Alcohol Beverage Services (ABS) Store-Level Performance, Baskets, and Risk

Course: DATA 205

CRN: 202620

Student: Arash Nateghian

Professor: Lori Perine

December 10, 2025

1. Introduction and Sponsor Context

Montgomery County Alcohol Beverage Services (ABS) operates a county-owned retail system with a dual mandate:

- serve the public and protect public safety, and
- generate as much profit as possible for the county budget.

Because all off-premises spirits sales run through this controlled system, even modest improvements in store-level performance, replenishment, or product mix can translate into meaningful revenue gains and public benefit.

The sponsor's stated goal is not simply to "produce charts," but to:

- determine effective approaches to store-level analysis using available methods, and
- gain "a better understanding of the metrics to observe each individual store."

ABS is also seeking guidance on replenishment strategies by location how to translate observed buying patterns into smarter purchasing decisions instead of treating all stores as if they behave identically.

This report uses detailed transactions, store, and demographic data to:

- construct store-level KPIs for sales, traffic, basket value, and sales per square foot
- quantify how category mix and customer behavior drive performance
- identifying over- and under-performing locations relative to a multivariate benchmark; and
- use market-basket analysis and bootstrapping to understand baskets and uncertainty around key metrics

The objective is not a one-off analysis, but a reusable monitoring framework that ABS can incorporate into ongoing operations.

2. Data and Methods

2.1 Data Sources

Three main datasets underpin the analysis:

1. **TransactionData.csv** – complete POS line-item data, including STORENAME, TRANSACTIONID, DESCRIPTION, CLASSIFICATIONDEPARTMENT, CLASSIFICATIONTYPE, SIZE, PACKUNIT, LINEQTY, NETAMOUNT, and TRANSDATE.
2. **Designation.csv** – store-level attributes, especially **SquareFootage**.

3. **Demographics.csv** – basic trade-area characteristics, including **TotalPopulation**, **Older25Years**, and **PovertyLevel**.

Two fiscal years are covered:

- **FY23–24**: July 1, 2023 – June 30, 2024
- **FY24–25**: July 1, 2024 – June 30, 2025

2.2 Core Cleaning and Standardization

The raw transaction data contains a variety of legacy codes, inconsistent strings, and non-retail records. The core cleaning steps include:

- Trimming whitespace and standardizing case for CLASSIFICATIONDEPARTMENT, CLASSIFICATIONTYPE, STORENAME, SIZE, PACKUNIT, TAGDESC, and DESCRIPTION.
- Converting TRANSDATE to a proper Date object.
- Normalizing **store names** (e.g., renaming "White Oak" to **"White Oak Town Center"**).
- Normalizing legacy **size labels** (e.g., 5LTR → 5L, 3LTR → 3L, 1LTR → 1L).
- Normalizing **PACKUNIT** values:
    - pure numeric codes are converted to Btl or xpk (e.g., "6" → "6pk"),
    - padded codes like "06pk" are simplified to "6pk",
    - "01pk" and "1" are treated as single bottles.
- Backfilling missing CLASSIFICATIONTYPE values for **BEER** rows by setting them to the department when missing or "NULL".

Non-retail and invalid rows are then filtered out:

- Dropping store **"Westwood"**, donations, non-inventory rows, and records where SIZE is "UNIT" or missing.
- Restricting to TAGDESC == "STOCK" and excluding "NULL"/NA TAGDESC values.
- Enforcing basic numeric validity:
    - NETAMOUNT and LINEQTY must be finite.
    - LINEQTY > 0.
    - TRANSACTIONID, STORENAME, and CLASSIFICATIONDEPARTMENT must be non-empty and non-missing.

A **UnitType** variable is defined as:

- PACKUNIT for **BEER**,
- SIZE for all other departments.

Refunds are flagged (IsRefund = NETAMOUNT < 0), and a **unit price** (FEE) is calculated as NETAMOUNT / LINEQTY only for non-refund rows. Final description strings are cleaned to uppercase alphanumeric plus spaces, with extra whitespace removed.

This process yields a clean transaction dataset suitable for KPI construction, modeling, and market-basket analysis.

2.3 Fiscal Years, Holidays, and Day Types

For calendar analysis, a clone of the transaction data (transactionsFS) is used. Each TRANSDATE is tagged with a **DayType**:

- **Holiday** – based on a custom holiday calendar that includes major U.S. holidays and two New Year holiday ranges for both FY23–24 and FY24–25.
- **Weekend** – Saturdays and Sundays.
- **Weekday** – all remaining days.

Daily total sales are computed across all stores, then averaged by DayType. The counts in transactionsFS indicate:

- Holidays: 59,073 transactions
- Weekdays: 7,111,358 transactions
- Weekends: 3,019,959 transactions

Average daily sales are markedly higher for **weekends and holidays** than weekdays, reinforcing the need to treat calendar position as a first-order driver for staffing and inventory decisions.

3. Store Performance – Revenue and Ticket Size

Total revenue and average basket value are profiled by store using **non-refund** transactions.

- **TotalSales** per store are computed and visualized in a descending bar chart.
- **Average ticket size** (AvgTicket = TotalSales / Transactions) is calculated and displayed by store.

Top revenue stores include high-traffic, high-capacity locations such as **Montrose**, **Hampden Lane**, and other major sites. Average basket values typically lie in the **$30–$45** range, but as shown later by the bootstrap analysis, these averages are inflated by a minority of very large baskets.

4. Top Items by Quantity and Revenue

Two perspectives are used:

1. **Top 20 items by units sold** – dominated by high-frequency, often mid-priced items, especially in LIQ and BEER.
2. **Top 20 items by revenue** – dominated by higher-priced items and core spirits SKUs with both high price and reasonable volume.

These views highlight which SKUs drive volume vs. revenue and help identify candidates for priority replenishment and potential promotional focus.

5. Category-Level EDA: Departments and Sizes

At the **classification type** level, the top revenue types are identified and plotted. At the **department** level:

- LIQ generates the majority of total sales,
- WINE and BEER follow,
- MISC is a small but non-zero share.

Within departments, the **UnitType** analysis shows:

- BEER is dominated by **6-packs**, followed by single bottles and 12-packs.
- LIQ is dominated by **750ml** and **1.75L** bottles.
- A long tail of niche sizes (minis, nonstandard pack sizes) contributes relatively small shares of revenue and quantity.

A faceted plot of TotalSales by **UnitType** within each department reinforces that a subset of core formats drives most of the business. This is important for replenishment strategy:

- Maintain depth and reliability in core formats.
- Evaluate low-volume formats for potential rationalization or targeted use.

A separate "bubble plot" of **size/pack frequencies** by department shows the density of SKUs by usage frequency. It provides a different lens on clutter and long-tail complexity.

6. Fiscal Year Comparison and Calendar Alignment

To compare **FY23–24** and **FY24–25**, daily sales are aligned by **weekday within month**:

- For each year, daily sales are aggregated and then "weekday-aligned" using a DayAligned index that corrects for which day of the week the month starts on.
- Line plots by month show how day-of-week patterns drive spikes around weekends and holidays in both years.

A separate monthly comparison aggregates transactions by YearMonth:

- Two fiscal-year series (July–June for 23–24 and 24–25) are plotted together.
- This reveals which months in 24–25 are ahead or behind the prior fiscal year, after controlling for calendar-month alignment.

Overall, 24–25 typically tracks or exceeds 23–24 in many months, but the variation across months and holidays underscores the need for seasonally-aware planning rather than relying on simple year-over-year totals.

7. Store-Level KPIs: Sales, Traffic, Baskets, and SqFt

For FY24–25, a core set of **store-level KPIs** is constructed:

From transactions (non-refund) and square footage:

- **TotalSales**
- **TotalTransactions**
- **TotalBaskets** (distinct TRANSACTIONID)
- **TotalSalesPerSqFt = TotalSales / SquareFootage**
- **TotalTransactionsPerSqFt = TotalTransactions / SquareFootage**
- **TotalBasketsPerSqFt = TotalBaskets / SquareFootage**
- **Avg_BasketValue = TotalSales / TotalBaskets**

Basket-level metrics are then built:

- For each STORENAME × TRANSACTIONID, compute BasketTotal and Items (sum of LINEQTY).
- For each store, compute **AvgBasketValue** and **AvgItemsPerBasket** from these baskets.

These metrics are merged into a final **store_summary** table.

Key findings:

- SalesPerSqFt varies widely, from roughly **$440** on the low end up to over **$3,100**.
- The median store is around **$1,525** per square foot, while the mean is around **$1,629**, reflecting skewness toward high performers.
- Stores like **Montrose**, **Hampden Lane**, **Potomac**, and **Kensington** appear at the very top in SalesPerSqFt.
- At the lower end are stores such as **Cloverly**, **Burtonsville**, and other locations with weaker traffic and/or smaller basket values.

These KPIs form the backbone for later quadrant and modeling work.

8. Department and Day-Type Mix by Store

Two sets of mix variables are constructed:

1. **Category mix** (Share_BEER, Share_LIQ, Share_WINE, Share_MISC)
    - For each store, department-level sales shares are computed.
    - These shares highlight whether a store is liquor-dominant, wine-heavy, beer-heavy, or has a relatively large MISC component.
2. **Day-type mix** (Share_Holiday, Share_Weekday, Share_Weekend)
    - For each store, sales are aggregated by DayType and normalized as shares.
    - A check confirms that each store's shares sum to 1, ensuring the mix variables are consistent.

These mix variables later enter the regression model as drivers of SalesPerSqFt and help differentiate stores that depend heavily on weekends and holidays from those that rely more on steady weekday trade.

9. Demographics and Sales Relationships

Demographic attributes from **Demographics.csv** are merged into the store KPIs:

- **TotalPopulation**
- **Older25Years** (population 25+)
- **PovertyLevel**

Scatterplots with linear fits are used to compare **TotalSales** with:

- **SquareFootage** – positive but weak correlation (~0.28). Larger stores tend to sell more, but there is wide variation.
- **TotalPopulation** – very weak correlation (~0.09). Trade-area population alone does not meaningfully explain store sales.
- **Older25Years** – similarly weak correlation (~0.12).
- **PovertyLevel** – slightly negative correlation (~–0.16) but not strongly predictive.

Conclusion: demographics and store size function as weak background controls rather than strong drivers of productivity in this dataset.

10. Quadrant Analyses: Sales, Traffic, and Baskets

A combined **store_final** dataset is built by merging:

- Store KPIs
- Demographics
- Category mix
- Day-type mix

Then, medians of **TotalSalesPerSqFt**, **TotalTransactionsPerSqFt**, and **TotalBasketsPerSqFt** are used to define quadrant systems:

1. **Quad_SalesTraffic**
    - High Spend & High Traffic
    - High Spend & Low Traffic
    - Low Spend & High Traffic
    - Low Spend & Low Traffic
2. **Quad_SalesBaskets**
    - High Spend & High Baskets
    - High Spend & Low Baskets
    - Low Spend & High Baskets
    - Low Spend & Low Baskets

3. **Quad_TrafficBaskets**
    - High Traffic & High Baskets
    - High Traffic & Low Baskets
    - Low Traffic & High Baskets
    - Low Traffic & Low Baskets

Three scatterplots visualize:

- SalesPerSqFt vs TransactionsPerSqFt
- SalesPerSqFt vs BasketsPerSqFt
- TransactionsPerSqFt vs BasketsPerSqFt

Each store is labeled and colored by quadrant, with dashed lines marking the median cutoffs.

These quadrants provide a **structural map** of the system:

- High Spend & High Traffic stores are the core engine: they combine strong demand with high productivity per square foot.
- Low Spend & Low Traffic stores are structurally weak: they lack both traffic and revenue intensity.
- Mixed cases (e.g. High Spend & Low Traffic, Low Spend & High Traffic) suggest more nuanced issues either underutilized space (low traffic but high spend) or poor conversion/value (high traffic but low spend).

11. Store-Level Performance Drivers – OLS Model

A modeling dataset (**model_data**) is created with one row per store and the following variables:

- **Outcome**:
    - SalesPerSqFt = TotalSalesPerSqFt
    - log_SalesPerSqFt = log (SalesPerSqFt)
- **Predictors**:
    - SquareFootage
    - TotalPopulation, Older25Years, PovertyLevel
    - Avg_BasketValue, AvgItemsPerBasket

- TotalTransactionsPerSqFt
- Share_BEER, Share_LIQ, Share_WINE
- Share_Weekday, Share_Weekend

OLS fits on the log scale:

log (SalesPerSqFt) ~ sim SquareFootage + TotalPopulation + Older25Years + PovertyLevel + Avg_BasketValue + AvgItemsPerBasket + TotalTransactionsPerSqFt + Share_BEER + Share_LIQ + Share_WINE + Share_Weekday + Share_Weekend

Key statistics (27 stores, 12 predictors):

- **Adjusted R^2 ≈ 0.955**
- **Residual SE ≈ 0.109** (log scale)
- **F-statistic ≈ 46.6**, p ≈ 3.7e–09

Main findings:

- **Traffic intensity dominates.**
    - TotalTransactionsPerSqFt is highly significant and has a strong positive coefficient. Stores that process more transactions per square foot are consistently more productive in SalesPerSqFt.
- **Category mix matters.**
    - Share_BEER, Share_LIQ, and Share_WINE all have positive, significant coefficients relative to the omitted **MISC** category.
    - Interpretation: stores that concentrate revenue in core alcohol (beer, liquor, wine) rather than miscellaneous items tend to be more productive per square foot. Because these are shares that sum to 1, coefficients should be read as **contrasts vs MISC**, not independent levers.
- **Basket value has a smaller, borderline effect.**
    - Avg_BasketValue is positive with a p-value around 0.06. It helps, but not nearly as strongly as traffic and core-mix.
- **Demographics and size are weak controls.**

- SquareFootage, TotalPopulation, Older25Years, and PovertyLevel are not statistically strong. They are retained as controls but are not primary drivers of store productivity.

Given the small sample (27 stores) relative to the number of predictors, this OLS model is **descriptive**, not a production-level forecasting tool. Its value lies in clarifying relative drivers and providing a benchmark for over/under-performance.

12. Penalized Models: Ridge, LASSO, and Elastic Net

To test whether regularization improves predictive performance or simplified interpretation, three penalized models on log_SalesPerSqFt are fitted:

- **Ridge (α = 0)** – shrinks all coefficients but keeps them non-zero.
- **LASSO (α = 1)** – shrinks some coefficients exactly to zero (variable selection).
- **Elastic Net (α = 0.5)** – a compromise between Ridge and LASSO.

All are tuned with cv.glmnet.

12.1 Elastic Net (Full Sample)

Using leave-one-out style cross-validation:

- Selected λ (1se) ≈ 0.15.
- Resulting model is extremely sparse:
    - Only **TotalTransactionsPerSqFt** retains a non-zero coefficient; all other predictors are shrunk to zero.

In-sample performance:

- **OLS**: RMSE ≈ 144, MAE ≈ 99.5
- **Elastic Net**: RMSE ≈ 245, MAE ≈ 173

Interpretation: when heavily penalized, the model identifies **traffic intensity** as the single robust predictor but fit deteriorates significantly. This confirms traffic's importance but shows that mix and basket variables still contribute meaningful explanatory power.

12.2 Ridge and LASSO (Train/Test Split)

With a simple random split (18 training, 9 test stores):

- **Ridge** (λ tuned): test RMSE ≈ 483, MAE ≈ 388
- **LASSO** (λ tuned): test RMSE ≈ 363, MAE ≈ 249

For comparison:

- **OLS (full)**: RMSE ≈ 144, MAE ≈ 99.5
- **OLS (train/test)**: test RMSE ≈ 153, MAE ≈ 128

None of the penalized models outperform OLS on this dataset. With such a small N, the variance reduction from regularization does not offset the information loss.

12.3 Takeaway

- Penalized models reinforce that **TotalTransactionsPerSqFt** is the only predictor that always survives heavy shrinkage.
- OLS remains the best-performing model in terms of in-sample and simple out-of-sample metrics.
- The overarching message: current modeling is **descriptive and diagnostic**, not predictive. More stores and richer features would be needed before penalized methods yield real gains.

13. Over- and Under-Performance vs OLS

Using the OLS model, store-level residuals are computed:

Residual_OLS = SalesPerSqFt - PredictedSalesPerSqFt

Interpretation:

- **Positive residuals** → store is outperforming its modeled potential.
- **Negative residuals** → store is underperforming relative to what the model expects, given its fundamentals.

From the residual ranking:

- Top **over-performers** (largest positive residuals) include **Cabin John**, **Downtown Rockville**, **Goshen Crossing**, **Darnestown**, and **Silver Spring**. These stores convert their traffic, mix, and space into SalesPerSqFt exceptionally well.
- Top **under-performers** (most negative residuals) include **Montrose**, **Kingsview**, **Seneca Meadows**, **Burtonsville**, and **Kensington**. Some of these, like **Montrose**, are still top performers in absolute terms, but the model indicates that they could be even more productive given their underlying strengths.

The residual ranking is not a “good vs bad” list; it is a **potential lens**. It shows where a store stands relative to what a multivariate benchmark suggests it should be delivering.

14. Market-Basket Analysis (MBA)

MBA is applied to **Y24–25 non-refund transactions** as follows:

- Identify the **top 200 items** by transaction frequency.
- Convert each transaction into a basket of unique item descriptions (one row per TRANSACTIONID).
- Represent baskets as an arules transactions object.

An **Apriori** model is then run with:

- Support ≥ 0.0002
- Confidence ≥ 0.20
- Rule length 2–3 items
- Target = "rules"

Rules are filtered to **lift > 1.05** and sorted by lift.

Findings:

- The strongest rules form a **tight cluster** around **New Amsterdam 50ml flavored vodkas** (Apple, Peach, Passion Fruit, Grapefruit).
- Customers who buy either of these flavors have a very high probability of buying a third, with very high lifts.
- Another notable rule links **L MARCA PRO ROSE 750ML** and **L MARCA PROSECCO 750ML**, suggesting frequent co-purchase.

Interpretation:

- There are real, strong associations, but they are **micro-patterns** within specific flavor families and brands, especially among 50ml minis.
- At the current support/confidence thresholds, MBA does not produce a broad, system-wide map of complementary items.

For MBA to become operationally powerful, ABS would likely need to:

- Run **category-specific** MBAs (e.g., within LIQ or WINE only).
- Possibly lower support thresholds within specific categories or stores.

- Focus on rules that involve **high-margin** or strategic items, not just high-volume flavored minis.

Right now, MBA is **interesting but narrowly** useful for confirming some intuitive relationships, but not yet a comprehensive cross-merchandising tool.

15. Bootstrapping

Two bootstrapping exercises quantify uncertainty around key metrics.

15.1 Median Basket Value

From all non-refund transactions:

- Per-transaction basket totals are computed from NETAMOUNT grouped by TRANSACTIONID.
- There are over **6 million baskets** in total, so a random subsample of **50,000** baskets is drawn for computational stability.
- A non-parametric bootstrap with **1,000 resamples** is used to estimate the distribution of the median basket value.

Results:

- **95% percentile CI** for median basket value ≈ **$23.99–$24.97**.

Interpretation:

- The **typical ABS basket is about $24**, which is significantly lower than the store-level **average** basket values (often $30–$45).
- This confirms that a relatively small number of large baskets pull up the mean; the median customer is making a modest purchase.

15.2 Quadrant Mean SalesPerSqFt (High–High vs Low–Low)

Using the SalesPerSqFt vs TransactionsPerSqFt quadrants:

- One group: **High Spend & High Traffic** stores.
- Another group: **Low Spend & Low Traffic** stores.

For each group, a bootstrap with **2,000 resamples** of the mean SalesPerSqFt is performed.

Results (approximate):

- **High Spend & High Traffic**:

    - Mean SalesPerSqFt ≈ **$2,223**
    - 95% CI ≈ **[$1,947, $2,512]**
- **Low Spend & Low Traffic**:
    - Mean SalesPerSqFt ≈ **$1,015**
    - 95% CI ≈ **[$819, $1,177]**

The confidence intervals are well separated; **high–high stores deliver roughly double the SalesPerSqFt** of low–low stores, and this gap is not attributable to random sampling variation alone.

This confirms that quadrant membership has **real, economically meaningful consequences**: moving a store from Low Spend & Low Traffic into High Spend & High Traffic is a step-change in productivity, not a marginal improvement.

16. Store Typology and Action Flags

A **store_typology** table combines:

1. Quadrant positions:
    - Quad_SalesTraffic (High/Low Spend × High/Low Traffic)
    - Quad_SalesBaskets (High/Low Spend × High/Low Baskets)
2. OLS residuals:
    - Residual_OLS = SalesPerSqFt – Pred_Sales_OLS
3. Performance and opportunity flags:
    - **PerfFlag**:
        - ≥ +75: **Strong over-performer**
        - +25 to +74: **Moderate over-performer**
        - –25 to –74: **Moderate under-performer**
        - ≤ –75: **Strong under-performer**
        - Otherwise: **Near model expectation**
    - **Traffic_Opportunity** (based on Quad_SalesTraffic):
        - High Spend & Low Traffic → “Increase traffic (awareness/location/selection)”

    - Low Spend & Low Traffic → “Fundamental issue: traffic + value”
    - Low Spend & High Traffic → “Improve conversion / basket value”
    - otherwise → “Traffic OK”
  - **Basket_Opportunity** (based on Quad_SalesBaskets):
    - High Spend & Low Baskets → “Increase visit frequency / basket count”
    - Low Spend & High Baskets → “Increase basket value (mix, pricing, upsell)”
    - Low Spend & Low Baskets → “Structural weakness (frequency + value)”
    - otherwise → “Basket density OK”

This framework clarifies:

- **Quadrants** describe a store’s structural position in terms of spend vs traffic vs baskets.
- **Residuals** describe whether the store is doing better or worse than expected, given its fundamentals.
- **Flags** translate this into concrete guidance on whether to prioritize **traffic**, **basket value**, or both.

17. Priority Store List

A **priority_stores** table ranks stores by Residual_OLS (worst first) and assigns an **ActionPriority**:

- **High**: strong or moderate under-performer
- **Medium**: near expectation
- **Low**: over-performer

Among the **High-priority** stores, the top seven include:

- **Montrose**
- **Kingsview**
- **Seneca Meadows**
- **Burtonsville**
- **Kensington**
- **Cloverly**

- **Olney**

These are not simply "bad" stores; some, like Montrose, have **excellent** raw SalesPerSqFt and ranking. The point is that, relative to their strong traffic, mix, and space, they appear to be under-delivering and thus offer the largest upside if issues are addressed.

The typology plus priority ranking is the **bridge from analytics to action**: it tells ABS which stores to focus on first and which levers (traffic, basket value, both) to pull at each location.

18. Limitations and Next Steps

Key Limitations

- **Small N** (27 stores).

  With 12 predictors, even a well-behaved OLS can overfit noise. The high R^2 should be interpreted cautiously.

- **Omitted variables.**

  Important factors are not observed: rent, competition, store hours, staffing levels and quality, parking, neighborhood commercial mix, detailed assortment depth, etc. These omissions limit explanatory power and could bias some effects.

- **Collinearity in mix variables.**

  Category shares must sum to 1. Coefficients on Share_BEER, Share_LIQ, and Share_WINE are best interpreted as contrasts against the omitted **MISC** share.

- **MBA scope.**

  At current thresholds, MBA mostly reveals tight brand-specific flavor clusters and does not deliver a broad cross-merchandising map.

Concrete Next Steps

1. **Richer modeling**
    - Add more years of data and incorporate store-level features such as competition density, rent, staffing, and hours.
    - Consider hierarchical or Bayesian models to stabilize estimates with small N.
2. **Category-specific analyses**
    - View quadrants and residuals at the department level (e.g., liquor-only SalesPerSqFt) to see where stores are especially strong or weak.

3. **Stronger MBA**
    - Run MBAs within categories or store clusters.
    - Allow lower support thresholds in targeted segments.
    - Prioritize rules involving high-margin or strategic items.
4. **Operational action plan**
    - Turn the typology and priority list into a formal action plan with owners and timelines.
    - Monitor SalesPerSqFt, traffic, and basket metrics before and after interventions at high-priority stores.

19. References and Acknowledgements

References

- ABS internal datasets: *TransactionData.csv*, *Designation.csv*, *Demographics.csv* (FY23–24 and FY24–25).
- R Core Team (2025). *R: A Language and Environment for Statistical Computing.*
- Packages used: **dplyr**, **tidyr**, **stringr**, **lubridate**, **ggplot2**, **readr**, **purrr**, **broom**, **glmnet**, **boot**, **scales**, **ggrepel**, **arules**, **caret**.
- OpenAI (2025). *ChatGPT (version 5.1)* – used as an assistant for troubleshooting code, clarifying errors, and improving the analytical workflow.

Acknowledgements

I thank **Montgomery County Alcohol Beverage Services (ABS)** for providing the data and framing the analytical questions, and **[Instructor's Name]** for guidance throughout DATA 205. The course emphasis on reproducibility and clear communication directly shaped both the analysis and its presentation.