Data Analysis of Marketing Campaigns

Roberto Torres

Udacity Institute of AI/Woolf

Project 1 – Master in AI Capstone

Udacity

Feb, 2026

Overview

This project implements a reproducible exploratory data analysis workflow using a marketing campaign performance dataset sourced from Kaggle. The objective was to explore relationships between marketing spend, engagement, and conversion outcomes

Dataset Description

The dataset contains several hundred rows of campaign-level performance data, including variables such as spend, impressions, clicks, conversions, and ROI. Both numeric and categorical variables enable meaningful exploratory analysis and visualization.

Workflow Description

The workflow consisted of data ingestion, cleaning, exploratory analysis, visualization, and interpretation. Functions were used to standardize column names, handle missing data, and compute summary statistics to support reproducibility.

This work makes use of coding to draw conclusions from a specific marketing dataset as exemplified by Danchev (2022).

Key Decisions and Assumptions

Median imputation was used for missing numeric values to reduce the influence of outliers, a common best practice in exploratory analysis (McKinney, 2010). Visualizations were selected to highlight relationships between spend and performance metrics.

Normalization was used to force another perspective from raw data as it seemed not to offer a significant insight at first.

This analysis assumes consistent metric definitions across campaigns. The dataset lacks contextual and temporal granularity, limiting causal interpretation. Normalization and aggregation choices may also reduce the visibility of extreme observations.

Results and Interpretation

The results indicate diminishing returns in some high-spend campaigns and substantial differences in ROI across campaign types. These findings suggest that reallocating budget toward more efficient campaigns could improve overall performance.

Additional exploratory analysis was conducted using a normalized time-series line chart and a normalized heatmap to account for differing metric scales (Figures 3, 4 and 5 in data workflow notebook). These visualizations enable clearer comparative interpretation across time and campaign segments.

The normalized time-series visualization (Figure 4) indicates that acquisition cost remains relatively stable, while conversion rate and ROI show greater variability. This suggests that efficiency metrics are more sensitive to campaign dynamics than to absolute spending levels.

The normalized heatmap (Figure 5) highlights relative performance differences across campaign types. Normalization reveals subtle but consistent variations in ROI and conversion efficiency that are not apparent in raw average comparisons.

The extended EDA demonstrates the importance of normalization in marketing analytics. These findings provide a strong foundation for subsequent statistical analysis and predictive modeling.

Responsible Practice (Bias and Data Quality)

Potential bias could arise from uneven representation of campaign types or channels, though current dataset maintains a reasonable even balance mitigating that possibility. Future work could incorporate temporal data and additional contextual variables to extinguish these possible limitations.

Reproducibility

Reproducibility is supported through version control, a requirements.txt file, and a structured notebook workflow. All steps can be rerun using the provided instructions and dependencies. Clone master branch in GitHub source at: https://github.com/bobcctorres/ai-programming-foundations-project.

References

Danchev, Valentin. (2022). Reproducible Data Science with Python: An Open Learning Resource. Journal of Open Source Education. 5. 156. 10.21105/jose.00156.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a.