Group 22

A Data-Driven Playbook

for U.S. Car Retail Investment

Final Report

1. Executive summary

Core Conclusions: A tier-based analysis of 2,527 U.S. counties reveals that new-car dealership density is primarily income-driven but reverses with city size—rural areas often exceed large metros. This finding indicates that market realities are shaped not by national averages but by tier-specific dynamics. New and used-car formats as complements rather than substitutes, and metropolitan markets consistently exhibit a structural "low density × high demand" gap, particularly pronounced for new-car dealerships. In contrast, small-city and rural markets present scalable opportunities for used-car expansion. Using within-tier Z-scores combined with mobility-based demand validation, this study identifies **31 verified** (confirming both scarcity and demand) new-car white markets and **8 verified** used-car white markets, leading to the definition of **3 priority** new-car targets and **3 priority** used-car strategic markets for focused investment planning and execution.

Investment Strategy Principles: The range of average visitors per dealership (Dealer demand intensity; Appx. 1) varies by a factor of 584× in the new-car market and 2,053× in the used-car market, underscoring a pronounced supply–demand imbalance and substantial market opportunity. The guiding principle is to establish a Multi-Criteria Decision Analysis (MCDA) framework tailored to market characteristics to guide format–market alignment, ensuring disciplined and success-driven capital deployment with sustained returns under a structured risk governance framework.

2. Research Framework

2.1 Research Focus

This research develops a **data-driven blueprint for the used-car retail sector**, benchmarked against new-car dealership dynamics. It centers on three strategic research questions: **(1)** key drivers of dealership distribution, **(2)** market structure and white-space analysis, and **(3)** investment strategy and risk governance. Collectively, these analyses deliver actionable insights to guide investment decisions, optimize market entry timing, and enhance portfolio performance.

2.2 Data Sources and Integration

The research leverages multiple tables from the SafeGraph database, particularly integrating key fields from the **visits** table to construct a demand intensity index, which verifies whether supply-short counties genuinely demonstrate strong consumer demand, effectively filtering out false white-space markets where supply and demand remain weak.

2.3 Analytical Principle: Tier-stratified standardization framework (P5)

A unified **P5 framework (Appx. 2.1-2.4)** is adopted to isolate factor effects and ensure cross-regional comparability. The research employes a multi-dimensional normalization strategy: **(P1)** defines counties **(FIPS)** as the fundamental investment unit to capture regulatory and market heterogeneity; **(P2)** measures dealership density per 10K residents for baseline comparability; **(P3)** adds household-based density (per 1K households) to reflect demand driven by household ownership; **(P4)** adjusts for purchasing power using the share of mid- to high-income households; and **(P5)** benchmarks within tiers, large metro, mid-sized, and small-city markets. This multi-level standardization framework ensures both **comparability and sensitivity to market heterogeneity**.

3. Research Focus, Methodology, and Key Business Insights

3.1 How do demographic and economic characteristics shape the spatial distribution of new-car and used-car dealerships across U.S. communities?

3.1.1 Analytical Methods

(1) Stratified Descriptive Statistics Segmented markets into different groups to isolate the independent and interactive effects of income and city size on dealership density. **(2) Tier-Stratified Correlation Analysis** Measured the strength of relationships between dealer density and demographic factors (income, education, ect.) within each metropolitan tier to determine whether predictive power varies by market size. **(3) Multicollinearity Diagnostics Examined** inter-correlations among demographic predictors to identify redundant variables.

(2) Key Findings and Business Implications

The analysis reveals strong structural patterns linking income, city scale, and dealership density **(Appx. 2.1-2.4)**. **Income effect:** New-car density is highly income-driven (0.172 vs. 0.012 per 1k HH; 14.3× difference), while used-car density is less sensitive (0.318 vs. 0.032; 9.9×). **City-size inversion:** Smaller markets show higher density. Tier 4 rural counties (0.182) exceed Tier 1 metros (0.075; 2.43×); used-car density also peaks in Tier 3 (0.339 vs. 0.231; 1.47×). **Education and homeownership** show weak explanatory power. Education ((0.179 vs 0.130; 1.38×) is collinear with income and city effects; homeownership mainly proxies urbanization.

At the interaction level, **income × city** tier creates a Simpson's paradox: national correlations ≈ 0, but within-tier effects reappear, strongest in Tier 4 (14.3× income-driven density gap (high-income vs low-income counties) persists within Tier 4) and weaker in Tier 3 as scale effects dilute income impact. **A supply–demand inversion** is also observed: per-store demand is highest in Tier 1 metros (858 vs 379; 2.27×) despite lower density (0.075 vs 0.182), indicating structural "low-density × high-demand" gaps in metro markets.

Business Action plan: (1) Tier-first screening. Judge markets against their own tier norms and flag materially under-served counties within tier. **(2) New-car strategy.** Lead with under-supplied metros where demand is robust; enter only when the income base is strong and a clear supply gap exists. **(3) Used car strategy.** Roll out broadly across smaller cities and rural markets; screen by supply–demand imbalance, treating income as a supportive (not gating)

factor. **(4) Scorecard.** Keep city tier, income, and demand intensity; drop education and homeownership to avoid noisy, collinear signals.

3.2 Research Question: Do new and used car dealerships exhibit substitute or complementary relationships, and where are areas of undersupply ("white market")?

3.2.1 Analytical Methods

Three methods are applied: (1) **correlation analysis** to assess whether new- and used-car formats compete or complement; (2) **tier-stratified Z-score (Appx. 4) benchmarking** (P5) to isolate true supply gaps while controlling for city-size effects; and (3) **supply–demand validation** combining gap metrics with mobility data to confirm real demand and pinpoint actionable white-space opportunities.

3.2.2 Key Findings and Business Implications

Market Relationship: Analysis confirms that new-car and used-car dealerships **operate as complements**, not substitutes. Nationwide correlation is r = 0.384, strengthening with city size—Tier 1: 0.755; Tier 2: 0.735; Tier 3: 0.538; Tier 4: 0.336—forming a 2.25× gradient that validates the P5 tier framework. Larger markets exhibit stronger co-presence and functional integration, supporting a unified ecosystem perspective.

White-space detection integrates tier-stratified Z-scores with mobility-based demand validation to pinpoint markets that are dealer-light but demand-heavy. Counties flagged as supply-deficient are filtered by population and opportunity tiers to remove low-activity areas, then validated through visitor data to confirm real demand. This process eliminates about **15%** of false white spaces, avoiding markets where low dealer count reflects weak demand rather than unmet potential. The final screening yields **31 new-car** and **8 used-car** verified white-space opportunities, with **84%** and **63%** respectively concentrated in **Tier 3** small metros)—markets with genuine under-supply and proven demand **(Appx. 6)**.

Business Implications: (1) New-car and used-car dealerships exhibit a complementary relationship, sharing customer ecosystems and generating traffic spillovers. In high-correlation markets ($r > 0.7$), integrated operations enable cross-selling opportunities—such as trade-ins and financing—and significantly reduce customer acquisition costs. **(2)** White-space mismatches are most prevalent in Tier 3 counties, average Z-scores (≤ -1) indicate structural under-supply, while visitor data confirm strong, underserved demand. With lower operating costs and weaker competition than Tier 1/2, Tier 3 offers the most scalable and capital-efficient growth opportunity.

3.3 Research Question: Given identified market gaps, how should investors determine optimal markets, dealership formats, and risk levels?

3.3.1 Analytical Methods

(1) Multi-Criteria Decision Analysis (MCDA) (**Appendix 4.)** Developed an optimized scoring system integrating market size (20%), purchasing power (20%), supply gap severity (30%), demand intensity (20%), and market stability (10%) to rank investment opportunities

objectively. (**2) Rule-Based Format Recommendation** (New/Used/Integrated). Weights were determined through initial sensitivity analysis. **(3) Multi-Dimensional Risk Assessment (Appx. 9)** builds a four-dimensional "Confidence–Stability–Purchasing Power–Demand Validation" risk matrix, quantifying white-space investment risk through complementarity (r-value), retention proxy (homeownership), income–format fit, and SafeGraph-based demand certainty.

3.3.2 Key Findings and Business Implications

(1) New-car Top 3 (Appx. 7)

FIPS:12119 (82, Low–Medium risk): standout stability (88.6% homeownership) with a monopoly structure, which is best for certainty-seeking capital; store traffic runs ~31% above median. **FIPS:40131** (82, Medium–High): the only Tier-1 undersupplied market among 2,527 counties; massive scale (1.43M population; 128K high-income HHs) but a short window (≈12–18 months). **FIPS:13223** (82, High): elite affluence (34.0% high-income) but a data anomaly (new-car undersupply while used is balanced); requires of due diligence.

(2) Used-car Top 3 (Appx. 8)

FIPS:72137 (83, Low–Medium): near-ideal fit (77.6% middle-income) customer base and the most severe supply gap (Z = −1.51). **FIPS:40131** (76): dual validation (both new and used score 10/10 on activity) supports an integrated new + CPO + used strategy**. FIPS:22071** (72): heavy competition (17 dealers) but extraordinary demand (traffic ~5× median); low homeownership (49.8%) can be converted via a Value + BHPH play.

3.3.3 Portfolio recommendations and priorities

Priority I Strategy: FIPs: 12119 new-car and 72137 used-car, simple market structure, low–medium risk, fast to stand up as stable anchors.

Priority II Strategy: FIPs: 40131 unique Tier-1 white space and the only market in both Top-2 lists; pursue an integrated path (launch new-car, reserve land, add CPO, then scale used) with rapid corridor scouting to secure sites before rivals.

Priority III (conditional) Strategy: FIPs: 13223 new-car (clear the anomaly via feasibility; then pursue a premium brand thesis) and 22071 used-car (needs competitive study to position a Value + BHPH model).

4. Conclusion

This research analyzed key drivers and market relationships, identified and validated white-space opportunities, and formulated data-driven investment recommendations based solely on SafeGraph data.

To maintain rigor, markets with incomplete data coverage (particularly in the visits table) were excluded from validation. Overall, the proposed framework and principles remain scalable and adaptable for broader datasets and future investment assessments.

Appendix 1 - Comparison Between Used-Car and New-Car dealer demand intensity

Metric	Used-Car Market	New-Car Market	Difference / Ratio
Sample Size	2,027 counties	1,868 counties	Used-car +159 counties
Average Demand Intensity	139.92	477.87	New-car ×3.4
Standard Deviation	111.66	276.93	New-car ×2.5
Coefficient of Variation (CV)	79.80%	58.00%	Used-car +38%
Maximum Value	1,995.67	3,607.11	New-car ×1.8
Minimum Value	0.97	6.17	Used-car lower
Range Ratio (Max / Min)	2,053×	584×	Used-car ×3.5

Appendix 2.1 - P2. Metro Tier (Urban Scale Classification)

Tier	Population Threshold	English Definition	Sample Size	Share	Typical City Examples
Tier 1	≥ 1,000,000	Large Metropolitan	48 counties	1.90%	Los Angeles County, Cook County (Chicago), Harris County (Houston)
Tier 2	250,000 – 999,999	Mid-size Metropolitan	240 counties	9.50%	Phoenix Metro, San Diego County, Denver Metro
Tier 3	50,000 – 249,999	Small Metropolitan	741 counties	29.30%	County-level cities, college towns, regional small cities
Tier 4	< 50,000	Rural	1,498 counties	59.30%	Agricultural counties, remote areas, small towns

Appendix 2.2 - P3. Income Level Classification Four-Category Scheme

Category	Threshold Range	English Definition	Market Characteristics	Typical Vehicle Segment
1 High_Income	≥ 15%	High Income	Affluent markets, strong demand for luxury vehicles	Premium new cars, luxury brands
2 _Mid_High_Income	10% – 15%	Mid-High Income	Upper-middle markets, mainstream new-car segment	Mid- to high-end new cars
3 _Middle_Income	7% – 10%	Middle Income	Balanced market for new and used cars	Mass-market new & used cars

4 _Low_Income	< 7%	Low Income	Price-sensitive markets dominated by used cars	Budget used cars

Data Source: Demographic Statistical Atlas of the United States — "Household Income (United States)" via StatisticalAtlas.com

Appendix 2.3 - P4. Education Level Classification

Category	Threshold Range	English Definition	Market Characteristics	Correlation with Income
1 _High_Edu	≥ 40%	High Education	Knowledge-intensive areas (e.g., Silicon Valley)	Strong positive (r = 0.744***)
2 _Mid_Edu	25% – 40%	Mid Education	College towns, white-collar communities	Positive
3 _Low_Mid_Edu	15% – 25%	Low-Mid Education	Ordinary cities, blue-collar communities	Moderate
4 _Low_Edu	< 15%	Low Education	Agricultural or industrial towns	Low income markets

Appendix 2.4 - P5. Homeownership Rate Classification

Category	Threshold Range	English Definition	Market Meaning	Proxy Implication
1 _High_Homeown	≥ 75%	High Homeownership	Suburban/rural areas, stable residents	High market stability
2 _Mid_Homeown	60% – 75%	Mid Homeownership	Typical urban residential zones	Moderate stability
3 _Low_Mid_Homeown	45% – 60%	Low-Mid Homeownership	Inner cities with high mobility	Lower stability
4 _Low_Homeown	< 45%	Low Homeownership	Rental and student zones in large cities	High mobility

Appendix 3.1 - Dealership Density by Income Group

Income Group	Affluence Rate (%)	Sample Size (Counties)	New-Car Density*	Used-Car Density*	New/Used Ratio
High Income (≥25%)	29.1	229	0.225	0.38	0.59

Mid-High Income (15–25%)	19.1	482	0.183	0.339	0.54
Middle Income (7–15%)	10.4	1,086	0.134	0.289	0.46
Low Income (<7%)	4.6	318	0.016	0.137	0.12

Density unit: dealerships per 1,000 households

New-car density is highly sensitive to income level. Used-car density varies more moderately (2.8× difference).

Appendix 3.2 - Dealership Density by Metropolitan Tier

Tier	Population Scale	Sample Size (Counties)	New-Car Density	Used-Car Density	P3 Density Characteristics
Tier 1 – Large Metro	≥ 1 million	48	0.075	0.231	Lowest
Tier 2 – Mid Metro	250K – 1M	240	0.115	0.29	Low–Medium
Tier 3 – Small Metro	50K – 250K	741	0.164	0.339	Medium–High
Tier 4 – Rural	< 50K	1,498	0.182	0.314	Highest

New-car density exhibits an *"inverted pyramid"* pattern. Smaller markets show higher dealership density per capita.

Appendix Table 3.3 - Cross-Analysis: Metropolitan Tier × Income Group

Tier	Income Group	Sample Size	New-Car Density	Used-Car Density	Income Effect
Tier 1	High Income	48	0.075	0.231	–
Tier 2	High Income	239	0.115	0.29	2.88×
Tier 2	Mid–Low Income	10	0.04	0	(vs. High Income)
Tier 3	High Income	729	0.164	0.339	16.4×
Tier 3	Mid–Low Income	12	0.01	0.003	(vs. High Income)
Tier 4	High Income	1,380	0.182	0.314	5.2×
Tier 4	Mid–Low Income	118	0.035	0.148	(vs. High Income)

Income effect is strongest in Tier 3 small cities (16.4×). In large metros (Tier 1), income variation is masked by city-scale effects. In rural areas (Tier 4), income still exerts a significant influence.

Appendix Table 3.4 - Dealership Density by Education Level

Education Group	Bachelor's Degree Rate (%)	Sample Size (Counties)	New-Car Density	Used-Car Density	Affluence Rate (%)
High Education (≥ 40%)	42.9	214	0.13	0.268	27.98
Mid Education (25–40%)	31.3	857	0.162	0.322	28.06
Low–Mid Education (15–25%)	20.1	1,424	0.179	0.334	19.35
Low Education (< 15%)	11.8	381	0.126	0.244	14.83

Effect size: 1.42× (Low–Mid Education vs. Low Education). Shows an *inverted U-shaped* relationship—nonlinear pattern. High-education counties exhibit lower density due to urban-scale effects. Education–Income correlation: r = 0.744 (strong collinearity).

Appendix Table 3.5 Dealership Density by Homeownership Rate

Homeownership Group	Ownership Rate (%)	Sample Size	P3 New-Car Density**	P4 New-Car Density***	Affluence Rate (%)	Education Rate (%)
Low Homeownership (< 45%)	34.8	732	0.102	0.261	29.36	41.03
Low–Mid Homeownership (45–60%)	55.3	721	0.136	0.244	23.01	29.83
Mid Homeownership (60–75%)	69.3	1,396	0.169	0.244	21.61	23.29
High Homeownership (≥ 75%)	79	288	0.157	0.199	23.7	21.94

Effect size: 1.31× (inverse relationship under P4 metric). P3 and P4 produce opposite orderings in density ranking. Homeownership acts as a proxy for urbanization, not an independent driver.

Appendix 4 - Score Calculation Formulas

Calculation Type	Formula	Variable Definition
Within-Tier Z-Score (PRIMARY)	$Z_{tier} = (X - \mu_{tier}) / \sigma_{tier}$	X = county-level dealership density μ_{tier} = mean within tier σ_{tier} = standard deviation within tier
National Z-Score (REFERENCE)	$Z_{national} = (X - \mu_{national}) / \sigma_{national}$	X = county-level dealership density $\mu_{national}$ = national mean $\sigma_{national}$ = national standard deviation
Reverse-Derived Density Formula	$Density_{threshold} = \mu_{tier} + (Z \times \sigma_{tier})$	Used to calculate density thresholds based on target Z-values

Appendix 5 - Z-Score to Investment Score Conversion Table (MCDA Scoring)

Z-Score Range	Supply Gap Score (Max 30 pts)	Score Ratio	Market Characteristic
Z ≤ −1.5	30 pts	100%	Extreme shortage — top-tier opportunity
−1.5 < Z ≤ −1.3	28 pts	93%	Very strong shortage
−1.3 < Z ≤ −1.2	26 pts	87%	Strong shortage
−1.2 < Z ≤ −1.1	24 pts	80%	Significant shortage
−1.1 < Z ≤ −1.0	22 pts	73%	Moderate shortage (minimum recommendation)
−1.0 < Z ≤ −0.8	18 pts	60%	Mild shortage (caution advised)
−0.8 < Z ≤ −0.6	14 pts	47%	Slight shortage (not recommended)
Z > −0.6	10 pts	33%	Balanced or oversupplied (avoid)

Appendix 6. White Market Comparison by Tier

Dimension	New-Car	Used-Car	Difference
Tier 3 Share	84% (26/31)	62.5% (5/8)	−21.5%
Tier 1/2 Share	16% (5/31)	37.5% (3/8)	21.50%
Tier 4 Share	0%	0%	Same

Appendix 7. top3_new_car_final_v2

Field	Priority #1	Priority #2	Priority #3
Priority Rank	1	2	3
County FIPS	12119	40131	13223
Metro Tier	Tier3_SmallMetro	Tier1_LargeMetro	Tier3_SmallMetro

Total Population	129938	1426790	164440
Total Households	59076	501182	54993
High-Income % (â‰¥$150K)	23.1	25.6	34
Homeownership %	88.6	55.4	77.6
College %	32	28	23.9
Current Dealers	1	3	3
Dealers per 1k HH	0.02	0.01	0.05
Supply Shortage Z-Score	-1.43	-1.1	-1.13
New Supply Status	Moderate_Undersupply	Moderate_Undersupply	Moderate_Undersupply
Undersupply Severity	Severe	Severe	Severe
Demand Intensity (visits/dealer/mo)	956	740	657
Visitors per 1k HH	19	7	33
Avg Visitors per Dealer	1129	1236	602
Activity Score (0“10)	10	10	10
Market Size Score (0“20)	10	20	12

Purchasing Power Score (0“20)	14	16	18
Supply Gap Score (0“30)	28	22	24
Demand Intensity Score (0“20)	20	20	20
Stability Score (0“10)	10	4	8
TOTAL MCDA Score (0“100)	82	82	82
White-Market Confidence	Medium_Confidence	High_Confidence	Not_Applicable (anomaly)
New-Market Quality	True_White_Market	True_White_Market	True_White_Market
Data Quality Flag	OK	OK	OK
Evidence-Based Recommendation (clean)	Severe supply gap with validated demand and high stability; single incumbent indicates capacity constraints.	Only validated Tier1 white market; dual-market demand validated; significant scale with undersupply.	New-car undersupply validated; used-car appears balanced”complementarity anomaly requires investigation.

Investment Decision (clean)	APPROVED “ Immediate Execution	APPROVED “ Expedite Execution	CONDITIONAL APPROVAL “ Pending Feasibility Study
Recommended Next Steps (clean)	30“60 day reconnaissance; assess competition, sites, franchise fit; launch within ~6 months if confirmed.	7“30 day corridor reconnaissance; secure prime sites; design multi-brand integrated strategy; phased rollout.	60“90 day study: verify used-car structure, EV/direct sales impact, consumer behavior; proceed if confirmed.

Appendix 8. top3_used_car_final_v2

Field	Priority #1	Priority #2	Priority #3
Priority Rank	1	2	3
County FIPS	72137	40131	22071
Metro Tier	Tier3_SmallMetro	Tier1_LargeMetro	Tier2_MidMetro
Total Population	79600	1426790	94328
Total Households	27915	501182	15486
High-Income %	77.6	43.1	54.5
Homeownership %	73.5	55.4	48.9
College %		28	17
Current Dealers	1	3	1
Dealers per 1k HH	0.04	0.02	0.11

Supply Shortage Z-Score	-1.51	-1.1	-1.04
Undersupply Severity	Severe_Undersupply	Moderate_Undersupply	Moderate_Undersupply
Demand Intensity (visits/dealer/mo)	216	213	870
Visitors per 1k HH	9	4	96
Avg Visitors per Dealer	241	238	872
Activity Score (0“10)	10	10	10
Market Size Score (0“20)	10	20	16
Purchasing Power Score (0“20)	10	16	12
Supply Gap Score (0“30)	28	24	22
Demand Intensity Score (0“20)	20	20	20
Stability Score (0“10)	7	4	2
Total Score (0“100)	83	76	72

White-Market Confidence	Medium_Confidence	High_Confidence	High_Confidence
New-Market Quality	True_White_Market	True_White_Market	True_White_Market
Data Quality	OK	OK	OK
Evidence Summary	Severe supply shortage validated by demand and mobility data; strong entry case.	Dual-market validation confirms undersupply; only Tier1 metro with proven demand.	Moderate undersupply with strong demand signals and low competition; mid-metro scalability potential.
Investment Decision	APPROVED “ Immediate Execution	APPROVED “ Strategic Entry	APPROVED “ Conditional Entry
Recommended Next Steps	30“60 day local validation; verify site, demand continuity, and franchise readiness.	7“14 day corridor scan; secure prime sites; phase rollout under multi-brand strategy.	60“90 day feasibility validation; confirm sustainability and site economics.

Appendix 7. Dimension Weights, Measurement Indicators, Theoretical Basis, and Scoring Ranges

Dimension	Weight	Measurement Indicator	Theoretical Basis	Scoring Range
Market Size	20%	total_population	Urban Economics – Scale Effect	6–20 points
Purchasing Power	20%	high_income_pct (New Car) / middle_income_pct (Used Car)	Income Elasticity of Demand	4–20 points
Supply Gap	30%	Tier-level Z-score	Supply–Demand Gap Theory	15–30 points
Demand Intensity	20%	market_activity_score (SafeGraph Visitor Data)	Revealed Preference Theory	0–20 points
Market Stability	10%	homeownership_rate	Customer Retention Proxy	2–10 points

Appendix 9 - Risk Scoring Criteria

Dimension	Metric & Data Source	Low Risk (9–10 pts)	Medium Risk (6–8 pts)	High Risk (2–5 pts)	Weight
1. White Space Confidence (Complementarity Match: New vs. Used)	*Step 2.1–2.3; Correlation coefficient (r-value)*	**r > 0.73** – Strong complementarity; Tier 1–2 joint shortage (New + Used both Z < −1.0)	**r = 0.54** – Moderate complementarity; Tier 3 single-side shortage (only one Z < −1.0)	**r < 0.40 or contradictory signal** – Tier 1–2 single-side shortage or inconsistent complementarity	**30%**
2. Market Stability (Homeownership Rate %)	*Step 1.1.4; U.S. Census Data*	**70–80%** – Suburban/small-town, stable family base ("sweet spot")	**60–70% or > 80%** – Moderate stability or aging urban population	**< 60%** – Rental-dominant, high mobility, fast customer churn	**25%**
3. Purchasing Power (Income Composition by Segment)	*Step 1.1.1; Income Tier Data*	**> P75** – New car: > 25% high-income; Used car: > 70% middle-income → ample customer base	**P25–P75** – New car: 15–25%; Used car: 50–70% → moderate purchasing power	**< P25** – New car: < 15%; Used car: < 50% → insufficient purchasing power	**20%**
4. Demand Validation (Traffic Certainty / Demand Intensity Percentile)	*Step 2.5; SafeGraph visits per dealer per month*	**> P75** – Activity 9–10/10; clear demand validation, no data anomalies	**P50–P75** – Activity 7–8/10; moderate demand, data credible	**< P50 or data anomaly** – Activity < 7/10; uncertain demand	**25%**

Group 22

Group 22

Appendix 9 - Query log

Query Log Directory

-- Step 0.1: County-level Master Table (FIXED VERSION)

-- verify: Verify Step 0.1 output

-- Step 0.2: Multi-dimensional Density Metrics Table

-- verify: Verify Step 0.2 output - Check data completeness and ranges

-- verify: View sample records with all density metrics

-- Step 0.3: Tiered Z-Score Standardization Table

-- verify: Verify Step 0.3 output - Check data completeness and z-score distribution

-- verify: Verify tier-level z-score distribution

-- Step 0.4: County-level Visitor Traffic Metrics Table

-- verify: Step 0.4 output - Check data coverage and basic statistics

-- verify: Compare coverage: step0_2 counties vs step0_4 counties with visitor data

-- verify: Check visitor data coverage and demand intensity by metro tier

-- Step 1.1.1: Descriptive Statistics by Income Level (P4 Purchasing Power Perspective)

-- verify: View the complete descriptive statistics by income level

-- verify: Create a comparison view highlighting key differences

-- verify: Compare P2 (population-based) vs P3 (household-based) metrics

-- Step 1.1.2: Descriptive Statistics by Metro Tier × Income Level (P5×P4)

-- verify: Prepare data for heatmap visualization (metro_tier × income_level)

-- Step 1.1.3: Descriptive Statistics by Education Level

-- verify: View complete descriptive statistics by education level

-- verify: Calculate education effect by comparing highest vs lowest education groups

-- verify: education-income correlation

-- Step 1.1.4: Descriptive Statistics by Homeownership Rate

-- verify: View complete descriptive statistics by homeownership rate

-- verify: Calculate percentage difference from lowest homeownership group

-- verify: Calculate homeownership effect by comparing highest vs lowest groups

Group 22

-- verify: Compare P3 (all households) vs P4 (homeowner households) metrics

-- verify: Create correlation matrix: homeownership vs income vs education

-- Step 1.2.1: National-level Correlation Matrix (Enhanced)

-- verify: View complete national-level correlation matrix

-- Step 1.2.2: Tier-stratified Correlation Analysis (Enhanced)

-- verify: View tier-stratified correlation comparison

-- verify: Compare national vs tier-level correlations to validate P5 stratification

-- verify: Check if correlations are statistically significant based on sample size

-- Step 1.3.1: Export Regression Data (Unified)

-- Step 1.2.1 found r(income,education) = 0.744

-- Step 1.1 showed these have high CV and unstable patterns

-- Verify exported regression data

-- Step 2.1: National-level Correlation - New Car vs Used Car

-- verify: View the complete correlation analysis results

-- verify: Compare with RQ1 Step 1.2.1 correlation results

-- Step 2.1 Supplementary: Tier-stratified Correlation

-- verify: View tier-stratified correlation results

-- Step 2.2: Tier-Stratified Correlation Analysis

-- verify: View complete Step 2.2 results with all metrics

-- Step 2.3: Income-Stratified Correlation Analysis

-- verify: View complete Step 2.3 results

-- Step 2.3 Supplementary: Income × Tier Interaction

-- verify: -- Compare P3 (household) vs P2 (population) correlation strength

-- Step 2.4 FINAL: Market Supply-Demand Classification with Tier-Stratified Thresholds

-- verify: Validation 1: Verify all tiers are appropriately classified

-- verify: Validation 2: Key success metrics

-- verify: Validation 3: Top 20 high opportunity counties for new car dealers

-- verify: : Opportunity Level × Confidence Level matrix

-- Step 2.5 REVISED: Integrated White Market Classification with Demand Validation

-- Step 2.6a: Export New Car True White Markets (Demand-Validated)

-- Step 2.6b: Export New Car Potential White Markets (Medium Demand)

-- Step 2.6c: Export Used Car True White Markets

-- Step 2.6d: Summary Statistics for Reporting

-- Step 3.1: Filter High-Quality White Markets for Investment Evaluation

-- Step 3.2: Calculate 5-Criteria MCDA Investment Scores

-- Step 3.2: Calculate 5-Criteria MCDA Investment Scores

-- Step 3.3: Business Format Recommendation and Risk Assessment

```
-- ============================================
-- Step 0.1: County-level Master Table (FIXED VERSION)
-- Purpose: Integrate dealer locations and demographic data at county level
-- Fix: Standardize FIPS codes to ensure successful joins
-- Output: group22-fa25-mgmt58200-final.safegraph2.step0_1_county_master
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step0_1_county_master` AS

WITH
-- CTE 1: Aggregate dealer counts by county FIPS
dealer_counts AS (
  SELECT
    SUBSTR(v.poi_cbg, 1, 5) AS county_fips,
    COUNT(DISTINCT CASE WHEN p.naics_code = 441110 THEN p.safegraph_place_id END) AS new_car_dealers,
```

```
  COUNT(DISTINCT CASE WHEN p.naics_code = 441120 THEN p.safegraph_place_id
END) AS used_car_dealers
 FROM `group22-fa25-mgmt58200-final.safegraph2.places` p
 INNER JOIN `group22-fa25-mgmt58200-final.safegraph2.visits` v
  ON p.safegraph_place_id = v.safegraph_place_id
 WHERE p.naics_code IN (441110, 441120)
  AND v.poi_cbg IS NOT NULL
  AND LENGTH(v.poi_cbg) >= 5
 GROUP BY county_fips
),

-- CTE 2: Aggregate demographic data by county FIPS
demographics_agg AS (
 SELECT
  SUBSTR(cbg, 1, 5) AS county_fips,

  -- P2: Population data
  SUM(pop_total) AS total_population,

  -- P3: Household data
  SUM(inc_total) AS total_households,
  SUM(value_total) AS homeowner_hh,
  SUM(rent_total) AS renter_hh,

  -- P4: Income stratification
  SUM(COALESCE(inc_lt10, 0) + COALESCE(`inc_10-15`, 0) + COALESCE(`inc_15-20`,
0) +
    COALESCE(`inc_20-25`, 0) + COALESCE(`inc_25-30`, 0)) AS low_income_hh,
  SUM(COALESCE(`inc_30-35`, 0) + COALESCE(`inc_35-40`, 0) +
```

```
    COALESCE(`inc_40-45`, 0) + COALESCE(`inc_45-50`, 0)) AS
lower_mid_income_hh,
  SUM(COALESCE(`inc_50-60`, 0) + COALESCE(`inc_60-75`, 0) +
    COALESCE(`inc_75-100`, 0)) AS upper_mid_income_hh,
  SUM(COALESCE(`inc_100-125`, 0) + COALESCE(`inc_125-150`, 0) +
    COALESCE(`inc_150-200`, 0) + COALESCE(inc_gte200, 0)) AS high_income_hh,

  -- Education data
  SUM(COALESCE(edu_coll_bach, 0) + COALESCE(edu_coll_mast, 0) +
    COALESCE(edu_coll_prof, 0) + COALESCE(edu_coll_doc, 0)) AS college_plus,
  SUM(edu_total) AS edu_total

 FROM `group22-fa25-mgmt58200-final.safegraph2.cbg_demographics`
 WHERE cbg IS NOT NULL
  AND LENGTH(cbg) >= 5
 GROUP BY county_fips
),

-- CTE 3: Get county geographic information
county_geo AS (
 SELECT DISTINCT
  county_fips,
  county,
  state
 FROM `group22-fa25-mgmt58200-final.safegraph2.cbg_fips`
 WHERE county_fips IS NOT NULL
)

-- Main query: Join with FULL OUTER JOIN to see what's missing
```

```
SELECT
 -- Geographic identifiers (prefer geo table)
 COALESCE(g.county_fips, demo.county_fips, d.county_fips) AS county_fips,
 g.state,
 g.county,

 -- Dealer counts
 COALESCE(d.new_car_dealers, 0) AS new_car_dealers,
 COALESCE(d.used_car_dealers, 0) AS used_car_dealers,

 -- Population statistics (P2)
 demo.total_population,

 -- Household statistics (P3)
 demo.total_households,
 demo.homeowner_hh,
 demo.renter_hh,

 -- Income stratification (P4)
 demo.low_income_hh,
 demo.lower_mid_income_hh,
 demo.upper_mid_income_hh,
 demo.high_income_hh,

 -- Education data
 demo.college_plus,
 demo.edu_total,
```

```
  -- Derived metrics: Income percentages
  ROUND(SAFE_DIVIDE(demo.high_income_hh, demo.total_households) * 100, 2) AS high_income_pct,
  ROUND(SAFE_DIVIDE(demo.low_income_hh + demo.lower_mid_income_hh, demo.total_households) * 100, 2) AS low_mid_income_pct,

  -- Derived metrics: Homeownership rate
  ROUND(SAFE_DIVIDE(demo.homeowner_hh, demo.total_households) * 100, 2) AS homeownership_rate,

  -- Derived metrics: College education rate
  ROUND(SAFE_DIVIDE(demo.college_plus, demo.edu_total) * 100, 2) AS college_rate,

  -- P5: Metropolitan tier classification
  CASE
    WHEN demo.total_population >= 1000000 THEN 'Tier1_LargeMetro'
    WHEN demo.total_population >= 250000 THEN 'Tier2_MidMetro'
    WHEN demo.total_population >= 50000 THEN 'Tier3_SmallMetro'
    ELSE 'Tier4_Rural'
  END AS metro_tier

FROM demographics_agg demo
FULL OUTER JOIN county_geo g
  ON demo.county_fips = g.county_fips
LEFT JOIN dealer_counts d
  ON COALESCE(g.county_fips, demo.county_fips) = d.county_fips

-- Apply filters: must have demographics data
WHERE demo.total_population IS NOT NULL
  AND demo.total_population >= 10000
```

```
  AND demo.total_households > 0

ORDER BY state, county;



-- verify:  Verify Step 0.1 output
SELECT
  COUNT(*) AS total_counties,
  COUNT(DISTINCT state) AS total_states,
  SUM(new_car_dealers) AS total_new_dealers,
  SUM(used_car_dealers) AS total_used_dealers,
  ROUND(AVG(total_population), 0) AS avg_population
FROM `group22-fa25-mgmt58200-final.safegraph2.step0_1_county_master`;




-- ============================================
-- Step 0.2: Multi-dimensional Density Metrics Table
-- Purpose: Calculate standardized density metrics based on P2/P3/P4 principles
-- Input: step0_1_county_master (all fields)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step0_2_county_density` AS

SELECT
  -- ============================================
  -- Preserve all fields from step0_1_county_master
```

```
-- ============================================
county_fips,
state,
county,
new_car_dealers,
used_car_dealers,
total_population,
total_households,
homeowner_hh,
renter_hh,
low_income_hh,
lower_mid_income_hh,
upper_mid_income_hh,
high_income_hh,
college_plus,
edu_total,
high_income_pct,
low_mid_income_pct,
homeownership_rate,
college_rate,
metro_tier,


-- ============================================
-- P2: Population-based density (auxiliary metrics)
-- ============================================
-- New car dealers per 10,000 population
ROUND(SAFE_DIVIDE(new_car_dealers, total_population) * 10000, 2) AS
new_per_10k_pop,
-- Used car dealers per 10,000 population
```

```
 ROUND(SAFE_DIVIDE(used_car_dealers, total_population) * 10000, 2) AS
used_per_10k_pop,


 -- ============================================

 -- P3: Household-based density ( PRIMARY METRICS)

 -- ============================================

 -- New car dealers per 1,000 households

 ROUND(SAFE_DIVIDE(new_car_dealers, total_households) * 1000, 2) AS
new_per_1k_hh,

 -- Used car dealers per 1,000 households

 ROUND(SAFE_DIVIDE(used_car_dealers, total_households) * 1000, 2) AS
used_per_1k_hh,


 -- ============================================

 -- P4: Purchasing power adjusted density

 -- ============================================

 -- New car dealers per 1,000 high-income households (>$100K)

 ROUND(SAFE_DIVIDE(new_car_dealers, high_income_hh) * 1000, 2) AS
new_per_1k_high_income,


 -- Used car dealers per 1,000 low-to-mid income households (<$50K)

 ROUND(SAFE_DIVIDE(used_car_dealers, low_income_hh + lower_mid_income_hh) *
1000, 2) AS used_per_1k_low_mid_income,


 -- New car dealers per 1,000 homeowner households

 ROUND(SAFE_DIVIDE(new_car_dealers, homeowner_hh) * 1000, 2) AS
new_per_1k_homeowner,


 -- Used car dealers per 1,000 homeowner households

 ROUND(SAFE_DIVIDE(used_car_dealers, homeowner_hh) * 1000, 2) AS
used_per_1k_homeowner
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step0_1_county_master`

-- No filtering needed - keep all counties from step0_1
ORDER BY state, county;



-- verify: Verify Step 0.2 output - Check data completeness and ranges
SELECT
  COUNT(*) AS total_counties,

  -- Check P2 metrics (population-based)
  ROUND(AVG(new_per_10k_pop), 2) AS avg_new_per_10k_pop,
  ROUND(AVG(used_per_10k_pop), 2) AS avg_used_per_10k_pop,

  -- Check P3 metrics (household-based, PRIMARY)
  ROUND(AVG(new_per_1k_hh), 2) AS avg_new_per_1k_hh,
  ROUND(AVG(used_per_1k_hh), 2) AS avg_used_per_1k_hh,

  -- Check P4 metrics (purchasing power adjusted)
  ROUND(AVG(new_per_1k_high_income), 2) AS avg_new_per_1k_high_income,
  ROUND(AVG(used_per_1k_low_mid_income), 2) AS
avg_used_per_1k_low_mid_income,

  -- Check for NULL values (should be minimal)
  COUNTIF(new_per_1k_hh IS NULL) AS null_count_new_hh,
  COUNTIF(used_per_1k_hh IS NULL) AS null_count_used_hh

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density`;
```

```
-- verify:  View sample records with all density metrics
SELECT
 county_fips,
 state,
 county,
 metro_tier,

 -- Original counts
 new_car_dealers,
 used_car_dealers,
 total_population,
 total_households,

 -- P2 metrics
 new_per_10k_pop,
 used_per_10k_pop,

 -- P3 metrics (PRIMARY)
 new_per_1k_hh,
 used_per_1k_hh,

 -- P4 metrics
 new_per_1k_high_income,
 used_per_1k_low_mid_income

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density`
ORDER BY total_population DESC
```

```
LIMIT 20;



-- ============================================
-- Step 0.3: Tiered Z-Score Standardization Table
-- Purpose: Calculate national and tier-based standardized scores based on P3 and P5 principles
-- Input: step0_2_county_density (all fields, focus on new_per_1k_hh, used_per_1k_hh, metro_tier)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore
-- ============================================


CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore` AS


WITH
-- CTE 1: Calculate national-level statistics for P3 household density metrics
national_stats AS (
  SELECT
    -- National average and standard deviation for new car dealer density
    AVG(new_per_1k_hh) AS national_avg_new,
    STDDEV(new_per_1k_hh) AS national_sd_new,
    -- National average and standard deviation for used car dealer density
    AVG(used_per_1k_hh) AS national_avg_used,
    STDDEV(used_per_1k_hh) AS national_sd_used
  FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density`
  WHERE new_per_1k_hh IS NOT NULL
    AND used_per_1k_hh IS NOT NULL
),
```

```
-- CTE 2: Calculate tier-level statistics for P5 stratified analysis
tier_stats AS (
  SELECT
    metro_tier,
    -- Tier-specific average and standard deviation for new car dealer density
    AVG(new_per_1k_hh) AS tier_avg_new,
    STDDEV(new_per_1k_hh) AS tier_sd_new,
    -- Tier-specific average and standard deviation for used car dealer density
    AVG(used_per_1k_hh) AS tier_avg_used,
    STDDEV(used_per_1k_hh) AS tier_sd_used,
    -- Also track sample size per tier for reference
    COUNT(*) AS tier_county_count
  FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density`
  WHERE new_per_1k_hh IS NOT NULL
    AND used_per_1k_hh IS NOT NULL
  GROUP BY metro_tier
)

-- Main query: Join with original data and calculate z-scores
SELECT
  -- ============================================
  -- Preserve all fields from step0_2_county_density
  -- ============================================
  d.county_fips,
  d.state,
  d.county,
  d.new_car_dealers,
  d.used_car_dealers,
```

```
d.total_population,
d.total_households,
d.homeowner_hh,
d.renter_hh,
d.low_income_hh,
d.lower_mid_income_hh,
d.upper_mid_income_hh,
d.high_income_hh,
d.college_plus,
d.edu_total,
d.high_income_pct,
d.low_mid_income_pct,
d.homeownership_rate,
d.college_rate,
d.metro_tier,

-- P2 density metrics
d.new_per_10k_pop,
d.used_per_10k_pop,

-- P3 density metrics (PRIMARY)
d.new_per_1k_hh,
d.used_per_1k_hh,

-- P4 density metrics
d.new_per_1k_high_income,
d.used_per_1k_low_mid_income,
d.new_per_1k_homeowner,
```

```
  d.used_per_1k_homeowner,


  -- ============================================
  -- National Z-scores (compare to national average)
  -- Formula: (value - mean) / standard_deviation
  -- ============================================
  -- New car dealer density z-score vs national benchmark
  ROUND(SAFE_DIVIDE(d.new_per_1k_hh - n.national_avg_new, n.national_sd_new), 3)
AS new_zscore_national,


  -- Used car dealer density z-score vs national benchmark
  ROUND(SAFE_DIVIDE(d.used_per_1k_hh - n.national_avg_used, n.national_sd_used),
3) AS used_zscore_national,


  -- ============================================
  -- P5: Tier-based Z-scores (CORE METRIC - compare within same tier)
  -- This controls for city size effects - large metros vs rural counties
  -- ============================================
  -- New car dealer density z-score vs tier benchmark
  ROUND(SAFE_DIVIDE(d.new_per_1k_hh - t.tier_avg_new, t.tier_sd_new), 3) AS
new_zscore_tier,


  -- Used car dealer density z-score vs tier benchmark
  ROUND(SAFE_DIVIDE(d.used_per_1k_hh - t.tier_avg_used, t.tier_sd_used), 3) AS
used_zscore_tier


FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density` d
CROSS JOIN national_stats n
LEFT JOIN tier_stats t
  ON d.metro_tier = t.metro_tier
```

```
-- Keep all counties from step0_2
ORDER BY d.state, d.county;


-- verify:  Verify Step 0.3 output - Check data completeness and z-score distribution
SELECT
 COUNT(*) AS total_counties,

 -- Check that z-scores are properly centered around 0
 ROUND(AVG(new_zscore_national), 4) AS avg_new_zscore_national,
 ROUND(AVG(used_zscore_national), 4) AS avg_used_zscore_national,
 ROUND(AVG(new_zscore_tier), 4) AS avg_new_zscore_tier,
 ROUND(AVG(used_zscore_tier), 4) AS avg_used_zscore_tier,

 -- Check standard deviation (should be close to 1.0)
 ROUND(STDDEV(new_zscore_national), 3) AS sd_new_zscore_national,
 ROUND(STDDEV(new_zscore_tier), 3) AS sd_new_zscore_tier,

 -- Count undersupplied markets (z-score < -1.0) - potential white markets
 COUNTIF(new_zscore_tier < -1.0) AS undersupplied_new_count,
 COUNTIF(used_zscore_tier < -1.0) AS undersupplied_used_count,

 -- Count severely undersupplied markets (z-score < -1.5)
 COUNTIF(new_zscore_tier < -1.5) AS severely_undersupplied_new,
 COUNTIF(used_zscore_tier < -1.5) AS severely_undersupplied_used,

 -- Count oversupplied markets (z-score > 1.0)
```

```
  COUNTIF(new_zscore_tier > 1.0) AS oversupplied_new_count,
  COUNTIF(used_zscore_tier > 1.0) AS oversupplied_used_count,

  -- Check for NULL values
  COUNTIF(new_zscore_tier IS NULL) AS null_zscore_count

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`;


-- verify: Verify tier-level z-score distribution
SELECT
  metro_tier,
  COUNT(*) AS county_count,

  -- Average z-scores by tier (should be close to 0 within each tier)
  ROUND(AVG(new_zscore_tier), 4) AS avg_new_zscore_tier,
  ROUND(AVG(used_zscore_tier), 4) AS avg_used_zscore_tier,

  -- Count undersupplied markets by tier
  COUNTIF(new_zscore_tier < -1.0) AS undersupplied_new,
  COUNTIF(used_zscore_tier < -1.0) AS undersupplied_used,

  -- Average actual density values
  ROUND(AVG(new_per_1k_hh), 3) AS avg_new_density,
  ROUND(AVG(used_per_1k_hh), 3) AS avg_used_density

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
GROUP BY metro_tier
```

```
ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;



 -- ============================================
-- Step 0.4: County-level Visitor Traffic Metrics Table
-- Purpose: Integrate visits data to calculate visitor traffic and demand intensity metrics
-- This provides demand-side validation for RQ2 and RQ3
-- Input: visits + places + step0_2_county_density
-- Output: group22-fa25-mgmt58200-final.safegraph2.step0_4_county_visitor_metrics
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step0_4_county_visitor_metrics` AS

WITH
-- CTE 1: Join visits and places, filter for car dealerships in Jan-Apr 2020
visits_with_naics AS (
 SELECT
  v.safegraph_place_id,
  v.poi_cbg,
```

```
  v.raw_visitor_counts,
  v.median_dwell,
  p.naics_code
 FROM `group22-fa25-mgmt58200-final.safegraph2.visits` v
 INNER JOIN `group22-fa25-mgmt58200-final.safegraph2.places` p
  ON v.safegraph_place_id = p.safegraph_place_id
 WHERE p.naics_code IN (441110, 441120)  -- Only new and used car dealerships
  AND v.poi_cbg IS NOT NULL
  AND LENGTH(v.poi_cbg) >= 5
  AND EXTRACT(YEAR FROM v.date_range_start) = 2020
  AND EXTRACT(MONTH FROM v.date_range_start) BETWEEN 1 AND 4  -- Jan-Apr 2020
),

-- CTE 2: Aggregate visitor metrics by county FIPS and dealer type
visitor_aggregates AS (
 SELECT
  SUBSTR(poi_cbg, 1, 5) AS county_fips,

  -- New car dealer (NAICS 441110) visitor metrics
  SUM(CASE WHEN naics_code = 441110 THEN raw_visitor_counts ELSE 0 END) AS
new_total_visitors,
  COUNT(CASE WHEN naics_code = 441110 THEN 1 END) AS new_visit_records,
  AVG(CASE WHEN naics_code = 441110 THEN median_dwell END) AS new_avg_dwell,

  -- Used car dealer (NAICS 441120) visitor metrics
  SUM(CASE WHEN naics_code = 441120 THEN raw_visitor_counts ELSE 0 END) AS
used_total_visitors,
  COUNT(CASE WHEN naics_code = 441120 THEN 1 END) AS used_visit_records,
  AVG(CASE WHEN naics_code = 441120 THEN median_dwell END) AS used_avg_dwell
```

```
 FROM visits_with_naics
 GROUP BY county_fips
)

-- Main query: Join with step0_2_county_density and calculate demand metrics
SELECT
 va.county_fips,

 -- ============================================
 -- Raw visitor data (aggregated from visits table)
 -- ============================================
 va.new_total_visitors,
 va.used_total_visitors,
 va.new_visit_records,
 va.used_visit_records,
 ROUND(va.new_avg_dwell, 2) AS new_avg_dwell,
 ROUND(va.used_avg_dwell, 2) AS used_avg_dwell,

 -- ============================================
 -- P3: Standardized visitor density (per 1,000 households)
 -- Normalizes visitors by market size
 -- ============================================
 ROUND(SAFE_DIVIDE(va.new_total_visitors, d.total_households) * 1000, 2) AS
new_visitors_per_1k_hh,
 ROUND(SAFE_DIVIDE(va.used_total_visitors, d.total_households) * 1000, 2) AS
used_visitors_per_1k_hh,

 -- ============================================
```

```
 -- Average visitors per dealer
 -- Shows how busy each dealer is on average
 -- ============================================
 ROUND(SAFE_DIVIDE(va.new_total_visitors, d.new_car_dealers), 2) AS
new_avg_visitors_per_dealer,
 ROUND(SAFE_DIVIDE(va.used_total_visitors, d.used_car_dealers), 2) AS
used_avg_visitors_per_dealer,


 -- ============================================
 -- ⭐ Demand Intensity Index (CORE METRIC)
 -- Formula: (visitors per 1k households) / (dealers per 1k households)
 -- High value = strong demand (few dealers serving many visitors = supply shortage)
 -- Low value = weak demand (many dealers but few visitors = oversupply)
 -- ============================================
 ROUND(SAFE_DIVIDE(
  SAFE_DIVIDE(va.new_total_visitors, d.total_households) * 1000,
  d.new_per_1k_hh
 ), 3) AS demand_intensity_new,

 ROUND(SAFE_DIVIDE(
  SAFE_DIVIDE(va.used_total_visitors, d.total_households) * 1000,
  d.used_per_1k_hh
 ), 3) AS demand_intensity_used,

 -- ============================================
 -- Data quality indicators
 -- Coverage = avg months of data per dealer (ideal = 1.0 for 4 months)
 -- ============================================
```

```
 ROUND(SAFE_DIVIDE(va.new_visit_records / 4.0, d.new_car_dealers), 2) AS
new_data_coverage,
 ROUND(SAFE_DIVIDE(va.used_visit_records / 4.0, d.used_car_dealers), 2) AS
used_data_coverage

FROM visitor_aggregates va
INNER JOIN `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density` d
 ON va.county_fips = d.county_fips

-- Keep only counties with at least one dealer
WHERE d.new_car_dealers > 0 OR d.used_car_dealers > 0

ORDER BY va.county_fips;



-- verify:  Step 0.4 output - Check data coverage and basic statistics
SELECT
 -- Total counties with visitor data
 COUNT(*) AS counties_with_visitor_data,

 -- Average visitor counts
 ROUND(AVG(new_total_visitors), 0) AS avg_new_visitors,
 ROUND(AVG(used_total_visitors), 0) AS avg_used_visitors,

 -- Average visitors per dealer
 ROUND(AVG(new_avg_visitors_per_dealer), 0) AS avg_new_visitors_per_dealer,
 ROUND(AVG(used_avg_visitors_per_dealer), 0) AS avg_used_visitors_per_dealer,

 -- ⭐ Average demand intensity (CORE METRIC)
```

```
 ROUND(AVG(demand_intensity_new), 2) AS avg_demand_intensity_new,
 ROUND(AVG(demand_intensity_used), 2) AS avg_demand_intensity_used,
 ROUND(STDDEV(demand_intensity_new), 2) AS sd_demand_intensity_new,

 -- Data quality check
 ROUND(AVG(new_data_coverage), 2) AS avg_new_data_coverage,
 ROUND(AVG(used_data_coverage), 2) AS avg_used_data_coverage,

 -- Check for missing data
 COUNTIF(new_total_visitors = 0) AS counties_no_new_visitors,
 COUNTIF(used_total_visitors = 0) AS counties_no_used_visitors,
 COUNTIF(demand_intensity_new IS NULL) AS null_demand_intensity_new,
 COUNTIF(demand_intensity_used IS NULL) AS null_demand_intensity_used,

 -- Extreme values check
 MAX(demand_intensity_new) AS max_demand_intensity_new,
 MIN(demand_intensity_new) AS min_demand_intensity_new

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_4_county_visitor_metrics`;


-- verify: Compare coverage: step0_2 counties vs step0_4 counties with visitor data
WITH coverage_stats AS (
 SELECT
  (SELECT COUNT(*) FROM `group22-fa25-mgmt58200-
final.safegraph2.step0_2_county_density`) AS total_counties_step0_2,
  (SELECT COUNT(*) FROM `group22-fa25-mgmt58200-
final.safegraph2.step0_4_county_visitor_metrics`) AS counties_with_visitors_step0_4
)
```

```
SELECT
 total_counties_step0_2,
 counties_with_visitors_step0_4,
 ROUND(counties_with_visitors_step0_4 * 100.0 / total_counties_step0_2, 1) AS
coverage_percentage,
 total_counties_step0_2 - counties_with_visitors_step0_4 AS
counties_missing_visitor_data
FROM coverage_stats;



-- verify: Check visitor data coverage and demand intensity by metro tier
SELECT
 d.metro_tier,
 COUNT(DISTINCT d.county_fips) AS total_counties_in_tier,
 COUNT(DISTINCT vm.county_fips) AS counties_with_visitor_data,
 ROUND(COUNT(DISTINCT vm.county_fips) * 100.0 / COUNT(DISTINCT d.county_fips),
1) AS coverage_pct,

 -- Average demand intensity by tier
 ROUND(AVG(vm.demand_intensity_new), 2) AS avg_demand_intensity_new,
 ROUND(AVG(vm.demand_intensity_used), 2) AS avg_demand_intensity_used,

 -- Average visitors per dealer by tier
 ROUND(AVG(vm.new_avg_visitors_per_dealer), 0) AS avg_visitors_per_new_dealer,
 ROUND(AVG(vm.used_avg_visitors_per_dealer), 0) AS avg_visitors_per_used_dealer

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_2_county_density` d
LEFT JOIN `group22-fa25-mgmt58200-
final.safegraph2.step0_4_county_visitor_metrics` vm
 ON d.county_fips = vm.county_fips
```

```
GROUP BY d.metro_tier
ORDER BY
 CASE d.metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;



 -- ============================================
-- Step 1.1.1: Descriptive Statistics by Income Level (P4 Purchasing Power Perspective)
-- Purpose: Compare dealer density differences across counties with different income levels
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_1_1_desc_by_income
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step1_1_1_desc_by_income` AS

SELECT
 -- ============================================
 -- Income category classification (P4 principle)
 -- ============================================
 CASE
  WHEN high_income_pct >= 15 THEN '1_High_Income'
  WHEN high_income_pct >= 10 THEN '2_Mid_High_Income'
  WHEN high_income_pct >= 7 THEN '3_Middle_Income'
```

```
  ELSE '4_Low_Income'
 END AS income_category,

 -- ============================================
 -- Sample size and population
 -- ============================================
 COUNT(*) AS n_counties,
 ROUND(SUM(total_population) / 1000000.0, 2) AS total_pop_millions,

 -- ============================================
 -- P3: Household density metrics (PRIMARY - average and standard deviation)
 -- ============================================
 -- New car dealer density
 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_per_1k_hh,
 ROUND(STDDEV(new_per_1k_hh), 3) AS sd_new,

 -- Used car dealer density
 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_per_1k_hh,
 ROUND(STDDEV(used_per_1k_hh), 3) AS sd_used,

 -- ============================================
 -- New vs Used dealer ratio
 -- ============================================
  ROUND(SUM(new_car_dealers) * 1.0 / NULLIF(SUM(used_car_dealers), 0), 3) AS
new_to_used_ratio,

 -- ============================================
 -- P2: Population density metrics (AUXILIARY - for comparison)
 -- ============================================
```

```
 ROUND(AVG(new_per_10k_pop), 3) AS avg_new_per_10k_pop,
 ROUND(AVG(used_per_10k_pop), 3) AS avg_used_per_10k_pop

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`

-- Group by income category
GROUP BY income_category

-- Order from highest to lowest income
ORDER BY income_category;

-- verify: View the complete descriptive statistics by income level
SELECT
 income_category,
 n_counties,
 total_pop_millions,

 -- P3 household density (primary metrics)
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,
 sd_new,
 sd_used,

 -- Dealer ratio
 new_to_used_ratio,

 -- P2 population density (comparison)
 avg_new_per_10k_pop,
```

```
  avg_used_per_10k_pop,


  -- Calculate percentage difference between highest and lowest income groups
  ROUND((avg_new_per_1k_hh - FIRST_VALUE(avg_new_per_1k_hh) OVER (ORDER BY
income_category DESC)) /
      FIRST_VALUE(avg_new_per_1k_hh) OVER (ORDER BY income_category DESC) * 100,
1) AS pct_diff_from_lowest


FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_1_desc_by_income`
ORDER BY income_category;



-- verify: Create a comparison view highlighting key differences
WITH stats AS (
  SELECT
    income_category,
    n_counties,
    avg_new_per_1k_hh,
    avg_used_per_1k_hh,
    new_to_used_ratio
  FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_1_desc_by_income`
),
high_income AS (
  SELECT avg_new_per_1k_hh AS high_new, avg_used_per_1k_hh AS high_used
  FROM stats WHERE income_category = '1_High_Income'
),
low_income AS (
  SELECT avg_new_per_1k_hh AS low_new, avg_used_per_1k_hh AS low_used
  FROM stats WHERE income_category = '4_Low_Income'
```

```
)

SELECT
 s.income_category,
 s.n_counties,
 s.avg_new_per_1k_hh,
 s.avg_used_per_1k_hh,
 s.new_to_used_ratio,

 -- Calculate ratio compared to low income counties
 ROUND(s.avg_new_per_1k_hh / l.low_new, 2) AS new_density_multiple_vs_low,
 ROUND(s.avg_used_per_1k_hh / l.low_used, 2) AS used_density_multiple_vs_low,

 -- Show absolute difference from high income counties
 ROUND(s.avg_new_per_1k_hh - h.high_new, 3) AS new_diff_from_high,
 ROUND(s.avg_used_per_1k_hh - h.high_used, 3) AS used_diff_from_high

FROM stats s
CROSS JOIN high_income h
CROSS JOIN low_income l
ORDER BY s.income_category;



-- verify: Compare P2 (population-based) vs P3 (household-based) metrics
-- This validates whether household density is more relevant
SELECT
 income_category,
```

```
 -- P3 metrics (household-based)
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,

 -- P2 metrics (population-based)
 avg_new_per_10k_pop,
 avg_used_per_10k_pop,

 -- Calculate coefficient of variation (SD/Mean) for P3
 ROUND(sd_new / NULLIF(avg_new_per_1k_hh, 0), 2) AS cv_new_p3,
 ROUND(sd_used / NULLIF(avg_used_per_1k_hh, 0), 2) AS cv_used_p3,

 -- Show density ratio pattern
 ROUND(avg_new_per_1k_hh / NULLIF(avg_used_per_1k_hh, 0), 2) AS
new_to_used_density_ratio_p3

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_1_desc_by_income`
ORDER BY income_category;



-- ============================================
-- Step 1.1.2: Descriptive Statistics by Metro Tier × Income Level (P5×P4)
-- Purpose: Test if income effects differ across city sizes (stratification principle)
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_1_2_desc_by_tier_income
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step1_1_2_desc_by_tier_income` AS
```

```
SELECT
 -- ============================================
 -- P5: Metro tier classification
 -- ============================================
 metro_tier,

 -- ============================================
 -- P4: Income level (binary classification for cross-tabulation)
 -- ============================================
 CASE
  WHEN high_income_pct >= 10 THEN 'High_Income'
  ELSE 'Low_Mid_Income'
 END AS income_level,

 -- ============================================
 -- Sample size
 -- ============================================
 COUNT(*) AS n_counties,

 -- ============================================
 -- P3: Household density metrics (average per 1,000 households)
 -- ============================================
 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_density,
 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_density,

 -- ============================================
 -- P4: Purchasing power adjusted density
```

```
 -- New car dealers per 1,000 high-income households
 -- ============================================
 ROUND(AVG(new_per_1k_high_income), 3) AS avg_new_per_affluent,

 -- ============================================
 -- P4: Purchasing power adjusted density
 -- Used car dealers per 1,000 low-to-mid income households
 -- ============================================
 ROUND(AVG(used_per_1k_low_mid_income), 3) AS avg_used_per_mid_low

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`

-- Group by both dimensions
GROUP BY metro_tier, income_level

-- Order by tier and income level for readability
ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
    WHEN 'Tier3_SmallMetro' THEN 3
    WHEN 'Tier4_Rural' THEN 4
  END,
  income_level DESC;


 -- View complete cross-tabulation results
SELECT
```

```
 metro_tier,
 income_level,
 n_counties,

 -- P3 metrics
 avg_new_density,
 avg_used_density,

 -- P4 metrics
 avg_new_per_affluent,
 avg_used_per_mid_low,

 -- Calculate density ratio (new/used)
 ROUND(avg_new_density / NULLIF(avg_used_density, 0), 2) AS new_to_used_ratio

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_2_desc_by_tier_income`
ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END,
 income_level DESC;


 -- Calculate income effect (High vs Low) within each metro tier
WITH cross_tab AS (
```

```
 SELECT
  metro_tier,
  income_level,
  n_counties,
  avg_new_density,
  avg_used_density
 FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_2_desc_by_tier_income`
)

SELECT
 h.metro_tier,

 -- Sample sizes
 h.n_counties AS high_income_counties,
 l.n_counties AS low_mid_income_counties,

 -- Density values
 h.avg_new_density AS high_income_new_density,
 l.avg_new_density AS low_mid_income_new_density,

 -- ⭐ Income effect on new car density (High / Low ratio)
 ROUND(h.avg_new_density / NULLIF(l.avg_new_density, 0), 2) AS
new_density_income_effect,

 -- Absolute difference
 ROUND(h.avg_new_density - l.avg_new_density, 3) AS new_density_abs_diff,

 -- Used car comparison
 h.avg_used_density AS high_income_used_density,
```

```
 l.avg_used_density AS low_mid_income_used_density,


 -- ⭐ Income effect on used car density
 ROUND(h.avg_used_density / NULLIF(l.avg_used_density, 0), 2) AS
used_density_income_effect


FROM cross_tab h
INNER JOIN cross_tab l
 ON h.metro_tier = l.metro_tier
 AND h.income_level = 'High_Income'
 AND l.income_level = 'Low_Mid_Income'


ORDER BY
 CASE h.metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;



 -- Analyze tier effects: Does density decrease with tier size?
WITH tier_stats AS (
 SELECT
  metro_tier,
  income_level,
  avg_new_density,
  avg_used_density,
  n_counties
```

```
 FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_2_desc_by_tier_income`
)

SELECT
 income_level,

 -- Tier1 (Large Metro)
 MAX(CASE WHEN metro_tier = 'Tier1_LargeMetro' THEN avg_new_density END) AS tier1_new,
 MAX(CASE WHEN metro_tier = 'Tier1_LargeMetro' THEN avg_used_density END) AS tier1_used,

 -- Tier2 (Mid Metro)
 MAX(CASE WHEN metro_tier = 'Tier2_MidMetro' THEN avg_new_density END) AS tier2_new,
 MAX(CASE WHEN metro_tier = 'Tier2_MidMetro' THEN avg_used_density END) AS tier2_used,

 -- Tier3 (Small Metro)
 MAX(CASE WHEN metro_tier = 'Tier3_SmallMetro' THEN avg_new_density END) AS tier3_new,
 MAX(CASE WHEN metro_tier = 'Tier3_SmallMetro' THEN avg_used_density END) AS tier3_used,

 -- Tier4 (Rural)
 MAX(CASE WHEN metro_tier = 'Tier4_Rural' THEN avg_new_density END) AS tier4_new,
 MAX(CASE WHEN metro_tier = 'Tier4_Rural' THEN avg_used_density END) AS tier4_used,

 -- ⭐ Calculate tier gradient (Tier1 / Tier4 ratio)
```

```
  ROUND(
    MAX(CASE WHEN metro_tier = 'Tier1_LargeMetro' THEN avg_new_density END) /
    NULLIF(MAX(CASE WHEN metro_tier = 'Tier4_Rural' THEN avg_new_density END), 0)
  , 2) AS tier_gradient_new

FROM tier_stats
GROUP BY income_level
ORDER BY income_level DESC;

-- verify:  Prepare data for heatmap visualization (metro_tier × income_level)
SELECT
  metro_tier,
  income_level,
  n_counties,
  avg_new_density AS new_car_density,
  avg_used_density AS used_car_density,

  -- Normalize to 0-100 scale for heatmap coloring
  ROUND((avg_new_density - MIN(avg_new_density) OVER ()) /
      NULLIF(MAX(avg_new_density) OVER () - MIN(avg_new_density) OVER (), 0) * 100, 1)
    AS new_density_normalized,

  ROUND((avg_used_density - MIN(avg_used_density) OVER ()) /
      NULLIF(MAX(avg_used_density) OVER () - MIN(avg_used_density) OVER (), 0) * 100,
1)
    AS used_density_normalized

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_2_desc_by_tier_income`
ORDER BY
```

```
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END,
 income_level DESC;



 -- ============================================
-- Step 1.1.3: Descriptive Statistics by Education Level
-- Purpose: Verify the relationship between education and dealer distribution
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_1_3_desc_by_education
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step1_1_3_desc_by_education` AS

SELECT
 -- ============================================
 -- Education category classification (college degree completion rate)
 -- ============================================
 CASE
  WHEN college_rate >= 40 THEN '1_High_Edu'
  WHEN college_rate >= 25 THEN '2_Mid_Edu'
  WHEN college_rate >= 15 THEN '3_Low_Mid_Edu'
  ELSE '4_Low_Edu'
 END AS education_category,
```

```
-- ============================================
-- Sample size
-- ============================================
COUNT(*) AS n_counties,

-- ============================================
-- P3: Household density metrics (average per 1,000 households)
-- ============================================
ROUND(AVG(new_per_1k_hh), 3) AS avg_new_per_1k_hh,
ROUND(AVG(used_per_1k_hh), 3) AS avg_used_per_1k_hh,

-- Standard deviation for variability assessment
ROUND(STDDEV(new_per_1k_hh), 3) AS sd_new,
ROUND(STDDEV(used_per_1k_hh), 3) AS sd_used,

-- ============================================
-- Average affluent percentage (to check education-income correlation)
-- ============================================
ROUND(AVG(high_income_pct), 2) AS avg_affluent_pct,

-- ============================================
-- New to used dealer ratio
-- ============================================
ROUND(AVG(new_per_1k_hh) / NULLIF(AVG(used_per_1k_hh), 0), 3) AS
new_to_used_ratio

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
```

```
-- Group by education category
GROUP BY education_category

-- Order from highest to lowest education
ORDER BY education_category;



-- verify:  View complete descriptive statistics by education level
SELECT
 education_category,
 n_counties,

 -- Density metrics
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,
 sd_new,
 sd_used,

 -- Income correlation
 avg_affluent_pct,

 -- Dealer type preference
 new_to_used_ratio,

 -- Calculate percentage difference from lowest education group
 ROUND((avg_new_per_1k_hh - FIRST_VALUE(avg_new_per_1k_hh) OVER (ORDER BY
education_category DESC)) /
     FIRST_VALUE(avg_new_per_1k_hh) OVER (ORDER BY education_category DESC) *
100, 1) AS pct_diff_from_lowest_edu
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_3_desc_by_education`
ORDER BY education_category;


-- verify: Calculate education effect by comparing highest vs lowest education groups
WITH edu_stats AS (
  SELECT
    education_category,
    n_counties,
    avg_new_per_1k_hh,
    avg_used_per_1k_hh,
    avg_affluent_pct
  FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_3_desc_by_education`
),
high_edu AS (
  SELECT avg_new_per_1k_hh AS high_new, avg_used_per_1k_hh AS high_used,
avg_affluent_pct AS high_income
  FROM edu_stats WHERE education_category = '1_High_Edu'
),
low_edu AS (
  SELECT avg_new_per_1k_hh AS low_new, avg_used_per_1k_hh AS low_used,
avg_affluent_pct AS low_income
  FROM edu_stats WHERE education_category = '4_Low_Edu'
)

SELECT
  e.education_category,
  e.n_counties,
```

```
 e.avg_new_per_1k_hh,
 e.avg_used_per_1k_hh,
 e.avg_affluent_pct,

 --  Calculate density multiple compared to low education counties
 ROUND(e.avg_new_per_1k_hh / l.low_new, 2) AS new_density_multiple_vs_low_edu,
 ROUND(e.avg_used_per_1k_hh / l.low_used, 2) AS
used_density_multiple_vs_low_edu,

 -- Show absolute difference from high education counties
 ROUND(e.avg_new_per_1k_hh - h.high_new, 3) AS new_diff_from_high_edu,
 ROUND(e.avg_used_per_1k_hh - h.high_used, 3) AS used_diff_from_high_edu,

 -- Income correlation
 ROUND(e.avg_affluent_pct - l.low_income, 2) AS income_diff_from_low_edu

FROM edu_stats e
CROSS JOIN high_edu h
CROSS JOIN low_edu l
ORDER BY e.education_category;



-- verify:  education-income correlation
SELECT
 education_category,
 n_counties,
 avg_affluent_pct,

 -- Density metrics
```

```
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,

 -- Check if education and income are correlated
 CASE
  WHEN avg_affluent_pct >= 15 THEN 'High_Income'
  WHEN avg_affluent_pct >= 10 THEN 'Mid_High_Income'
  WHEN avg_affluent_pct >= 7 THEN 'Mid_Income'
  ELSE 'Low_Income'
 END AS implied_income_level,

 -- Coefficient of variation
 ROUND(sd_new / NULLIF(avg_new_per_1k_hh, 0), 2) AS cv_new,
 ROUND(sd_used / NULLIF(avg_used_per_1k_hh, 0), 2) AS cv_used

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_3_desc_by_education`
ORDER BY education_category;

-- ============================================
-- Step 1.1.4: Descriptive Statistics by Homeownership Rate
-- Purpose: Verify the relationship between homeownership and dealer distribution
-- Hypothesis: Counties with higher homeownership have higher dealer density
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-
final.safegraph2.step1_1_4_desc_by_homeownership
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step1_1_4_desc_by_homeownership` AS
```

```
SELECT
 -- ============================================
 -- Homeownership category classification
 -- ============================================
 CASE
  WHEN homeownership_rate >= 75 THEN '1_High_Homeown'
  WHEN homeownership_rate >= 60 THEN '2_Mid_Homeown'
  WHEN homeownership_rate >= 45 THEN '3_Low_Mid_Homeown'
  ELSE '4_Low_Homeown'
 END AS homeownership_category,

 -- ============================================
 -- Sample size
 -- ============================================
 COUNT(*) AS n_counties,

 -- ============================================
 -- Average homeownership rate for reference
 -- ============================================
 ROUND(AVG(homeownership_rate), 2) AS avg_homeownership_rate,

 -- ============================================
 -- P3: Standard household density (for comparison)
 -- ============================================
 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_per_1k_hh,
 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_per_1k_hh,
```

```
-- ============================================
-- P4: Homeowner-adjusted density (PRIMARY METRIC)
-- Dealers per 1,000 homeowner households
-- ============================================
ROUND(AVG(new_per_1k_homeowner), 3) AS avg_new_per_1k_homeowner,
ROUND(AVG(used_per_1k_homeowner), 3) AS avg_used_per_1k_homeowner,

-- Standard deviation for variability assessment
ROUND(STDDEV(new_per_1k_homeowner), 3) AS sd_new_homeowner,
ROUND(STDDEV(used_per_1k_homeowner), 3) AS sd_used_homeowner,

-- ============================================
-- Correlation check: income and education levels
-- ============================================
ROUND(AVG(high_income_pct), 2) AS avg_affluent_pct,
ROUND(AVG(college_rate), 2) AS avg_college_rate

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`

-- Group by homeownership category
GROUP BY homeownership_category

-- Order from highest to lowest homeownership
ORDER BY homeownership_category;


-- verify:  View complete descriptive statistics by homeownership rate
SELECT
```

```
 homeownership_category,
 n_counties,
 avg_homeownership_rate,

 -- P3 standard density (all households)
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,

 -- ⭐ P4 homeowner-adjusted density (PRIMARY)
 avg_new_per_1k_homeowner,
 avg_used_per_1k_homeowner,
 sd_new_homeowner,
 sd_used_homeowner,

 -- Correlation with other factors
 avg_affluent_pct,
 avg_college_rate,

-- verify: Calculate percentage difference from lowest homeownership group
 ROUND((avg_new_per_1k_homeowner - FIRST_VALUE(avg_new_per_1k_homeowner)
OVER (ORDER BY homeownership_category DESC)) /
     FIRST_VALUE(avg_new_per_1k_homeowner) OVER (ORDER BY
homeownership_category DESC) * 100, 1)
   AS pct_diff_from_lowest_homeown

FROM `group22-fa25-mgmt58200-
final.safegraph2.step1_1_4_desc_by_homeownership`
ORDER BY homeownership_category;
```

```
-- verify: Calculate homeownership effect by comparing highest vs lowest groups
WITH homeown_stats AS (
  SELECT
    homeownership_category,
    n_counties,
    avg_homeownership_rate,
    avg_new_per_1k_homeowner,
    avg_used_per_1k_homeowner,
    avg_affluent_pct,
    avg_college_rate
  FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_4_desc_by_homeownership`
),
high_homeown AS (
  SELECT
    avg_new_per_1k_homeowner AS high_new,
    avg_used_per_1k_homeowner AS high_used,
    avg_affluent_pct AS high_income,
    avg_college_rate AS high_edu
  FROM homeown_stats
  WHERE homeownership_category = '1_High_Homeown'
),
low_homeown AS (
  SELECT
    avg_new_per_1k_homeowner AS low_new,
    avg_used_per_1k_homeowner AS low_used,
    avg_affluent_pct AS low_income,
    avg_college_rate AS low_edu
  FROM homeown_stats
```

```
  WHERE homeownership_category = '4_Low_Homeown'
)

SELECT
  h.homeownership_category,
  h.n_counties,
  h.avg_homeownership_rate,
  h.avg_new_per_1k_homeowner,
  h.avg_used_per_1k_homeowner,

  -- ⭐ Calculate density multiple compared to low homeownership counties
  ROUND(h.avg_new_per_1k_homeowner / l.low_new, 2) AS
new_density_multiple_vs_low_homeown,
  ROUND(h.avg_used_per_1k_homeowner / l.low_used, 2) AS
used_density_multiple_vs_low_homeown,

  -- Show absolute difference from high homeownership counties
  ROUND(h.avg_new_per_1k_homeowner - hi.high_new, 3) AS
new_diff_from_high_homeown,
  ROUND(h.avg_used_per_1k_homeowner - hi.high_used, 3) AS
used_diff_from_high_homeown,

  -- Check correlation with income and education
  h.avg_affluent_pct,
  h.avg_college_rate,
  ROUND(h.avg_affluent_pct - l.low_income, 2) AS income_diff_from_low_homeown,
  ROUND(h.avg_college_rate - l.low_edu, 2) AS edu_diff_from_low_homeown

FROM homeown_stats h
CROSS JOIN high_homeown hi
```

```
CROSS JOIN low_homeown l
ORDER BY h.homeownership_category;

-- verify: Compare P3 (all households) vs P4 (homeowner households) metrics
-- This shows if homeowner-adjustment changes the patterns
SELECT
 homeownership_category,
 n_counties,
 avg_homeownership_rate,

 -- P3: All households basis
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,

 -- P4: Homeowner households basis
 avg_new_per_1k_homeowner,
 avg_used_per_1k_homeowner,

 -- ⭐ Calculate adjustment factor (P4/P3 ratio)
 ROUND(avg_new_per_1k_homeowner / NULLIF(avg_new_per_1k_hh, 0), 2) AS
new_adjustment_factor,
 ROUND(avg_used_per_1k_homeowner / NULLIF(avg_used_per_1k_hh, 0), 2) AS
used_adjustment_factor,

 -- Coefficient of variation
 ROUND(sd_new_homeowner / NULLIF(avg_new_per_1k_homeowner, 0), 2) AS cv_new,
 ROUND(sd_used_homeowner / NULLIF(avg_used_per_1k_homeowner, 0), 2) AS
cv_used
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_4_desc_by_homeownership`
ORDER BY homeownership_category;


-- verify: Create correlation matrix: homeownership vs income vs education
SELECT
 homeownership_category,
 n_counties,

 -- Core metrics (normalized to 0-100 for comparison)
 ROUND(avg_homeownership_rate, 1) AS homeownership_pct,
 ROUND(avg_affluent_pct, 1) AS income_pct,
 ROUND(avg_college_rate, 1) AS education_pct,

 -- Density metrics
 avg_new_per_1k_homeowner,
 avg_used_per_1k_homeowner,

 -- Rank by each dimension
 RANK() OVER (ORDER BY avg_homeownership_rate DESC) AS homeown_rank,
 RANK() OVER (ORDER BY avg_affluent_pct DESC) AS income_rank,
 RANK() OVER (ORDER BY avg_college_rate DESC) AS edu_rank,
 RANK() OVER (ORDER BY avg_new_per_1k_homeowner DESC) AS density_rank

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_1_4_desc_by_homeownership`
ORDER BY homeownership_category;
```

```
-- ============================================
-- Step 1.2.1: National-level Correlation Matrix (Enhanced)
-- Purpose: Quantify relationship strength between P3 household density and demographic indicators
-- Enhancement: Added multicollinearity checks based on Step 1.1 findings
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_2_1_corr_national
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step1_2_1_corr_national` AS

SELECT
  'National' AS scope,
  COUNT(*) AS n_counties,

  -- ============================================
  -- PRIMARY: New car dealer correlations
  -- ============================================
  ROUND(CORR(new_per_1k_hh, high_income_pct), 3) AS new_income_corr,
  ROUND(CORR(new_per_1k_hh, college_rate), 3) AS new_edu_corr,
  ROUND(CORR(new_per_1k_hh, homeownership_rate), 3) AS new_homeown_corr,

  -- ============================================
  -- PRIMARY: Used car dealer correlations
  -- ============================================
  ROUND(CORR(used_per_1k_hh, high_income_pct), 3) AS used_income_corr,
  ROUND(CORR(used_per_1k_hh, low_mid_income_pct), 3) AS used_lowmid_corr,
```

```
 ROUND(CORR(used_per_1k_hh, college_rate), 3) AS used_edu_corr,
 ROUND(CORR(used_per_1k_hh, homeownership_rate), 3) AS used_homeown_corr,

 -- ============================================
 -- ENHANCEMENT 1: New vs Used correlation
 -- ============================================
 ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS new_used_corr,

 -- ============================================
 -- ENHANCEMENT 2: Multicollinearity checks
 -- Income vs Education (Step 1.1.3 found strong correlation)
 -- ============================================
 ROUND(CORR(high_income_pct, college_rate), 3) AS income_edu_corr,

 -- ============================================
 -- ENHANCEMENT 3: Homeownership vs other factors
 -- (Step 1.1.4 found negative correlation = urban proxy)
 -- ============================================
 ROUND(CORR(homeownership_rate, high_income_pct), 3) AS
homeown_income_corr,
 ROUND(CORR(homeownership_rate, college_rate), 3) AS homeown_edu_corr,

 -- ============================================
 -- ENHANCEMENT 4: Low-mid income correlation check
 -- (Alternative hypothesis for used cars)
 -- ============================================
 ROUND(CORR(new_per_1k_hh, low_mid_income_pct), 3) AS new_lowmid_corr

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`;
```

```
-- verify: View complete national-level correlation matrix
SELECT
 scope,
 n_counties,

 -- New car correlations
 new_income_corr,
 new_edu_corr,
 new_homeown_corr,
 new_lowmid_corr,

 -- Used car correlations
 used_income_corr,
 used_lowmid_corr,
 used_edu_corr,
 used_homeown_corr,

 -- Cross-type correlation
 new_used_corr,

 -- Multicollinearity checks
 income_edu_corr,
 homeown_income_corr,
 homeown_edu_corr

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_1_corr_national`;
```

```
-- ============================================
-- Step 1.2.2: Tier-stratified Correlation Analysis (Enhanced)
-- Purpose: Verify if correlations differ by city size (P5 principle validation)
-- Enhancement: More comprehensive correlation set based on Step 1.1 findings
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_2_2_corr_by_tier
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step1_2_2_corr_by_tier` AS

SELECT
  metro_tier,
  COUNT(*) AS n_counties,

  -- ============================================
  -- PRIMARY: New car dealer correlations by tier
  -- ============================================
  ROUND(CORR(new_per_1k_hh, high_income_pct), 3) AS new_income_corr,
  ROUND(CORR(new_per_1k_hh, college_rate), 3) AS new_edu_corr,
  ROUND(CORR(new_per_1k_hh, homeownership_rate), 3) AS new_homeown_corr,

  -- ============================================
  -- PRIMARY: Used car dealer correlations by tier
  -- ============================================
  ROUND(CORR(used_per_1k_hh, high_income_pct), 3) AS used_income_corr,
  ROUND(CORR(used_per_1k_hh, low_mid_income_pct), 3) AS used_lowmid_corr,
```

```
 ROUND(CORR(used_per_1k_hh, college_rate), 3) AS used_edu_corr,
 ROUND(CORR(used_per_1k_hh, homeownership_rate), 3) AS used_homeown_corr,

 -- ============================================
 -- ENHANCEMENT 1: New vs Used correlation by tier
 -- ============================================
 ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS new_used_corr,

 -- ============================================
 -- ENHANCEMENT 2: Multicollinearity checks by tier
 -- ============================================
 ROUND(CORR(high_income_pct, college_rate), 3) AS income_edu_corr,
 ROUND(CORR(homeownership_rate, high_income_pct), 3) AS homeown_income_corr

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`

GROUP BY metro_tier

ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;
```

```
-- verify: View tier-stratified correlation comparison
SELECT
  metro_tier,
  n_counties,

  -- New car correlations
  new_income_corr,
  new_edu_corr,
  new_homeown_corr,

  -- Used car correlations
  used_income_corr,
  used_lowmid_corr,
  used_edu_corr,
  used_homeown_corr,

  -- Cross-type
  new_used_corr,

  -- Multicollinearity
  income_edu_corr,
  homeown_income_corr

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_2_corr_by_tier`
ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
```

```
   WHEN 'Tier3_SmallMetro' THEN 3
   WHEN 'Tier4_Rural' THEN 4
  END;



-- verify:  Compare national vs tier-level correlations to validate P5 stratification
WITH national AS (
  SELECT
   'National_Average' AS tier_label,
   new_income_corr,
   used_income_corr,
   income_edu_corr
  FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_1_corr_national`
),
tier_range AS (
  SELECT
   'Tier_Range' AS tier_label,
   MAX(new_income_corr) - MIN(new_income_corr) AS new_income_range,
   MAX(used_income_corr) - MIN(used_income_corr) AS used_income_range,
   MAX(income_edu_corr) - MIN(income_edu_corr) AS income_edu_range
  FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_2_corr_by_tier`
)

SELECT
  n.tier_label,
  n.new_income_corr AS national_new_income,
  n.used_income_corr AS national_used_income,
  n.income_edu_corr AS national_income_edu,
```

```
  NULL AS range_value
FROM national n

UNION ALL

SELECT
  t.tier_label,
  t.new_income_range AS new_income_range,
  t.used_income_range AS used_income_range,
  t.income_edu_range AS income_edu_range,
  NULL
FROM tier_range t;



-- verify: Check if correlations are statistically significant based on sample size
-- Rule of thumb: |r| > 2/sqrt(n) for significance at p<0.05
SELECT
  metro_tier,
  n_counties,

  -- Calculate critical value for significance
  ROUND(2.0 / SQRT(n_counties), 3) AS critical_value_005,

  -- New car correlations with significance flags
  new_income_corr,
  CASE WHEN ABS(new_income_corr) > 2.0 / SQRT(n_counties) THEN '***' ELSE 'ns'
END AS new_income_sig,

  new_edu_corr,
```

```
  CASE WHEN ABS(new_edu_corr) > 2.0 / SQRT(n_counties) THEN '***' ELSE 'ns' END AS
new_edu_sig,


  new_homeown_corr,
  CASE WHEN ABS(new_homeown_corr) > 2.0 / SQRT(n_counties) THEN '***' ELSE 'ns'
END AS new_homeown_sig,


  -- Used car correlations with significance flags
  used_income_corr,
  CASE WHEN ABS(used_income_corr) > 2.0 / SQRT(n_counties) THEN '***' ELSE 'ns'
END AS used_income_sig,


  -- Multicollinearity flags (high correlation warning)
  income_edu_corr,
  CASE WHEN ABS(income_edu_corr) > 0.7 THEN ' ⚠️ High' ELSE 'OK' END AS
multicollinearity_warning


FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_2_corr_by_tier`
ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
    WHEN 'Tier3_SmallMetro' THEN 3
    WHEN 'Tier4_Rural' THEN 4
  END;




  -- ============================================
-- Step 1.3.1: Export Regression Data (Unified)
```

```
-- Purpose: Prepare clean dataset for regression analysis
-- Based on Step 1.2 findings: income is key, multicollinearity needs checking
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step1_3_1_regression_data
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step1_3_1_regression_data` AS

SELECT
  -- ============================================
  -- Geographic identifiers
  -- ============================================
  county_fips,
  state,
  county,

  -- ============================================
  -- P5: Metro tier (categorical control variable)
  -- ============================================
  metro_tier,

  -- ============================================
  -- Dependent variables (both models use same sample)
  -- ============================================
  new_per_1k_hh AS new_density,
  used_per_1k_hh AS used_density,

  -- ============================================
```

```
-- P4: Income variables (primary predictors)
-- ============================================
high_income_pct,
low_mid_income_pct,

-- ============================================
-- Education (potential multicollinearity with income)
-- Step 1.2.1 found r(income,education) = 0.744 ⚠️
-- ============================================
college_rate,

-- ============================================
-- Homeownership (proxy for urban/rural, Step 1.1.4)
-- May have multicollinearity with metro_tier
-- ============================================
homeownership_rate,

-- ============================================
-- Size controls (log transformation will be done in Python)
-- ============================================
total_population,
total_households,

-- ============================================
-- Additional useful variables for diagnostics
-- ============================================
new_car_dealers,
used_car_dealers,
```

```
 new_zscore_tier,
 used_zscore_tier

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`

-- ============================================
-- Filter: Exclude very small counties for statistical stability
-- Step 1.1 showed these have high CV and unstable patterns
-- ============================================
WHERE total_population >= 50000

-- Order for easier inspection
ORDER BY metro_tier, state, county;

-- Verify exported regression data
SELECT
 COUNT(*) AS n_counties,
 COUNT(DISTINCT metro_tier) AS n_tiers,

 -- Check for missing values
 COUNTIF(new_density IS NULL) AS null_new,
 COUNTIF(used_density IS NULL) AS null_used,
 COUNTIF(high_income_pct IS NULL) AS null_income,
 COUNTIF(college_rate IS NULL) AS null_edu,

 -- Descriptive statistics
 ROUND(AVG(new_density), 3) AS avg_new_density,
 ROUND(AVG(used_density), 3) AS avg_used_density,
```

```
  ROUND(AVG(high_income_pct), 2) AS avg_income_pct,
  ROUND(AVG(college_rate), 2) AS avg_edu_rate,

  -- Sample size by tier
  COUNTIF(metro_tier = 'Tier1_LargeMetro') AS tier1_n,
  COUNTIF(metro_tier = 'Tier2_MidMetro') AS tier2_n,
  COUNTIF(metro_tier = 'Tier3_SmallMetro') AS tier3_n,
  COUNTIF(metro_tier = 'Tier4_Rural') AS tier4_n

FROM `group22-fa25-mgmt58200-final.safegraph2.step1_3_1_regression_data`;



-- ============================================
-- Step 2.1: National-level Correlation - New Car vs Used Car
-- Purpose: Determine if the two business types are substitutes or complements
-- Based on RQ1 findings, we expect moderate complementary relationship
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_national
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step2_1_correlation_national` AS

WITH correlation_calc AS (
  SELECT
    'National' AS scope,
    COUNT(*) AS n_counties,

    -- Calculate Pearson correlation between new and used car dealer density
```

```
  CORR(new_per_1k_hh, used_per_1k_hh) AS density_correlation,

  -- Additional statistics for context
  AVG(new_per_1k_hh) AS avg_new_density,
  AVG(used_per_1k_hh) AS avg_used_density,
  STDDEV(new_per_1k_hh) AS sd_new_density,
  STDDEV(used_per_1k_hh) AS sd_used_density

 FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
 WHERE new_per_1k_hh IS NOT NULL
  AND used_per_1k_hh IS NOT NULL
)

SELECT
 scope,
 n_counties,
 ROUND(density_correlation, 3) AS density_correlation,

 -- Classify relationship type based on correlation strength
 CASE
  WHEN density_correlation > 0.5 THEN 'Strong_Complementary'
  WHEN density_correlation > 0.2 THEN 'Moderate_Complementary'
  WHEN density_correlation > -0.2 THEN 'Independent'
  WHEN density_correlation > -0.5 THEN 'Moderate_Substitute'
  ELSE 'Strong_Substitute'
 END AS relationship_type,

 -- Add descriptive interpretation
```

```
  CASE
    WHEN density_correlation > 0.5 THEN 'Strong positive correlation: high co-location, clear complementary relationship'
    WHEN density_correlation > 0.2 THEN 'Moderate positive correlation: tend to co-locate, likely complementary'
    WHEN density_correlation > -0.2 THEN 'Weak/no correlation: markets operate independently'
    WHEN density_correlation > -0.5 THEN 'Moderate negative correlation: some substitution effects'
    ELSE 'Strong negative correlation: clear substitute relationship'
  END AS interpretation,

  -- Context statistics
  ROUND(avg_new_density, 3) AS avg_new_density,
  ROUND(avg_used_density, 3) AS avg_used_density,
  ROUND(sd_new_density, 3) AS sd_new_density,
  ROUND(sd_used_density, 3) AS sd_used_density

FROM correlation_calc;


-- verify:  View the complete correlation analysis results
SELECT
  scope,
  n_counties,
  density_correlation,
  relationship_type,
  interpretation,
  avg_new_density,
  avg_used_density
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_national`;

-- verify:  Compare with RQ1 Step 1.2.1 correlation results
-- This validates consistency across analyses
SELECT
 'RQ2_Step2.1' AS source,
 density_correlation AS new_used_corr,
 relationship_type
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_national`

UNION ALL

SELECT
 'RQ1_Step1.2.1' AS source,
 new_used_corr,
 CASE
  WHEN new_used_corr > 0.5 THEN 'Strong_Complementary'
  WHEN new_used_corr > 0.2 THEN 'Moderate_Complementary'
  WHEN new_used_corr > -0.2 THEN 'Independent'
  WHEN new_used_corr > -0.5 THEN 'Moderate_Substitute'
  ELSE 'Strong_Substitute'
 END AS relationship_type
FROM `group22-fa25-mgmt58200-final.safegraph2.step1_2_1_corr_national`;


-- ============================================
-- Step 2.1 Supplementary: Tier-stratified Correlation
-- Purpose: Show that relationship strength varies by metro tier
```

```
-- Based on RQ1 finding: correlation ranges from 0.336 (Tier4) to 0.755 (Tier1)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_by_tier
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_by_tier` AS

SELECT
  metro_tier AS scope,
  COUNT(*) AS n_counties,

  -- Calculate correlation
  ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS density_correlation,

  -- Classify relationship type
  CASE
    WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.5 THEN 'Strong_Complementary'
    WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.2 THEN 'Moderate_Complementary'
    WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.2 THEN 'Independent'
    WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.5 THEN 'Moderate_Substitute'
    ELSE 'Strong_Substitute'
  END AS relationship_type,

  -- Context statistics
  ROUND(AVG(new_per_1k_hh), 3) AS avg_new_density,
  ROUND(AVG(used_per_1k_hh), 3) AS avg_used_density

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
```

```
WHERE new_per_1k_hh IS NOT NULL
  AND used_per_1k_hh IS NOT NULL

GROUP BY metro_tier

ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
    WHEN 'Tier3_SmallMetro' THEN 3
    WHEN 'Tier4_Rural' THEN 4
  END;



-- verify:  View tier-stratified correlation results
SELECT
  scope AS metro_tier,
  n_counties,
  density_correlation,
  relationship_type,
  avg_new_density,
  avg_used_density,

  -- Calculate the ratio of correlation to understand tier effect
  ROUND(density_correlation /
    (SELECT MAX(density_correlation)
     FROM `group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_by_tier`), 2)
    AS correlation_relative_to_max
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_1_correlation_by_tier`
ORDER BY
 CASE scope
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;

-- ============================================
-- Step 2.2: Tier-Stratified Correlation Analysis
-- Purpose: Validate that relationship strength varies by city size (P5 principle)
-- Compare P3 (household density) vs P2 (population density) correlations
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_2_correlation_by_tier
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step2_2_correlation_by_tier` AS

SELECT
 metro_tier,
 COUNT(*) AS n_counties,

 -- P3: Household density correlation (primary metric)
 ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS correlation_p3_household,

 -- P2: Population density correlation (comparison)
```

```
 ROUND(CORR(new_per_10k_pop, used_per_10k_pop), 3) AS
correlation_p2_population,


 -- Calculate difference to show P3 superiority

 ROUND(CORR(new_per_1k_hh, used_per_1k_hh) - CORR(new_per_10k_pop,
used_per_10k_pop), 3) AS p3_advantage,


 -- Descriptive statistics for household density (P3)

 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_per_1k_hh,

 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_per_1k_hh,

 ROUND(STDDEV(new_per_1k_hh), 3) AS sd_new_per_1k_hh,

 ROUND(STDDEV(used_per_1k_hh), 3) AS sd_used_per_1k_hh,


 -- Descriptive statistics for population density (P2) - for comparison

 ROUND(AVG(new_per_10k_pop), 3) AS avg_new_per_10k_pop,

 ROUND(AVG(used_per_10k_pop), 3) AS avg_used_per_10k_pop,


 -- Classify relationship type based on P3 correlation

 CASE

   WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.5 THEN 'Strong_Complementary'

   WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.2 THEN
'Moderate_Complementary'

   WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.2 THEN 'Independent'

   WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.5 THEN 'Moderate_Substitute'

   ELSE 'Strong_Substitute'

 END AS relationship_type,


 -- Calculate coefficient of variation to assess heterogeneity

 ROUND(STDDEV(new_per_1k_hh) / NULLIF(AVG(new_per_1k_hh), 0), 3) AS
cv_new_density,
```

```
 ROUND(STDDEV(used_per_1k_hh) / NULLIF(AVG(used_per_1k_hh), 0), 3) AS
cv_used_density


FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
WHERE new_per_1k_hh IS NOT NULL
 AND used_per_1k_hh IS NOT NULL
 AND new_per_10k_pop IS NOT NULL
 AND used_per_10k_pop IS NOT NULL

GROUP BY metro_tier

ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;



-- verify: View complete Step 2.2 results with all metrics
SELECT
 metro_tier,
 n_counties,

 -- Correlations
 correlation_p3_household,
 correlation_p2_population,
 p3_advantage,
```

```
 -- Relationship
 relationship_type,

 -- P3 Household density stats
 avg_new_per_1k_hh,
 avg_used_per_1k_hh,
 sd_new_per_1k_hh,
 sd_used_per_1k_hh,

 -- Variability
 cv_new_density,
 cv_used_density

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_2_correlation_by_tier`

ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
    WHEN 'Tier3_SmallMetro' THEN 3
    WHEN 'Tier4_Rural' THEN 4
  END;



-- ============================================
-- Step 2.3: Income-Stratified Correlation Analysis
```

```
-- Purpose: Examine if relationship varies by income level (P4 principle)
-- Test interaction between P4 (purchasing power) and market relationship
-- Input: step0_3_county_zscore
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_3_correlation_by_income
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step2_3_correlation_by_income` AS

WITH income_classified AS (
  SELECT
    *,
    -- Create income groups based on high_income_pct
    CASE
      WHEN high_income_pct >= 12 THEN 'High_Income'
      WHEN high_income_pct >= 8 THEN 'Middle_Income'
      ELSE 'Low_Income'
    END AS income_group

  FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
  WHERE new_per_1k_hh IS NOT NULL
    AND used_per_1k_hh IS NOT NULL
    AND high_income_pct IS NOT NULL
)

SELECT
  income_group,
  COUNT(*) AS n_counties,
```

```
 -- Calculate correlation
 ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS correlation,

 -- Classify relationship type
 CASE
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.5 THEN 'Strong_Complementary'
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.2 THEN
'Moderate_Complementary'
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.2 THEN 'Independent'
  WHEN CORR(new_per_1k_hh, used_per_10k_pop) > -0.5 THEN 'Moderate_Substitute'
  ELSE 'Strong_Substitute'
 END AS relationship_type,

 -- Descriptive statistics
 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_density,
 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_density,
 ROUND(STDDEV(new_per_1k_hh), 3) AS sd_new_density,
 ROUND(STDDEV(used_per_1k_hh), 3) AS sd_used_density,

 -- Income statistics
 ROUND(AVG(high_income_pct), 1) AS avg_income_pct,
 ROUND(MIN(high_income_pct), 1) AS min_income_pct,
 ROUND(MAX(high_income_pct), 1) AS max_income_pct,

 -- Metro tier distribution within each income group
 COUNTIF(metro_tier = 'Tier1_LargeMetro') AS n_tier1,
 COUNTIF(metro_tier = 'Tier2_MidMetro') AS n_tier2,
 COUNTIF(metro_tier = 'Tier3_SmallMetro') AS n_tier3,
 COUNTIF(metro_tier = 'Tier4_Rural') AS n_tier4
```

```
FROM income_classified

GROUP BY income_group

ORDER BY
 CASE income_group
  WHEN 'High_Income' THEN 1
  WHEN 'Middle_Income' THEN 2
  WHEN 'Low_Income' THEN 3
 END;


-- verify: View complete Step 2.3 results
SELECT
 income_group,
 n_counties,
 correlation,
 relationship_type,
 avg_new_density,
 avg_used_density,
 avg_income_pct,
 min_income_pct,
 max_income_pct,

 -- Tier distribution
 n_tier1,
 n_tier2,
```

```
 n_tier3,
 n_tier4,

 -- Calculate tier composition percentages
 ROUND(100.0 * n_tier1 / n_counties, 1) AS pct_tier1,
 ROUND(100.0 * n_tier4 / n_counties, 1) AS pct_tier4

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_3_correlation_by_income`

ORDER BY
 CASE income_group
  WHEN 'High_Income' THEN 1
  WHEN 'Middle_Income' THEN 2
  WHEN 'Low_Income' THEN 3
 END;



-- ============================================
-- Step 2.3 Supplementary: Income × Tier Interaction
-- Purpose: Examine if relationship varies by BOTH income and tier
-- This reveals the most nuanced market patterns
-- Output: group22-fa25-mgmt58200-
final.safegraph2.step2_3_correlation_income_tier
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step2_3_correlation_income_tier` AS

WITH income_classified AS (
```

```
 SELECT
  *,
  CASE
   WHEN high_income_pct >= 12 THEN 'High_Income'
   WHEN high_income_pct >= 8 THEN 'Middle_Income'
   ELSE 'Low_Income'
  END AS income_group

 FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`
 WHERE new_per_1k_hh IS NOT NULL
  AND used_per_1k_hh IS NOT NULL
  AND high_income_pct IS NOT NULL
)

SELECT
 metro_tier,
 income_group,
 COUNT(*) AS n_counties,

 -- Correlation
 ROUND(CORR(new_per_1k_hh, used_per_1k_hh), 3) AS correlation,

 -- Relationship type
 CASE
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.5 THEN 'Strong_Complementary'
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > 0.2 THEN
'Moderate_Complementary'
  WHEN CORR(new_per_1k_hh, used_per_1k_hh) > -0.2 THEN 'Independent'
  ELSE 'Substitute'
```

```
 END AS relationship_type,

 -- Densities
 ROUND(AVG(new_per_1k_hh), 3) AS avg_new_density,
 ROUND(AVG(used_per_1k_hh), 3) AS avg_used_density,

 -- Income
 ROUND(AVG(high_income_pct), 1) AS avg_income_pct

FROM income_classified

GROUP BY metro_tier, income_group
HAVING COUNT(*) >= 10  -- Only groups with sufficient sample size

ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END,
 CASE income_group
  WHEN 'High_Income' THEN 1
  WHEN 'Middle_Income' THEN 2
  WHEN 'Low_Income' THEN 3
 END;
```

```
-- verify: -- Compare P3 (household) vs P2 (population) correlation strength
SELECT
  metro_tier,
  n_counties,

  -- P3 vs P2 comparison
  correlation_p3_household AS p3_correlation,
  correlation_p2_population AS p2_correlation,
  p3_advantage,

  -- Show which is stronger
  CASE
    WHEN p3_advantage > 0 THEN 'P3_Stronger'
    WHEN p3_advantage < 0 THEN 'P2_Stronger'
    ELSE 'Equal'
  END AS stronger_metric,

  -- Percentage improvement
  ROUND(100.0 * p3_advantage / NULLIF(ABS(correlation_p2_population), 0), 1) AS
pct_improvement

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_2_correlation_by_tier`

ORDER BY
  CASE metro_tier
    WHEN 'Tier1_LargeMetro' THEN 1
    WHEN 'Tier2_MidMetro' THEN 2
    WHEN 'Tier3_SmallMetro' THEN 3
```

```
  WHEN 'Tier4_Rural' THEN 4
 END;


-- ============================================
-- Step 2.4 FINAL: Market Supply-Demand Classification with Tier-Stratified Thresholds
-- Purpose: Classify counties by supply status and opportunity level
-- Key Innovation: Tier-specific population thresholds to include all tiers appropriately
--
-- CRITICAL DESIGN DECISIONS:
--   1. NO state filter - state is NULL for all records but doesn't affect county-level analysis
--   2. Tier-stratified thresholds: Tier1/2 (100K), Tier3 (50K), Tier4 (15-20K)
--   3. Separate opportunity categories for Tier4 to enable demand validation priority
--   4. Confidence levels based on Step 2.1 complementarity findings
--
-- INSIGHTS FROM PRIOR ANALYSES:
--   - Step 2.1: Complementarity varies by tier (r=0.755 in Tier1 to r=0.336 in Tier4)
--   - Step 2.2: Tier4 has highest variability (CV=0.92) = largest opportunity pool
--   - Step 2.3: Income effects are tier-dependent, requiring stratified approach
--
-- Input: step0_3_county_zscore (all fields, consistent with P3 household density metric)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final
-- ============================================


CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final` AS
```

```
SELECT
 -- ============================================
 -- Part 1: Core Identifiers and Geographic Info
 -- ============================================
 county_fips,       -- Primary key: unique county identifier
 state,             -- Descriptive label (NULL in this dataset, but retained for future use)
 county,            -- Descriptive label
 metro_tier,        -- Critical stratification variable (Tier1-4)

 -- ============================================
 -- Part 2: Dealer Counts and Density Metrics
 -- ============================================
 new_car_dealers,   -- Raw count of new car dealerships
 used_car_dealers,  -- Raw count of used car dealerships

 -- P3 Household-based density (primary metric, consistent with all RQ1 and RQ2 analyses)
 new_per_1k_hh,     -- New car dealers per 1,000 households
 used_per_1k_hh,    -- Used car dealers per 1,000 households

 -- P2 Population-based density (reference only, shown to be similar to P3 in Step 2.2)
 new_per_10k_pop,   -- New car dealers per 10,000 population
 used_per_10k_pop,  -- Used car dealers per 10,000 population

 -- ============================================
 -- Part 3: Tier-Stratified Z-Scores (from Step 0.3)
 -- ============================================
 new_zscore_tier,   -- Z-score within metro tier (enables fair comparison)
 used_zscore_tier,  -- Z-score within metro tier
```

```
-- ============================================
-- Part 4: Demographics
-- ============================================
total_population,
total_households,
high_income_pct,      -- Percentage of households earning $150K+
low_mid_income_pct,   -- Percentage of households earning $35K-$150K
college_rate,         -- Percentage with bachelor's degree or higher
homeownership_rate,   -- Percentage of owner-occupied housing units

-- ============================================
-- Part 5: NEW CAR Supply Status Classification
-- Based on tier-stratified Z-scores
-- Thresholds: -1.5 (severe), -1.0 (moderate), -0.5 (slight)
-- ============================================
CASE
  WHEN new_zscore_tier < -1.5 THEN 'Severe_Undersupply'
  WHEN new_zscore_tier < -1.0 THEN 'Moderate_Undersupply'
  WHEN new_zscore_tier < -0.5 THEN 'Slight_Undersupply'
  WHEN new_zscore_tier > 1.5 THEN 'Oversupply'
  WHEN new_zscore_tier > 0.5 THEN 'Slight_Oversupply'
  ELSE 'Balanced'
END AS new_supply_status,

-- Simplified binary classification for easier filtering
CASE
  WHEN new_zscore_tier < -0.5 THEN 'Undersupply'
```

```
  WHEN new_zscore_tier > 0.5 THEN 'Oversupply'
  ELSE 'Balanced'
END AS new_supply_binary,

-- ============================================
-- Part 6: USED CAR Supply Status Classification
-- Same logic as new car
-- ============================================
CASE
  WHEN used_zscore_tier < -1.5 THEN 'Severe_Undersupply'
  WHEN used_zscore_tier < -1.0 THEN 'Moderate_Undersupply'
  WHEN used_zscore_tier < -0.5 THEN 'Slight_Undersupply'
  WHEN used_zscore_tier > 1.5 THEN 'Oversupply'
  WHEN used_zscore_tier > 0.5 THEN 'Slight_Oversupply'
  ELSE 'Balanced'
END AS used_supply_status,

CASE
  WHEN used_zscore_tier < -0.5 THEN 'Undersupply'
  WHEN used_zscore_tier > 0.5 THEN 'Oversupply'
  ELSE 'Balanced'
END AS used_supply_binary,

-- ============================================
-- Part 7: JOINT Supply Pattern (New + Used)
-- Critical for confidence assessment based on Step 2.1 complementarity findings
-- ============================================
CASE
```

```
  -- Both undersupplied: Most credible white market signal
  -- Especially important in Tier1/2 where complementarity is strong (r>0.73)
  WHEN new_zscore_tier < -0.5 AND used_zscore_tier < -0.5
   THEN 'Both_Undersupply'

  -- Only new undersupplied: Acceptable in Tier4 (weak complementarity r=0.34)
  -- Suspicious in Tier1/2 (contradicts strong complementarity)
  WHEN new_zscore_tier < -0.5 AND used_zscore_tier >= -0.5
   THEN 'New_Only_Undersupply'

  -- Only used undersupplied: Less relevant for new car dealer opportunities
  WHEN new_zscore_tier >= -0.5 AND used_zscore_tier < -0.5
   THEN 'Used_Only_Undersupply'

  -- Both oversupplied: Saturated market, avoid
  WHEN new_zscore_tier > 0.5 AND used_zscore_tier > 0.5
   THEN 'Both_Oversupply'

  -- Mixed or balanced: Neither clearly undersupplied
  ELSE 'Mixed_or_Balanced'
END AS joint_supply_pattern,

-- ============================================
-- Part 8: NEW CAR Opportunity Level (TIER-STRATIFIED THRESHOLDS)
--
-- DESIGN RATIONALE:
-- Different tiers require different population thresholds due to:
--   1. Market dynamics (urban vs rural dealer viability)
```

```
--   2. Average population by tier (Tier1: 1.8M, Tier4: 25K)
--   3. Complementarity strength (affects confidence in white market identification)
--
-- THRESHOLD LOGIC:
--   Tier1/2: 100K population (high-value markets, proven demand)
--   Tier3:   50K population (small metro markets)
--   Tier4:   15-20K population (rural markets, lower per-dealer revenue but viable)
--
-- TIER4 SEPARATE CATEGORIES:
--   Purpose: Enable different demand validation thresholds in Step 2.5
--   Tier4 has highest variability (CV=0.92) → needs stricter demand verification
-- ============================================
CASE
  -- HIGH OPPORTUNITY: Tier1/2 large metros + strong undersupply
  -- Highest priority: Large population ensures viable market size
  -- Strong complementarity (r>0.73) provides confidence
  WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
      AND new_zscore_tier < -1.0
      AND total_population >= 100000
   THEN 'High_Opportunity'

  -- MEDIUM-HIGH OPPORTUNITY: Any tier large city + moderate undersupply
  -- Large population (100K+) ensures sufficient market regardless of tier
  -- Captures large Tier3 cities and additional Tier1/2 markets
  WHEN total_population >= 100000
      AND new_zscore_tier < -0.7
   THEN 'Medium_High_Opportunity'
```

```
-- MEDIUM OPPORTUNITY: Tier3 small metros + moderate undersupply
-- Traditional small city markets (50K-100K population)
-- Established infrastructure, moderate competition
WHEN metro_tier = 'Tier3_SmallMetro'
    AND new_zscore_tier < -0.7
    AND total_population >= 50000
 THEN 'Medium_Opportunity'

-- TIER4-HIGH OPPORTUNITY: Rural counties + strong undersupply + sufficient size
-- NEW CATEGORY: Captures high-potential rural markets
-- 20K threshold: Ensures minimum viable market size for rural dealership
-- Weak complementarity (r=0.34) allows independent new car market identification
-- REQUIRES: Demand validation in Step 2.5 due to high variability (CV=0.92)
WHEN metro_tier = 'Tier4_Rural'
    AND new_zscore_tier < -1.0
    AND total_population >= 20000
 THEN 'Tier4_High_Opportunity'

-- TIER4-MEDIUM OPPORTUNITY: Rural counties + moderate undersupply
-- NEW CATEGORY: Broader rural opportunity pool
-- 15K threshold: Captures ~75% of Tier4 counties with undersupply
-- Lower threshold justified by:
--   (1) Lower operating costs in rural areas
--   (2) Less competition (higher market share per dealer)
--   (3) Higher customer loyalty in rural communities
-- REQUIRES: More stringent demand validation in Step 2.5
WHEN metro_tier = 'Tier4_Rural'
    AND new_zscore_tier < -0.7
```

```
      AND total_population >= 15000
    THEN 'Tier4_Medium_Opportunity'

  -- NOT RECOMMENDED: All other cases
  -- Includes: balanced/oversupplied markets, very small counties, positive z-scores
  ELSE 'Not_Recommended'
END AS new_opportunity_level,

-- Numeric score for ranking and prioritization (0-5 scale)
-- Higher score = higher priority for market entry
CASE
  WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
      AND new_zscore_tier < -1.0
      AND total_population >= 100000
    THEN 5  -- Highest: Large metros, proven demand, high certainty

  WHEN total_population >= 100000
      AND new_zscore_tier < -0.7
    THEN 4  -- High: Large cities, moderate undersupply

  WHEN metro_tier = 'Tier3_SmallMetro'
      AND new_zscore_tier < -0.7
      AND total_population >= 50000
    THEN 3  -- Medium: Small metros, established markets

  WHEN metro_tier = 'Tier4_Rural'
      AND new_zscore_tier < -1.0
      AND total_population >= 20000
```

```
    THEN 2  -- Lower: Rural high-potential (needs demand validation)

  WHEN metro_tier = 'Tier4_Rural'
     AND new_zscore_tier < -0.7
     AND total_population >= 15000
    THEN 1  -- Lowest: Rural medium-potential (careful evaluation needed)

  ELSE 0  -- Not recommended
END AS new_opportunity_score,

-- ============================================
-- Part 9: USED CAR Opportunity Level
-- Parallel structure to new car classification
-- Same tier-stratified thresholds apply
-- ============================================
CASE
  WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
     AND used_zscore_tier < -1.0
     AND total_population >= 100000
    THEN 'High_Opportunity'

  WHEN total_population >= 100000
     AND used_zscore_tier < -0.7
    THEN 'Medium_High_Opportunity'

  WHEN metro_tier = 'Tier3_SmallMetro'
     AND used_zscore_tier < -0.7
     AND total_population >= 50000
```

```
    THEN 'Medium_Opportunity'

  WHEN metro_tier = 'Tier4_Rural'
     AND used_zscore_tier < -1.0
     AND total_population >= 20000
   THEN 'Tier4_High_Opportunity'

  WHEN metro_tier = 'Tier4_Rural'
     AND used_zscore_tier < -0.7
     AND total_population >= 15000
   THEN 'Tier4_Medium_Opportunity'

  ELSE 'Not_Recommended'
END AS used_opportunity_level,

-- Numeric score for used car opportunities (same 0-5 scale)
CASE
  WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
     AND used_zscore_tier < -1.0
     AND total_population >= 100000
   THEN 5

  WHEN total_population >= 100000
     AND used_zscore_tier < -0.7
   THEN 4

  WHEN metro_tier = 'Tier3_SmallMetro'
     AND used_zscore_tier < -0.7
```

```
      AND total_population >= 50000
    THEN 3

  WHEN metro_tier = 'Tier4_Rural'
      AND used_zscore_tier < -1.0
      AND total_population >= 20000
    THEN 2

  WHEN metro_tier = 'Tier4_Rural'
      AND used_zscore_tier < -0.7
      AND total_population >= 15000
    THEN 1

  ELSE 0
 END AS used_opportunity_score,


 -- ============================================
 -- Part 10: White Market Confidence Level
 -- Based on Step 2.1 complementarity findings
 --
 -- CONFIDENCE LOGIC VALIDATION:
 --   High Confidence: Tier1/2 with both undersupplied
 --     → Aligns with strong complementarity (r=0.73-0.76)
 --     → Both new and used markets confirm undersupply
 --
 --   Medium Confidence: Tier3/4 scenarios
 --     → Tier3: Both undersupplied (moderate complementarity r=0.54)
 --     → Tier4: New only undersupplied (weak complementarity r=0.34 allows
independence)
```

```
--
--   Low Confidence: Tier1/2 with only new undersupplied
--    → Contradicts strong complementarity → suspicious pattern
--    → May indicate data issues or special market conditions
-- ============================================
CASE
 -- HIGH CONFIDENCE: Tier1/2 with both new and used undersupplied
 -- Strong complementarity (r=0.73-0.76) makes this a reliable signal
 -- Both markets independently confirm the opportunity
 WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
    AND new_zscore_tier < -0.5
    AND used_zscore_tier < -0.5
  THEN 'High_Confidence'

 -- MEDIUM CONFIDENCE: Tier4 with new undersupplied
 -- Weak complementarity (r=0.34) means new car market can operate independently
 -- Used car market status is less informative for new car opportunities
 WHEN metro_tier = 'Tier4_Rural'
    AND new_zscore_tier < -0.5
  THEN 'Medium_Confidence'

 -- MEDIUM CONFIDENCE: Tier3 with both undersupplied
 -- Moderate complementarity (r=0.54) suggests both should be undersupplied
 -- Joint undersupply provides reasonable confidence
 WHEN metro_tier = 'Tier3_SmallMetro'
    AND new_zscore_tier < -0.5
    AND used_zscore_tier < -0.5
  THEN 'Medium_Confidence'
```

```
  -- LOW CONFIDENCE: Tier1/2 with only new undersupplied
  -- Contradicts strong complementarity → suspicious
  -- Possible explanations: data issues, Tesla/direct-sales dominance, timing mismatch
  -- Requires additional investigation before investment decision
  WHEN metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
      AND new_zscore_tier < -0.5
      AND used_zscore_tier >= -0.5
    THEN 'Low_Confidence'

  -- NOT APPLICABLE: No undersupply detected
  ELSE 'Not_Applicable'
 END AS white_market_confidence,

-- ============================================
-- Part 11: Market Characterization Tags
-- Quick filters for analysis and reporting
-- ============================================

-- Undersupply severity for new car market
-- Enables prioritization by urgency of market need
CASE
  WHEN new_zscore_tier < -1.5 THEN 'Critical'    -- Extreme undersupply, urgent opportunity
  WHEN new_zscore_tier < -1.0 THEN 'Severe'     -- Strong undersupply, high priority
  WHEN new_zscore_tier < -0.5 THEN 'Moderate'   -- Noticeable undersupply, worth investigating
  ELSE 'None'                        -- Balanced or oversupplied
```

```
 END AS new_undersupply_severity,

 -- Tier variability flag (from Step 2.2 CV analysis)
 -- High variability tiers = more heterogeneous markets = higher due diligence needed
 -- Tier4 CV=0.92, Tier1 CV=0.81 → highest variability
 -- Tier2 CV=0.60, Tier3 CV=0.60 → moderate variability
 CASE
  WHEN metro_tier IN ('Tier4_Rural', 'Tier1_LargeMetro')
    THEN 'High_Variability_Tier'
  ELSE 'Moderate_Variability_Tier'
 END AS tier_variability_flag,

 -- ============================================
 -- Part 12: Data Quality Flag
 -- Identifies potentially problematic records for manual review
 -- NOTE: State is NULL for all records but this is NOT a data quality issue
 --      (county_fips provides unique identification at county level)
 -- ============================================
 CASE
  -- Suspicious pattern: Large population but zero dealers
  -- May indicate: new suburban county, data lag, or genuine white market
  WHEN new_car_dealers = 0
     AND used_car_dealers = 0
     AND total_population > 500000
   THEN 'Suspicious_Zero_Dealers'

  -- Missing critical identifier
  WHEN county_fips IS NULL
```

```
    THEN 'Missing_FIPS'

  -- Missing population data (required for threshold logic)
  WHEN total_population IS NULL
    THEN 'Missing_Population'

  -- All checks passed
  ELSE 'OK'
 END AS data_quality_flag

FROM `group22-fa25-mgmt58200-final.safegraph2.step0_3_county_zscore`;




-- verify:  Validation 1: Verify all tiers are appropriately classified
-- Success criteria:
--   - Tier1/2 should have High_Opportunity counties (50-70 expected)
--   - Tier3 should have Medium_Opportunity counties (150-200 expected)
--   - Tier4 should have Tier4_High/Medium_Opportunity (300-400 expected)
--   - Not all tiers should be 100% Not_Recommended


SELECT
 metro_tier,
 new_opportunity_level,
 COUNT(*) AS n_counties,
```

```
 -- Percentage within tier
 ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER (PARTITION BY metro_tier), 1) AS
pct_of_tier,

 -- Average characteristics
 ROUND(AVG(new_zscore_tier), 2) AS avg_zscore,
 ROUND(AVG(new_per_1k_hh), 3) AS avg_density,
 ROUND(AVG(total_population), 0) AS avg_population,
 ROUND(AVG(high_income_pct), 1) AS avg_income_pct,

 -- Data quality check
 COUNTIF(data_quality_flag != 'OK') AS n_data_issues

FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

GROUP BY metro_tier, new_opportunity_level

ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END,
 CASE new_opportunity_level
  WHEN 'High_Opportunity' THEN 1
  WHEN 'Medium_High_Opportunity' THEN 2
  WHEN 'Medium_Opportunity' THEN 3
```

```
  WHEN 'Tier4_High_Opportunity' THEN 4
  WHEN 'Tier4_Medium_Opportunity' THEN 5
  WHEN 'Not_Recommended' THEN 6
 END;



-- verify:  Validation 2: Key success metrics
-- Compare against original Step 2.4 where Tier4 was 100% excluded
-- ============================================


SELECT
 'Total Counties' AS metric,
 CAST(COUNT(*) AS STRING) AS count,
 '2527' AS expected,
 CASE WHEN COUNT(*) = 2527 THEN '✓' ELSE '✗' END AS status
FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Total Opportunity Counties',
 CAST(COUNTIF(new_opportunity_level != 'Not_Recommended') AS STRING),
 '500-700',
 CASE
  WHEN COUNTIF(new_opportunity_level != 'Not_Recommended') BETWEEN 500 AND
700
  THEN '✓' ELSE '✗'
```

```
  END
FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Tier1/2 High Opportunity',
 CAST(COUNTIF(metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
        AND new_opportunity_level = 'High_Opportunity') AS STRING),
 '50-70',
 CASE
  WHEN COUNTIF(metro_tier IN ('Tier1_LargeMetro', 'Tier2_MidMetro')
         AND new_opportunity_level = 'High_Opportunity') BETWEEN 50 AND 70
  THEN '✓' ELSE '✗'
 END
FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Tier3 Medium Opportunity',
 CAST(COUNTIF(metro_tier = 'Tier3_SmallMetro'
        AND new_opportunity_level = 'Medium_Opportunity') AS STRING),
 '150-200',
 CASE
  WHEN COUNTIF(metro_tier = 'Tier3_SmallMetro'
         AND new_opportunity_level = 'Medium_Opportunity') BETWEEN 150 AND 200
```

```
  THEN '✓' ELSE '✗'
 END
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Tier4 Total Opportunity',
 CAST(COUNTIF(metro_tier = 'Tier4_Rural'
        AND new_opportunity_level IN ('Tier4_High_Opportunity', 'Tier4_Medium_Opportunity')) AS STRING),
 '300-400',
 CASE
  WHEN COUNTIF(metro_tier = 'Tier4_Rural'
         AND new_opportunity_level IN ('Tier4_High_Opportunity', 'Tier4_Medium_Opportunity'))
     BETWEEN 300 AND 400
  THEN '✓' ELSE '✗'
 END
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'High Confidence Counties',
 CAST(COUNTIF(white_market_confidence = 'High_Confidence') AS STRING),
 '60-70',
 CASE
```

```
  WHEN COUNTIF(white_market_confidence = 'High_Confidence') BETWEEN 60 AND
70
  THEN '✓' ELSE '✗'
 END
FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Medium Confidence Counties',
 CAST(COUNTIF(white_market_confidence = 'Medium_Confidence') AS STRING),
 '400-600',
 CASE
  WHEN COUNTIF(white_market_confidence = 'Medium_Confidence') BETWEEN 400
AND 600
  THEN '✓' ELSE '✗'
 END
FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

UNION ALL

SELECT
 'Data Quality Issues',
 CAST(COUNTIF(data_quality_flag != 'OK') AS STRING),
 '<50',
 CASE
  WHEN COUNTIF(data_quality_flag != 'OK') < 50
  THEN '✓' ELSE '✗'
```

```
  END
FROM `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final`;



-- verify: Validation 3: Top 20 high opportunity counties for new car dealers
-- Should now include Tier1/2 counties (previously excluded by state filter)
-- Ranked by opportunity score (5=highest) then population
-- ========================================

SELECT
  county_fips,
  county,
  metro_tier,

  -- Supply classification
  new_supply_status,
  used_supply_status,
  joint_supply_pattern,

  -- Opportunity assessment
  new_opportunity_level,
  new_opportunity_score,
  white_market_confidence,

  -- Key metrics
  ROUND(new_zscore_tier, 2) AS new_zscore,
  ROUND(new_per_1k_hh, 3) AS new_density,
```

```
 new_car_dealers,
 total_population,
 ROUND(high_income_pct, 1) AS income_pct,

 -- For validation: used car comparison
 ROUND(used_zscore_tier, 2) AS used_zscore,
 used_car_dealers,

 -- Data quality
 data_quality_flag

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final`

WHERE new_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity')

ORDER BY
 new_opportunity_score DESC,
 total_population DESC

LIMIT 20;




-- verify: : Opportunity Level × Confidence Level matrix
-- Validates that confidence logic aligns with complementarity findings
-- ========================================
```

```
SELECT
  new_opportunity_level,
  white_market_confidence,
  COUNT(*) AS n_counties,

  -- Average metrics
  ROUND(AVG(new_zscore_tier), 2) AS avg_new_zscore,
  ROUND(AVG(used_zscore_tier), 2) AS avg_used_zscore,
  ROUND(AVG(total_population), 0) AS avg_population,

  -- Tier composition
  COUNTIF(metro_tier = 'Tier1_LargeMetro') AS n_tier1,
  COUNTIF(metro_tier = 'Tier2_MidMetro') AS n_tier2,
  COUNTIF(metro_tier = 'Tier3_SmallMetro') AS n_tier3,
  COUNTIF(metro_tier = 'Tier4_Rural') AS n_tier4,

  -- Tier4 percentage (should be high for Tier4 categories)
  ROUND(100.0 * COUNTIF(metro_tier = 'Tier4_Rural') / COUNT(*), 1) AS pct_tier4

FROM `group22-fa25-mgmt58200-
final.safegraph2.step2_4_market_classification_final`

WHERE new_opportunity_level != 'Not_Recommended'

GROUP BY new_opportunity_level, white_market_confidence

ORDER BY
  CASE new_opportunity_level
    WHEN 'High_Opportunity' THEN 1
```

```
  WHEN 'Medium_High_Opportunity' THEN 2
  WHEN 'Medium_Opportunity' THEN 3
  WHEN 'Tier4_High_Opportunity' THEN 4
  WHEN 'Tier4_Medium_Opportunity' THEN 5
 END,
 CASE white_market_confidence
  WHEN 'High_Confidence' THEN 1
  WHEN 'Medium_Confidence' THEN 2
  WHEN 'Low_Confidence' THEN 3
  WHEN 'Not_Applicable' THEN 4
 END;


-- ============================================
-- Step 2.5 REVISED: Integrated White Market Classification with Demand Validation
-- Purpose: Merge supply-demand classification with visitor-based demand validation
--
-- CRITICAL DATA LIMITATION ACKNOWLEDGED:
--   SafeGraph visitor data coverage: Only 7.1% of undersupply counties (50/706)
--   - Tier1/2: 7-8% coverage (primarily due to 60xxx suspicious records)
--   - Tier3: 30% coverage (best tier for validation)
--   - Tier4: 0.6% coverage (rural areas lack SafeGraph coverage)
--
-- REVISED CLASSIFICATION APPROACH:
--   - Only counties WITH visitor data can be classified as True/Potential/Low_Demand
--   - Counties WITHOUT visitor data are marked as 'Cannot_Validate_Demand'
--   - This distinction is critical for honest reporting and investment decisions
--
```

```
-- EXPECTED OUTPUT:
--   - True White Markets: ~31 counties (validated with demand data)
--   - Potential White Markets: ~10 counties (validated with medium demand)
--   - Cannot Validate: ~659 counties (require alternative validation methods)
--
-- Input Tables:
--   - step2_4_market_classification_final: Supply-demand classification
--   - step0_4_county_visitor_metrics: Visitor-based demand indicators
--
-- Output: group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final
-- ============================================


CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final` AS


WITH
-- ============================================
-- CTE 1: Calculate demand intensity percentiles
-- ============================================
demand_percentiles AS (
  SELECT
    APPROX_QUANTILES(demand_intensity_new, 100)[OFFSET(25)] AS new_demand_p25,
    APPROX_QUANTILES(demand_intensity_new, 100)[OFFSET(50)] AS new_demand_p50,
    APPROX_QUANTILES(demand_intensity_new, 100)[OFFSET(75)] AS new_demand_p75,
    APPROX_QUANTILES(demand_intensity_used, 100)[OFFSET(25)] AS used_demand_p25,
    APPROX_QUANTILES(demand_intensity_used, 100)[OFFSET(50)] AS used_demand_p50,
```

```
  APPROX_QUANTILES(demand_intensity_used, 100)[OFFSET(75)] AS
used_demand_p75
 FROM `group22-fa25-mgmt58200-final.safegraph2.step0_4_county_visitor_metrics`
 WHERE demand_intensity_new IS NOT NULL
  AND demand_intensity_used IS NOT NULL
),

-- ============================================
-- CTE 2: Main data integration (LEFT JOIN to preserve all counties)
-- ============================================
integrated_data AS (
 SELECT
  -- All fields from market classification (Step 2.4)
  mc.county_fips,
  mc.state,
  mc.county,
  mc.metro_tier,
  mc.new_car_dealers,
  mc.used_car_dealers,
  mc.new_per_1k_hh,
  mc.used_per_1k_hh,
  mc.new_per_10k_pop,
  mc.used_per_10k_pop,
  mc.new_zscore_tier,
  mc.used_zscore_tier,
  mc.total_population,
  mc.total_households,
  mc.high_income_pct,
  mc.low_mid_income_pct,
```

```
  mc.college_rate,
  mc.homeownership_rate,
  mc.new_supply_status,
  mc.new_supply_binary,
  mc.used_supply_status,
  mc.used_supply_binary,
  mc.joint_supply_pattern,
  mc.new_opportunity_level,
  mc.new_opportunity_score,
  mc.used_opportunity_level,
  mc.used_opportunity_score,
  mc.white_market_confidence,
  mc.new_undersupply_severity,
  mc.tier_variability_flag,
  mc.data_quality_flag,

  -- Visitor metrics from Step 0.4 (NULL if no data)
  vm.demand_intensity_new,
  vm.demand_intensity_used,
  vm.new_visitors_per_1k_hh,
  vm.used_visitors_per_1k_hh,
  vm.new_avg_visitors_per_dealer,
  vm.used_avg_visitors_per_dealer,

  -- Visitor data availability flag
  CASE
   WHEN vm.county_fips IS NOT NULL THEN TRUE
   ELSE FALSE
```

```
  END AS has_visitor_data,

  -- Percentile references
  dp.new_demand_p25,
  dp.new_demand_p50,
  dp.new_demand_p75,
  dp.used_demand_p25,
  dp.used_demand_p50,
  dp.used_demand_p75

 FROM `group22-fa25-mgmt58200-final.safegraph2.step2_4_market_classification_final` mc
 LEFT JOIN `group22-fa25-mgmt58200-final.safegraph2.step0_4_county_visitor_metrics` vm
  ON mc.county_fips = vm.county_fips
 CROSS JOIN demand_percentiles dp
)

-- ============================================
-- Main SELECT: Add market quality classifications
-- ============================================
SELECT
 *,

 -- ============================================
 -- Part 4: NEW CAR Market Quality Classification (REVISED)
 --
 -- KEY CHANGE: 'No_Visitor_Data' → 'Cannot_Validate_Demand'
 -- More accurate terminology that reflects data limitation
```

```
--
-- CLASSIFICATION LOGIC:
--   True White Market: Supply shortage + High demand (> P50) + HAS DATA
--   Potential White Market: Supply shortage + Medium demand (P25-P50) + HAS DATA
--   Low Demand Market: Supply shortage + Low demand (< P25) + HAS DATA
--   Cannot Validate Demand: Supply shortage + NO DATA (93% of cases)
--   Saturated Market: Adequate supply (regardless of demand)
-- ==========================================
CASE
  -- TRUE WHITE MARKET: Supply shortage + Strong demand + VERIFIED
  -- These are the ONLY counties we confidently recommend
  -- Investment risk: LOW (both supply and demand validated)
  WHEN new_zscore_tier < -1.0
     AND demand_intensity_new > new_demand_p50
     AND demand_intensity_new IS NOT NULL
   THEN 'True_White_Market'

  -- POTENTIAL WHITE MARKET: Supply shortage + Medium demand + VERIFIED
  -- Secondary targets requiring more due diligence
  -- Investment risk: MEDIUM (demand is moderate, not strong)
  WHEN new_zscore_tier < -1.0
     AND demand_intensity_new >= new_demand_p25
     AND demand_intensity_new <= new_demand_p50
     AND demand_intensity_new IS NOT NULL
   THEN 'Potential_White_Market'

  -- LOW DEMAND MARKET: Supply shortage + Weak demand + VERIFIED
  -- Supply shortage does NOT indicate opportunity when demand is weak
```

```
  -- Investment risk: HIGH - AVOID
  -- This validates our approach: not all undersupply = opportunity
  WHEN new_zscore_tier < -1.0
      AND demand_intensity_new < new_demand_p25
      AND demand_intensity_new IS NOT NULL
   THEN 'Low_Demand_Market'

  -- CANNOT VALIDATE DEMAND: Supply shortage but NO VISITOR DATA
  -- REVISED NAME: More accurate than "No_Visitor_Data"
  -- These counties CANNOT be recommended without alternative validation
  -- Investment risk: UNKNOWN (supply shortage exists, but demand unverified)
  --
  -- This category includes 93% of undersupply counties due to:
  --   1. SafeGraph coverage limitations in rural areas (Tier4: 99%)
  --   2. Counties with zero dealers have no POIs to track visitors
  --   3. Data quality issues (60xxx/40xxx FIPS codes)
  WHEN new_zscore_tier < -1.0
      AND demand_intensity_new IS NULL
   THEN 'Cannot_Validate_Demand'

  -- SATURATED MARKET: Adequate or excessive supply
  -- Not an opportunity regardless of demand level
  WHEN new_zscore_tier >= -0.5
   THEN 'Saturated_Market'

  -- OTHER: Edge cases
  ELSE 'Other'
 END AS new_market_quality,
```

```
-- ==========================================
-- Part 5: USED CAR Market Quality Classification (REVISED)
-- Same logic as new car
-- ==========================================
CASE
 WHEN used_zscore_tier < -1.0
    AND demand_intensity_used > used_demand_p50
    AND demand_intensity_used IS NOT NULL
  THEN 'True_White_Market'

 WHEN used_zscore_tier < -1.0
    AND demand_intensity_used >= used_demand_p25
    AND demand_intensity_used <= used_demand_p50
    AND demand_intensity_used IS NOT NULL
  THEN 'Potential_White_Market'

 WHEN used_zscore_tier < -1.0
    AND demand_intensity_used < used_demand_p25
    AND demand_intensity_used IS NOT NULL
  THEN 'Low_Demand_Market'

 WHEN used_zscore_tier < -1.0
    AND demand_intensity_used IS NULL
  THEN 'Cannot_Validate_Demand'

 WHEN used_zscore_tier >= -0.5
  THEN 'Saturated_Market'
```

```
  ELSE 'Other'
 END AS used_market_quality,

 -- ============================================
 -- Part 6: Market Activity Score (0-10 scale for RQ3)
 -- Only meaningful when visitor data exists
 -- ============================================
 CASE
  WHEN demand_intensity_new >= new_demand_p75 THEN 10
  WHEN demand_intensity_new >= new_demand_p50 THEN 8
  WHEN demand_intensity_new >= new_demand_p25 THEN 6
  WHEN demand_intensity_new IS NOT NULL THEN 4
  ELSE 0  -- No data = cannot assess activity
 END AS market_activity_score_new,

 CASE
  WHEN demand_intensity_used >= used_demand_p75 THEN 10
  WHEN demand_intensity_used >= used_demand_p50 THEN 8
  WHEN demand_intensity_used >= used_demand_p25 THEN 6
  WHEN demand_intensity_used IS NOT NULL THEN 4
  ELSE 0
 END AS market_activity_score_used

FROM integrated_data;


-- ============================================
```

```
-- Step 2.6a: Export New Car True White Markets (Demand-Validated)
-- Purpose: Export only markets with VERIFIED strong demand
-- Filtering: Requires both supply shortage AND visitor data validation
--
-- Expected Output: ~20 counties (primarily from Tier3, some from Tier2)
-- Quality: All recommendations are evidence-based, not speculation
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_6a_new_car_true_white_top20` AS

SELECT
 -- ============================================
 -- Part 1: Geographic Identifiers
 -- ============================================
 county_fips,
 county,
 state,
 metro_tier,

 -- ============================================
 -- Part 2: Population and Demographics
 -- ============================================
 total_population,
 total_households,
 ROUND(high_income_pct, 1) AS high_income_pct,
 ROUND(college_rate, 1) AS college_rate,
 ROUND(homeownership_rate, 1) AS homeownership_rate,
```

```
-- ============================================
-- Part 3: Supply Status (Z-score based)
-- ============================================
new_supply_status,
ROUND(new_zscore_tier, 2) AS new_zscore,
new_car_dealers AS actual_dealers,
ROUND(new_per_1k_hh, 3) AS dealers_per_1k_hh,

-- ============================================
-- Part 4: Demand Validation Metrics (KEY ADDITION)
-- These metrics prove demand exists
-- ============================================
ROUND(demand_intensity_new, 0) AS demand_intensity,
ROUND(new_visitors_per_1k_hh, 0) AS visitors_per_1k_hh,
ROUND(new_avg_visitors_per_dealer, 0) AS avg_visitors_per_dealer,
market_activity_score_new,

-- ============================================
-- Part 5: Classification and Confidence
-- ============================================
new_opportunity_level,
new_opportunity_score,
new_market_quality,  -- Should all be 'True_White_Market'
white_market_confidence,

-- ============================================
-- Part 6: Data Quality Flags
-- ============================================
```

```
  data_quality_flag,
  has_visitor_data,  -- Should all be TRUE

  -- ============================================
  -- Part 7: Market Type Label (for reporting)
  -- ============================================
  'Demand_Validated_True_White_Market' AS market_type

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final`

WHERE
  -- CRITICAL FILTERS: Only verified opportunities
  new_market_quality = 'True_White_Market'  -- Strong demand validated
  AND has_visitor_data = TRUE  -- Must have demand data
  AND data_quality_flag = 'OK'  -- Exclude 33 suspicious records
  AND new_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity',
'Medium_Opportunity')
  AND total_population >= 50000  -- Minimum viable market size

ORDER BY
  -- Prioritization logic
  new_opportunity_score DESC,  -- Tier priority (5=Tier1/2, 4=Medium-High, 3=Medium)
  market_activity_score_new DESC,  -- Demand strength (10=top 25%, 8=top 50%)
  demand_intensity_new DESC,  -- Absolute demand level
  total_population DESC  -- Market size

LIMIT 20;
```

```
-- ============================================
-- Step 2.6b: Export New Car Potential White Markets (Medium Demand)
-- Purpose: Secondary targets with moderate demand (P25-P50)
-- Risk level: MEDIUM (require more careful evaluation)
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_6b_new_car_potential_top10` AS

SELECT
  county_fips,
  county,
  state,
  metro_tier,
  total_population,
  total_households,
  ROUND(high_income_pct, 1) AS high_income_pct,
  ROUND(college_rate, 1) AS college_rate,
  new_supply_status,
  ROUND(new_zscore_tier, 2) AS new_zscore,
  new_car_dealers AS actual_dealers,
  ROUND(new_per_1k_hh, 3) AS dealers_per_1k_hh,

  -- Demand metrics (will show medium demand: P25-P50)
  ROUND(demand_intensity_new, 0) AS demand_intensity,
  ROUND(new_visitors_per_1k_hh, 0) AS visitors_per_1k_hh,
  market_activity_score_new,

  new_opportunity_level,
```

```
  new_market_quality,  -- Should all be 'Potential_White_Market'
  white_market_confidence,
  data_quality_flag,

  'Demand_Validated_Potential_White_Market' AS market_type

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final`

WHERE
  new_market_quality = 'Potential_White_Market'
  AND has_visitor_data = TRUE
  AND data_quality_flag = 'OK'
  AND new_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity',
'Medium_Opportunity')
  AND total_population >= 50000

ORDER BY
  new_opportunity_score DESC,
  market_activity_score_new DESC,
  demand_intensity_new DESC

LIMIT 10;



-- ============================================
-- Step 2.6c: Export Used Car True White Markets
-- Same logic as new car, but for used car markets
-- ============================================
```

```
CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step2_6c_used_car_true_white_top20` AS

SELECT
 county_fips,
 county,
 state,
 metro_tier,
 total_population,
 total_households,
 ROUND(low_mid_income_pct, 1) AS low_mid_income_pct,  -- Used car target demographic
 ROUND(homeownership_rate, 1) AS homeownership_rate,

 used_supply_status,
 ROUND(used_zscore_tier, 2) AS used_zscore,
 used_car_dealers AS actual_dealers,
 ROUND(used_per_1k_hh, 3) AS dealers_per_1k_hh,

 -- Used car demand metrics
 ROUND(demand_intensity_used, 0) AS demand_intensity,
 ROUND(used_visitors_per_1k_hh, 0) AS visitors_per_1k_hh,
 ROUND(used_avg_visitors_per_dealer, 0) AS avg_visitors_per_dealer,
 market_activity_score_used,

 used_opportunity_level,
 used_opportunity_score,
 used_market_quality,  -- Should all be 'True_White_Market'
 white_market_confidence,
```

```
  data_quality_flag,

  'Demand_Validated_True_White_Market_Used' AS market_type

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final`

WHERE
  used_market_quality = 'True_White_Market'
  AND has_visitor_data = TRUE
  AND data_quality_flag = 'OK'
  AND used_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity',
'Medium_Opportunity')
  AND total_population >= 50000

ORDER BY
  used_opportunity_score DESC,
  market_activity_score_used DESC,
  demand_intensity_used DESC

LIMIT 20;



-- ============================================
-- Step 2.6d: Summary Statistics for Reporting
-- Purpose: Provide aggregate numbers for research report
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step2_6d_white_market_summary` AS
```

```
SELECT
  metro_tier,

  -- New car markets
  COUNTIF(new_market_quality = 'True_White_Market') AS new_true_white,
  COUNTIF(new_market_quality = 'Potential_White_Market') AS new_potential,
  COUNTIF(new_market_quality = 'Low_Demand_Market') AS new_low_demand,
  COUNTIF(new_market_quality = 'Cannot_Validate_Demand') AS
new_cannot_validate,

  -- Used car markets
  COUNTIF(used_market_quality = 'True_White_Market') AS used_true_white,
  COUNTIF(used_market_quality = 'Potential_White_Market') AS used_potential,
  COUNTIF(used_market_quality = 'Low_Demand_Market') AS used_low_demand,
  COUNTIF(used_market_quality = 'Cannot_Validate_Demand') AS
used_cannot_validate,

  -- Total opportunity counties (supply shortage)
  COUNTIF(new_zscore_tier < -1.0) AS total_undersupply,

  -- Validation rates
  ROUND(100.0 * COUNTIF(new_market_quality = 'True_White_Market')
      / NULLIF(COUNTIF(new_zscore_tier < -1.0), 0), 1) AS pct_validated_true,
  ROUND(100.0 * COUNTIF(new_market_quality = 'Cannot_Validate_Demand')
      / NULLIF(COUNTIF(new_zscore_tier < -1.0), 0), 1) AS pct_cannot_validate,

  -- Average characteristics of true white markets
  ROUND(AVG(CASE WHEN new_market_quality = 'True_White_Market'
```

```
        THEN demand_intensity_new END), 0) AS avg_demand_true_white,
 ROUND(AVG(CASE WHEN new_market_quality = 'True_White_Market'
        THEN total_population END), 0) AS avg_pop_true_white

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final`

GROUP BY metro_tier

ORDER BY
 CASE metro_tier
  WHEN 'Tier1_LargeMetro' THEN 1
  WHEN 'Tier2_MidMetro' THEN 2
  WHEN 'Tier3_SmallMetro' THEN 3
  WHEN 'Tier4_Rural' THEN 4
 END;


 -- ============================================
-- Step 3.1: Filter High-Quality White Markets for Investment Evaluation
-- Purpose: Select only demand-validated white markets for MCDA scoring
--
-- CRITICAL FILTERING CRITERIA:
--   1. Market Quality: Only 'True_White_Market' (strong demand validated)
--   2. Data Quality: Only counties with 'OK' data quality flag
--   3. Visitor Data: Only counties with visitor data available
--   4. Population: Minimum 50K for market viability
--
-- RATIONALE:
--   - 93% of counties cannot be validated due to SafeGraph coverage limits
```

```
--   - We ONLY recommend markets where both supply shortage AND strong demand are verified
--   - This ensures honest, evidence-based investment recommendations
--
-- Input: step2_5_white_market_final (2,527 counties)
-- Expected Output: ~39 counties (31 new car + 8 used car true white markets)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step3_1_white_markets_for_scoring
-- ============================================


CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step3_1_white_markets_for_scoring` AS


SELECT
 *,


 -- ============================================
 -- Add Investment Eligibility Flags
 -- These flags determine which markets enter new car vs used car scoring
 -- ============================================


 -- New car investment eligible
 CASE
  WHEN new_market_quality = 'True_White_Market'
     AND has_visitor_data = TRUE
     AND data_quality_flag = 'OK'
     AND total_population >= 50000
     AND new_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity', 'Medium_Opportunity')
    THEN TRUE
```

```
  ELSE FALSE
 END AS new_car_investment_eligible,

 -- Used car investment eligible
 CASE
  WHEN used_market_quality = 'True_White_Market'
     AND has_visitor_data = TRUE
     AND data_quality_flag = 'OK'
     AND total_population >= 50000
     AND used_opportunity_level IN ('High_Opportunity', 'Medium_High_Opportunity',
'Medium_Opportunity')
    THEN TRUE
  ELSE FALSE
 END AS used_car_investment_eligible,

 -- Combined eligibility (for integrated format consideration)
 CASE
  WHEN new_market_quality = 'True_White_Market'
     AND used_market_quality = 'True_White_Market'
     AND has_visitor_data = TRUE
     AND data_quality_flag = 'OK'
     AND total_population >= 100000  -- Higher threshold for integrated
    THEN TRUE
  ELSE FALSE
 END AS integrated_format_eligible

FROM `group22-fa25-mgmt58200-final.safegraph2.step2_5_white_market_final`

WHERE
```

```
 -- At least one market type must be True_White_Market
 (new_market_quality = 'True_White_Market' OR used_market_quality =
'True_White_Market')

 -- Must have visitor data for demand validation
 AND has_visitor_data = TRUE

 -- Must pass data quality checks
 AND data_quality_flag = 'OK'

 -- Minimum viable market size
 AND total_population >= 50000;



 -- ============================================
-- Step 3.2: Calculate 5-Criteria MCDA Investment Scores
-- Purpose: Comprehensive scoring for investment decision-making
--
-- SCORING FRAMEWORK (Total 100 points):
--   Criterion 1: Market Size (20 points) - P2 Population base
--   Criterion 2: Purchasing Power (20 points) - P4 Income indicators
--   Criterion 3: Supply Gap (30 points) - P5 Tier-stratified Z-scores
--   Criterion 4: Demand Intensity (20 points) - Visitor-based validation (DOUBLED
WEIGHT)
--   Criterion 5: Market Stability (10 points) - Homeownership as proxy
--
-- KEY INNOVATION: Demand Intensity weight increased from 10% to 20%
--   - Reflects importance of demand validation (Step 2.5 findings)
```

```
--   - Prevents investment in supply-shortage-but-low-demand markets
--   - 82% validation success rate proves effectiveness
--
-- Input: step3_1_white_markets_for_scoring (~39 counties)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step3_2_investment_scores
-- ============================================


CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step3_2_investment_scores` AS


SELECT
 *,


 -- ============================================
 -- CRITERION 1: Market Size Score (20 points)
 -- Rationale: Larger population = larger addressable market
 -- Threshold adjustment: Recognize Tier3 opportunities (lower thresholds)
 -- ============================================
 CASE
  WHEN total_population >= 500000 THEN 20  -- Very large metro
  WHEN total_population >= 300000 THEN 18  -- Large city
  WHEN total_population >= 200000 THEN 16  -- Mid-large city
  WHEN total_population >= 150000 THEN 14  -- Mid city
  WHEN total_population >= 100000 THEN 12  -- Small metro (Tier3 sweet spot)
  WHEN total_population >= 75000 THEN 10   -- Small city
  WHEN total_population >= 50000 THEN 8    -- Minimum viable
  ELSE 5
 END AS market_size_score,
```

```
-- ==========================================
-- CRITERION 2A: New Car Purchasing Power (20 points)
-- Based on high_income_pct (households earning $150K+)
-- Rationale: New car buyers typically high-income
-- ==========================================
CASE
  WHEN high_income_pct >= 30 THEN 20  -- Affluent market (>30%)
  WHEN high_income_pct >= 20 THEN 18  -- High income (20-30%)
  WHEN high_income_pct >= 15 THEN 16  -- Above average (15-20%)
  WHEN high_income_pct >= 12 THEN 14  -- Average (12-15%)
  WHEN high_income_pct >= 10 THEN 12  -- Below average (10-12%)
  WHEN high_income_pct >= 8 THEN 10   -- Lower income (8-10%)
  WHEN high_income_pct >= 6 THEN 8    -- Low income (6-8%)
  ELSE 5                      -- Very low income (<6%)
END AS new_purchasing_power_score,

-- ==========================================
-- CRITERION 2B: Used Car Purchasing Power (20 points)
-- Based on low_mid_income_pct (households earning $35K-$150K)
-- Rationale: Used car buyers typically middle-income
-- ==========================================
CASE
  WHEN low_mid_income_pct >= 65 THEN 20  -- Strong middle class (>65%)
  WHEN low_mid_income_pct >= 60 THEN 18  -- Large middle class (60-65%)
  WHEN low_mid_income_pct >= 55 THEN 16  -- Healthy middle class (55-60%)
  WHEN low_mid_income_pct >= 50 THEN 14  -- Average middle class (50-55%)
  WHEN low_mid_income_pct >= 45 THEN 12  -- Below average (45-50%)
  WHEN low_mid_income_pct >= 40 THEN 10  -- Small middle class (40-45%)
```

```
  ELSE 8                    -- Very small middle class (<40%)
END AS used_purchasing_power_score,

-- ============================================
-- CRITERION 3A: New Car Supply Gap Score (30 points)
-- Based on tier-stratified Z-score (from Step 0.3)
-- Rationale: Stronger shortage = larger opportunity
-- ============================================
CASE
  WHEN new_zscore_tier <= -1.5 THEN 30   -- Severe undersupply (critical shortage)
  WHEN new_zscore_tier <= -1.3 THEN 28   -- Very strong undersupply
  WHEN new_zscore_tier <= -1.2 THEN 26   -- Strong undersupply
  WHEN new_zscore_tier <= -1.1 THEN 24   -- Significant undersupply
  WHEN new_zscore_tier <= -1.0 THEN 22   -- Moderate undersupply
  WHEN new_zscore_tier <= -0.8 THEN 18   -- Slight undersupply
  WHEN new_zscore_tier <= -0.6 THEN 14   -- Minor undersupply
  ELSE 10                    -- Minimal or balanced
END AS new_supply_gap_score,

-- ============================================
-- CRITERION 3B: Used Car Supply Gap Score (30 points)
-- Same logic as new car
-- ============================================
CASE
  WHEN used_zscore_tier <= -1.5 THEN 30
  WHEN used_zscore_tier <= -1.3 THEN 28
  WHEN used_zscore_tier <= -1.2 THEN 26
  WHEN used_zscore_tier <= -1.1 THEN 24
```

```
  WHEN used_zscore_tier <= -1.0 THEN 22
  WHEN used_zscore_tier <= -0.8 THEN 18
  WHEN used_zscore_tier <= -0.6 THEN 14
  ELSE 10
END AS used_supply_gap_score,

-- ============================================
-- CRITERION 4A: New Car Demand Intensity Score (20 points)
-- DOUBLED WEIGHT from original 10 points
-- Based on market_activity_score_new (0-10 from Step 2.5)
--
-- RATIONALE FOR WEIGHT INCREASE:
--   - Step 2.5 validation showed 82% success rate (41/50 viable)
--   - Successfully filtered 6 low-demand markets (12% false positive)
--   - Direct measure of consumer interest (visits to existing dealers)
--   - More accurate than population-based proxies
--
-- SCORING LOGIC:
--   market_activity_score_new ranges 0-10 (calculated in Step 2.5)
--   - 10 = Top 25% demand (P75+)
--   - 8 = Top 50% demand (P50-P75)
--   - 6 = Bottom 50% demand (P25-P50)
--   - 4 = Bottom 25% demand (<P25)
--   - 0 = No visitor data
--
--   We double this to 0-20 scale for MCDA
-- ============================================
market_activity_score_new * 2 AS new_demand_intensity_score,  -- 0-20 points
```

```
-- ============================================
-- CRITERION 4B: Used Car Demand Intensity Score (20 points)
-- Same logic as new car
-- ============================================
market_activity_score_used * 2 AS used_demand_intensity_score,  -- 0-20 points

-- ============================================
-- CRITERION 5: Market Stability Score (10 points)
-- Based on homeownership_rate
-- Rationale: High homeownership = stable, committed residents = repeat customers
-- ============================================
CASE
  WHEN homeownership_rate >= 80 THEN 10  -- Very stable (>80% homeowners)
  WHEN homeownership_rate >= 75 THEN 9   -- Highly stable (75-80%)
  WHEN homeownership_rate >= 70 THEN 8   -- Stable (70-75%)
  WHEN homeownership_rate >= 65 THEN 7   -- Moderately stable (65-70%)
  WHEN homeownership_rate >= 60 THEN 6   -- Average stability (60-65%)
  WHEN homeownership_rate >= 55 THEN 5   -- Below average (55-60%)
  WHEN homeownership_rate >= 50 THEN 4   -- Low stability (50-55%)
  ELSE 3                         -- Very low stability (<50%)
END AS stability_score,

-- ============================================
-- TOTAL INVESTMENT SCORES (100 points maximum)
-- ============================================

-- New Car Investment Score
```

```
(
  -- Criterion 1: Market Size (20%)
  CASE
    WHEN total_population >= 500000 THEN 20
    WHEN total_population >= 300000 THEN 18
    WHEN total_population >= 200000 THEN 16
    WHEN total_population >= 150000 THEN 14
    WHEN total_population >= 100000 THEN 12
    WHEN total_population >= 75000 THEN 10
    WHEN total_population >= 50000 THEN 8
    ELSE 5
  END
  +
  -- Criterion 2: Purchasing Power (20%)
  CASE
    WHEN high_income_pct >= 30 THEN 20
    WHEN high_income_pct >= 20 THEN 18
    WHEN high_income_pct >= 15 THEN 16
    WHEN high_income_pct >= 12 THEN 14
    WHEN high_income_pct >= 10 THEN 12
    WHEN high_income_pct >= 8 THEN 10
    WHEN high_income_pct >= 6 THEN 8
    ELSE 5
  END
  +
  -- Criterion 3: Supply Gap (30%)
  CASE
    WHEN new_zscore_tier <= -1.5 THEN 30
```

```
    WHEN new_zscore_tier <= -1.3 THEN 28
    WHEN new_zscore_tier <= -1.2 THEN 26
    WHEN new_zscore_tier <= -1.1 THEN 24
    WHEN new_zscore_tier <= -1.0 THEN 22
    WHEN new_zscore_tier <= -0.8 THEN 18
    WHEN new_zscore_tier <= -0.6 THEN 14
    ELSE 10
  END
  +
  -- Criterion 4: Demand Intensity (20% - DOUBLED)
  market_activity_score_new * 2
  +
  -- Criterion 5: Stability (10%)
  CASE
    WHEN homeownership_rate >= 80 THEN 10
    WHEN homeownership_rate >= 75 THEN 9
    WHEN homeownership_rate >= 70 THEN 8
    WHEN homeownership_rate >= 65 THEN 7
    WHEN homeownership_rate >= 60 THEN 6
    WHEN homeownership_rate >= 55 THEN 5
    WHEN homeownership_rate >= 50 THEN 4
    ELSE 3
  END
 ) AS new_car_investment_score,

 -- Used Car Investment Score (parallel calculation)
 (
  CASE
```

```
    WHEN total_population >= 500000 THEN 20
    WHEN total_population >= 300000 THEN 18
    WHEN total_population >= 200000 THEN 16
    WHEN total_population >= 150000 THEN 14
    WHEN total_population >= 100000 THEN 12
    WHEN total_population >= 75000 THEN 10
    WHEN total_population >= 50000 THEN 8
    ELSE 5
  END
  +
  CASE
    WHEN low_mid_income_pct >= 65 THEN 20
    WHEN low_mid_income_pct >= 60 THEN 18
    WHEN low_mid_income_pct >= 55 THEN 16
    WHEN low_mid_income_pct >= 50 THEN 14
    WHEN low_mid_income_pct >= 45 THEN 12
    WHEN low_mid_income_pct >= 40 THEN 10
    ELSE 8
  END
  +
  CASE
    WHEN used_zscore_tier <= -1.5 THEN 30
    WHEN used_zscore_tier <= -1.3 THEN 28
    WHEN used_zscore_tier <= -1.2 THEN 26
    WHEN used_zscore_tier <= -1.1 THEN 24
    WHEN used_zscore_tier <= -1.0 THEN 22
    WHEN used_zscore_tier <= -0.8 THEN 18
    WHEN used_zscore_tier <= -0.6 THEN 14
```

```
    ELSE 10
  END
  +
  market_activity_score_used * 2
  +
  CASE
    WHEN homeownership_rate >= 80 THEN 10
    WHEN homeownership_rate >= 75 THEN 9
    WHEN homeownership_rate >= 70 THEN 8
    WHEN homeownership_rate >= 65 THEN 7
    WHEN homeownership_rate >= 60 THEN 6
    WHEN homeownership_rate >= 55 THEN 5
    WHEN homeownership_rate >= 50 THEN 4
    ELSE 3
  END
 ) AS used_car_investment_score

FROM `group22-fa25-mgmt58200-
final.safegraph2.step3_1_white_markets_for_scoring`;


-- ============================================
-- Step 3.2: Calculate 5-Criteria MCDA Investment Scores
-- Purpose: Comprehensive scoring for investment decision-making
--
-- SCORING FRAMEWORK (Total 100 points):
--   Criterion 1: Market Size (20 points)
--   Criterion 2: Purchasing Power (20 points)
--   Criterion 3: Supply Gap (30 points)
--   Criterion 4: Demand Intensity (20 points) - DOUBLED from original
```

```
--   Criterion 5: Market Stability (10 points)
--
-- Input: step3_1_white_markets_for_scoring (36 counties)
-- Output: group22-fa25-mgmt58200-final.safegraph2.step3_2_investment_scores
-- ============================================

CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-
final.safegraph2.step3_2_investment_scores` AS

SELECT
 *,

 -- ============================================
 -- CRITERION 1: Market Size Score (20 points)
 -- ============================================
 CASE
  WHEN total_population >= 1000000 THEN 20  -- Mega metro (1M+)
  WHEN total_population >= 500000 THEN 18   -- Very large (500K-1M)
  WHEN total_population >= 300000 THEN 16   -- Large (300K-500K)
  WHEN total_population >= 200000 THEN 14   -- Mid-large (200K-300K)
  WHEN total_population >= 150000 THEN 12   -- Mid (150K-200K)
  WHEN total_population >= 100000 THEN 10   -- Small metro (100K-150K)
  WHEN total_population >= 75000 THEN 8     -- Small city (75K-100K)
  WHEN total_population >= 50000 THEN 6     -- Minimum viable (50K-75K)
  ELSE 4
 END AS market_size_score,

 -- ============================================
 -- CRITERION 2A: New Car Purchasing Power (20 points)
```

```
-- ============================================
CASE
  WHEN high_income_pct >= 40 THEN 20   -- Very affluent (>40%)
  WHEN high_income_pct >= 30 THEN 18   -- Affluent (30-40%)
  WHEN high_income_pct >= 25 THEN 16   -- High income (25-30%)
  WHEN high_income_pct >= 20 THEN 14   -- Above average (20-25%)
  WHEN high_income_pct >= 15 THEN 12   -- Average (15-20%)
  WHEN high_income_pct >= 12 THEN 10   -- Below average (12-15%)
  WHEN high_income_pct >= 10 THEN 8    -- Lower income (10-12%)
  WHEN high_income_pct >= 8 THEN 6     -- Low income (8-10%)
  WHEN high_income_pct >= 6 THEN 4     -- Very low (6-8%)
  ELSE 2                       -- Minimal (<6%)
END AS new_purchasing_power_score,

-- ============================================
-- CRITERION 2B: Used Car Purchasing Power (20 points)
-- ============================================
CASE
  WHEN low_mid_income_pct >= 70 THEN 20   -- Very strong middle class (>70%)
  WHEN low_mid_income_pct >= 65 THEN 18   -- Strong middle class (65-70%)
  WHEN low_mid_income_pct >= 60 THEN 16   -- Large middle class (60-65%)
  WHEN low_mid_income_pct >= 55 THEN 14   -- Healthy middle class (55-60%)
  WHEN low_mid_income_pct >= 50 THEN 12   -- Average (50-55%)
  WHEN low_mid_income_pct >= 45 THEN 10   -- Below average (45-50%)
  WHEN low_mid_income_pct >= 40 THEN 8    -- Small middle class (40-45%)
  WHEN low_mid_income_pct >= 35 THEN 6    -- Very small (35-40%)
  ELSE 4                        -- Minimal (<35%)
END AS used_purchasing_power_score,
```

```
-- ==========================================
-- CRITERION 3A: New Car Supply Gap Score (30 points)
-- ==========================================
CASE
  WHEN new_zscore_tier <= -1.6 THEN 30   -- Extreme (critical shortage)
  WHEN new_zscore_tier <= -1.5 THEN 29   -- Severe+
  WHEN new_zscore_tier <= -1.4 THEN 28   -- Severe
  WHEN new_zscore_tier <= -1.3 THEN 27   -- Very strong
  WHEN new_zscore_tier <= -1.2 THEN 26   -- Strong+
  WHEN new_zscore_tier <= -1.1 THEN 24   -- Strong
  WHEN new_zscore_tier <= -1.0 THEN 22   -- Significant
  WHEN new_zscore_tier <= -0.9 THEN 20   -- Moderate+
  WHEN new_zscore_tier <= -0.8 THEN 18   -- Moderate
  ELSE 15                      -- Slight
END AS new_supply_gap_score,

-- ==========================================
-- CRITERION 3B: Used Car Supply Gap Score (30 points)
-- ==========================================
CASE
  WHEN used_zscore_tier <= -1.7 THEN 30
  WHEN used_zscore_tier <= -1.6 THEN 29
  WHEN used_zscore_tier <= -1.5 THEN 28
  WHEN used_zscore_tier <= -1.4 THEN 27
  WHEN used_zscore_tier <= -1.3 THEN 26
  WHEN used_zscore_tier <= -1.2 THEN 25
  WHEN used_zscore_tier <= -1.1 THEN 24
```

```
  WHEN used_zscore_tier <= -1.0 THEN 22
  WHEN used_zscore_tier <= -0.9 THEN 20
  ELSE 18
END AS used_supply_gap_score,

-- ============================================
-- CRITERION 4A: New Car Demand Intensity Score (20 points)
-- DOUBLED WEIGHT - Key differentiator from supply-only analysis
-- ============================================
market_activity_score_new * 2 AS new_demand_intensity_score,  -- 0-20 points

-- ============================================
-- CRITERION 4B: Used Car Demand Intensity Score (20 points)
-- ============================================
market_activity_score_used * 2 AS used_demand_intensity_score,  -- 0-20 points

-- ============================================
-- CRITERION 5: Market Stability Score (10 points)
-- ============================================
CASE
  WHEN homeownership_rate >= 85 THEN 10   -- Very stable (>85%)
  WHEN homeownership_rate >= 80 THEN 9    -- Highly stable (80-85%)
  WHEN homeownership_rate >= 75 THEN 8    -- Stable (75-80%)
  WHEN homeownership_rate >= 70 THEN 7    -- Moderately stable (70-75%)
  WHEN homeownership_rate >= 65 THEN 6    -- Average (65-70%)
  WHEN homeownership_rate >= 60 THEN 5    -- Below average (60-65%)
  WHEN homeownership_rate >= 55 THEN 4    -- Low stability (55-60%)
  WHEN homeownership_rate >= 50 THEN 3    -- Very low (50-55%)
```

```
  ELSE 2                    -- Minimal (<50%)
 END AS stability_score,

 -- ============================================
 -- TOTAL INVESTMENT SCORES (100 points)
 -- ============================================

 -- New Car Investment Score
 (
  CASE
   WHEN total_population >= 1000000 THEN 20
   WHEN total_population >= 500000 THEN 18
   WHEN total_population >= 300000 THEN 16
   WHEN total_population >= 200000 THEN 14
   WHEN total_population >= 150000 THEN 12
   WHEN total_population >= 100000 THEN 10
   WHEN total_population >= 75000 THEN 8
   WHEN total_population >= 50000 THEN 6
   ELSE 4
  END
  +
  CASE
   WHEN high_income_pct >= 40 THEN 20
   WHEN high_income_pct >= 30 THEN 18
   WHEN high_income_pct >= 25 THEN 16
   WHEN high_income_pct >= 20 THEN 14
   WHEN high_income_pct >= 15 THEN 12
   WHEN high_income_pct >= 12 THEN 10
```

```
  WHEN high_income_pct >= 10 THEN 8
  WHEN high_income_pct >= 8 THEN 6
  WHEN high_income_pct >= 6 THEN 4
  ELSE 2
 END
 +
 CASE
  WHEN new_zscore_tier <= -1.6 THEN 30
  WHEN new_zscore_tier <= -1.5 THEN 29
  WHEN new_zscore_tier <= -1.4 THEN 28
  WHEN new_zscore_tier <= -1.3 THEN 27
  WHEN new_zscore_tier <= -1.2 THEN 26
  WHEN new_zscore_tier <= -1.1 THEN 24
  WHEN new_zscore_tier <= -1.0 THEN 22
  WHEN new_zscore_tier <= -0.9 THEN 20
  WHEN new_zscore_tier <= -0.8 THEN 18
  ELSE 15
 END
 +
 market_activity_score_new * 2
 +
 CASE
  WHEN homeownership_rate >= 85 THEN 10
  WHEN homeownership_rate >= 80 THEN 9
  WHEN homeownership_rate >= 75 THEN 8
  WHEN homeownership_rate >= 70 THEN 7
  WHEN homeownership_rate >= 65 THEN 6
  WHEN homeownership_rate >= 60 THEN 5
```

```
    WHEN homeownership_rate >= 55 THEN 4
    WHEN homeownership_rate >= 50 THEN 3
    ELSE 2
  END
 ) AS new_car_investment_score,

 -- Used Car Investment Score
 (
  CASE
    WHEN total_population >= 1000000 THEN 20
    WHEN total_population >= 500000 THEN 18
    WHEN total_population >= 300000 THEN 16
    WHEN total_population >= 200000 THEN 14
    WHEN total_population >= 150000 THEN 12
    WHEN total_population >= 100000 THEN 10
    WHEN total_population >= 75000 THEN 8
    WHEN total_population >= 50000 THEN 6
    ELSE 4
  END
  +
  CASE
    WHEN low_mid_income_pct >= 70 THEN 20
    WHEN low_mid_income_pct >= 65 THEN 18
    WHEN low_mid_income_pct >= 60 THEN 16
    WHEN low_mid_income_pct >= 55 THEN 14
    WHEN low_mid_income_pct >= 50 THEN 12
    WHEN low_mid_income_pct >= 45 THEN 10
    WHEN low_mid_income_pct >= 40 THEN 8
```

```
  WHEN low_mid_income_pct >= 35 THEN 6
  ELSE 4
 END
 +
 CASE
  WHEN used_zscore_tier <= -1.7 THEN 30
  WHEN used_zscore_tier <= -1.6 THEN 29
  WHEN used_zscore_tier <= -1.5 THEN 28
  WHEN used_zscore_tier <= -1.4 THEN 27
  WHEN used_zscore_tier <= -1.3 THEN 26
  WHEN used_zscore_tier <= -1.2 THEN 25
  WHEN used_zscore_tier <= -1.1 THEN 24
  WHEN used_zscore_tier <= -1.0 THEN 22
  WHEN used_zscore_tier <= -0.9 THEN 20
  ELSE 18
 END
 +
 market_activity_score_used * 2
 +
 CASE
  WHEN homeownership_rate >= 85 THEN 10
  WHEN homeownership_rate >= 80 THEN 9
  WHEN homeownership_rate >= 75 THEN 8
  WHEN homeownership_rate >= 70 THEN 7
  WHEN homeownership_rate >= 65 THEN 6
  WHEN homeownership_rate >= 60 THEN 5
  WHEN homeownership_rate >= 55 THEN 4
  WHEN homeownership_rate >= 50 THEN 3
```

```
    ELSE 2
  END
 ) AS used_car_investment_score


FROM `group22-fa25-mgmt58200-final.safegraph2.step3_1_white_markets_for_scoring`;


-- ============================================
-- Step 3.3: Business Format Recommendation and Risk Assessment
-- Purpose: Determine optimal dealership format and investment risk level
--
-- BUSINESS FORMAT LOGIC (Based on Step 2.1 Complementarity Findings):
--   - Tier1/2: Strong complementarity (r=0.73-0.76) → require both undersupply
--   - Tier3: Moderate complementarity (r=0.54) → flexible format options
--   - High income + new undersupply → New car primary
--   - Middle income + used undersupply → Used car primary
--   - Both undersupply + large market → Integrated format
--
-- RISK ASSESSMENT LOGIC:
--   - White market confidence level (from Step 2.4)
--   - Homeownership rate (stability proxy)
--   - Income level (purchasing power security)
--   - Data quality (certainty of estimates)
--
-- Input: step3_2_investment_scores
-- Output: group22-fa25-mgmt58200-final.safegraph2.step3_3_format_and_risk
-- ============================================
```

```
CREATE OR REPLACE TABLE `group22-fa25-mgmt58200-final.safegraph2.step3_3_format_and_risk` AS

SELECT
 *,

 -- ==========================================
 -- BUSINESS FORMAT RECOMMENDATION
 -- ==========================================
 CASE
  -- Format 1: NEW CAR PRIMARY
  -- Conditions: High income + strong new car undersupply + new car true white market
  WHEN new_car_investment_eligible = TRUE
    AND high_income_pct >= 12
    AND new_zscore_tier <= -1.0
    AND new_car_investment_score >= 75
   THEN 'New_Car_Dealership_Primary'

  -- Format 2: USED CAR PRIMARY
  -- Conditions: Strong middle income + strong used car undersupply + used car true white market
  WHEN used_car_investment_eligible = TRUE
    AND low_mid_income_pct >= 50
    AND used_zscore_tier <= -1.0
    AND used_car_investment_score >= 75
   THEN 'Used_Car_Dealership_Primary'

  -- Format 3: INTEGRATED (NEW + USED)
  -- Conditions: Large market + both markets undersupplied + both eligible
```

```
  WHEN integrated_format_eligible = TRUE
     AND total_population >= 150000
     AND new_zscore_tier <= -0.8
     AND used_zscore_tier <= -0.8
     AND (new_car_investment_score + used_car_investment_score) / 2 >= 75
   THEN 'Integrated_Dealership_Both'

  -- Format 4: NEW CAR SECONDARY
  -- Conditions: Moderate income + new car opportunity better than used
  WHEN new_car_investment_eligible = TRUE
     AND new_car_investment_score > used_car_investment_score
     AND new_car_investment_score >= 70
   THEN 'New_Car_Dealership_Secondary'

  -- Format 5: USED CAR SECONDARY
  -- Conditions: Used car opportunity better than new
  WHEN used_car_investment_eligible = TRUE
     AND used_car_investment_score > new_car_investment_score
     AND used_car_investment_score >= 70
   THEN 'Used_Car_Dealership_Secondary'

  -- Default: No clear recommendation
  ELSE 'Requires_Further_Analysis'
END AS recommended_format,

-- ============================================
-- RISK LEVEL ASSESSMENT
-- Considers multiple risk dimensions:
```

```
-- 1. White market confidence (supply-demand alignment)
-- 2. Homeownership (market stability)
-- 3. Income level (purchasing power security)
-- 4. Demand validation (data certainty)
-- ========================================
CASE
 -- LOW RISK: All factors favorable
 WHEN white_market_confidence = 'High_Confidence'
    AND homeownership_rate >= 70
    AND high_income_pct >= 12
    AND market_activity_score_new >= 8
  THEN 'Low_Risk'

 -- LOW-MEDIUM RISK: Most factors favorable
 WHEN white_market_confidence IN ('High_Confidence', 'Medium_Confidence')
    AND homeownership_rate >= 65
    AND (high_income_pct >= 10 OR low_mid_income_pct >= 50)
    AND market_activity_score_new >= 6
  THEN 'Low_Medium_Risk'

 -- MEDIUM RISK: Mixed factors
 WHEN white_market_confidence IN ('High_Confidence', 'Medium_Confidence')
    AND homeownership_rate >= 60
    AND market_activity_score_new >= 6
  THEN 'Medium_Risk'

 -- MEDIUM-HIGH RISK: Some concerning factors
 WHEN white_market_confidence = 'Low_Confidence'
```

```
      OR homeownership_rate < 60
      OR market_activity_score_new < 6
    THEN 'Medium_High_Risk'

  -- HIGH RISK: Multiple concerning factors
  ELSE 'High_Risk'
 END AS risk_level,

 -- ============================================
 -- INVESTMENT PRIORITY RANKING
 -- Lower number = higher priority
 -- ============================================
 ROW_NUMBER() OVER (
  PARTITION BY
   CASE
    WHEN new_car_investment_eligible THEN 'new_car'
    WHEN used_car_investment_eligible THEN 'used_car'
    ELSE 'other'
   END
  ORDER BY
   CASE
    WHEN new_car_investment_eligible THEN new_car_investment_score
    WHEN used_car_investment_eligible THEN used_car_investment_score
    ELSE 0
   END DESC,
   total_population DESC
 ) AS investment_priority_rank
```

```
FROM `group22-fa25-mgmt58200-final.safegraph2.step3_2_investment_scores`

WHERE new_car_investment_eligible = TRUE
  OR used_car_investment_eligible = TRUE;
```

Group 22

Appendix 10 - AI Support Statement

How we used AI in this project

Throughout the project, we used AI tools to enhance our workflow in several key areas:

1. Understanding data structures and relationships – AI was used to assist in interpreting the relationships among databases, clarifying variable definitions, and ensuring consistent understanding across data sources.

2. Clarifying technical terminology – We consulted AI to explain specialized terms and concepts related to the datasets and analytical procedures.

3. Improving documentation and communication – AI tools helped summarize meeting discussions, organize notes, and ensure that key decisions and next steps were clearly documented.

4. Supporting analytical reasoning – When completing specific analytical or data preparation tasks, we used AI to explore applicable methods, understand their underlying principles, and identify the conditions under which each approach would be most suitable.

5. Language refinement – AI assisted in polishing the language of our written materials, including reports and presentations, to make them clearer, more concise, and more professional.

Platform used

All AI assistance was provided through OpenAI's ChatGPT platform (GPT-5 model).

Example prompts used:

- "Explain the meaning and possible relationship of these database variables."
- "Summarize the main discussion points and action items from this meeting."
- "Clarify this technical term in the context of data analysis."
- "Improve the clarity and professionalism of this paragraph."
- "Help me write this SQL code to implement the functionality of XXX (a very complex logic) and save the results in the database."

Ethical use statement

AI was used as a supplementary tool to enhance understanding and communication. All analytical decisions, coding, and interpretations were conducted and verified by the research team.