BRSM Report 2

Descriptive Statistics, Visualization and Inferential Statistics

25 April 2025

Contents

Name Vinit Mehta
Roll Number 2022111001
Mini Project 5 (Social Media Advertising Dataset)

1 Introduction

Name of Dataset: Social Media Advertising Dataset
Link to Dataset: kaggle
About Dataset:

- This dataset contains detailed information about various social media advertising campaigns across platforms such as Facebook, Instagram, Pinterest, and Twitter.
- The dataset includes 16 features like *Impressions, Clicks, Spend, Conversion Rate, ROI, Engagement Score, Target Audience, Location, Channel Used, Campaign Goal*, and others to evaluate campaign effectiveness.
- These features represent user engagement metrics, demographic targeting, platform choices, and financial aspects of advertising campaigns.
- Total number of datapoints = 3,00,000 and there are no missing values in any column.

Background: Advertising has played a pivotal role in commerce since ancient times, evolving into a dominant force in capitalist economies by the mid-19th century, largely driven by newspapers and magazines [1]. The 20th century witnessed a rapid expansion of advertising across emerging media, including direct mail, radio, television, the internet, and mobile platforms. In the United States, advertising expenditure consistently averaged 2.2% of the Gross Domestic Product (GDP) between 1919 and 2007 [1].

Advertising has evolved over time, beginning with the first newspaper ad in the Boston News-Letter in 1704 [2], followed by the first radio ad in 1922 [3] and the first paid TV ad in 1941 [4]. The 1990s marked the rise of digital marketing with search engines and targeted customer databases [5]. The advent of the internet and social media led to platforms like Facebook hosting the first social media ad by JPMC in 2006 [6], fueling the growth of social media advertising ever since.

This dataset, consisting of 300,000 entries, provides a comprehensive view of social media advertising campaigns conducted across platforms such as Facebook, Instagram, Pinterest, and Twitter. It captures a wide range of campaign attributes, including impressions, clicks, spend, engagement scores, conversion rates, as well as contextual information like campaign duration, language, location, and target audience characteristics. These features enable a multifaceted exploration of how digital ads perform across different demographic and platform-specific segments. However, despite the richness of the data, advertisers often face challenges in interpreting how specific campaign elements influence key performance indicators like ROI, engagement, and conversion outcomes. While most existing studies focus on building predictive models, they frequently neglect the use of inferential statistics, which are essential for identifying the significance and causal impact of individual factors. This study bridges that gap by leveraging descriptive statistics, data visualizations, and inferential techniques to uncover meaningful relationships between campaign metrics and performance, ultimately providing data-driven insights to improve advertising effectiveness and strategic decision-making.

Note

To ensure efficient analysis, a sample of 8,000 data points was drawn from the original 300,000-entry dataset. The sampling strategy aimed to maintain the original distribution of key categorical features using stratified sampling. Rare classes were grouped under an "Other" category, and weighted random sampling was used when stratification wasn't feasible. This approach, ensures the sampled data remains representative of the full dataset. Distribution plots before and after sampling were also used to verify consistency.

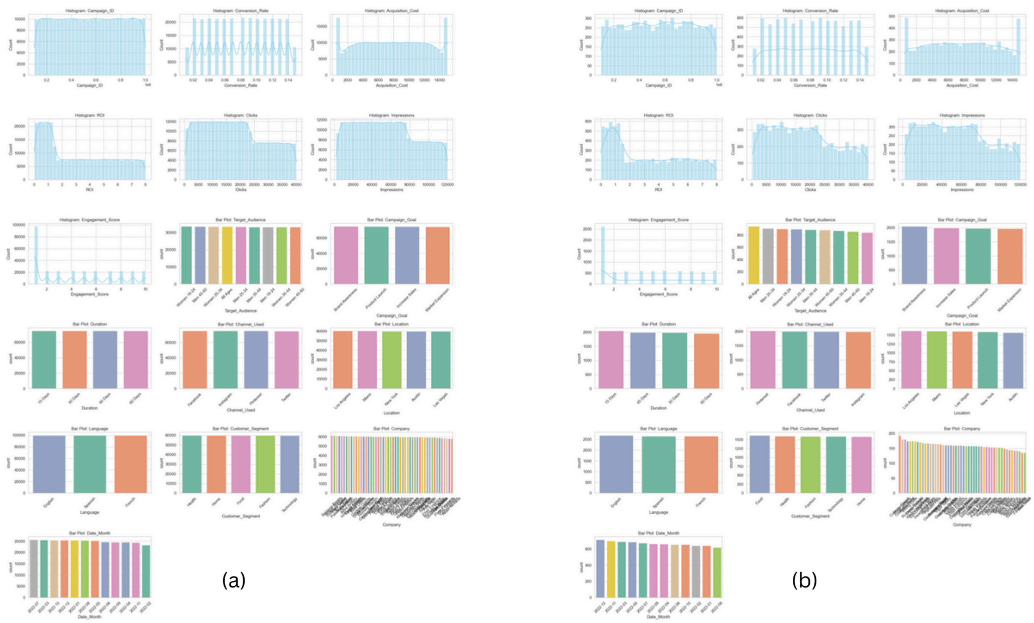

Figure 1: (a) Distribution of original data; (b) Distribution of sampled data.

2 Descriptive Statistics

Numerical Variables

Metric	Campaign_ID	Conversion Rate	Acquisition Cost	ROI	Clicks	Impressions	Engagement Score
Count	8000.0	8000.0	8000.0	8000.0	8000.0	8000.0	8000.0
Mean	550234.14	0.0798	7679.27	3.2091	17984.41	55520.12	4.3714
Std Dev	257829.13	0.0406	4313.13	2.4598	11054.68	32671.58	3.1565
Min	100066.0	0.01	500.0	0.0	293.0	1937.0	1.0
25%	331569.5	0.04	4040.44	0.9581	8542.5	27728.5	1.0
50%	556482.0	0.08	7659.95	2.74	17042.5	53552.0	4.0
75%	770410.0	0.11	11281.15	5.38	26605.5	80325.5	7.0
Max	999726.0	0.15	15000.0	8.0	39989.0	119968.0	10.0

Table 1: Statistical Summary of Sampled Social Media Advertising Dataset

Categorical Variables

- **Target Audience:** Men 35-44, Women 18-24, Women 25-34, Men 18-24, Men 45-60, All Ages, Women 45-60, Men 25-34, Women 35-44.
- **Campaign Goal:** Product Launch, Market Expansion, Brand Awareness, Increase Sales.

- **Duration:** 30 Days, 60 Days, 45 Days, 15 Days.
- **Channel Used:** Facebook, Instagram, Twitter, Pinterest.
- **Location:** New York, Las Vegas, Austin, Los Angeles, Miami.
- **Language:** Spanish, French, English.
- **Customer Segment:** Technology, Health, Food, Home, Fashion.
- **Date:** 2022-01-01 to 2022-12-31.
- **Company:** Ex: Innovate Infinity, Fitness Front, Gourmet Grove, NexGen Nerds, Dine Divine (total of 50 companies).

3 Visualizing the dataset

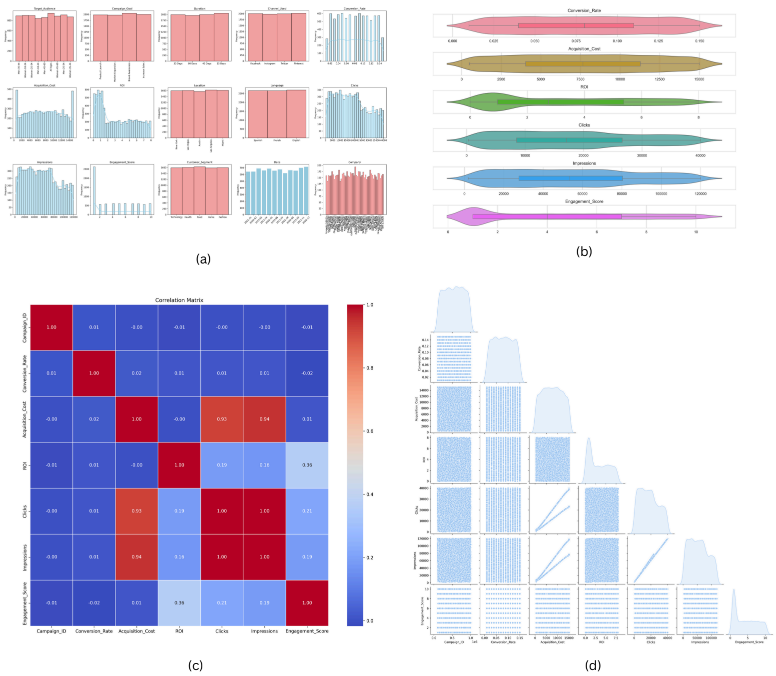

Figure 2: **(a) Histogram:** Shows feature-wise value distribution; **(b) Box-Violin Plot:** Depicts data spread, central tendency, and outliers; **(c) Correlation Matrix:** Highlights linear relationships among features; **(d) Pair Plot:** Explores pairwise feature interactions and distributions.

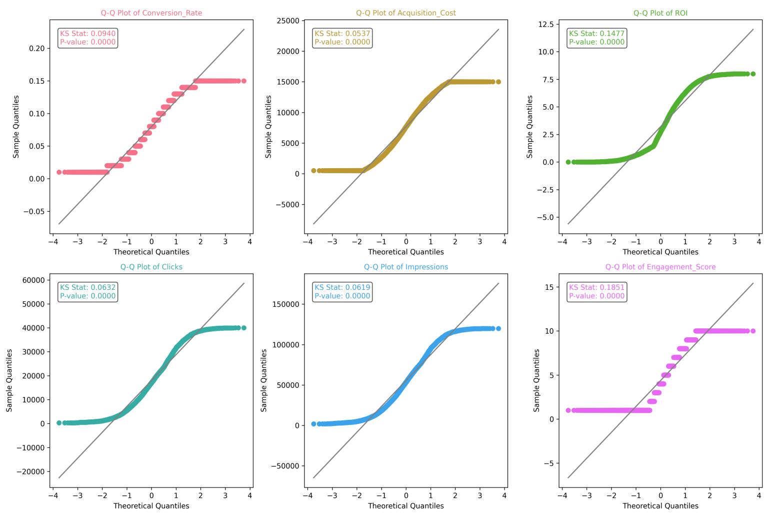

Figure 3: Checking for normality of various numerical features. From the Q-Q plot and KS test we can say that none of them are normally distributed.

4 Statistical Methodology Overview

1. **Hypothesis Formulation:** Clearly define the statistical hypotheses before conducting any analysis. The *null hypothesis* (H_0) typically assumes no difference or no effect between groups, while the *alternative hypothesis* (H_1) proposes a significant difference or effect.
2. **Data Visualization:** Utilize various graphical tools such as histograms, boxplots, and scatterplots to visually assess the distribution of the data and explore potential relationships between variables.
3. **Assumption Checks:** Prior to selecting appropriate statistical tests, verify key assumptions like *normality: Q-Q Plot + Kolmogorov-Smirnov test - as large number of samples* and *homogeneity of variances: Levene's test* which are prerequisites for parametric methods such as ANOVA. Q-Q plots are employed to visually evaluate the assumption of normality.
4. **Choice of Statistical Test:** Depending on the results of assumption checks:
   - Use **Analysis of Variance (ANOVA)** when assumptions are met (parametric).
   - Use the **Kruskal–Wallis H Test** when assumptions are violated (non-parametric).
5. **Post Hoc Analysis:** If the ANOVA or Kruskal–Wallis test reveals significant differences among groups, conduct pairwise comparisons to identify where the differences lie:
   - Use **Independent t-tests** (parametric) or **Mann–Whitney U tests** (non-parametric) for pairwise group comparisons.
   - Apply appropriate multiple testing correction methods such as the **Bonferroni correction** or the **Benjamini–Hochberg procedure** to adjust p-values and control for Type I error.
6. **Additional Analyses:** Employ other relevant techniques where applicable:
   - Use **Regression Analysis**, including techniques like *stepwise regression* and checking for multicollinearity via the **Variance Inflation Factor (VIF)**, for predictive modeling.
   - Use **Permutation Tests** as a non-parametric alternative when distributional assumptions are not satisfied.
7. **Reporting Results:** Conclude by summarizing all key findings. Report the corresponding **p-values** along with the **effect sizes** to assess not only statistical significance but also the practical importance of the results.

Note

Unless stated otherwise assume the level of significance, $\alpha = 0.05$.

Q1 Social Media Platform vs Engagement Levels

Hypothesis Formulation

- **Null Hypothesis (H_0):** There is no significant difference in engagement levels across different social media platforms. That is, the choice of platform does not impact engagement.
- **Alternative Hypothesis (H_a):** There is a significant difference in engagement levels across different social media platforms. That is, at least one platform has a different average engagement level compared to others.

Data Visualization

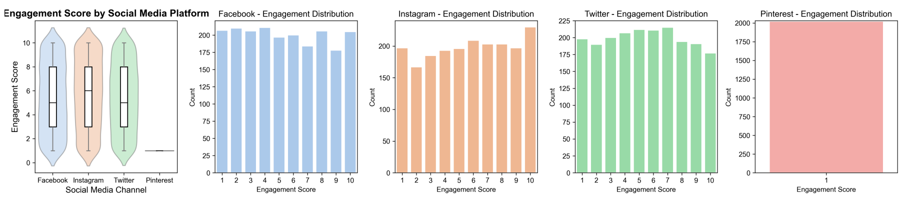

Figure 4: The first figure presents box plots overlaid within violin plots, illustrating the distribution of engagement scores for each social media platform. Following this, a series of histograms display the frequency distribution of different engagement levels.

We can clearly see that the data for pinterest drastically differs from the rest. So moving forward we will consider 2 scenarios for testing - one considering pinterest and one without pinterest to make sure pinterest is not undermining any other statistically significant difference that exists.

Assumptions Check

We assume independence of all 3 social media platforms as each campaign was done only on a single social media platform, and we test for normality of residuals **(Q-Q plot + KS test)** of engagement scores and homogeneity of variances across groups **(Levene's test)**.
Q-Q plot + KS test results:

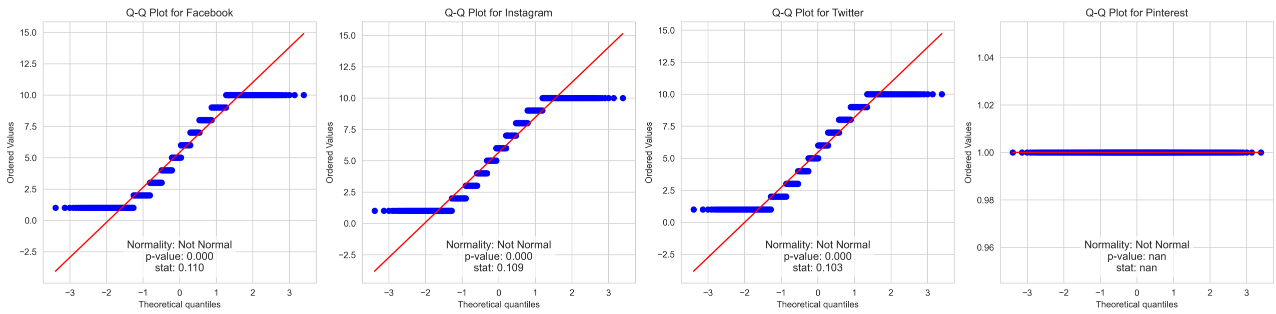

Figure 5: Q-Q plots with KS test results for engagement scores for each of the channels.

Observations

None of the group follow Normal distribution. Normality assumption not met.

Levene's Test Results:

- **Test 1: All Channels Included**
  Test Statistic: 1822.1527
  p-value: 0.0000
- **Test 2: Pinterest Removed (due to constant values)**
  Test Statistic: 1.1954
  p-value: 0.3026

Observations

Since p-value $< \alpha$ variances are not homogeneous **with Pinterest**. Since p-value $> \alpha$ variances are homogeneous **without considering Pinterest**.

Statistical Test

Since the assumption of normality was violated for all groups, we proceeded with the non-parametric **Kruskal-Wallis H test**, which does not require the normality assumption, regardless of whether the assumption of homogeneity of variances is satisfied.

- **With Pinterest Included:**
  Kruskal-Wallis test statistic yielded a **p-value = 0.000**, indicating a statistically significant difference in engagement levels across at least one of the social media platforms.
- **With Pinterest Excluded:**
  Kruskal-Wallis test statistic yielded a **p-value = 0.0154**, which also indicates a statistically significant difference in engagement levels across the remaining platforms.

Post Hoc Test & Effect size

Since the significance test in the previous subsection revealed a statistically significant difference, we proceed with post hoc analyses to identify the specific groups that differ from one another. Also we use rank-biserial correlation to measure the effect size of significance.

$$r_{rb} = 1 - \frac{2U}{n_1 n_2}$$

Full Set (Including Pinterest):

Channel 1	**Channel 2**	**p-value**	**Significance**	**Effect Size (r_{rb})**
Facebook	Instagram	0.0082	Significant	0.0481
Facebook	Twitter	0.6994	Not Significant	0.0070
Facebook	Pinterest	0.0000	Significant	-0.8967
Instagram	Twitter	0.0200	Significant	-0.0424
Instagram	Pinterest	0.0000	Significant	-0.9005
Twitter	Pinterest	0.0000	Significant	-0.9008

Table 2: Pairwise Mann-Whitney U Test Results Across All Channels (with Effect Size)

Subset (Excluding Pinterest):

Channel 1	**Channel 2**	**p-value**	**Significance**	**Effect Size (r_{rb})**
Facebook	Instagram	0.0082	Significant	0.0481
Facebook	Twitter	0.6994	Not Significant	0.0070
Instagram	Twitter	0.0200	Significant	-0.0424

Table 3: Pairwise Mann-Whitney U Test Results (Excluding Pinterest) with Effect Size

Observations

Engagement scores on Pinterest (practically important difference - high effect size) and Instagram (practically unimportant difference - low effect size) significantly differs from the rest. Facebook and Twitter has no significant difference in their engagement scores.

Results

Based on the non-parametric statistical analysis (Kruskal-Wallis and Mann-Whitney U tests), we draw the following conclusions:

- We **reject the null hypothesis** H_0 for the choice of social media platform, as it **significantly impacts** engagement levels.
- **Pinterest**, with its constant engagement scores, shows statistically distinct behavior, causing a violation of variance homogeneity. We **reject the null hypothesis** for Pinterest's engagement scores, suggesting that they do not follow the same distribution as other platforms.
- Even after **removing Pinterest** from the analysis, we still **reject the null hypothesis** between platforms, reinforcing that platform choice has a significant effect on engagement.
- Pairwise comparisons between **Instagram vs. Facebook** and **Instagram vs. Twitter** result in **rejecting the null hypothesis** in both cases, showing consistent significant differences in engagement.
- However, for **Facebook vs. Twitter**, we **fail to reject the null hypothesis**, indicating no significant difference in engagement levels between these two platforms.

Therefore, **Instagram** emerges as the most effective platform for generating engagement (although practically the difference is not much), while **Pinterest** is the least reliable due to its consistently constant engagement scores.

Q2 Duration of Campaign vs Engagement Scores/Conversion Rates

Hypothesis Formulation

- **Null Hypothesis ($H_{0,Engagement}$):** There is no significant difference in engagement scores across different campaign durations.
- **Alternative Hypothesis ($H_{a,Engagement}$):** There is a significant difference in engagement scores across at least one campaign duration compared to others.
- **Null Hypothesis ($H_{0,Conversion}$):** There is no significant difference in conversion rates across different campaign durations.
- **Alternative Hypothesis ($H_{a,Conversion}$):** There is a significant difference in conversion rates across at least one campaign duration compared to others.

Data Visualization

Figure 6: Histograms illustrating the frequency distribution of engagement scores and conversion rates across different campaign durations.

Assumptions Check

We assume independence of the campaigns conducted for different durations. We tested for normality **(Q-Q plot + KS test)** and homogeneity of variances across groups **(Levene's test)**.
Q-Q plot + KS test results:

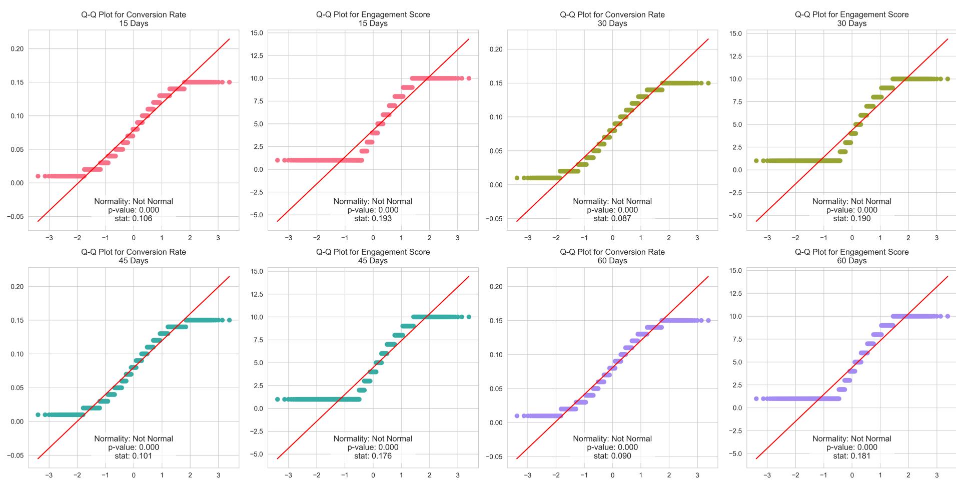

Figure 7: Q-Q plots with KS test results for engagement scores and conversion rates across different campaign durations.

Observations

None of the groups follow a Normal distribution. Normality assumption not met for both Engagement Score and Conversion Rate.

Levene's Test Results:

- **Levene's test for Conversion Rate:**
  Test Statistic: 0.8023
  p-value: 0.4924
- **Levene's test for Engagement Score:**
  Test Statistic: 0.4922
  p-value: 0.6877

Observations

Since p-value $>$ α variances are homogeneous for Conversion Rate. Since p-value $>$ α variances are homogeneous for Engagement Score.

Statistical Test

Since the assumption of normality was violated for all groups, we proceeded with the non-parametric **Kruskal-Wallis H test**, which does not require the normality assumption.

- **Kruskal-Wallis Test: Engagement Score vs Duration**
  p-value: 0.3088
  No statistically significant difference in Engagement Score across durations.
- **Kruskal-Wallis Test: Conversion Rate vs Duration**
  p-value: 0.3011
  No statistically significant difference in Conversion Rate across durations.

Post Hoc Test & Effect size

Since the significance tests in the previous subsection did not reveal any statistically significant differences, no post hoc analyses were performed.

Results

We fail to reject the null hypotheses $H_{0,Engagement}$ and $H_{0,Conversion}$. The duration of a campaign does not have a statistically significant impact on either engagement scores or conversion rates. Based on the available data, campaign duration—whether short or long—does not meaningfully influence the outcomes in terms of engagement or conversion success.

Q3 Location/Language vs Engagement Scores/Conversion Rates

Hypothesis Formulation

- **Null Hypothesis ($H_{0,Engagement\_Location}$):** There is no significant difference in engagement scores across different geographic locations.
- **Alternative Hypothesis ($H_{a,Engagement\_Location}$):** There is a significant difference in engagement scores across at least one geographic location compared to others.
- **Null Hypothesis ($H_{0,Conversion\_Location}$):** There is no significant difference in conversion rates across different geographic locations.
- **Alternative Hypothesis ($H_{a,Conversion\_Location}$):** There is a significant difference in conversion rates across at least one geographic location compared to others.

- **Null Hypothesis ($H_{0,Engagement\_Language}$):** There is no significant difference in engagement scores across different languages.
- **Alternative Hypothesis ($H_{a,Engagement\_Language}$):** There is a significant difference in engagement scores across at least one language compared to others.
- **Null Hypothesis ($H_{0,Conversion\_Language}$):** There is no significant difference in conversion rates across different languages.
- **Alternative Hypothesis ($H_{a,Conversion\_Language}$):** There is a significant difference in conversion rates across at least one language compared to others.

Data Visualization

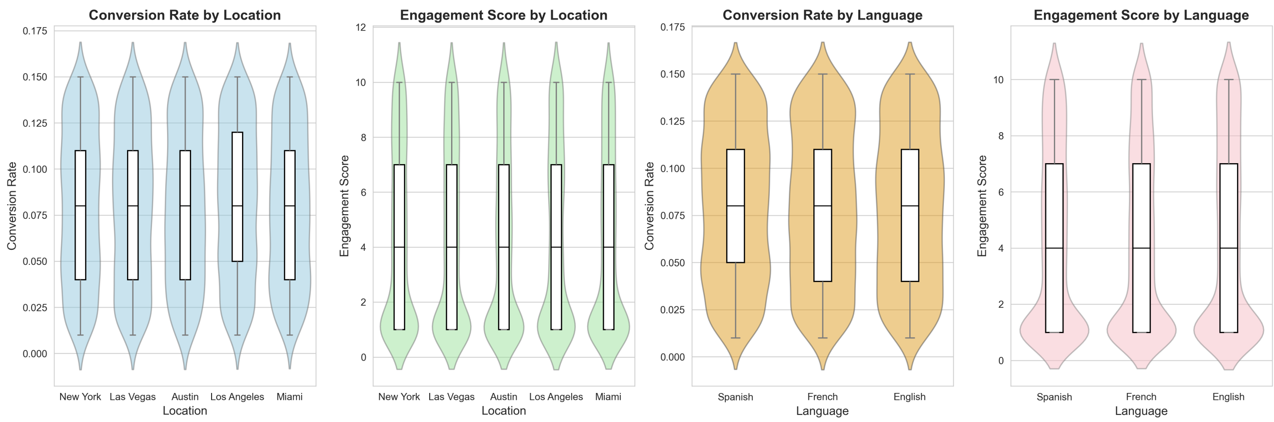

Figure 8: Histograms illustrating the frequency distribution of engagement scores and conversion rates across different geographic locations and languages.

Assumptions Check

We assume independence of the campaigns conducted in different geographic locations and languages. We tested for normality **(Q-Q plot + KS test)** and homogeneity of variances across groups **(Levene's test)**.
Q-Q plot + KS test results:

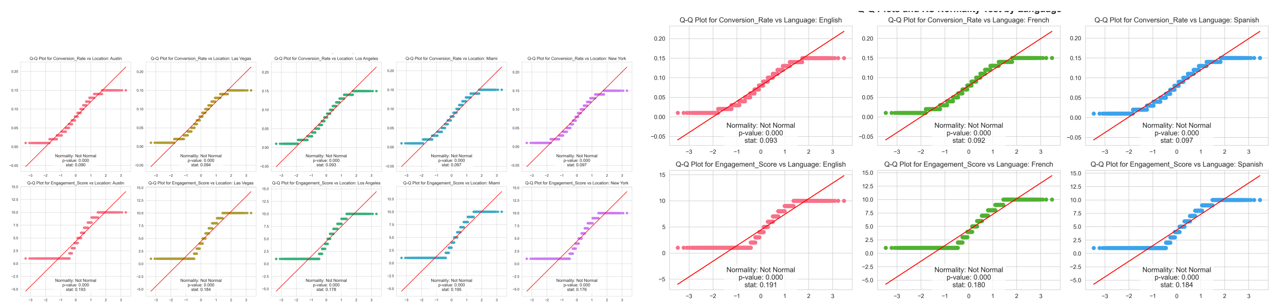

Figure 9: Q-Q plots with KS test results for engagement scores and conversion rates across different geographic locations (left) and languages (right).

Observations

None of the groups were normally distributed. Normality assumption not met for both Engagement Score and Conversion Rate across both Location and Language.

Levene's Test Results:

- **Levene's test for Conversion Rate based on Locations:**
  Test Statistic: 0.6816
  p-value: 0.6046

- **Levene's test for Engagement Score based on Locations:**
  Test Statistic: 0.7834
  p-value: 0.5357
- **Levene's test for Conversion Rate based on Languages:**
  Test Statistic: 0.3173
  p-value: 0.7282
- **Levene's test for Engagement Score based on Languages:**
  Test Statistic: 4.9331
  p-value: 0.0072

Observations

Since p-value $> \alpha$ variances are homogeneous for Conversion Rate based on Locations. Since p-value $> \alpha$ variances are homogeneous for Engagement Score based on Locations. Since p-value $> \alpha$ variances are homogeneous for Conversion Rate based on Languages. Since p-value $< \alpha$ variances are not homogeneous for Engagement Score based on Languages.

Statistical Test

Since the assumption of normality was violated for all groups, we proceeded with the non-parametric **Kruskal-Wallis H test**, which does not require the normality assumption.

- **Kruskal-Wallis Test: Engagement Score vs Location**
  p-value: 0.1212
  No statistically significant difference in Engagement Score across Locations.
- **Kruskal-Wallis Test: Conversion Rate vs Location**
  p-value: 0.5028
  No statistically significant difference in Conversion Rate across Locations.
- **Kruskal-Wallis Test: Engagement Score vs Language**
  p-value: 0.8735
  No statistically significant difference in Engagement Score across Languages.
- **Kruskal-Wallis Test: Conversion Rate vs Language**
  p-value: 0.7446
  No statistically significant difference in Conversion Rate across Languages.

Post Hoc Test & Effect size

Since the significance tests in the previous subsection did not reveal any statistically significant differences, no post hoc analyses were performed.

Results

We fail to reject the null hypothesis of homogeneous variances for both conversion rates and engagement scores across different geographic locations. This allows for the application of comparative statistical tests, such as the Kruskal-Wallis test, to assess potential median differences in these metrics across locations without concerns regarding unequal variance.

Regarding language-based differences, the analysis reveals homogeneous variances in conversion rates, suggesting consistency in conversion outcomes across the languages studied. However, a statistically significant heterogeneity in variance is observed for engagement scores across different languages. This indicates that the dispersion of engagement scores varies considerably depending on the language of the campaign. Consequently, when comparing engagement scores across languages, statistical methods robust to unequal variances, such as Welch's ANOVA or non-parametric alternatives like the Kruskal-Wallis test, should be employed to ensure the validity of the comparisons.

Q4 Engagement Score vs Target Audience, Campaign Goal

Hypothesis Formulation

- **Null Hypothesis ($H_{0,Audience}$):** There is no significant difference in engagement scores across different target audiences.
- **Alternative Hypothesis ($H_{a,Audience}$):** There is a significant difference in engagement scores across at least one target audience compared to others.
- **Null Hypothesis ($H_{0,Goal}$):** There is no significant difference in engagement scores across different campaign goals.
- **Alternative Hypothesis ($H_{a,Goal}$):** There is a significant difference in engagement scores across at least one campaign goal compared to others.

Data Visualization

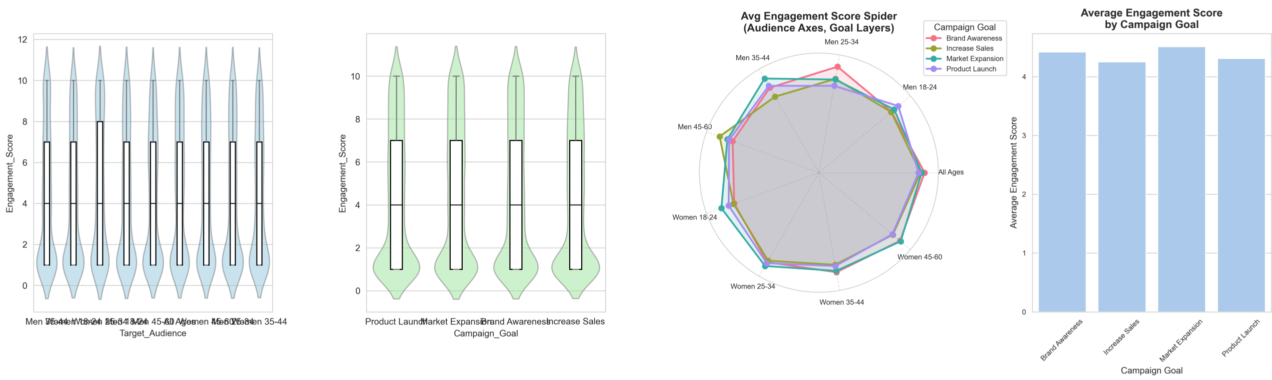

Figure 10: Visualizations of engagement scores across different target audiences and campaign goals. The first plot shows box plots within violin plots for engagement scores by target audience. The second plot displays similar plots for engagement scores by campaign goal. The third plot is a spider plot illustrating the average engagement score for each campaign goal across different target audience segments. The fourth plot is a bar graph showing the average engagement score for each campaign goal.

Assumptions Check

We assume independence of the engagement scores for different target audiences and campaign goals. We tested for normality **(Q-Q plot + KS test)** and homogeneity of variances across groups **(Levene's test)**.
Q-Q plot + KS test results:

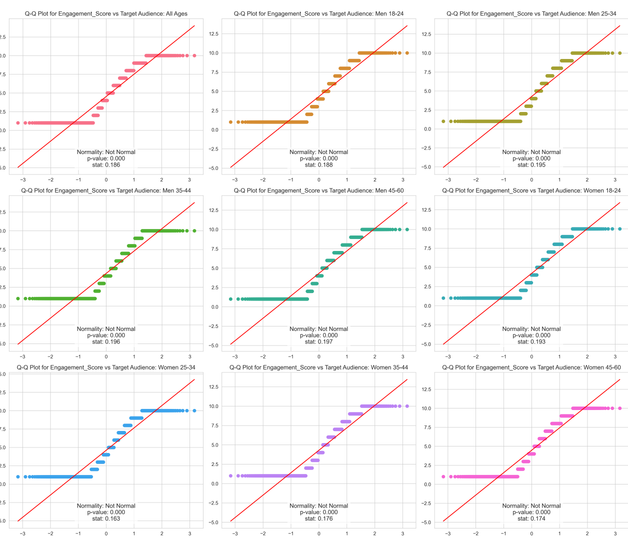

Figure 11: Q-Q plots with KS test results for engagement scores across different target audiences.

Figure 12: Q-Q plots with KS test results for engagement scores across different campaign goals.

Observations

None of the groups are normally distributed. Normality assumption not met for Engagement Score across both Target Audience and Campaign Goal.

Levene's Test Results:

- **Levene's test for Engagement Score based on Target Audience:**
  Test Statistic: 0.8168
  p-value: 0.5876
- **Levene's test for Engagement Score based on Campaign Goal:**
  Test Statistic: 0.7200
  p-value: 0.5399

Observations

Since p-value $> \alpha$ variances are homogeneous for Engagement Score across Target Audiences. Since p-value $> \alpha$ variances are homogeneous for Engagement Score across Campaign Goals.

Statistical Test

Since the assumption of normality was violated for all groups, we proceeded with the non-parametric **Kruskal-Wallis H test**, which does not require the normality assumption.

- **Kruskal-Wallis Test: Engagement Score vs Target Audience**
  p-value: 0.1230
  No statistically significant difference in Engagement Score across Target Audiences.
- **Kruskal-Wallis Test: Engagement Score vs Campaign Goal**
  p-value: 0.0529
  No statistically significant difference in Engagement Score across Campaign Goals.

Post Hoc Test & Effect size

Since the significance tests in the previous subsection did not reveal any statistically significant differences, no post hoc analyses were performed.

Results

Non-parametric Kruskal-Wallis tests were used due to non-normal engagement scores. Levene's test confirmed homogeneous variances for both *Target Audience* ($p = 0.5876$) and *Campaign Goal* ($p = 0.5399$).
The Kruskal-Wallis test revealed no significant difference in median engagement scores based on *Target Audience* ($H = 4.321$, $p = 0.1230$) or *Campaign Goal* ($H = 5.851$, $p = 0.0529$).
Thus, this analysis indicates that neither *Target Audience* nor *Campaign Goal* significantly impacted user engagement scores in this dataset.

Q5 Which factors can be used to predict the ROI?

Hypothesis Formulation

Since ROI is a continuous variable and this is a prediction task we will use multi-linear regression analysis.

- **Null Hypothesis ($H_{0,Regression}$):** There is no statistically significant linear relationship between ROI and the set of independent variables. That is, all regression coefficients are zero.
- **Alternative Hypothesis ($H_{a,Regression}$):** At least one independent variable has a statistically significant linear relationship with ROI. That is, at least one regression coefficient is not zero.

Data Visualization

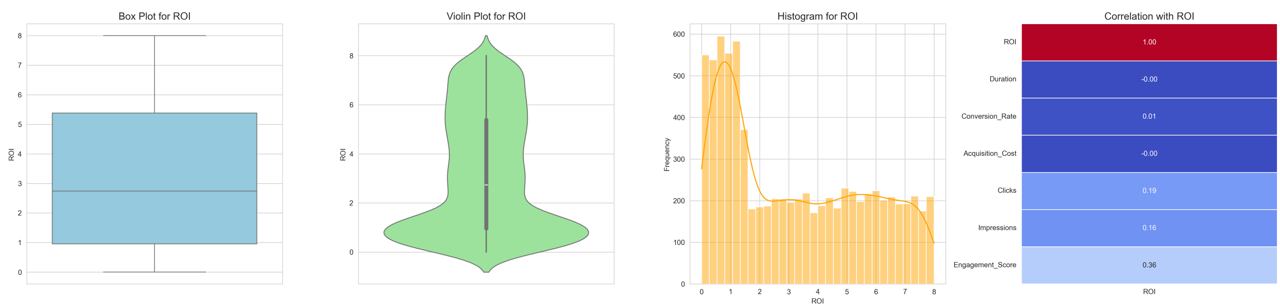

Figure 13: (Left to Right) Box-plot showing the median and quartiles of the ROI data distribution, violin plot showing the density distribution of ROI data across it's range, histogram depicting the frequency of each bucket of the data bin and finally the correlation plot showing the correlation value between each variable and ROI.

From the correlation values we can see that ROI has little to no correlation with any of the other numerical values this suggests that we won't be able to infer anything useful from the regression analysis.

Performing Regression

We begin with multiple independent variables to perform a multivariate regression analysis with Return on Investment (ROI) as the dependent variable. However, one of the common challenges in multivariate regression is multicollinearity, where independent variables are highly correlated with one another. This redundancy can distort the model by inflating variance and reducing the interpretability of coefficients. To address this, we employ Variance Inflation Factor (VIF) analysis, which helps identify and eliminate highly collinear features, ensuring that only informative and non-redundant variables are retained for building a robust regression model.

Starting VIF elimination process...

- Max VIF: 50613.833361870034 for feature `Impressions`
- Removing `Impressions` with VIF: 50613.833361870034
- Max VIF: 22.85847008957904 for feature `Acquisition_Cost`
- Removing `Acquisition_Cost` with VIF: 22.85847008957904
- Max VIF: 6.643896945408754 for feature `Clicks`
- All VIFs are below threshold, stopping elimination

Final VIF values:

Variable	VIF
`const`	14.302846
`Clicks`	6.643897
`Duration`	6.348047
`Engagement_Score`	1.263444
`Conversion_Rate`	1.000628

Table 4: Final VIF values after feature elimination

Assumptions Check

1. **Linearity:** The relationship between the dependent variable and the independent variables is linear.
2. **Independence of Errors:** The residuals (errors) are independent of each other. This is often checked using the Durbin-Watson test.
3. **Homoscedasticity:** The variance of the residuals is constant across all levels of the independent variables.
4. **No Multicollinearity:** Independent variables are not highly correlated with each other. This is usually tested using the Variance Inflation Factor (VIF).
5. **Normality of Residuals:** The residuals are normally distributed, which can be checked using Q-Q plots or statistical tests like the Shapiro-Wilk test.
6. **No Autocorrelation:** Particularly in time series data, the residuals should not be correlated with each other over time.
7. **Measurement Level:** The dependent variable is continuous, and the independent variables are either continuous or categorical (with appropriate encoding).

We assume the independence of errors, we have already mitigated multicollinearity using VIF reduction, we assume that there is no autocorrelation, and both independent and dependent variables are numerical, we will test for the assumption of normality, linearity and hooscedasticity in this section.

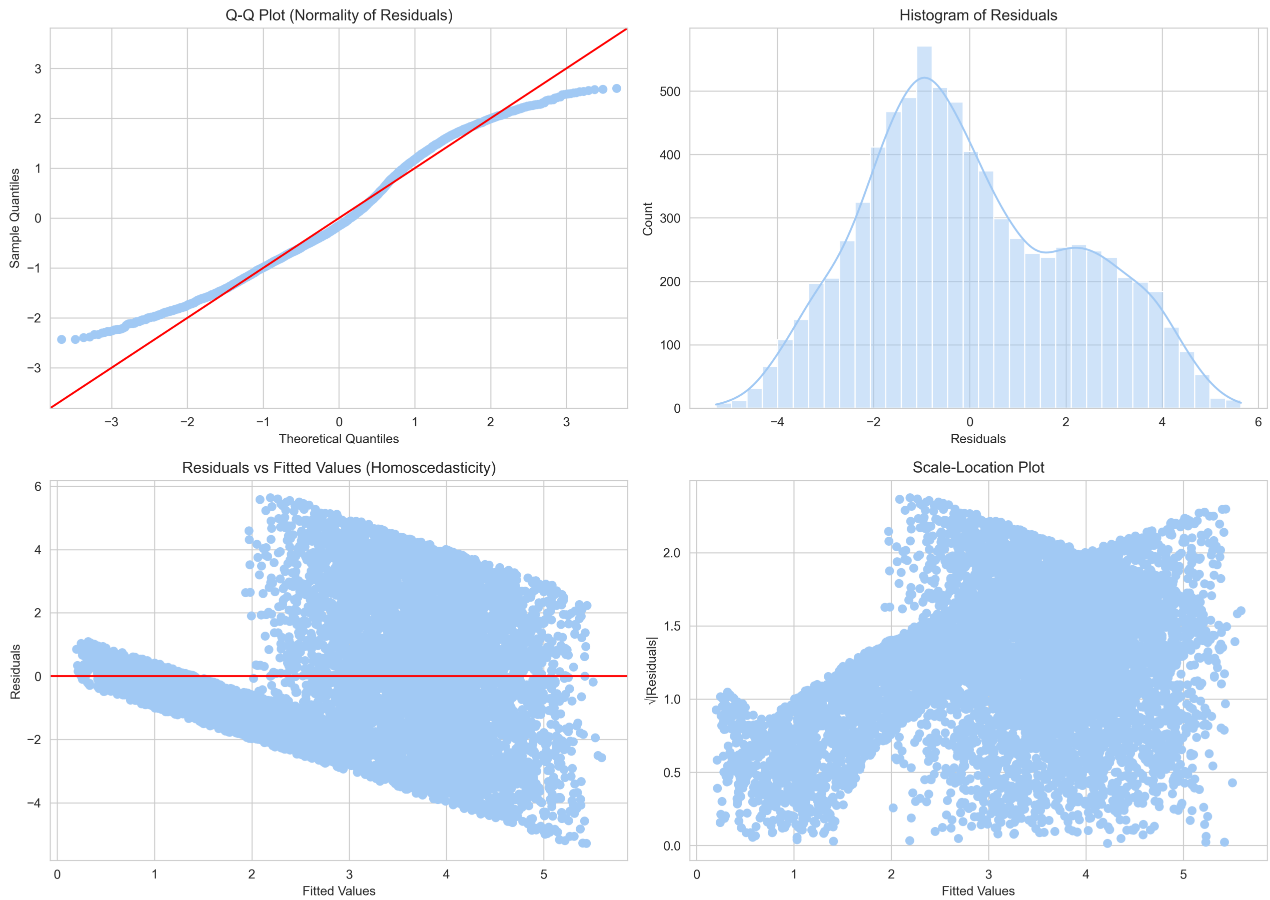

Figure 14: This figure includes multiple diagnostic plots: the Q-Q plot and histogram assess the normality of residuals, while the residuals vs. fitted values plot is used to detect patterns indicating non-linearity or heteroscedasticity. Additionally, the scale-location plot helps evaluate the homoscedasticity (constant variance) of residuals across fitted values.

Table 5: Statistical Test Results

Test	Statistic	p-value
Shapiro-Wilk Test for Normality	0.9795	0.0000
Breusch-Pagan Test for Homoscedasticity	625.0105	0.0000

Interpretation: Residuals are not normally distributed ($p < 0.05$); heteroscedasticity is present ($p < 0.05$).

Table 6: Prediction Metrics for VIF-Reduced Model

Metric	Value
R^2 Score	0.2218
Mean Squared Error (MSE)	4.7082
Root Mean Squared Error (RMSE)	2.1698
Mean Absolute Error (MAE)	1.7893

Hence we can see that the assumptions of normality and homoscedasticity are violated.

Predictions

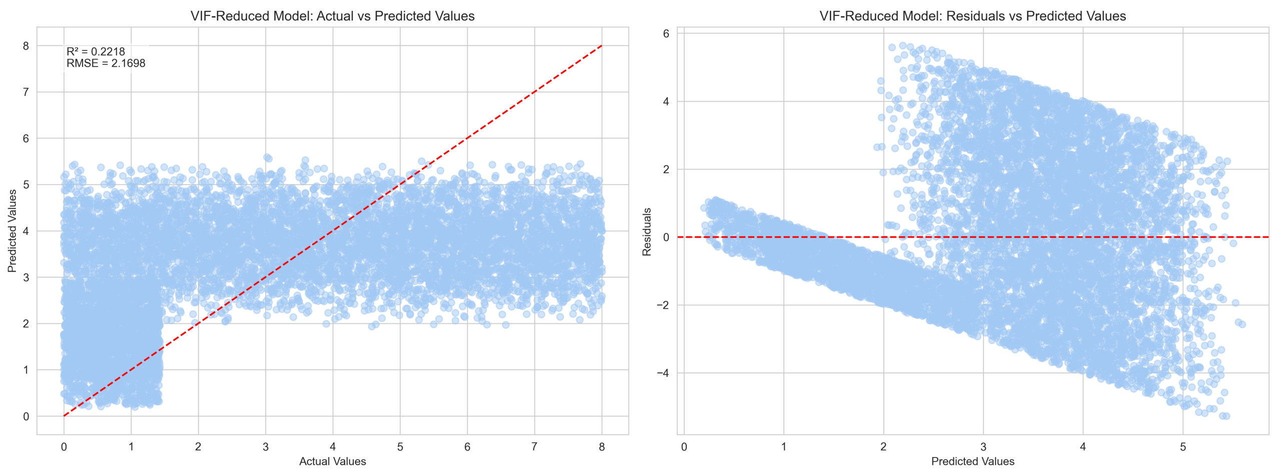

Figure 15: Predictions vs Actual values plots.

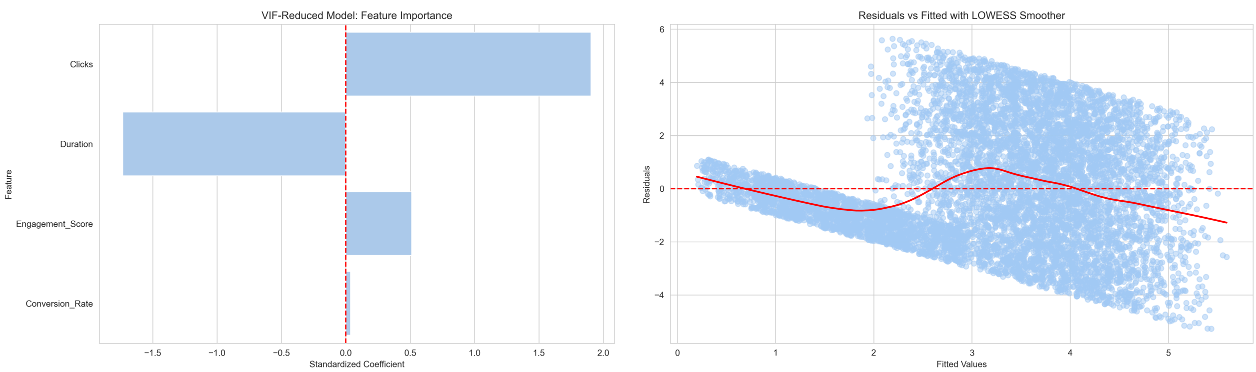

Figure 16: Importance of features (left) plot and LOWESS smoother fitter residuals plot (right).

From the figures we can clearly see that the best fit line is not a good predictor for the ROI hence we fail to reject the null hypothesis.

Results

The multiple linear regression analysis did not yield meaningful insights due to weak relationships between ROI and the independent variables. Diagnostic tests revealed violations of key assumptions such as normality and homoscedasticity, and the model demonstrated low predictive power overall. As a result, we fail to reject the null hypothesis, indicating that the regression does not provide sufficient evidence to support a statistically significant linear relationship between ROI and the selected features.

Q6 Does the duration of the campaign have any effect on impressions?

Hypothesis Formulation

To determine if the duration of a campaign has a statistically significant effect on the number of impressions, we will employ a permutation test. This non-parametric method is suitable for comparing groups when the assumptions of traditional parametric tests might not be met.

- **Null Hypothesis ($H_{0,Permutation}$):** There is no statistically significant difference in the mean number of impressions across different campaign durations.
- **Alternative Hypothesis ($H_{a,Permutation}$):** There is a statistically significant difference in the mean number of impressions across at least one pair of different campaign durations.

Data Visualization

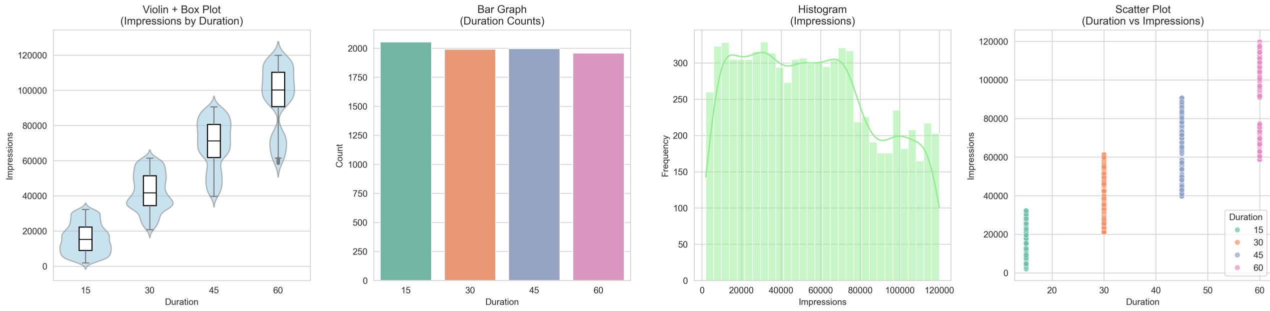

Figure 17: Visualization of Campaign Duration and Impressions. (Left to Right) A combined box plot and violin plot displaying the distribution of impressions for different campaign durations, a bar graph showing the frequency of various campaign durations, a histogram illustrating the frequency distribution of impression counts, and a scatter plot depicting the relationship between campaign duration and impressions.

The initial data visualization provides an overview of the distribution of impressions across different campaign durations. The scatter plot, in particular, can give a preliminary visual indication of any potential relationship between these two variables.

Permutation Test

To formally test our hypothesis, we performed a permutation test to assess whether the observed differences in impressions across campaign durations are likely to have occurred by chance. The observed statistic, which measures the between-group variance, was calculated, and then compared to the distribution of this statistic obtained by randomly shuffling the impression data across the different duration groups many times.

Permutation Test Results:

- Observed Statistic (Between-group variance): 7.170×10^{12}
- Permutation Test p-value: 0.0000

The extremely low p-value (0.0000) from the permutation test indicates that the observed difference in impressions across campaign durations is highly unlikely to have occurred by random chance. Therefore, we reject the null hypothesis.

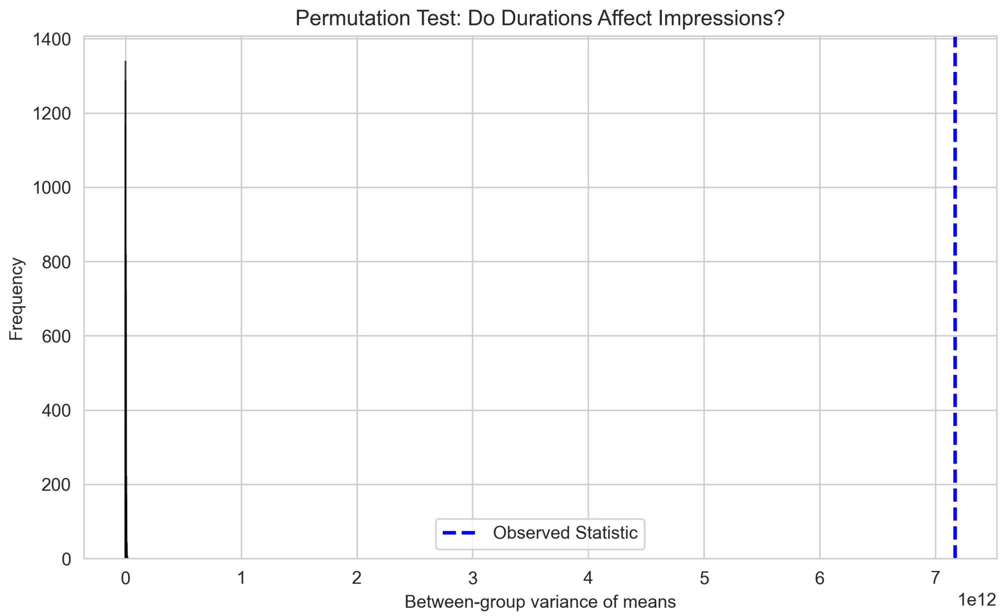

Figure 18: The result shows how the observed value of difference between groups is far away from the distribution due to random shuffling.

Post Hoc Analysis

Since the permutation test revealed a significant overall difference, we conducted a post hoc analysis to identify which specific pairs of campaign durations exhibit statistically significant differences in their mean impression counts. We performed pairwise comparisons between all unique pairs of campaign durations using permutation tests.

Pairwise Permutation Test Results:

```
Comparison: 30 Days vs 60 Days | p-value: 0.0000
Comparison: 30 Days vs 45 Days | p-value: 0.0000
Comparison: 30 Days vs 15 Days | p-value: 0.0000
Comparison: 60 Days vs 45 Days | p-value: 0.0000
Comparison: 60 Days vs 15 Days | p-value: 0.0000
Comparison: 45 Days vs 15 Days | p-value: 0.0000
```

The results of the pairwise permutation tests show statistically significant differences (p ¡ 0.05) in the mean number of impressions for all pairs of campaign durations.

Correction for Multiple Comparisons

To account for the increased risk of Type I errors (false positives) due to performing multiple comparisons, we applied the Bonferroni correction. This method adjusts the significance level (α) for each test by dividing it by the total number of comparisons made.

Pairwise Permutation Test Results (Bonferroni Corrected):

```
Comparison: 30 Days vs 60 Days | p-value (corrected): 0.0000
Comparison: 30 Days vs 45 Days | p-value (corrected): 0.0000
Comparison: 30 Days vs 15 Days | p-value (corrected): 0.0000
Comparison: 60 Days vs 45 Days | p-value (corrected): 0.0000
Comparison: 60 Days vs 15 Days | p-value (corrected): 0.0000
Comparison: 45 Days vs 15 Days | p-value (corrected): 0.0000
```

Even after applying the Bonferroni correction, all pairwise comparisons remain statistically significant (corrected p ¡ 0.05).

Results

The permutation test revealed a highly significant difference in the number of impressions across different campaign durations ($p < 0.0001$). Furthermore, the post hoc pairwise permutation tests, even after applying the conservative

Bonferroni correction for multiple comparisons, indicate that the mean number of impressions is significantly different for every pair of campaign durations (15, 30, 45, and 60 days). Therefore, we reject the null hypothesis and conclude that the duration of the campaign has a statistically significant effect on the number of impressions received. Campaigns with different durations tend to generate different numbers of impressions.

Conclusion

The statistical analyses conducted across various dimensions of social media campaigns yield the following key insights:

- **Social Media Platform vs. Engagement Levels:** The choice of social media platform significantly influences engagement levels. Instagram consistently shows higher engagement compared to Facebook and Twitter, while Pinterest exhibits anomalous behavior due to its constant engagement scores. Pairwise comparisons reveal that Instagram significantly outperforms both Facebook and Twitter, although the difference is not practically large. No significant difference is observed between Facebook and Twitter.
- **Campaign Duration vs. Engagement and Conversion:** The duration of a campaign does not have a statistically significant impact on either engagement scores or conversion rates. This suggests that both short-term and long-term campaigns perform similarly in these aspects.
- **Geographic Location and Language vs. Engagement and Conversion:** No significant variance is found in conversion rates and engagement scores across different locations. For language, while conversion rates remain consistent, engagement scores show significant variance, indicating the necessity for variance-robust statistical methods when comparing engagement across languages.
- **Engagement Score vs. Target Audience and Campaign Goal:** Neither the target audience nor the campaign goal significantly affects engagement scores. Homogeneous variances were confirmed, but the Kruskal-Wallis test showed no statistically significant differences.
- **Predictors of ROI:** Multiple linear regression analysis failed to identify any strong predictors of ROI. The model violated assumptions of normality and homoscedasticity and showed low predictive power, suggesting no statistically significant linear relationship between ROI and the selected campaign features.
- **Campaign Duration vs. Impressions:** In contrast to engagement and conversion, campaign duration has a statistically significant impact on impressions. Permutation tests reveal that each campaign duration (15, 30, 45, and 60 days) generates a significantly different number of impressions, indicating duration as a critical factor for visibility.

Overall, platform choice and campaign duration (in terms of impressions) emerge as significant factors in digital campaign performance, whereas other factors such as campaign goal, target audience, and duration (in terms of engagement or conversion) exhibit limited statistical influence.

References

[1] Wikipedia. *History of Advertising*. `https://en.wikipedia.org/wiki/History_of_advertising/`. Accessed: 20-Feb-2025.

[2] Power Direct Marketing. *A Brief History of Print Advertising*. `https://powerdirect.net/history-print-advertising/`. Accessed: 20-Feb-2025.

[3] Spotify Editorial Team. *The history of radio advertising and the state of audio today*. `https://ads.spotify.com/en-US/news-and-insights/history-of-radio-advertising/`. Accessed: 20-Feb-2025.

[4] Wikipedia. *Television advertisement*. `https://en.wikipedia.org/wiki/Television_advertisement/`. Accessed: 20-Feb-2025.

[5] Wikipedia. *Digital Marketing*. `https://en.wikipedia.org/wiki/Digital_marketing/`. Accessed: 20-Feb-2025.

[6] AmberCreative. *The Evolution Of Social Media Ads: From Banners To TV... Again*. `https://ambercreative.sg/blog/evolution-of-social-media-ads`. Accessed: 24-Apr-2025. 2020.