Optimizing Craft Beer Production and Sales: Analyzing Factors Affecting Quality ,Efficiency, and Alcohol Content

By

Charan Kumar Pathakamuri,

Srinivasan Lakkala,

Sherin Feno Kumaresan

Data Overview

- Data Source: The data set was taken from Kaggle. Size of the dataset is 2.4 GB (9545009 X 20).
- Link: Brewery Operations and Market Analysis Dataset
- Columns: Batch ID, Brew Date, Beer Style, SKU, Location, Fermentation Time, Temperature, PH Level, Gravity, Alcohol Content, Bitterness, Color, Ingredient Ratio, Volume Produced, Total Sales, Quality Score, Brewhouse Efficiency, Loss During Brewing , Loss During Fermentation, Loss During Bottling Kegging.

Introduction

- This Dataset has Various Potential Analysis.
- We have united different splits (10) of Data into single Data Frame.
- Data Cleaning
- Feature Engineering: We have derived some columns which are useful for Data Analysis.
- Machine Learning: Model Training and evaluating
- Drawing Insights.

Problem Statement

Alcohol trends and Market requirements change Rapidly.

Addressing the Efficiency of Brewery.

Finding the different losses and which is most effecting.

Finding the most efficient beer style.

PROJECT EXECUTION OVERVIEW

Data Collection and Cleaning

- Process: Gathered NIST data on environmental metrics and solar panel performance.
- Cleaning: Addressed missing values, outliers, and inconsistencies to ensure data quality and reliability.

Exploratory Data Analysis (EDA)

- Objective: Analyzed the data to uncover patterns, trends, and relationships that inform the model's development.
- Tools Used: Utilized statistical tools and visualization libraries to examine data distributions and correlations

.Machine Learning

- Approach: Developed predictive models to estimate optimal solar panel placement based on historical data.
- Regression MOdel: Chosen for their robustness and effectiveness in handling varied and complex data structures.

Pipeline Technique

- Implementation: Created a streamlined process using pipeline techniques to automate the flow of data through various preprocessing and modeling steps.
- Benefits: Ensured consistency in data handling and model application, enhancing reproducibility and efficiency.

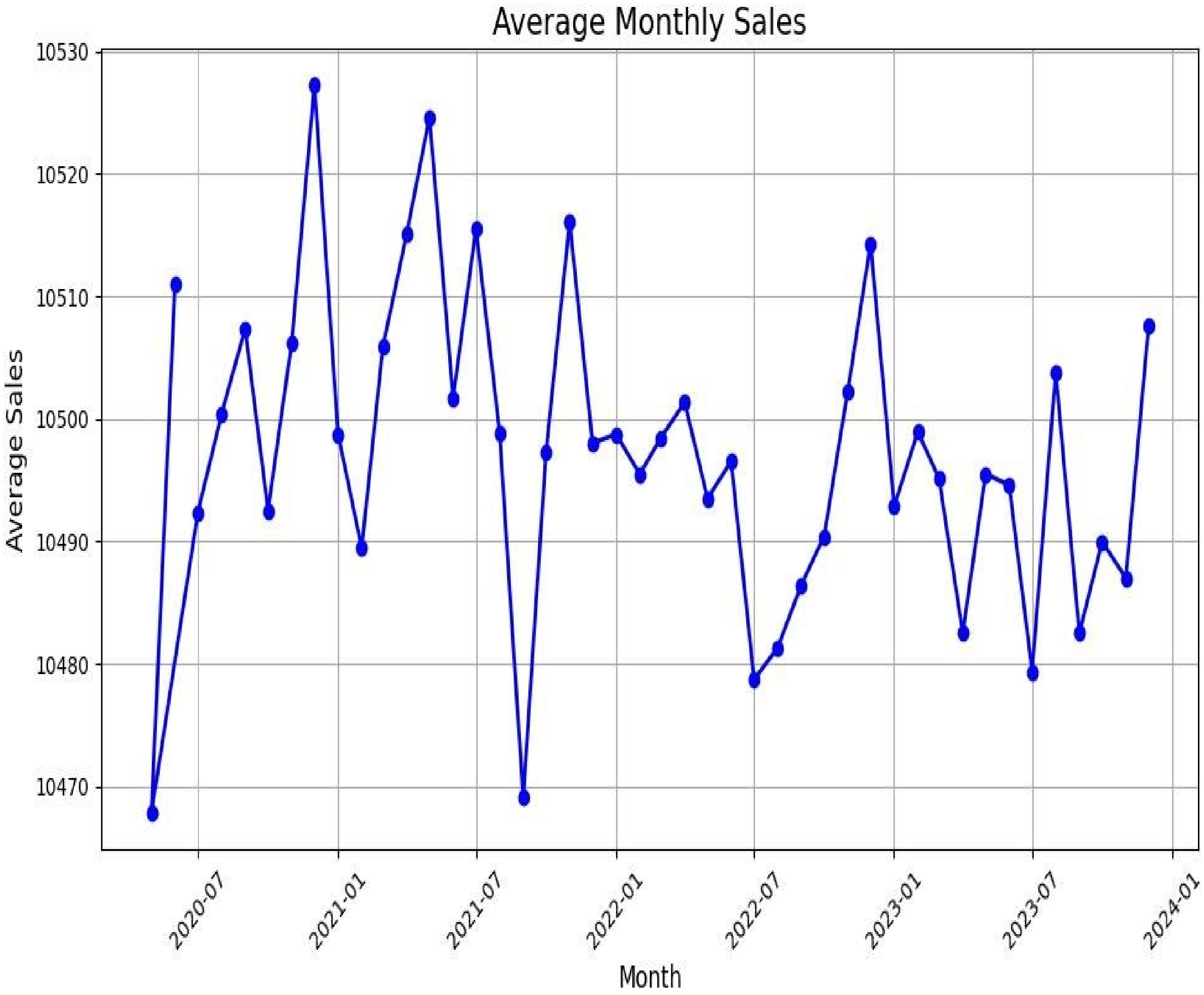
Objective

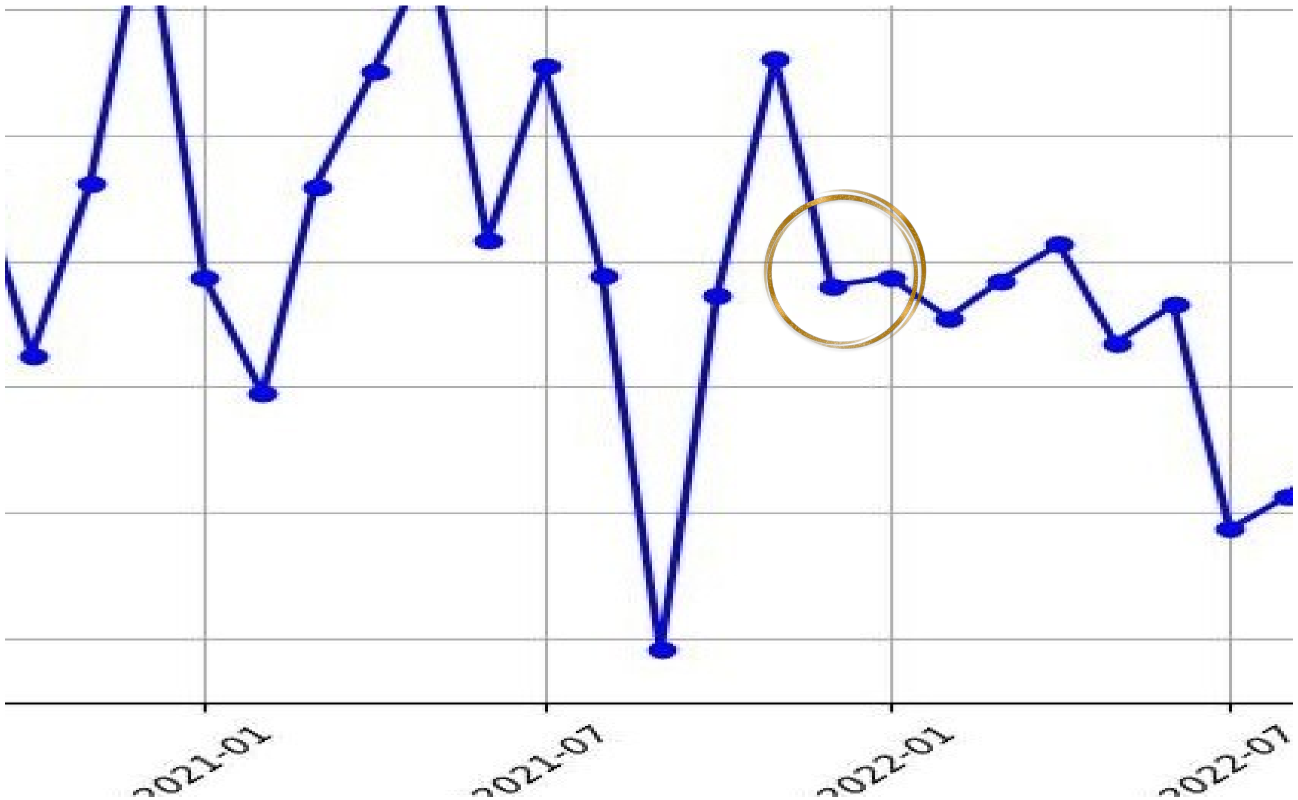
- The objective of this project is to uncover actionable insights to improve the brewing process, optimize resource utilization, enhance product quality, and maximize profitability.
- To build predictive model for estimating Quality.
- Giving ideas to Manufacturers.

Logistic Regression Model

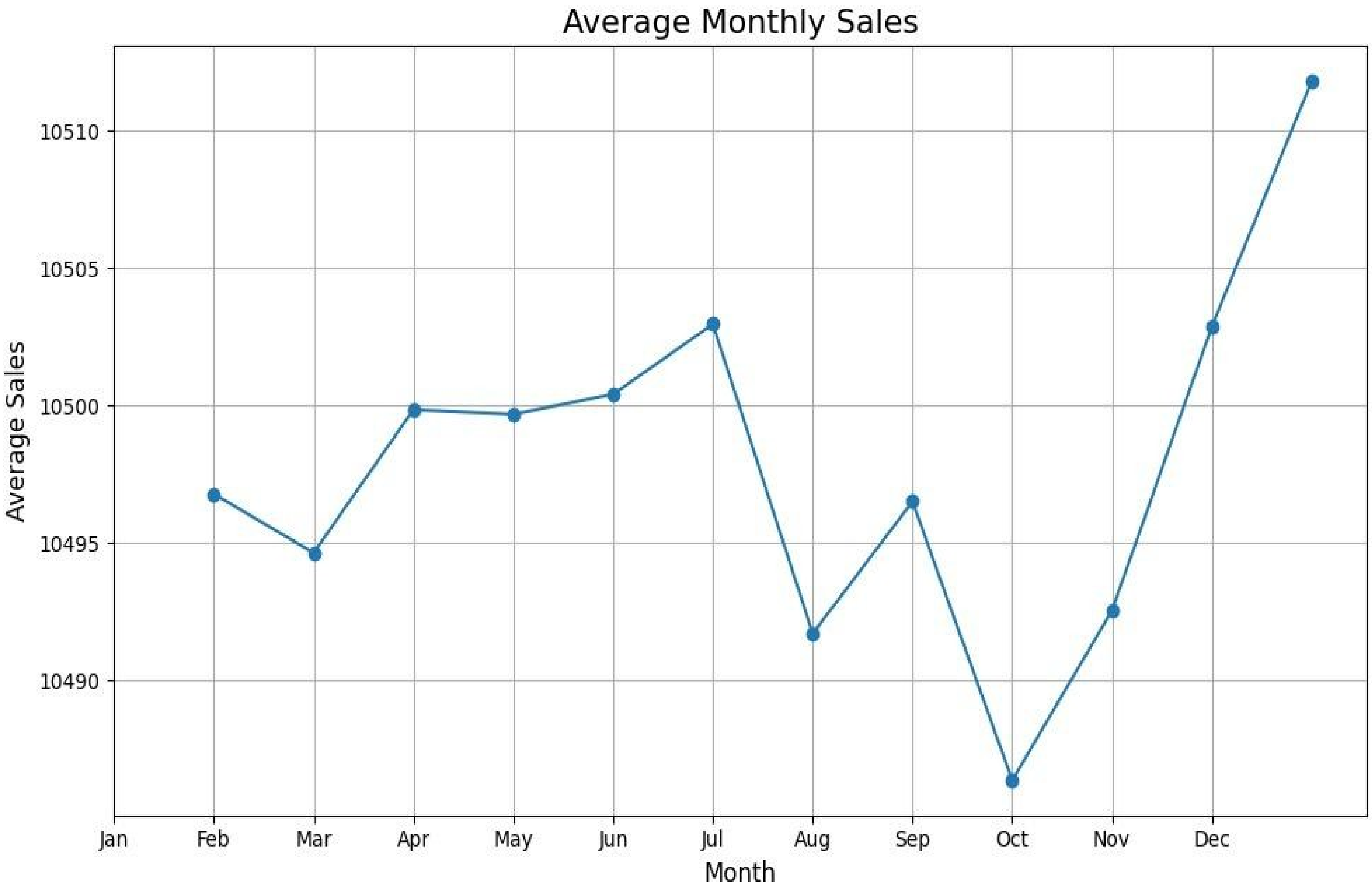
- Accuracy: 0.5000787857927920
- Root Mean Squared Error (RMSE) on test data = 0.4240858211120456

Insights: Market Analysis and Trends

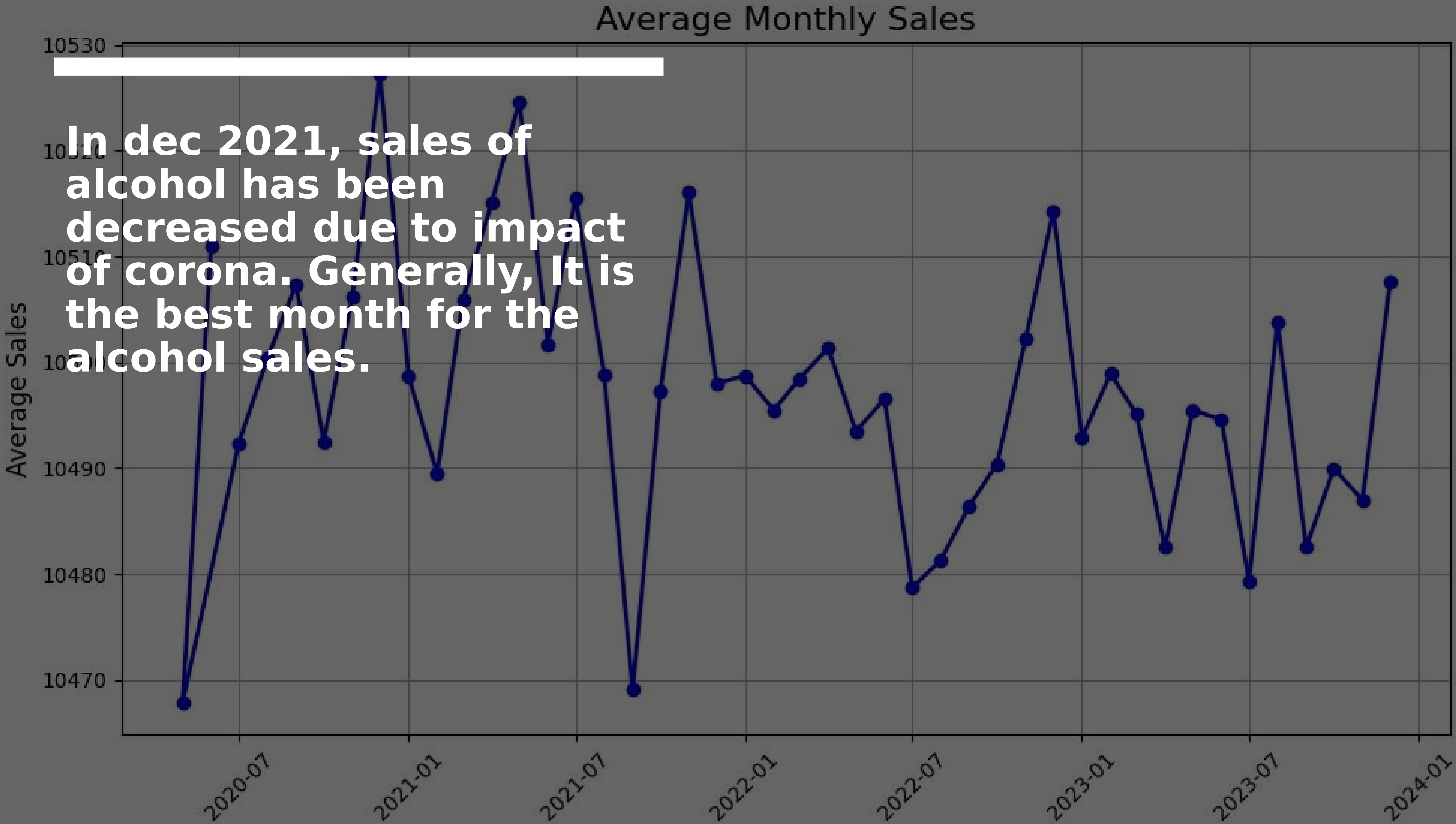

Average Monthly Sales
10530
10520
10510
10500
10490
10480
10470
Average Sales
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
Month


2021-01
2021-07
2022-01
2022-07

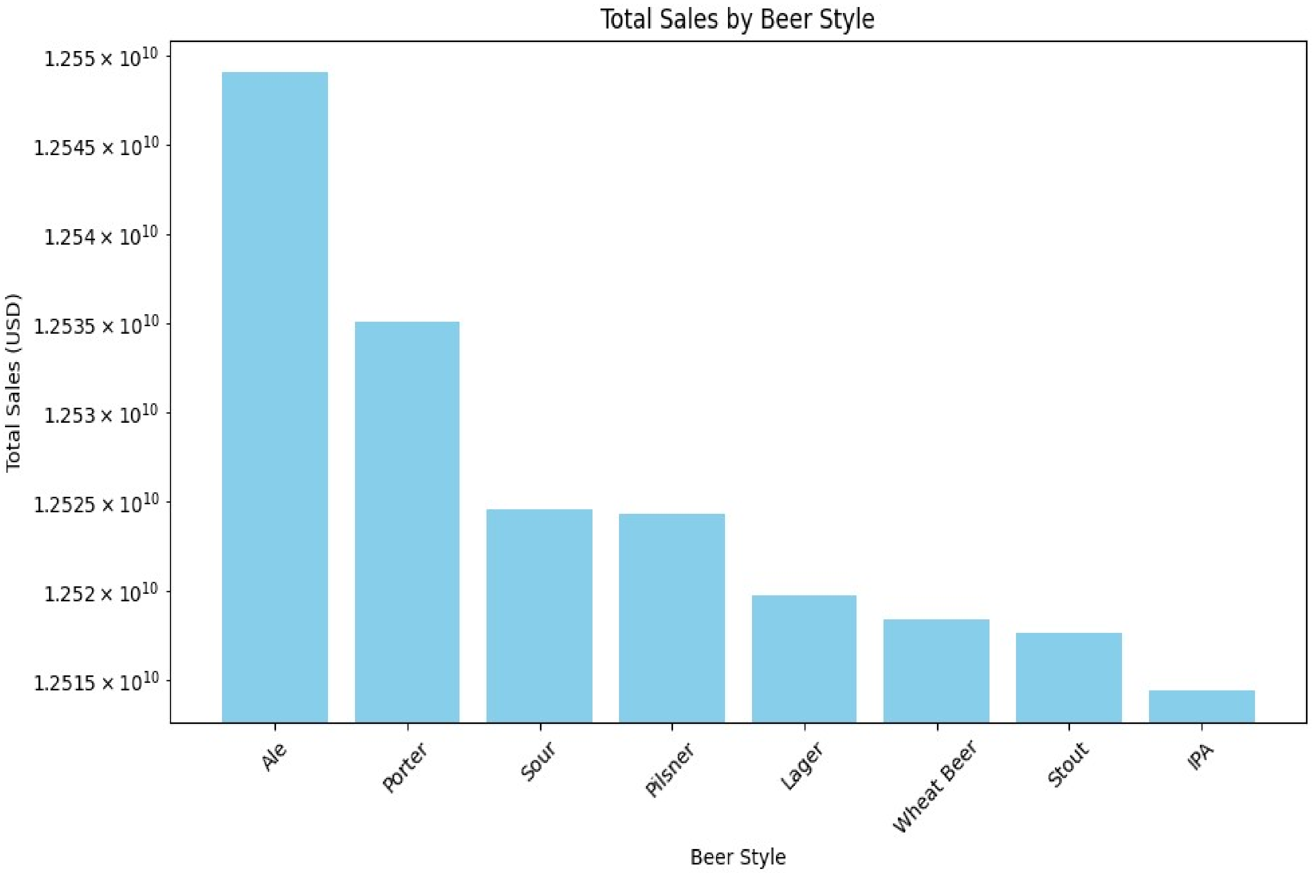
[redacted]


Average Monthly Sales
Average Sales
10510
10505
10500
10495
10490
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

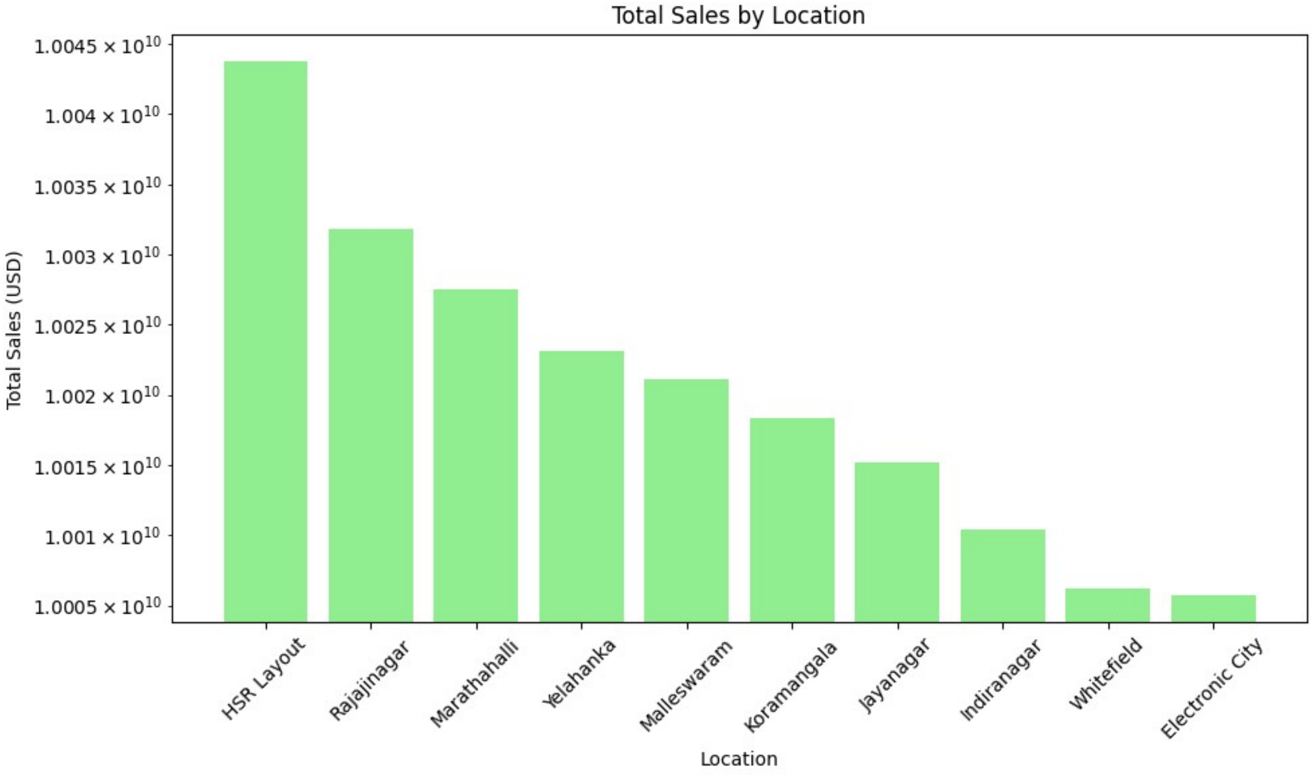
In dec 2021, sales of alcohol has been decreased due to impact of corona. Generally, It is the best month for the alcohol sales.



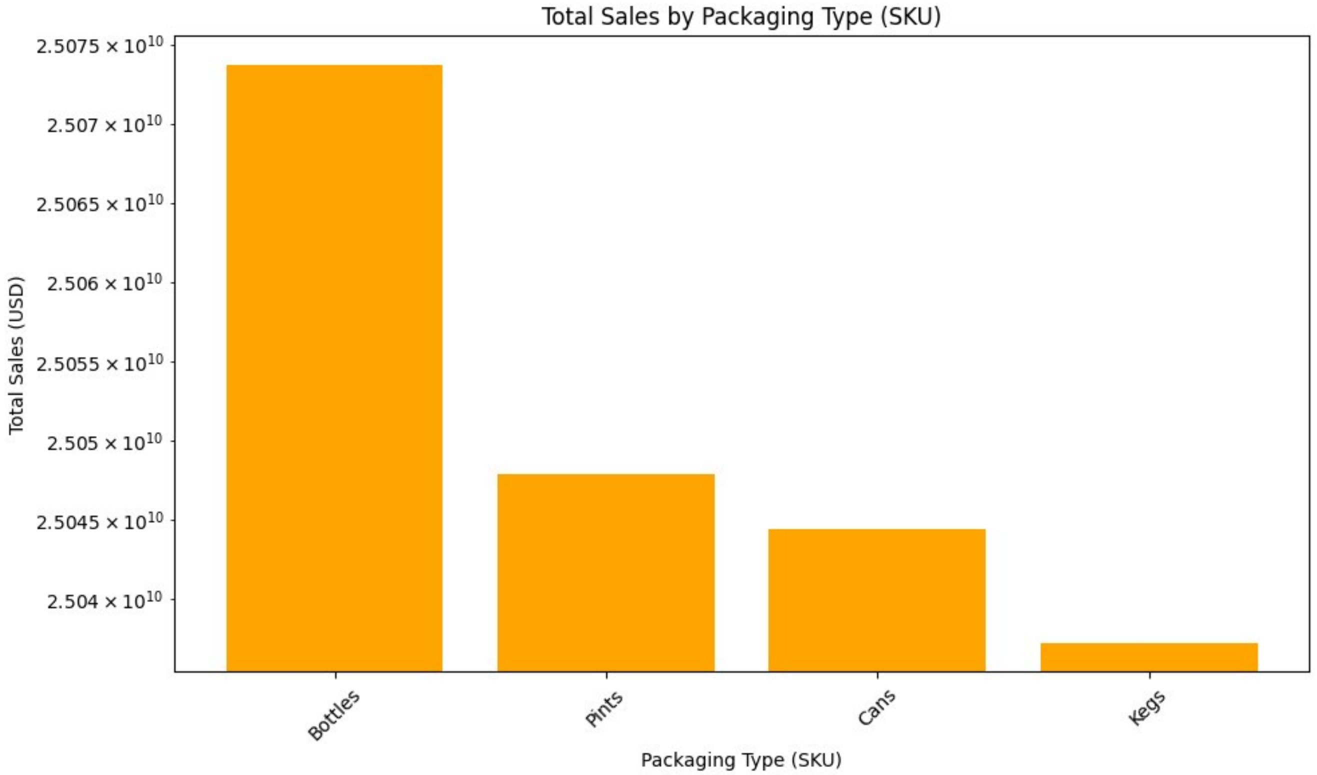
Trend Analysis: Ale is the most sold beer style.



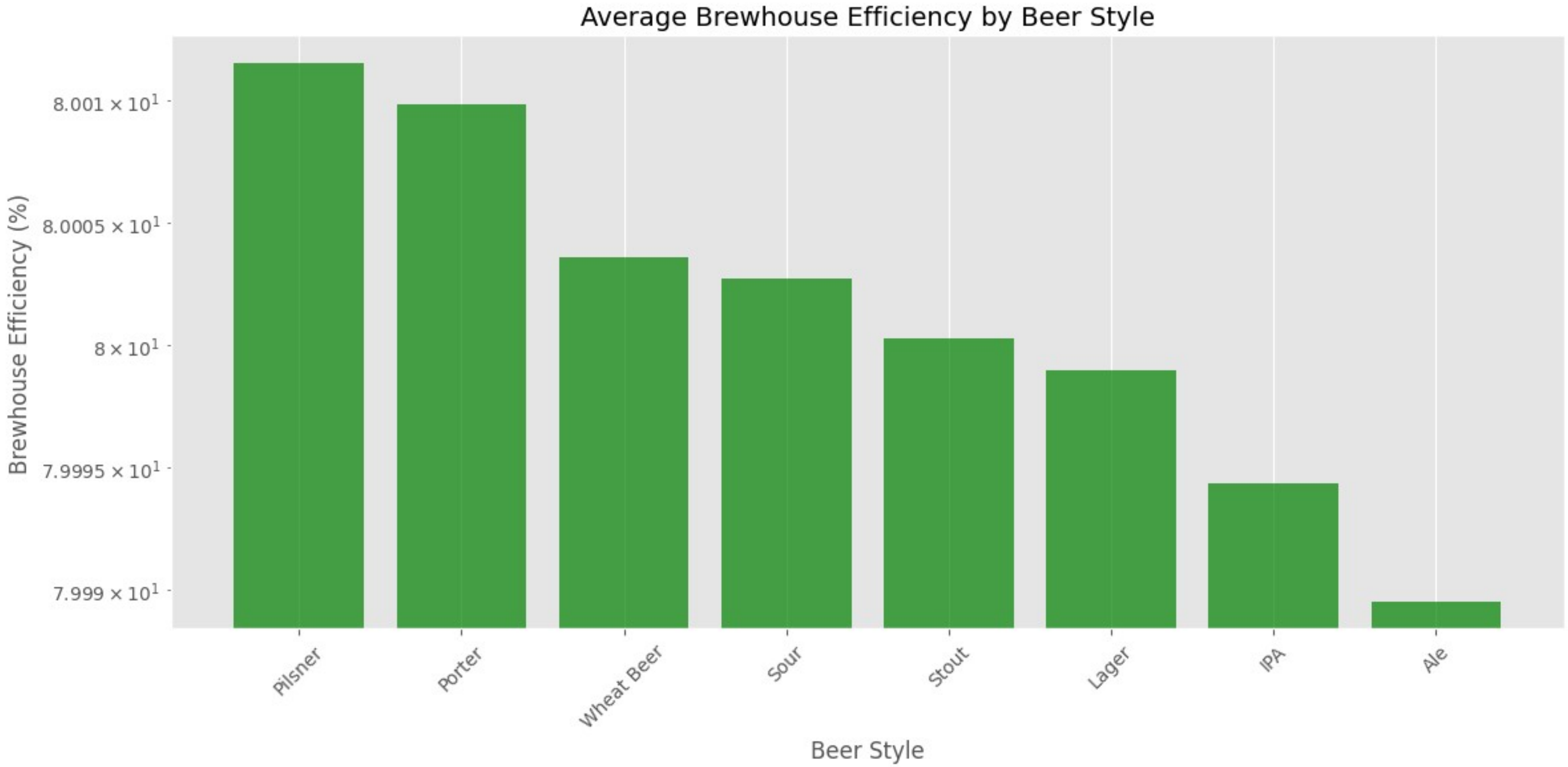
Total sales by location



Total Sales by Packaging Type



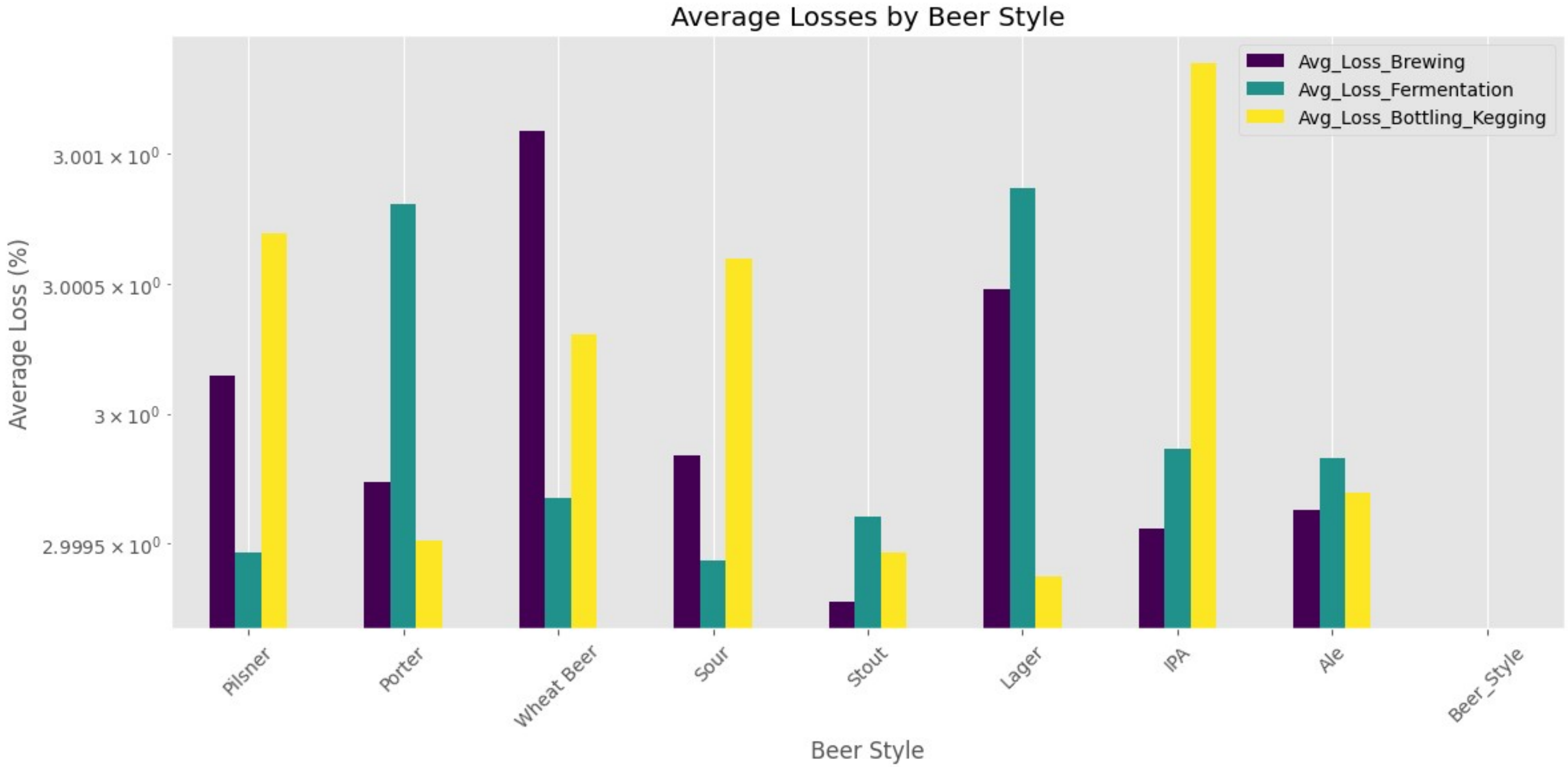
Average Brew House efficiency by Beer Style



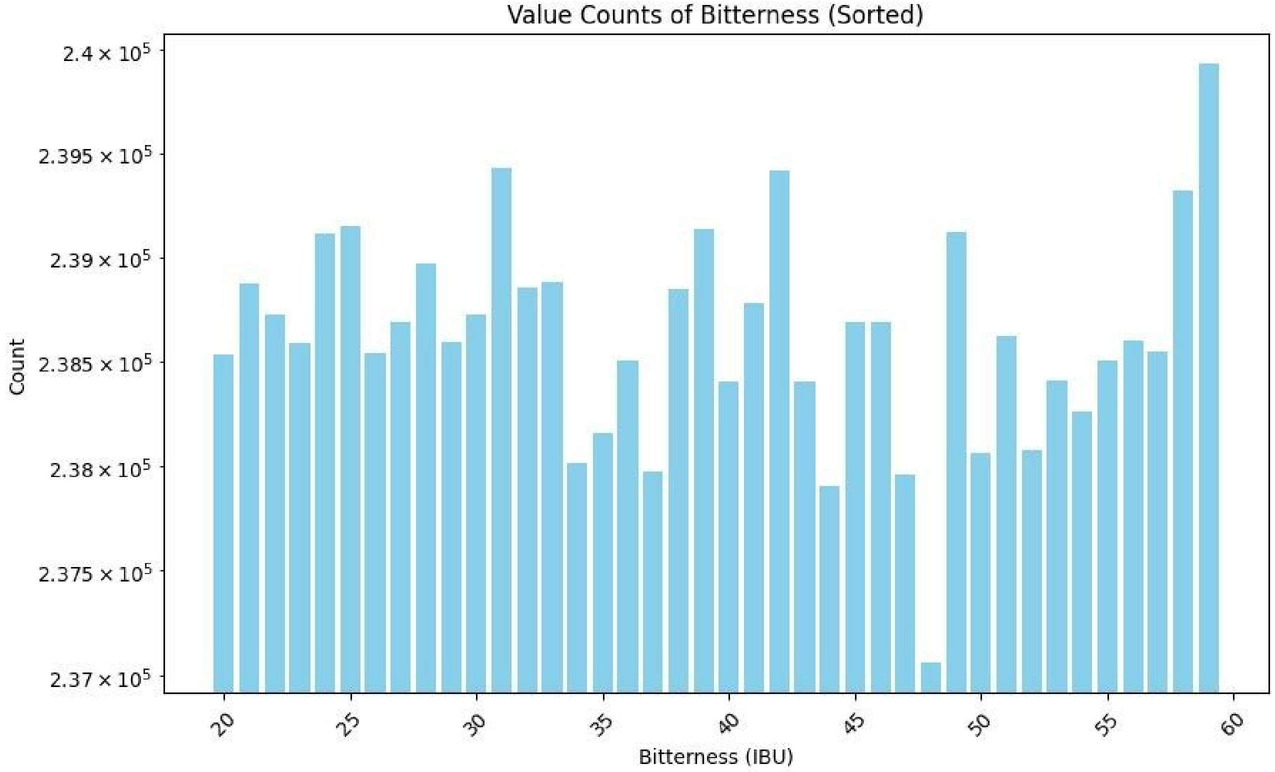
Idea 1

- Ale is the least efficient beer style, but it is most sold beer style so, manufacturers need to come with new style of manufacturing methods to increase the efficiency

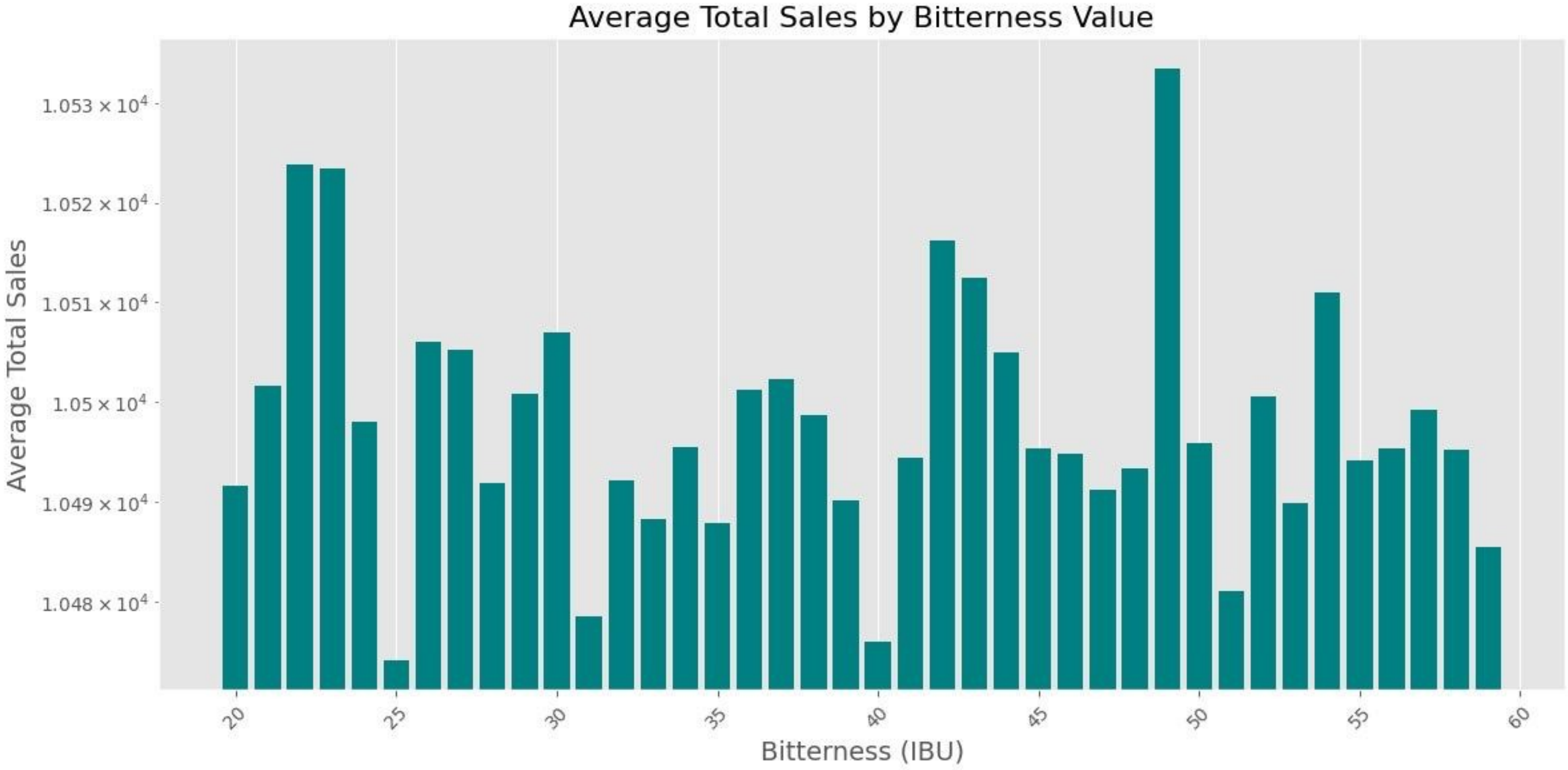
Average kinds of losses by Beer Style



Distribution of beer production based on bitterness



Average Sales by Bitterness



Idea 2:

- Manufacturing of level 60 bitterness is very high, but sales of sales are not about to that point.
- Producers can increase the level 40 bitterness since it has sales.

Conclusion

Conclusion

- The brewing industry thrives on efficiency, precision, and innovation. Through our analysis of the extensive dataset, we uncovered key insights that can significantly impact both the quality of the product and the profitability of the operations.
- By leveraging Apache Spark on Databricks, we achieved scalability, speed, and a seamless collaborative environment for big data analytics and machine learning.