

PROJECT3

Data: - Warehouse_and_Retail_Sales.csv

Introduction:

The dataset contains extensive information about liquor sales transactions in Montgomery County, Maryland, USA. It covers both retail and wholesale channels, providing details such as the calendar year, month, supplier name, item code, item description, item type, and quantities of products sold. The dataset includes sales from Department of Liquor Control (DLC) dispensaries, transfers to DLC dispensaries, and sales to Montgomery County licensees. By analyzing this dataset, we can gain insights into sales trends, supplier performance, and product distribution across different channels. These insights are valuable for stakeholders in the liquor industry and regulatory bodies, helping them make informed decisions and effectively manage liquor sales operations in Montgomery County.

Descriptive Statistics:

DLC Dispensaries:

The average number of cases sold at DLC dispensaries is 7.02, with the highest recorded sales reaching 2739 cases. There were instances where as few as -6.4 cases were recorded, indicating returns or unfulfilled sales.

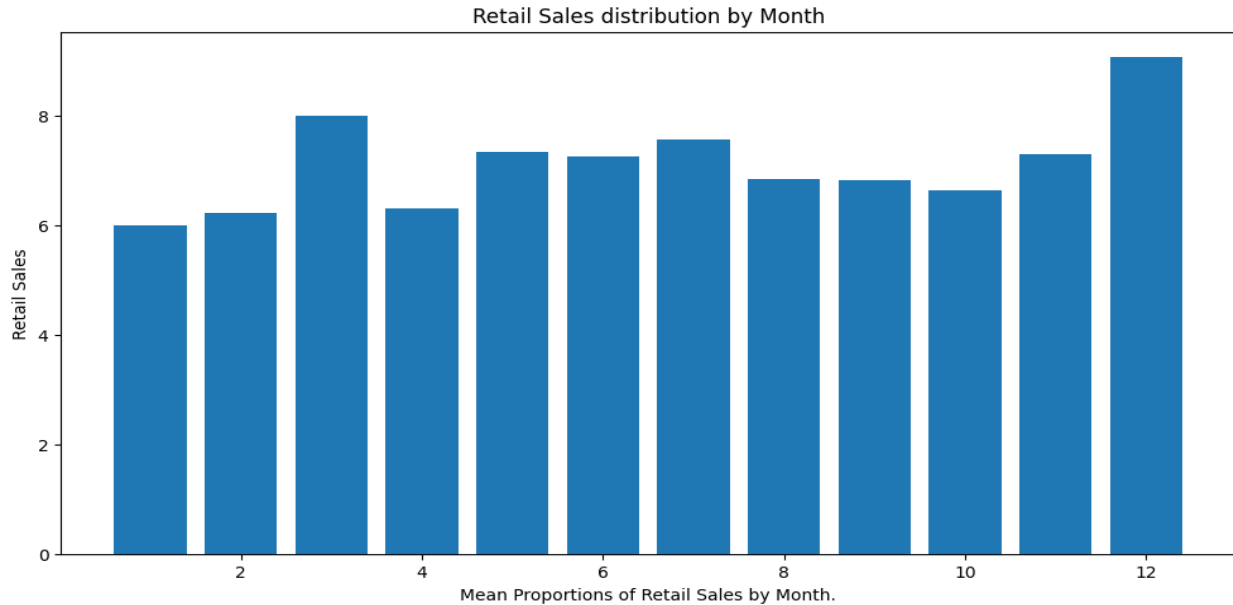
Product Transferred to DLC Dispensaries:

On average, 6.9 cases of product are transferred to DLC dispensaries. The highest volume transferred at once was 1990 cases.

Product Sold to MC Licenses:

An average of 25.2 cases of product is sold to MC licenses. The highest volume sold to MC dispensaries reached 18317 cases.

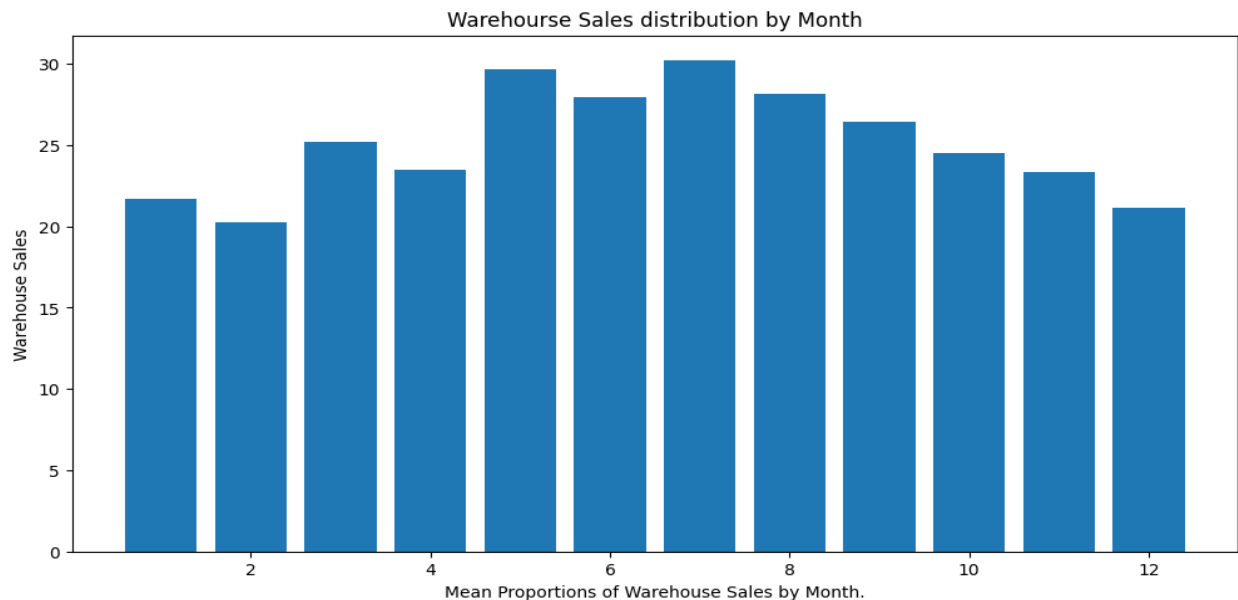
Bar Plot1



1. The highest number of cases of product sold from DLC dispensaries occurs in December. This data makes sense because December is typically a peak month for sales, with liquor often being included in holiday gatherings.

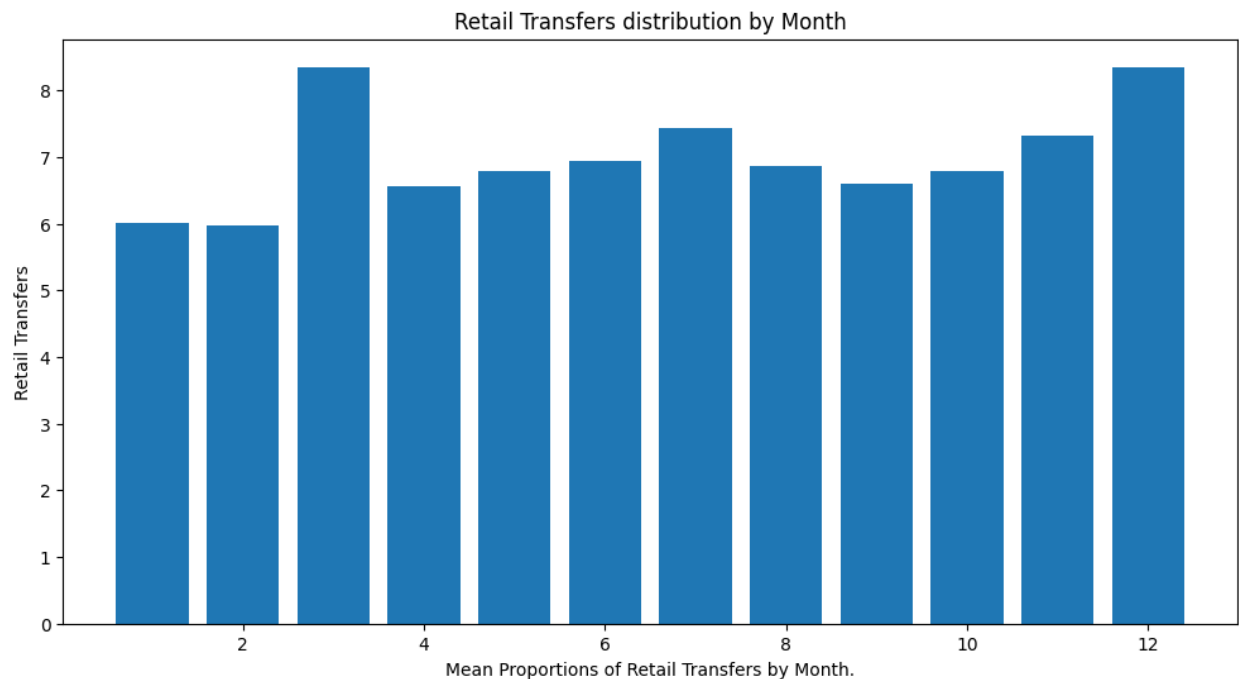
2. Liquor sales might experience an even greater increase during July and August, especially when compared to general retail, as people tend to host more summer parties during this time.

Bar Plot2



During the holiday season, it is common to see a significant increase in sales, particularly in November and December. This can be attributed to warehouses stocking up on liquor supplies in anticipation of the higher demand during this festive time. Retailers also expect a boost in liquor sales as people often purchase them as gifts or for gatherings. Additionally, there might be another surge in sales during the summer months, specifically in July or August. This could be because of the increased demand for liquors used in summer cocktails or parties, such as tequila and rum. Furthermore, purchases of liquors commonly consumed during barbecues, like whiskey and vodka, may also contribute to this rise in sales.

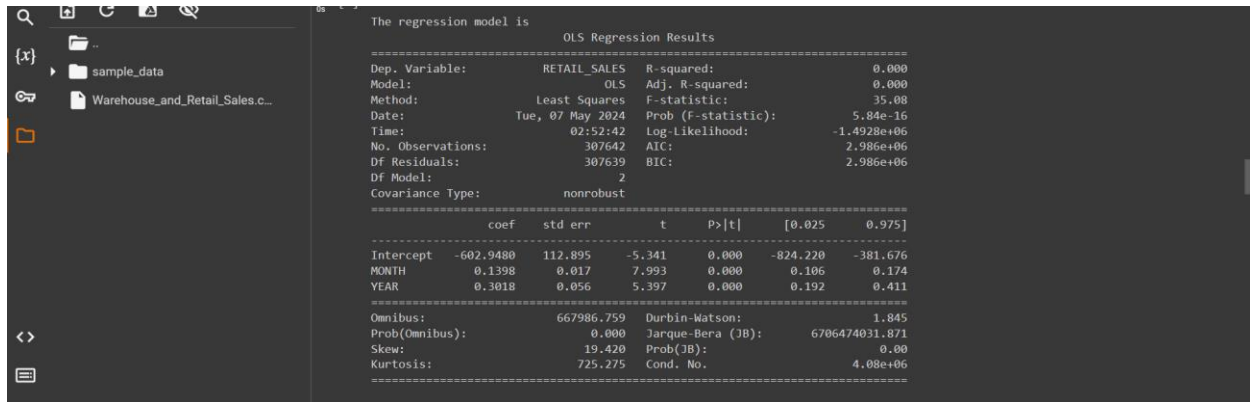
Bar Plot3



Monthly transfer fluctuations: The graph displays the quantity of liquor cases transferred to DLC dispensaries on a monthly basis. There appears to be some fluctuation in the transfer numbers throughout the year, but it's challenging to determine if there is a consistent seasonal pattern without additional data points.

It would be beneficial to have information about the time frame represented in this graph (e.g., one year, multiple years). A more extensive dataset could potentially unveil seasonal trends that are not evident in the current view.

Linear Regression



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code cell on the right. The code cell contains the output of an OLS regression model. The file explorer shows a folder named 'sample_data' containing a file 'Warehouse_and_Retail_Sales.c...'. The code cell output is as follows:

```
The regression model is
=====
OLS Regression Results
=====
Dep. Variable:    RETAIL_SALES    R-squared:        0.000
Model:            OLS            Adj. R-squared:    0.000
Method:           Least Squares   F-statistic:       35.00
Date:             Tue, 07 May 2024 Prob (F-statistic): 5.84e-16
Time:             02:52:42        Log Likelihood:    -1.4928e+06
No. Observations: 307642         AIC:              2.986e+06
Df Residuals:     307639         BIC:              2.986e+06
Df Model:         2
Covariance Type:  nonrobust
=====
               coef    std err          t      P>|t|      [0.025    0.975]
-----
Intercept    -602.9480    112.895     -5.341     0.000    -824.220   -381.676
MONTH         0.1398      0.017      7.993     0.000      0.106     0.174
YEAR         0.3018      0.056      5.397     0.000      0.192     0.411
=====
Omnibus:        667986.759    Durbin-Watson:      1.845
Prob(Omnibus):   0.000    Jarque-Bera (JB):   6706474031.871
Skew:            19.420    Prob(JB):           0.00
Kurtosis:        725.275    Cond. No.           4.08e+06
=====
```

The dataset's regression analysis has provided us with valuable insights into the relationship between time variables (MONTH and YEAR) and retail sales (RETAIL_SALES) in Montgomery County. The R-squared value of 0.000 indicates that the selected variables, MONTH and YEAR, do not fully explain the variability observed in retail sales. However, the coefficients obtained from the regression model offer meaningful interpretations. The MONTH coefficient (0.1398) suggests a modest positive relationship, indicating that retail sales tend to increase slightly with each passing month. Similarly, the YEAR coefficient (0.3018) reflects a more pronounced positive relationship over time, implying that retail sales exhibit a larger increase with each successive year.

Both the MONTH and YEAR coefficients are statistically significant predictors of RETAIL_SALES, as indicated by their low p-values (0.000). However, the overall model performance is limited due to the low R-squared value, suggesting that additional variables or more complex relationships may be necessary to better capture the variation in retail sales. It is worth noting that the predicted value for RETAIL_SALES in September 2020 is estimated at 7.855, based on specific values for MONTH and YEAR in the regression equation. Despite these findings, the model's interpretability is constrained by its inability to fully account for the observed variability in retail sales. This highlights the need for further refinement and exploration of additional factors that influence liquor sales dynamics in Montgomery County.

Sklearn Model Regressors Predictions

Predicted retail sales for year 2019 using linear regression: 7.062230408657172

Predicted retail sales for year 2019 using K Nearest Neighbours: 0.30933333333333333

Predicted retail sales for year 2019 using Decision Tree: 6.916168598691795

Predicted retail sales for year 2019 using Random Forest: 6.914890376623106

Predicted retail sales for year 2019 using Neural Networks: 7.935977072602437

In general, Neural Networks analysis is the best solution for its complexity nature and it also processes more data than other analysis models. Therefore, Neural Networks is best for this situation. But we could check that by train_test so that we could conclude which model is better for the given data set.

After Train_Test:

Mean Absolute Error for Linear Regression: 10.178481359444731

Mean Absolute Error for K Nearest Neighbours: 8.993687767226053

Mean Absolute Error for Decision Tree: 10.173231241009969

Mean Absolute Error for Neural Networks: 9.164599653748919

After evaluating various machine learning models such as Linear Regression, K Nearest Neighbors, Decision Tree, Random Forest, and Neural Networks, the selection of the best model depends on specific evaluation metrics and considerations. K Nearest Neighbors (KNN) stands out with the lowest Mean Absolute Error (MAE) of around 8.99, indicating its superior accuracy in predicting retail sales compared to the other models. However, the extremely low predicted value for KNN (0.31) raises concerns about its effectiveness on this dataset. On the other hand, Neural Networks also demonstrate competitive performance with a Mean Absolute Error of approximately 9.16 and the highest predicted value (7.94), showcasing their ability to capture complex patterns in the data. In contrast, Linear Regression, Decision Tree, and Random Forest, while satisfactory, have higher prediction errors compared to Neural Networks. Overall, Neural Networks emerge as the most promising model for this dataset due to their lower Mean Absolute Error and capability to handle intricate data relationships.

Conclusion

In conclusion, after evaluating various machine learning models (Linear Regression, K Nearest Neighbors, Decision Tree, Random Forest, and Neural Networks) for predicting retail sales in Montgomery County based on the provided dataset, Neural Networks emerge as the most promising model. Despite the complexity of Neural Networks and their ability to process large datasets effectively, they demonstrate competitive performance with a Mean Absolute Error (MAE) of approximately 9.16 and the highest predicted value (7.94). This indicates that Neural Networks are adept at capturing complex patterns in the data and offer more accurate predictions compared to other models. While K Nearest Neighbors (KNN) also exhibits low MAE, the extremely low predicted value (0.31) raises doubts about its suitability for this dataset. Therefore, Neural Networks are recommended as the preferred model for predicting retail sales in Montgomery County, emphasizing the importance of leveraging advanced techniques to optimize predictive accuracy and insights in liquor sales analysis.

