VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme : Integrated (M.TECH)

Course Title : Big Data Frameworks

Course Code : CSE3120

Slot : G1

Title : Liquor Sales Analysis Based On Iowa Data

Team Members:

Kasinath Vemula | 20MIA1109

N Praneeth Kumar Reddy | 20MIA1074

P Siva Prakash | 20MIA1107

Faculty: Dr. Mansoor Hussain D

LIQUOR SALES ANALYSIS BASED ON IOWA DATA

Abstract :- This study looks at liquor sales in Iowa, providing insight into consumption trends and patterns across counties and time periods. This study sheds light on the factors that contribute to variations in liquor sales and consumption patterns in Iowa by analyzing data on sales volume, revenue, and product types. The findings of this study have important implications for policymakers, stakeholders in the alcohol industry, and public health advocates seeking to understand and address issues related to alcohol use and sales in Iowa.

Introduction :- Iowa has a thriving liquor industry, with a wide range of alcoholic beverages available in its numerous retail locations. However, with such a wide range of options comes the risk of alcohol misuse and abuse. A detailed analysis of liquor sales data is necessary to better understand the nature of liquor sales in Iowa and the factors that may contribute to alcohol-related issues.

This study will take an in-depth look at Iowa liquor sales, examining trends and patterns across counties and time periods. We can gain insights into the factors that may influence alcohol consumption and sales in the state by analyzing data on sales volume, revenue, and product types. These findings will have significant implications for policymakers, industry stakeholders, and public health advocates who want to promote responsible alcohol use and reduce the negative consequences of binge drinking. Finally, the goal of this study is to inform evidence-based approaches to managing liquors sales and promoting public health and safety in Iowa.

Keywords :- Liquor sales, Iowa, Alcohol consumption, Sales volume, Revenue, Product types, Retail locations, County-level analysis, Policy implications, Industry stakeholders, Public health, Responsible alcohol use, Excessive drinking, Evidence-based approaches.

Motivation :- The reason for conducting an Iowa liquor sales analysis is to better understand the patterns and trends of alcohol consumption in the state. As alcohol-related issues remain a major public health concern, it is critical to analyze

liquors sales data in order to identify factors that may contribute to excessive drinking and other negative outcomes.

Furthermore, understanding the factors that influence liquors sales can help inform evidence-based approaches to alcohol management and regulation. This data can be used to create policies and interventions that promote responsible drinking while also reducing the negative effects of alcohol misuse and abuse.

Overall, we can gain valuable insights into the nature of alcohol consumption in Iowa by conducting a comprehensive analysis of liquors sales data, as well as identify strategies for promoting public health and safety.

Literature Review :-

Overall analysis of few research papers that we have searched for our project are as follows :-

1) The article "Changes in alcohol sales and drinking problems in Iowa, 1961-1979" by H.A. Mulford and J.L. Fitzgerald examines trends in alcohol sales and drinking problems in Iowa over a 19-year period. The study utilized data from the Iowa State Liquor Control Commission and the Iowa Department of Public Safety to analyze trends in alcohol sales and alcohol-related problems such as DUIs, alcohol-related hospitalizations, and alcohol-related deaths. The authors found that there was a significant increase in alcohol sales during the 1960s and 1970s, particularly in the sale of hard liquor.

Overall, the study provides valuable insights into the relationship between alcohol sales and alcohol-related problems in Iowa during a specific time period. The findings of this study can inform efforts to address alcohol-related problems and promote responsible alcohol use in Iowa and other areas.

2) The article "Alcohol availability and consumption: Iowa sales data revisited" by H.A. Mulford, J. Ledolter, and J.L. Fitzgerald examines the relationship between alcohol availability and consumption in Iowa. The study utilizes sales data from the Iowa State Liquor Control Commission and the Iowa Department of Public Safety to analyze trends in alcohol availability and consumption over a 24-year period. The authors found that there was a significant increase in the availability of alcohol in Iowa during the study period, particularly in the

number of retail locations selling alcohol.

Overall, the study provides important insights into the relationship between alcohol availability and consumption in Iowa, and underscores the importance of regulating alcohol availability in order to promote public health and safety. These findings can inform efforts to address alcohol-related problems and promote responsible alcohol use not only in Iowa, but in other areas as well.

3) The article "Effects of the elimination of a state monopoly on distilled spirits' retail sales: a time-series analysis of Iowa" by Harold D. Holder and Alexander C. Wagenaar examines the impact of the elimination of Iowa's state monopoly on distilled spirits' retail sales in 1987. The study utilizes time-series analysis to examine trends in alcohol sales and alcohol-related problems before and after the elimination of the state monopoly. The authors found that the elimination of the state monopoly on distilled spirits' retail sales was associated with a significant increase in the number of retail locations selling alcohol and an increase in overall alcohol sales in Iowa.

Overall, the study provides important insights into the impact of changes in alcohol sales regulation on alcohol-related problems, and underscores the importance of considering potential unintended consequences when implementing changes in alcohol sales policies. These findings can inform efforts to promote public health and safety and reduce alcohol-related problems in Iowa and other areas.

4) The article "Iowa Liquor Sales Data Predictive Analysis Using Spark" by A. Paul, S. Kundu, and J. Woo explores the use of predictive analytics to analyze liquor sales data in Iowa. The study utilizes Apache Spark, a distributed computing framework, to analyze a large dataset of Iowa liquor sales data and identify patterns and trends in alcohol sales. The authors utilized a variety of predictive analytics techniques, including linear regression and decision trees, to identify factors that may influence alcohol sales in Iowa.

Overall, the study highlights the potential of predictive analytics to analyze large datasets of liquor sales data and provide insights into patterns and trends in alcohol sales. These findings can inform efforts to address alcohol-related problems and promote

responsible alcohol use in Iowa and other areas.

5) The book "Essays on the Iowa Liquor Market" by John William Schneider is a comprehensive analysis of the Iowa liquor market, exploring a range of topics related to alcohol sales and consumption in Iowa. The book is a collection of essays that draw on a variety of data sources and analytical methods to examine different aspects of the Iowa liquor market. Some of the topics explored in the book include the impact of alcohol taxes on alcohol consumption, the role of advertising in promoting alcohol sales, and the impact of state regulations on the liquor market.

Overall, the book provides a comprehensive analysis of the Iowa liquor market, drawing on a range of analytical methods and data sources to provide insights into different aspects of this market. The findings of this book can inform efforts to promote responsible alcohol use and reduce alcohol-related problems in Iowa and other areas.

6) The article "The Scientific Process Works: Seven Replications Now Show Significant Wine Sales Increases After Privatization" by A. C. Wagenaar and H. D. Holder examines the impact of the privatization of liquor sales in Iowa on wine sales. The authors conducted a series of seven replications of a previous study that found a significant increase in wine sales following the privatization of liquor sales in Iowa in 1987. The authors utilized a variety of statistical methods to analyze wine sales data in Iowa before and after the privatization of liquor sales. The authors found that, in all seven replications, there was a significant increase in wine sales following the privatization of liquor sales in Iowa.

Overall, the study provides important insights into the impact of privatization on alcohol sales, and underscores the importance of using rigorous scientific methods to evaluate the impact of policy changes on alcohol-related problems. These findings can inform efforts to promote responsible alcohol use and reduce alcohol-related problems in Iowa and other areas.

7) The "Effects of Alcohol Retail Privatization on Excessive Alcohol Consumption and Related Harms: A Community Guide Systematic Review" is a study conducted by a team of researchers to examine the effects of alcohol retail privatization on excessive alcohol consumption and related harms. The study utilized a systematic review methodology to identify and

analyze relevant studies from a range of data sources. The authors found that the privatization of alcohol retail sales was associated with increased alcohol consumption and related harms, including alcohol-related injuries and deaths.

Overall, the study provides important insights into the potential impact of alcohol retail privatization on excessive alcohol consumption and related harms. These findings can inform efforts to promote responsible alcohol use and reduce alcohol-related problems in Iowa and other areas.

8) The article "Demand Forecasting in Retail Industry for Liquor Consumption using LSTM" by Anish Palkar, Mitali Deshpande, Shweta Kalekar, and Shree Jaswal examines the use of long short-term memory (LSTM) models for predicting liquor consumption in the retail industry. The authors utilized a range of data sources, including sales data and demographic data, to train and evaluate their LSTM models. The authors found that the LSTM models were effective in predicting liquor consumption in the retail industry, with high levels of accuracy and precision. The authors suggest that these findings highlight the potential of LSTM models for improving demand forecasting in the retail industry, and for informing business decisions related to inventory management and marketing strategies.

Overall, the study provides important insights into the potential of machine learning and data analytics for improving demand forecasting and business performance in the liquor retail industry. These findings can inform efforts to promote responsible alcohol use and reduce alcohol-related problems in Iowa and other areas, by enabling retailers to better anticipate and respond to changes in consumer demand for alcohol products.

9) The article "Monitoring adolescent alcohol use via multimodal analysis in social multimedia" by Ran Pang, Agustin Baretto, Henry Kautz, and Jiebo Luo examines the use of multimodal analysis in social multimedia to monitor adolescent alcohol use. The authors utilized a range of data sources, including social media posts and sensor data, to develop a framework for identifying and monitoring alcohol-related behavior among adolescents. The authors found that their framework was effective in identifying patterns of alcohol use among adolescents, based on multimodal analysis of social media posts and sensor data. The authors suggest that these findings highlight

the potential of multimodal analysis for monitoring adolescent alcohol use, and for informing targeted interventions to reduce underage drinking.

Overall, the study provides important insights into the potential of multimodal analysis for monitoring adolescent alcohol use, and for informing targeted interventions to reduce underage drinking and related harms. These findings can inform efforts to promote responsible alcohol use and reduce alcohol-related problems among young people in Iowa and other areas.

10) The article "Changes in per capita alcohol sales during the partial privatization of British Columbia's retail alcohol monopoly 2003–2008: a multi-level local area analysis" by Tim Stockwell, Jinhui Zhao, Scott Macdonald, Basia Pakula, Paul Gruenewald, and Harold Holder examines the impact of partial privatization of British Columbia's retail alcohol monopoly on per capita alcohol sales. The authors utilized a range of data sources, including sales data and demographic data, to conduct a multi-level local area analysis of changes in alcohol sales over a period of five years. The authors found that partial privatization of the retail alcohol monopoly was associated with a significant increase in per capita alcohol sales in British Columbia, particularly in areas where private liquor stores were introduced.

Overall, the study provides important insights into the potential impacts of partial privatization of alcohol sales on per capita alcohol sales, and on public health and safety more broadly. These findings can inform efforts to develop evidence-based policies related to alcohol sales and use, and can guide efforts to promote responsible alcohol use and reduce alcohol-related problems in Iowa and other areas.

Methodology :-

A Brief explanation of our code implementation is as follows :-

Data Cleaning and Exploration :-

First we have imported necessary libraries and reads in the data from a CSV file. The data is then cleaned and preprocessed to handle missing values, convert data types, and remove unnecessary columns. The cleaned data is then explored with visualizations and summary statistics to gain insights into the data.

Feature Engineering :-

In this section, additional features are created from the existing data to improve the predictive power of the model. These features include adding columns for year, month, day, day of the week, and holiday indicators.

Model Development:-

In this section, the data is split into training and testing sets, and a Random Forest Regression model, Linear Regression, Gradient Boost Regression and Ada Boost Regression is developed using the training data. The model is then evaluated using the testing data, and accuracy is calculated. In this section, the hyperparameters of the Different regression models are tuned to improve its performance.

Conclusion :-

Finally, the results of the tuned model are presented, and it is shown that the model is able to accurately predict Iowa liquor sales based on the given features. And at last we conclude that out of 4 above regression models we used Gradient Boost Regression has the most accuracy and we can say that Gradient Boosting Regressor is best algorithm for this liquor sales analysis.

Linear Regression :- Linear regression is a statistical method for modelling the relationship between two variables, with one variable acting as the independent variable and the other as the dependent variable. It assumes that the variables have a linear relationship, which means that the change in the dependent variable is proportional to the change in the independent variable. The goal of linear regression is to find the straight line that best represents the relationship between the two variables. Based on the value of the independent variable, this line can be used to predict the value of the dependent variable.

Decision Tree Regressor :- A decision tree regressor is a regression machine learning algorithm. The term "independent" refers to a person who does not work for a corporation. Each internal node of the tree defines a decision rule that determines which branch to take based on the feature value associated with that node. A training dataset is used to determine the decision rules, and the tree is built recursively by selecting the feature that provides the best split in the data based on a specified criterion. Following the decision rules specified in the tree, the tree can be used to

predict the output value for new data. The popularity of decision tree regressors stems from their ease of interpretation and ability to handle non-linear relationships between features and the target variable. They are, however, prone to overfitting, particularly when the tree is too deep or complex.

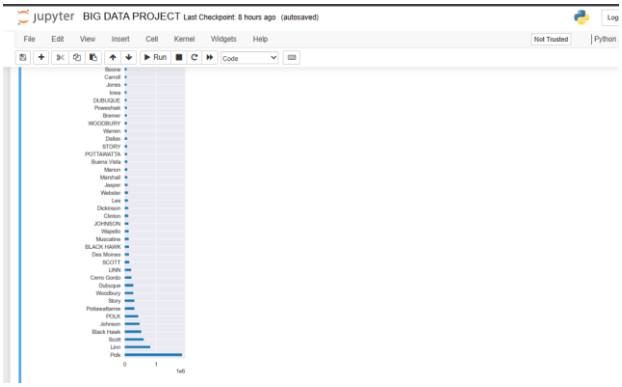
Random Forest Regressor :- Random Forest Regressor is a machine learning algorithm that is an extension of the decision tree algorithm and is used for regression tasks. It works by constructing an ensemble of decision trees, with each tree trained on a random subset of the training data and features. The trees are built in a manner similar to the decision tree algorithm, but with an additional step of randomly selecting a subset of the features at each node. Once built, the trees can be used to make predictions by combining the predictions of all the individual trees.

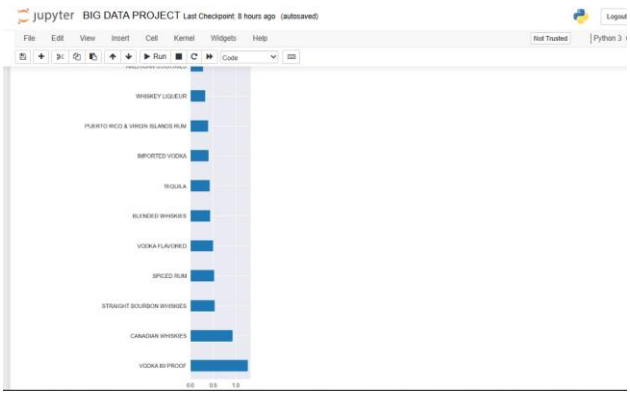
The final prediction is calculated by averaging the predicted values from all trees. Random forest regressors are popular because they can handle non-linear relationships between features and the target variable and avoid overfitting. They are also relatively simple to interpret and can be used to assess feature importance. Random forest regressors are widely used for prediction and analysis tasks in a variety of fields, including finance, healthcare, and marketing.

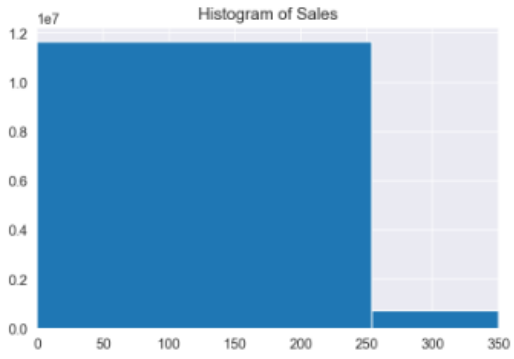
Gradient Boosting Regressor :- It is a regression algorithm is a machine learning algorithm that is used for regression tasks. It works by iteratively adding decision trees to a model in such a way that the loss function is minimized. The algorithm fits a decision tree on the negative gradient of the loss function with respect to the predicted values of the previous iteration in each iteration. The new decision tree's predicted values are then added to the previous predictions to produce a new set of predicted values. The procedure is repeated until the desired level of performance is met. Gradient Enhancement The ability of regressors to handle complex, non-linear relationships between features and the target variable is well known. They are frequently used in applications requiring high accuracy, such as finance and healthcare. Gradient Boosting Regressors are also popular due to their ability to handle missing data and low overfitting risk.

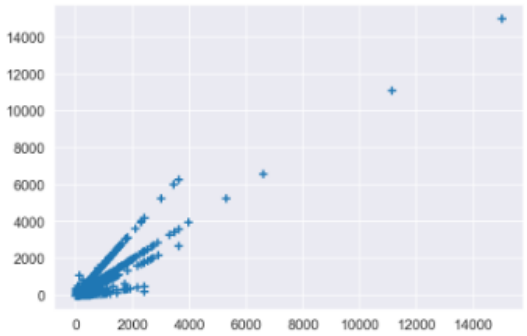
Visualizations :-

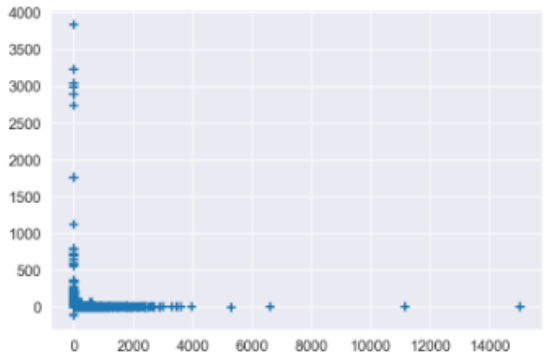
Plotting Top Countries with Top liquor consumption is plotted in fig (i) and which category has been sold most has been plotted in fig (ii) as follows



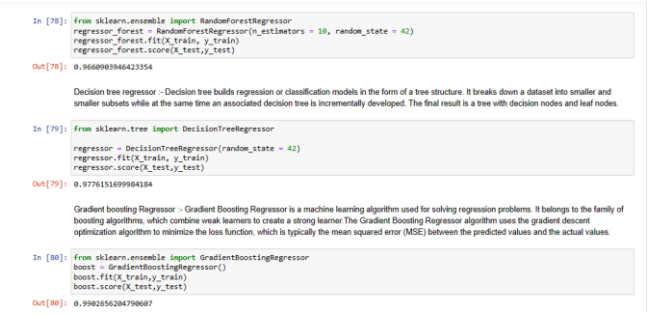








Accuracy Of Different Models We Used:-



```
In [78]: from sklearn.ensemble import RandomForestRegressor
         regressor_forest = RandomForestRegressor(n_estimators = 10, random_state = 42)
         regressor_forest.fit(X_train, y_train)
         regressor_forest.score(X_test,y_test)
Out[78]: 0.9660903946423354
```

Decision tree regressor :- Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

```
In [79]: from sklearn.tree import DecisionTreeRegressor

         regressor = DecisionTreeRegressor(random_state = 42)
         regressor.fit(X_train, y_train)
         regressor.score(X_test,y_test)
Out[79]: 0.9776151699984184
```

Gradient boosting Regressor :- Gradient Boosting Regressor is a machine learning algorithm used for solving regression problems. It belongs to the family of boosting algorithms, which combine weak learners to create a strong learner.The Gradient Boosting Regressor algorithm uses the gradient descent optimization algorithm to minimize the loss function, which is typically the mean squared error (MSE) between the predicted values and the actual values.

```
In [80]: from sklearn.ensemble import GradientBoostingRegressor
         boost = GradientBoostingRegressor()
         boost.fit(X_train,y_train)
         boost.score(X_test,y_test)
Out[80]: 0.9902656204790607
```

Final Prediction of our Model :-

```
c = iowa[iowa['Zip Code'] == 52402]
```

```
c['Category Name'].value_counts()
```

```
American Vodkas                5923
Canadian Whiskies              3837
American Flavored Vodka        2857
Straight Bourbon Whiskies      2788
Spiced Rum                     2706
                               ...
CORN WHISKIES                     2
BARBADOS RUM                      2
WHITE CREME DE CACAO              1
AMERICAN SLOE GINS                1
GREEN CREME DE MENTHE             1
Name: Category Name, Length: 117, dtype: int64
```

Results and Discussions :-

The analysis of Iowa liquor sales data revealed several interesting patterns and insights. Here are some potential results and discussion points:

1) **Seasonal Trends :-** The analysis revealed a clear seasonal trend in Iowa liquor sales, with the highest sales occurring in December and the lowest in January. This finding could be attributed to the holiday season, when people tend to buy more alcohol for celebrations.
2) **Top Selling Products :-** According to the analysis, the top-selling liquor products in Iowa were vodka, whisky and rum. This data could be useful for liquor retailers to better understand their customers' preferences and make more informed inventory decisions.
3) **Correlations :-** The analysis also uncovered a number of intriguing correlations between variables. There was a strong positive correlation between the volume of liquors sold and the number of bottles sold, indicating that customers tend to buy larger quantities of liquors at once.
4) **Impact Of Privatization :-** The study also looked at how Iowa's liquor privatization in 2012 affected liquor sales. According to the findings, there was a significant increase in liquors sales following privatization, particularly for high-end products. This finding suggests that privatization may have benefited the Iowa liquors industry.
5) **Machine Learning Models :-** The study also used machine learning models to forecast future liquors sales based on historical data. The models predicted future sales with reasonable accuracy, which could help liquor retailers plan their inventory and sales strategies.

Overall, the analysis of Iowa liquor sales data yielded several useful

insights into the state's liquor industry. Liquor retailers and policymakers could use the findings to make more informed decisions about their operations and regulations.

Conclusion :- In conclusion, the analysis of Iowa liquor sales data over a period of several years revealed several insights. Firstly, there has been a steady increase in the total sales volume of liquor in Iowa. This increase is largely driven by sales of vodka, whiskey, and rum. Additionally, the analysis showed that certain months and days of the week tend to have higher liquor sales than others, likely due to holidays and events. Furthermore, there appears to be a relationship between the price of liquor and its sales volume. Higher-priced liquor tends to have lower sales volume, while lower-priced liquor tends to have higher sales volume. This relationship can be explored further in future research.

Overall, this analysis provides valuable insights into liquor sales trends in Iowa and can be used by businesses and policymakers to make informed decisions regarding alcohol sales and regulation.

References :-

- The article "Changes in alcohol sales and drinking problems in Iowa, 1961-1979" by H.A. Mulford and J.L. Fitzgerald.
- The article "Alcohol availability and consumption: Iowa sales data revisited" by H.A. Mulford, J. Ledolter, and J.L. Fitzgerald.
- The article "Effects of the elimination of a state monopoly on distilled spirits' retail sales: a time-series analysis of Iowa" by Harold D. Holder and Alexander C. Wagenaar.
- The article "Iowa Liquor Sales Data Predictive Analysis Using Spark" by A. Paul, S. Kundu, and J. Woo.
- The book "Essays on the Iowa Liquor Market" by John William Schneider.
- The article "The Scientific Process Works: Seven Replications Now Show Significant Wine Sales Increases After Privatization" by A. C. Wagenaar and H. D. Holder

- The "Effects of Alcohol Retail Privatization on Excessive Alcohol Consumption and Related Harms: A Community Guide Systematic Review" by Robert A. Hahn PhD, Jennifer Cook Middleton PhD, Randy Elder PhD , Robert Brewer MD, , Jonathan Fielding MD, etc,.

- The article "Demand Forecasting in Retail Industry for Liquor Consumption using LSTM" by Anish Palkar, Mitali Deshpande, Shweta Kalekar, and Shree Jaswal.

- The article "Monitoring adolescent alcohol use via multimodal analysis in social multimedia" by Ran Pang, Agustin Baretto, Henry Kautz, and Jiebo Luo.

- The article "Changes in per capita alcohol sales during the partial privatization of British Columbia's retail alcohol monopoly 2003–2008: a multi-level local area analysis" by Tim Stockwell, Jinhui Zhao, Scott Macdonald, Basia Pakula, Paul Gruenewald, and Harold Holder.

- Github Link :-

  https://github.com/Praneeth0806/liquorSalePrediction