**CAPSTONE INTERIM REPORT**

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT Chennai July 22 |
| Team members | 1. SWETHA R VE 2. NARASHIMMHAN 3. MUTHURAM PANDIAN 4. NANDAKUMAR A K 5. JEEVA ANAND B |
| Domain of Project | SALES ANALYSIS |
| Proposed project title | Unlocking Sales Potential in Lowa Liquor through Data Analytics. |
| Group Number | Team 04 |
| Team Leader | SWETHA R VE |
| Mentor Name | Pratik Sonar |

**Date:09/06/2023**

**Signature of the Mentor Signature of the Team Leader**

## TABLE OF CONTENT

|  |  |
| --- | --- |
| **TOPIC** | **PAGE NO** |
| **1.BUSINESS UNDERSTANDING** | **3** |
| 1.1 BUSINESS PROBLEM STATEMENT | 3 |
| 1.2 TOPIC SURVEY | 3 |
| 1.3 CRITICAL ASSESSMENT OF TOPIC SURVEY | 4 |
| **2.DATA UNDERSTANDING** | **4** |
| 2.1 DATA DICTIONARY | 6 |
| 2.2 VARIABLE CATEGORIZATION | 7 |
| 2.3 NULL VALUE TREATMENT | 8 |
| 2.4 DISTRIBUTION OF VARIABLES | 9 |
| 2.5 CORRELATION BETWEEN THE VARIABLES | 10 |
| **3. EXPLORE DATA ANALYSIS** | **11** |
| **4. DATA PRE-PROCESSING FOR MODEL BUILDING** | 13 |
| **5. BASE MODEL** | 16 |
| **6. FITTING MULTIPLE REGRESSION MODEL** | 18 |
| **7. CONCLUSION** | 19 |

# **BUSINESS UNDERSTANDING:**

As we know that the liquor sales is one of the worlds biggest business market all around the world where it can generate a large impact on the revenue .

Revenue in the Alcoholic Drinks market amounts to US$1,609.00bn in recent years. The market is expected to grow annually by 5.42% . In global comparison, most revenue is generated in China

In relation to total population figures, per person revenues of US$209.40 are generated.

So that as per the studies and data people all around the world use different types of liquor and each country is getting taxes and other benefits through the sales. There are different regulations in this industry as we know that consumption of alcohol is injuries to the health and people may get addicted to this habit.Keeping all these factors we can analyse the Lowa Liquor sales and different market studies and how to increase the sales by proper marketing and personalised advertisements.

## BUSINESS PROBLEM STATEMENT:

Lowa Liquor is a retail store that specializes in selling various types of alcoholic beverages. The store has been facing a decline in sales over the past year, and the management team is concerned about the reasons behind this decline. The store wants to identify the factors that are contributing to the decline in sales and find ways to improve the sales performance

**Business Objective:**

The objective of the business is to identify the factors that are causing the decline in sales and develop strategies to increase sales revenue. The business wants to analyze sales data and customer behavior to identify patterns and trends that can help them make informed decisions about how to improve their business operations

## TOPIC SURVEY :

* + 1. **Problem understanding:**

The problem is that Lowa Liquor, a retail store specializing in selling alcoholic beverages, has experienced a decline in sales over the past year. The management team is concerned about the reasons behind this decline and wants to identify the factors contributing to it.

## Current solution to the problem:

There is currently no specific solution in place to address the decline in sales at Lowa Liquor. The store may be implementing general strategies such as 5 | P a g e marketing and promotion campaigns, but there is no evidence that these strategies are effective.

## Proposed solution to the problem:

The proposed solution is to use data analysis and machine learning techniques to identify the factors contributing to the decline in sales and develop strategies to improve sales revenue. This may involve analyzing sales data and customer behavior, identifying patterns and trends, and using this information to make datadriven decisions about pricing, product mix, promotions, and inventory management.

# DATA UNDERSTANDING:

## DATA DICTIONARY:

|  |  |  |
| --- | --- | --- |
| **S.No** | **Feature Name** | **Feature Description** |
| **1.** | Invoice and item number | Invoice number for the purchased product |
| **2.** | Date | Date of the product purchase |
| **3.** | Store number | Product sold store number |
| **4.** | Store name | Product sold store name |
| **5.** | Address | Product sold store Address |
| **6.** | City | Product sold store city |
| **7.** | Zip code | Product sold store zip code |
| **8.** | Store location | Product sold store location |
| **9.** | County number | Product sold country number |
| **10.** | County | Product sold country number |
| **11.** | Category | Category number of Product sold |
| **12.** | Category name | Category name of Product sold |
| **13.** | Vendor number | Vendor number for the product distributed to the stores |
| **14.** | Vendor name | Vendor name for the product |
| **15.** | Item number | item number for the product |
| **16.** | Item description | Description of the item sold |
| **17.** | Pack | Number of bottles in a pack |
| **18.** | Bottle volume (ml) | Quantity per bottle |
| **19.** | State bottle cost | Cost of the bottle state wise (whole sale) |
| **20.** | State bottle retail | Cost of the bottle retail |
| **21.** | Bottle sold | Number bottle bought |
|  | Sales in dollar | Price in dollar |
|  | Volume sold in litres | Quantity sold in liters |
|  | Volume sold in gallons | Quantity sold in gallons |

## VARIABLE CATEGORIZATION :

### Independent variables:

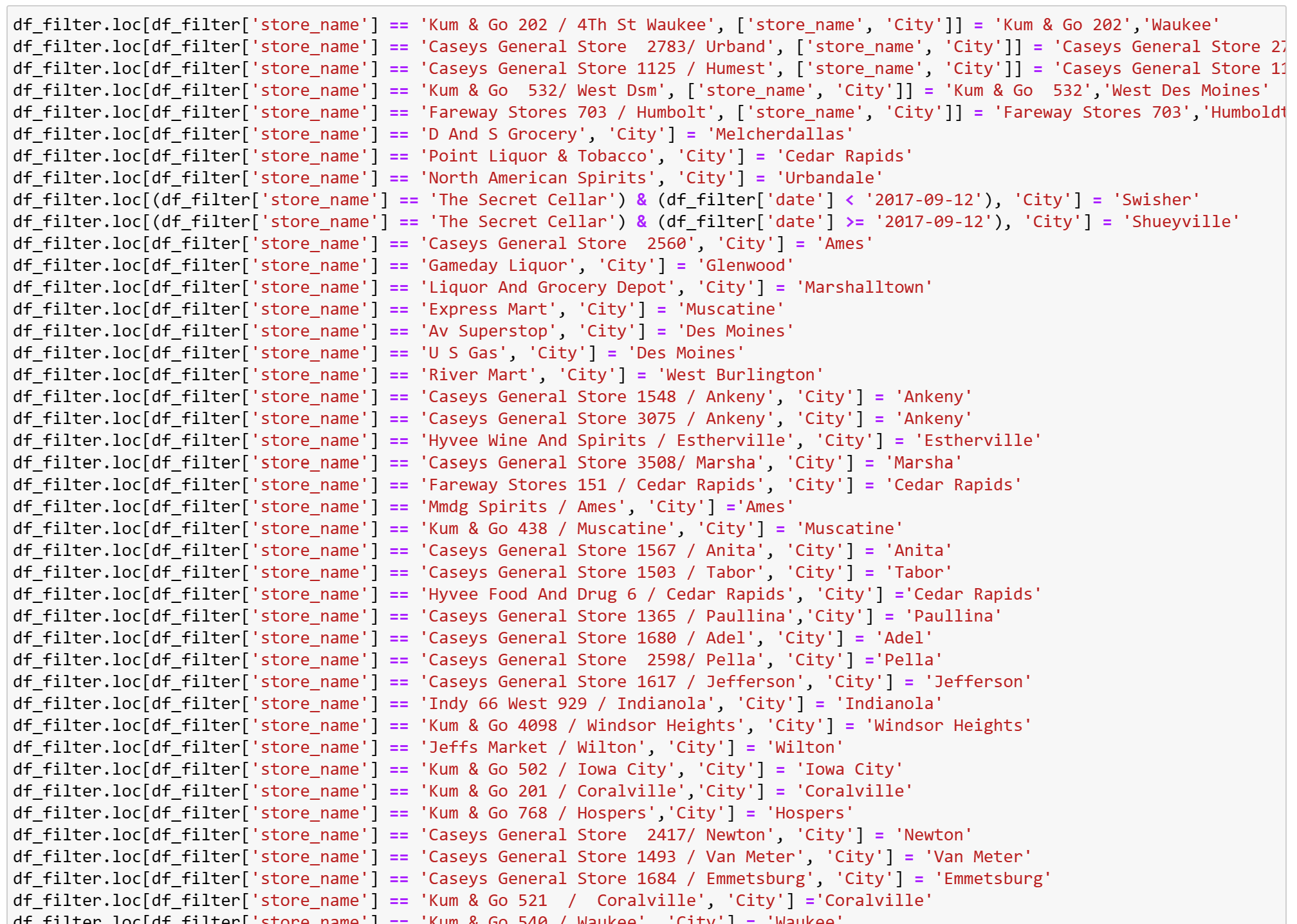
Numerical column: 14

Categorical column: 10

### Target variable:

Quantity sold in litres : Numerical

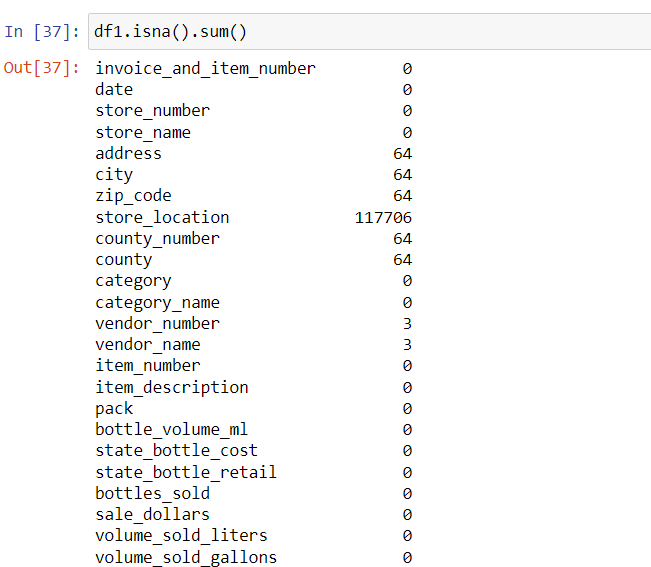
# 3.DATA PREPROCESSING:





## NULL VALUE TREATMENT:

Null value treatment is essential to building most of the commonly used machine learning classification models such as logistic regression, decision tree, KNN, and others. To infer that we have used isnull() function the null values from the dataset.



From the above figure, it is evident that the maximum of missing value is **117706** which is observed only in store location column. Since we have store address ,city name and zip code we will be dropping the column store location.

Missing values in columns **address,city ,zip code,**county number and county were represented as null . We had replaced it with NaN for the ease of processing.

## DISTRIBUTION OF VARIABLES:

The Lowa Liquor dataset which we had selected have 1048575 rows and 24 columns. The data consists of Numerical and Categorical data. While further analyzing the data we find that there is 14 numerical data and 10 categorical data. We found that there is 8 columns which have the presence of null variable in which 7 of them can be negligible but the column store there is about 117706 null values which need to be treated or the column need to be ruled out. The numerical features have different scales, which may be a problem for some machine learning algorithms. The features should be rescaled to have similar scale

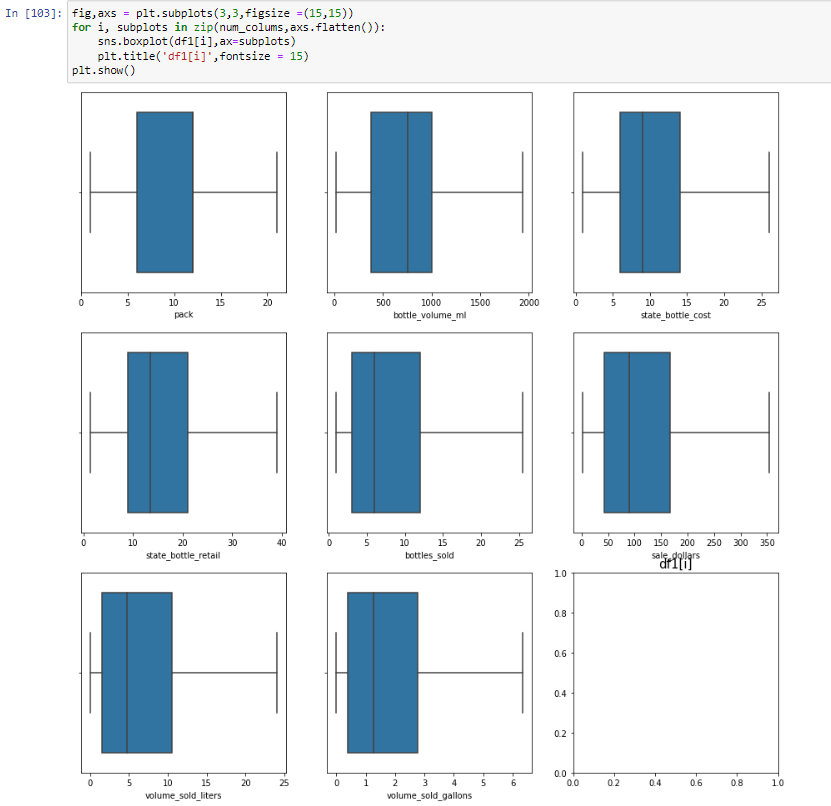
**Distribution of Numeric Variables Original Data:**

As we are analyzing the sales we will be mainly dealing with the numerical data more than the categorical one. So that as a primary step we will be sorting the numerical columns separately for analyzing the data.

Mainly we are taking 8 numerical columns for the analysis of the sales and the distribution of the numerical variables is here:



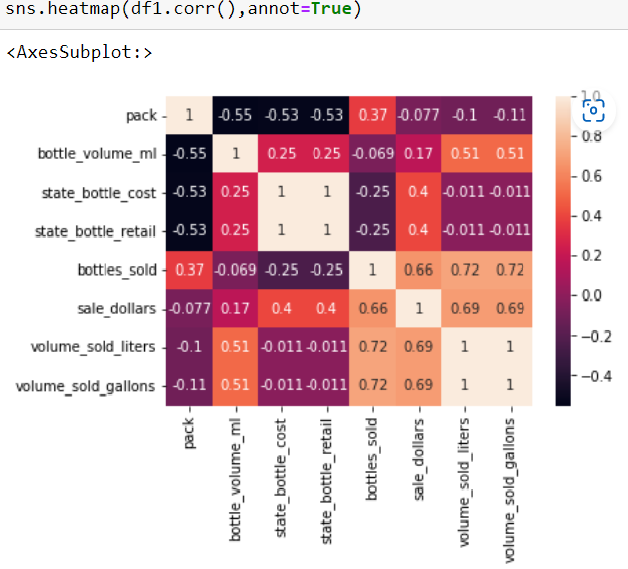
**Outliers of Numeric Variables Original Data:**



## Correlation between the variables.

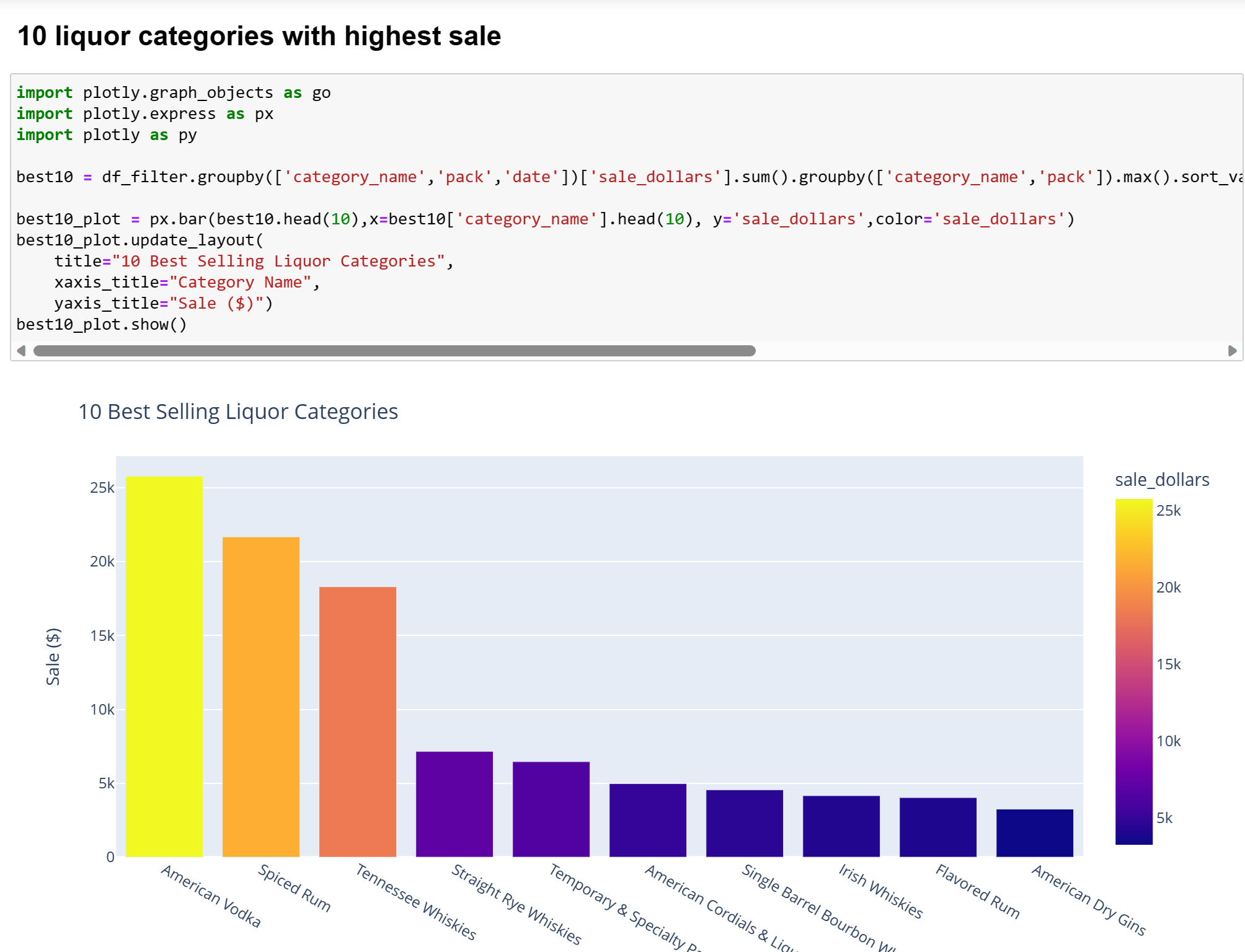
## As we are considering the dependent numerical variable we need to look into the correlation between the variables for better analysis.

## Here is the heat map :

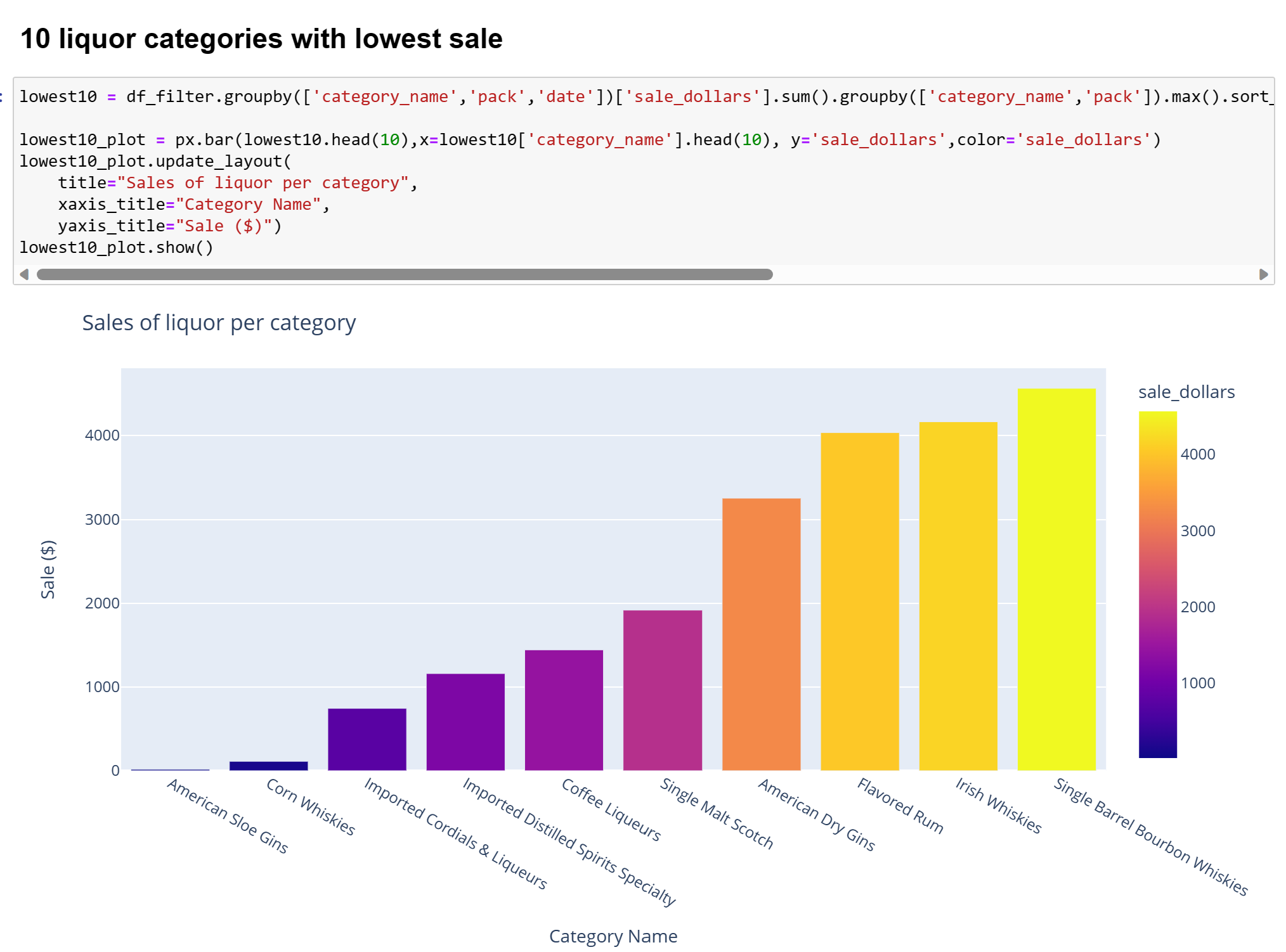
****

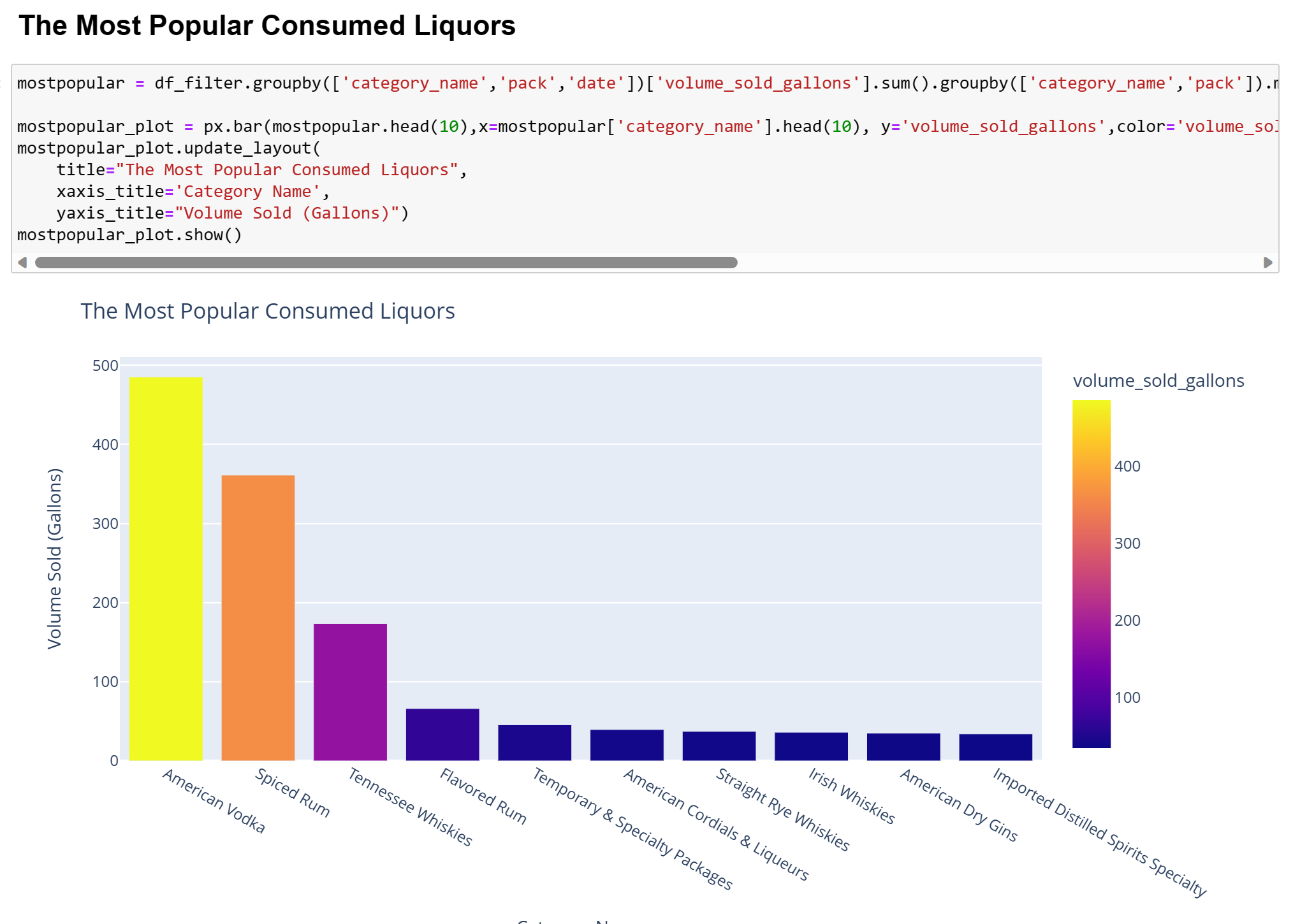
**Vendor number, store number, county number , pre\_icu\_los\_days** – Since these feautures have no impact on the future prediction of the volume of liquor sold we will be dropping this feautures(store name,vendor name,county name is already mentioned in dataset)

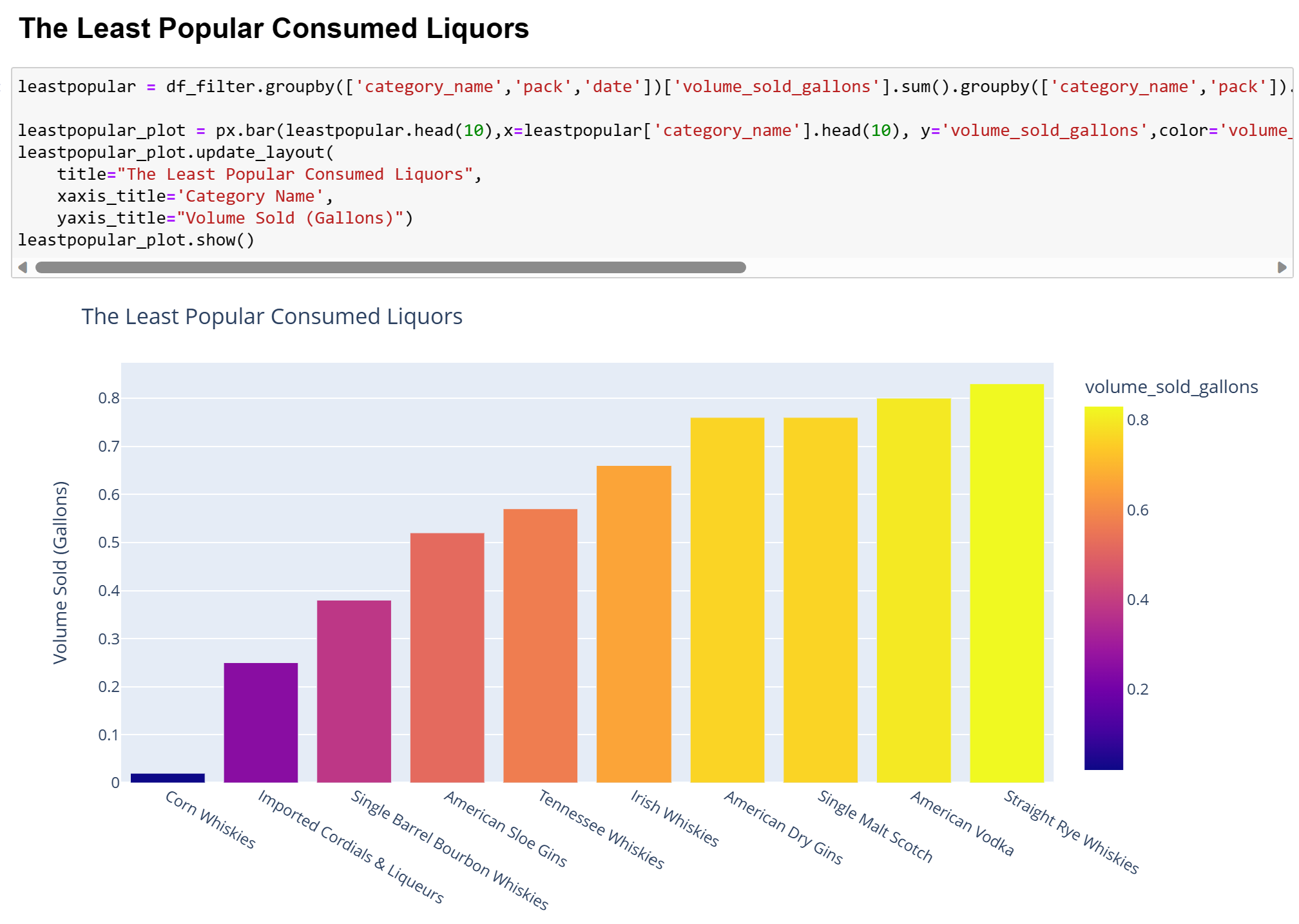
# Explore Data Analysis:

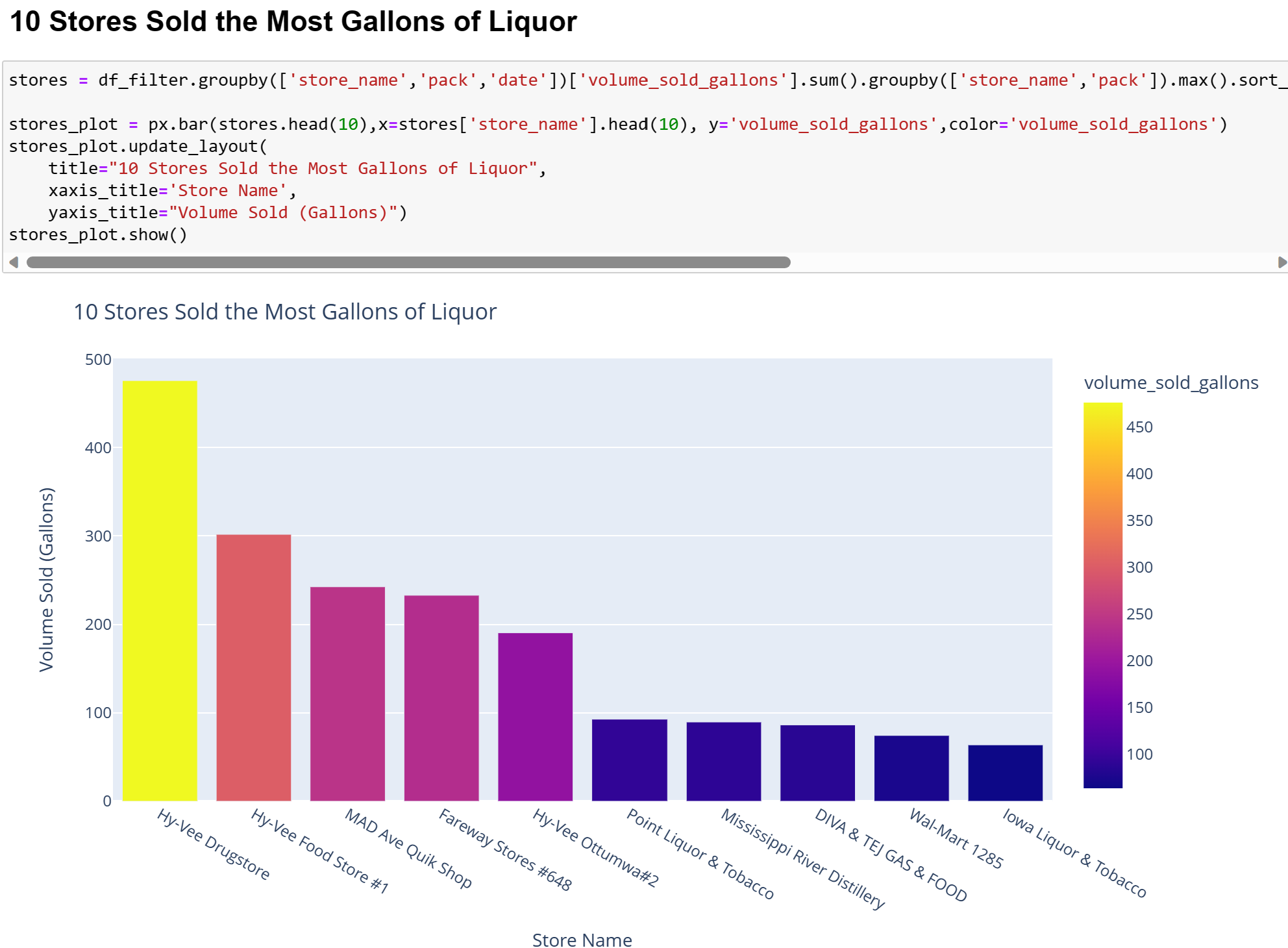
****

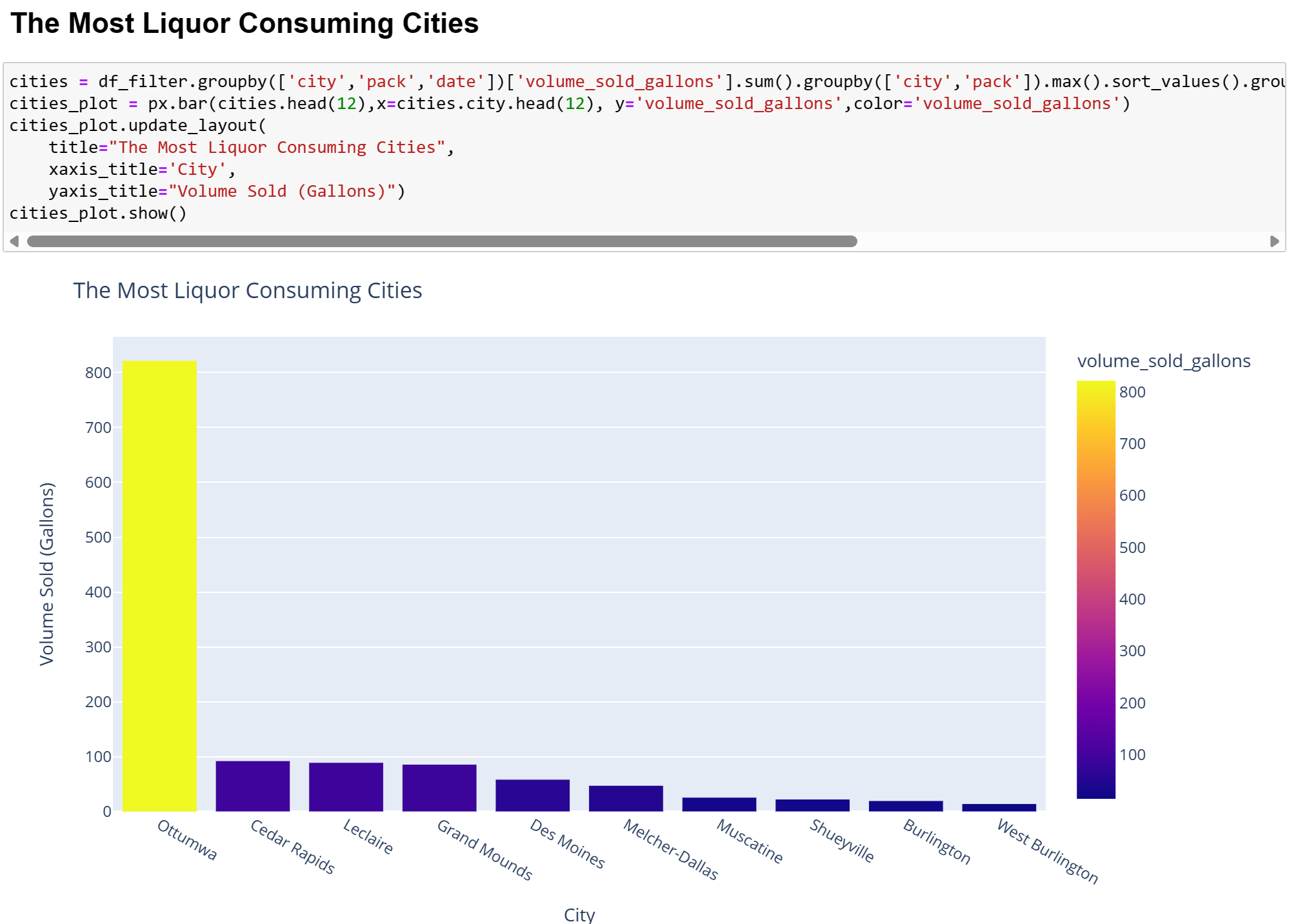
Above Plot shows top 10 best selling liquor, In that we can see American Vodka contributing more.

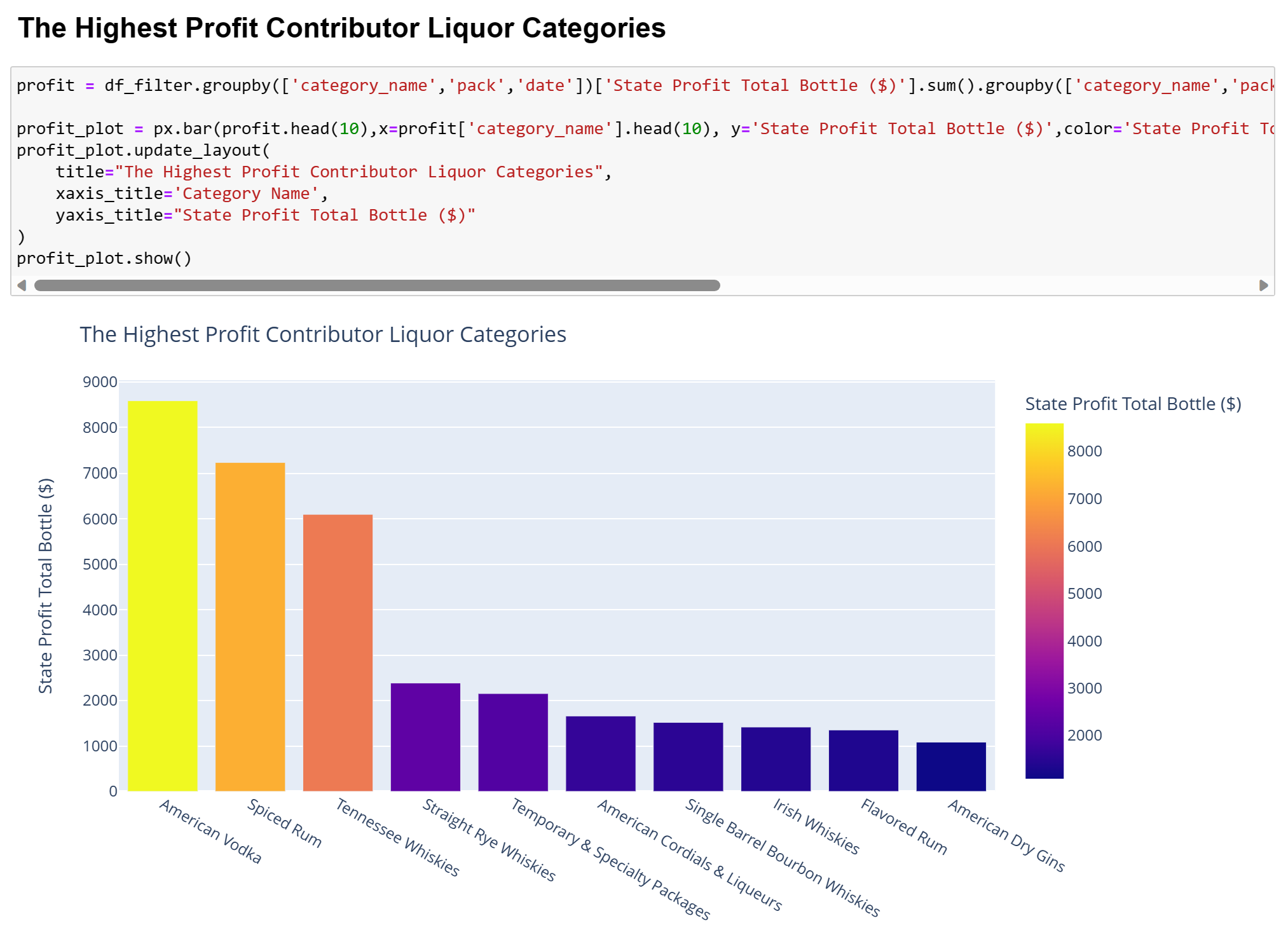
****

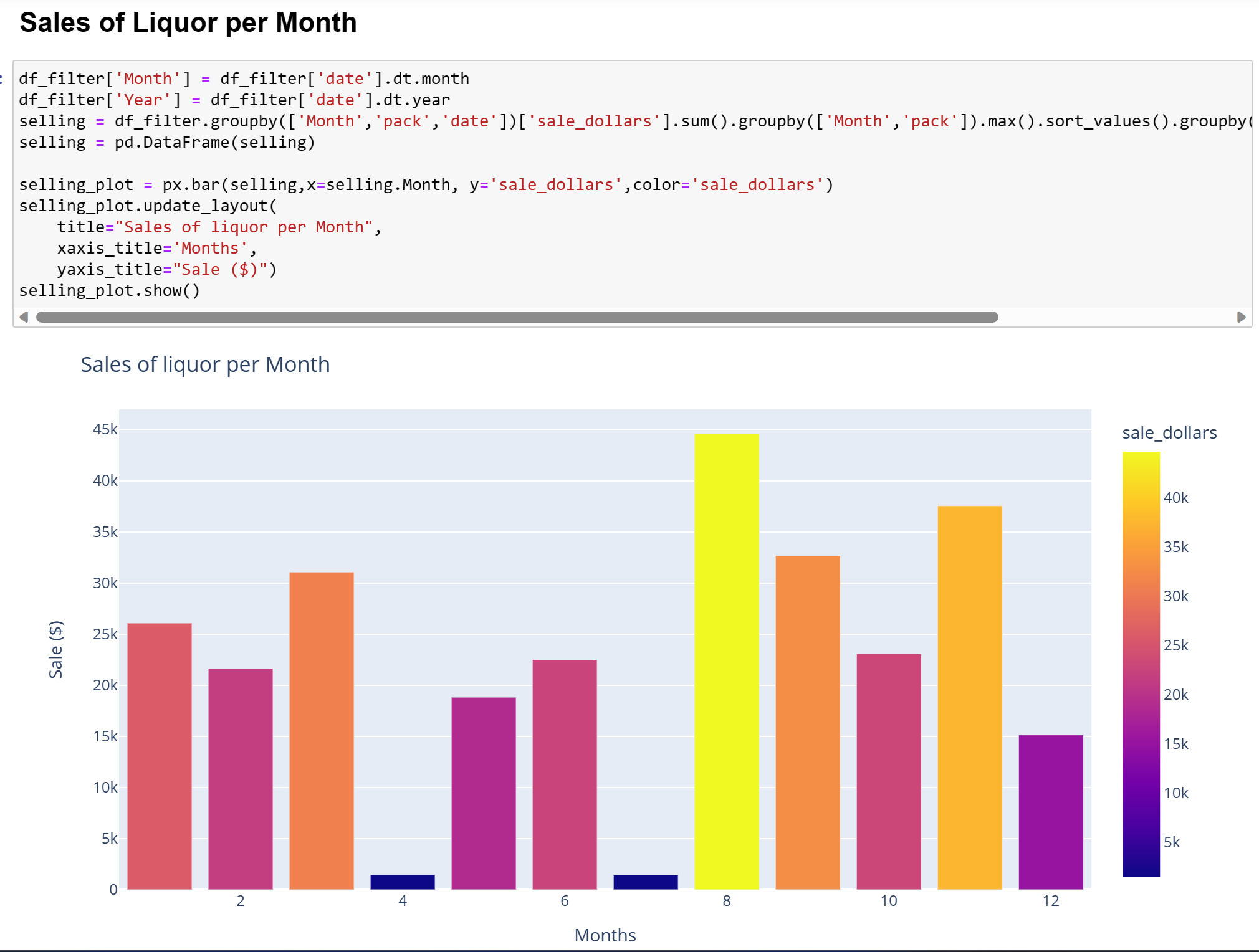
****

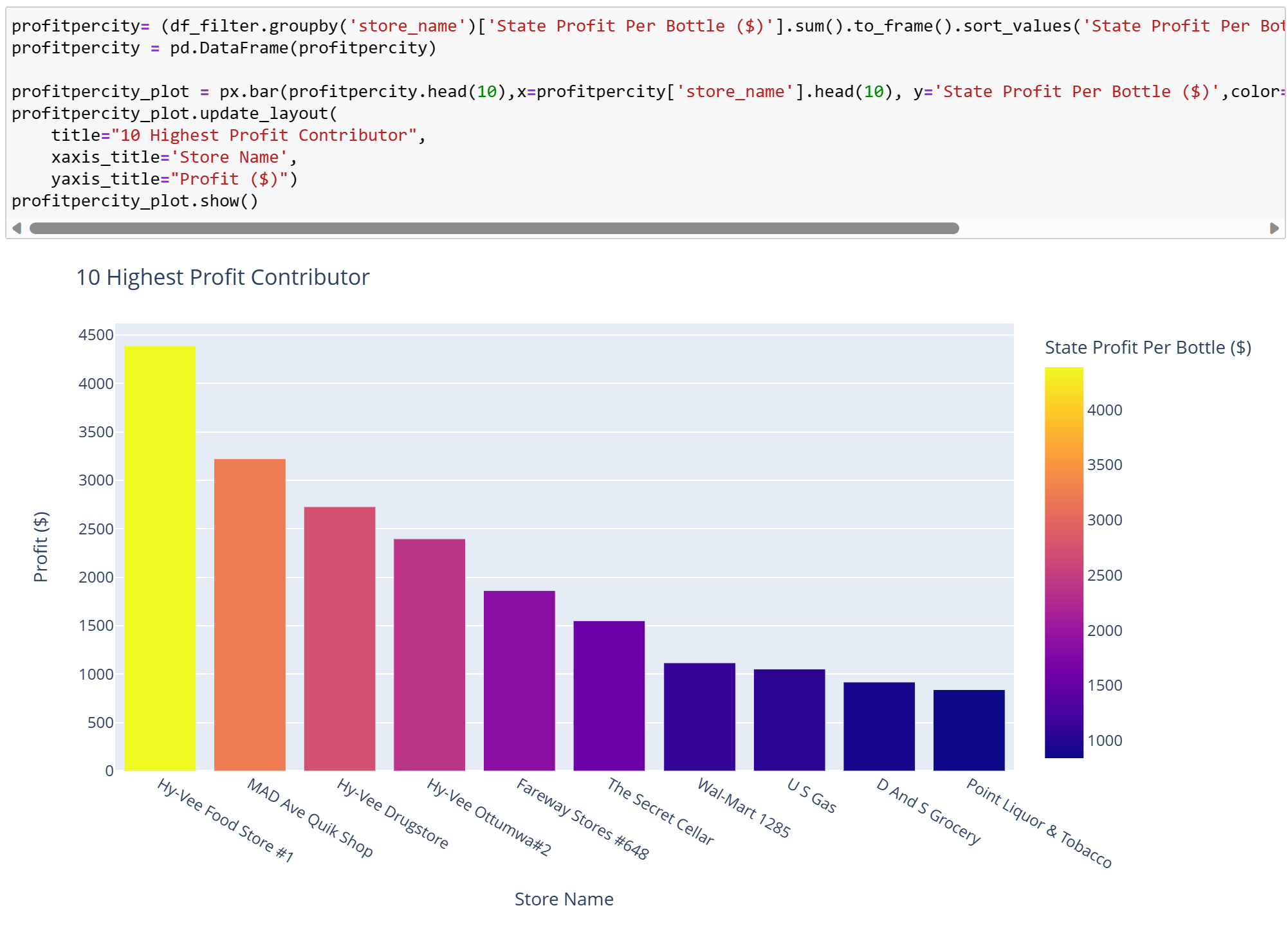
****

****

****

****

****

****

****

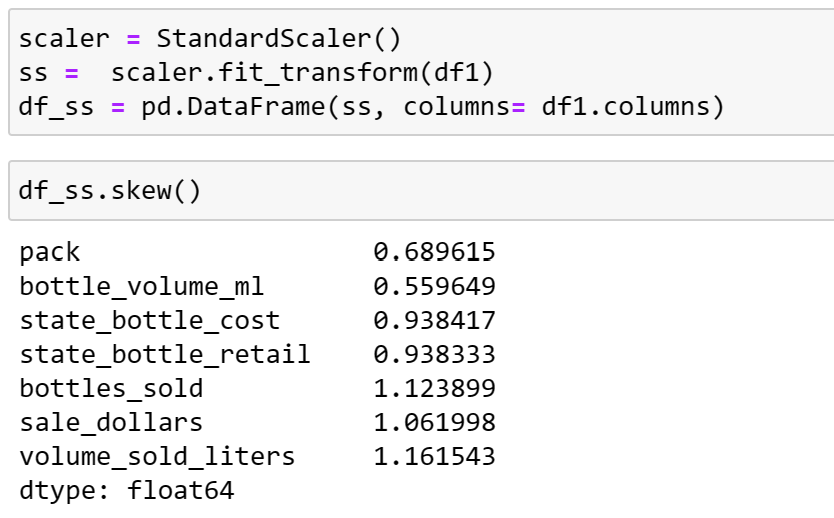
# Data Pre-processing for Model Building

Dropping the columns which is not significant for model building.

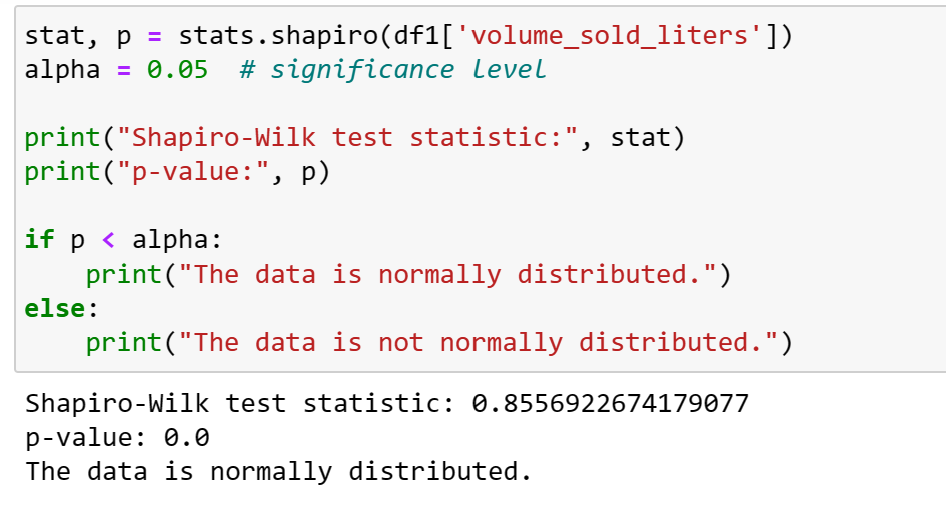
****

****

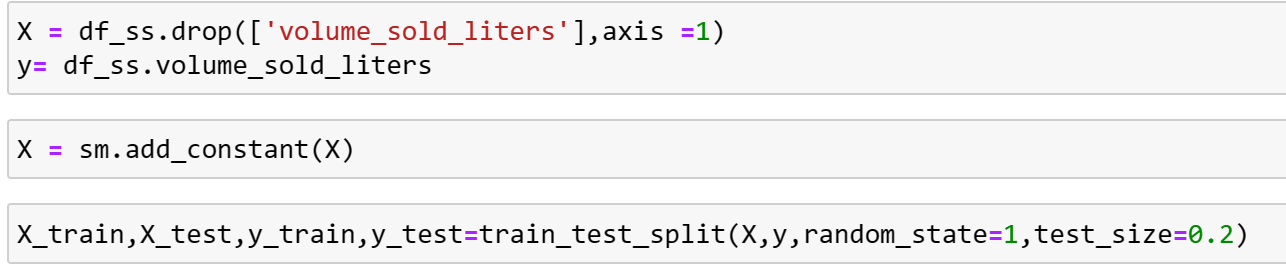
**Data is Slightly skewed**

****

**Statistical test on dependent variable**

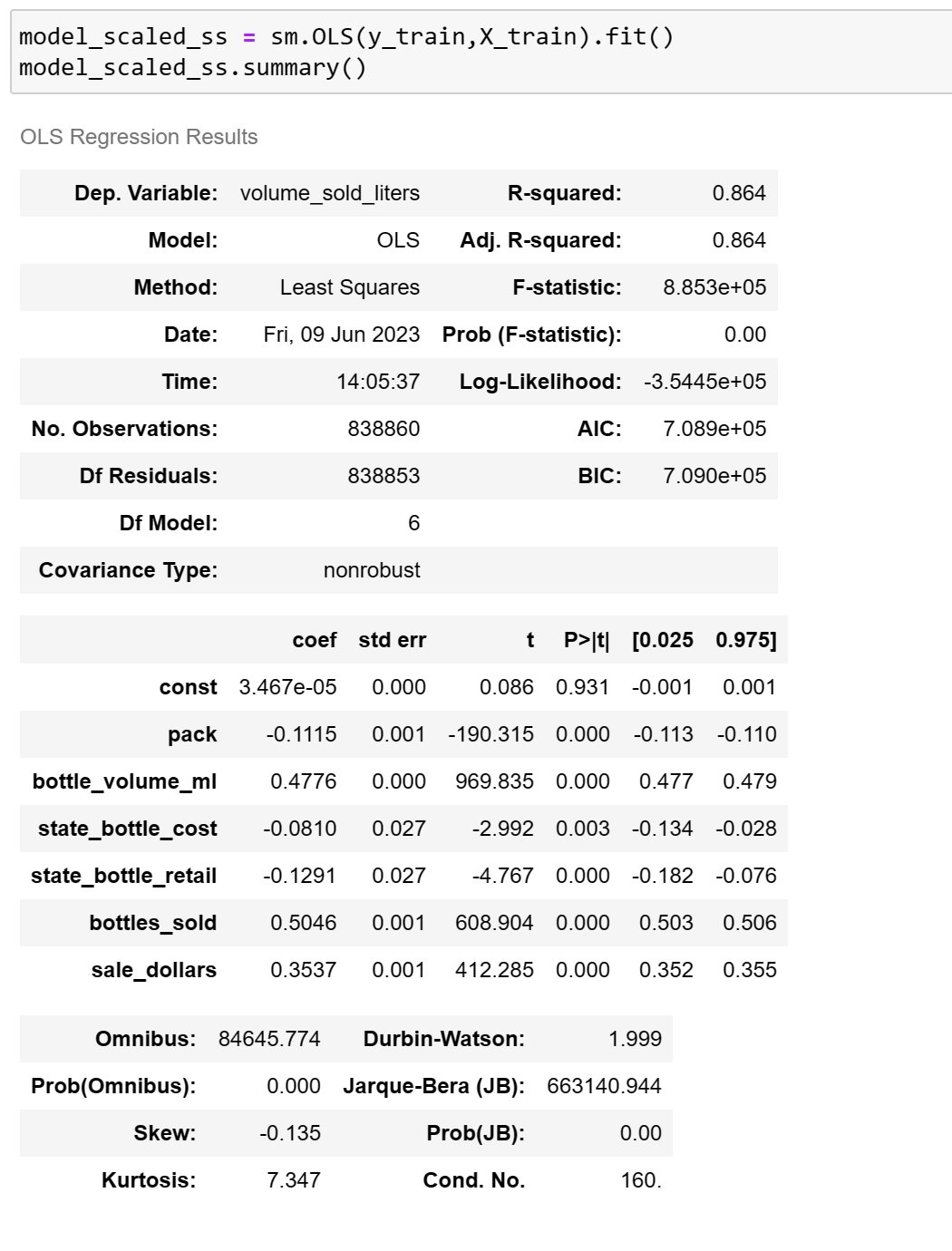
****

**Splitting the Test and Train data (80:20)**

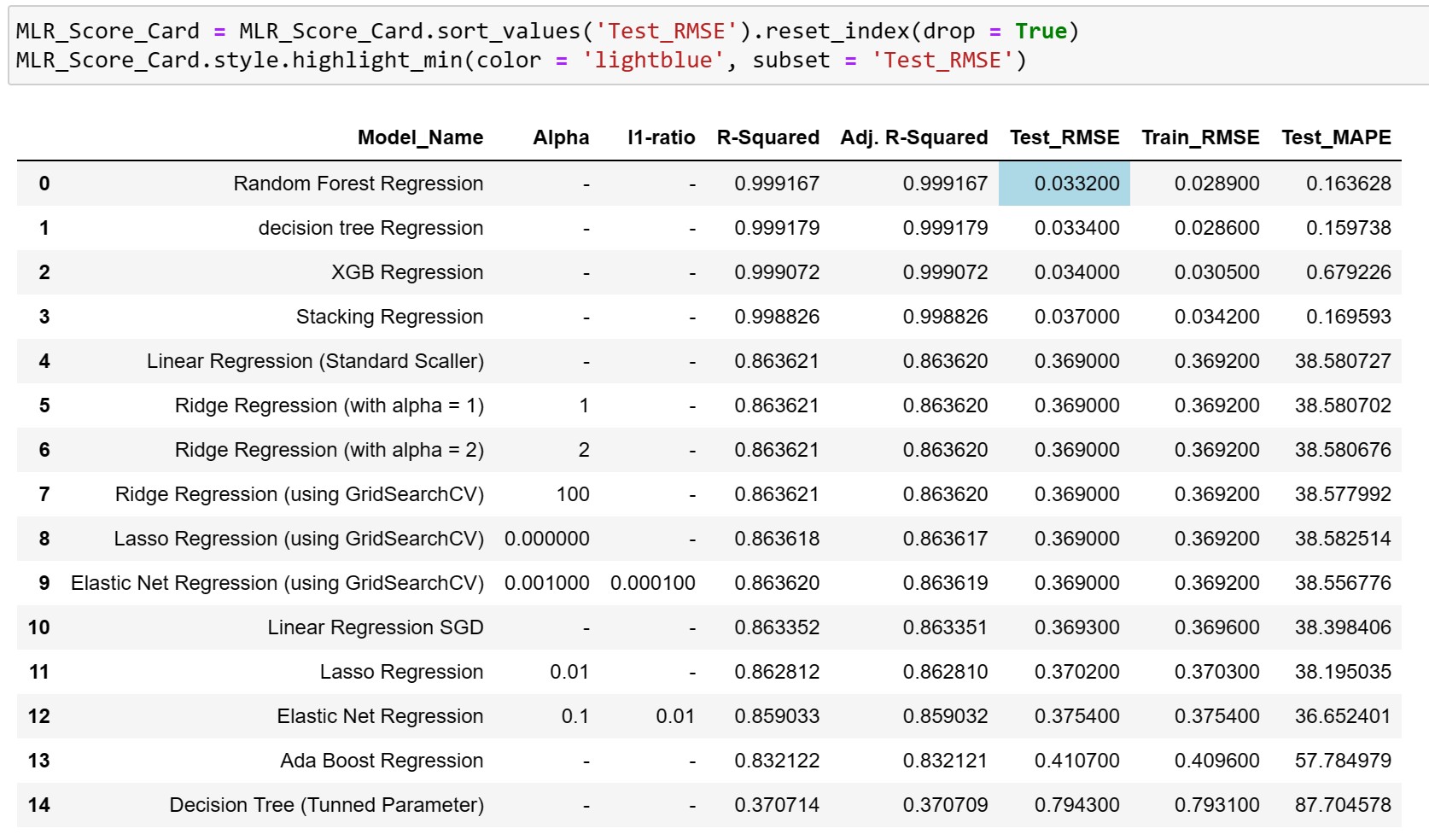
****

## Base Model

Fitting the base model using OLS method

****

## Fitting Multiple Regression Model:

****

## CONCLUSION:

Based on the Above information, the random forest regression model seems to have achieved excellent performance based on several evaluation metrics:

1. R-squared: The R-squared value of 0.99916 indicates that the model explains approximately 99.916% of the variance in the target variable. A high R-squared value suggests that the model fits the data very well, and the majority of the variability in the target variable is captured by the model.

2. Adjusted R-squared: The adjusted R-squared value of 0.99916 is identical to the R-squared value in this case. This suggests that the model contains no unnecessary variables or overfitting issues, as the adjusted R-squared is usually lower than the R-squared when there are excessive variables in the model.

3. Test RMSE: The test root mean squared error (RMSE) of 0.0332 indicates that, on average, the model's predictions have an error of approximately 0.0332 units when applied to unseen test data. A lower RMSE value suggests better predictive performance, so the provided RMSE value is relatively low.

4. Train RMSE: The train RMSE of 0.0289 represents the average error of the model's predictions on the training data. A lower train RMSE suggests that the model is fitting the training data well, with small discrepancies between the predicted values and the actual values.

5. MAPE: The Mean Absolute Percentage Error (MAPE) of 0.163 indicates the average percentage difference between the predicted and actual values. A lower MAPE indicates better accuracy, and the provided MAPE value is relatively low.

Based on these metrics, the random forest regression model appears to be performing exceptionally well, demonstrating a high level of accuracy and predictive power.

Hence RANDOM FOREST REGRESSOR is predicting ‘Volume sold litre’ WELL compare to other Regression model.