Final Report

Team 13: Fiona Romanoschi, Alec Barrett, Macy Cripe, Ethan Hulke, Brianna Campbell, Jarukit Ketjoy, and Brendan Tynan

GitHub repository with code and data: https://github.com/alec10barrett/SDS322E_project

Introduction

Our project dives into the fascinating world of Mixed Beverage Gross Receipts, exploring how sales vary across different ZIP codes. This exploration is especially intriguing as it reveals how wealth and community diversity impact consumer choices in these areas. Our project aims to help restaurants and bars fine-tune their marketing, pricing, and operations. Understanding the trends in mixed beverage sales can be a game changer in boosting profits, efficiently using resources, and enhancing customer satisfaction.

This study is more than just numbers and graphs. We're passionate about giving businesses the tools they need to predict future sales more accurately. By sculpting a precise predictive model for mixed beverage sales, we can help establishments predict future demands more accurately. This foresight is essential to ensuring that resources are judiciously allocated, thus safeguarding businesses from the pitfalls of profit erosion and operational inefficiencies.

Data

Our initial dataset was the Mixed Beverage Gross Receipts compiled by the Texas Comptroller of Public Accounts, detailing the list of taxpayers, their names, amounts reported and other public information required under Tax Code Chapter 183, Subchapter B. The features of this original dataset are listed below, but we mainly chose to focus on Responsibility Duration, Location Zip, Liquor Receipts, Wine Receipts, Beer Receipts and Total Receipts. To clean the data, we converted the explanatory variables to be numeric, and tried a couple of normalization techniques but scrapped it because it did not impact the modeling of the data as much.

At the beginning of this project, we intended to use the geographical components of this dataset, mainly the location address to gleam insights about alcohol receipts. Unfortunately, due to API paywalls, we were unable to convert the addresses to geographical coordinates, meaning we required additional data to act as explanatory variables for this project. Fiona aimed to resolve that issue by hand-compiling another dataset* that compiled census information for each ZIP code, selecting a few demographic variables of interest listed as the Demographic Data Set Features below. Then, she joined the datasets based on the ZIP code and finalized the merged dataset by typecasting several columns for analysis and modeling.

Dataset was formally compiled from Austin's demographic data library, and more specifically all of the separate spreadsheets from the American Community Survey 2017 Profiles for ZIP Codes.

Mixed Beverage Gross Receipts Features: Taxpayer Number, Taxpayer Name, Taxpayer Address, Taxpayer City, Taxpayer State, Taxpayer Zip, Taxpayer County, Location Number, Location Name, Location Address, Location City, Location State, Location Zip, Location County, Inside/Outside City Limits, TABC Permit Number, Responsibility Begin Date, Responsibility End Date, Obligation End Date, Liquor Receipts, Wine Receipts, Beer Receipts, Cover Charge Receipts, and Total Receipts

Demographics Data Set Features:

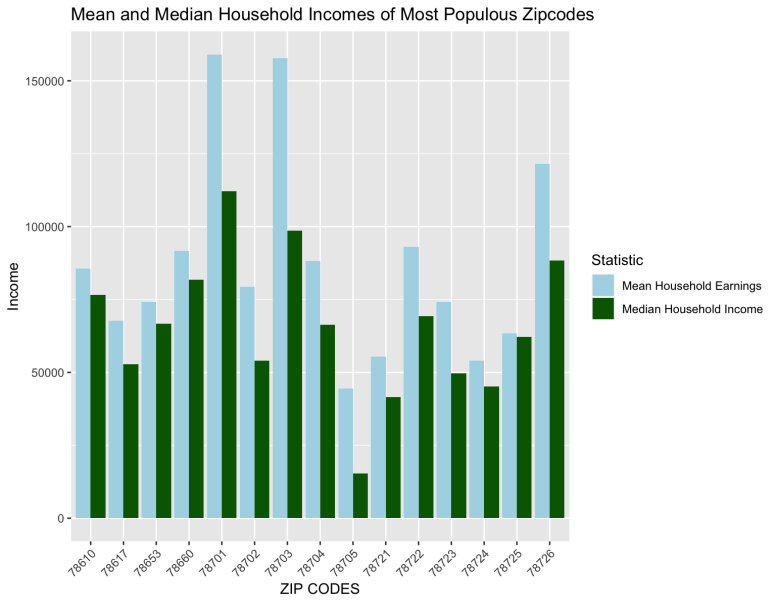
ZIP CODES, Median Age in Years, Total Population, 21 years and over population, Male Count, Hispanics or Latino of Any Race, White Alone, Black Alone, Asian Alone, Median Household Income, Mean Household Earnings, and Total Housing Units.

Exploratory Analysis

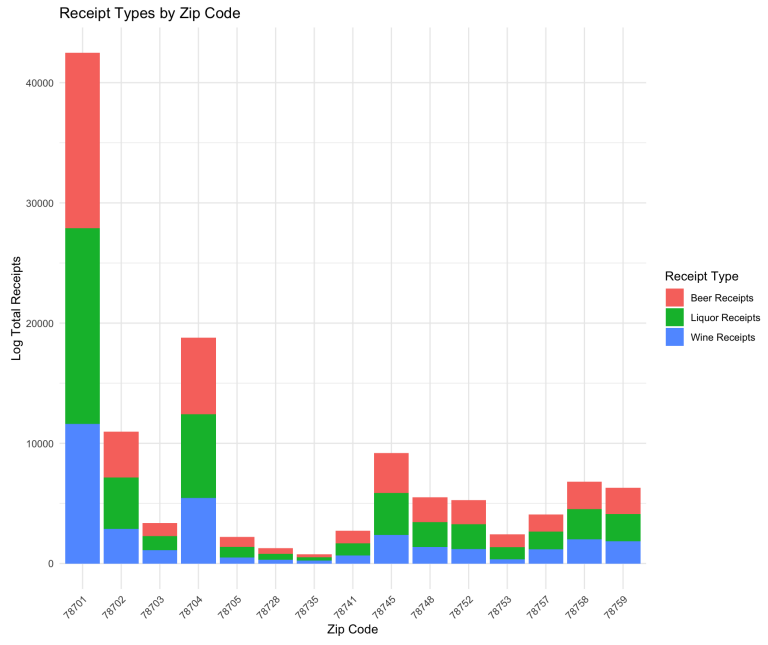
Our beverage_receipts dataset at a glance:

# of Locations in Austin that submitted Alcohol Tax Receipts	**195,682**
Mean # of **Liquor** Receipts	**31,828**
Mean # of **Wine** Receipts	**8,836**
Mean # of **Beer** Receipts	**15,455**
Mean # of **Cover Charge** Receipts	**119**
Mean # of **Total Receipts** (All Alcohol and Cover Charge)	**55,841**

We first made some preliminary visualizations to get familiar with our dataset and look for interesting trends.

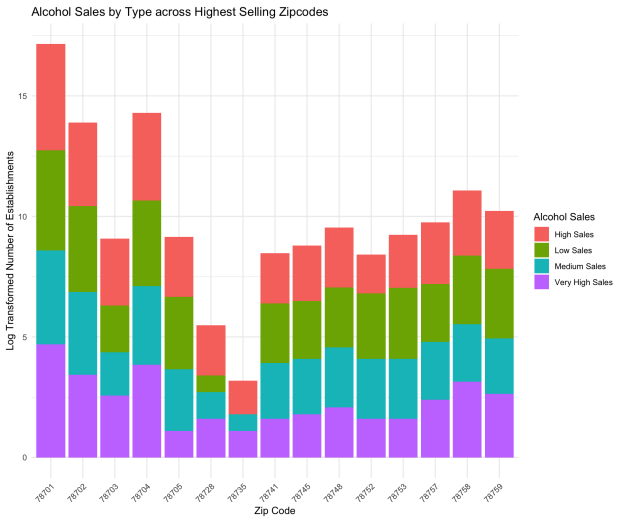


The first thing we wanted to look into was income levels across the 15 most populous zip codes in Austin. We visualized the average mean and median incomes for the top 15 zip codes, and found the mean household income was predictably skewed by outliers with very high incomes. Neither mean nor median household income ended up being significant for predicting total sales, but median household income would have been the most accurate measure.



The next preliminary visualization we looked at was the alcohol sale type across the 15 highest selling zip codes. Even after logarithmically transforming the total receipts, the data is still very skewed – this can be explained by the zip code 78701 encompassing 6th street and Rainey street,

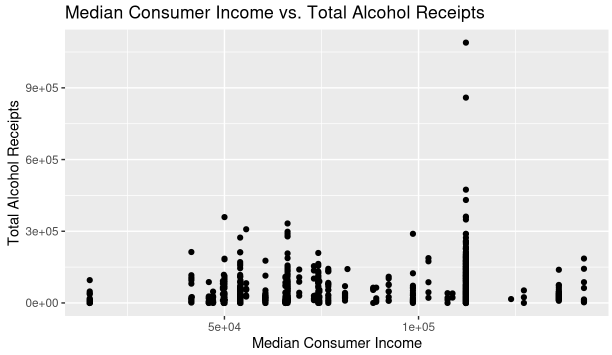
which have the highest sales by far. This visualization gave us the idea to investigate if beer is the highest selling type of alcohol in Austin for one of our hypotheses.



Our last preliminary visualization utilized our categorical variable of alcohol sale level. We have 4 categories – Very High Sales, High Sales, Medium Sales, and Low Sales. These categories were created from the 4 quartiles of our dataset, with Very High Sales equating to the top 25% and low sales to the bottom 25%. We ended up using this variable to determine the accuracy of our machine learning models.

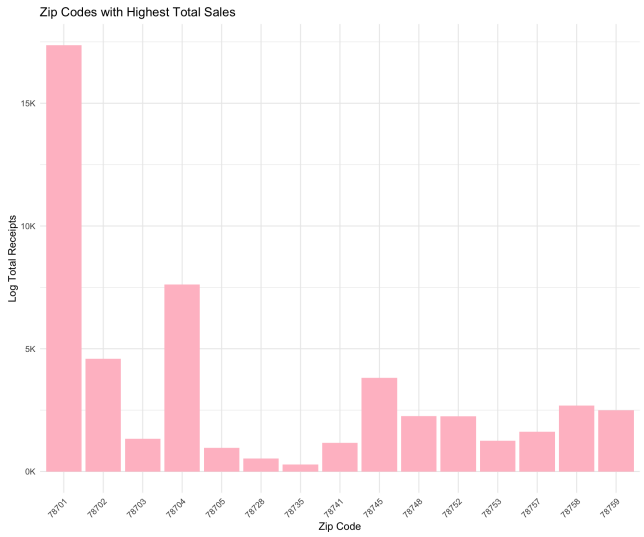
After exploring our dataset, we came up with the following hypotheses.

Our first hypothesis was that higher consumer income correlated to higher ticket alcohol receipts. This stemmed from our preliminary visualizations that seemed to show some type of relationship between income and receipts. In order to visualize this a correlation map was created between the two variables.



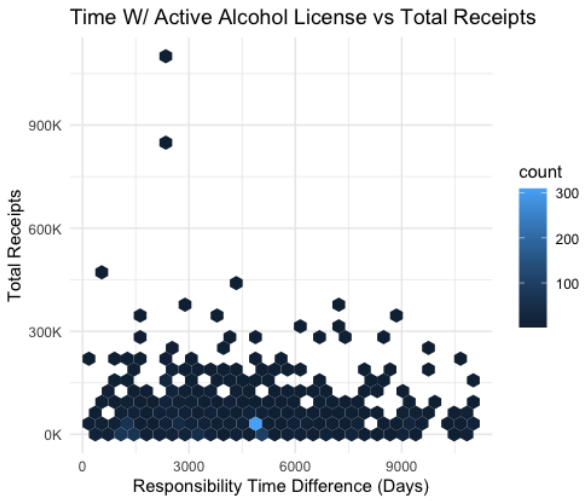
From this visualization there is no pattern or correlation between the two variables. When trying to transform the data using log, the graph looked almost exactly alike. Another method used was non linear regression, but this also led to the same conclusion as the original data unmodified. The correlation coefficient calculated using the original data set came to be 0.186. This is extremely low indicating a very weak relationship between consumer income and ticket alcohol receipts. Due to these findings the hypothesis was not supported. One factor to consider that mistakenly leads to this belief that there is any possibility of a relationship, is population density and location. The areas in the initial visualizations that showed high consumer income and high alcohol receipts was in downtown Austin where the population density is much larger than other areas and much more expensive to live in.

Our second hypothesis was that downtown zip codes of large cities will have the highest number of total alcohol sales. We examined this by visualizing total alcohol receipts across the 15 zip codes with the highest sales, and looked for which zip codes were in downtown areas.



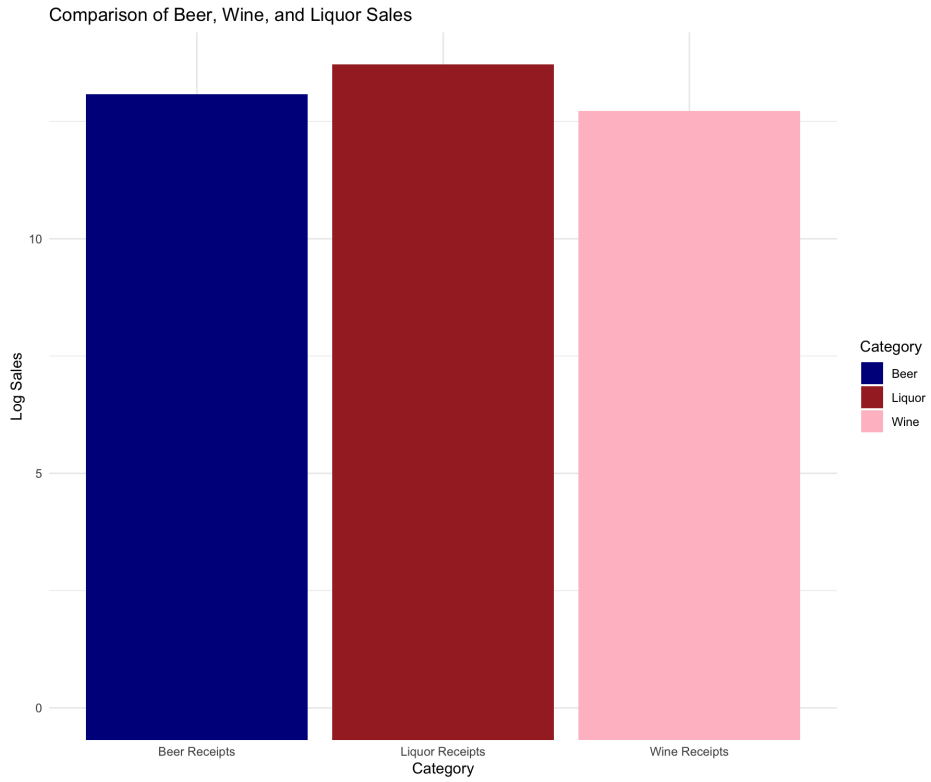
Even after logarithmically transforming the data, the graph is still skewed. This aligns with what we saw in our second preliminary visualization, and confirms our hypothesis. The top selling zip codes are 78701, which is the true ‘downtown’ Austin zip code and contains 6th street and Rainey, which both are huge tourist attractions with tons of bars and restaurants selling alcohol. The next highest selling zip code is 78704, which contains South Lamar, Barton Springs & Zilker, and South Congress. This zipcode is also considered downtown Austin, and so it confirms our hypothesis. Downtown areas get the most traffic and have the most vibrant nightlife scenes, so it makes sense they have the highest alcohol sales.

Our third hypothesis was that a longer active alcohol license would result in more sales, measured by higher levels of total receipts.



To examine this visually we used a hexbin plot, which is useful for visualizing large datasets with many data points with the same values. The lighter the hexagon, the more data points are contained within, shown by the legend. This hypothesis was tested in depth using an AMIRA model explained below, but both the model and the visualization show the same conclusion – there is no true correlation between total sales and the length of an active alcohol license. We made this hypothesis because we believed the longer a place was open the more customers would come as they had time to develop a reputation and regular clientele, but the data does not support this. Our hypothesis was wrong in this case, and total sales cannot be accurately predicted by how long an establishment has had an active liquor license.

Our fourth and final hypothesis was that beer would be the highest selling alcohol across Austin.



To examine this, we looked at the aggregate data of total sales across Austin (after a logarithmic transformation) by receipt type. Before transforming the data, it appeared that liquor was by far the highest selling alcohol, but after the transformation we see total sales are much more evenly distributed across the 3 types of alcohol. While liquor is still the highest selling alcohol, which disproves our hypothesis, beer is the second highest selling alcohol, selling more than wine. We believed beer would be the top selling alcohol because of the large number of breweries in Austin, and the prevalence of custom brewed IPAs at popular bars and restaurants across town. We see that beer is more popular than wine, but not liquor. This is probably due to the presence of 6th street and Rainey street, which have many bars with custom cocktails containing liquor.

Modeling

Fiona's Approach

Once merging the two datasets based on their zip codes, I filtered for the Location City to be in Austin, returning me 195,682 different alcohol receipt venues. The basis of our task was investigating how a variety of features (*Responsibility Time Difference (Days), Median Age in Years, Total Population, 21+ population, Male Count, Hispanics or Latino of Any Race, White Alone, Black Alone, Asian Alone, Median Household Income, Mean Household Earnings, Total Housing Units*) related to predicting our target variable of *Total Alcohol Receipts*.

Initially, I attempted regression as our modeling task, but after trying several normalization techniques (MinMax, Log, etc. scaling) to minimize the mean-squared error, I scrapped the idea and decided to do classification instead. So, I found the quartile ranges for the *Total Alcohol Receipts* for the entire dataset, and labeled each data point on the quartile it fell into (*Alcohol Sales* column : [Low Sales, Medium Sales, High Sales and Very High Sales]). My rationale aimed to classify the general level of alcohol sales a certain location had in relation to the features of influence rather than precisely predicting the tax receipts on their own. In hindsight, classification also suited the breadth of this project, as zip codes were many times too large to distinguish specific differences between locations. Meaning, to properly conduct a regressive task, we would need to narrow our geographical scope, which was not available in our collected data. A breakdown of the classifications within each sector was shown in the EDA section as the last preliminary visualization, showing a roughly balanced classification, with each classification range representing a fourth within each zip code. This is good as it lessens a geographical imbalance despite 78701 being the zip code with the most alcohol vending locations.

Moving forward in our classification task, I re-cleaned/formatted the dataset (dropping NAs and typecasting all columns to be numeric) and used a commonly used train-test-split ratio, with 80% of our dataset comprising the training set and 20% for the validation / testing set.

Based on what was introduced in class, I chose to evaluate several classifications against each other, using a Decision Tree, Random Forest, SVM (Support Vector Machine), K-Nearest Neighbors and Logistic Classifiers. Though all of these classifiers were utilized using open-source libraries, here's a brief summary of how each algorithm functions below.

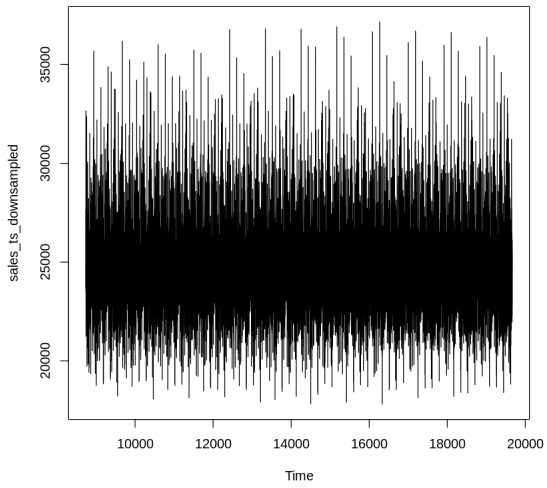
1. Decision Tree
    a. A Decision Tree classifies data by recursively splitting it into subsets based on the most significant feature at each node. The process continues until the data is partitioned into homogenous groups, creating a tree-like structure where each leaf node represents a distinct class.
2. Random Forest

    a. Random Forest is an ensemble method that builds multiple Decision Trees and combines their predictions. Each tree is constructed using a random subset of the data and features, and the final prediction is determined by aggregating the individual predictions, typically through a majority voting mechanism.
3. SVM
    a. SVM classifies data by "finding the hyperplane that maximally separates different classes in a high-dimensional space" (Mathworks). It aims to create a decision boundary with a maximum margin between classes.
4. K-Nearest Neighbors
    a. KNN classifies data based on the majority class of its k nearest neighbors in the feature space. It relies on the assumption that similar data points share the same class, using a distance-related similarity measure to distinguish classes.
5. Logistic Classification
    a. It models the probability that an instance belongs to a particular class using the logistic function. The algorithm estimates coefficients for input features, and the sigmoid-shaped curve of the logistic function maps the linear combination of these features to a probability between 0 and 1, determining the class label.

Ethan's Approach

In addition to the more classical approach taken above, we attempted a more extra-curricular approach using a few packages outside the scope of the class. The focus of this portion was to build a model that could predict the number of receipts given the duration of operation. To do this we used a library called "forecast" and wrote a time-series forecasting code that employs the ARIMA (AutoRegressive Integrated Moving Average) model. The ARIMA model is popular for time-series analysis, capturing temporal dependencies and trends. The selected order (0,0,0) indicates no autoregressive or moving average components, making it essentially a mean model.

The dataset, Mixed_Beverage_Gross_Receipts, is loaded, containing information about various establishments, their locations, and associated total receipts. Rows with missing values are removed, and date columns are converted to the appropriate Date format. To manage the large dataset, a 5% random sample (sample_frac(0.05)) is taken for analysis. From there a time series is constructed and scaled to match the frequency of the dataset. The said time series is shown below. a reminder that this is showing roughly 60k points (5% of our 3.1 million rows).



Then the data is split, with 80% being used for training and 20% being for testing. The model is chosen automatically with auto.arima() from the 'forecast' library mentioned above. Lastly, a forecast is generated based on the chosen model using forecast(). Additionally, the accuracy of this forecast (and model) can be assessed using accuracy(), the outputs of which as discussed in the discussion section below.

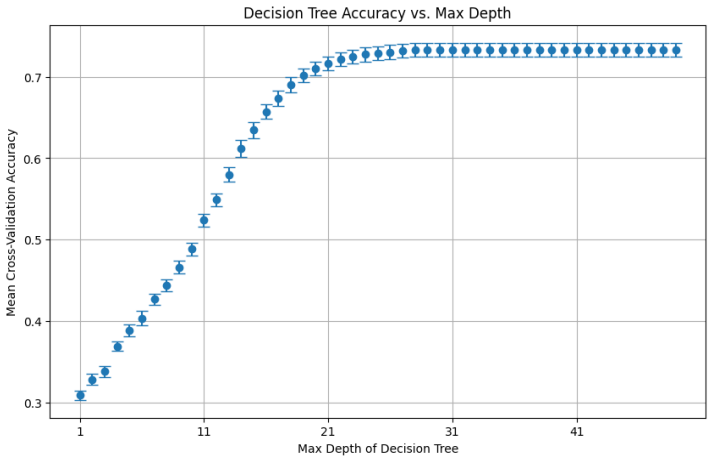
Discussion, Evaluation Metrics and Results:

Fiona's Approach

For the aforementioned classifiers, I chose to evaluate them based on their accuracy (the number of correct predictions divided by the total number of predictions across all classes) as it was the most standard comparison, and our topic interest did not require minimizing for Type I or Type II errors such as in medical contexts. The accuracy of each model is displayed in the table below.

Model	Accuracy
Decision Tree	0.7347
Random Forest	0.7340
SVM	0.35
K-Nearest	0.69
Logistic	0.33

The Decision Tree and Random Forest Classifiers were almost identical in their accuracy. Although, for this classification task, I chose to go forward with the Decision Tree classifier as it is the easiest model to interpret and optimize. Similar to the lab exercise, I found the maximum depth of the decision tree that would lead to the highest mean-cross-validation accuracy, plateauing after a value of 31 levels, leading to a final accuracy of 0.73474.

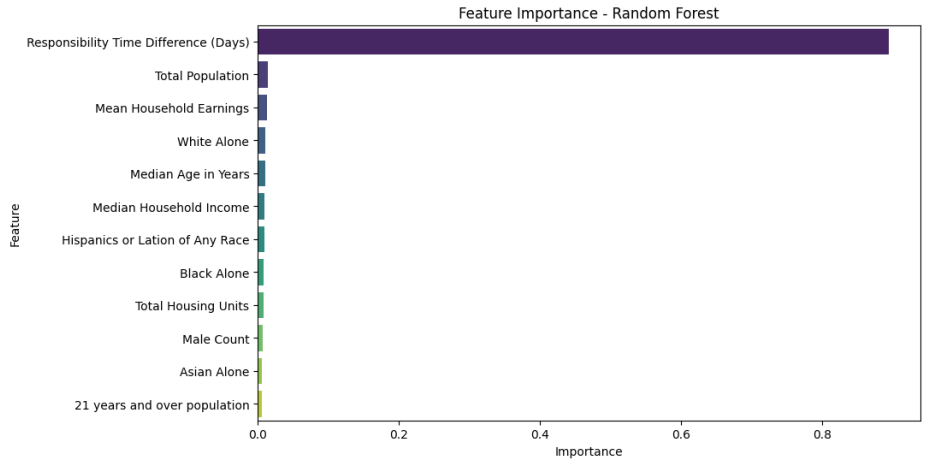


Looking at the features themselves, it is overwhelmingly clear that the Responsibility Time Difference (how long a placer has an active alcohol license) plays the largest role in classification for both the Decision Tree and Random Forest Models. Though correlation is not causation, this slightly supports our initial conjecture that the longer a place was open, the more alcohol it would sell.

Decision Tree Feature Importance:

1. Responsibility Time Difference (Days): 0.7708210262162251
2. Median Age in Years: 0.023920088862575472
3. Total Population: 0.0399389616249138
4. 21 years and over population: 0.005884332529497651
5. Male Count: 0.010830214356953397
6. Hispanics or Latino of Any Race: 0.016764541655760815
7. White Alone: 0.021969880483929807
8. Black Alone: 0.033247819893109605
9. Asian Alone: 0.016196664533149996
10. Median Household Income: 0.03595960250631818
11. Mean Household Earnings: 0.013759883595239343
12. Total Housing Units: 0.010706983742326958

Random Forest Feature Importance



Reflecting on this classification task, I encountered several issues in the compilation of the data and eventual modeling. Firstly, Google Colab is incredibly finicky with large dataframes, especially using the general UT student drive account that limits one's storage to 5 gigabytes. This was particularly difficult as I was re-running portions of the code and obtaining different results though I controlled for the random seed setting. I figured out that whenever a dataset is loaded locally, it can still execute a call on partially loaded data. Meaning if I ran pd.read_csv(), all executions would be run without notifying the data was incomplete. This took several days for me to figure out and finalize our results with the entire Austin dataset. Additionally, it explains why there is a difference in model accuracy between our previous presentation and this report, as I was able to fully load the absolute entirety of the data. Funnily enough, this once again supports the commonly held belief that with more data, one can train a more accurate model. Secondly, SVM takes a particularly long while to run on large datasets, making it the most inefficient classifier for its accuracy. Lastly, I think the particular challenge within this entire project was rounding out the initial barren dataset as I could not convert the addresses to geographical coordinates. This required me to compile a demographics dataset by hand and creatively merge them based on the information that was logical and available. I am certain there are better explanatory variables that exist for predicting alcohol sales, requiring much more geographical contextualization than a demographic makeup of a zip code. Moving forward, I would suggest more specific geographical features for a regression task.

Ethan's Approach

As mentioned in the Modeling section, we used a package called "forecast" to build and test our ARIMA model. That package can also evaluate the performance of our model. It's worth mentioning that the primary rationale behind all of the applications and use of ARIMA and its code was after a mix of online research and dialogue with ChatGPT about effective time-series model-building packages.

The accuracy of our ARIMA model is shown in the table below. This table is generated by the accuracy() function on the forecast and test_data.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	1.3582e-11	2797.52	2169.744	-1.12763	8.38459	0.656560	0.05068	NA
Test set	-1.0875e+03	2677.70	2149.898	-5.39679	9.01991	0.650555	-0.05119	0.76645

The acronyms are as follows: Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Autocorrelation of Forecast Errors (ACF1)

The ARIMA(0,0,0) model captures the mean behavior of the total receipts over time. The low ACF1 values and relatively small errors on the training set suggest a reasonable fit to the training data. However, the model's performance on the test set indicates limitations, as evidenced by increased errors, especially in terms of MAPE. The potential limitations were that the simplicity of the ARIMA(0,0,0) model might not capture more intricate patterns in the data. Exploring different ARIMA configurations and parameters could enhance model performance.

In conclusion, while the ARIMA(0,0,0) model provides a baseline, further exploration of model complexity and parameter tuning may yield improved forecasting results.

Ethics

Anticipate: People Affected

The implementation of a machine learning model presents businesses with a valuable opportunity to optimize pricing and marketing strategies, potentially enhancing their overall efficiency. However, there exists a concern regarding the potential negative impacts on the general public. While the insights derived from the model enable businesses to adjust pricing based on predicted high-demand periods or locations, a careful balance must be struck to avoid disadvantaging consumers. It is imperative to consider the broader societal implications of changes in business strategies and ensure that consumer interests are adequately safeguarded.

Reflect: Diversity, Equity, and Inclusion

A critical consideration in the application of machine learning models is the absence of demographic data in the training dataset. The oversight in incorporating characteristics of consumers raises ethical concerns about fairness and the perpetuation of existing inequalities. Without accounting for diverse demographics, the model may inadvertently perpetuate biases, leading to disparate impacts on certain demographic groups. This ethical dilemma underscores the importance of reflecting on the potential consequences of model outputs, not only for business outcomes but also for the broader principles of diversity, equity, and inclusion.

Engage: Under-represented

The nature of the variables included in the data introduces a potential bias towards locations with a higher concentration of establishments selling mixed beverages. This bias may raise questions of representation, particularly concerning regions with fewer such establishments. The risk of underrepresentation in these areas could result in certain demographics or businesses being either overrepresented or underrepresented in the model's analysis, thereby influencing accuracy and fairness. To address this issue, it is crucial to engage with the potential disparities in representation and strive for a more balanced and equitable model.

Act: Continuous Improvement

Given the dynamic nature of the mixed beverage industry, continuous improvement of the machine learning model is paramount. Adaptation to changes in market regulations or state laws is essential to maintain the model's relevance, accuracy, and ethical use over time. Neglecting to address external factors, such as regulatory changes, introduces exogenous variability in the data and may lead to unintended biases that disproportionately affect those impacted by regulatory shifts. Therefore, a proactive approach to monitoring and addressing changes in relevant variables is necessary to ensure the ongoing effectiveness and ethical application of the model.

Conclusion

Our investigation into the correlation between demographics and alcohol sales has yielded nuanced findings. Despite initial expectations, the relationship between demographic indicators, such as income and mean age of a ZIP code, and alcohol receipts proved more complex than anticipated. While our models demonstrated decent accuracy with forecasting and classification, particularly with Random Forest at 0.734 and Decision Tree at 0.73474, the overall strength of A to B correlations in our dataset was limited. Notably, the null hypothesis that higher consumer income correlates to higher alcohol receipts was not disproven, and our

exploration into factors like the duration of alcohol license and sales by alcohol type did not reveal conclusive patterns.

Our assumptions and justifications, grounded in data from the American Community Survey and the accuracy of the alcohol dataset, highlight the challenges in drawing definitive conclusions in this complex domain. Despite these challenges, our investigation did confirm the hypothesis that downtown ZIP codes in larger cities, such as 78701 and 78704, tend to have the most significant number of alcohol receipts. In summary, while some hypotheses were validated, many others were not, emphasizing the intricate relationship between demographics, location, operational factors, and alcohol sales.

It is important to acknowledge the complexity of the relationship between demographics and alcohol sales. The nuanced nature of our findings, coupled with the limitation in the overall strength of correlations, highlights the intricacies of this domain. Despite inconclusive results, our study provides insights with practical implications. For instance, the observation for downtown ZIP codes in larger cities remains a noteworthy trend. This information could guide strategic decisions for businesses looking to establish or expand their presence in specific geographic areas. Moreover, while our exploration into factors such as the duration of alcohol licenses and sales by alcohol type did not reveal conclusive patterns, it lays the groundwork for future research.

Businesses and policymakers can use this as a starting point for more targeted investigations, recognizing the potential impact of these factors on their strategies. This validates the importance of data-driven research and offers strategic insights. Although some patterns were not definitive, the study contributes to the broader knowledge base, suggesting directions for future research to unravel further the intricate connections between demographics, consumer behavior, and market trends.

Bibliography

Mixed beverage sales tax

Accounts, Texas Comptroller of Public. “Who Is Responsible for This Tax?” *Mixed Beverage Sales Tax*, Glenn Hegar, comptroller.texas.gov/taxes/mixed-beverage/sales.php. Accessed 5 Dec. 2023.

Mixed beverage gross receipts data

"Mixed Beverage Gross Receipts." *State of Texas Open Data Portal*, Texas Open Data Portal, data.texas.gov/dataset/Mixed-Beverage-Gross-Receipts/naix-2893/data. Accessed 5 Dec. 2023.

List of Taxpayers for mixed beverages.

Texas Comptroller of Public Accounts. "Mixed Beverage Gross Receipts." *Https://Data.Texas.Gov/*, 23 Nov. 2023, data.texas.gov/dataset/Mixed-Beverage-Gross-Receipts/naix-2893.

SVM Mathworks

MathWorks. (2023). *Support Vector Machines for Binary Classification*. https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

Acknowledgements and Group Breakdown:

If points were split evenly amongst 7 group members : 100/7 = 14.28571429

Instead of taking the max contribution, we decided to gauge it relative to the overall project split, where if someone contributed more than the even split, they would receive a 100 and it would be relative to the split rather than the max.

Task	Data Preparation	Graphs (EDA)	ML	Presentation Prep	Presentation	Report Writing	Website		Total	Contribution	**FINAL GRADE**
Task %	15	15	15	10	20	15	10		100		
Fiona	8	0	7	1	1	2	0		19	133	**100**
Alec	2	2	1	1.42	3.95	3	0		13.37	93.59	**94**
Macy	2	6	0	1.6	2.95	2	0		14.55	101.85	**100**
Kit	1	0	1	1.12	2.45	1.5	5		12.07	84.49	**85**
Brendan	1	0	0	1.84	3.05	2.5	5		13.39	93.73	**94**
Brianna	0	6	0	1.6	3.35	2	0		12.95	90.65	**91**
Ethan	1	1	6	1.42	2.95	2	0		14.37	100.59	**100**
Total	15	15	15	10	19.7	15	10	Total	99.7		