Retail Demand Forecasting and Stockout Prediction: Analyzing Iowa's Liquor Sales

Alaina McKnight

Old Dominion University

BNAL 415/515: Advanced Business Analytics with Big Data Applications

Weiyong Zhang, Ph.D.

December 4, 2024

Contents

Retail Demand Forecasting and Stockout Prediction: Analyzing Iowa's Liquor Sales

Executive Summary

This report addresses the challenges of retail demand forecasting and stockout prediction in Iowa's liquor market, emphasizing the importance of accurate sales forecasting for effective inventory management. Utilizing a dataset of Iowa's Class "E" liquor license sales from January 2012 to the present, the report highlights the impact of data quality issues such as skewness and outliers, which were resolved through Python-based filtering. Time-series forecasting models, specifically ARIMA and Exponential Smoothing (Holt-Winters), were employed to predict future liquor sales, revealing promising results in both models with high predictive accuracy. These findings provide valuable insights for optimizing stock levels and planning for future demand, ultimately improving business efficiency in the state's unique liquor distribution system.

Introduction

Retail demand forecasting and stockout prediction are critical challenges for retailers and distributors in the State of Iowa's liquor market. Despite the availability of detailed and comprehensive sales data, inefficiencies in predicting demand often lead to stockouts or overstocking, which can result in lost sales, increased costs, and dissatisfied customers. This issue is particularly relevant in Iowa, where the state controls wholesale liquor distribution, making the following dataset a unique and valuable resource for addressing these challenges.

The Dataset: An Overview

The dataset that will be utilized in this project contains the spirits purchase information of Iowa Class "E" liquor licenses by product and date of purchase from January 1, 2012 to current. While searching for a suitable dataset for this project (which did come with obstacles), I came across a reddit page where users could suggest interesting datasets for practicing classification, regression, etc. This dataset caught my eye. Initially, it contained 30,305,765 items. I utilized BigQuery from Google Cloud to create a sample subset containing 150,000 randomly selected items for further analysis. The dataset contains wholesale orders of liquor by all grocery stores, liquor stores, convenience stores, etc., with details about the store and location, the exact liquor

brand and size, and the number of bottles ordered. It originally consisted of the following 24 variables:

invoice_line_no	vendor_no
date	vendor_name
store	itemno
name	item_desc
address	pack
city	bottle_volume_ml
zipcode	state_bottle_cost
store_location	state_bottle_retail
county_number	bottles_sold
county	sale_dollars
category	volume_sold_liters
category_name	volume_sold_gallons

The original dataset was comprised of 15 categorical variables and 8 continuous variables. The sample size was sufficiently large enough to allow for a meaningful exploration of patterns.

Data Quality Assessment

Initially, the dataset was relatively dense, but a small proportion of items required filtering. Several hundred missing values were identified in the county, category, and category_name attributes, which were excluded using OpenRefine prior to further analysis. Additionally, the county_number and store_number attributes were removed due to redundancy.

After loading the dataset as an Excel file into SPSS Modeler, the initial data audit revealed that although the values and fields were complete, many of the attributes were extremely skewed.

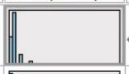
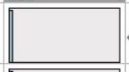
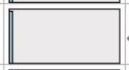
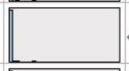
Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Median	Mode	Unique	Valid
pack		Continuous	1.000	288.000	12.162	7.881	3.043	12.000	12.000	--	149141
bottle_volume_ml		Continuous	20.000	225000.000	874.585	774.423	162.780	750.000	750.000	--	149141
state_bottle_cost		Continuous	0.000	5500.000	10.789	17.058	224.941	8.500	8.250	--	149141
state_bottle_retail		Continuous	0.000	8250.000	16.193	25.587	224.953	12.750	15.000	--	149141
bottles_sold		Continuous	-60.000	2304.000	10.979	30.034	23.946	6.000	12.000	--	149141
sale_dollars		Continuous	-1125.000	63336.000	147.734	516.509	34.455	77.400	90.000	--	149141
volume_sold_liters		Continuous	-45.000	4032.000	9.299	37.199	35.015	4.800	9.000	--	149141
volume_sold_gallo...		Continuous	-11.880	1065.140	2.454	9.827	35.014	1.260	2.770	--	149141

Figure 1. Initial data audit using unbalanced dataset

For example, the skewness of both state_bottle_cost and state_bottle_retail were well over 200. The Quality Tab revealed that many of the fields contained large amounts of outliers and extreme values.

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes
county	Nominal	--	--
category_na...	Nominal	--	--
pack	Continuous	4218	8
bottle_volum...	Continuous	5	14
state_bottle_...	Continuous	213	226
state_bottle_...	Continuous	213	226
bottles_sold	Continuous	536	600
sale_dollars	Continuous	400	553
volume_sold...	Continuous	268	476
volume_sold...	Continuous	268	476

Figure 2. Outliers and extreme values using unbalanced dataset

To address the issue and balance the dataset, Python was used to identify and remove outliers and extreme values. A standard deviation threshold of 3 was applied to detect outliers, while extreme values were identified using a threshold of 5. Additionally, any negative values were

removed as they were deemed invalid for this context. After filtering out negative values, the cleaned DataFrame contained 132,752 items and 22 rows. The complete code for this process can be accessed via the link provided in the References.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Median	Mode	Unique	Valid
pack		Continuous	1.000	30.000	11.414	5.004	1.222	12.000	12.000	--	132752
bottle_volume_ml		Continuous	20.000	3000.000	872.493	484.707	0.703	750.000	750.000	--	132752
state_bottle_cost		Continuous	0.000	29.200	9.748	5.386	1.086	8.330	8.250	--	132752
state_bottle_retail		Continuous	0.000	43.800	14.633	8.079	1.085	12.500	15.000	--	132752
bottles_sold		Continuous	1.000	96.000	7.499	6.350	2.134	6.000	12.000	--	132752
sale_dollars		Continuous	0.000	603.840	97.625	84.840	1.661	72.000	90.000	--	132752
volume_sold_liters		Continuous	0.020	21.000	6.106	4.797	0.700	4.500	9.000	--	132752
volume_sold_gallo...		Continuous	0.000	5.550	1.610	1.267	0.701	1.190	2.770	--	132752

Figure 3. Final data audit of the balanced dataset

A final data audit of the balanced dataset showed significantly reduced skewness. While some outliers remained, the most critical objective—removing extreme values—was successfully achieved.

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes
category_na...	Nominal	--	--
pack	Continuous	4	0
bottle_volum...	Continuous	23	0
state_bottle_...	Continuous	1943	0
state_bottle_...	Continuous	1947	0
bottles_sold	Continuous	140	508
sale_dollars	Continuous	1878	196
volume_sold...	Continuous	2777	0
volume_sold...	Continuous	2777	0

Figure 4. Outliers and extreme values using balanced dataset

Describing the Data

Before diving deeper into the analysis, a brainstorming session was conducted to generate key questions that would guide the exploration of the data. A few questions to be answered are:

What does the Regional Distribution of Sales look like?

Which period/season is most profitable/What are the seasonal trends in sales?

What types of trends in percent markup can be found from the data?

What will the next few years of total sales look like?

To answer these questions, SPSS Modeler and Python were used to create distributions for analysis. The distribution of sales by region can provide insights on which counties purchase the most liquor, and thus which areas need to be stocked more frequently. It can also provide insights on which counties need more marketing investment.

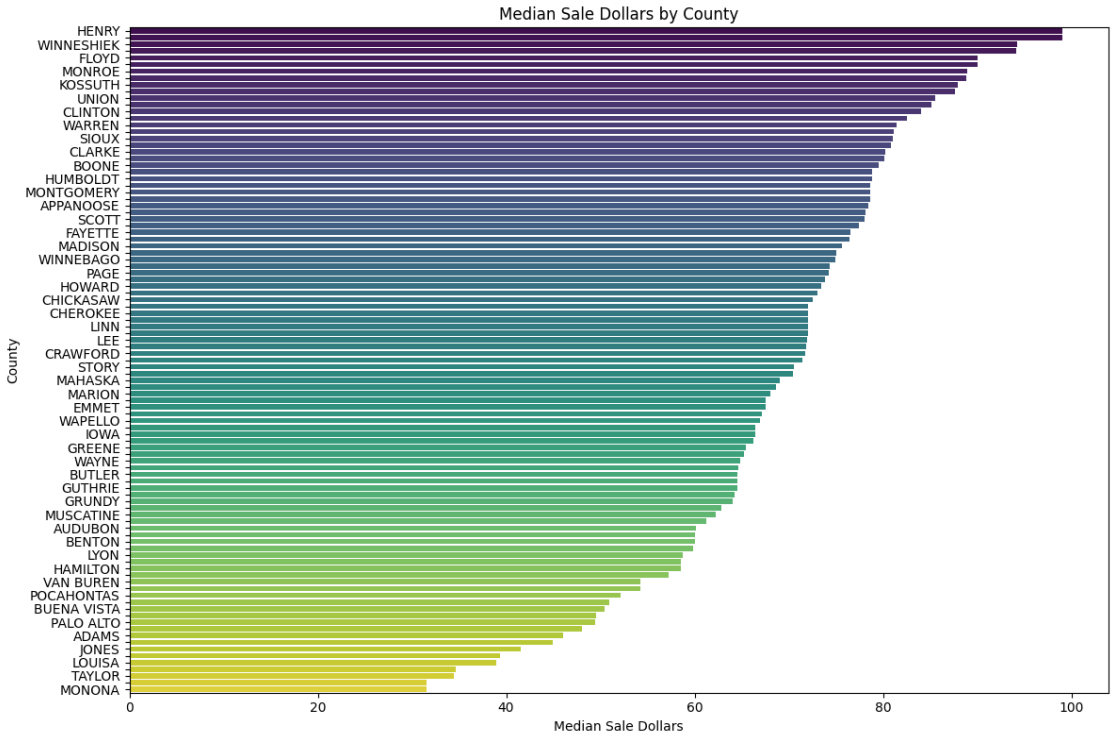


Figure 5. Overall median sale dollars by county in Iowa 2012-2024

The plot above displays the overall median sales in dollars per county in Iowa. We can infer that the Henry and Winneshiek populations purchase and drink more liquor than the populations of Taylor and Monona.

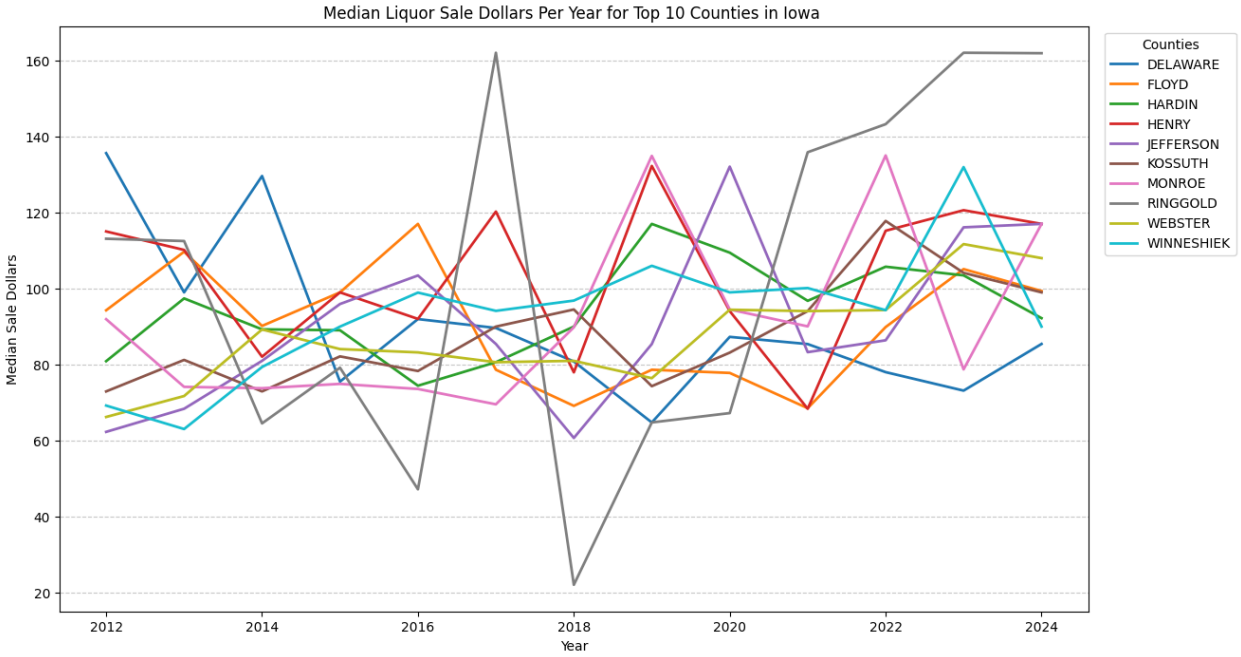


Figure 6. Median liquor sales in dollars per year for top 10 counties in Iowa

To gather further insights, a distribution was created to show the median liquor sales in dollars per year for the top 10 counties in Iowa. From this visual, we can gather that Ringgold County had the greatest fluctuation of sales over the years.

Next, a distribution was created to observe any seasonal trends in the data. The data was grouped by quarter, using Quarter 1 (Q1): January, February, March, Quarter 2 (Q2): April, May, June, Quarter 3 (Q3): July, August, September, and Quarter 4 (Q4): October, November, December.

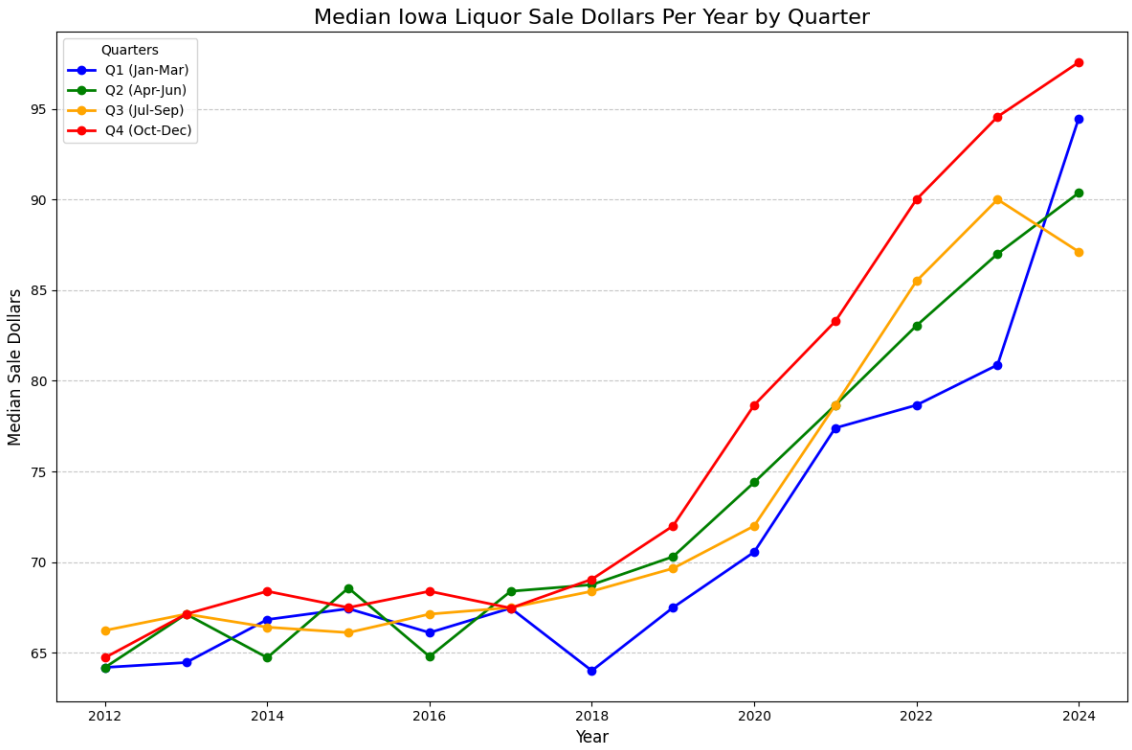


Figure 7. Median Iowa liquor sales in dollars per year by quarter

It is easy to observe that the most sales occur from October to December, but there are some interesting fluctuations from 2012 to 2018. Another observation to note is the steep upward trend in median sales from 2018 to current.

Observing the State Bottle Cost versus the State Bottle Retail gathered insights on price points over the years.

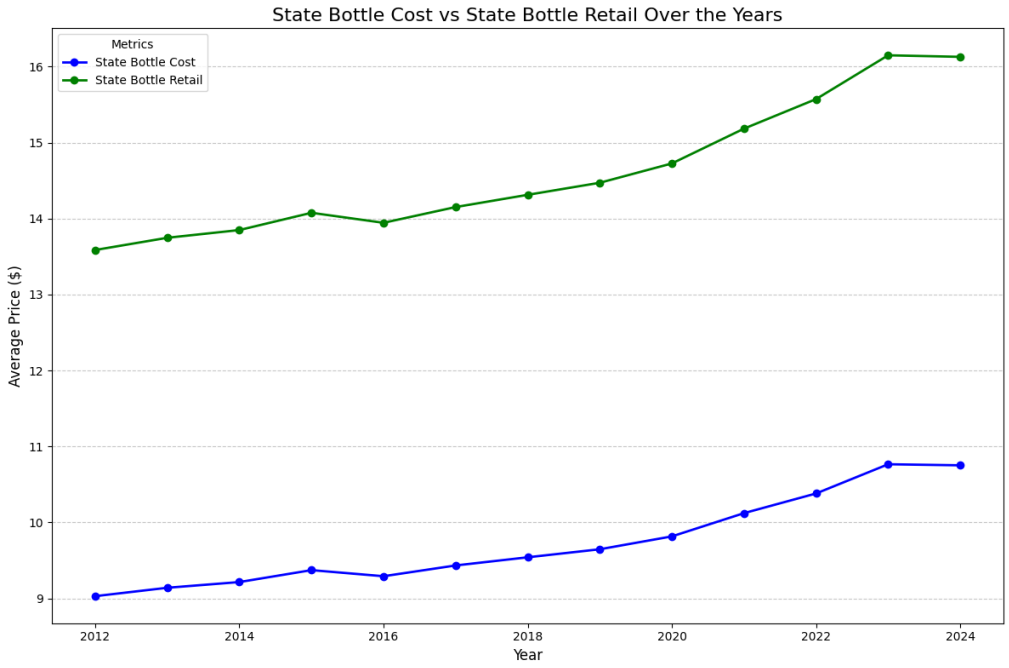


Figure 8. State bottle cost versus state bottle retail 2012-2024

The percent markup was calculated observed using the following formula:

$$\text{Markup Percentage} = \frac{\text{Selling Price - Cost}}{\text{Cost}} \times 100$$

Figure 9. Markup calculator formula

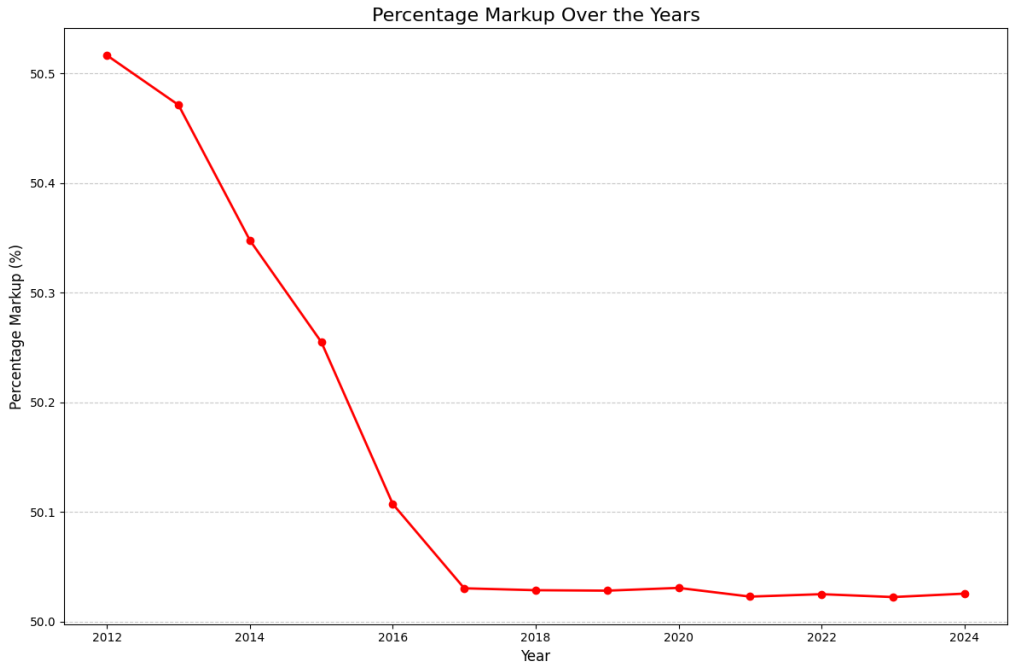


Figure 10. Markup percentage 2012-2024

This graphic is slightly misleading because visually there is a large decline, however, the actual greatest change was less than 0.5% from 2012 to 2017. From 2017 on there is minimal fluctuation.

Sales Predictions Using Time Series Forecasting Models

Given the sequential nature of sales data, time-series forecasting methods are particularly well-suited for demand prediction. For this analysis, the ARIMA (AutoRegressive Integrated Moving Average) model and Exponential Smoothing (Holt-Winters) model were selected to predict future sales based on the provided dataset.

The following seven variables were incorporated as inputs for the time-series model:

- Pack
- Bottle Volume (ml)
- State Bottle Cost
- State Bottle Retail
- Bottles Sold
- Volume Sold (liters)

- Volume Sold (gallons)

The sale_dollars variable was utilized as the target for prediction. Monthly intervals were established as the time unit, and two separate time series models were executed, one for the ARIMA model and the other for the Holt-Winters Additive method.

ARIMA Model

Target: sale_dollars

Model Information

Model Building Method		ARIMA
		Non-seasonal p=0,d=0,q=0; Seasonal p=0,d=1,q=1
Number of Predictors		4
Model Fit	MSE	3,558,846.146
	RMSE	1,886.490
	RMSPE	3.598
	MAE	1,435.789
	MAPE	2.794
	MAXAE	5,938.511
	MAXAPE	12.044
	AIC	2,102.674
	BIC	2,120.281
	R-Squared	0.931
	Stationary R-Squared	0.812
Ljung-Box Q(#)	Statistic	32.403
	df	17.0
	Significance	0.0

Figure 11. ARIMA model building method using sales in dollars as the target variable

The ARIMA model delivered impressive results, achieving an R-squared value of 0.931, which reflects a strong fit to the sales data. Additionally, a high Stationary R-squared value of 0.812 indicates that the model successfully captured both the underlying trends and the stationary components of the dataset. The Mean Absolute Percentage Error (MAPE) of 2.794 signifies strong predictive accuracy, suggesting that the ARIMA model provides reliable and precise forecasts for future sales.

Exponential Smoothing (Holt-Winters) Model

Target: sale_dollars

Model Information

Model Building Method		Exponential Smoothing Winters' additive
Number of Predictors		1
Model Fit	MSE	10,274,955.476
	RMSE	3,205.457
	RMSPE	6.354
	MAE	2,604.118
	MAPE	5.169
	MAXAE	7,688.272
	MAXAPE	19.073
	AIC	2,489.334
	BIC	2,498.445
	R-Squared	0.811
	Stationary R-Squared	0.730
Ljung-Box Q(#)	Statistic	50.432
	df	15.0
	Significance	0.0

Figure 12. Exponential Smoothing (Holt-Winters) model building method using sales in dollars as the target variable

The Exponential Smoothing (Holt-Winters) model also demonstrated high predictive accuracy, with an R-squared of 0.811 and a Stationary R-squared of 0.730, signifying a solid capture of the sales data's underlying patterns. The MAPE of 5.169, while slightly higher than that of the ARIMA model, still indicates a satisfactory level of forecast accuracy, making it a valuable alternative for sales prediction.

Results and Future Trends

To visualize the results, a Time Series plot was generated to display the predicted sales data over the next 24 months. Both models indicate a continued upward trend in sales,

suggesting that demand will persistently grow. However, based on the trends observed in the forecasts, this growth may not continue at the same steep rate, pointing to a potential stabilization in sales growth moving forward.

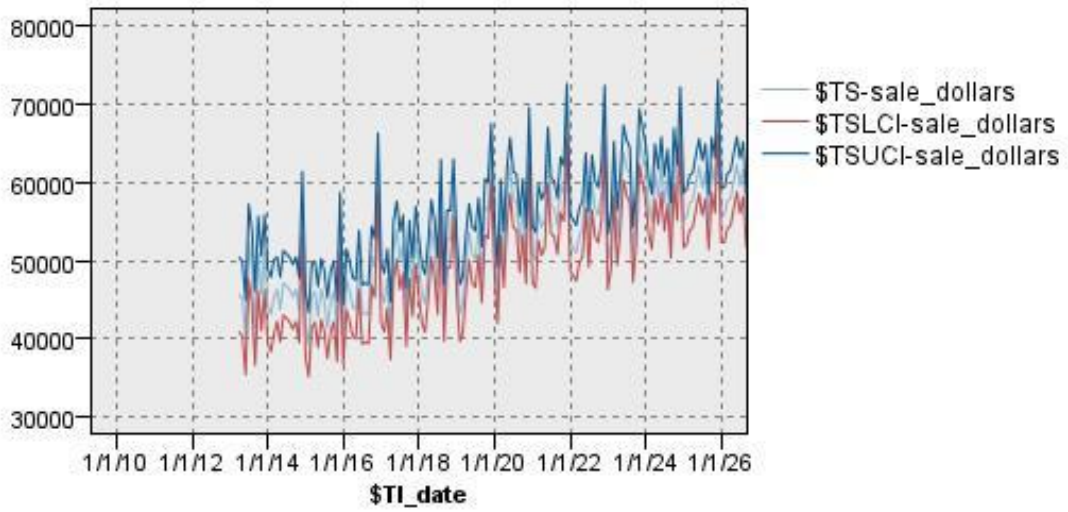


Figure 13. Time-series plot using predicted data from ARIMA forecasting model

These results underscore the effectiveness of time-series forecasting models in predicting future sales trends, offering valuable insights for business planning and decision-making.

Conclusion

This analysis demonstrates the potential of advanced forecasting techniques to address the challenges of demand prediction and stockout prevention in Iowa's liquor market. By leveraging time-series models such as ARIMA and Exponential Smoothing, the study provides strong insights into future sales trends, highlighting a sustained upward trajectory with potential stabilization in growth. The data cleaning process, including handling outliers and skewness, significantly improved the quality of the dataset, ensuring the reliability of the forecasts. These findings underscore the importance of using data-driven approaches for strategic decision-making in the retail sector, offering valuable insights for stakeholders in Iowa's liquor distribution system.

Iowa Liquor Sales Data Dictionary

1. **Invoice/Item Number:** A unique identifier for each transaction.
2. **Date:** The date the transaction occurred.
3. **Store Number:** A unique identifier for each store.
4. **Store Name:** The name of the store making the purchase.
5. **Address:** The address of the store.
6. **City:** The city where the store is located.
7. **Zip Code:** The postal code for the store's location.
8. **Store Location:** The latitude and longitude coordinates of the store.
9. **County Number:** A unique numeric code representing the county.
10. **County:** The name of the county where the store is located.
11. **Category:** A numeric code representing the liquor category.
12. **Category Name:** The descriptive name of the liquor category.
13. **Vendor Number:** A unique identifier for the vendor.
14. **Vendor Name:** The name of the vendor supplying the liquor.
15. **Item Number:** A unique identifier for the liquor product.
16. **Item Description:** A detailed description of the liquor product.
17. **Pack:** The number of bottles per case.
18. **Bottle Volume (ml):** The volume of each bottle in milliliters.
19. **State Bottle Cost:** The cost of a bottle to the state.
20. **State Bottle Retail:** The retail price of a bottle set by the state.
21. **Bottles Sold:** The number of bottles sold in the transaction.
22. **Sale (Dollars):** The total sale amount in dollars for the transaction.
23. **Volume Sold (Liters):** The total volume of liquor sold in liters.
24. **Volume Sold (Gallons):** The total volume of liquor sold in gallons.

References

1. Iowa Liquor Sales. (n.d.). *Iowa Liquor Sales Dataset*. Iowa Data. Retrieved from https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy/data
2. Iowa Liquor Sales. (n.d.). *Iowa Liquor Sales on Google Cloud Marketplace*. Iowa Department of Commerce. Retrieved from https://console.cloud.google.com/marketplace/details/iowa-department-of-commerce/iowa-liquor-sales?pli=1&project=proven-impact-421222
3. Iowa Liquor Sales. (n.d.). *Data Dictionary: Iowa Liquor Sales*. Iowa Data. Retrieved from https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy/about_data
4. Corporate Finance Institute. (n.d.). *Markup Calculator Formula*. Retrieved from https://corporatefinanceinstitute.com/resources/financial-modeling/markup-calculator-formula/
5. EDA Notebook: https://colab.research.google.com/drive/1qLn5kNlYFgsuTcScWDrMbOBgXTaaM82w?usp=sharing
6. IBM SPSS Modeler Final Model: