**Data Scientist Test**

**Report: Media Mix Modeling**

**Objective**

The primary goal was to build a Media Mix Model (MMM) to understand how media channels, control variables, and regional factors affect weekly sales. Ultimately, the model aims to serve as a statistical foundation for ROI-driven budget allocation decisions across time and geography.

**1. Methodology Overview**

The analysis was conducted using Python 3.12.4, and key libraries for data handling, visualization, and modeling included: pandas, numpy, matplotlib, seaborn, and statsmodels.

**a. Data Preparation**

* The dataset consisted of 6,240 weekly observations across 40 different geographical regions.
* There were no missing values or duplicated rows.
* The “date” column was converted into datetime format, and from it, new temporal features were extracted: year, month, quarter, and week number.

**b. Feature Engineering**

* **Geographical Encoding**  
  The categorical variable “geo” was transformed into binary dummy variables to capture fixed effects across regions. This helped reflect baseline demand variations, regional marketing culture, and infrastructure differences.
* **Temporal Transformation**  
  Seasonal components were modeled using sinusoidal transformations for the “month” variable. This approach captures seasonality as a circular phenomenon rather than a linear trend, thus avoiding the common pitfall of assuming December and January are far apart in time.

**2. Correlation Analysis**

* **Heatmap of Variables**  
  A Pearson correlation matrix was used to explore linear dependencies among numeric variables.
* **Top Correlations with Sales**

Most Positive:

* + Population (correlation: 0.7987) – makes intuitive sense: larger markets naturally have higher sales potential.
  + Search Spend (0.4689)
  + TV Spend (0.4137)

Most Negative:

* + Competitor Sales Control (–0.1167)
  + Sentiment Score Control (–0.0094)
* **Interpretation:**  
  The population variable shows the strongest positive correlation with sales, which is expected due to market size. Competitor activity has a negative effect on our sales, as shown by the correlation of the control variable. The sentiment score has minimal effect, likely due to spatial dilution or weak signal.

**3. Seasonality and Temporal Patterns**

* **Outlier Detection in Sales**  
  Outliers in the sales distribution were visualized via boxplots. These help identify weeks of unusually high or low sales, possibly linked to local promotions or data inconsistencies.
* **Time-Series Decomposition**  
  A seasonal decomposition was performed on the weekly sales of Geo0. The trend component showed a mild decline from mid-2022 to late 2023. The seasonal component revealed strong weekly cycles. Residuals were fairly balanced but revealed occasional sales spikes—these could be tied to promotional campaigns or seasonal events.

**4. Modeling: Structure and Transformations**

* **Adstock and Saturation Transformations**

To model the actual effect of advertising spend over time, two transformations were applied:

* + **Adstock (decay = 0.5):** simulates the carryover effect of advertising by assuming part of last week’s impact continues influencing the current week.
  + **Logarithmic Saturation:** applied a log1p transformation to simulate diminishing returns, a known effect in marketing economics.

Impressions were deliberately excluded as predictors due to high collinearity with spend. In advertising, impressions scale linearly with budget, so their inclusion would add noise and distort the coefficients. Furthermore, spend reflects strategic allocation decisions better than impressions, which are often just by-products.

**5. Model Estimation and Results**

An Ordinary Least Squares (OLS) regression model was trained using:

* + Transformed media spends (after adstock and saturation)
  + Temporal signals (month\_sin and month\_cos)
  + Control variables (population, competitor\_sales\_control, sentiment\_score\_control)
  + Regional dummy variables (geo\_\*)
* **Model Performance:**
  + R-squared: 0.696
  + Adjusted R-squared: 0.694
  + F-statistic: 295.2
  + Observations: 6,240
* **Key Findings:**
  + TV, Print, and Search spend were statistically significant with positive coefficients.
  + Social spend was not statistically significant at a global level.
  + The population variable had a strong and significant positive effect.
  + Competitor sales showed a strong negative effect on our own sales.
  + Sentiment score was statistically significant but had a small negative effect.

**6. ROI and Marginal Contribution**

For each media channel, the following metrics were calculated:

* + Contribution = coefficient × average transformed spend
  + ROI = contribution / average real spend
* **Results:**

TV: ROI = 6.95, Contribution = 45,166  
Print: ROI = 4.66, Contribution = 9,025  
Social: ROI = 3.05, Contribution = 23,329  
Radio: ROI = 2.73, Contribution = 13,701  
Search: ROI = 0.13, Contribution = 1,828

* **Interpretation:**
  + TV delivered the highest ROI, making it the most efficient channel.
  + Search, despite statistical significance, yielded low ROI, possibly due to overinvestment or poor campaign design.
  + Print and Social showed healthy returns relative to spend.

**7. Model Evaluation**

To ensure the reliability and robustness of the regression model, several statistical tests and diagnostic plots were conducted:

* **Residual Analysis**
  + A plot of residuals versus fitted values revealed a mild funnel pattern, suggesting **heteroskedasticity**, i.e., the variance of the errors is not constant.
  + The residuals were mostly centered around zero, indicating no major bias in predictions.
* **Autocorrelation Check (Ljung–Box Test)**
  + Result: p-value = 0.118 (lag=12)
  + Interpretation: No significant autocorrelation detected in residuals. This supports the model's validity over time.
* **Heteroskedasticity Check (Breusch–Pagan Test)**
  + Result: p-value < 0.001
  + Interpretation: Heteroskedasticity is present. This suggests that variance in errors may increase with fitted values. In future modeling stages, robust standard errors or Generalized Least Squares could be applied to address this.
* **ACF and PACF Plots**
  + Partial autocorrelation and autocorrelation plots showed no significant spikes, confirming that residuals behave randomly over time, which supports the temporal stability of the model.

Overall, the model meets the key assumptions of linear regression, with the exception of heteroskedasticity, which is manageable with further enhancements.

**8. Executive Summary – Stakeholder Translation**

Our Media Mix Model offers a data-backed understanding of how marketing investments and external factors influence sales across both time and geography. Here's what it reveals in practical business terms:

**1. The model quantifies the value of each media channel**

It shows how much each channel contributes to sales, based on historical trends.

* + **TV** spending consistently generates the **highest ROI**, both globally and across most regions.
  + **Print** shows **strong regional performance** — particularly effective in specific times and places, such as **Geo9 during Q4 2022**.
  + **Search** demonstrates **limited return**, which suggests possible oversaturation or inefficiency in campaign execution.

**2. The model captures regional and seasonal performance differences**

For every combination of geography, channel, and quarter, the model computes:

* + Estimated sales contribution
  + ROI (sales per peso invested)

This enables **precise, tactical decisions**:

* + "In Geo23, radio works well in Q1, but TV outperforms in Q4."
  + "Print underperforms globally but shines in a few key geographies."

**3. The model helps answer key business questions**

| Business Question | Model Insight |
| --- | --- |
| What’s the most profitable channel overall? | **TV**, with highest ROI and contribution |
| Which region responds best to Print? | **Geo9** and **Geo12**, particularly in late Q4 |
| Is Social worth the investment? | Yes — **moderate ROI**, useful in multi-channel mix |
| Can we reallocate spend more efficiently? | Yes — **Print** and **Radio** deserve more in select regions |

**9. How to validate the model?**

* **Sensitivity and Coefficient Sanity Checks**

We can simulate marginal increases in spend (without actually changing campaigns) and examine whether:

* Marginal ROI decreases (saturation is working)
* Less effective channels show low response
* Population and competitor controls behave as expected

This checks for **coefficient realism** and validates model logic before using it for resource reallocation decisions.

* **Backtesting with Historical Campaigns**

The model can be used to **simulate past periods** with known outcomes (e.g., Q4 2022 campaigns). If the model correctly identifies high-ROI channels and matches known sales spikes (e.g., TV or Print in specific geos), this reinforces its explanatory power and alignment with business intuition.

* **Strategic Monitoring (Without Forecasting)**

Although the model is not predictive, it can still support production decision-making through **interactive scenario dashboards** (e.g., Streamlit or Power BI):

* Allow users to simulate different budget distributions
* Visualize expected sales contribution and ROI by channel
* Compare performance across regions and quarters

This offers prescriptive value, even in the absence of real-time forecasting or automated predictions.

* **Controlled Geo-Time Budget Testing (Last Resort)**

As a final and optional step, a **limited geo-based test** could be designed to reallocate spend using model recommendations (e.g., boost Print in Geo9 during Q4). However, this should only occur **after all prior validation steps confirm the model's reliability**, given the real financial implications of misallocation.

**10. What Would Be Done with More Time**

Given additional time and resources, several areas could be enhanced to improve both model precision and practical utility:

* **Modeling Enhancements**
  + **Bayesian Hierarchical Modeling:**  
    Use a probabilistic framework (e.g., PyMC) to estimate **geo-level variation** with **credible intervals**. This would help quantify uncertainty and tailor strategies by location.
  + **Advanced Saturation Modeling:**  
    Replace the logarithmic transformation with the **Hill function**, which allows more flexibility in modeling **nonlinear saturation dynamics**.
  + **Channel-Specific Adstock Tuning:**  
    Rather than applying a fixed decay factor (0.5) across all channels, perform a **grid search** or optimization procedure to find the optimal **decay per channel**, reflecting different media lifespans (e.g., TV vs. Social).
* **Feature Engineering**

To enrich the dataset and improve predictive accuracy:

* + **Weather Data:**  
    Integrate **temperature or weather anomalies** per geography, which may affect consumer behavior and sales seasonality.
  + **Spatial Features:**  
    Add features like **distance between geographies**, to model **regional influence spillover** or cross-geo exposure effects.
  + **Event Indicators:**  
    Incorporate **campaign flags, holidays, product launches**, or other exogenous events that drive short-term sales spikes or depressions.
* **Production and Monitoring**
  + **Hold-Out and Rolling Forecasting:**  
    Implement **temporal hold-out testing** and **rolling windows** to assess out-of-sample generalizability and detect model drift over time.
  + **Deployment Dashboard:**  
    Build a **dashboard** (using tools like **Streamlit** or **Power BI**) for **marketing and strategy teams** to interact with ROI results and simulate budget allocation.
  + **Modular Pipeline Architecture:**  
    Design a robust end-to-end pipeline, split into the following reusable components:
    - 1. **ETL (Extract, Transform, Load):** Load and preprocess data for modeling.
      2. **Model Training:** Run model with tunable configurations and logging.
      3. **Scenario Simulation & Reporting:** Allow “what-if” planning and generate ROI reports.
* **MLOps Integration for Versioning and Experimentation:**
  + - Use **MLflow** to log:
      * Model parameters (e.g., decay rates, saturation curves)
      * Performance metrics (R², residual tests)
      * ROI outputs per run
    - Use **DVC** (Data Version Control) to:
      * Version control raw data and feature sets
      * Ensure reproducibility of transformations and training pipelines
      * Track experiments and rollback when needed