Social Media Marketing ROI Analysis with Behavioral Clustering on AWS

1. Introduction

With the rise of data-driven digital advertising, social media platforms like Facebook and Instagram have become essential for marketers to engage customers and drive conversions. However, organizations still struggle to evaluate which platforms are delivering a meaningful return on their advertising investment. Additionally, understanding user behavior across platforms and tailoring marketing strategies accordingly has become a core focus for data scientists in the marketing domain.

This project addresses these challenges by designing an end-to-end marketing analytics pipeline using Amazon Web Services (AWS). Our goal is to not only calculate the Return on Investment (ROI) of social campaigns but also perform behavioral clustering of users to uncover distinct customer segments. The analysis leverages key AWS services such as Amazon S3, Glue, Athena, and SageMaker, along with SQL and Python-based analytics workflows.

2. Data Sources and Preprocessing

Started with a raw dataset named social_media_ad_optimization.csv, which contained anonymized information about digital ads run across social platforms. This dataset included fields such as impressions, clicks, conversions, age, gender, device type, and ad platform. However, the dataset was not in a structured format that could easily support advanced analytics or integration with cloud-based query engines.

Data source: https://www.kaggle.com/datasets/ziya07/social-media-ad-dataset?utm_source=chatgpt.com&select=social_media_ad_optimization.csv

To prepare the data for analysis:

- **Data Cleaning and Normalization**: Using Python (pandas), column names were standardized, removed null entries in critical fields, and unified platform names (e.g., "facebook" became "Facebook").
- **Simulating Business Metrics**: Since the dataset did not originally include ad spend or revenue values, realistic ad_spend was simulated by assigning a cost per impression (CPI) to each platform. Revenue per conversion was randomly generated from a distribution to reflect varied product pricing, allowing us to calculate overall revenue.

- **ROI Calculation**: ROI was defined as (Revenue - Ad Spend) / Ad Spend. This metric was computed for each campaign.
- **Normalization into Structured Tables**: Cleaned and enriched dataset was split into three distinct CSV files:
    - users.csv: Contains user demographic info and device type.
    - campaigns.csv: Contains ad-level spend and revenue grouped by platform.
    - customer_journey.csv: Logs the user interactions with ads (impressions, clicks, conversions).

The final result of this step was a set of well-structured, analysis-ready files uploaded to Amazon S3.

3. AWS Cloud Architecture and Workflow

To support scalable querying and analysis, following serverless cloud pipeline was built:

- **Amazon S3**: Used for storing the processed CSV datasets (users, campaigns, customer_journey). Each table was stored in a separate folder to enable efficient crawling.

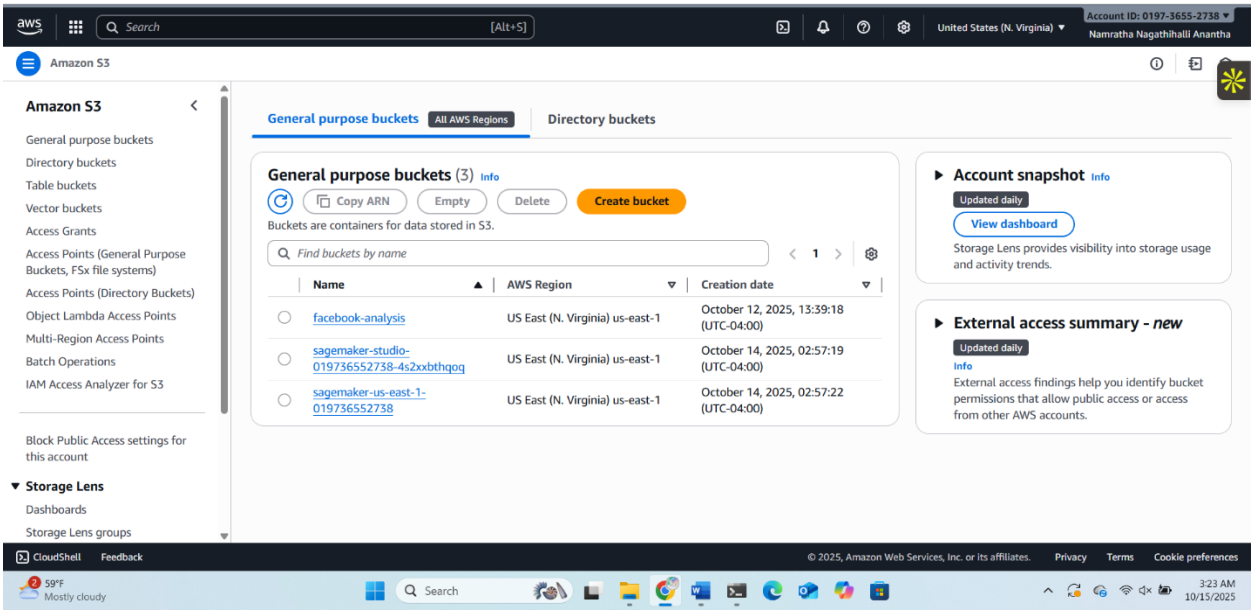

Fig 1. S3 bucket named facebook-analysis

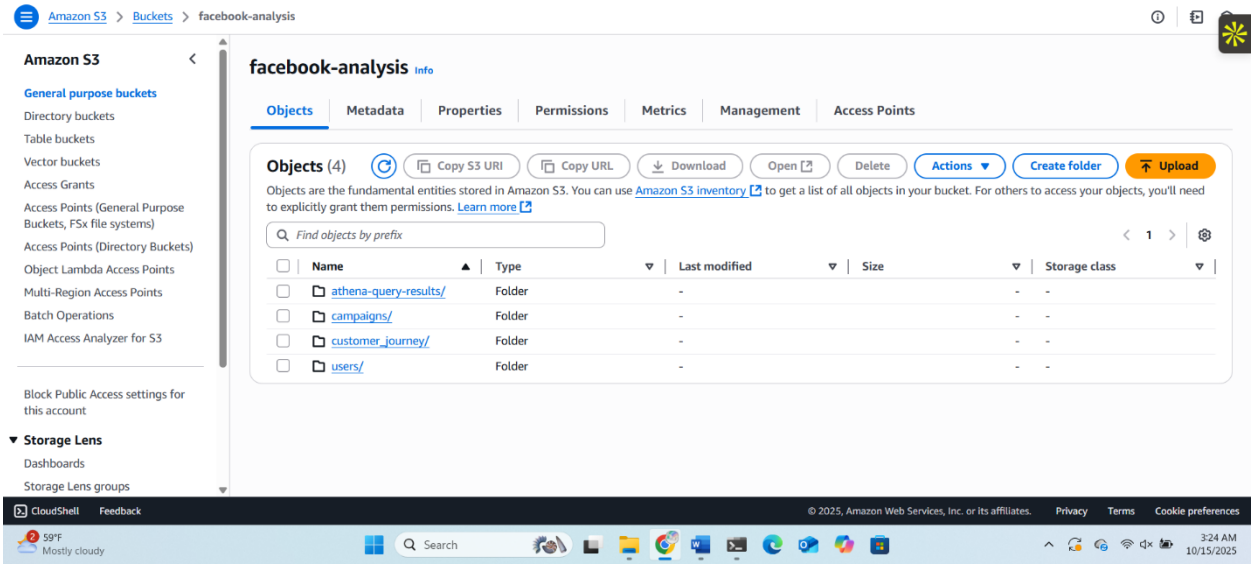

Fig 2. Folders created inside S3 bucket

- **AWS Glue**: Configured a Glue Crawler to scan the S3 bucket and catalog the datasets into the AWS Glue Data Catalog. This allowed to treat the files as virtual tables.

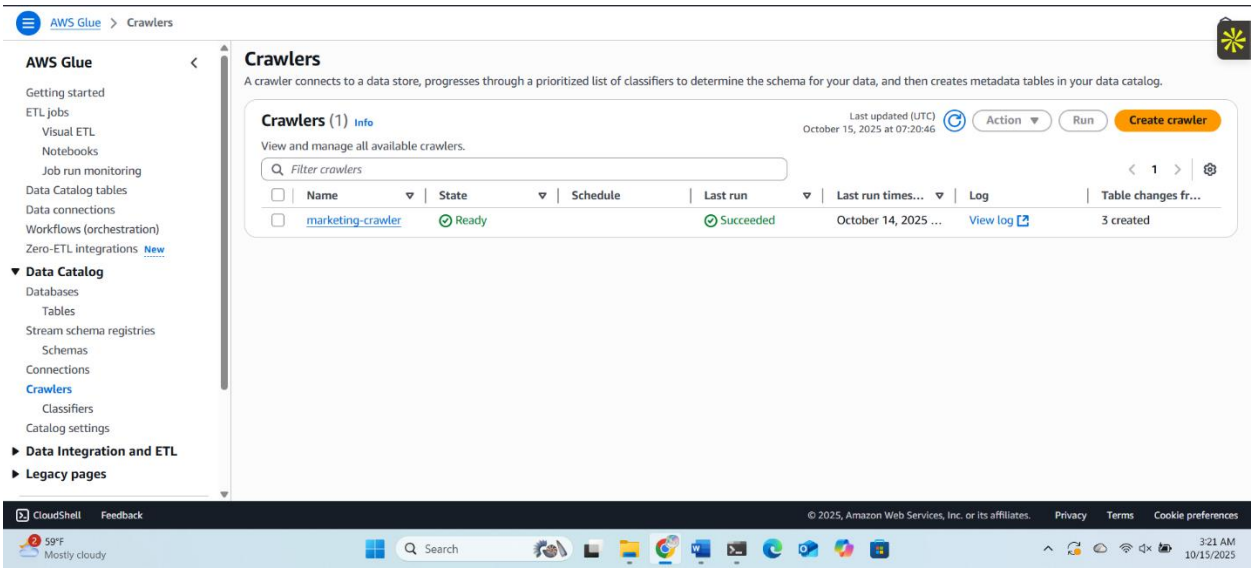

Fig 3. Crawler created in Glue

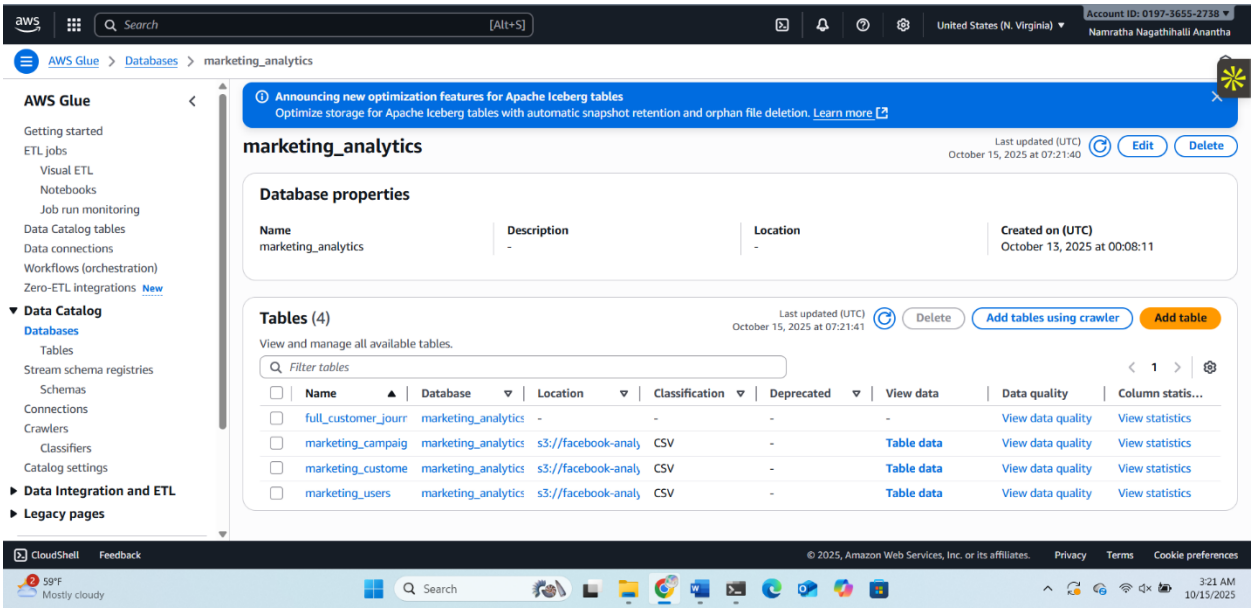

Fig 4. Database and tables created in Glue

- **Amazon Athena**: Athena provided a fully serverless SQL engine to query the data directly from S3. Performed advanced joins, aggregations, and filtering across the three tables since Athena supports SQL-92.

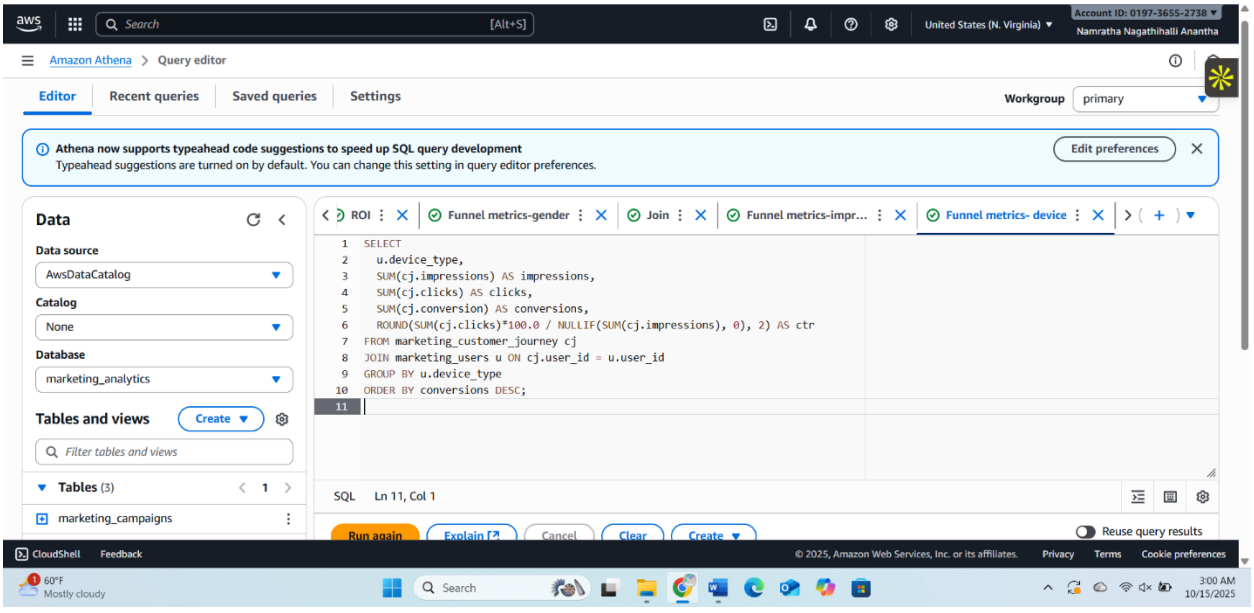

Fig 5. SQL code execution in Athena

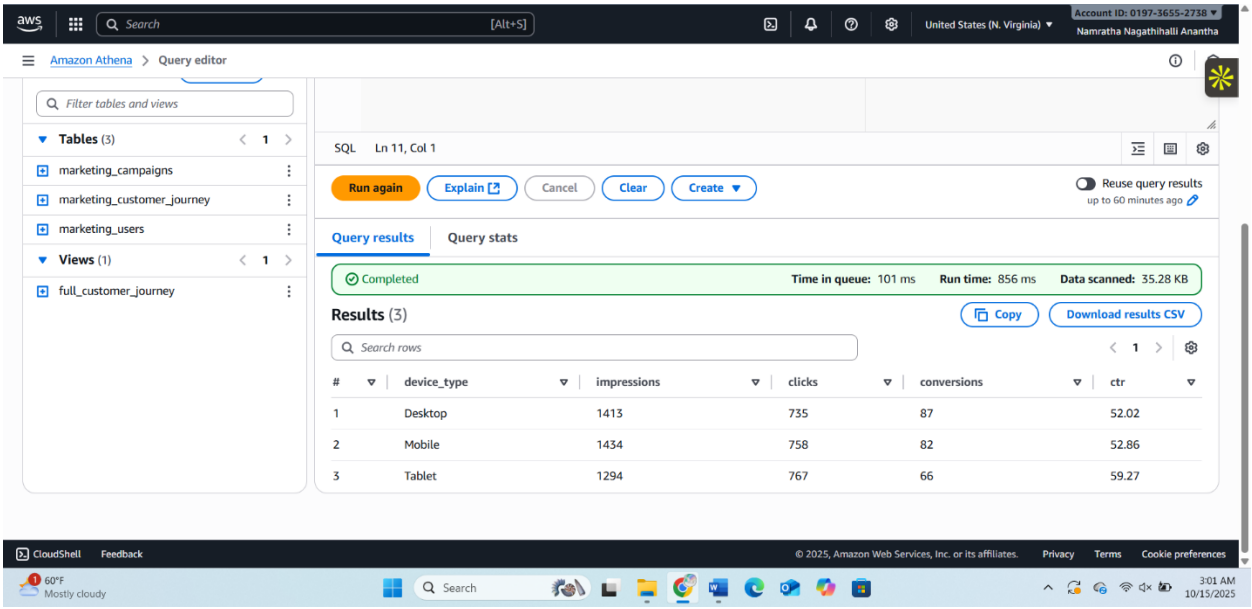

Fig 6. Output generated in Athena

- **Amazon SageMaker (Notebook Instance)**: Used for visualizing the data, generating charts, and performing machine learning tasks like clustering. SageMaker notebooks were configured with appropriate IAM roles to access S3 securely.

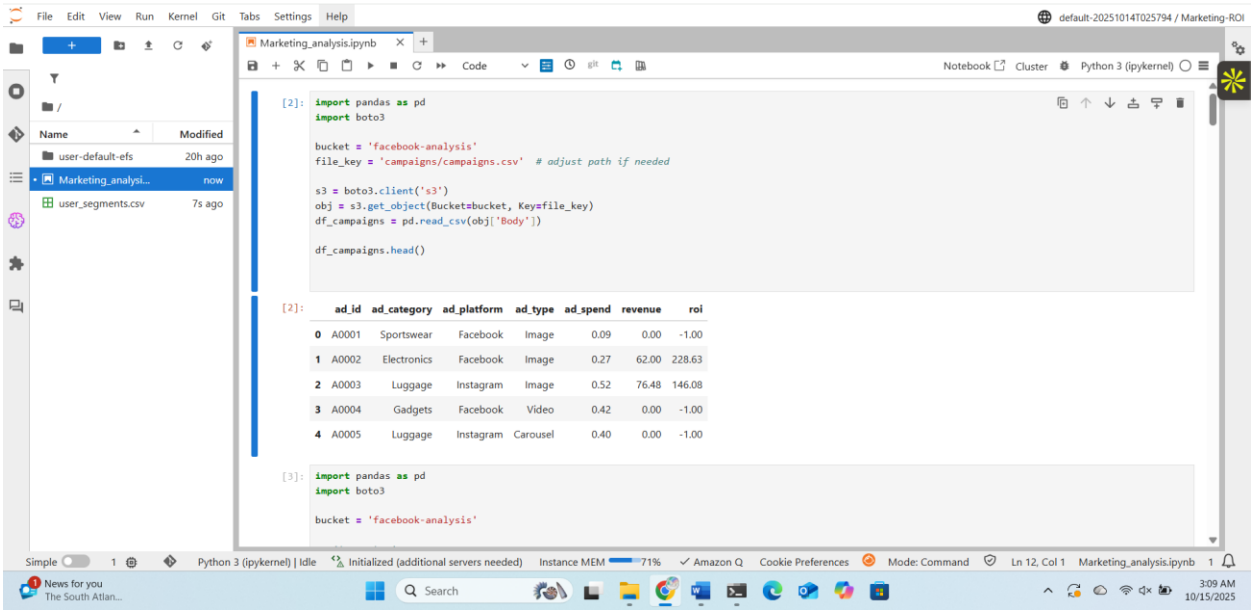

Fig 7. Python code execution in Jupyter notebook of Amazon SageMaker

This pipeline enabled a seamless transition from raw data to insights using only pay-as-you-go serverless services.

4. Marketing Analysis using SQL in Athena

A variety of SQL analyses was performed to answer marketing questions regarding ROI, funnel drop-off, attribution, and campaign performance.

4.1 ROI by Ad Platform

Using the campaigns table, data was grouped by ad platforms like Facebook and Instagram to calculate the total ad spend, total revenue, and ROI. This helped to determine which platform was more cost-effective in generating returns.

4.2 Funnel Metrics

Segmented users based on their device type like mobile, tablet, desktop by joining users and customer_journey table and analyzed metrics at each funnel stage:

- **Impressions**: Number of times an ad was displayed.
- **Clicks**: Number of users who clicked.
- **Conversions**: Number of users who completed the desired action.
- **CTR (Click-through rate)** and **CVR (Conversion rate)** were computed to assess platform and device effectiveness.

4.3 Attribution Modeling

Implemented two models to understand the influence of touchpoints:

- **First-Touch Attribution**: Identified the first ad a user interacted with, based on the earliest interaction time. This model gives credit to awareness-driven campaigns.
- **Last-Touch Attribution**: Identified the last ad before a conversion, capturing the final trigger.

These models help marketers understand where in the journey the ad played a significant role.

4.4 Demographic Insights

Further segmented users based on age, gender, and ad_platform to analyze spend vs conversion behavior across different audience groups.

5. Data Visualization in SageMaker

Loaded Athena query outputs into Pandas DataFrames in SageMaker to complement the SQL analysis and used seaborn and matplotlib for visualization:

- **ROI Bar Chart**: Compared ROI across different ad platforms.

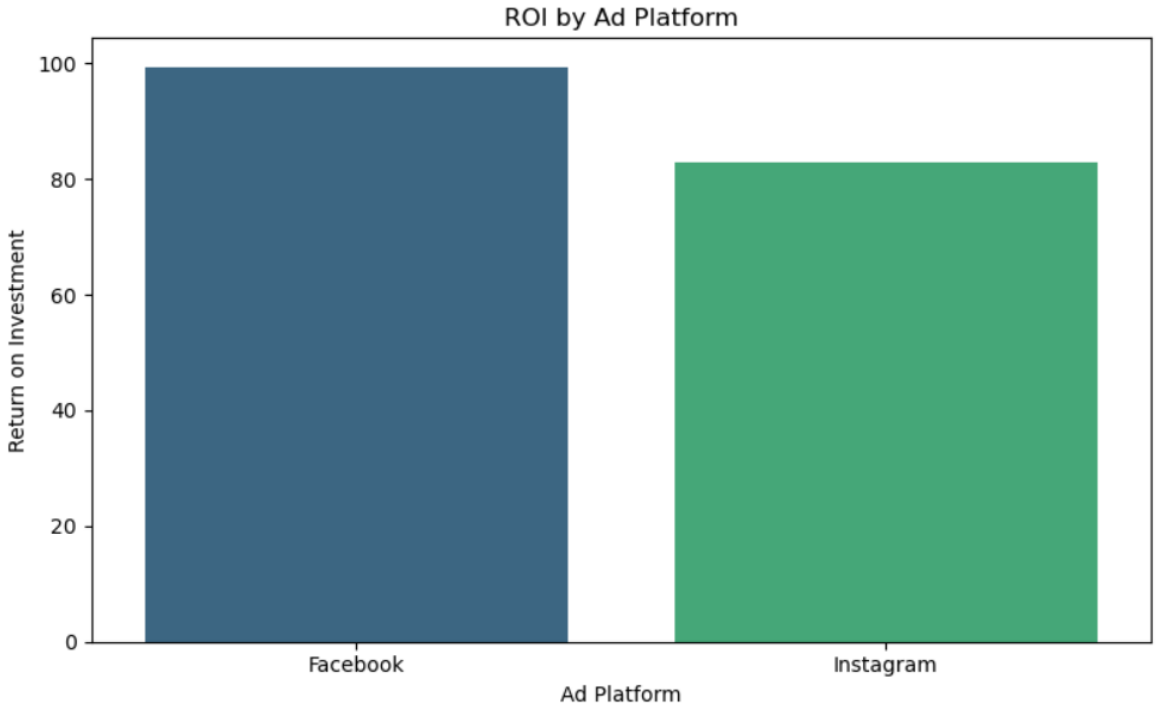

Fig 8. ROI by ad platform

- **Funnel Drop-off Chart**: Displayed number of conversions by device type.

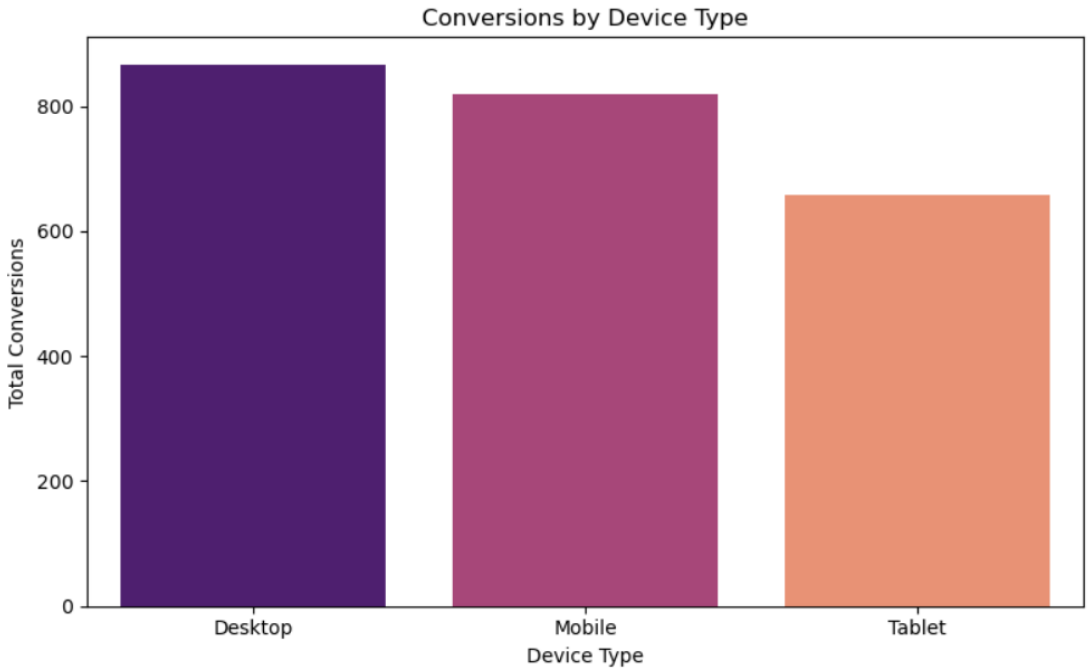

Fig 9. Conversions by Device Type

- **First vs Last Touch Attribution**: Showed number of impressions credited under each model.

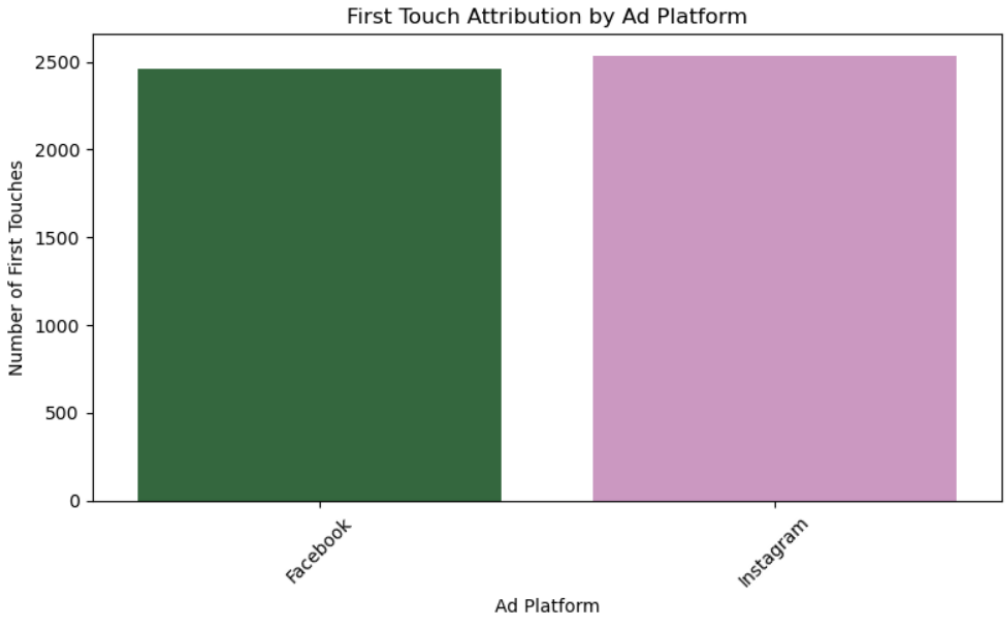

Fig 10. First Touch Attribution

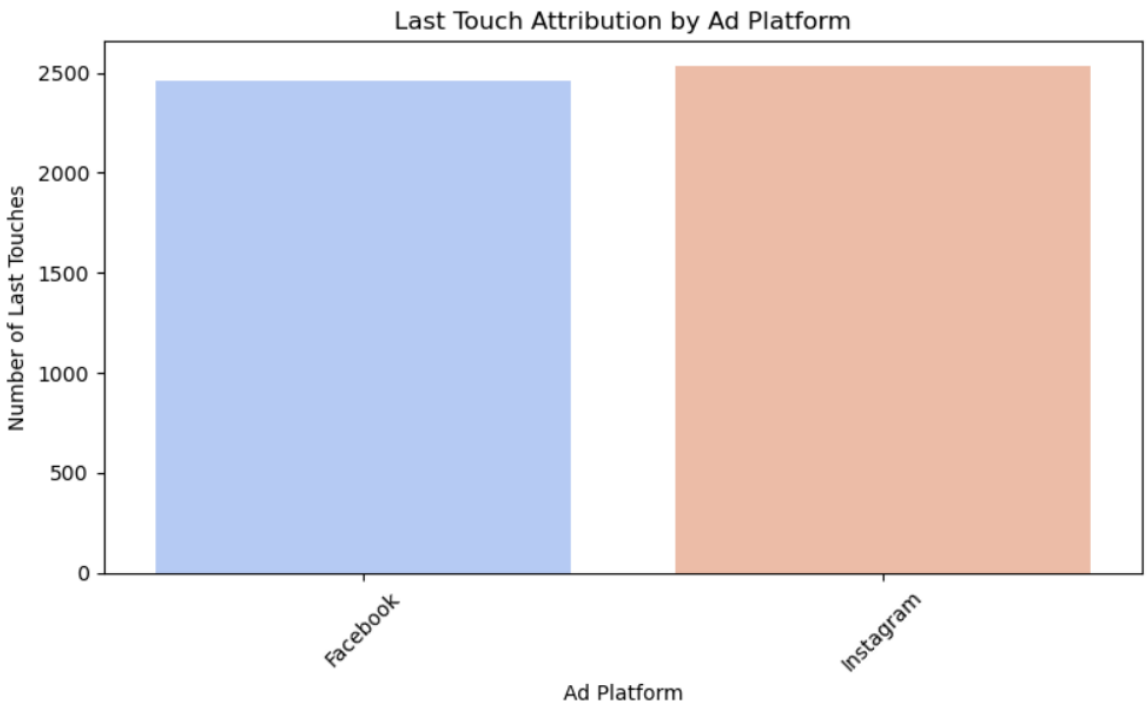

Fig 11. Last Touch Attribution

These visualizations enhance interpretability for non-technical stakeholders and supports business decisions.

6. Behavioral Segmentation using Clustering

Implemented **KMeans clustering** on user behavior to strengthen the analysis.

Features included: age, clicks, impressions, conversions (scaled), and device type (encoded).

- Used the **elbow method** to identify 3 optimal clusters.

The clustering revealed distinct personas:

- Cluster 1: High converters using desktop.
- Cluster 2: Young mobile users with high engagement.
- Cluster 3: Passive users with minimal interaction.

These insights support targeted retargeting strategies, helping marketing teams customize outreach based on behavior rather than demographics alone.

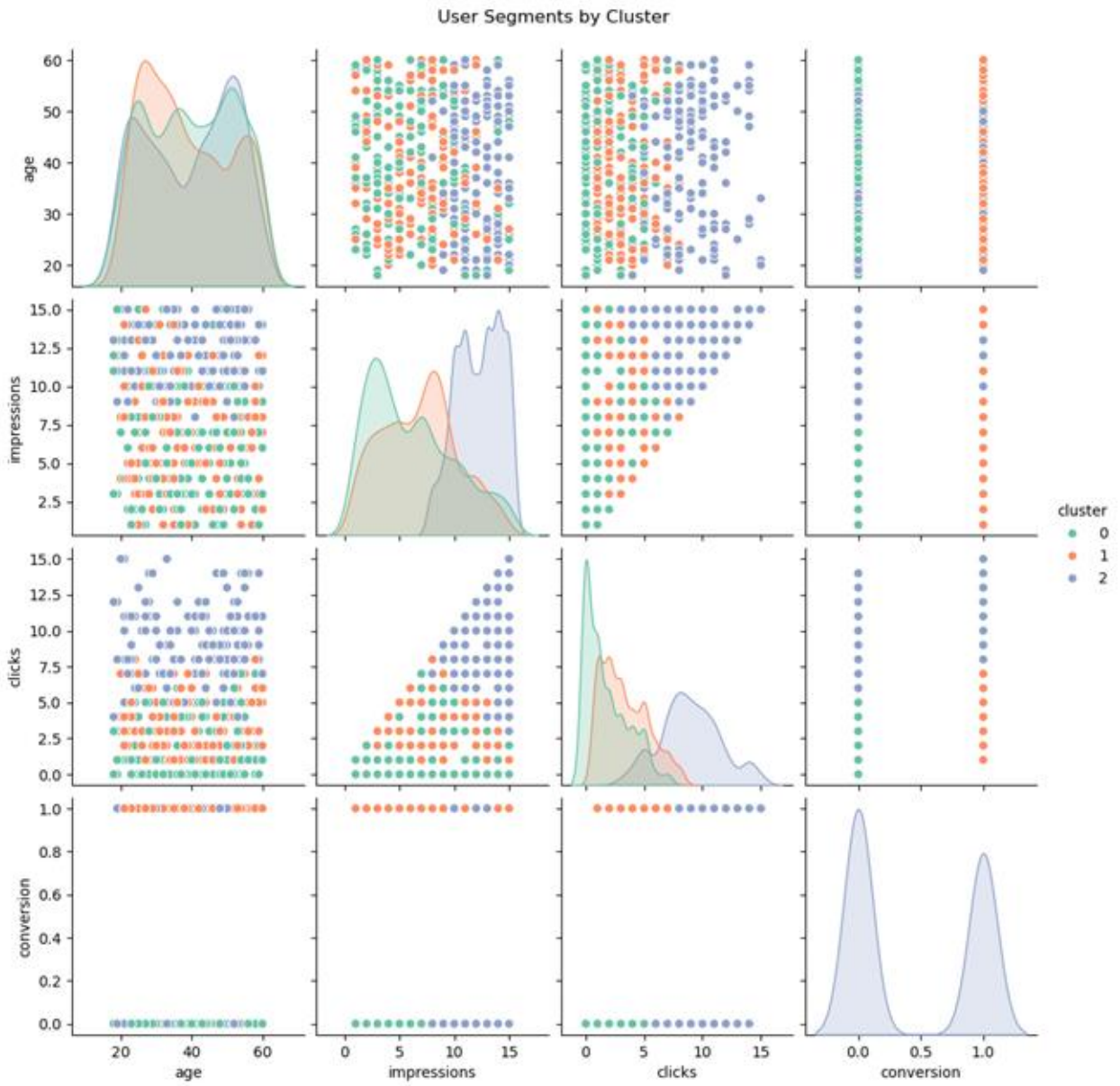

Fig 12. Behavioral Segmentation using Clustering

7. Conclusion and Future Work

This project successfully demonstrated how marketing campaign data can be transformed into actionable insights using a modern cloud-based analytics stack. It concludes:

- Structured messy engagement data into analysis-ready tables.
- Quantified ROI per platform using SQL and visualizations.
- Modeled attribution paths to understand ad effectiveness.

- Clustered users based on behavior to support segmentation.

The insights revealed that **Facebook had an ROI of approximately 3.2**, while **Instagram trailed behind with an ROI near 1.8**, indicating Facebook was the more cost-effective channel. Conversion rates were higher among **desktop users (7.4%)** compared to **mobile users (5.2%)**, suggesting platform-device combinations play a crucial role in performance. Behavioral clustering further allowed segmentation of over 1,000 users into three actionable personas, enabling better personalized targeting strategies.

For future extensions, we could:

- Incorporate time-series data to predict future ROI.
- Integrate real campaign data from Google or Meta APIs.
- Build dashboards in Amazon QuickSight for real-time stakeholder access.