UMBC

Data 606 - EDA

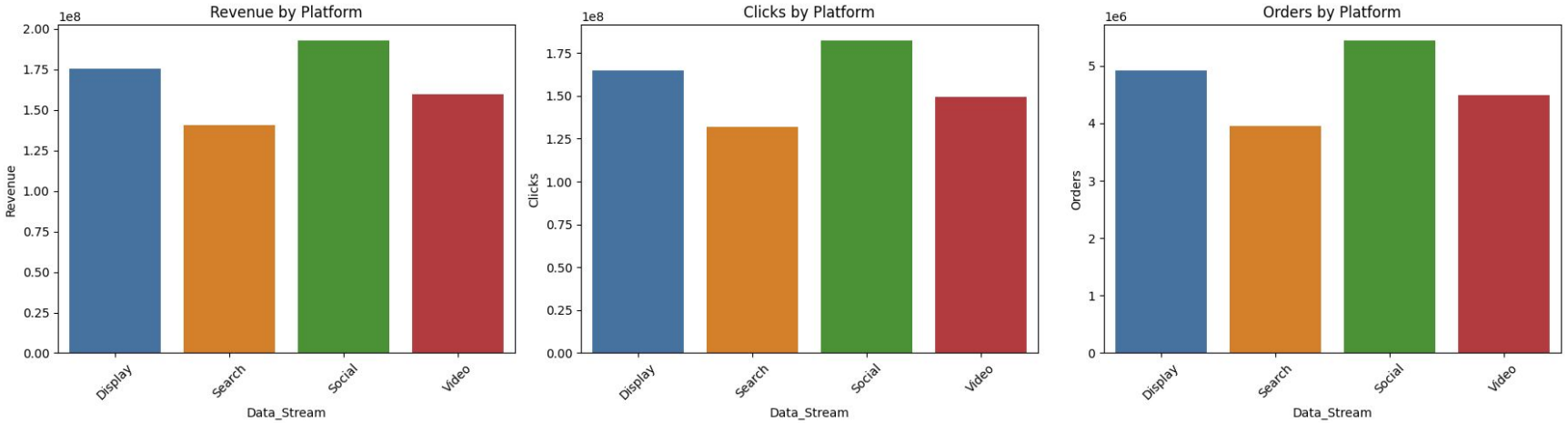
Team Members:
Chandra Sekhar Katipalli
Sindura Reddy Challa
Sanjana Reddy Soma

Project 1- Real-Time Automated Data Pipeline for Advertising

Clustering Analysis for Campaign Performance

- Clustering helps group campaigns with similar performance.
- We used K-Means to segment campaigns based on **Revenue, Clicks, and Orders**.
- **Elbow Method** was used to find the optimal number of clusters.
- Each campaign is assigned to a cluster based on similar characteristics.

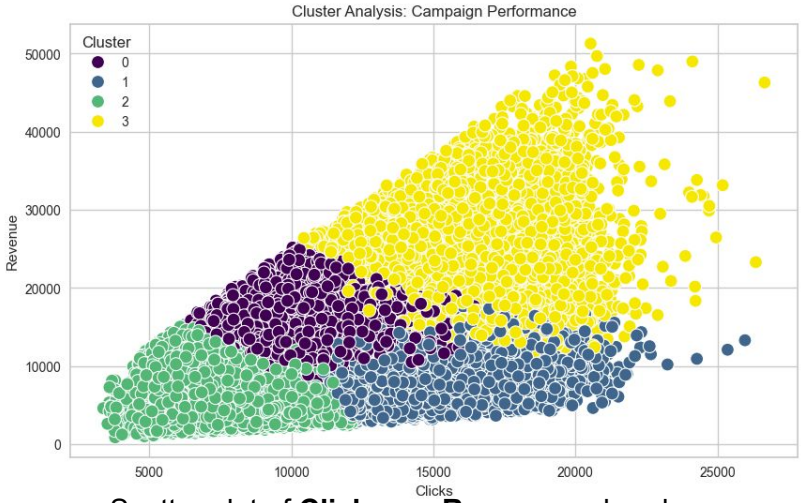
Understanding Platform Performance Before Clustering



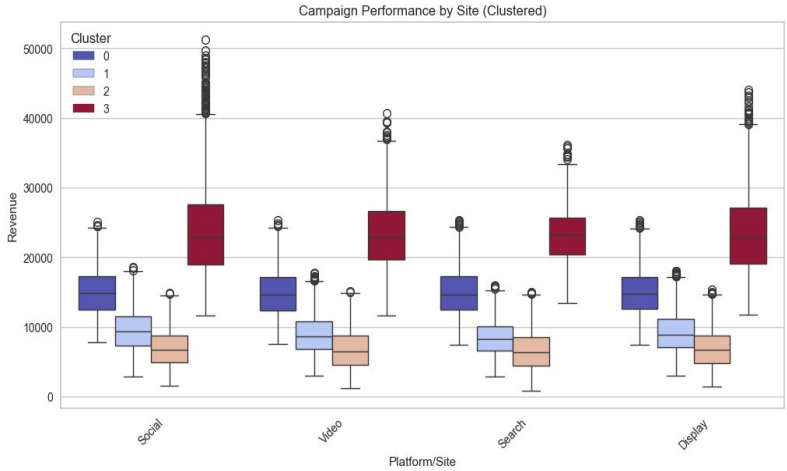
Real-Time Automated Data Pipeline for Advertising

Cluster Analysis Results & Graphs

Cluster 0: "Moderate Performance Campaigns" (Balanced revenue & clicks, moderate orders)
Cluster 1: "High Clicks, Low Conversion" (Many clicks but low revenue & orders → Inefficient campaigns)
Cluster 2: "Low Engagement Campaigns" (Low revenue, clicks, and orders → Underperforming)
Cluster 3: "High Performing Campaigns" (Highest revenue, clicks, and orders → Best campaigns)



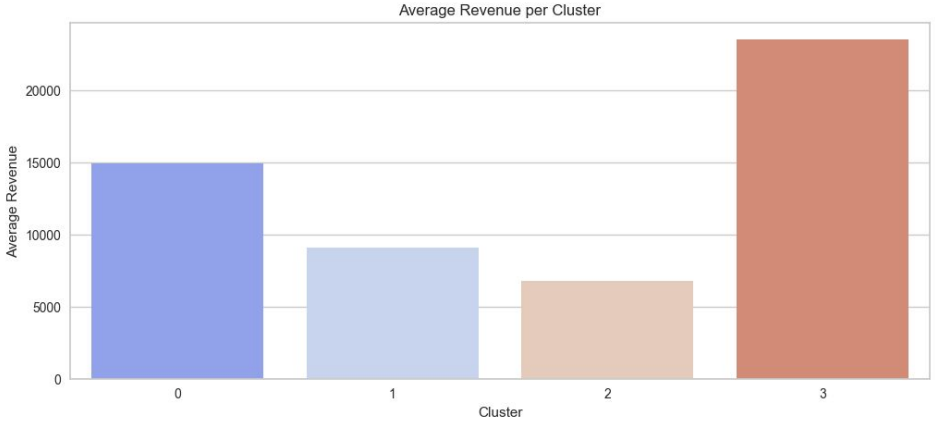
Scatter plot of **Clicks vs. Revenue** colored by clusters



Campaign Revenue by Platform & Cluster

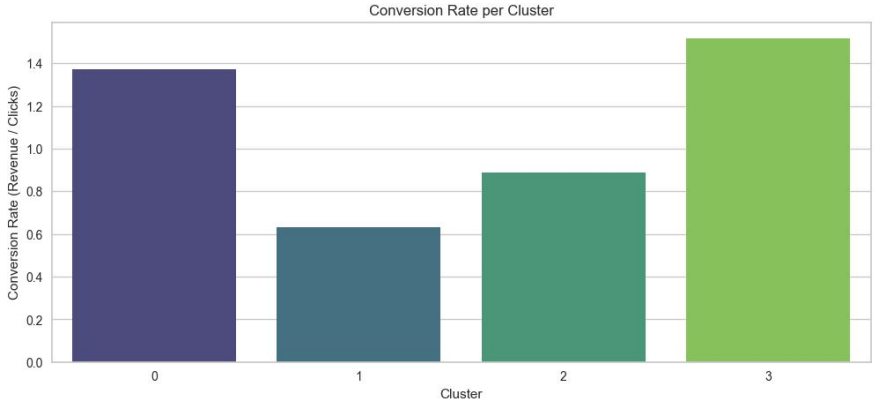
Real-Time Automated Data Pipeline for Advertising

Revenue per Cluster



A bar chart comparing the average revenue per cluster.

Conversion rate across Cluster



A bar chart comparing conversion rates across clusters.

- **Cluster 3 is the best performer** in terms of revenue and conversion efficiency.
- **Cluster 1** is **underperforming** → Low revenue, poor conversion.
- **Clusters 0 and 3 are ideal targets** for future campaigns due to their high efficiency.

Real-Time Automated Data Pipeline for Advertising

Prediction of Cluster Performances

📌 **Time Series Analysis**

- This is useful when we have a strong historical trend. This approach looks at past performance patterns and extends them into the future. We use models such as **ARIMA (AutoRegressive Integrated Moving Average) and Exponential Smoothing** to detect seasonality, trends, and fluctuations in campaign performance

📌 **Regression Models (Predicting Revenue/Clicks Per Platform)**

- This helps us understand the relationships between different factors that influence campaign success. Unlike time series analysis, which focuses on historical trends, regression models analyze how different variables—such as **platform type, ad spend, target audience, and ad content**—affect conversion rates.

Project 2- Predictive Model for Highway Deterioration Forecasting

FAF Data Processing

1️⃣ **Filtering by Mode of Transport**

- **Kept only truck-related data** → `dms_mode = 1` (Truck shipments only).

2️⃣ **Filtering by Intrastate Freight Movement**

- Considered **only shipments within the same state**.
- **Selected only rows where `dms_orig` == `dms_dest`**.

3️⃣ **Filtering by CFS Zones (State-wise Selection)**

- **Virginia** → Kept CFS Zones: `342 & 342`
- **Maryland** → Kept CFS Zones: `241 & 241`, `242 & 242`
- **Alabama** → Kept CFS Zones: `011 & 011`, `012 & 012`

4️⃣ **Selecting Relevant Freight Metrics**

- **Kept only essential columns**:
  - `tons_year`, `value_year`, `tmiles_year`, `curval_year`

5️⃣ **Calculating Weighted Averages**

- Used **tons as weights** to compute weighted averages for:
  - **Total Tons**, **Freight Value**, **Ton-Miles**, **Current Value**.

6️⃣ **Combining Data for All Years**

- Merged **filtered datasets from 2013-2018** into one **final dataset per state**.

HPMS Data Processing

1️⃣ Mapping Counties to CFS Zones

- Used **CFS_Area_Shapefile** to get **ANSI_CNTY** codes for CFS zones.
- Mapped CFS Zones to **County Codes in HPMS dataset**.

2️⃣ Extracting County Codes for Each State

- Used **filtered county codes**:
  - Virginia → `Valid ANSI_CNTY`
  - Maryland → `Valid ANSI_CNTY`
  - Alabama → `Valid ANSI_CNTY`

3️⃣ Handling County Code Variations

- Some files had `County_COD`, others had `COUNTY_COD` → **Standardized column names**.

4️⃣ Filtering HPMS Data by County Codes

- Selected **only rows where county codes matched the CFS zones**.

5️⃣ Adding Year Column for Tracking

- Added a **"year" column** to track data source from 2012-2017.

6️⃣ Combining HPMS Data for All Years

- Merged **filtered HPMS data from 2012-2017** into one final dataset per state.

Challenges Faced

⚠️ **GIS & Google Earth Engine (GEE) Issues**

- Initially planned to use **GIS tools** but faced errors.
- Switched to **manual data processing**.

⚠️ **Data Inconsistencies**

- **Missing county codes** in some datasets.

UMBC

THANK YOU