Course Project Final Report

Title: *Income and Liquor Spending Behavior Across Iowa Counties: A Data-Driven Policy Analysis*
Prepared by: *Shreyansh Budhia, William Brewster*, and *Ramzi Nia*

1. Introduction (5 pts)

Our analysis was conducted for a **state-level public policy decision maker in Iowa**, such as the Iowa Department of Public. These organizations operate in a decision environment that balances **economic revenue generation**, **public health concerns**, and **community well-being**. Liquor sales represent a major source of state revenue, but alcohol consumption is also associated with long-term public health and social costs.

The broader decision climate is influenced by:

- Rising healthcare expenditures,
- Public concerns over substance abuse,
- And debates over how alcohol access and taxation should be structured across income groups and geographic regions.

Research Question

Is there a meaningful relationship between median household income and liquor spending per capita across Iowa counties?

Policy Relevance

Understanding whether higher-income counties consume more alcohol on a per-person basis helps decision makers:

- Evaluate the **equity of alcohol taxation policies**,
- Identify whether alcohol consumption burdens fall disproportionately on certain income groups,
- And determine where **education, prevention, or regulatory resources** should be targeted.

Policy decisions based on this analysis affect:

- State revenue collection,
- Public health planning,
- And long-term community outcomes related to alcohol misuse.

2. Data Summary (5 pts)

Liquor Sales Dataset (2020)

This dataset contains county-level liquor sales information broken down by product category. Key variables include:

- county
- sale.dollars
- sale.volume
- category

Each observation represents aggregated sales for a specific product category within a county. Because counties vary widely in population size, direct comparison of total sales alone would introduce **population bias**.

American Community Survey Dataset (2019)

The ACS dataset provides socioeconomic indicators at the county level. Variables used include:

- county
- income (median household income)
- population

Methodological Implications

Because the two datasets come from different sources and different years:

- We assumed **relative short-term income stability** between 2019 and 2020.
- We transformed total liquor sales into **per-capita values** to ensure fair comparisons.
- An **inner join by county** ensured consistent geographic alignment.

Descriptive Summary

Across Iowa counties:

- Median household incomes show clear regional variation.
- Population ranges from small rural counties to large urban counties.
- Liquor sales totals vary substantially, largely driven by population size before normalization.

3. Data Analytics (50 pts)

Aggregation and Feature Engineering

We first aggregated liquor sales at the county level:

$$\text{Total County Liquor Sales} = \sum \text{sale.dollars}$$

We then constructed a **per-capita spending variable**:

$$\text{Liquor Dollars Per Capita} = \frac{\text{Total County Liquor Sales}}{\text{Population}}$$

This transformation was essential to remove population-driven distortion.

Merging Process

The sales and ACS datasets were merged by county using:

- SQL for initial validation,
- R for data cleaning and aggregation,
- Python for final analysis, correlation testing, and visualization.

All three tools produced **consistent numerical results**, confirming the reliability of the workflow.

Correlation Analysis

We computed the **Pearson correlation coefficient** between:

- Median household income
- Liquor dollars per capita

Final Result

$$r = -0.106$$

Interpretation

This value indicates:

- A **very weak negative relationship**

- Income explains **almost none** of the variation in liquor spending per capita
- Higher-income counties do **not** appear to consume more liquor per person
- Consitent coefficient using R and Python

The sign of the correlation suggests a slight tendency for lower-income counties to spend marginally more on liquor per capita, but the magnitude is too small to imply a strong behavioral pattern.

Visualization Insights

Our scatter plot shows:

- A wide vertical dispersion of liquor spending across all income levels
- No visible upward or downward trend
- A nearly flat regression line, confirming the statistical result

This visual reinforces the conclusion that:

Liquor consumption behavior in Iowa does not scale predictably with income.

Suggested Analytic Extension (Excursion)

A meaningful next extension would be to:

- Decompose liquor spending by **product type** (beer, wine, spirits),
- Control for **urbanization, tourism, and age distribution**,
- And apply **panel data across multiple years**.

These additional variables would allow for causal modeling rather than simple correlation.

Advantages & Challenges of the Software Tools

Tool	Advantages	Challenges
SQL	Excellent for large-scale aggregation and joins	Limited visualization, less flexible transformation
R	Powerful statistical tools and good for cleaning data	Requires scripting fluency
Python	Best balance of modeling, automation, and visualization	File path management and environments can be complex

Using all three tools allowed us to:

- Cross-validate results,
- Leverage each platform's strengths,
- And improve overall methodological robustness.

4. Conclusion (10 pts)

This project applied a full analytical pipeline:

1. Data ingestion and aggregation
2. Normalization through per-capita transformation
3. Correlation testing
4. Visualization and interpretation
5. Cross-platform validation

Final Answer to the Research Question

Median household income is not a meaningful predictor of liquor spending per capita across Iowa counties.

While small statistical noise exists, there is no economically significant relationship between income and individual liquor consumption at the county level.

Limitations

- Single-year analysis
- No behavioral controls (age, tourism, density)
- No causal inference
- County-level averages mask within-county inequality

Recommended Future Work

- Multi-year panel analysis
- Product-specific consumption modeling
- Incorporation of public health outcomes

During our initial SQL-based analysis, we observed a positive relationship between median household income and total liquor sales at the county level. This result reflects the fact that higher-income counties also tend to have larger populations, greater commercial density, and higher overall transaction volume. As a result, these counties naturally generate more total liquor revenue.

However, when we normalized sales by population in Python and R to compute liquor spending on a per-capita basis, this positive relationship disappeared. The per-capita correlation between income and liquor spending was approximately –0.106, indicating a very weak and slightly negative relationship.

This difference is not a contradiction but rather a change in the underlying unit of analysis. Total sales capture market size effects, while per-capita spending captures individual consumption behavior. Together, these findings suggest that while higher-income counties produce more total liquor revenue, higher-income individuals do not consume meaningfully more liquor than lower-income individuals.

5. Policy Recommendation (10 pts)

Policy Decision Facing the State

The decision maker is evaluating whether to:

- Adjust alcohol taxes based on income distribution, or
- Implement targeted prevention programs in specific counties.

Data-Driven Recommendation

Because income is **not strongly linked** to liquor consumption per capita, we recommend:

Targeting alcohol policy interventions based on consumption intensity and public health risk rather than income levels alone.

This suggests:

- Taxation policy should remain **uniform rather than income-adjusted**
- Prevention resources should focus on **high per-capita consumption counties, regardless of income**

First-Order Effects

- More efficient allocation of prevention funding
- Reduced misclassification of high-risk counties
- More equitable regulatory enforcement

Second-Order Effects

- Lower long-term healthcare costs
- Reduced alcohol-related crime and injuries
- Improved workforce productivity

Benefits

- Evidence-based targeting
- Reduced waste in public spending
- Improved health outcomes in high-risk regions

Risks

- Public resistance if high-consumption counties feel singled out
- Short-term revenue volatility in affected counties

These risks can be mitigated through:

- Transparent data communication
- Gradual implementation
- Community-based education programs