MULTI-SOURCE ANALYTICS PROJECT

REFLECTION DOCUMENT

STUDENT NAME: HEMANT BORANA

COURSE: DATA MODELING AND VISUALIZATION (DMV)

MODULE: 8 - MULTI-SOURCE ANALYTICS WITH R INTEGRATION

DATE: DECEMBER 2025

Contents

Project Overview

This project involved building a comprehensive multi-channel marketing analytics system by integrating data from four different source types: CSV, JSON, XML, and text logs. The goal was to analyze digital marketing performance across channels and provide actionable business insights through advanced analytics and visualization.

Objectives Achieved

- Successfully integrated heterogeneous data sources into a unified analytical framework
- Implemented a robust ETL pipeline with data quality controls
- Performed multi-channel attribution analysis using four different models
- Created 8 professional visualizations (5 static, 3 interactive)
- Generated an automated R Markdown report with business recommendations

Technical Approach

Data Generation

Created realistic marketing data simulating a real e-commerce company's digital campaigns. The data included:

- **10 campaigns** across 8 channels (Email, Google Ads, Facebook, Instagram, etc.)
- **7 social media posts** with engagement metrics
- **50 web analytics sessions** with device and traffic source data
- **77 customer journey touchpoints** from 15 users

All data sources were interconnected with realistic relationships and business logic (e.g., campaigns appearing in multiple sources, realistic conversion rates, appropriate CTR by channel).

ETL Pipeline Development

The ETL process handled several data quality challenges:

Extraction:

- Used appropriate R packages for each format (base R for CSV, jsonlite for JSON, XML package for XML)
- Parsed unstructured text logs using string manipulation

Transformation:

- Fixed timestamp errors in customer journey logs (years were incorrect)
- Standardized date formats across all sources
- Handled missing values (converted "none" to NA)
- Created derived metrics (cost per conversion, traffic type classification)
- Calculated engagement totals and normalized metrics

Loading:

- Created 6 integrated datasets using left joins to preserve all records
- Aggregated data by campaign for unified analysis
- Generated user-level journey summaries

Analytics Implementation

Multi-Channel Attribution: Implemented four attribution models to understand campaign value:

1. **First-Touch Attribution** - Credits the initial campaign touchpoint
2. **Last-Touch Attribution** - Credits the final campaign before conversion

3. **Linear Attribution** - Equal credit across all touchpoints
4. **Time Decay Attribution** - Recent touchpoints weighted higher using exponential decay

Key finding: Linear attribution revealed that CMP001 and CMP008 contribute to conversions even though they're not first or last touch, demonstrating the importance of multi-touch analysis.

Customer Journey Analysis:

- Mapped conversion funnels showing 25% drop-off from impressions to clicks
- Found converters had 40% more touchpoints (6.67 vs 4.75) than non-converters
- Identified that 46.67% of users abandoned carts with $777.94 in potential revenue

Statistical Modeling:

- Correlation analysis showed engagement rate and conversions as strongest revenue predictors
- Channel performance analysis revealed Email (22.4x ROAS) and Google Ads (7.67x ROAS) as top performers
- Campaign efficiency matrix identified 4 overperforming and 4 underperforming campaigns

Visualization Suite

Created 8 professional visualizations addressing different analytical needs:

Static Visualizations (ggplot2):

1. Channel Performance Comparison - Faceted bar charts showing ROAS, CTR, and conversion rates
2. Attribution Model Comparison - Stacked bars comparing revenue attribution
3. Conversion Funnel - Bar chart with percentages showing customer journey drop-offs
4. Customer Journey Timelines - Timeline plot tracking converters' touchpoint sequences
5. Revenue vs Budget by Channel - Dual-axis combo chart with bars and lines

Interactive Visualizations (plotly): 6. ROAS vs Budget Scatter - Interactive bubble chart with hover details 7. Campaign Efficiency Heatmap - Normalized performance metrics with color scale 8. Campaign Dashboard - Multi-panel view of clicks and conversions

All visualizations used professional styling, appropriate color palettes, and clear labeling for business presentation.

Key Findings and Insights

Channel Performance

- **Email campaigns** delivered the highest ROAS (22.4x) with minimal budget
- **Google Ads** generated the most absolute revenue ($51,207) but required higher investment
- **Display advertising** showed poor performance (1.05x ROAS) and should be reconsidered
- **Instagram** demonstrated strong engagement (9.86% rate) and solid conversions

Customer Behavior

- Average customer journey spans **246 days** with **5.13 touchpoints**
- Converters engage significantly more (**6.67 touchpoints**) than non-converters (4.75)
- **20% conversion rate** from tracked users suggests effective targeting
- Cart abandonment at **46.67%** presents recovery opportunity of $116.69

Attribution Insights

- Different models assign credit differently, highlighting the limitation of single-touch attribution
- CMP001 and CMP008 contribute to the conversion path but don't receive credit in first/last-touch models
- Time decay attribution shows recent touchpoints (last 7 days) drive most conversion value

Campaign Efficiency

- **4 campaigns overperforming**: CMP010 (237%), CMP002 (214%), CMP001 (134%), CMP004 (124%)
- **4 campaigns underperforming**: CMP005 (17%), CMP008 (42%), CMP007 (73%), CMP009 (74%)
- Overall marketing ROAS of **6.38x** indicates profitable campaigns

Business Recommendations

1. Budget Reallocation

Increase investment in:

- Email campaigns (proven highest ROAS at 22.4x)
- Google Shopping (CMP010 showing 237% efficiency)
- Instagram influencer campaigns (strong engagement and conversion)

Reduce or optimize:

- Display network retargeting (only 1.05x ROAS)
- YouTube pre-roll ads (72% efficiency, below target)
- LinkedIn B2B campaign (42% efficiency, needs strategy revision)

Estimated impact: **20-30% improvement** in overall ROAS through reallocation.

2. Cart Abandonment Recovery

Implement automated recovery campaigns targeting the 46.67% abandonment rate:

- Email reminders within 24 hours (personalized product images)
- Retargeting ads on social platforms
- Time-limited discount codes (10-15% off)
- Exit-intent popups on product pages

Expected recovery: **$116.69 in additional revenue** at 15% recovery rate.

3. Multi-Touch Attribution Strategy

Move away from last-click attribution to linear or time-decay models:

- Recognize value of awareness campaigns (CMP001, CMP008)
- Allocate budget based on full customer journey contribution
- Don't cut campaigns that assist conversions even if not last-touch

4. Customer Journey Optimization

Increase touchpoint frequency to match converter behavior:

- Implement nurture email sequences (5-7 touches)
- Cross-channel retargeting campaigns
- Content marketing to create additional engagement opportunities
- Social media remarketing for users who clicked but didn't convert

Technical Challenges and Solutions

Challenge 1: Timestamp Errors in Journey Logs

Problem: Customer journey logs had incorrect years (2091, 2047 instead of 2024) due to data generation logic.

Solution: Created custom parsing function to extract month/day and reconstruct with correct year 2024. Used string manipulation to split timestamp and rebuild properly.

```
date_parts <- strsplit(timestamp_parts[1], "-")[[1]]

correct_date <- paste0("2024-", date_parts[2], "-", date_parts[3])
```

Challenge 2: Missing Values Across Sources

Problem: Not all campaigns appeared in all data sources (e.g., only social campaigns had posts).

Solution: Used left joins to preserve all campaign records and filled missing values appropriately:

- Numeric fields: Set to 0 where no data exists
- Categorical fields: Converted "none" to NA for proper handling

Challenge 3: JSON Structure Complexity

Problem: Initial JSON parsing created nested lists instead of flat dataframe.

Solution: Used simplifyDataFrame = TRUE parameter in fromJSON() to automatically flatten structure into usable dataframe format.

Challenge 4: R Markdown Path Issues

Problem: Image paths failed when knitting report from different directories.

Solution: Set working directory in setup chunk using knitr::opts_knit$set(root.dir = "...") and used relative paths from that root.

Challenge 5: Interactive HTML in Report

Problem: Plotly HTML widgets needed to be embedded in R Markdown output.

Solution: Used htmltools::includeHTML() to embed saved HTML files directly into the rendered report.

Skills Demonstrated

Technical Proficiency

- Multi-format data handling (CSV, JSON, XML, text)
- ETL pipeline development with R
- Advanced data manipulation using dplyr
- Statistical analysis and correlation studies
- Data visualization with ggplot2 and plotly
- Automated reporting with R Markdown

Analytical Thinking

- Multi-channel attribution modeling
- Customer journey mapping
- Conversion funnel analysis
- Campaign efficiency assessment
- Pattern recognition in user behavior

Business Acumen

- Marketing metrics interpretation (ROAS, CTR, conversion rate)
- Budget optimization recommendations
- Customer lifetime value considerations
- Channel mix strategy
- ROI-focused decision making

Communication Skills

- Clear data visualization design
- Executive summary writing
- Technical documentation
- Actionable insight presentation
- Storytelling with data

Learning Outcomes

What Went Well

1. **Data Integration:** Successfully combined heterogeneous sources into unified model
2. **Attribution Modeling:** Implemented multiple models providing different perspectives
3. **Visualizations:** Created publication-quality charts with professional styling
4. **Automation:** Built reproducible pipeline that can run on updated data
5. **Business Focus:** Maintained practical applicability throughout analysis

Areas for Improvement

1. **Real-Time Processing:** Could implement streaming data concepts more deeply
2. **Predictive Modeling:** Could add forecasting for future campaign performance
3. **A/B Testing Framework:** Could incorporate statistical testing for campaign comparison
4. **API Integration:** Could connect to live data sources instead of simulated files
5. **Dashboard Deployment:** Could deploy interactive dashboard using Shiny

Key Takeaways

- **Multi-source integration** is complex but critical for comprehensive analysis
- **Attribution modeling** dramatically changes understanding of campaign value
- **Data quality** issues are inevitable and require robust handling
- **Visualization choice** matters for communicating different insights
- **Business context** must drive technical decisions

Time Management

Actual Time Allocation

- **Hour 1:** Data generation and initial exploration (as planned)
- **Hour 2:** ETL pipeline development with debugging (slightly over)
- **Hour 3:** Advanced analytics and attribution models (as planned)
- **Hour 4:** Visualization creation and R Markdown report (as planned)

Efficiency Notes

- Data generation took less time than expected due to vectorized R operations
- ETL debugging took extra time fixing timestamp issues
- Visualization creation was faster with reusable ggplot themes
- R Markdown report compiled smoothly after fixing paths

Reproducibility

All code is fully reproducible with the following steps:

1. **Setup:** Install required packages (dplyr, ggplot2, plotly, jsonlite, XML, DT)
2. **Generate Data:** Run generate_data.R to create source files
3. **Explore:** Run data_exploration.R to inspect data
4. **Transform:** Run etl_pipeline.R to clean and integrate
5. **Analyze:** Run advanced_analytics.R for attribution and insights
6. **Visualize:** Run visualizations.R to create charts
7. **Report:** Knit marketing_analytics_report.Rmd for final output

All file paths are relative and clearly documented. No manual data manipulation required.

Conclusion

This project successfully demonstrated end-to-end multi-source analytics capabilities, from raw data integration through advanced analysis to automated reporting. The combination of technical execution and business insight creation showcases the practical value of data science in marketing analytics.

The most valuable learning was understanding how different attribution models can lead to different strategic decisions. This reinforces the importance of choosing analytical methods that align with business objectives rather than defaulting to convenient single-touch models.

The automated R Markdown reporting pipeline ensures this analysis can be refreshed with new data, making it a sustainable solution for ongoing marketing performance monitoring.

Final Assessment: Project objectives fully achieved with deliverables exceeding minimum requirements. Ready for production use with minor enhancements for real-time data integration.

THANK YOU

Thank you for reviewing my project.

CONTACT INFORMATION
Email: hemantpb123@gmail.com
Phone: 9284494154

DATE: DECEMBER 2025