## Group 33

## Declaration on Plagiarism

| | |
|---|---|
| **Name/s:** | Teena Sharma and Teenu Prathyush |
| **Student Number/s:** | 21261593 and 21262966 |
| **Programme:** | MSc. in Computing (Data Analytics) |
| **Module Code:** | CA660 |
| **Assignment Title:** | Statistical Data Analysis Assignment (CA660) |
| **Submission Date:** | 01-12-2021 |
| **Module Coordinator:** | Marija Bezbradica |

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion, or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines.

Name: Teena Sharma Date: 01-12-2021

Name: Teenu Prathyush Date: 01-12-2021

# Iowa Liquor Sales Analysis and Prediction using Statistical Methods and Machine Learning

ABSTRACT:

These days, sales trend analysis is in vogue. The analysis of old sales data and identifying patterns are becoming increasingly important. This can be highly beneficial for business growth. The purpose of this study is to analyze sales data on liquor collected from Iowa in order to do exploratory analysis. Specifically, the data consists of spirits purchases in the state of Iowa. In order to compare an area according to sales, we first found out what was happening in the top two cities, 'Des Moines' and 'Cedar Rapids'. We perform exploratory data analysis on sales in these two cities, then we perform hypothesis testing to answer certain questions that arise from the exploratory analysis. Finally, we use a random forest regressor model for accurately predicting liquor sales in these two cities. The random forest regressor model is then tested with metrics such as Mean Absolute error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to validate the accuracy of the model.

*Keywords:* Random Forest Regressor model, Z-test, Mean absolute error (MAE), Mean squared error (MSE), and Root mean square error (RMSE).

## I. INTRODUCTION

In today's market, predictions about any sale are in high demand. The accuracy of sales predictions is very important for all businesses to maintain their organization and workforce. An evaluation tool used to predict future performance is sales prediction. It uses past and current sales data to do so. It is helpful for companies to know about their customers' likes and dislikes, as well as which products generate more revenue. By analysing sales data, we allow companies to focus on what is most relevant for their growth. Consequently, marketing techniques can be utilized to target customers in a selective way[1].

The aim of this work is to use data collected from liquor sales in Iowa to understand the factors that will increase profits for the company, such as the most popular liquor brands or cities that make the most profit, etc., so that retailers can make more money. To achieve an accurate prediction, a large amount of data is necessary. Our current analysis is based on both old and recent data, so we can accurately predict and analyse the mentioned factors.

We used data from Iowa, which maintains a monopoly on the wholesale of alcohol throughout the entire state due to its alcohol beverage control law. Consequently, private retailers must purchase their alcohol from the state before they can sell it to individual consumers. Therefore, our prediction will assist retailers in choosing which products are highly popular and will bring them more profit.

## II. RELATED WORK

We have explored relevant research papers and articles in order to gain an understanding of the various predictions and analyses made in the area of liquor sales. Michael Salmon published an article on 3rd April 2017, that was quite informative. In their project, they used 2015 sales data to train a model that predicted total 2016 sales based on Q1 2016 sales. Upon completing their analysis, they concluded that the model for 2016 predicted total liquor sales of $292 million, a 3% increase from 2015 ($284 million). For 2016, the state estimated

[1] G. T. G. U. Shreya Kohli, Sales Prediction Using Linear and KNN Regression "Springer Link," 26 July 2020.

that it would profit at $98 million, up from $33 million ($95 million) in 2015 [1].

There was one research paper that had significant references, which gave us valuable insights. There was a detailed explanation of liquor consumption in [2]. The researchers implemented various algorithms like Multiple Linear Regression, Support Vector Machines, and Long Short-Term Memory in their research to generate forecasts with maximum accuracy. They achieved excellent results by implementing LSTM. A neural approach is employed in LSTM, which is observed to be faster than SVM and regression techniques. Keras, a python deep-learning library, is used to implement their LSTM algorithm. In order to support Keras, they used TensorFlow as a backend.

## III. DATASET AND EXPLORATORY ANALYSIS

*DATASET:* Our dataset was obtained from the Iowa Open Data website. The Iowa Liquor Sales data contains all spirits purchases made by Iowa Class "E" liquor licenses between January 1, 2012, and October 31, 2021. There are 22.5M rows and 24 columns in the dataset, where each row represents a product purchase. The dataset contains a lot of values and as a result, we cannot open the complete dataset on Excel as it has a limit on the number of rows and columns. Therefore, we have used RStudio for cleaning and processing the dataset.

After cleaning and processing our data in RStudio, we used the following columns to carry out the exploratory data analysis: Date,Store_Number,Store_Name,City_Number, City, Category_Number, Category_Name, Vendor_Number,Vendor_Name,Item_Number, Item_Name, Pack, Bottle_Volume_in_Liters, State_Bottle_Retail,Bottles_Sold, Volume_Sold_in_Liters and Sale_in_Dollars. The above columns are mostly of date, numeric, and character types.

*EXPLORATORY DATA ANALYSIS:*

The dataset is filtered to only include sales from the top two cities in Iowa (in terms of population) - 'Des Moines' and 'Cedar Rapids'. We will be performing exploratory analysis on sales in these two cities from Jan. 2017 - Oct. 2021.

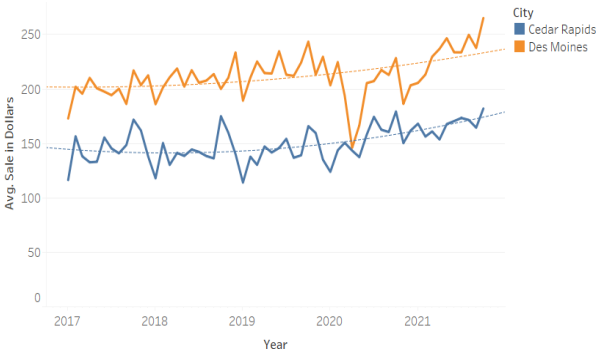


Figure 1: Average monthly sales from 2017 to 2021

From the above figure, it is observed that average liquor sales in Des Moines are significantly higher than in Cedar Rapids. It is also observed that liquor sales in both cities follow a similar trend with October having the highest average sales and January having the lowest average sales in most cases. There are some months where average sales drop significantly, for example, Des Moines recorded the lowest average sales in the month of April 2020. This could be a result of the pandemic where lockdowns contributed to a decrease in sales. Liquor sales have gradually increased in the last year, in particular, Des Moines, which has seen a rapid increase in its Liquor sales.

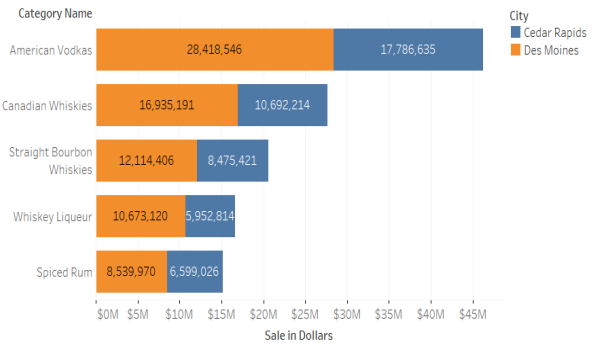


Figure 2: Top 5 categories with highest sales

The above figure shows the top 5 categories of liquor in regards to total sales from 2017 to 2021 in the respective cities. American Vodkas generate the highest revenue in both Des Moines and Cedar Rapids.

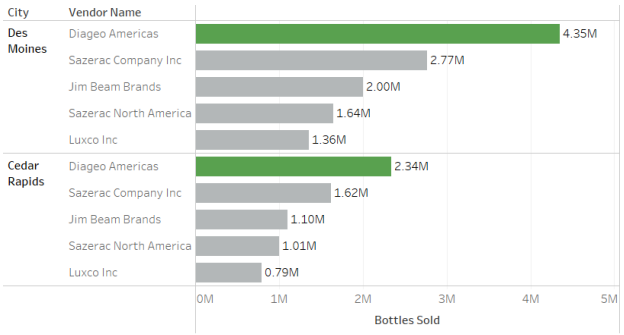


Figure 3:Top 5 vendors with the highest bottles sold

The above figure shows the top 5 vendors in both cities from 2017 to 2021. Diageo Americas account for a total of nearly 6.7 million bottle sales in both cities.

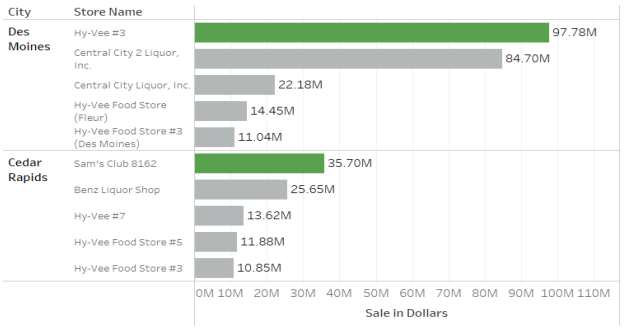


Figure 4: Top 5 Stores with the highest revenue

The above figure shows the top 5 stores in both cities from 2017 to 2021. Hy-Vee #3 has generated a revenue of 97.78 million dollars, followed by Central City 2 Liquor Inc., while there is a huge difference to the next highest revenue-generating store in Des Moines. The top 5 stores in Cedar Rapids haven't generated revenue anywhere nearly as much as the top 2 stores in Des Moines, however, Sam's Club has recorded the highest revenue in Cedar Rapids with 35.7 million dollars over 5 years.

## IV. HYPOTHESES AND OR RESEARCH QUESTIONS

Upon doing further exploratory data analysis in figure 1, we see that there is a trend in the average sales during the months of June and July where the average sales are quite good during June but it goes down during the month of July in almost all the years. However, this is not the case in 2021 as we observe that average sales during June and July are almost the same. We would like to know if this is right, therefore, we carry out a hypothesis test to see if there is any significant difference between the average sales during these two months.

Null Hypothesis ($H_0$): There is no significant difference in the average sales between the months – June and July.

Alternative Hypothesis ($H_1$): There is a significant difference in the average sales between the months – June and July

We carry out a Two-Sample z-test for comparing the two means at a confidence interval of 95%. The formula is given below.

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where $\overline{x_1}$ and $\overline{x_2}$ are the sample means of June and July respectively. $\Delta$ is the hypothesized difference between the population means. n1 and n2 are the sample sizes of June and July respectively. $\sigma 1$ and $\sigma 2$ are the standard deviations of the two populations.

$\overline{x_1} = 233.70$, $\overline{x_2} = 233$

n1 = 18982, n2 = 17605

$\sigma 1 = 764.528$, $\sigma 2 = 2005.72$

We then substitute the values in the above formula to obtain a z-score of 2.22.

The z-score lies outside the 95% confidence interval region therefore we reject the null hypothesis. We conclude by stating that there is a significant difference in the average sales between the two months – June and July.

In the next section, we are going to train a machine learning model to accurately predict

liquor sales based on historic data present in our dataset.

## V. METHODS USED AND WHY

Our analysis is based on a random forest regression model. It operates by constructing several decision trees during training time and outputs the mean of the classes as the prediction of all the trees [4]. The reason behind choosing the Random Forest algorithm to implement our model is that it has the advantage of being simple and generally produces good predictions and it also handles large datasets efficiently [4].

To train our random forest regression model, we first removed all the unnecessary columns and included only the numerical columns that will be used in the model. The columns that have been used are Store_Number, City_Number, Category _Number, Vendor_Number, Item_Number, Pack, Bottle_Volume_in_Liters, State_Bottle_Retail, Bottles_Sold, Volume_Sold_in_Liters.

We then split the dataset into a training set (80%) and test set (20%) so that the model learns from the training set while the test set remains untouched and is only used for prediction purposes. We then find the correlation between the Independent variables and dependent variable (Sale_in_Dollars).

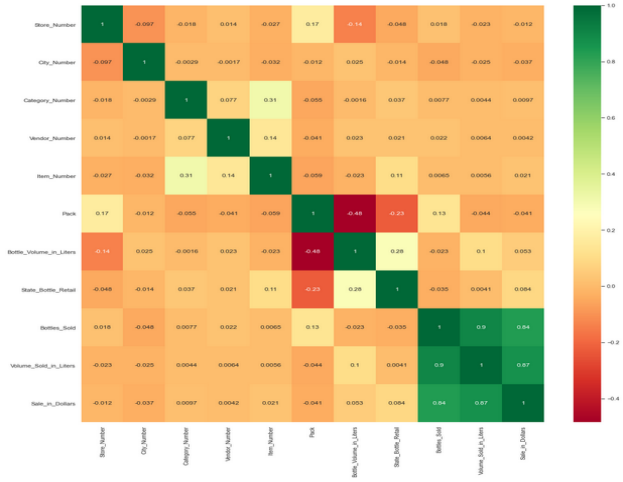

Figure 5: Correlation between independent variables and dependent variable

Next, we train the model using the RandomForestRegressor() function obtained from the sklearn module. Hyperparameter tuning is done before training the model with n_estimators = 500. The below figure displays the importance of each feature in training the model.

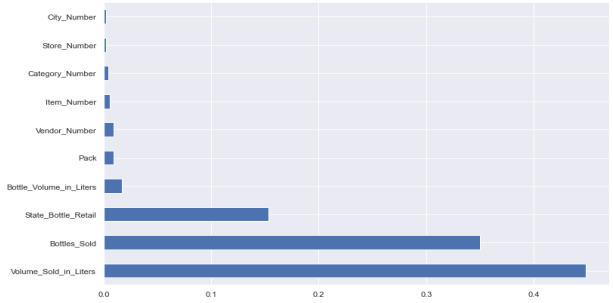


Figure 6: Feature Importance in training the ML model

## VI. RESULTS AND FINDINGS

After training, it is time to test the accuracy of the model. We obtained an accuracy of 96.83% upon testing the model on the test dataset. We also analyse the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values. The following table displays MAE, MSE, and RMSE values obtained from our machine learning model.

| Metric | Values Obtained |
|---|---|
| Mean Absolute Error (MAE) | 1.301 |
| Mean Squared Error (MSE) | 14523.591 |
| Root Mean Squared Error (RMSE) | 120.51 |

A general rule of thumb is that the lower the MAE and RMSE values are, the better is the

prediction capability of the machine learning model. From the above table, we have obtained very good results and the model is able to correctly predict liquor sales with an accuracy of 96.83% for the two cities.

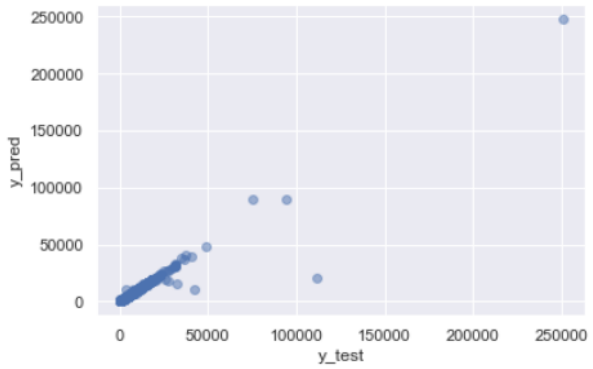


Figure 7: Scatter plot of test data and their predicted values

The above figure displays a comparison of the actual and predicted values of the test dataset.

## VII. CONCLUSIONS

Based on the liquor sales of the top two cities in Iowa - Des Moines and Cedar Rapids - between January 2017 and October 2021, it has been observed that the average liquor sales in Des Moines are higher than those in Cedar Rapids. Furthermore, sales are high during October, and sales are low during January in both cities. American Vodkas is the highest-selling liquor category in both cities. The vendor - Diageo Americas is responsible for the most number of liquor bottles sold in both cities. Finally, the analysis carried out on the stores showed that Hy-Vee #3 generated the highest revenue of 97.78 million dollars in Des Moines and Sam's Club 8162 generated the highest revenue of 35.70 million dollars in Cedar Rapids. In response to the analysis we performed in Figure 1, we carried out a hypothesis test to see if there is a significant decrease in the average sales between the months - June and July. Then, for predicting future liquor sales we trained our dataset with the random forest regressor model and we obtained an accuracy score of 99.7% Other metrics such as MAE, MSE, and RMSE was calculated to validate the performance of the model.

## VIII. REFERENCES

[1] M. Salmon, "towards data science," 3 April 2017. [Online]. Available: https://towards datascience.com/predictive-modeling-with-iowa-state-liquor-sales-data-e45342081b83.

[2] A. Palkar, M. Deshpande, S. Kalekar and S. Jaswal, "Demand Forecasting in Retail Industry for Liquor Consumption using LSTM," *IEEE,* 2020.

[3] G. T. G. U. Shreya Kohli, "Springer Link," 26 July 2020. [Online]. Available: https://link. springer.com/chapter/10.1007/978-981-15-5243-4_29.

[4] Bakshi C., "Random Forest Regression", 8 June 2020. [Online] Available: https://levelup .gitconnected.com/random-forest-regression-209c0f354c84.

[5] Acharya S., What are RMSE and MAE?. [Online] Available: https://towardsdatascience .com/what-are-rmse-and-mae-e405ce230383

[6] Salmon M., Predictive Modeling with Iowa State Liquor Sales Data. [Online] Available: https://towardsdatascience.com/predictive-modeling-with-iowa-state-liquor-sales-data-e45342081b83.

[7] CRAN, Analyzing Iowa Liquor Sales. [Online] Available: https://cran.r-project. org/web/packages/ialiquor/vignettes/b_analysis.html.

[8] CliffsNotes, Two-Sample z-test for Comparing Two Means. [Online] Available: https://www.cliffsnotes.com/study-guides/statistics/univariate-inferential-tests/two-sample-z-test-for-comparing-two-means.