I. Introduction:

It is apparent that social media has increasingly become inseparable entertainment activities in human lives. Statistics from Datareportal showed that at the start of July 2025, the number of social media users reached 5.41 billion users, accounting for 65.7% of total population worldwide (Datareportal, 2025). Among social media platforms, Facebook still dominates the market, accounting for approximately 67.32% of total social media usage across all devices (Statcounter, 2025). By leveraging data analytics tools, businesses can better understand audience behavior and preferences, evaluate campaign performance, and refine their strategies for improved outcomes.

This report indicates steps of analyzing social media marketing performance and generating insights from data. The dataset was collected from a public dataset in Kaggle. The analysis aims to:

- Determine factors that influence customer conversion by running linear regression model
- Evaluate the digital performance of each marketing campaign throughout exploratory data analysis (EDA).

- Suggest actionable recommendations to improve the performance and cost optimization.

Data features:

- **ad_id:** Unique identifier assigned to each advertisement.
- **xyzcampaignid:** Unique ID representing each marketing campaign run by the XYZ company.
- **fbcampaignid:** Identifier used by Facebook to track the corresponding campaign.
- **age:** Age group of the audience who viewed the advertisement.
- **gender:** Gender of the individual exposed to the ad.
- **interest:** Numerical code representing the interest category associated with the viewer, as listed in their Facebook profile.
- **Impressions:** Total number of times the advertisement was displayed to users.
- **Clicks:** Number of times users clicked on the advertisement.
- **Spent:** Amount of money paid by XYZ company to Facebook for displaying the ad.
- **Total_Conversion:** Number of users who expressed interest in the product after viewing the ad.
- **Approved_Conversion:** Number of users who completed a purchase after viewing the ad.

II. Limitation

1. Lack of timeframe information

The absence of specific time-related data significantly impacts the analysis. Without knowing the exact launch dates and durations of each campaign, it is not possible to accurately assess their performance over time or account for variations due to seasonal factors (e.g., New Year, Christmas). To maintain consistency and comparability, this analysis assumes that all three campaigns were executed concurrently over a standardized period of three months. As a result, any potential influence from external time-based factors has been excluded from the evaluation.

2. Insufficient financial data

The unavailability of critical financial metrics, such as total revenue, operational expenses, and human resource costs, limits the scope of this report. Without these inputs, it is not feasible to

assess the overall economic performance of the business or determine profitability. Consequently, this analysis focuses solely on digital marketing effectiveness. Specifically, it evaluates the cost of acquiring customers through ad spending and identifies which campaign demonstrates the most cost-effective performance.

III. Methodologies:

All the data analysis will be coded in Python stored in Colab Notebook.

For exploratory data analysis, I will apply several techniques to discover the insight, moving from a broad overview to a more detailed focus. First, I will check the overall performance of the company throughout three campaigns (a, b and c). Then, each feature will be examined in detail to identify differences across age and gender segments. In this stage, the most common graph used to evaluate the performance is multiple bar charts. To investigate the correlation between 2 variables, I will use scatter plots. Moreover, to understand the distribution of a variable (such as spent), I will use histogram charts.

To identify factors influencing conversion, a linear regression model will be developed. Linear regression analysis estimates the value of one variable based on another. The variable being predicted is the dependent variable, while the variable used for prediction is the independent variable (IBM). In this dataset, after quickly checking the correlation among variables through correlation matrix or scatter plot, **Total_Conversion** is chosen to be the target, and **Spent, Clicks, Age and Gender** are predictors. Since Linear Regression model only works with numerical data, categorical data such as age and gender need to be encoded into dummies data.

IV. Findings

A. Exploratory data analysis

1. **Research question 1:** What is the frequency distribution of ads regarding audience's demographic ?

Regarding Figure 1, there are 592 ads targeted to males, 41 ads higher than females (551 total ads). Therefore, it can be said that the company Figure 2 demonstrates that people aged 30-34

years old took a dominant proportion in ad distribution. This suggests that the company's digital strategies focus on young people (30-34 years old) and focus on males rather than females.

Figure 1.

Visualization of gender distribution

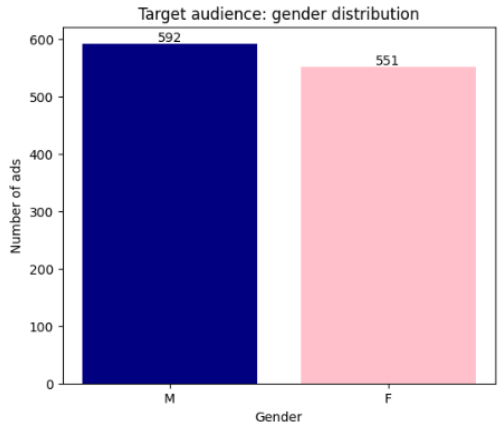
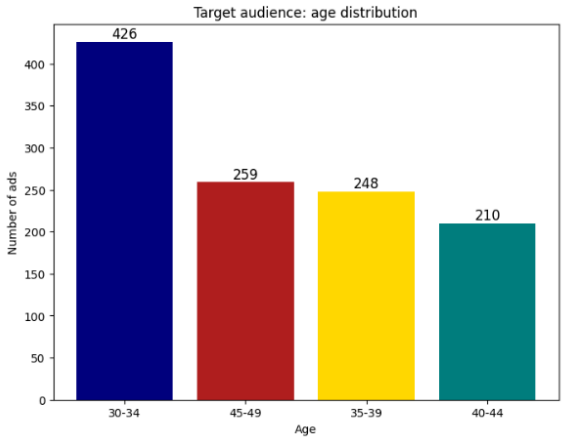


Figure 2.

Visualization of age distribution



2. **Research question 2:** What is the frequency distribution of ads and marketing spending regarding campaigns ?

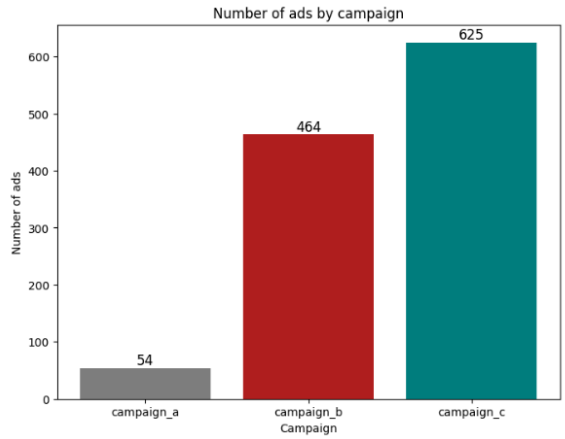
Campaign_c generated the largest number of ads (625) and was also the most expensive with a total of $55,662. Ranked second, campaign_b produced 464 ads at a cost of $2,893. However when comparing ad number and marketing spend of campaign_c with campaign_b, ad numbers of campaign_b comprised 80% of those in campaign_c. However, the spending was far behind, only 5% compared to the investment in campaign_c. From that finding, it can be concluded that campaign_b emphasizes organic performance and search engine optimization (SEO) rather than relying heavily on search engine marketing (SEM). In contrast, campaign_a had fewer ads (54 ads) and lower spending ($150), indicating it was likely a smaller-scale or niche campaign designed to optimize costs while targeting specific market segments.

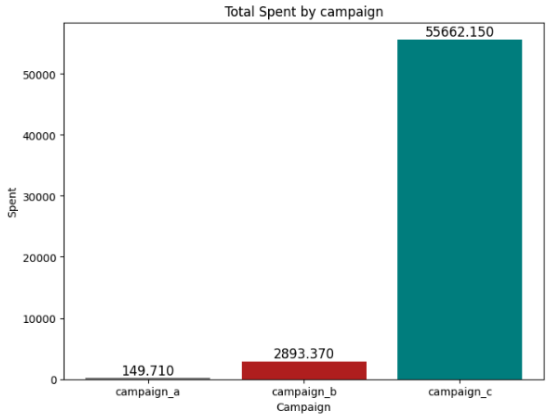
Figure 3.

Visualization of ads distribution

Figure 4.

Visualization of spending distribution





3. **Research question 3:** What is the target audience of each campaign ?

By looking at 2 above figures, the target audience of each campaign is discovered:

- **Campaign_a:** males aged from 30-34
- **Campaign_b:** females aged from 30-34
- **Campaign_c:** males aged from 30-34

Figure 5.

Visualization of gender distribution by campaign

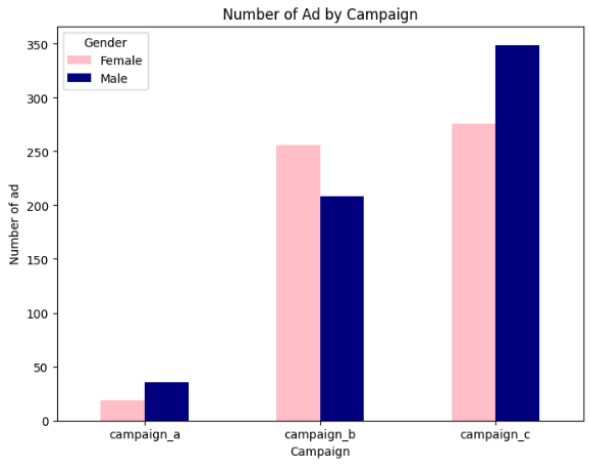
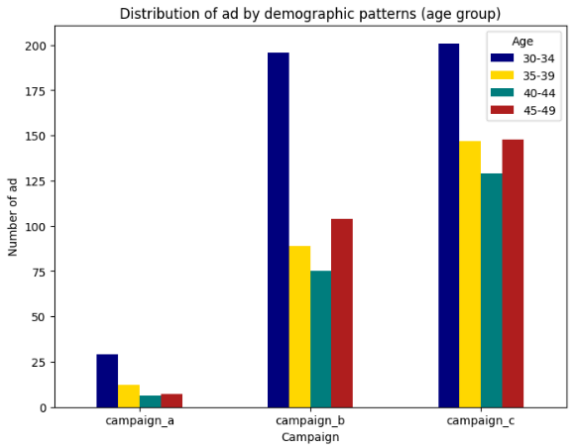


Figure 6.

Visualization of age distribution by campaign

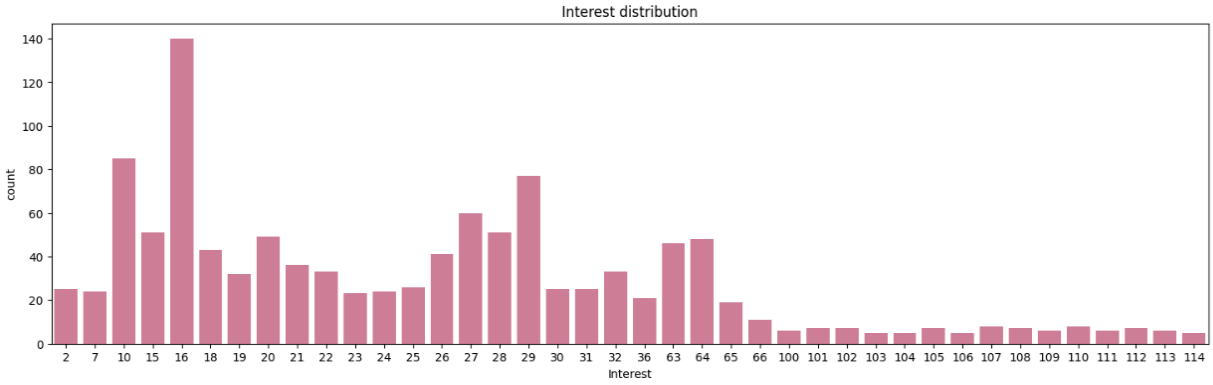


4. **Research question 4:** What are the top interests ?

Interest code #16, #10 and #29 generated the highest number of ads, indicating strong audience engagement in these interests.

Figure 7.

Visualization of interest distribution



5. **Research question 5:** How is the impression performance of each campaign ?

Ads targeting female audiences generated higher impressions, particularly in campaign_b and campaign_c. Similarly, ads aimed at individuals aged 45–49 consistently achieved the highest impression counts across all three campaigns.

Problem: This trend indicates a targeting inconsistency, as the campaigns attracted substantial engagement from audiences outside their intended primary segments, suggesting a misalignment between targeting strategy and actual audience reach.

Figure 8.

Visualization of impression distribution by campaign and age

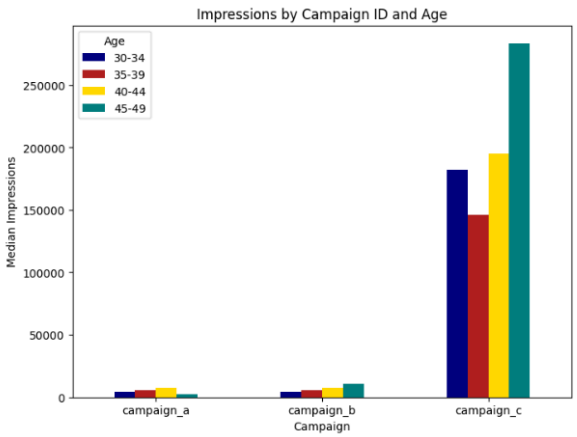
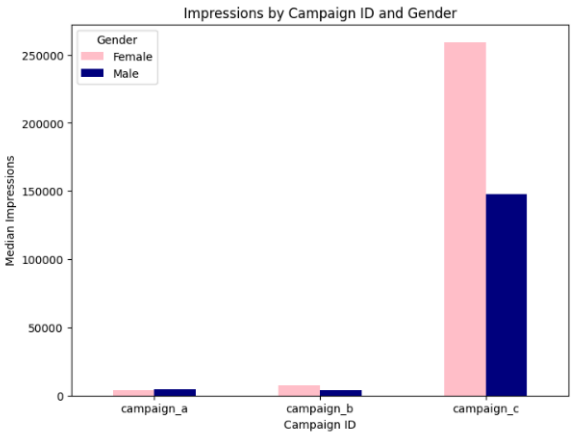


Figure 9.

Visualization of impression distribution by campaign and gender



6. **Research question 6:** How is the click performance of each campaign ?

Ads targeting individuals aged 45–49 recorded the highest number of clicks, outperforming all other age groups. In contrast, the intended target group (30–34 years old) generated the lowest click volume, indicating a gap between targeting intention and audience responsiveness. Additionally, ads directed toward female audiences consistently achieved significantly higher click rates than those targeting males, with the disparity being most pronounced in campaign_c.

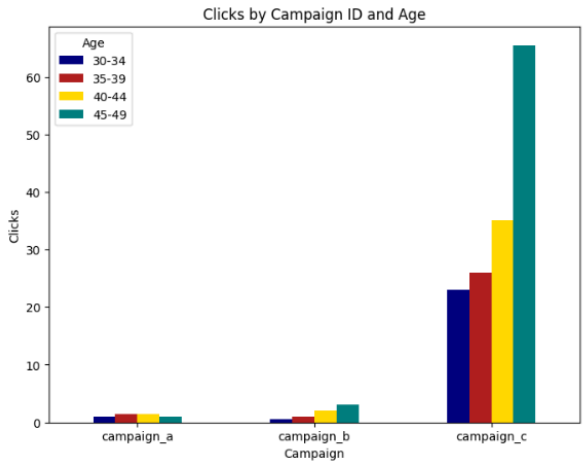
Problem: The company failed to attract clicks from the target audience group. Possibly because of poor targeting, irrelevant or uncompelling messages, lack of Call-to-Action (CTA), or ineffective keyword selection and bidding strategies not attractive enough to generate many clicks.

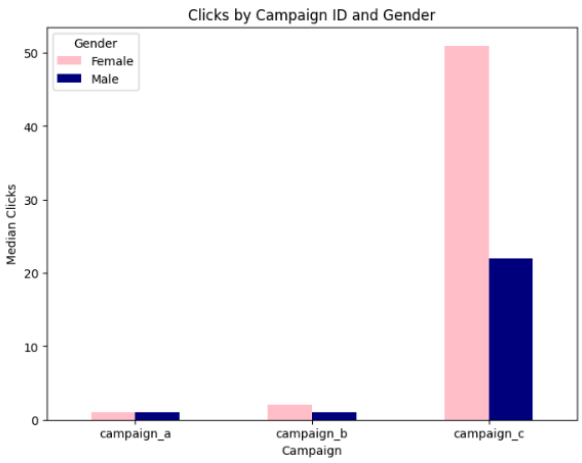
Figure 10.

Visualization of click distribution by campaign and age

Figure 11.

Visualization of click distribution by campaign and gender



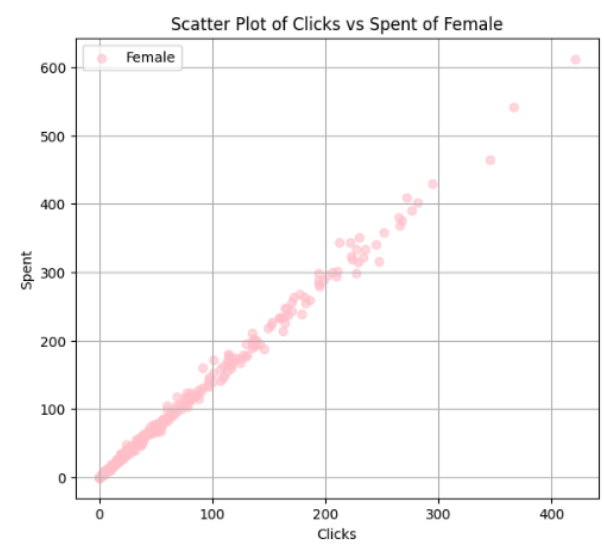


7. **Research question 7:** Is there any correlation between clicks and spending regarding gender and age ?

According to Figure 12 and 13, Spent and Clicks have strong correlation to each other. In other words, the company spent more money for ads targeted to women and people aged 45-49, contributing to low click in target audience.

Figure 12.

Correlation between clicks and spent by gender



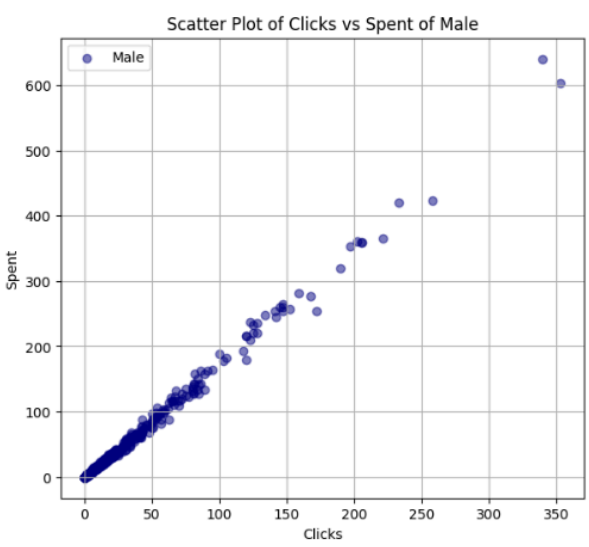
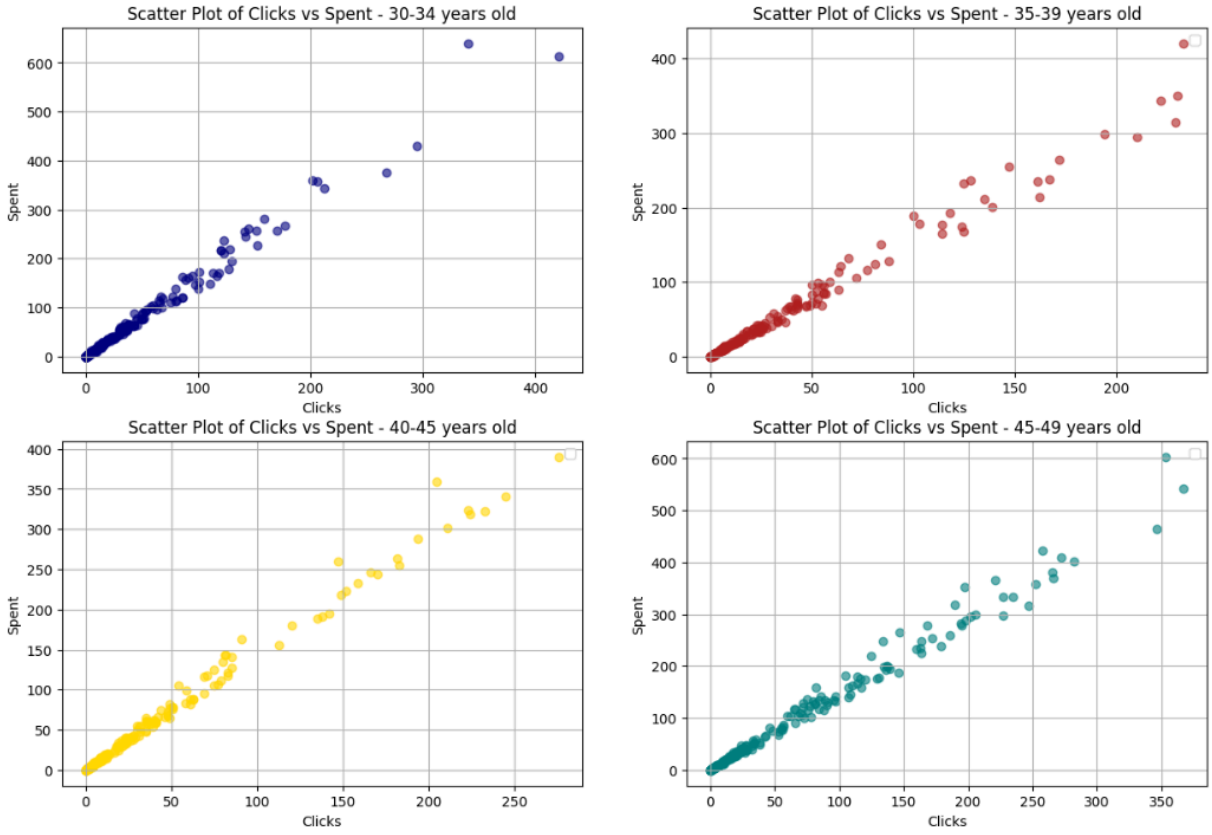


Figure 13.

Correlation between clicks and spent by age

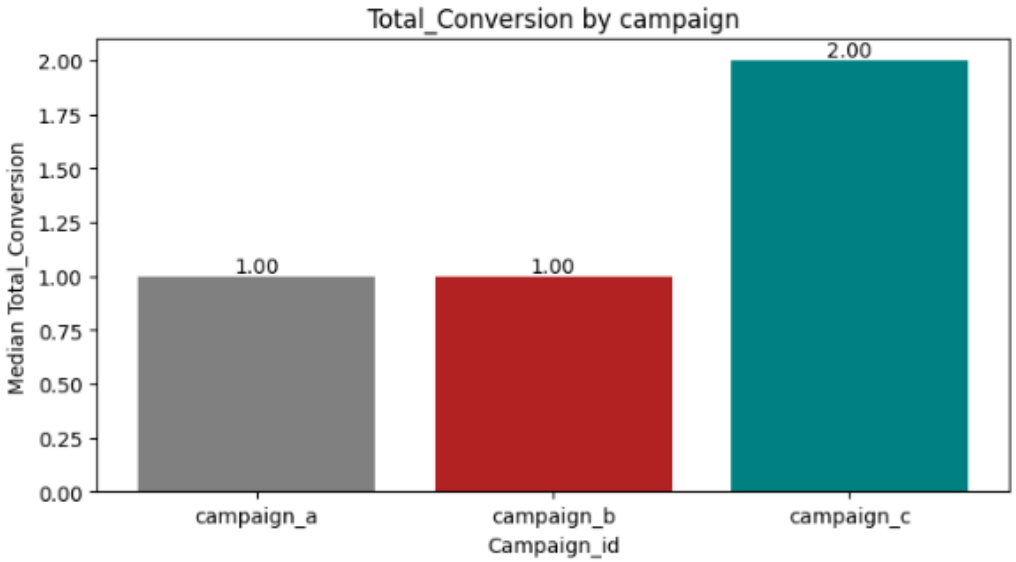


8. **Research question 8:** How is the total_conversion performance of each campaign ?

Overall, campaign_c achieved the highest number of conversions, around 2 conversions per ad. However, this ratio is still low, suggesting that the company failed to convert their audience effectively. Notably, despite its limited scale and lower investment, campaign_a achieved a total conversion rate comparable to campaign_b, highlighting the efficiency and effectiveness of its niche targeting strategy.

Figure 14.

Visualization of distribution of total_conversion by campaigns



9. **Research question 9:** How is the total_conversion performance of each campaign ?

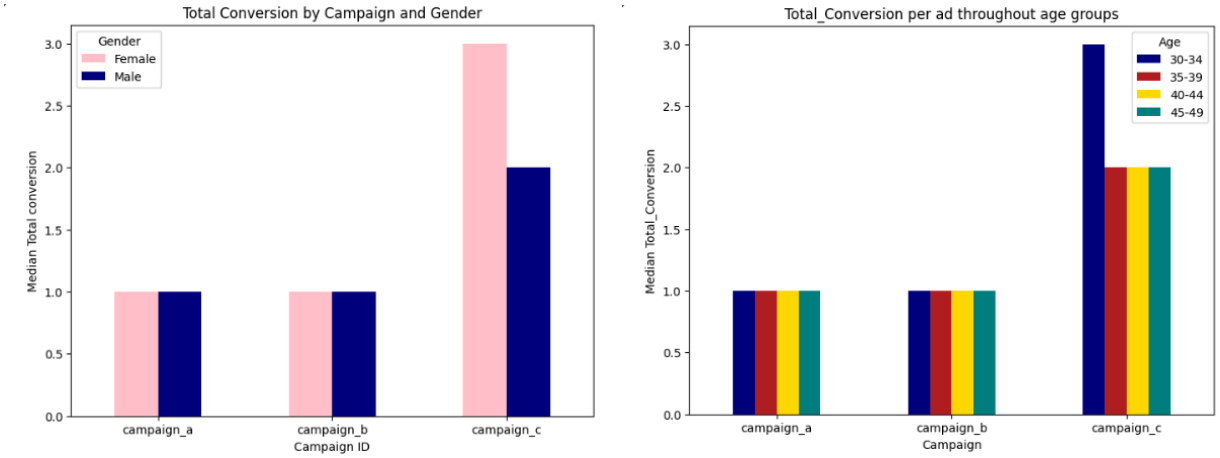
Although campaign_c achieved the highest number of conversions overall, the majority of these conversions came from female audiences (with 3 conversions per ad), indicating a misalignment between targeting strategy and intended male target audience. Despite recording fewer impressions and clicks, the 30–34 age group delivered the highest conversion rate, reinforcing the campaign's effectiveness in reaching its intended demographic segment.

Figure 15.

Visualization of click distribution by campaign

Figure 16.

Visualization of click distribution by campaign

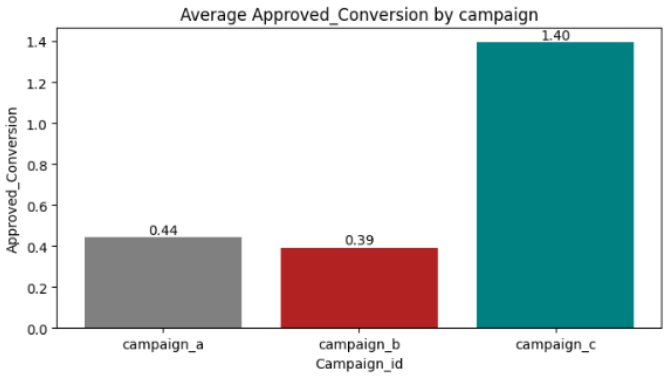


10. **Research question 10:** How is the approved_conversion performance of each campaign ?

The overall approved conversion rate remained relatively low (ranging from 0.39 to 1.4), indicating that the company converted only a small proportion of potential customers. Despite having lower ad density, engagement, and investment, campaign_a achieved a higher conversion rate than campaign_b, demonstrating its greater cost-effectiveness.

Figure 17.

Visualization of approved_conversion distribution by campaign

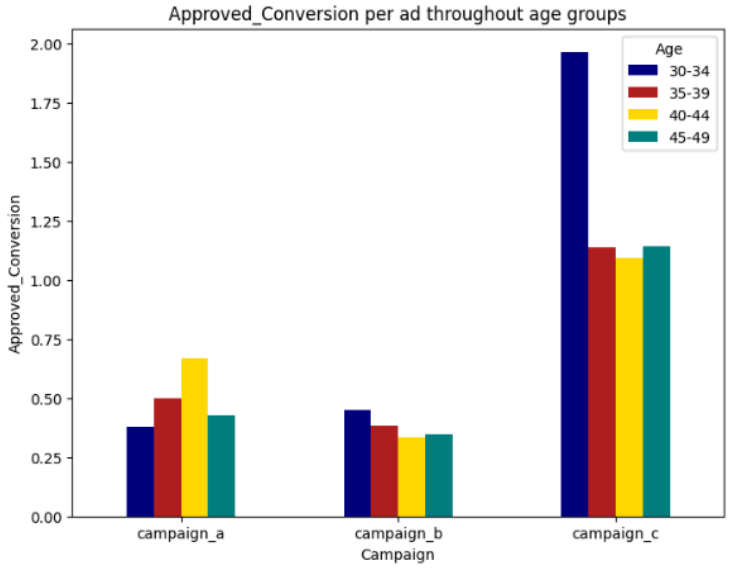


11. **Research question 11:** How is the approved_conversion performance of each campaign by age?

In terms of age segmentation, potential customers for campaign_b and campaign_c were primarily from the 30–34 age group, aligning with their intended target audience. Conversely, campaign_a attracted more conversions from the 40–44 age group, indicating an opportunity to focus future efforts on this demographic segment (Figure 18).

Figure 18.

Visualization of approved_conversion distribution by campaign



12. **Research question 11:** How is the approved_conversion performance of each campaign by gender?

There were no significant differences in approved conversions between male and female audiences. Therefore, the company does not need to design marketing strategies that cater for males and females separately.

Figure 19.

Visualization of approved_conversion distribution by gender

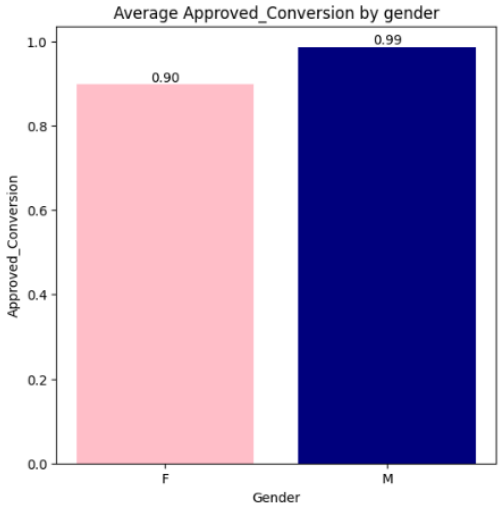
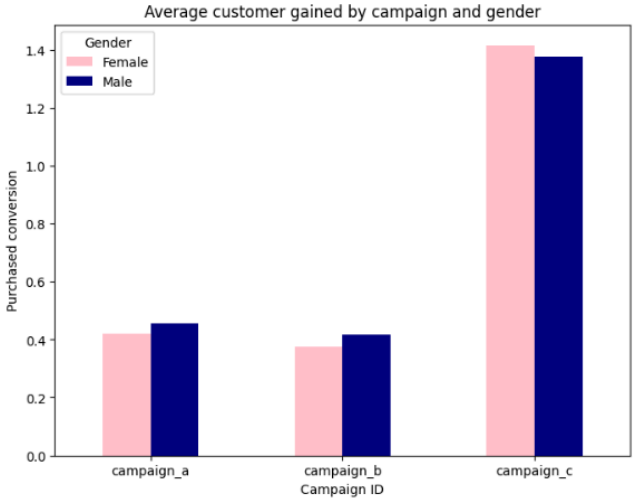


Figure 20.

Visualization of approved_conversion distribution by campaigns and gender



13. **Research question 13:** What are the relationships among features?

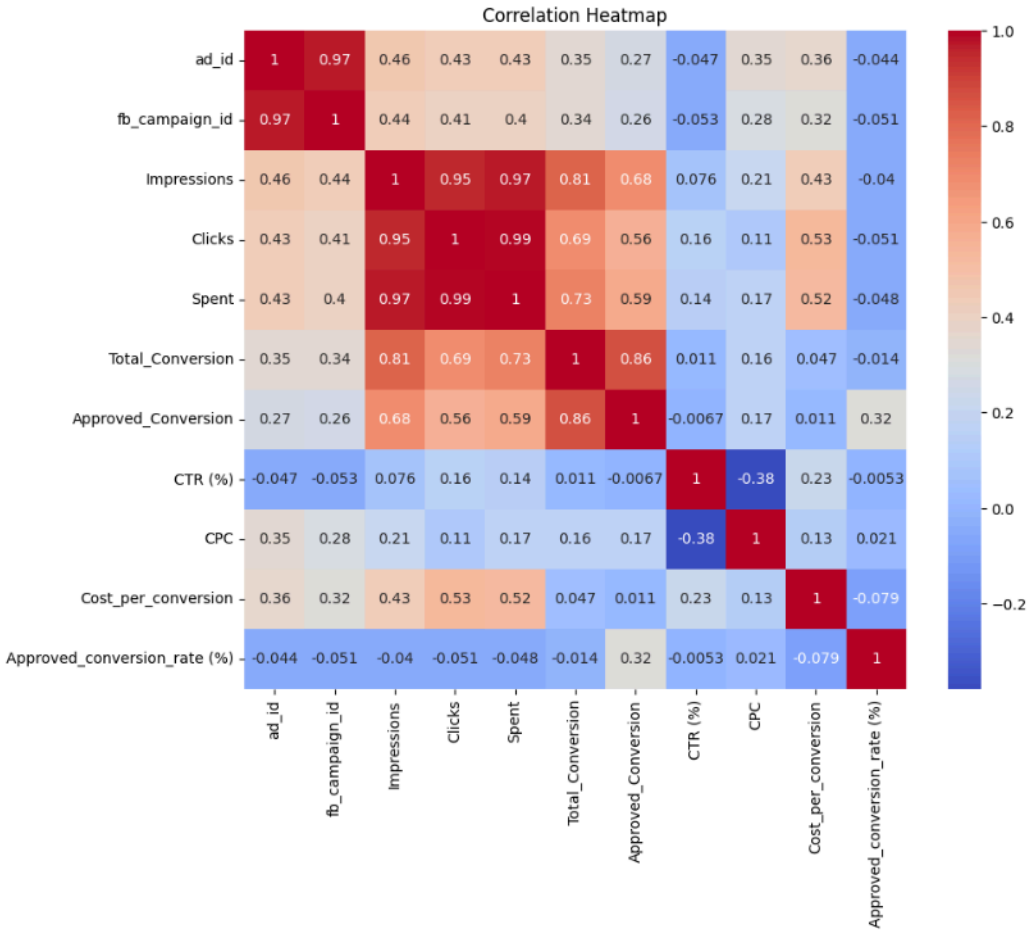
The figure 21 demonstrated correlation matrix of all numerical data in this dataset. The variables Impressions, Clicks, and Spent exhibit a strong positive correlation with one another (0.95–0.99), indicating that they convey largely overlapping information. Including all three in a regression model may therefore introduce multicollinearity issues.

A strong correlation between Total_Conversion and Impressions (0.81) suggests that greater ad exposure generally drives higher conversion outcomes. Similarly, Total_Conversion shows a moderate to strong correlation with Spent (0.73) and Clicks (0.69), indicating that increased ad spending and engagement contribute positively to conversion performance.

Meanwhile, the cost per click (CPC) negatively correlated (-0.38) with click-through rate (CTR), implying that higher CPC values are often associated with lower CTRs, possibly due to reduced cost efficiency in audience engagement.

Figure 21.

Correlation matrix



B. Data Modelling

1. **R-squared ($R^2 = 0.523$)**

This statistic, also known as the coefficient of determination, quantifies the proportion of variance in the **dependent variable** that is explained by the set of independent variables. In this model, an R^2 of 0.523 means that **52.3 %** of the variation in the outcome can be accounted for by the predictors. Given the complexity of human or consumer behavior (where many external factors are at play), an R^2 in this range is often considered reasonably strong and acceptable in practice (Jim, 2019).

2. **Adjusted R-squared (Adj. $R^2 = 0.520$)**:

In this case, an adjusted R^2 of 0.520—very close to the raw R^2—suggests the model maintains robustness even after accounting for model complexity.

3. **p-values of intercept and variables**:

For **Spent, Clicks, age_30–34, age_45–49, gender_F, gender_M**, the p-values are < 0.05 — indicating that those variables are **statistically significant** predictors (i.e. we can reject the null hypothesis that their coefficient is zero) at conventional confidence levels.

For **age_35–39** and **age_40–44**, p-values exceed 0.05, meaning that their effects are **not statistically significant** in this model (i.e. we do not have strong evidence that they influence the dependent variable directly in the presence of other variables).

When all independent variables are set to zero, the **level of total conversions** remains at **0.48**, indicating that even without marketing inputs, a small portion of conversions occurs, likely from organic or brand-loyal customers.

4. **coefficients:**

Spent ($\beta = 0.07$):
For every one unit increase in ad spending, the model predicts an average increase of 0.07 conversions, suggesting that higher investment positively influences conversion outcomes.

Clicks ($\beta = -0.06$):
Each additional click slightly reduces conversions by 0.06, implying potential inefficiency or unqualified traffic, possibly due to misleading ad content or poor audience targeting.

Age_30–34 ($\beta = 1.02$):
A one-unit increase in the likelihood of the ad being shown to individuals aged 30–34 is associated with a 1.02 increase in total conversions, the highest coefficient among all age groups, confirming this segment as the most responsive audience.

Age_45–49 (β = -0.63):

Each unit increase in exposure to the 45–49 age group decreases total conversions by 0.63, suggesting that this demographic is less receptive to the campaign and may not warrant heavy targeting investment.

Gender_F (β = 0.24):

Ads shown to female audiences increase conversions by 0.24, indicating a moderate positive response from this demographic.

Gender_M (β = 0.25):

Ads shown to male audiences increase conversions by 0.25, slightly higher than the female group, indicating that males generated more conversions overall.

Figure 22.

Linear regression output

OLS Regression Results

Dep. Variable:	Total_Conversion	R-squared:	0.523
Model:	OLS	Adj. R-squared:	0.520
Method:	Least Squares	F-statistic:	162.3
Date:	Wed, 08 Oct 2025	Prob (F-statistic):	5.32e-139
Time:	04:39:46	Log-Likelihood:	-1943.4
No. Observations:	894	AIC:	3901.
Df Residuals:	887	BIC:	3934.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4838	0.051	9.421	0.000	0.383	0.585
Spent	0.0706	0.010	7.038	0.000	0.051	0.090
Clicks	-0.0570	0.015	-3.719	0.000	-0.087	-0.027
age_30-34	1.0193	0.111	9.218	0.000	0.802	1.236
age_35-39	0.2376	0.129	1.837	0.067	-0.016	0.492
age_40-44	-0.1469	0.145	-1.015	0.310	-0.431	0.137
age_45-49	-0.6262	0.135	-4.640	0.000	-0.891	-0.361
gender_F	0.2361	0.085	2.788	0.005	0.070	0.402
gender_M	0.2477	0.080	3.091	0.002	0.090	0.405

Omnibus:	531.853	Durbin-Watson:	1.937
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9374.142
Skew:	2.353	Prob(JB):	0.00
Kurtosis:	18.150	Cond. No.	1.45e+18

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.92e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

C. Model Evaluation

```
mse: 3.1371935166547464
rmse: 1.7712124425530513
```

1. **Compare to the mean:**

- The mean of Clicks is 28.77. The RMSE of 1.77 is about 6.2% of the mean
- The mean of Spent is 44.15. The RMSE of 1.77 is about 4% of the mean

→ This indicates that the average error in the model's predictions is about 6.2% and 4% of the average value in the dataset.

2. **Compare to the standard deviation**

- The standard deviation of Clicks is 46.4. The RMSE is about 3.8% of the standard deviation.
- The standard deviation of Spent is 70.2. The RMSE is about 2.5% of the standard deviation.

→ This suggests that the RMSE is relatively small compared to the variability of the data, indicating good model performance.

V. Recommendations

1. Retargeting the Target Audience

- For niche marketing campaigns (like campaign_a), it is suggested for the company to focus on older demographics and apply personalized messaging to strengthen engagement within this segment.
- For mass-targeting campaigns (like campaign_c), prioritize the 30–34 age group across all genders, identified as the core conversion driver.

2. Organic and Cost-Efficient Strategy

- Leverage user-generated content (UGC) and community engagement initiatives to promote sustained organic growth.
- Enhance SEO performance and social media optimization to boost visibility without increasing ad spend.

3. Trend-Aligned Content Development

- Continuously monitor and integrate emerging trends into campaign content.
- Develop more materials aligned with top-performing interests (e.g., Interest #16, #10, and #29) to maximize content relevance and audience engagement.

4. Model Enhancement and Accuracy Improvement

- Expand the dataset with additional behavioral or contextual variables (e.g., device type, ad placement, or engagement time) to improve predictive precision.