DATA ANALYSIS
IOWA LIQUOR SALES (2012-2023)

Presented by:
Gabriel Ferreira

TABLE OF CONTENTS

1. ABSTRACT

This research explores the application of various data analytics and machine learning techniques to analyze Iowa liquor sales data, aiming to understand the liquor sales in the state of Iowa, and derive actionable insights for decision-making. We start by performing data curation and cleaning, improving the dataset quality for analysis. Key steps include encoding categorical variables and handling numerical features for both regression and classification tasks. A regression model is built to predict liquor sales in dollars using numerical and categorical features, followed by a classification model to categorize sales into liquor categories. Clustering techniques, including K-Means clustering with the elbow method, are employed to group the data based on similarities, helping to discover patterns and segmentation within the liquor sales market. Visualizations such as elbow plots, confusion matrices, and cluster visualizations are used to validate the models. The findings reveal significant insights into the relationship between liquor sales patterns and various factors such as volume, category, and sales trends. These insights could inform inventory management, marketing strategies, and regional pricing models for liquor stores.

2. INTRODUCTION

This study analyzes Iowa liquor sales data to derive actionable insights for optimizing store operations, inventory management, and marketing strategies. The dataset contains detailed information on liquor transactions across various stores in Iowa, including store locations, product categories, and sales figures.

The key variables in the dataset include:

- **Invoice/Item Number**: A unique identifier for each transaction.
- **Date**: The date of the transaction.
- **Store Details**: Including Store Number, Store Name, Address, City, Zip Code, and Store Location.
- **Geographical Data**: County Number and County.
- **Product Information**: Including Category, Category Name, Vendor Number, Vendor Name, Item Number, and Item Description.
- **Sales Data**: Including Pack, Bottle Volume (ml), State Bottle Cost, State Bottle Retail, Bottles Sold, Sale (Dollars), Volume Sold (Liters), and Volume Sold (Gallons).

The primary goal of this analysis is to leverage machine learning models to explore and derive insights from this dataset. The key questions we aim to answer include:

1. **Predicting Sales**: What factors influence the total sales in dollars, and how can we predict future sales for better inventory management?
2. **Categorizing Liquor**: How can we classify liquor sales into categories based on sales patterns and product features?
3. **Clustering Sales Patterns**: Can we identify distinct clusters in the sales data that represent different types of stores or customer behaviors?

To answer these questions, we employ various machine learning techniques, including regression analysis, classification, and clustering. Each method is applied strategically to different aspects of the dataset, with a focus on maximizing the interpretability and utility of the insights for business decision-makers.

3. DATA ANALYSIS

Before diving into these machine learning models, data exploration and cleaning were performed to ensure the dataset was suitable for analysis. For computational feasibility, the dataset was reduced to 30% of its original size, retaining enough data for meaningful analysis without sacrificing performance.

3.1. Data Exploration and Cleaning

The exploration was started by applying python functions such as head(), tail(), shape(), .columns, info(), to gain familiarity with the data and understand its structure and contents.

3.1.1. Data Sample

The dataset provided insights into various aspects of liquor sales, including store locations, transaction details, product descriptions, and sales metrics such as bottles sold and sale dollars. In this data sample, we have transactions in 99 Iowa counties.

Invoice/Item Number	Date	Store Number	Store Name	Address	City	Zip Code	Store Location	County Number	County	Category	Category Name	Vendor Number	Vendor Name	Item Number	Item Description	Pack	Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Bottles Sold	Sale (Dollars)	Volume Sold (Liters)	Volume Sold (Gallons)
RINV-05018600024	12/11/2023	4698	QUALITY QUICK STOP / FAIRFIELD	201 WEST BURLINGTIN AVE.	FAIRFIELD	52556	POINT (-91.965393979 41.006794)	NaN	JEFFERSON	1022200.0	100% AGAVE TEQUILA	619.0	CAMPARI AMERICA	87619	ESPOLON BLANCO	12	750	16.00	24.00	-6	-144.00	-4.50	-1.18
S27341700031	08/17/2015	2487	ANAMOSA FAMILY FOODS	402 EAST MAIN	ANAMOSA	52205	POINT (-91.281344 42.108275)	53.0	JONES	1062300.0	FLAVORED RUM	35.0	BACARDI U.S.A., INC.	43137	BACARDI LIMON	12	1000	9.50	14.25	3	42.75	3.00	0.79
S06461900038	07/09/2012	2622	HY-VEE FOOD STORE / IOWA CITY	1125 N DODGE ST	IOWA CITY	52240	POINT (-91.518868 41.676095)	52.0	JOHNSON	1012100.0	CANADIAN WHISKIES	65.0	JIM BEAM BRANDS	15248	WINDSOR CANADIAN PET	6	1750	8.92	13.38	6	80.28	10.50	2.77
INV-07548000052	10/02/2017	4559	OSAGE PAYLESS FOODS	633 CHASE ST	OSAGE	50461.0	POINT (-92.811539 43.285134)	66.0	MITCHELL	1041100.0	AMERICAN DRY GINS	55.0	SAZERAC NORTH AMERICA	30056	FLEISCHMANN'S GIN	12	750	3.32	4.98	1	4.98	0.75	0.20
S17673000080	03/03/2014	2588	HY-VEE FOOD AND DRUG #6 / CEDAR RAPI	4035 MT. VERNON ROAD S.E.	CEDAR RAPIDS	52403	POINT (-91.60978 41.976835)	57.0	LINN	1031200.0	VODKA FLAVORED	260.0	DIAGEO AMERICAS	41715	SMIRNOFF CRANBERRY VODKA	12	750	8.25	12.37	3	37.11	2.25	0.59

3.1.2. Data Dimensions

- Number of observations: 2535874
- Number of columns: 24

3.1.3. Fixing data types

Some columns were stored with incorrect data types. For example, the Date column was stored as an object and required conversion to a date-time format. Similarly, columns such as County Number, Category, and Vendor Number were floats but needed to be integers.

The table below illustrates the before and after for data type corrections:

Column	Data Type(Before)	Data Type (After)
Date	object	datetime64[ns]
County Number	float64	int64
Category	float64	int64
Vendor Number	float64	int64

3.1.4. Handling Missing data/nulls

Missing values were present in columns related to location and transaction details. Below is a summary of the missing data:

Column	Number of Missing Values
Address	7488
City	7488
Zip Code	7494
Store Location	225184
County Number	363685
County	14484
Category	1452
Category Name	2173
Vendor Number	1
Vendor Name	1
latitude	225184
longitude	225184

To address these missing values, the guidance of stakeholders was followed to fill them using the mode of each feature. This approach ensured that the most frequent values within the dataset were used to replace missing entries.

3.1.5. Standardizing Columns Name Format

To maintain consistency and readability across the dataset, the Python Enhancement Proposal (PEP 8) was followed for naming conventions, converting column names to lowercase and replacing spaces with underscores. The table below shows examples of the transformations made:

Before	After
Invoice/Item Number	invoice_item_number
Date	date
Store Number	store_number
Store Name	store_name
Address	address
City	city
Zip Code	zip_code
Store Location	store_location
County Number	county_number
County	county
Category	category
Category Name	category_name
Vendor Number	vendor_number
Vendor Name	vendor_name
Item Number	item_number
Item Description	item_description
Pack	pack

Bottle Volume (ml)	bottle_volume_ml
State Bottle Cost	state_bottle_cost
State Bottle Retail	state_bottle_retail
Bottles Sold	bottles_sold
Sale (Dollars)	sale_dollars
Volume Sold (Liters)	volume_sold_liters
Volume Sold (Gallons)	volume_sold_gallons

3.2. Feature Engineering

Feature engineering was a critical step in extracting meaningful insights from the dataset. By creating new features based on the existing columns, we added value to the analysis and improved model performance.

3.2.1. Date features

We extracted year, month, and day of the week from the original Date column to allow for temporal analysis of liquor sales trends. These features provided more granular control over time-based data and helped in the analysis of seasonal trends and daily patterns in sales.

3.2.2. Coordinates (Latitude and Longitude)

The *Store Location* column originally stored geographical coordinates in the format "POINT(latitude, longitude)" as a string. To make these coordinates usable for geospatial analysis, we split this column into separate latitude and longitude columns with float64 data types.

3.2.3. Sale Cost

The feature 'sale_cost' was computed as:

$$bottles\ sold\ \times\ state\ bottle\ cost.$$

This column provides insight into the store's expenditure for each transaction.

3.2.4. Profit Percentage

The feature 'profit_percentage' was computed as:

$$\left(\frac{(sale\ dollars - sale\ cost)}{sale\ cost}\right) \times 100.$$

This feature provided a metric for analyzing the profit margins for each sale, which was particularly useful in understanding store profitability across various transactions and product categories.

3.3. Data Visualization and Insights

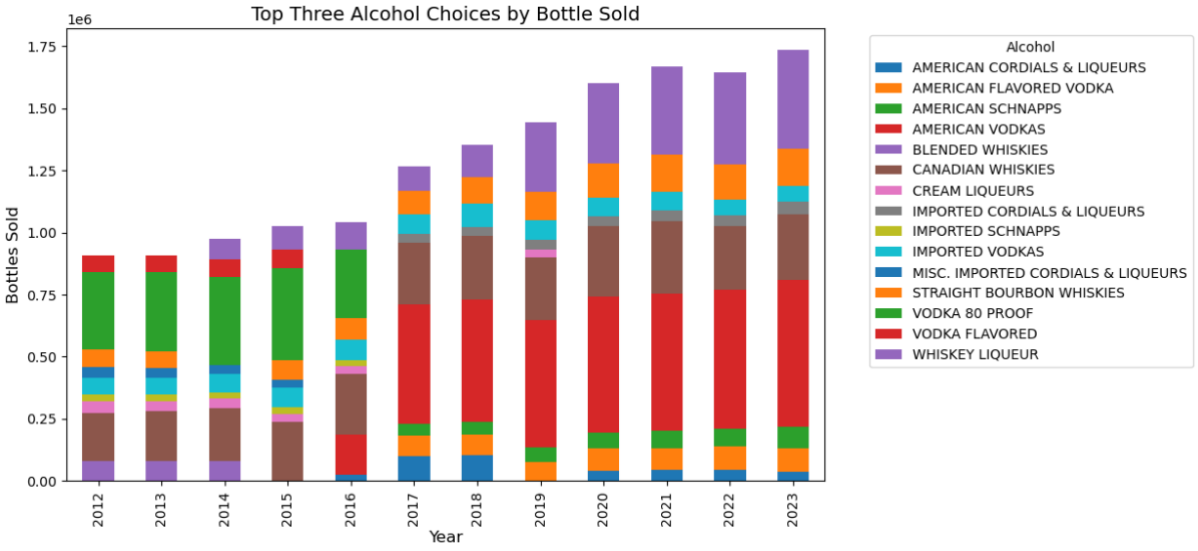
Visualizing the dataset provided insights into customer behavior and sales patterns in Iowa liquor stores. Here are some key insights derived from the data, supported by visualizations that illustrate important trends:

3.3.1. Profit Margin Insight

The first key insight revolves around the average profit margin for each sale in Iowa liquor stores. Based on the data, Iowa liquor stores maintain an average profit margin of **51%** per sale.

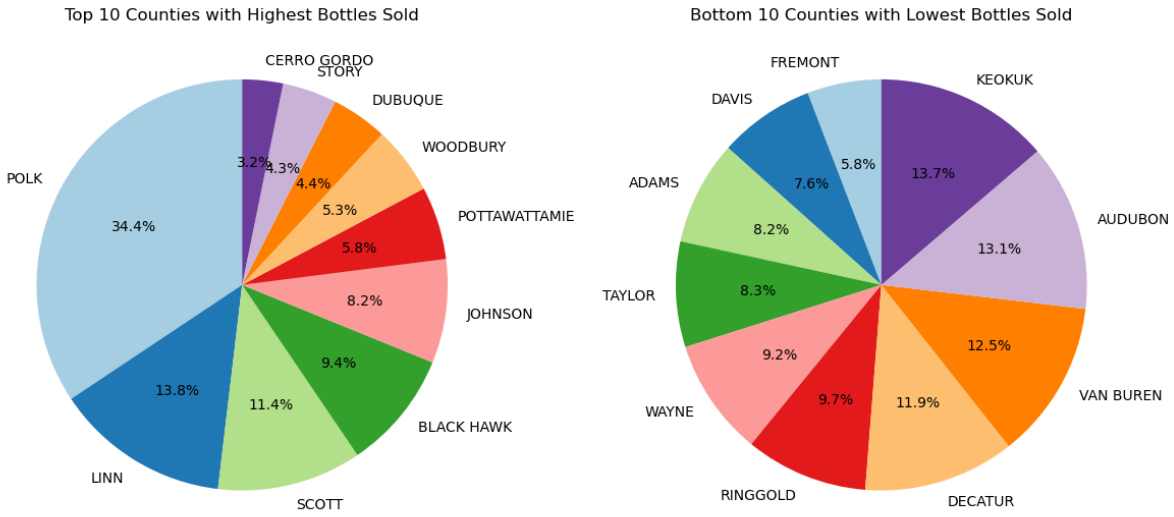
3.3.2. Top 3 Alcohol Choices by Bottle Sold

The bar chart visualizes the top three alcohol choices by bottles sold from 2012 to 2023. Each bar represents the total number of bottles sold each year, with the different colors indicating the contribution of each alcohol category. This stacked format allows for a comparison across both years and categories.



3.3.3. Top and Bottom 10 Counties with Highest and Lowest Bottles Sold

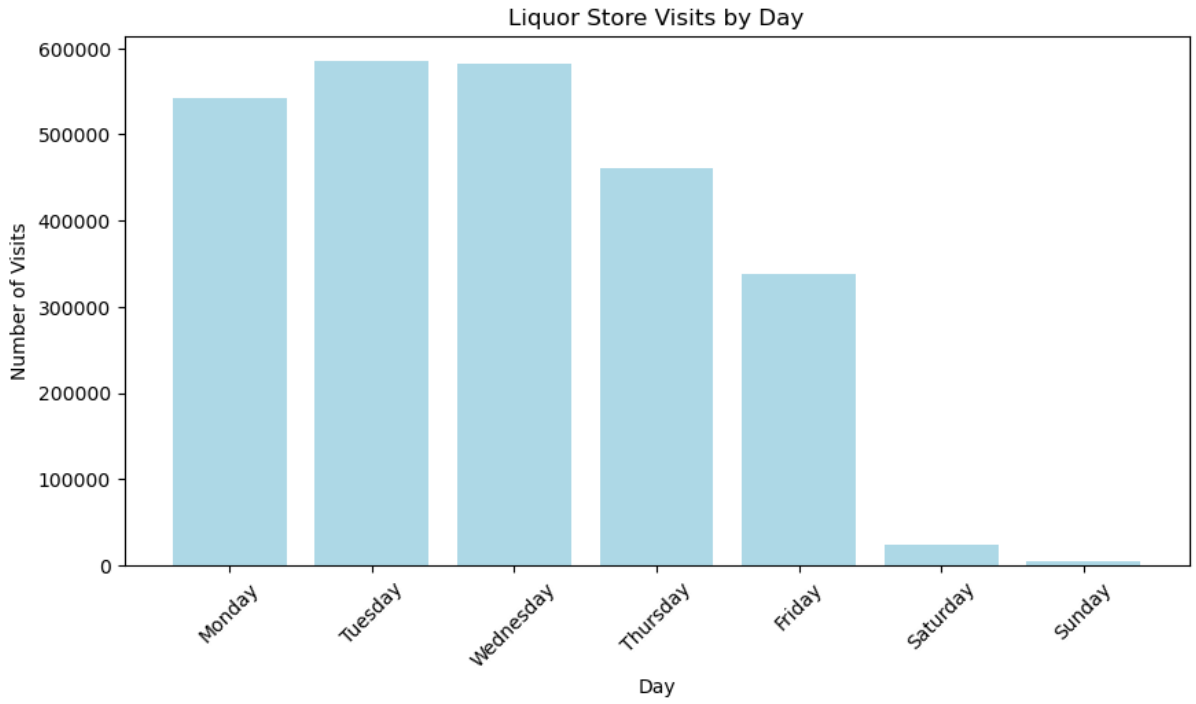
The pie charts visualize the top 10 and bottom 10 counties in Iowa based on the number of liquor bottles sold.



The significant difference in sales between the top and bottom counties suggests that urban and densely populated areas such as Polk, Linn, and Scott counties drive liquor consumption in Iowa. Meanwhile, more rural and less populated counties, such as Keokuk, Audubon, and Van Buren, experience considerably lower sales.

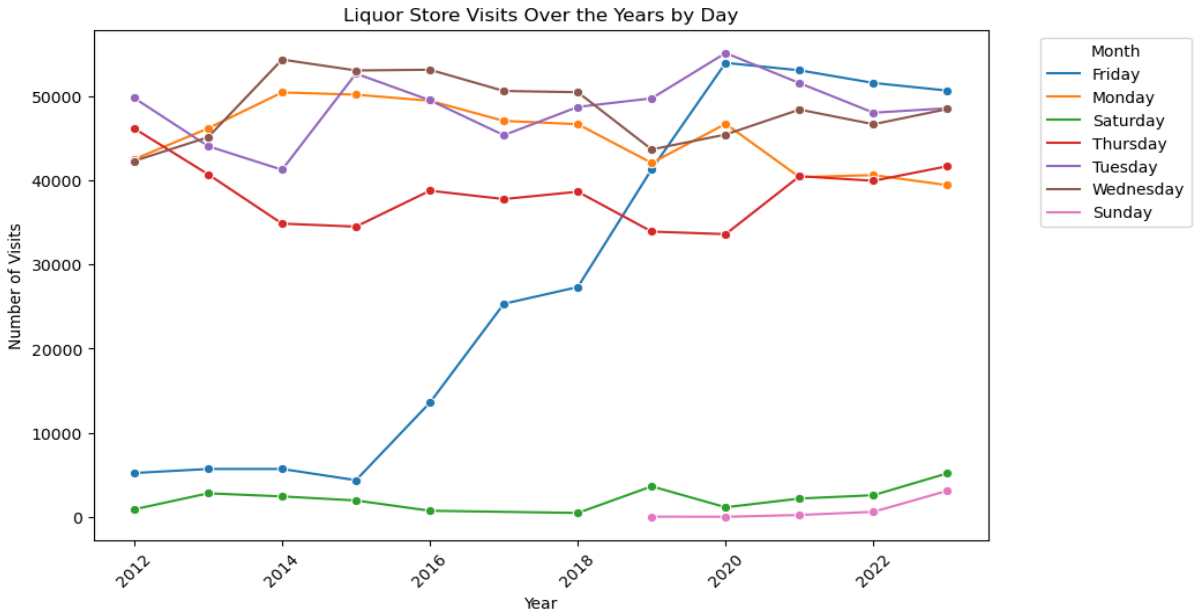
3.3.4. Day of the Week Insights

Over a 11-year period (2012-2023), analysis of liquor store visits by day of the week revealed that Monday, Tuesday and Wednesday consistently had the highest number of visits, as shown in the figure below.



These findings suggest that consumers tend to purchase alcohol earlier in the week, which could inform inventory planning and staffing schedules in liquor stores.

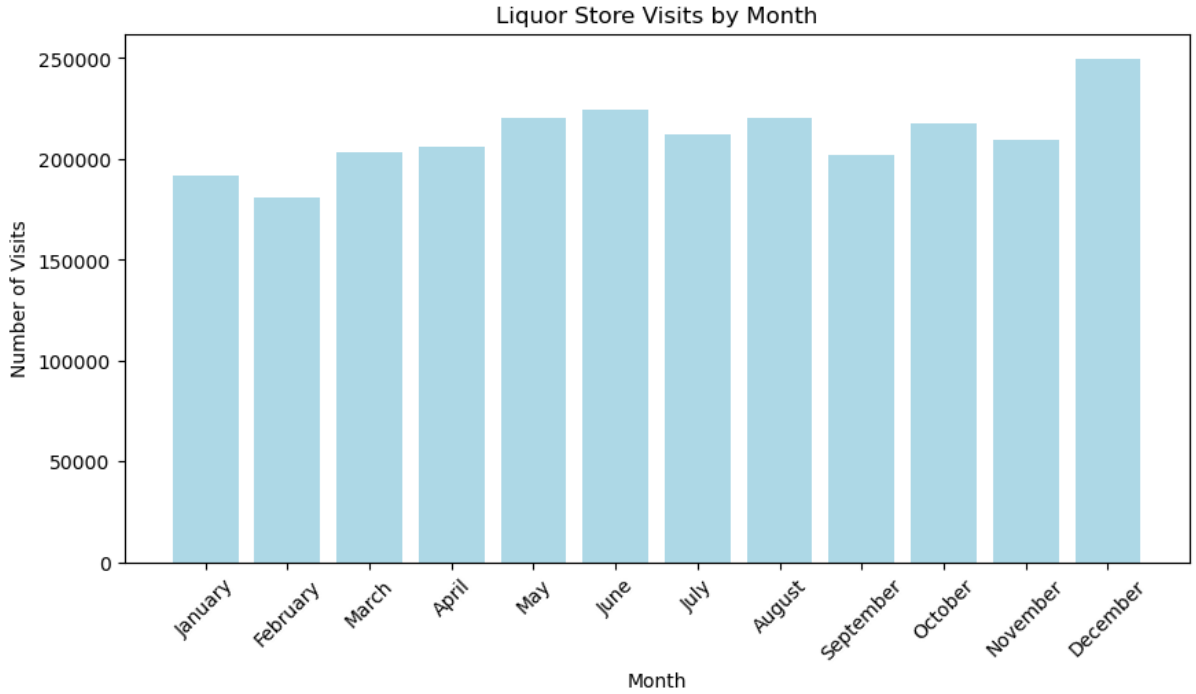
In the line plot illustrating liquor store visits over the years, broken down by day of the week, several interesting trends emerge between 2012 and 2023. This visualization confirms the initial insight that the beginning of the week consistently sees the highest number of visits,
with Tuesday and Wednesday remaining the busiest days throughout the entire decade.



Additionally, starting in 2015, Friday experienced a sharp and sustained increase in visits, with the trend continuing through 2023. This change in consumer behavior suggests that Friday has become a critical day for liquor sales, almost rivaling the mid-week traffic. As a result, liquor stores must ensure they are well-stocked and prepared for the high influx of customers at the end of the week.

3.3.5. Monthly Insights

The analysis of liquor store visits by month over the same 11-year period reveals clear seasonal trends. December consistently records the highest number of visits, exceeding 250,000, driven by the holiday season. Other months like May and November also show elevated traffic, likely due to Memorial Day and Thanksgiving celebrations. In contrast, January and February register the fewest visits, reflecting a post-holiday dip in consumer spending.

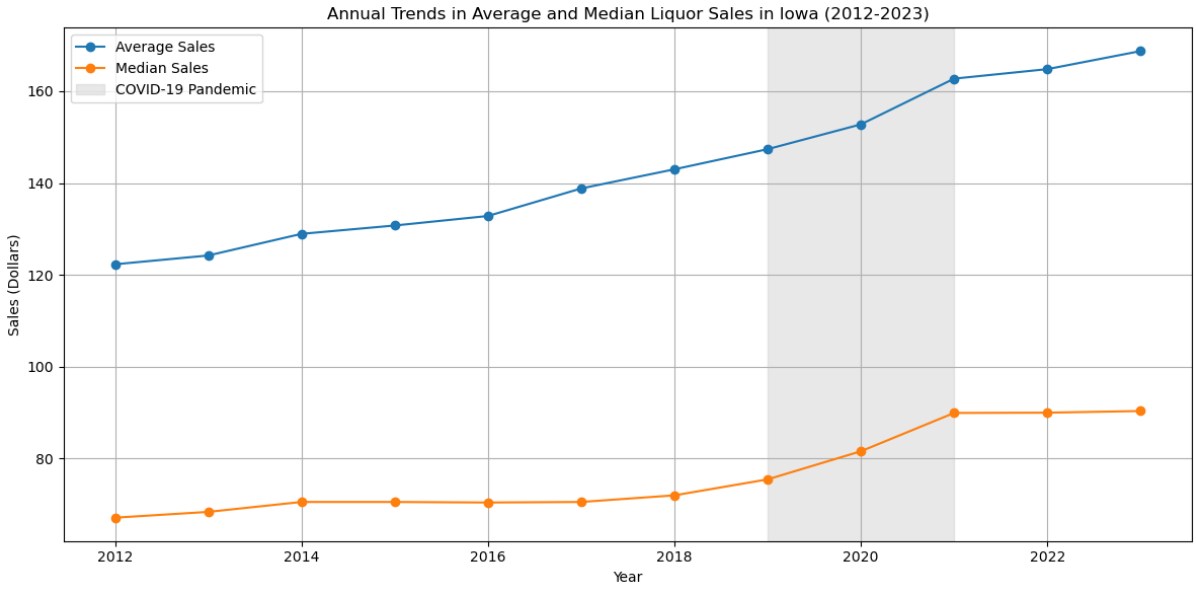


The line plot further illustrates that December has shown an upward trend in visits since 2016, reinforcing its importance for liquor store planning. June also displays a noticeable increase in visits, especially after 2015, possibly due to summer events. The relatively lower and more stable visits in January, February, and March confirm these as slower months for liquor sales.

In summary, these trends indicate that liquor stores should prioritize stocking and marketing efforts for December, with additional focus on May, November, and August. Meanwhile, the slower months provide an opportunity for targeted promotions to boost sales during off-peak periods.

3.3.6. COVID19 – Alcohol Consumption Impact Analysis

The line graph illustrates the annual trends in average and median liquor sales in Iowa from 2012 to 2023, with a specific focus on the period affected by the COVID-19 pandemic (highlighted in gray).



The data reveals a steady increase in both average and median liquor sales from 2012 to 2019, reflecting a consistent rise in alcohol consumption. However, during the pandemic years (2019-2021), there is a noticeable spike in average sales, suggesting that alcohol consumption significantly increased during the lockdown period.

This insight highlights how external factors like global crises can influence consumer behavior, making it important for businesses to adapt to such changes in demand.

4. Modeling

4.1. Train and Test Dataset Methodology

For the supervised learning tasks, the dataset is split into two parts:

- **Training Set (75%)**: This portion of the data is used to train the model, allowing the algorithm to learn patterns and relationships between the features and the target variable (sale_dollars).
- **Testing Set (25%)**: The remaining 25% is reserved for testing the model's performance, ensuring that the model generalizes well to unseen data by evaluating its accuracy and other metrics on this separate dataset.

This train-test split ensures that the model's performance is evaluated in a realistic setting and guards against overfitting.

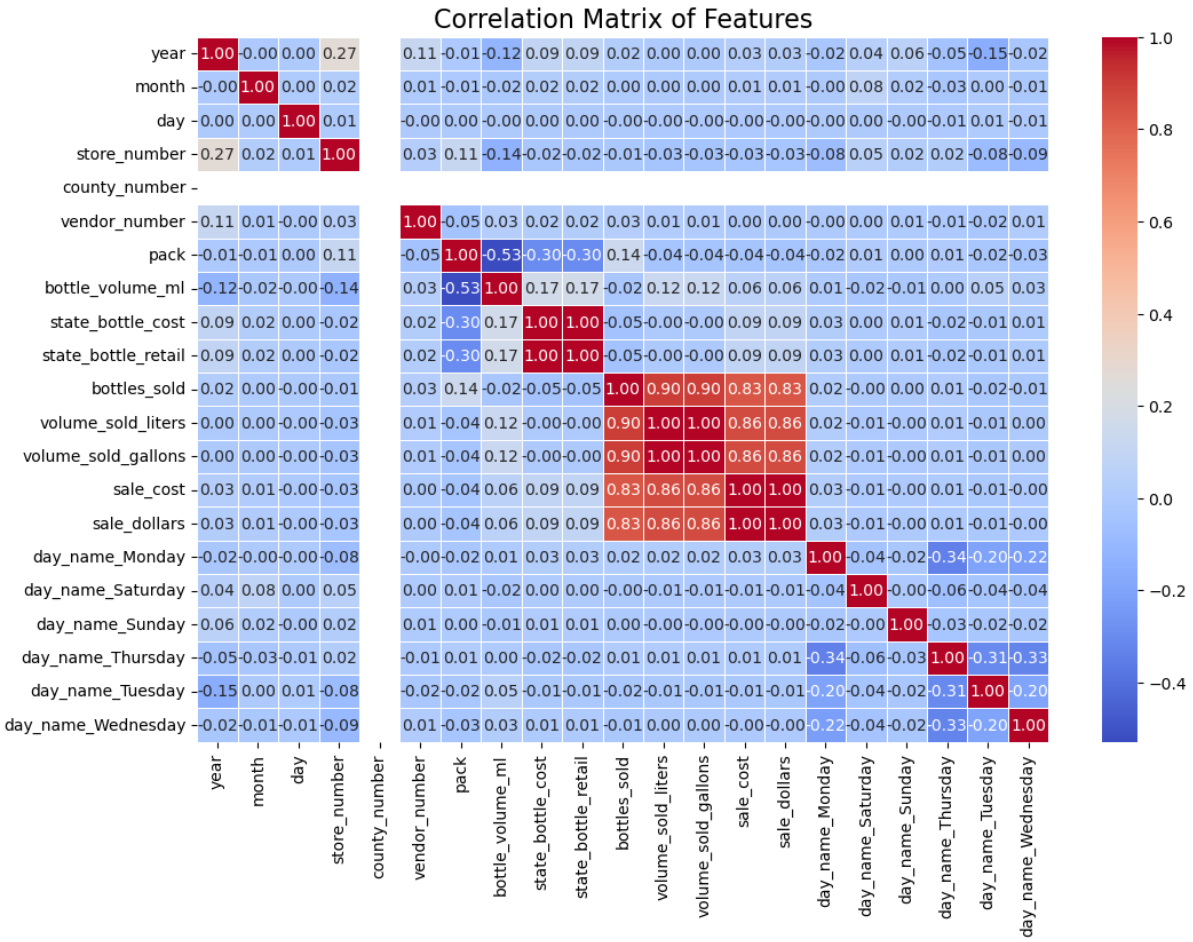
4.2. Regression Problem

4.2.1. Regression Target: 'sale_dollars'

The target variable for the regression model is 'sale_dollars', representing the total revenue generated per transaction.

4.2.2. Feature correlation for Selection

To identify relevant features for predicting this variable, we used a correlation matrix to analyze the relationships between input features and the target.



Features like 'bottles_sold', 'volume_sold_liters', 'volume_sold_gallons' are the most correlated to ‘sale_dollars’ with a strong positive correlation. These relationships indicate that larger quantities sold directly drive revenue and make they a key predictor for sales revenue.

4.2.3. Regression model

For this problem, we use the Random Forest Regressor model from the sklearn package.

4.2.4. Model Train

The model is trained using 75% of the dataset, ensuring that the data follows a chronological order to maintain temporal consistency. The model was fitted to this training set, and the performance on the training data is as follows:

- **Training R-squared**: 0.9926
- **Training RMSE**: 52.0804

4.2.5. Model Evaluation

The model's performance was then evaluated on the remaining 25% of the dataset (test set) to assess its ability to generalize to unseen data. The results on the test set were:

- **Testing R-squared**: 0.7372

- **Testing RMSE**: 502.9272

The drop in R-squared and increase in RMSE between the training and testing sets suggest that while the model performs well on the training data, its performance on the test data is lower, likely due to some overfitting.

4.3. Classification Problem

4.3.1. Classification Target: 'category_name'

The target variable for the regression model is 'category_name', representing the top 5 category names in the dataset. Here we used the label_encoder from the sklearn package to convert the categorical labels into numerical values for the classification task.

4.3.2. Feature Selection

The input features selected for this classification task are the same as those used in the regression model

4.3.3. Classification Model

For this classification problem, we use the Random Forest Classifier model from the sklearn package.

4.3.4. Model Train

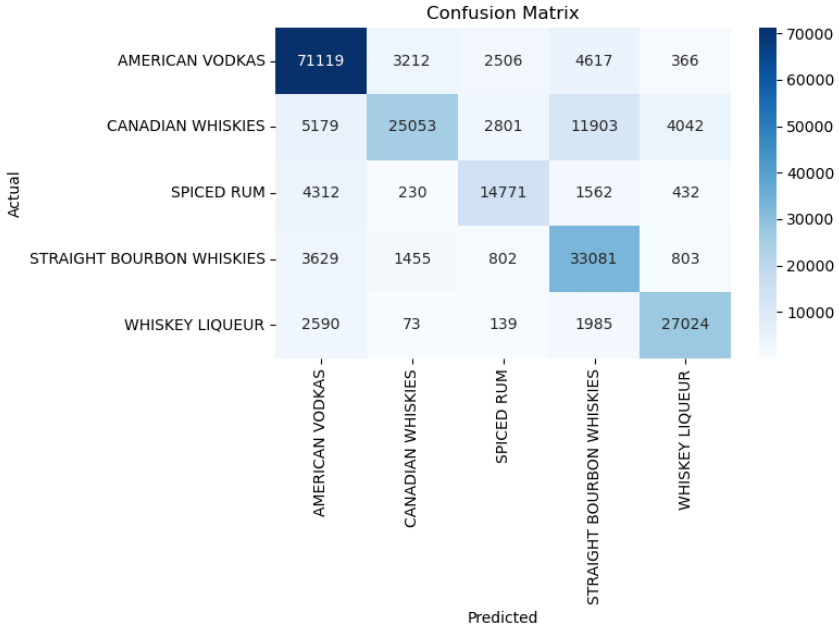
As in the regression problem, the model was trained using 75% of the dataset, ensuring that the data followed a chronological order to preserve temporal consistency. The performance on the training data was:

- **Training Accuracy:** 0.9941

4.3.5. Model Evaluation

The model was evaluated on the remaining 25% of the dataset (test set) to assess its ability to generalize to unseen data. The confusion matrix provides insight into the model's classification performance.

- **Testing Accuracy:** 0.7647



5. CONCLUSION

In this analysis, we explored various aspects of liquor sales in Iowa, focusing on trends across different time periods, consumer behavior, and the impact of external factors such as the COVID-19 pandemic. Using both regression and classification models, we identified key features that influence liquor sales, such as the number of bottles sold and volume of alcohol, which strongly correlate with revenue. The Random Forest models demonstrated strong performance, though some overfitting was observed in regression.

Additionally, the analysis of top-selling alcohol categories and the geographical distribution of sales revealed that urban counties such as Polk and Linn drive most of the liquor sales, while rural counties have significantly lower sales. The COVID-19 pandemic led to a noticeable spike in liquor consumption, with average sales rising sharply during 2020-2021. The higher demand in the beginning of the week, which is quite unexpected.

Overall, this analysis offers a comprehensive understanding of liquor sales trends in Iowa, providing actionable insights for inventory management, marketing strategies, and long-term planning for liquor stores and distributors.