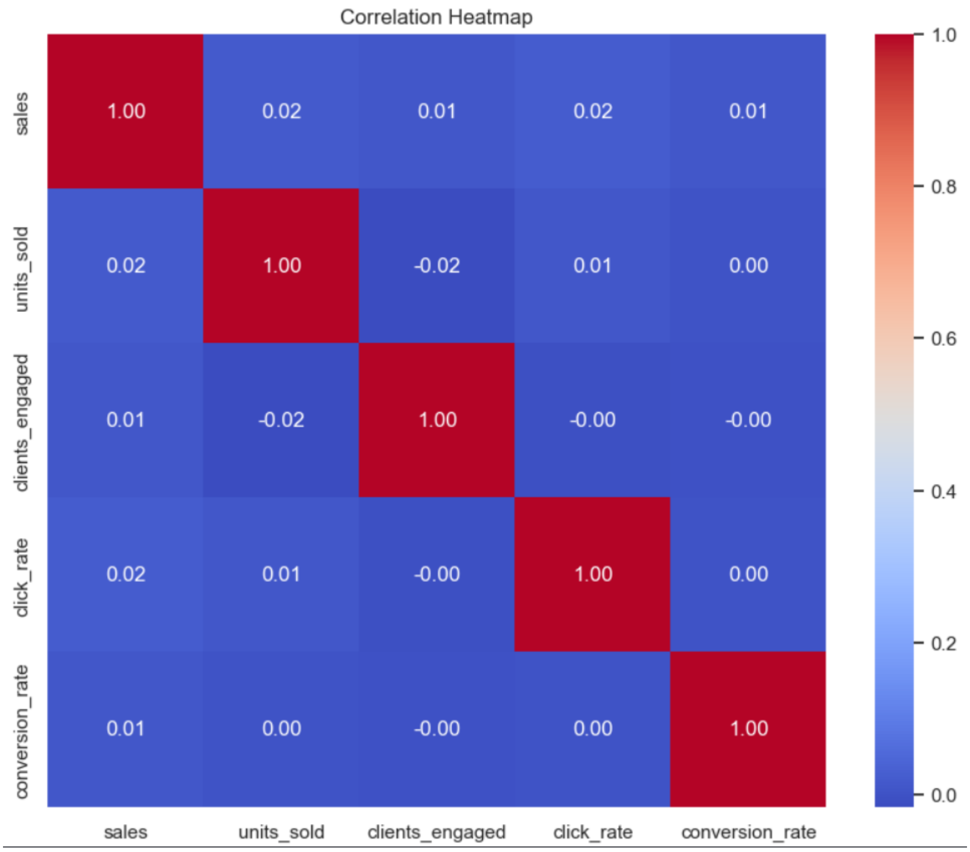


Figure 12: Correlation Heatmap

This heatmap shows the pairwise correlation coefficients between sales, units sold, clients engaged, click rate, and conversion rate. The values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. Positive values indicate a positive relationship between variables, while negative values indicate a negative relationship. Higher absolute values indicate stronger correlations.

7 Conclusion

In this project, we performed data cleaning, transformation, and visualization to analyze the effectiveness of marketing campaigns. One notable aspect of our analysis is that we utilized the entire dataset by concatenating both the training and test datasets. By leveraging the complete dataset, we were able to gain comprehensive insights into the performance of various campaign types, regional trends, customer demographics, and engagement metrics.

The insights derived from this analysis can serve as valuable inputs for enhancing future campaign strategies. By understanding the patterns and relationships within the data, marketers can make informed decisions to optimize campaign targeting, messaging, and allocation of resources. Additionally, the utilization of the entire dataset ensures that the findings are robust and representative of the overall campaign performance.

Overall, our comprehensive approach to data analysis, incorporating both data cleaning and visualization techniques, provides a solid foundation for refining marketing strategies and achieving better campaign outcomes in the future.

List of Figures