Aligning YouTube Interests, Demographics, and E-Commerce: Multi-Source Analysis of U.S. Consumer Trends

Yucheng Qin, Kyle Huang, Anchal Anchal, Karuna Kambalapadu Eediga

Executive Summary

This project examines how demographic structures and digital engagement influence U.S. e-commerce consumption, offering insights relevant to emerging markets. The analysis integrates state–month e-commerce transactions, demographic survey data, and YouTube trending engagement in U.S. from 2021 to 2022, merged into panel datasets to capture both structural and dynamic relationships.

Five research questions were addressed. RQ1 found that raw correlations between demographics and sales, such as the negative link with citizenship, disappear once fixed effects are applied. RQ2 showed that states with higher education consistently allocate more spending to technology, though short-term changes in education shares do not drive sales. RQ3 highlighted alignment between demographic improvements and YouTube preferences, particularly in technology and entertainment. RQ4 found little evidence that YouTube trends forecast e-commerce sales. RQ5 revealed stronger education elasticity for technology (0.65) and a negative elasticity for office supplies (–0.30), though not statistically significant.

Overall, three insights emerge: demographic effects are structural and long-run, education is the strongest driver of technology adoption, and YouTube trends reflect culture but lack predictive power for consumption. For stakeholders, this underscores technology as an education-driven growth opportunity, while highlighting the need for better forecasting tools.

Data preparation

Star Schema Design

To facilitate multifaceted analysis of e-commerce transactions, demographic profiles, and YouTube trending data, a dimensional data model was architected utilizing the star schema paradigm. This design was selected for its inherent simplicity and high query performance, which are optimal for business intelligence and online analytical processing (OLAP). The resultant schema (Figure 1) is composed of a central fact table linked to four descriptive dimension tables as follow:

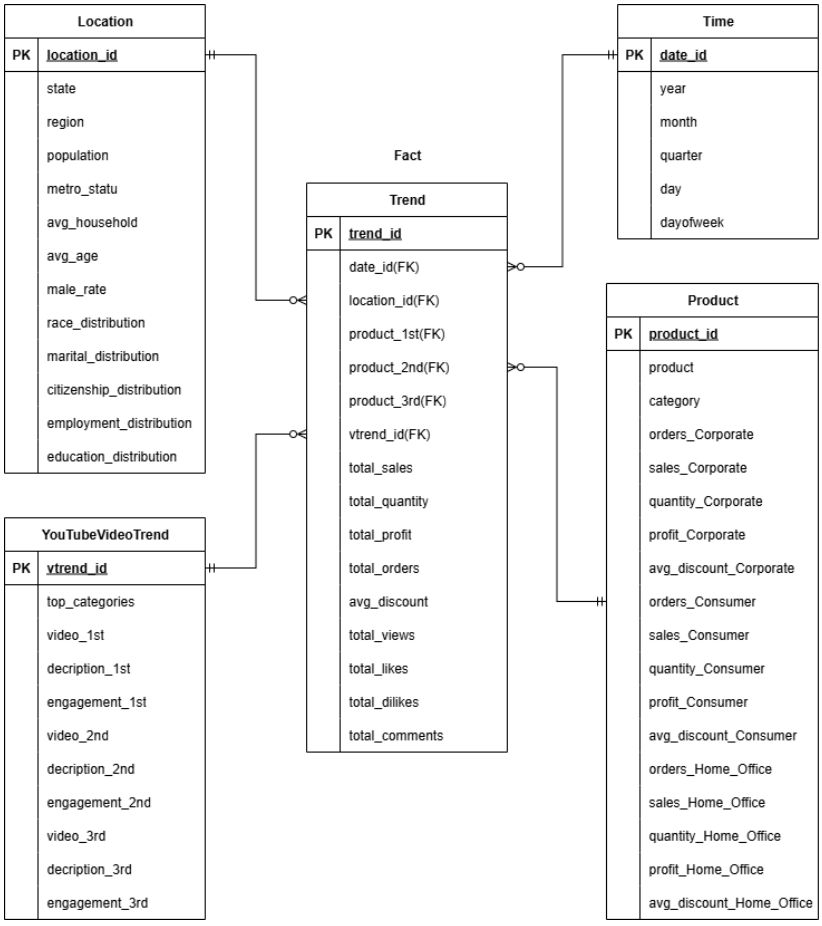

Figure 1: ER diagram of star schema

- **Fact Table**: The centre of the data warehouse is the `Trend` fact table, engineered to capture quantitative performance metrics from the e-commerce and YouTube source systems. The level of this table is set at a daily and state-wide level, meaning each individual row summarizes all measured activities (like total sales and total video views) that occurred in a single state on a single day. Each record represents daily top three products and video trend in YouTube in a perticular US state, assuming that every state has the same video trend as national one. The table is linked to the dimension tables through unique keys representing time, location, video trends, and the top three products.
- **Dimension Tables**:

  1. Time Dimension: This dimension table is derived from the date fields present in the source datasets. It provides the temporal context for the facts, enabling time-series analysis. It incor-

porates a natural chronological hierarchy with levels for Year, Quarter, Month, Day, and Day of Week. This structure permits analysts to examine trends at a daily granularity, aggregate the data to observe monthly, quarterly, or annual patterns, and also analyze trends based on the day of the week.

2. Location Dimension: This dimension provides the geographical context for the facts, containing descriptive demographic attributes for each location. It features a primary geographical hierarchy with levels for Region and State. Furthermore, it supports analytical roll-ups of its demographic distribution attributes, such as the top three categories for race, marital status, citizenship, employment, and education. This allows for the aggregation of detailed state-level demographic profiles to create summarized, region-level insights.
3. Product Dimension: This dimension provides detailed, descriptive information about each unique product. Attributes include the product's name and its assigned category, allowing for analysis of sales performance by individual items or by product groups.
4. YouTubeVideoTrend Dimension: This dimension is derived from the YouTube dataset. It captures daily cultural trends by storing qualitative and quantitative data about the highest-performing videos and categories for each day a video was trending. Due to large amount of videos trending daily, this dimension stores key information for only the top three performing videos to represent the day's trend. The ranking is determined by a calculated engagement score, where $engagement = views + likes - dislikes + comments$.

Data Warehouse Construction

An Extract-Load-Transform (ELT) architecture was implemented exclusively within the SQLite database engine, utilizing a single SQL script (`trend_dw.sql`). This methodology leverages the database's native processing power for all data manipulation, resulting in a self-contained and reproducible workflow.

Extract

The Extract phase consists of accessing the raw data from the three disparate source systems. In this project, the datasets were provided as Comma-Separated Values (CSV) files. The extraction process is therefore the initial read operation of these files from the local file system.

Load

The Load phase involves the ingestion of the raw, untransformed data directly into the SQLite database. The `.import` command-line utility was employed to create and populate three temporary tables: `demography`, `ecommerce`, and `youtube`. This procedure collocates the source data within the database environment, where it can be transformed efficiently using SQL.

```
# loading data through sqlite3 cmd
.import --csv --schema temp data\cps_00001.csv demography
.import --csv --schema temp data\Ecommerce_data.csv ecommerce
.import --csv --schema temp data\US_youtube_trending_data.csv youtube
```

Transform

The Transform phase constitutes the most complex stage of the ELT pipeline. It commences with an initial cleaning of the staged data, followed by a series of SQL transformations that populate the dimension tables first, and finally the central fact table to construct the star schema.

1. Data Cleaning

Before populating the dimensional model, the raw data within the demography, ecommerce, and youtube staging tables is cleaned and standardized. This involves using `UPDATE` statements with `CASE` expressions to transform ID numbers into human-readable text (e.g., converting `STATEFIP` codes to state names, `RACE` codes to race descriptions) and to standardize date formats across tables. This preparatory step ensures data consistency and quality before the main transformation logic is applied. (see Appendix or `trend_dw.sql`)

2. Schema Creation

Following the cleaning process, the logical structure of the data warehouse was built based on star schema. `DROP TABLE IF EXISTS` statement was run before creating tables in case of re-run. Subsequently, `CREATE TABLE` statements define the schema for the central Trend fact table and the four dimension tables, establishing primary keys, foreign keys, and data types, which is shown in Appendix and `trend_dw.sql` script. This creates the empty framework of the star schema, ready for data population.

3. Filling Dimension Tables

With the clean data and defined schema in place, each dimension table is populated through a dedicated `INSERT INTO ... SELECT` statement that transforms and aggregates data from the staging tables.

- `Time` Dimension: This table is populated by creating a complete daily calendar for the period of analysis. Although the `ecommerce` data defines the overall start and end dates for the analysis, it lacks entries for many days within that range. To create a continuous timeline, the dimension is built using the more consistent daily dates from the `youtube` dataset. These dates are filtered to fall strictly within the minimum and maximum order dates from the ecommerce data, ensuring a complete and relevant time dimension. Subsequently, the `strftime` function is used to parse each date into its constituent components.
- `Product` Dimension: This table is populated by grouping the `ecommerce` data by product. The query pivots the data based on the three distinct values in the `customer_segment` column (Corporate, Consumer, Home Office) to create specific columns for sales, quantity, and profit metrics for each segment.
- `Location` Dimension: This table is populated by grouping the cleaned `demography` data by state (`STATEFIP`). During this process, key demographic metrics like population are calculated. For distributional attributes, an analysis of the source data identified the most prevalent categories. Consequently, the transformation logic creates textual summaries by aggregating percentages for these primary groups: race (Black, White, Asian or Pacific Islander), marital status (Married, Single, Divorced), citizenship (Native-born, Naturalized, Non-Citizen), employment status (At work, Army, NILF - Not in Labor Force), and education level (Less than High School, High School, Bachelor or Higher). This method provides a concise yet representative demographic profile for each state.
- `YouTubeVideoTrend` Dimension: Populating this table is a multi-step process. First, a temporary table is created to calculate a composite engagement score for each video and rank them on a daily basis, as well as storing relationship between date and `YouTubeVideoTrend`.

Subsequently, this temporary table is queried, and the data is pivoted to create a single row for each day, containing the top three trending videos and their engagement scores.

4. Populating the Fact Table As the final step in the transformation process, the central `Trend` fact table is populated. This is accomplished SELECT statement that:

- Joins all the previously populated dimension tables with aggregated data from the staging tables.
- Calculates all quantitative measures by applying aggregate functions like `SUM` and `AVG`.
- Groups the results by `date_id` and `location_id` to ensure the data conforms to the defined granularity of the fact table.

Research Analysis

Research Questions

RQ1. Do demographic characteristics drive differences in e-commerce consumption patterns?

E-commerce markets are not uniform: consumer demand often reflects the underlying demographic structure of a region. Factors such as age distribution, education levels, and employment status can significantly influence what categories of products households are willing and able to purchase. For example, younger and more educated populations may allocate more spending toward technology products, while older demographics may prioritize household-related categories such as furniture.

This question is particularly important in the context of fast-growing but under-researched e-commerce markets such as New Zealand. Local stakeholders lack reliable intelligence on which demographic segments are driving demand, making it difficult to design targeted marketing strategies or region-specific resource allocation. By leveraging detailed U.S. data as a benchmark, it's possible to test whether clear demographic–consumption linkages emerge across product categories in technology, furniture and office Supplies.

Answering RQ1 provides two key insights.

The first one is, it highlights whether structural demographic variables explain regional heterogeneity in e-commerce consumption, beyond random fluctuations.

Second, it establishes a baseline for later questions, linking social shifts like education, employment, age, etc. with both consumption and digital engagement signals shown as YouTube trends.

Ultimately, RQ1 is about discovering systematic demographic drivers of online demand. For executives, this means moving from intuition like young people probably buy more tech toward evidence-based strategy, where investments in marketing, product focus, and regional planning are guided by clear demographic patterns.

RQ2. Is there a structural relationship between changes in educational attainment and consumption of high-value categories (Technology)?

Education has long been recognized as a critical driver of household consumption. Higher educational attainment is often associated with greater income potential, digital literacy, and preference for advanced products, particularly technology-related goods. If a strong structural link exists, regions with a more educated population may systematically allocate a larger share of their consumption to technology categories compared with traditional segments such as furniture or office supplies.

This question is especially relevant for fast-growing e-commerce markets where consumer intelligence is scarce. In contexts like New Zealand, reliable market research reports are limited, making it difficult for firms to identify which social shifts, such as increasing tertiary education translate into tangible changes in product demand. Leveraging U.S. data provides an opportunity to test whether education-driven structural patterns in technology adoption can be observed, offering a proxy for markets with fewer available insights. Addressing RQ2 provides two contributions.

First, it tests whether education levels are disproportionately associated with technology consumption compared to other categories, providing evidence for structural demand segmentation. Second, it links demographic progress in education attainment with digital consumption categories, which can guide firms in anticipating future demand growth as education levels rise. For executives, this means validating or rejecting the intuition that "better educated consumers buy more technology products," turning it into a measurable structural insight for market positioning and strategic investment.

RQ3. Are nationwide demographic trends consistent with YouTube video preference trends?

Digital media preferences are increasingly shaped by demographic shifts in society. Younger populations with higher educational attainment and stronger digital literacy are often more engaged in online platforms, particularly in categories such as Entertainment, Gaming, and Technology. At the same time, demographic variables such as employment and gender composition reflect broader social dynamics that can influence the adoption of digital content.

For emerging e-commerce and digital economies, this question holds significant importance. Understanding whether demographic change aligns with shifts in nationwide digital content preferences provides insight into how consumer structures translate into media engagement. In markets such as New Zealand, where detailed digital consumption intelligence is limited, analyzing United States data enables us to explore whether stable structural links exist between demographics and YouTube preferences.

Answering this question contributes to two objectives. First, it examines whether demographic developments, such as rising shares of young and educated populations, correspond to stronger engagement with digital categories like Gaming and Technology. Second, it provides a lens for connecting long-term demographic change with online cultural trends, thus allowing firms to anticipate shifts in consumer attention and digital marketing opportunities.

RQ4. Can YouTube trending videos predict e-commerce consumption trends?

RQ4 asks whether YouTube trending videos can serve as leading indicators for e-commerce consumption trends. The rationale is that digital engagement patterns, such as viewing and liking videos in categories like Technology, Gaming, or Entertainment, may anticipate shifts in consumer demand for related product categories. If such predictive links exist, firms could leverage YouTube metrics as a low-cost, real-time signal of consumer preferences, enhancing demand forecasting and marketing strategy.

RQ5. Does the demographic sensitivity of consumption behavior differ significantly across product categories?

Demographic factors such as education, employment, and age are widely recognized as important drivers of consumption demand. Yet their impact may vary substantially across product categories. Technology products are often linked to education levels and digital literacy, furniture demand is closely tied to household structure, and office supplies are associated with employment conditions. Understanding these differences is critical for firms, as it helps identify which categories are more responsive to demographic change and therefore deserve priority in resource allocation and strategic planning.

SQL for Business Analytics

RQ1. Do demographic characteristics drive differences in e-commerce consumption patterns?

To address Research Question 1, two foundational datasets are constructed. The first dataset, ecom_agg, aggregates e-commerce transactions at the state–month–category level, producing indicators such as total sales, total profit, and total quantity. The second dataset, demo_agg, aggregates demographic survey data at the state–month level, generating weighted measures including average age, share of population with undergraduate education or above, employment and unemployment shares, gender composition, and citizenship share. By merging these datasets, the panel dataset panel_rq1 is formed, aligning category-level consumption outcomes with corresponding demographic structures.

The analysis proceeds in two stages. The first stage applies raw correlation analysis. As shown in Figure 2, Pearson correlation coefficients, derived from the covariance formula, are computed between category-level sales and demographic variables. The results indicate that the citizenship share is consistently and strongly negatively correlated with sales, with values around –0.57 across categories. The unemployment share shows a moderate positive correlation with sales in Technology and Office Supplies, approximately 0.25. Other variables, such as average age and gender, exhibit weaker associations. These findings suggest that demographic structures appear linked to consumption differences, but such relationships may be influenced by persistent state-level heterogeneity or seasonal factors.

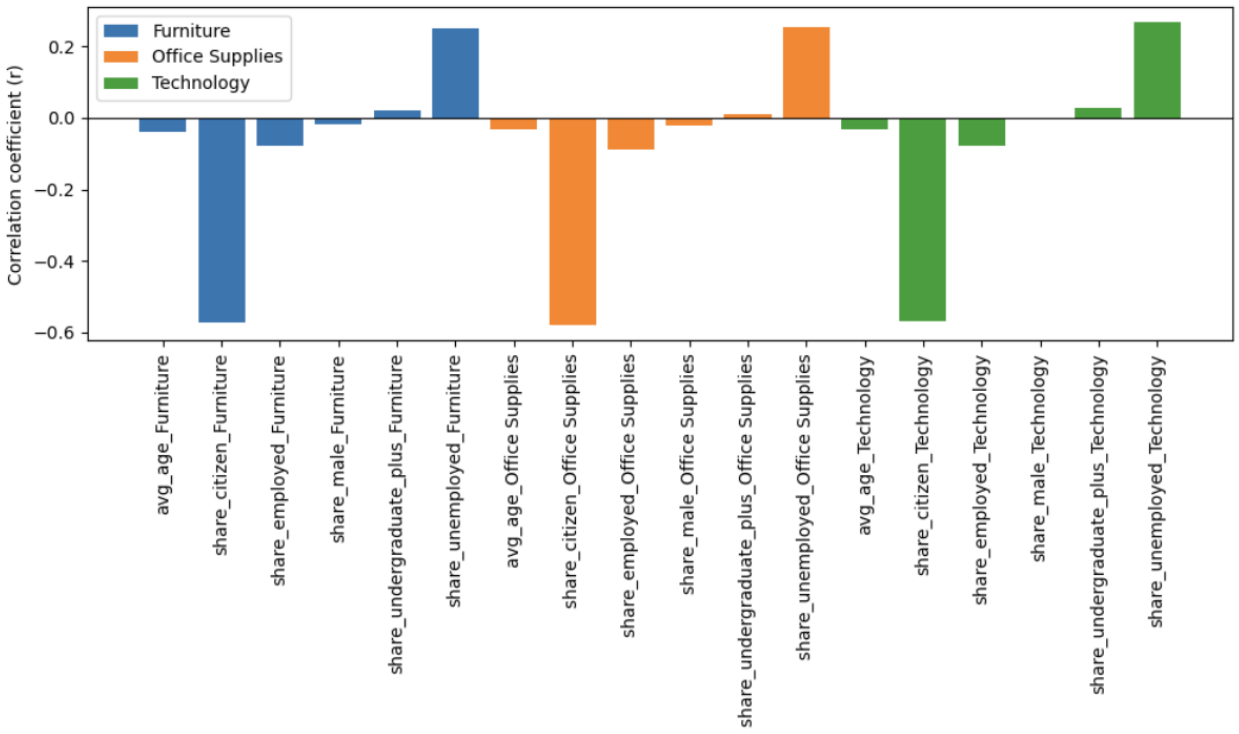

Figure 2: Raw correlation by category

The second stage introduces fixed effects adjustments. As shown in Figure 3, a Two-Way Fixed Effects demeaning procedure is applied, which controls both state and month averages. This approach eliminates long-term cross-state heterogeneity and common temporal shocks, isolating within-state and within-time variation. After adjustment, the correlations between demographic variables and sales diminish substantially, with all coefficients approaching zero and absolute values remaining below 0.03.

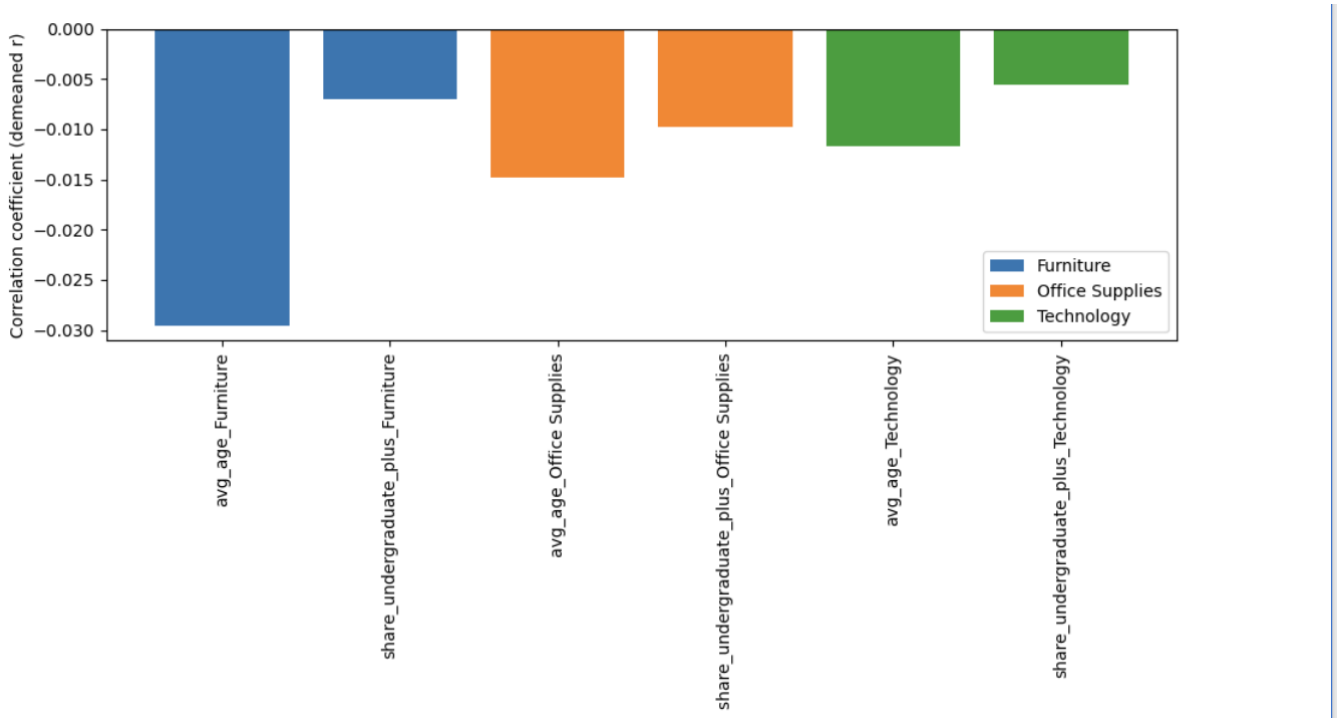

Figure 3: TWFE-adjusted correlation

In conclusion, demographic variables initially appear correlated with category-level sales, particularly citizenship and unemployment shares. However, these associations largely disappear once state and temporal fixed effects are controlled. This indicates that the observed relationships are primarily driven by structural differences across states rather than short-term demographic dynamics. From a managerial perspective, firms should regard regional baselines, such as states with systematically higher or lower citizenship shares, as structural demand characteristics rather than expecting short-run demographic fluctuations to materially influence consumption patterns.

RQ2. Is there a structural relationship between changes in educational attainment and consumption of high-value categories (Technology)?

To address Research Question 2, the dataset panel_rq2 was constructed, which integrates e-commerce aggregates (ecom_agg) with demographic aggregates (demo_agg). This allows category-level outcomes such as total sales to be analyzed alongside demographic measures, with particular focus on educational attainment, measured as the share of the population with undergraduate or higher qualifications. Interaction terms between education and product categories were created to isolate structural category-specific effects, especially for Technology. The analysis proceeded in two complementary directions. First, interaction structures was examined by comparing states with different educational profiles across categories. As shown in Figure 4, the cross-sectional relationship between education and technology consumption is weakly negative on average, with considerable dispersion across states. This indicates that the simple structural correlation is not as strong as expected, and that education may not serve as the sole driver of technology adoption. Second, the panel_rq2_diff dataset was used to track month-to-month changes in both education shares and sales. As shown in Figure 5, education shares remain nearly constant over time while technology sales fluctuate substantially, suggesting that short-term demographic changes do not align with short-term sales dynamics.

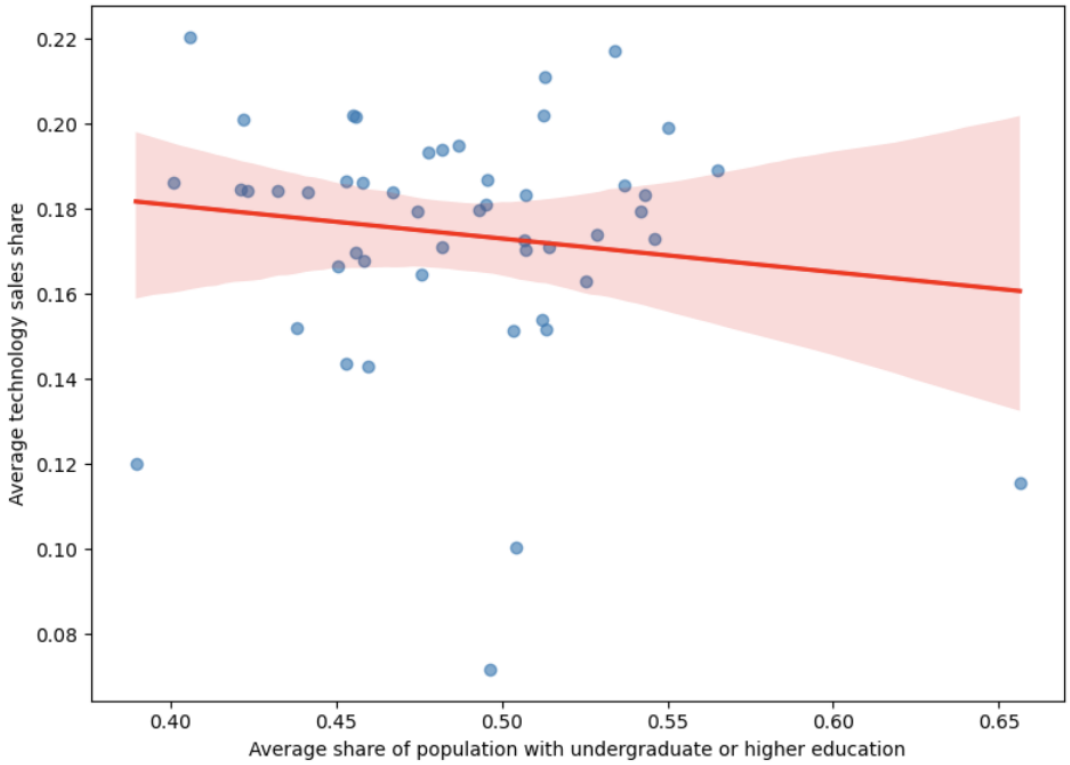

Figure 4: Education and Technology Consumption Share (State Level)

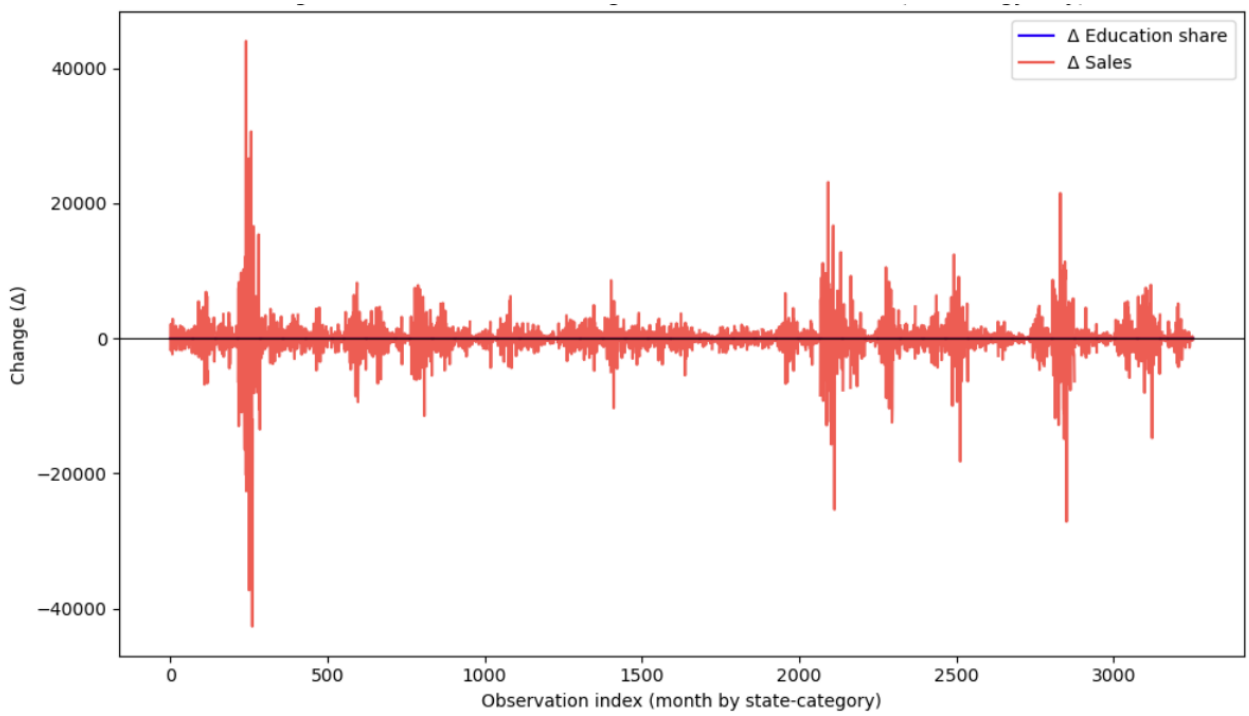

Figure 5: Month-to-Month Changes in Education adn Sales (Technology only)

The clustering analysis in state_cluster_rq2 further illustrates these insights. As shown in Figure 6, states differ substantially in their average education levels and corresponding technology sales shares, with some

states displaying both higher education attainment and higher consumption shares. However, the wide dispersion and presence of volatility imply that education interacts with other structural and cultural factors in shaping technology demand.

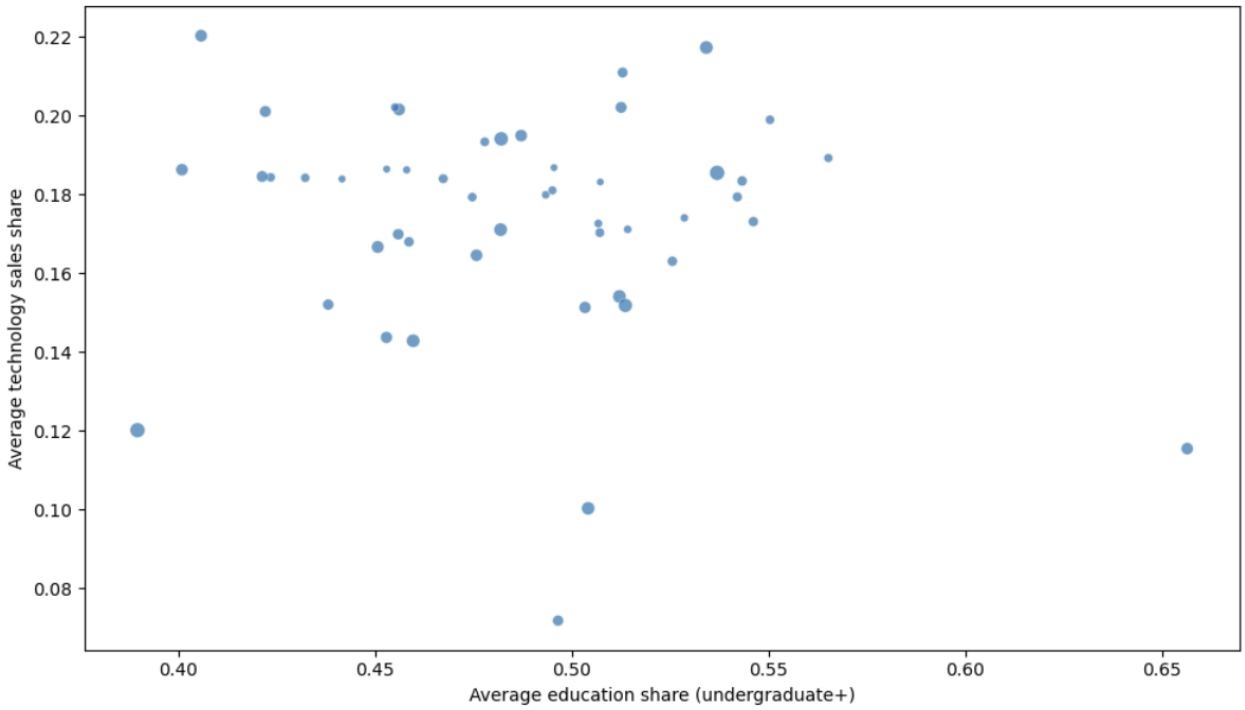

Figure 6: States by Education adn Technology Sales share (Bubble size = volatility)

In conclusion, the evidence suggests that the relationship between education and technology consumption is not uniformly strong across states. While short-term month-to-month changes in education shares do not predict immediate consumption shifts, states with higher educational attainment tend to be positioned toward the higher end of technology consumption, albeit with notable heterogeneity. From a managerial perspective, education may serve as one useful indicator when targeting regions for technology products, but firms should also account for other persistent state-specific characteristics that influence consumption patterns.

RQ3. Are nationwide demographic trends consistent with YouTube video preference trends?

To address Research Question 3, a nationwide panel dataset was constructed by merging monthly demographic aggregates (demo_trend) with YouTube category engagement shares (youtube_trend). The merged dataset (panel_rq3) provides a combined view of demographic structures and digital engagement preferences at the national level.

The analysis proceeds in two steps. First, demographic aggregates were generated by weighting state-level measures by survey weights, producing national averages of key indicators such as age, education attainment, employment status, gender distribution, and citizenship. Second, YouTube video records were reclassified into three consolidated categories: Entertainment, Gaming, and Technology. Monthly engagement was aggregated in terms of views, likes, comments, and video counts, and relative category shares were calculated to capture preference shifts over time.

As shown in Figure 7, national education attainment (share of undergraduate or higher population) moved broadly in line with technology consumption share and overall YouTube engagement (average views). This

alignment suggests that long-run demographic improvements in education coincide with greater orientation toward technology-related content and higher levels of digital engagement.

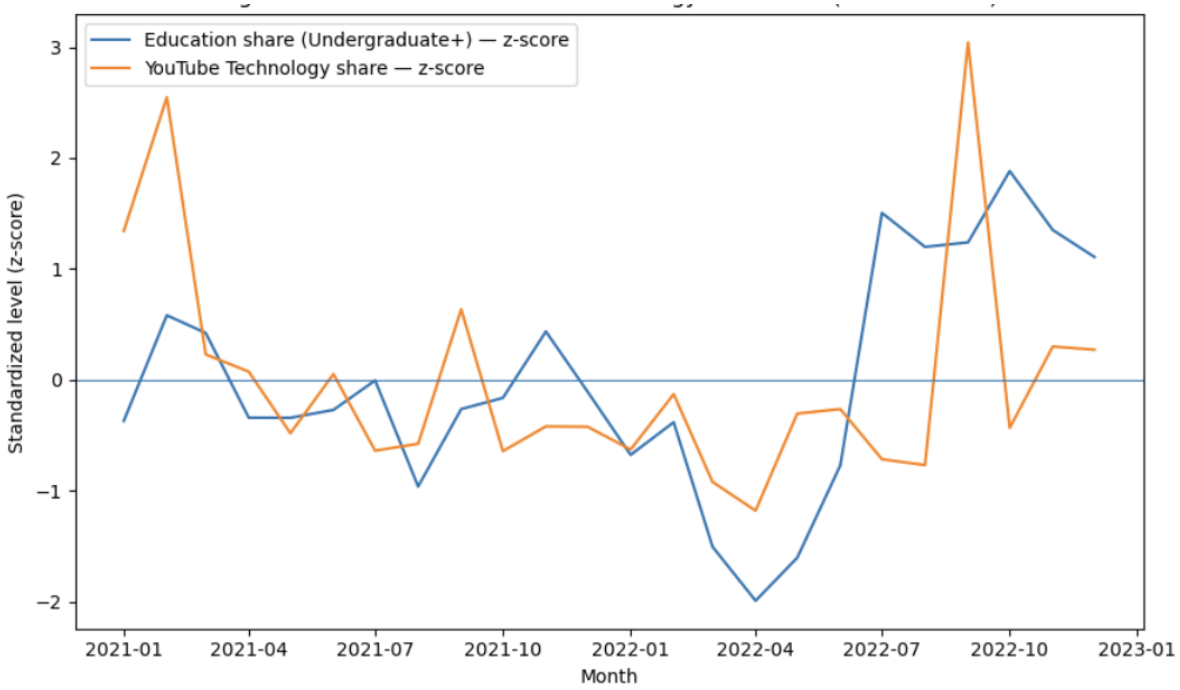

Figure 7: Education vs Youtube Technology Preference (Standardized)

Complementing this, Figure 8 compares national average age with YouTube's Entertainment share. The visual pattern indicates that periods of a relatively younger population are associated with higher attention to entertainment content, while older demographic shifts correspond with reduced entertainment engagement. This provides further evidence that demographic structures and online preferences evolve in tandem, albeit differently across content categories.

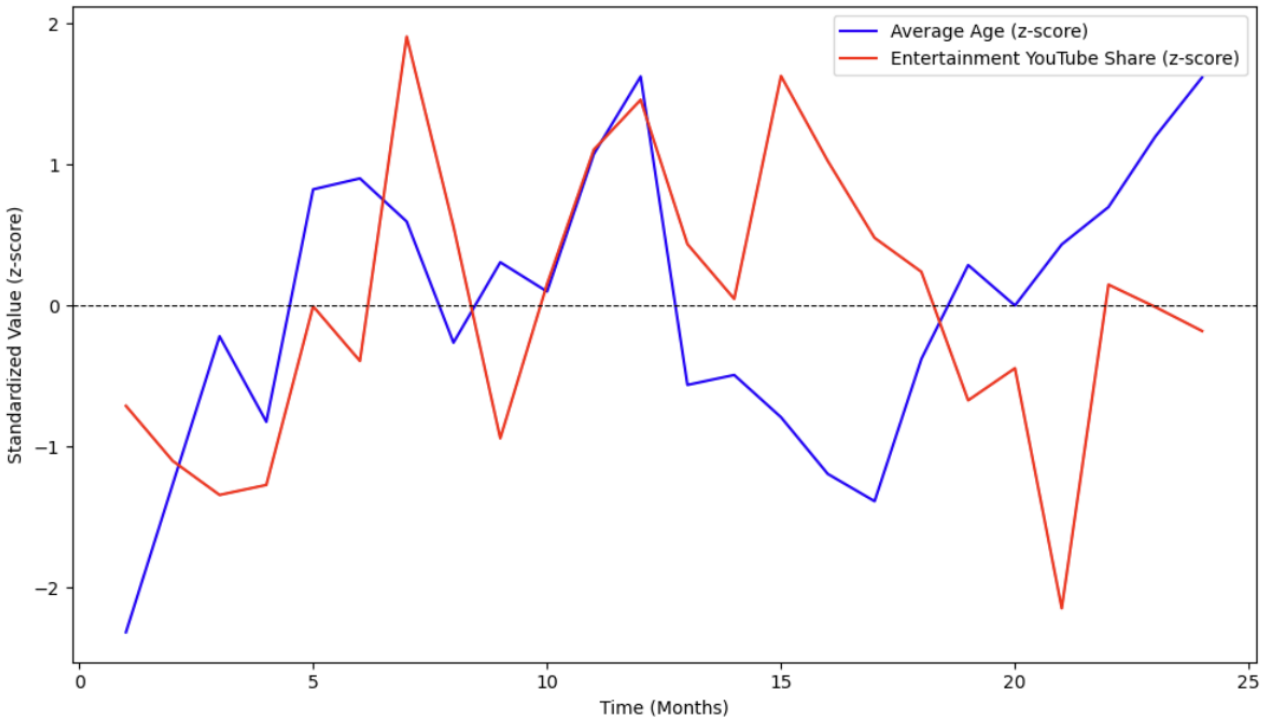

Figure 8: Average Age vs Entertainment YouTube Share

In conclusion, Figure 7 and Figure 8 illustrate that demographic trends, particularly education and age are mirrored in the evolution of online video engagement. While the alignment is visible in broad structural terms, the strength and timing of these co-movements will benefit from formal techniques such as Dynamic Time Warping or cross-correlation analysis. For managerial purposes, the findings imply that demographic trajectories shape not only offline consumption but also digital engagement landscapes, with education linked to technological adoption and age linked to entertainment demand.

RQ4. Can YouTube trending videos predict e-commerce consumption trends?

A panel dataset (panel_rq4) was constructed by combining national-level e-commerce sales with YouTube engagement shares across Technology, Furniture, and Office Supplies. Lag structures up to three months and first differences were applied to capture potential dynamic relationships. As shown in Figure 9, Technology sales and YouTube Technology share follow broadly similar long-term trajectories, but the co-movement reflects parallel trends rather than predictive power.

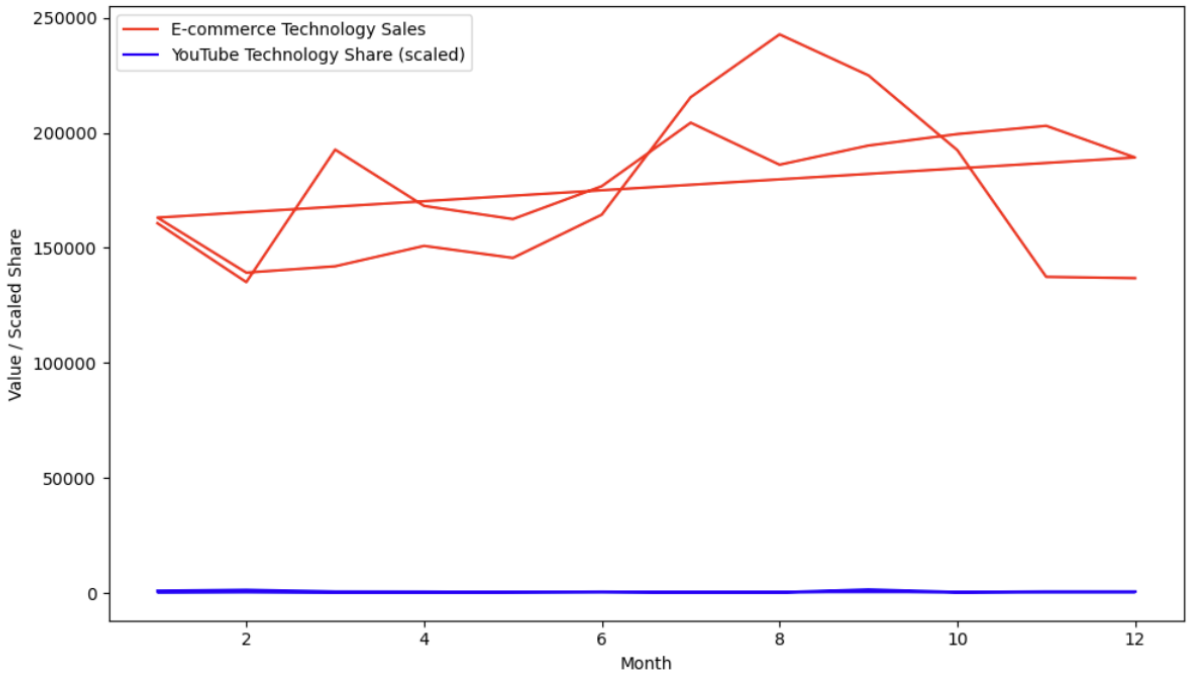

Figure 9: Technology Sales vs YouTube Technology Share (Long-trem Trend)

Short-run dynamics, illustrated in Figure 10, highlight sharp fluctuations in sales compared to relatively stable YouTube shares, with no clear contemporaneous alignment.

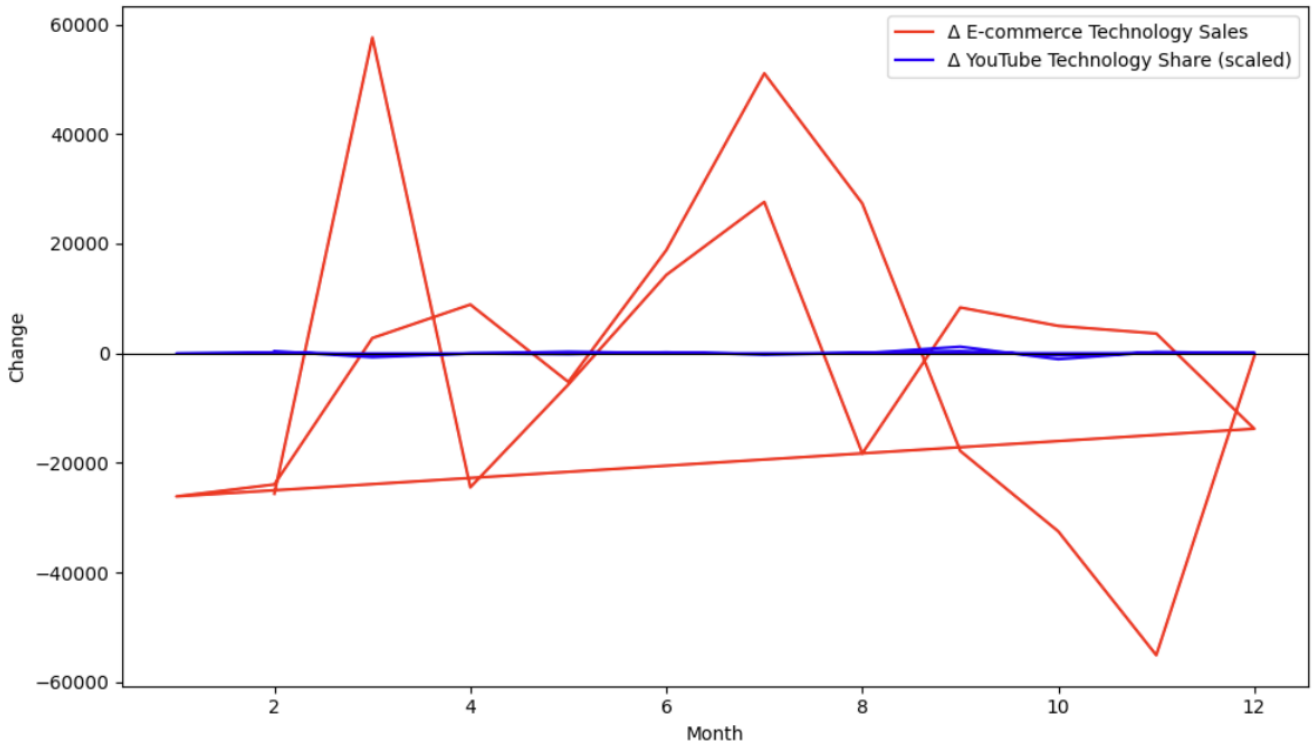

Figure 10: Month-to-Month Changes in Technology Sales and YouTube Share

Lagged correlation analysis in Figure 11 further confirms the absence of predictive strength: coefficients

remain small across all categories and lags, generally below 0.2, with no consistent direction.

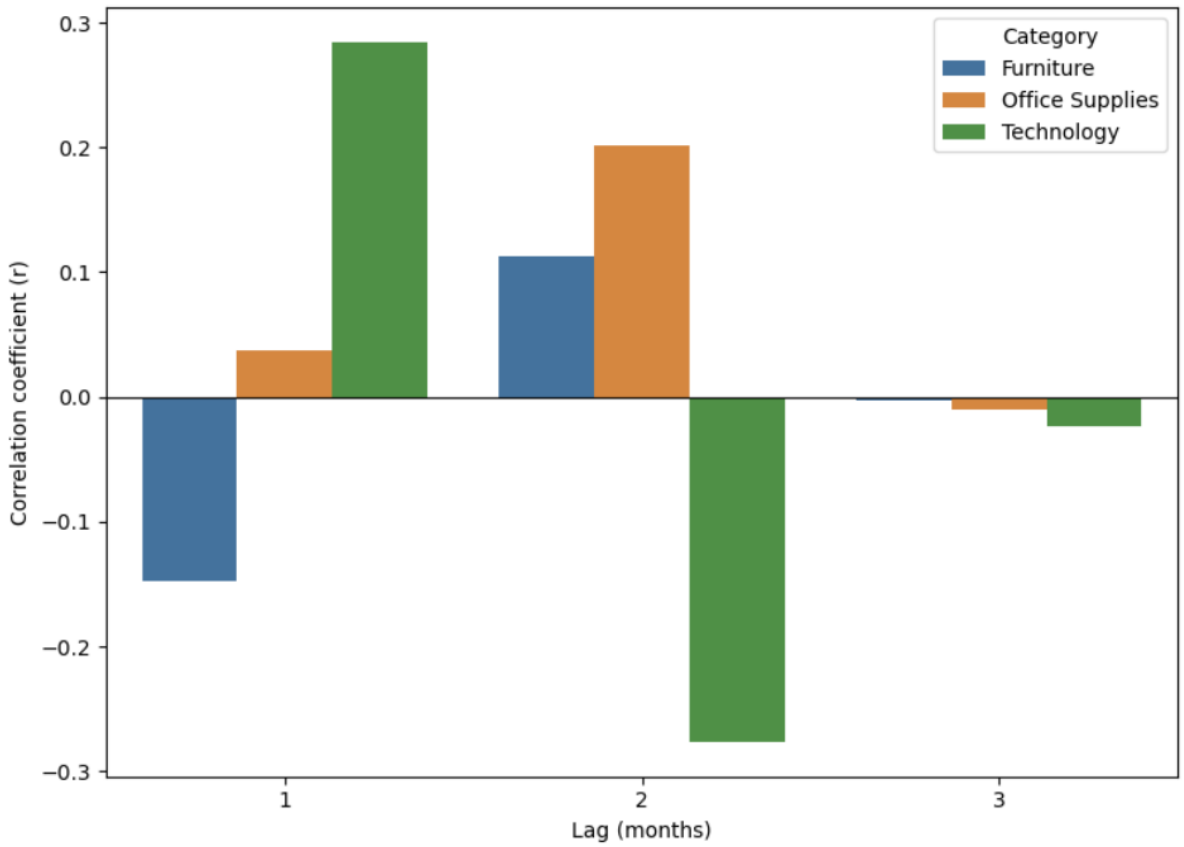

Figure 11: Lagged Correlations between YouTube and E-commerce by Category

Taken together, the findings indicate that YouTube trending metrics are not reliable predictors of near-term e-commerce sales. Instead, they should be viewed as indicators of cultural and entertainment interests rather than forecasting tools for consumption. From a managerial perspective, YouTube signals may help in understanding audience engagement and brand visibility but cannot substitute for demand forecasting models grounded in economic and market fundamentals.

RQ5. Does the demographic sensitivity of consumption behavior differ significantly across product categories?

To address this question, state–month–category level sales were merged with demographic measures, with log transformations applied to capture elasticity-style effects. The results highlight notable differences across categories.

As shown in Figure 12, technology sales are more responsive to higher education attainment, while furniture shows little association and office supplies trend slightly downward.

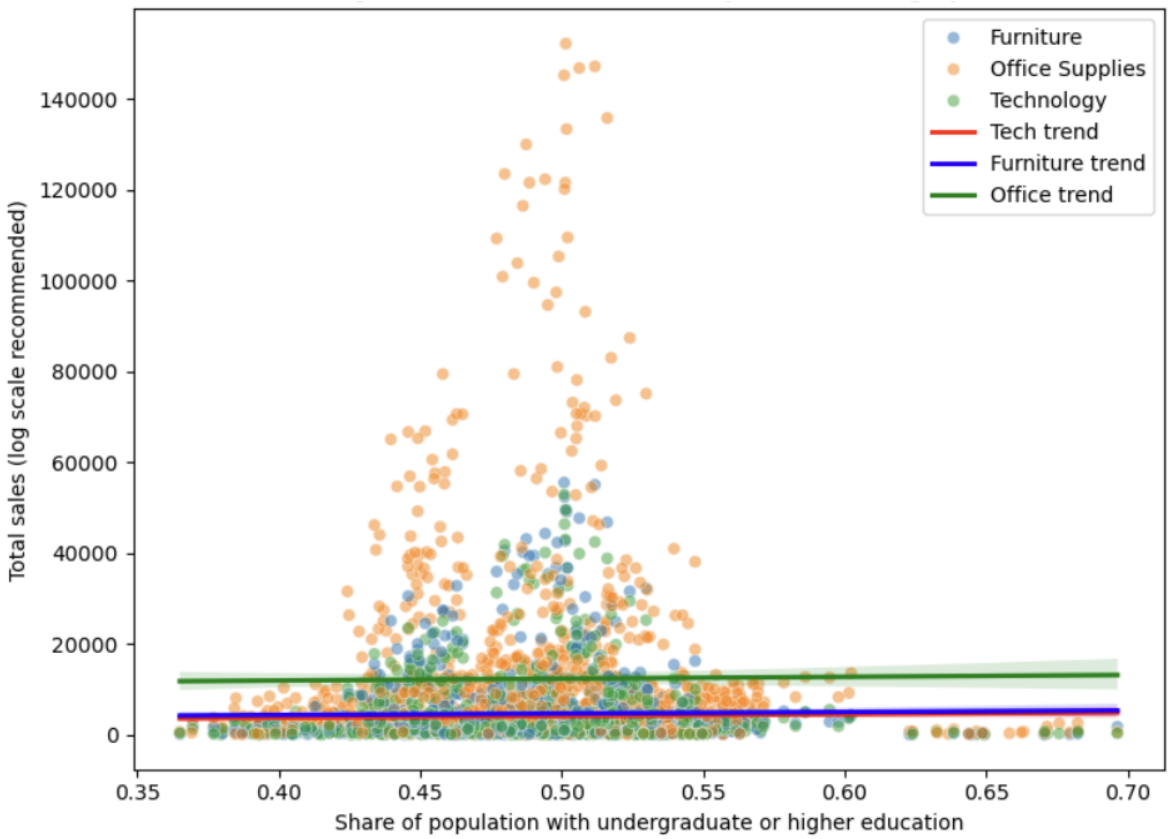

Figure 12: Scatterplot of Education and Sales by Product Category

Elasticity estimates in Table 1 confirm this pattern: technology exhibits a positive elasticity of about 0.65, compared with near-zero for furniture and a negative –0.30 for office supplies. However, t-statistics remain below conventional thresholds, suggesting limited statistical confidence.

Table 1: Elasticity of Sales with respect to Education (Undergraduate+ Share)

Category	n	r	Elasticity	t-stat	Significance
Furniture	1082	0.006	0.082	0.194	0
Office Supplies	1132	-0.022	-0.305	-0.729	0
Technology	1039	0.047	0.653	1.509	0

A parallel analysis with employment shares in Figure 13 and Table 2 yields uniformly negative elasticities, ranging from –1.57 for technology to –2.82 for office supplies, all statistically significant at conventional levels. This indicates that higher employment shares are associated with lower category sales, potentially reflecting substitution effects or measurement issues.

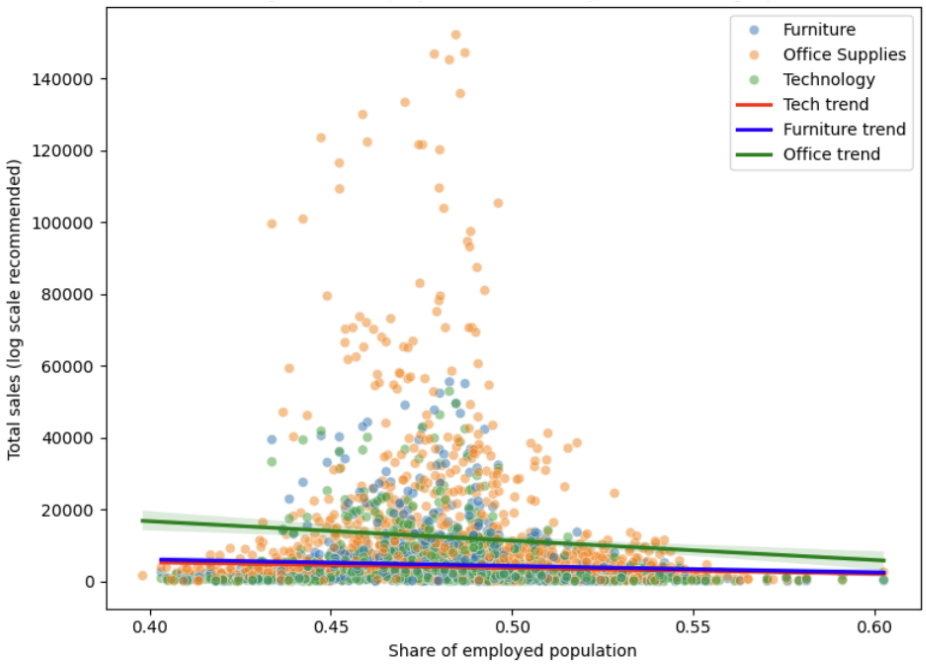

Figure 13: Employment vs Sales by Product Category

Table 2: Elasticity of Sales with respect to Employment Share

Category	n	r	Elasticity	t-stat	Significance
Furniture	1082	-0.124	-2.493	-4.113	1
Office Supplies	1132	-0.139	-2.821	-4.700	1
Technology	1039	-0.078	-1.567	-2.506	1

Finally, Figure 14 shows that the overall sales structure remains stable across states, with office supplies dominating regardless of demographic profiles.

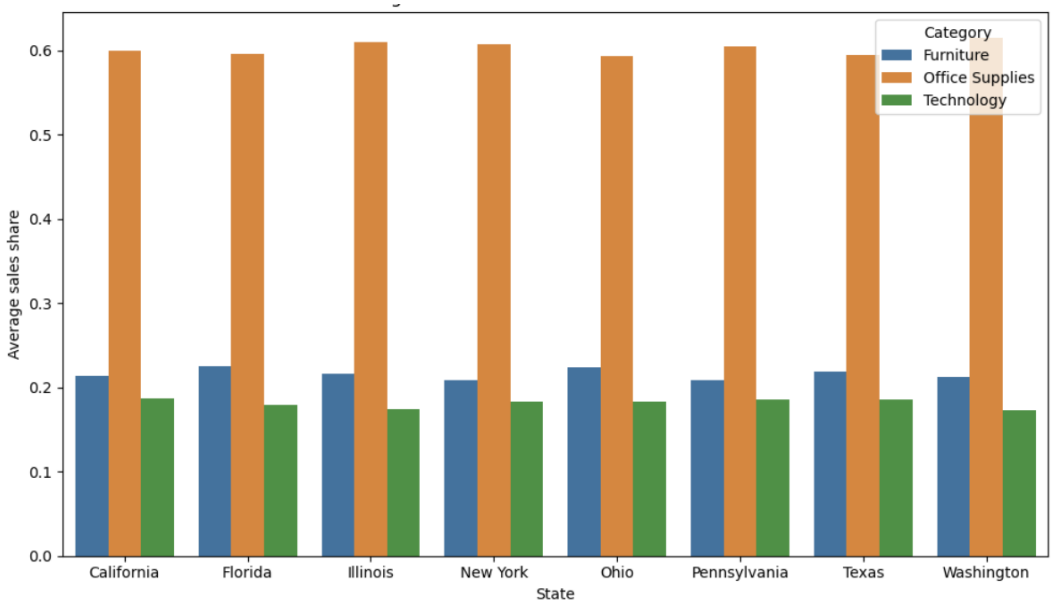

Figure 14: Sales Structure across States

Overall, evidence suggests education exerts the strongest relative influence on technology demand, while employment effects appear more consistently negative. These findings point to suggestive, though partly inconclusive, demographic sensitivities in category-specific consumption.

Insight and Future Work

The analysis of RQ1–RQ5 highlights several consistent insights. Demographic structures influence e-commerce demand primarily through long-run differences rather than short-term variation, as shown in RQ1 where fixed effects eliminated most raw correlations. Education emerged as the most important driver of technology consumption, both in RQ2 where higher attainment aligns with stronger technology demand, and in RQ5 where elasticity estimates indicate positive responsiveness to education. In contrast, employment and unemployment exhibit weaker or inconsistent associations, reinforcing that demographic sensitivity is category-specific. RQ3 further demonstrates that nationwide shifts in education and age are reflected in digital engagement, particularly in the growing share of technology-related YouTube content. However, RQ4 shows that YouTube engagement does not provide reliable short-run predictive power for e-commerce sales. Future work can strengthen these findings in three directions. First, applying more advanced causal methods, such as panel vector autoregression, could move beyond correlation to identify directionality of effects. Second, refining product categories, especially within technology, would help capture heterogeneity in education-driven demand. Third, validating the analysis with international or emerging market data would test the generalizability of structural patterns observed in the U.S. These extensions would provide more robust evidence to guide managerial decision-making in digital commerce and online engagement.

Appendix

1. RQ1_Code 1: SQL for ecom_agg

```sql
-- Ecom aggregation: State x Year x Month x category
DROP VIEW IF EXISTS ecom_agg;

CREATE VIEW ecom_agg AS
SELECT
    customer_state AS state_name,
    CAST(strftime('%Y', order_date) AS INT) AS year,
    CAST(strftime('%m', order_date) AS INT) AS month,
    category_name as category,
    SUM(sales_per_order)    AS total_sales,
    SUM(profit_per_order)   AS total_profit,
    SUM(order_quantity)     AS total_quantity
FROM ecommerce
GROUP BY state_name, year, month, category;
```

2. RQ1_Code2: SQL for demo_agg

```sql
-- Demo aggregation:  State x Year x Month
DROP VIEW IF EXISTS demo_agg;

CREATE VIEW demo_agg AS
SELECT
    CAST(YEAR AS INT)  AS year,
    CAST(MONTH AS INT) AS month,
    state_name,

    -- Total_weight
    SUM(HWTFINL) AS total_weight,

    -- Average ages
    SUM(HWTFINL * AGE) * 1.0 / SUM(HWTFINL) AS avg_age,

    -- Eduction
    SUM(CASE WHEN EDUC IN ('Undergraduate','Graduate and above') THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_undergraduate_plus,
    SUM(CASE WHEN EDUC = 'Undergraduate' THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_undergraduate,
    SUM(CASE WHEN EDUC = 'Graduate and above' THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_graduate_plus,

    -- Employment
    SUM(CASE WHEN EMPSTAT = 'At work' THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_employed,
    SUM(CASE WHEN EMPSTAT = 'Unemployed' THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_unemployed,

    -- Gender
    SUM(CASE WHEN SEX = 'Male' THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_male,

    -- Citizen
    SUM(CASE WHEN CITIZEN IN ('Born in U.S','Born in U.S. outlying','Born abroad of American parents','Naturalized citizen')
             THEN HWTFINL ELSE 0 END) * 1.0 / SUM(HWTFINL) AS share_citizen
FROM demography
GROUP BY year, month, state_name;
```

3. RQ1_Code3: SQL for panel_rq1

```sql
-- Panel Data with Differences (for RQ1)
DROP VIEW IF EXISTS panel_rq1;

CREATE VIEW panel_rq1 AS
SELECT
    e.year,
    e.month,
    e.state_name,
    e.total_sales,
    e.total_profit,
    e.total_quantity,
    e.category,
    d.avg_age,
    d.share_undergraduate_plus,
    d.share_undergraduate,
    d.share_graduate_plus,
    d.share_employed,
    d.share_unemployed,
    d.share_male,
    d.share_citizen
FROM ecom_agg e
LEFT JOIN demo_agg d
  ON e.year = d.year
 AND e.month = d.month
 AND e.state_name = d.state_name;
```

4. RQ1_Code4: Correlation results (raw)

```sql
-- Analytics RQ1_1: Raw correlation by category
DROP VIEW IF EXISTS rq1_corr_by_category;
CREATE VIEW rq1_corr_by_category AS
WITH base AS (
  SELECT category, total_sales,
         avg_age, share_undergraduate_plus, share_employed, share_unemployed, share_male, share_citizen
  FROM panel_rq1
  WHERE total_sales IS NOT NULL
)
SELECT
  category,
  -- Pearson correlation coefficient, calculated manually based on the covariance formula
  'avg_age'                AS var, (COUNT(*)*SUM(avg_age*total_sales)-SUM(avg_age)*SUM(total_sales)) * 1.0 /
                                   (SQRT((COUNT(*)*SUM(avg_age*avg_age)-SUM(avg_age)*SUM(avg_age)) *
                                         (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base
GROUP BY category
UNION ALL
SELECT category,'share_undergraduate_plus',
       (COUNT(*)*SUM(share_undergraduate_plus*total_sales)-SUM(share_undergraduate_plus)*SUM(total_sales)) * 1.0 /
       (SQRT((COUNT(*)*SUM(share_undergraduate_plus*share_undergraduate_plus)-SUM(share_undergraduate_plus)*SUM(share_undergraduate_plus)) *
             (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base GROUP BY category
UNION ALL
SELECT category,'share_employed',
       (COUNT(*)*SUM(share_employed*total_sales)-SUM(share_employed)*SUM(total_sales)) * 1.0 /
       (SQRT((COUNT(*)*SUM(share_employed*share_employed)-SUM(share_employed)*SUM(share_employed)) *
             (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base GROUP BY category
UNION ALL
SELECT category,'share_unemployed',
       (COUNT(*)*SUM(share_unemployed*total_sales)-SUM(share_unemployed)*SUM(total_sales)) * 1.0 /
       (SQRT((COUNT(*)*SUM(share_unemployed*share_unemployed)-SUM(share_unemployed)*SUM(share_unemployed)) *
             (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base GROUP BY category
UNION ALL
SELECT category,'share_male',
       (COUNT(*)*SUM(share_male*total_sales)-SUM(share_male)*SUM(total_sales)) * 1.0 /
       (SQRT((COUNT(*)*SUM(share_male*share_male)-SUM(share_male)*SUM(share_male)) *
             (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base GROUP BY category
UNION ALL
SELECT category,'share_citizen',
       (COUNT(*)*SUM(share_citizen*total_sales)-SUM(share_citizen)*SUM(total_sales)) * 1.0 /
       (SQRT((COUNT(*)*SUM(share_citizen*share_citizen)-SUM(share_citizen)*SUM(share_citizen)) *
             (COUNT(*)*SUM(total_sales*total_sales)-SUM(total_sales)*SUM(total_sales)))) AS r
FROM base GROUP BY category
ORDER BY category, var;
```

5. RQ1_Code5: Correlation results (TWFE)

```sql
-- Analytics RQ1_2: TWFE-adjusted correlation (controlling for state and month fixed effects)
DROP VIEW IF EXISTS rq1_twfe_corr;
CREATE VIEW rq1_twfe_corr AS
WITH b AS (
  SELECT * FROM panel_rq1 WHERE total_sales IS NOT NULL
),
by_state AS (
  SELECT state_name, AVG(total_sales) AS s_mean FROM b GROUP BY state_name
),
by_time AS (
  SELECT year, month, AVG(total_sales) AS t_mean FROM b GROUP BY year, month
),
demean AS (
  SELECT
    b.category, b.state_name, b.year, b.month,
    (b.total_sales - s.s_mean - t.t_mean + (SELECT AVG(total_sales) FROM b)) AS y_dm,
    (b.avg_age - AVG(b.avg_age) OVER (PARTITION BY b.state_name) - AVG(b.avg_age) OVER (PARTITION BY b.year,b.month)
      + AVG(b.avg_age) OVER ()) AS x_age_dm,
    (b.share_undergraduate_plus - AVG(b.share_undergraduate_plus) OVER (PARTITION BY b.state_name)
      - AVG(b.share_undergraduate_plus) OVER (PARTITION BY b.year,b.month) + AVG(b.share_undergraduate_plus) OVER ()) AS x_edu_dm
  FROM b b
  JOIN by_state s ON s.state_name=b.state_name
  JOIN by_time  t ON t.year=b.year AND t.month=b.month
)
SELECT
  category,
  'avg_age' AS var,
  (COUNT(*)*SUM(x_age_dm*y_dm)-SUM(x_age_dm)*SUM(y_dm)) * 1.0 /
  (SQRT((COUNT(*)*SUM(x_age_dm*x_age_dm)-SUM(x_age_dm)*SUM(x_age_dm)) *
        (COUNT(*)*SUM(y_dm*y_dm)-SUM(y_dm)*SUM(y_dm)))) AS r_dm
FROM demean
GROUP BY category
UNION ALL
SELECT
  category,
  'share_undergraduate_plus',
  (COUNT(*)*SUM(x_edu_dm*y_dm)-SUM(x_edu_dm)*SUM(y_dm)) * 1.0 /
  (SQRT((COUNT(*)*SUM(x_edu_dm*x_edu_dm)-SUM(x_edu_dm)*SUM(x_edu_dm)) *
        (COUNT(*)*SUM(y_dm*y_dm)-SUM(y_dm)*SUM(y_dm)))) AS r_dm
FROM demean
GROUP BY category
ORDER BY category, var;
```

6. RQ2_Code1: Education × Category Interactions & Category Dummy Variables

```sql
-- RQ2: Is there a structural relationship between changes in educational attainment and consumption of high-value categories (Technology)?
-- 1). Education × Category Interactions & Category Dummy Variables
DROP VIEW IF EXISTS panel_rq2;

CREATE VIEW panel_rq2 AS
SELECT
  e.year,
  e.month,
  e.state_name,
  e.category,

  -- Dependent variables (choose the one need for analysis from ecom_agg)
  e.total_sales,
  e.total_profit,
  e.total_quantity,

  -- Education and control variables (from demo_agg)
  d.share_undergraduate_plus,
  d.share_undergraduate,
  d.share_graduate_plus,
  d.share_employed,
  d.share_unemployed,
  d.avg_age,
  d.share_male,
  d.share_citizen,

  -- Category dummy variables
  CASE WHEN e.category='Technology'      THEN 1 ELSE 0 END AS is_tech,
  CASE WHEN e.category='Furniture'       THEN 1 ELSE 0 END AS is_furniture,
  CASE WHEN e.category='Office Supplies' THEN 1 ELSE 0 END AS is_office,

  -- Interaction terms: Education (Undergraduate+) × Category
  CASE WHEN e.category='Technology'      THEN d.share_undergraduate_plus ELSE 0 END AS edu_up_x_tech,
  CASE WHEN e.category='Furniture'       THEN d.share_undergraduate_plus ELSE 0 END AS edu_up_x_furn,
  CASE WHEN e.category='Office Supplies' THEN d.share_undergraduate_plus ELSE 0 END AS edu_up_x_office

FROM ecom_agg e
LEFT JOIN demo_agg d
  ON d.year = e.year
 AND d.month = e.month
 AND d.state_name = e.state_name;
```

7. RQ2_Code2: First Differences: Sales & Education

```sql
-- 2). First Differences: Sales & Education
DROP VIEW IF EXISTS panel_rq2_diff;

CREATE VIEW panel_rq2_diff AS
WITH base AS (
  SELECT
    e.year,
    e.month,
    e.state_name,
    e.category,
    e.total_sales,
    d.share_undergraduate_plus
  FROM ecom_agg e
  LEFT JOIN demo_agg d
    ON d.year = e.year AND d.month = e.month AND d.state_name = e.state_name
),
lagged AS (
  SELECT
    *,
    -- One-month lag of sales by state × category
    LAG(total_sales, 1) OVER (
      PARTITION BY state_name, category
      ORDER BY year, month
    ) AS sales_lag1,

    -- One-month lag of education share by state
    LAG(share_undergraduate_plus, 1) OVER (
      PARTITION BY state_name
      ORDER BY year, month
    ) AS edu_up_lag1
  FROM base
)
SELECT
  *,
  -- First difference of sales and education
  (total_sales - sales_lag1)               AS d_sales_m1,
  (share_undergraduate_plus - edu_up_lag1) AS d_edu_up_m1,

  -- Tech dummy for clustering or filtering
  CASE WHEN category='Technology' THEN 1 ELSE 0 END AS is_tech
FROM lagged;
```

8. RQ2_Code3: State-level Clustering Features: Technology Sales Share & Smoothing Metrics

```sql
-- 3). State-level Clustering Features: Technology Sales Share & Smoothing Metrics
DROP VIEW IF EXISTS state_cluster_rq2;

CREATE VIEW state_cluster_rq2 AS
WITH sales_mix AS (
  -- Technology sales share by state × month
  SELECT
    e.year, e.month, e.state_name,
    SUM(CASE WHEN e.category = 'Technology' THEN e.total_sales ELSE 0 END) * 1.0 /
    NULLIF(SUM(e.total_sales), 0) AS tech_sales_share
  FROM ecom_agg e
  GROUP BY e.year, e.month, e.state_name
),
sales_mix_lag AS (
  SELECT
    *,
    -- Lagged technology sales share
    LAG(tech_sales_share, 1) OVER (
      PARTITION BY state_name
      ORDER BY year, month
    ) AS tech_share_lag1
  FROM sales_mix
),
edu_series AS (
  SELECT year, month, state_name, share_undergraduate_plus
  FROM demo_agg
),
edu_lag AS (
  SELECT
    *,
    -- Lagged education share
    LAG(share_undergraduate_plus, 1) OVER (
      PARTITION BY state_name
      ORDER BY year, month
    ) AS edu_up_lag1
  FROM edu_series
)
SELECT
  s.state_name,
  -- Average level & month-to-month volatility of Tech share
  AVG(s.tech_sales_share)                       AS tech_share_mean,
  AVG(ABS(s.tech_sales_share - s.tech_share_lag1)) AS tech_share_mae,
  -- Education level and average first-difference
  AVG(e.share_undergraduate_plus)               AS edu_up_mean,
  AVG(e.share_undergraduate_plus - e.edu_up_lag1)  AS edu_up_d1_mean
FROM sales_mix_lag s
JOIN edu_lag e
  ON e.state_name = s.state_name
 AND e.year       = s.year
 AND e.month      = s.month
GROUP BY s.state_name;
```

9. RQ3_Code1: Nationwide demographic trends: month, weighted (from demo_agg)

```sql
-- RQ3: Are nationwide demographic trends consistent with YouTube video preference trends?
-- 1). Nationwide demographic trends: month, weighted (from demo_agg)
-- national weighted monthly demographics
DROP VIEW IF EXISTS demo_trend;

CREATE VIEW demo_trend AS
SELECT
  CAST(year AS INT)  AS year,
  CAST(month AS INT) AS month,

  SUM(total_weight) AS us_total_weight,

  -- weighted avg
  SUM(total_weight * avg_age)                * 1.0 / NULLIF(SUM(total_weight),0) AS us_avg_age,
  SUM(total_weight * share_undergraduate_plus) * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_undergraduate_plus,
  SUM(total_weight * share_undergraduate)      * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_undergraduate,
  SUM(total_weight * share_graduate_plus)      * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_graduate_plus,
  SUM(total_weight * share_employed)           * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_employed,
  SUM(total_weight * share_unemployed)         * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_unemployed,
  SUM(total_weight * share_male)               * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_male,
  SUM(total_weight * share_citizen)            * 1.0 / NULLIF(SUM(total_weight),0) AS us_share_citizen
FROM demo_agg
GROUP BY year, month
ORDER BY year, month;
```

10. RQ3_Code2: Map YouTube categories into 3 groups

```sql
-- 2). Map YouTube categories into 3 groups
DROP VIEW IF EXISTS youtube_mapped;

CREATE VIEW youtube_mapped AS
SELECT
  CAST(strftime('%Y', trending_date) AS INT)  AS year,
  CAST(strftime('%m', trending_date) AS INT)  AS month,
  CASE
    WHEN categoryId IN ('Entertainment','Film & Animation','People & Blogs','Comedy','Movies') THEN 'Entertainment'
    WHEN categoryId IN ('Gaming') THEN 'Gaming'
    WHEN categoryId IN ('Science & Technology','Technology') THEN 'Technology'
    ELSE 'Other'
  END AS std_category,
  COALESCE(view_count,0)    AS views,
  COALESCE(likes,0)         AS likes,
  COALESCE(comment_count,0) AS comments,
  1 AS videos,  --count every row as 1
  (COALESCE(view_count,0) + COALESCE(likes,0) + COALESCE(comment_count,0)) AS engagement
FROM youtube;
```

11. RQ3_Code3: Monthly aggregation by category

```sql
-- 3). Monthly aggregation by category
DROP VIEW IF EXISTS youtube_monthly;

CREATE VIEW youtube_monthly AS
SELECT
  year, month, std_category,
  SUM(views)       AS views,
  SUM(likes)       AS likes,
  SUM(comments)    AS comments,
  COUNT(videos)    AS videos,
  SUM(engagement)  AS eng
FROM youtube_mapped
GROUP BY year, month, std_category;

-- 4). Restrict to 3 target categories and compute shares
DROP VIEW IF EXISTS youtube_trend;
```

12. RQ_Code4: Restrict to 3 target categories and compute shares

```sql
-- 4). Restrict to 3 target categories and compute shares
DROP VIEW IF EXISTS youtube_trend;

CREATE VIEW youtube_trend AS
WITH three AS (
  SELECT year, month,
         SUM(CASE WHEN std_category='Entertainment' THEN eng ELSE 0 END) AS eng_entertainment,
         SUM(CASE WHEN std_category='Gaming'        THEN eng ELSE 0 END) AS eng_gaming,
         SUM(CASE WHEN std_category='Technology'    THEN eng ELSE 0 END) AS eng_technology,
         SUM(CASE WHEN std_category='Entertainment' THEN views ELSE 0 END) AS views_entertainment,
         SUM(CASE WHEN std_category='Gaming'        THEN views ELSE 0 END) AS views_gaming,
         SUM(CASE WHEN std_category='Technology'    THEN views ELSE 0 END) AS views_technology,
         SUM(CASE WHEN std_category='Entertainment' THEN likes ELSE 0 END) AS likes_entertainment,
         SUM(CASE WHEN std_category='Gaming'        THEN likes ELSE 0 END) AS likes_gaming,
         SUM(CASE WHEN std_category='Technology'    THEN likes ELSE 0 END) AS likes_technology,
         SUM(CASE WHEN std_category='Entertainment' THEN comments ELSE 0 END) AS comments_entertainment,
         SUM(CASE WHEN std_category='Gaming'        THEN comments ELSE 0 END) AS comments_gaming,
         SUM(CASE WHEN std_category='Technology'    THEN comments ELSE 0 END) AS comments_technology,
         SUM(CASE WHEN std_category='Entertainment' THEN videos ELSE 0 END) AS videos_entertainment,
         SUM(CASE WHEN std_category='Gaming'        THEN videos ELSE 0 END) AS videos_gaming,
         SUM(CASE WHEN std_category='Technology'    THEN videos ELSE 0 END) AS videos_technology
  FROM youtube_monthly
  WHERE std_category IN ('Entertainment','Gaming','Technology')
  GROUP BY year, month
)
SELECT
  year, month,
  eng_entertainment, eng_gaming, eng_technology,
  eng_entertainment * 1.0 / NULLIF(eng_entertainment + eng_gaming + eng_technology, 0) AS us_share_entertainment,
  eng_gaming       * 1.0 / NULLIF(eng_entertainment + eng_gaming + eng_technology, 0) AS us_share_gaming,
  eng_technology   * 1.0 / NULLIF(eng_entertainment + eng_gaming + eng_technology, 0) AS us_share_technology,
  views_entertainment, views_gaming, views_technology,
  likes_entertainment, likes_gaming, likes_technology,
  comments_entertainment, comments_gaming, comments_technology,
  videos_entertainment,   videos_gaming,   videos_technology
FROM three
ORDER BY year, month;
```

13. RQ3_Code5: Final panel for RQ3

```sql
-- 5). Final panel for RQ3
DROP VIEW IF EXISTS panel_rq3;

CREATE VIEW panel_rq3 AS
SELECT
  d.year, d.month,
  d.us_total_weight, d.us_avg_age,
  d.us_share_undergraduate_plus, d.us_share_employed, d.us_share_male,
  y.us_share_entertainment, y.us_share_gaming, y.us_share_technology,
  y.eng_entertainment, y.eng_gaming, y.eng_technology
FROM demo_trend d
JOIN youtube_trend y
  ON y.year  = d.year
 AND y.month = d.month
ORDER BY d.year, d.month;
```

14. RQ4_Code1: Base panel (monthly national): YT shares + E-com sales (pivoted)

```sql
--RQ4: Can YouTube trending videos predict e-commerce consumption trends?
-- 1). Base panel (monthly national): YT shares + E-com sales (pivoted)
DROP VIEW IF EXISTS panel_rq4;

CREATE VIEW panel_rq4 AS
SELECT
  y.year,
  y.month,

  -- YouTube category shares (relative among tracked cats)
  y.us_share_entertainment,
  y.us_share_gaming,
  y.us_share_technology,

  -- Absolute engagement (optional controls)
  y.eng_entertainment,
  y.eng_gaming,
  y.eng_technology,

  -- E-commerce sales by headline categories (pivoted)
  e.ec_technology,
  e.ec_furniture,
  e.ec_office
FROM youtube_trend y
JOIN (
  SELECT
    year, month,
    SUM(CASE WHEN category = 'Technology'       THEN total_sales ELSE 0 END) AS ec_technology,
    SUM(CASE WHEN category = 'Furniture'        THEN total_sales ELSE 0 END) AS ec_furniture,
    SUM(CASE WHEN category = 'Office Supplies'  THEN total_sales ELSE 0 END) AS ec_office
  FROM ecom_agg
  GROUP BY year, month
) e USING (year, month)
ORDER BY year, month;
```

15. RQ4_Code2: Add 1–3 month lags and first differences (Δ)

```sql
-- 2). Add 1-3 month lags and first differences (Δ)
DROP VIEW IF EXISTS panel_rq4_diff;

CREATE VIEW panel_rq4_diff AS
WITH lagged AS (
  SELECT
    p.*,

    -- Lags of YouTube shares
    LAG(us_share_entertainment,1) OVER (ORDER BY year,month) AS ent_lag1,
    LAG(us_share_entertainment,2) OVER (ORDER BY year,month) AS ent_lag2,
    LAG(us_share_entertainment,3) OVER (ORDER BY year,month) AS ent_lag3,

    LAG(us_share_gaming,1) OVER (ORDER BY year,month) AS gam_lag1,
    LAG(us_share_gaming,2) OVER (ORDER BY year,month) AS gam_lag2,
    LAG(us_share_gaming,3) OVER (ORDER BY year,month) AS gam_lag3,

    LAG(us_share_technology,1) OVER (ORDER BY year,month) AS tech_lag1,
    LAG(us_share_technology,2) OVER (ORDER BY year,month) AS tech_lag2,
    LAG(us_share_technology,3) OVER (ORDER BY year,month) AS tech_lag3,

    -- Lags of E-com sales
    LAG(ec_technology,1) OVER (ORDER BY year,month) AS ec_tech_lag1,
    LAG(ec_furniture,1)  OVER (ORDER BY year,month) AS ec_furn_lag1,
    LAG(ec_office,1)     OVER (ORDER BY year,month) AS ec_office_lag1
  FROM panel_rq4 p
)
SELECT
  year, month,

  -- Δ E-com sales (t - t-1)
  (ec_technology - ec_tech_lag1)   AS d_ec_technology,
  (ec_furniture  - ec_furn_lag1)   AS d_ec_furniture,
  (ec_office     - ec_office_lag1) AS d_ec_office,

  -- Δ YT shares (t - t-1)
  (us_share_entertainment - ent_lag1) AS d_ent_share,
  (us_share_gaming        - gam_lag1) AS d_gam_share,
  (us_share_technology    - tech_lag1) AS d_tech_share,

  -- Keep lagged YT share diffs for lead-lag checks
  LAG((us_share_entertainment - ent_lag1),1) OVER (ORDER BY year,month) AS d_ent_lag1,
  LAG((us_share_gaming - gam_lag1),1) OVER (ORDER BY year,month) AS d_gam_lag1,
  LAG((us_share_technology - tech_lag1),1) OVER (ORDER BY year,month) AS d_tech_lag1,

  LAG((us_share_entertainment - ent_lag1),2) OVER (ORDER BY year,month) AS d_ent_lag2,
  LAG((us_share_gaming - gam_lag1),2) OVER (ORDER BY year,month) AS d_gam_lag2,
  LAG((us_share_technology - tech_lag1),2) OVER (ORDER BY year,month) AS d_tech_lag2,

  LAG((us_share_entertainment - ent_lag1),3) OVER (ORDER BY year,month) AS d_ent_lag3,
  LAG((us_share_gaming - gam_lag1),3) OVER (ORDER BY year,month) AS d_gam_lag3,
  LAG((us_share_technology - tech_lag1),3) OVER (ORDER BY year,month) AS d_tech_lag3
FROM lagged
WHERE ent_lag1 IS NOT NULL
  AND ec_tech_lag1 IS NOT NULL
  AND ec_furn_lag1 IS NOT NULL
  AND ec_office_lag1 IS NOT NULL;
```

16. RQ4_Code3: SQL-only lead–lag summary: correlate ΔEC with lagged ΔYT shares

```sql
-- 3). SQL-only lead-lag summary: correlate ΔEC with lagged ΔYT shares
DROP VIEW IF EXISTS state_cluster_rq4;

CREATE VIEW state_cluster_rq4 AS
WITH pairs AS (
  -- Map categories to likely YT drivers (edit if needed)
  SELECT 'Technology' AS cat, 1 AS lag, d_ec_technology AS y, d_tech_lag1 AS x FROM panel_rq4_diff
  UNION ALL SELECT 'Technology', 2, d_ec_technology, d_tech_lag2 FROM panel_rq4_diff
  UNION ALL SELECT 'Technology', 3, d_ec_technology, d_tech_lag3 FROM panel_rq4_diff

  UNION ALL SELECT 'Furniture',  1, d_ec_furniture,  d_ent_lag1 FROM panel_rq4_diff
  UNION ALL SELECT 'Furniture',  2, d_ec_furniture,  d_ent_lag2 FROM panel_rq4_diff
  UNION ALL SELECT 'Furniture',  3, d_ec_furniture,  d_ent_lag3 FROM panel_rq4_diff

  UNION ALL SELECT 'Office Supplies', 1, d_ec_office, d_gam_lag1 FROM panel_rq4_diff
  UNION ALL SELECT 'Office Supplies', 2, d_ec_office, d_gam_lag2 FROM panel_rq4_diff
  UNION ALL SELECT 'Office Supplies', 3, d_ec_office, d_gam_lag3 FROM panel_rq4_diff
),
clean AS (
  SELECT cat, lag, x, y FROM pairs WHERE x IS NOT NULL AND y IS NOT NULL
),
stats AS (
  SELECT
    cat, lag,
    COUNT(*) AS n,
    SUM(x)   AS sx,  SUM(y) AS sy,
    SUM(x*x) AS sxx, SUM(y*y) AS syy,
    SUM(x*y) AS sxy
  FROM clean
  GROUP BY cat, lag
)
SELECT
  cat,
  lag,
  n,
  -- Pearson r
  (n*sxy - sx*sy) * 1.0 /
  (SQRT( (n*sxx - sx*sx) * (n*syy - sy*sy) )) AS r,
  -- Approx significance flag (|t|≈≥2)
  CASE
    WHEN n > 2 THEN
      CASE WHEN ABS( ((n*sxy - sx*sy) * 1.0 /
                      (SQRT( (n*sxx - sx*sx) * (n*syy - sy*sy) ))) *
                     SQRT( (n-2) * 1.0 /
                           (1 - POWER( (n*sxy - sx*sy) * 1.0 /
                                       (SQRT( (n*sxx - sx*sx) * (n*syy - sy*sy) )), 2) ) )
                   ) >= 2
           THEN 1 ELSE 0 END
  END AS signif_approx
FROM stats
ORDER BY cat, lag;
```

17. RQ5_Code1: Panel: state × month × category + demographics

```sql
-- RQ5:Does the demographic sensitivity of consumption behavior differ significantly across product categor
-- 1). Panel: state × month × category + demographics
DROP VIEW IF EXISTS panel_rq5;

CREATE VIEW panel_rq5 AS
SELECT
  e.year,
  e.month,
  e.state_name,
  e.category,

  -- Outcomes
  e.total_sales,
  e.total_profit,
  e.total_quantity,

  -- Demographics (shares in 0–1)
  d.avg_age,
  d.share_undergraduate_plus,
  d.share_employed,
  d.share_unemployed,
  d.share_male,
  d.share_citizen
FROM ecom_agg e
LEFT JOIN demo_agg d
  ON d.year  = e.year
 AND d.month = e.month
 AND d.state_name = e.state_name;
```

18. RQ5_Code2: Log transforms for elasticity-style analysis

```sql
-- 2). Log transforms for elasticity-style analysis
DROP VIEW IF EXISTS panel_rq5_log;

CREATE VIEW panel_rq5_log AS
SELECT
  *,
  ln(total_sales + 1.0) AS ln_sales,
  ln(total_quantity + 1.0) AS ln_qty,
  ln(NULLIF(avg_age,0)) AS ln_age,
  ln(NULLIF(share_undergraduate_plus,0) + 1e-6) AS ln_edu_up,
  ln(NULLIF(share_employed,0) + 1e-6) AS ln_emp,
  ln(NULLIF(share_unemployed,0) + 1e-6) AS ln_unemp,
  ln(COALESCE(share_male,0.5)) AS ln_male,
  ln(COALESCE(share_citizen,0.9)) AS ln_citizen
FROM panel_rq5;
```

19. RQ5_Code3_A: Category-level elasticities: ln_sales ~ ln_edu_up and ln_sales ~ ln_emp (per category) - Elasticity wrt Education (Undergraduate+)

```sql
-- 3). Category-level elasticities: ln_sales ~ ln_edu_up and ln_sales ~ ln_emp (per category)
-- A) Elasticity wrt Education (Undergraduate+)
DROP VIEW IF EXISTS rq5_elasticity_edu;

CREATE VIEW rq5_elasticity_edu AS
WITH base AS (
  SELECT category, ln_sales AS y, ln_edu_up AS x
  FROM panel_rq5_log
  WHERE ln_sales IS NOT NULL AND ln_edu_up IS NOT NULL
),
grp AS (
  SELECT
    category,
    COUNT(*) AS n,
    SUM(x) AS sx, SUM(y) AS sy,
    SUM(x*x) AS sxx, SUM(y*y) AS syy,
    SUM(x*y) AS sxy
  FROM base
  GROUP BY category
),
rstat AS (
  SELECT
    category, n, sx, sy, sxx, syy, sxy,
    (n*sxy - sx*sy) * 1.0 /
    (SQRT( (n*sxx - sx*sx) * (n*syy - sy*sy) )) AS r
  FROM grp
)
SELECT
  category,
  n,
  r,
  -- slope (elasticity) = r * (sd_y / sd_x)
  r * ( SQRT( (n*syy - sy*sy)*1.0 ) / SQRT(n) ) /
      ( SQRT( (n*sxx - sx*sx)*1.0 ) / SQRT(n) ) AS elasticity,
  -- approx t-stat for r (df=n-2)
  CASE WHEN n>2 THEN r * SQRT( (n-2) * 1.0 / (1 - r*r) ) END AS t_stat,
  CASE WHEN n>2 AND ABS(r * SQRT( (n-2) * 1.0 / (1 - r*r) )) >= 2 THEN 1 ELSE 0 END AS signif_approx
FROM rstat
ORDER BY category;
```

20. RQ5_Code3_B: Category-level elasticities: ln_sales ~ ln_edu_up and ln_sales ~ ln_emp (per category) - Elasticity wrt Employment

```sql
-- B) Elasticity wrt Employment
DROP VIEW IF EXISTS rq5_elasticity_emp;

CREATE VIEW rq5_elasticity_emp AS
WITH base AS (
  SELECT category, ln_sales AS y, ln_emp AS x
  FROM panel_rq5_log
  WHERE ln_sales IS NOT NULL AND ln_emp IS NOT NULL
),
grp AS (
  SELECT
    category,
    COUNT(*) AS n,
    SUM(x) AS sx, SUM(y) AS sy,
    SUM(x*x) AS sxx, SUM(y*y) AS syy,
    SUM(x*y) AS sxy
  FROM base
  GROUP BY category
),
rstat AS (
  SELECT
    category, n, sx, sy, sxx, syy, sxy,
    (n*sxy - sx*sy) * 1.0 /
    (SQRT( (n*sxx - sx*sx) * (n*syy - sy*sy) )) AS r
  FROM grp
)
SELECT
  category,
  n,
  r,
  r * ( SQRT( (n*syy - sy*sy)*1.0 ) / SQRT(n) ) /
      ( SQRT( (n*sxx - sx*sx)*1.0 ) / SQRT(n) ) AS elasticity,
  CASE WHEN n>2 THEN r * SQRT( (n-2) * 1.0 / (1 - r*r) ) END AS t_stat,
  CASE WHEN n>2 AND ABS(r * SQRT( (n-2) * 1.0 / (1 - r*r) )) >= 2 THEN 1 ELSE 0 END AS signif_approx
FROM rstat
ORDER BY category;
```