Data Bootcamp Final Project Report

What Factors can Increase Store Sales?

Yuhe Wu

Yw6800

Introduction

For small businesses like grocery store owner, or supermarket chain store boss gaining sales requires trying out different methods. Some may want to buy an existing store, or invest a new store in a new location, or try different advertisements, but a lot of money can be wasted if one method fails. Therefore, for small business owners, we need to know how to increase store sales effectively. With the data I have, I will analyse this problem with stores in the USA. This analysis can provide actionable insights for store owners and managers who face similar challenges, helping them make better investment and operational decisions.

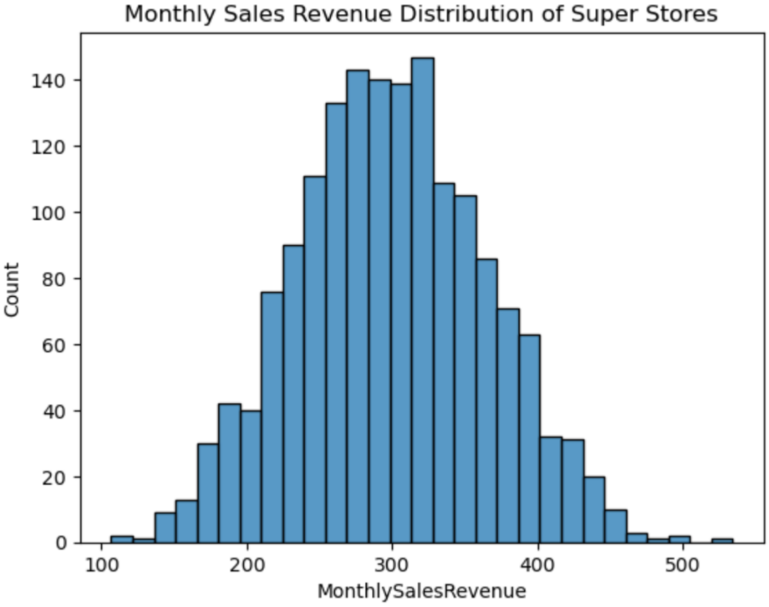
The objective is to predict monthly store sales revenue based on factors such as product variety, marketing spend, customer footfall, store size, promotions, and other relevant store-related variables. Additionally, I will classify stores as "profitable" or "non-profitable" based on their monthly sales crossing a specified threshold.

My main finding is that, product variety and store size were found to have the most significant positive impact on store sales. To increase sales, store owners should focus on expanding product variety, improving operational efficiency, and carefully managing promotional activities.

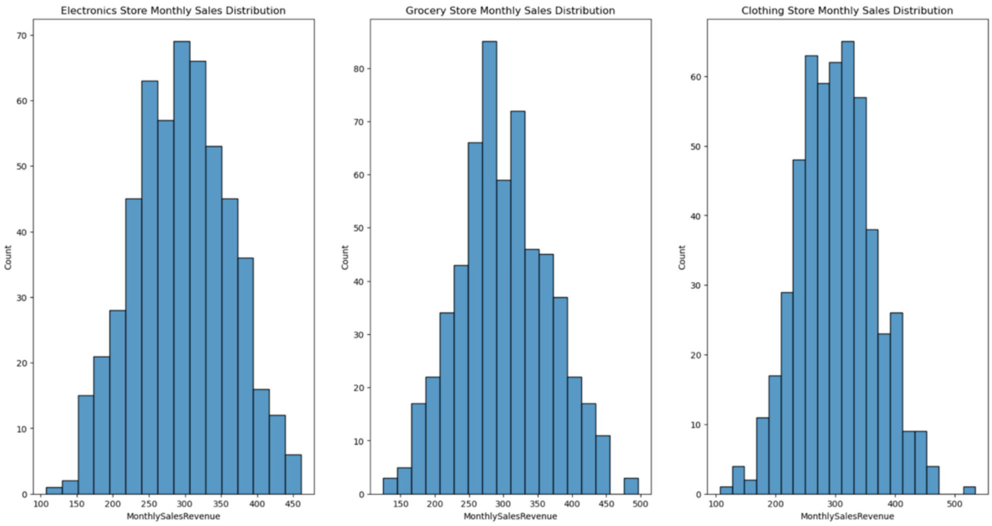
Data Description

The dataset I used is from Kaggle, that is cited by the author from an original project in UWM. It includes data from different type of stores from some cities in the USA, and included different influencing factors, such as Product Variety, Marketing Spend etc. There are in total 1650 store data. Two of them are object data type, and others are integers or floats. There were enough information to investigate on my research topic.

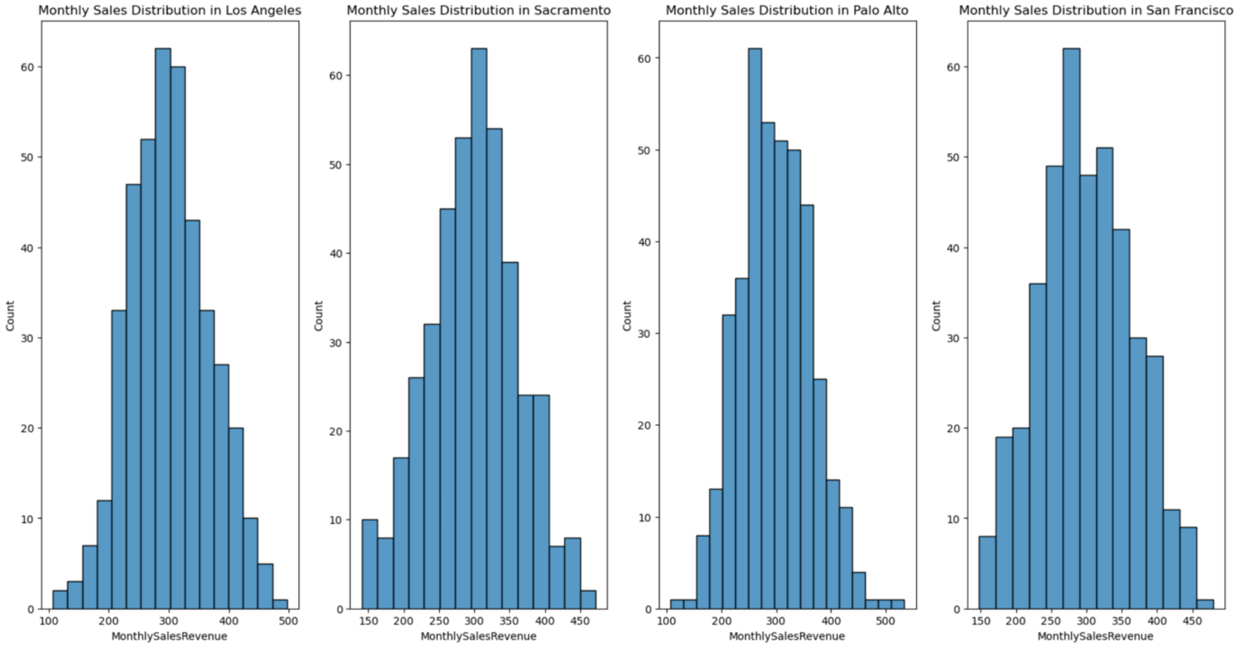
Importantly, our focus on is the range of "Monthly Sales Revenue." It ranges from 65.54 to 534.26 thousand dollars, and its average is 299.25 thousand dollars. The majority of the data concentrated in the 300–350 range.



After seeing the overall sales number distribution, I would like to see how monthly sales is distributed for the different types of stores. From the chart, we can tell that all three categories show relatively symmetric, bell-shaped distributions, hinting at potential normality. The central ranges are similar, with grocery stores having a slightly narrower spread, suggesting more consistent sales. Clothing stores exhibit the widest range, potentially reflecting greater variation in revenue influenced by external factors like fashion trends or promotional events. They are similar in general.



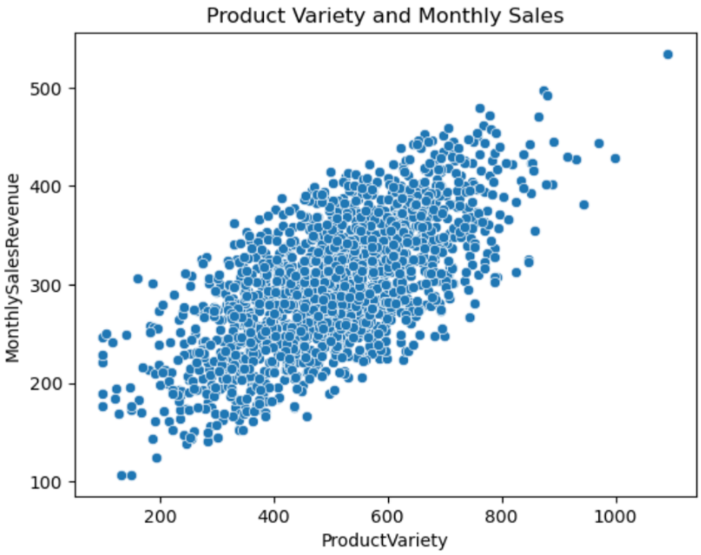
All cities exhibit approximately normal distributions in sales revenue. The central range (~250–350 units) is consistent across all locations, indicating that most stores perform similarly in revenue. Sacramento shows the narrowest range, indicating more consistency. Palo Alto and Los Angeles have slightly broader ranges, suggesting variability in performance.



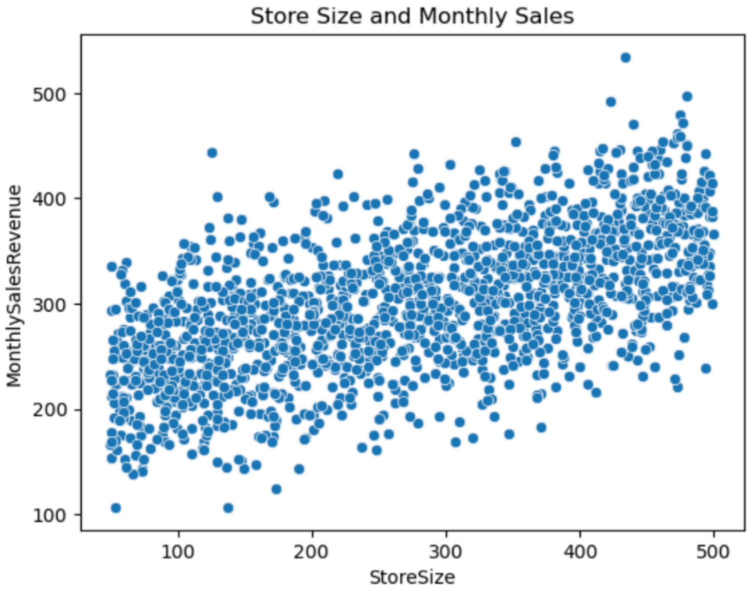
Data Visualization

In this part, I tried to visualize the relationship between variables to learn more about the problem I wanted to solve.

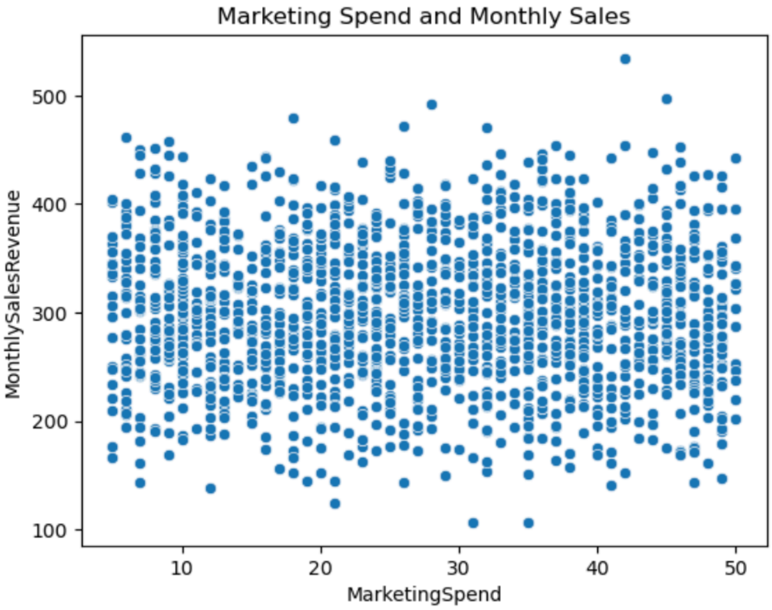
There seems to be a linear positive relationship between monthly sales and product variety.



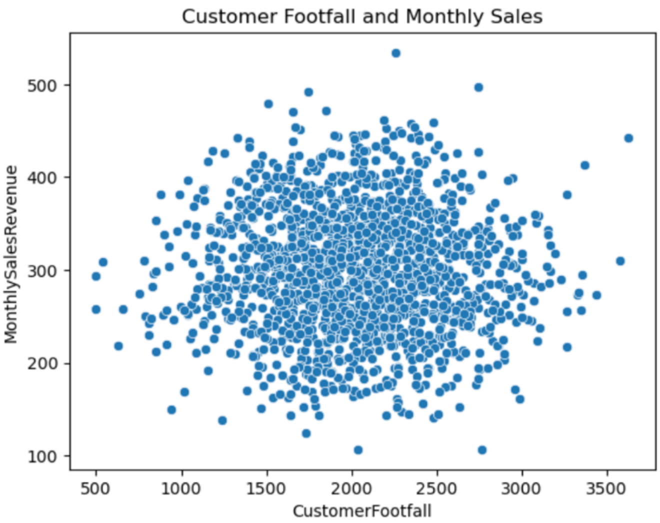
There is a clear positive trend between store size and monthly sales revenue too, suggesting that as Store Size increases, Monthly Sales Revenue also tends to increase. A few points deviate significantly from the trend, which need to be checked.

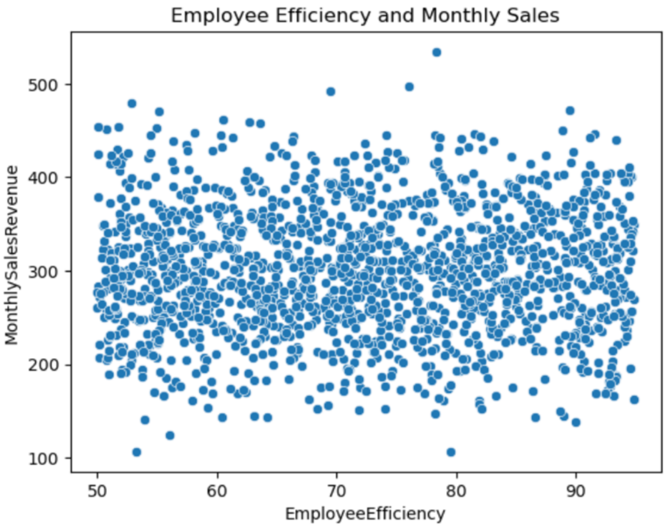


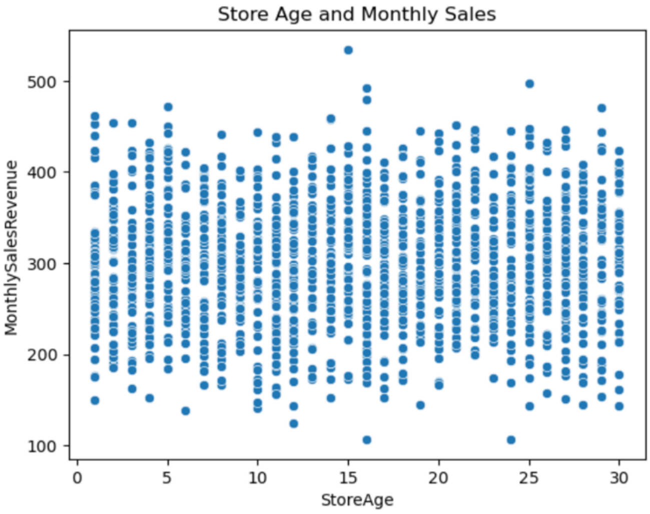
One thing to notice is that, from the scatter plot of marketing spend and monthly sales, I can see there might no strong relationship between monthly sales revenue and the marketing spend. This is interesting because we may assume that if we spend more money on marketing, we could attract more customers, and thus sell more. From the data standpoint, this may not be exactly true.

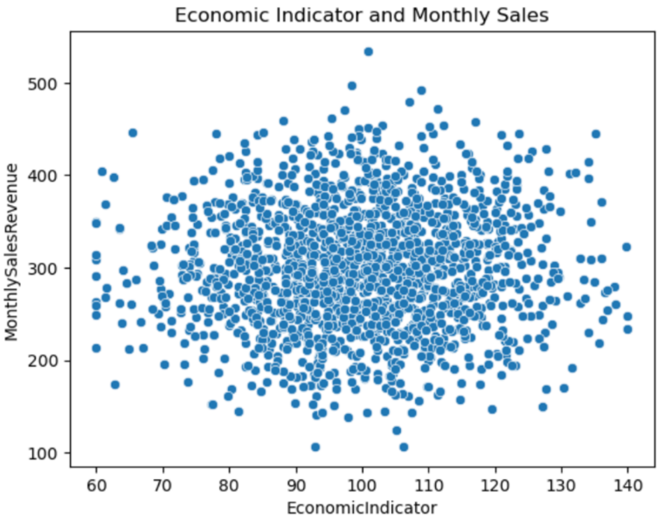


For the rest of the variables like customer footfall, employee efficiency, store age, and economic indicator, they mostly do not show clear relationship with monthly sales. They are likely less influential on store sales.

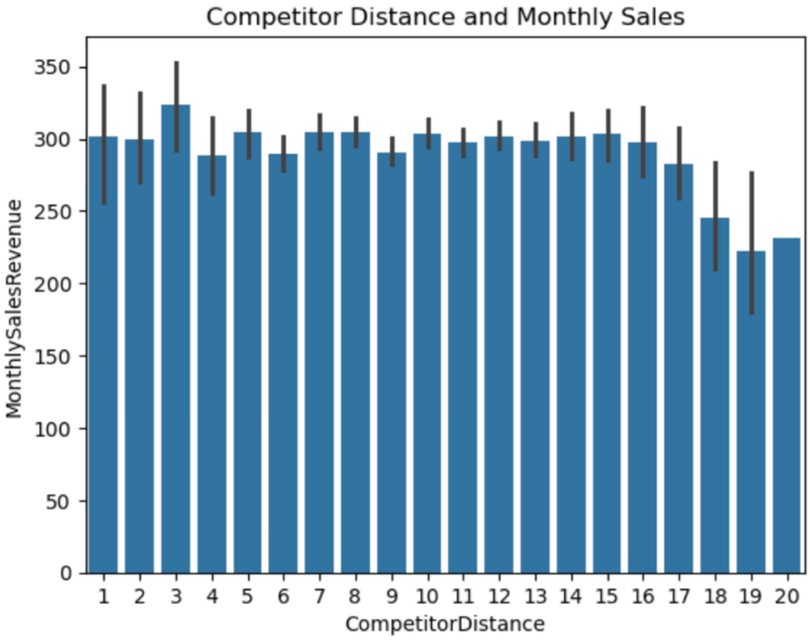




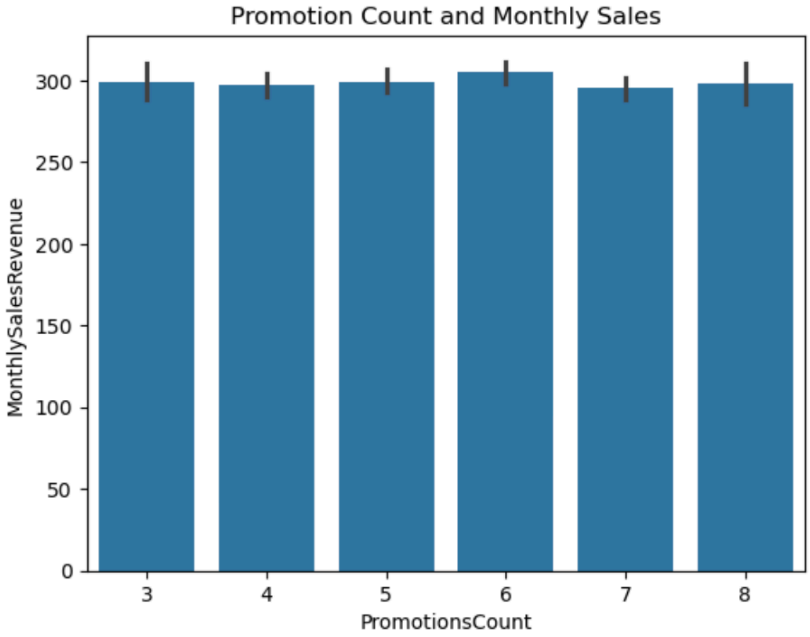




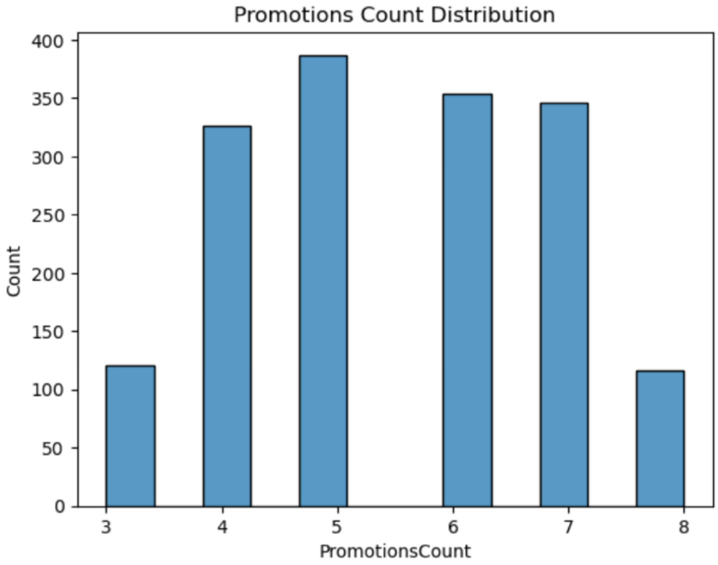
Surprisingly, competitor distance seems not important for monthly sales either. The competitors with distance below 15km do not show a clear influence to the store monthly sales number. Competitor distances greater than 15Km shows some decrease, this may mean the area is too far from downtown, and the population is less dense than the other areas, so the monthly sales could be low.



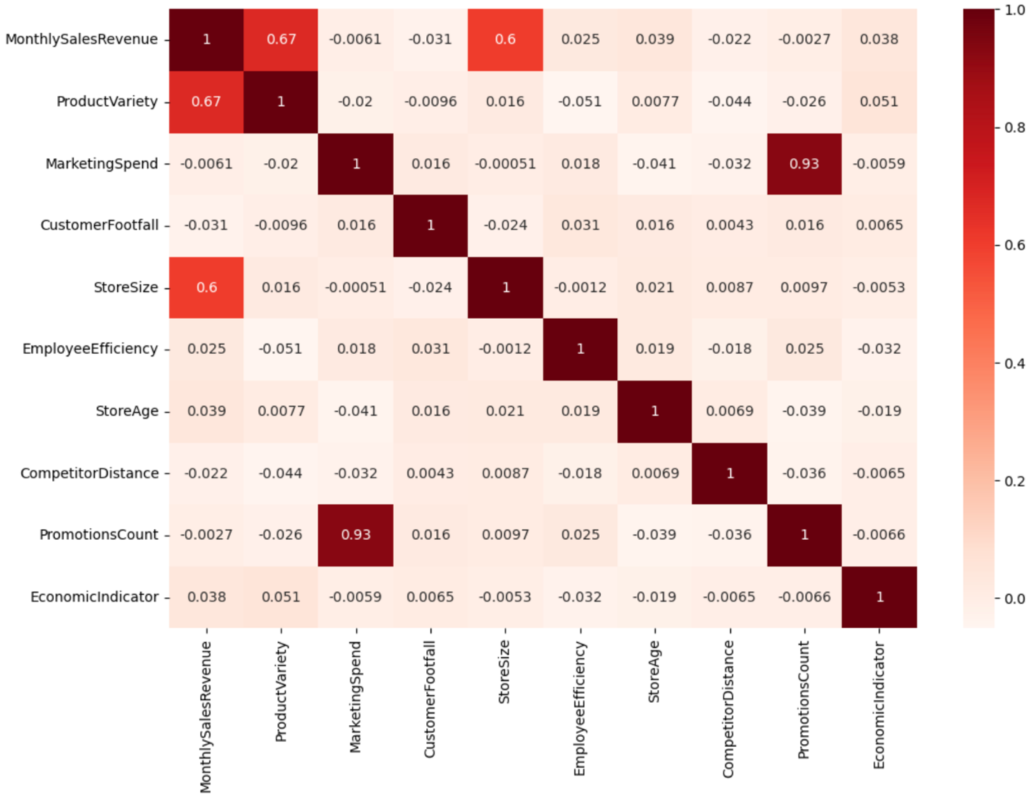
For sales and promotion counts, sales remain relatively stable across promotion counts, making the difference small. However, this does not mean we can simply cut off all promotion activities.



From this chart, we can see the promotion intensity are very similar. Most stores have 4-7 times promotion a month. Although we see "insignificant" influence of promotions count to monthly sales revenue, still if we cut the promotions, it may put us in a disadvantage in the competition. So, this might be a tricky decision.



To summarize, I will analyse the heat chart showing coefficients between variables. This is pretty in line with what we found in the individual scatter plots.



For monthly sales revenue, the strongest correlations are:

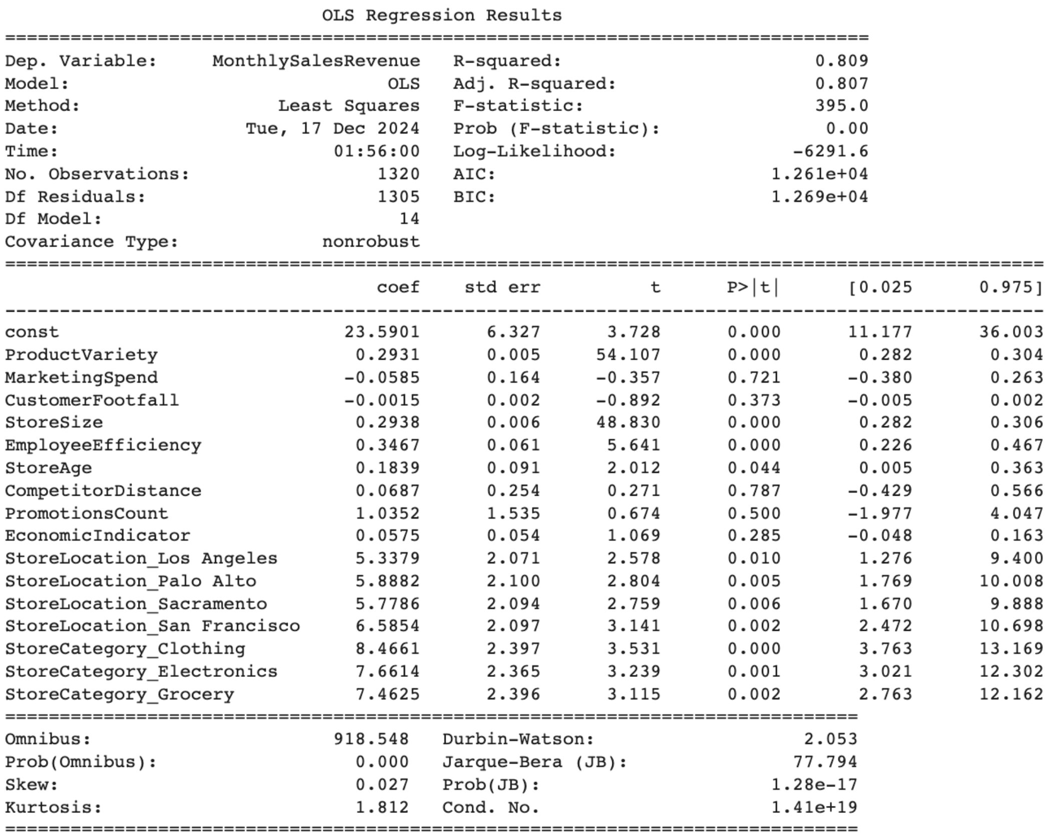
- ProductVariety has strong positive correlation (0.67) with monthly sales revenue. Stores offering a greater variety of products tend to achieve higher revenue.
- StoreSize has strong positive correlation of 0.6. Larger stores are associated with higher sales revenue, which aligns with earlier scatter plot observations.
- Other variables have minimal correlation with sales, indicating weak or no linear relationship.
- Product variety, store size, employee efficiency, store age and economic indicator positively impact revenue, reflecting their importance in operational success and influence of pre-determined factors.

Modelling and Interpretation

Stats Model (OLS)

I used stats model OLS to view more detailed summary of the regression results and check the validity.

From the regression table, we can tell ProductVariety, StoreSize, EmployeeEfficiency are the statistically significant variables, and they dominant predictors positively influencing Monthly Sales Revenue. Store Locations and Store Categories were also statistically significant. We can observe that discernible differences exist between locations and store categories, with San Francisco, Electronics, and Clothing stores showing higher performance. Other factors like MarketingSpend, CompetitorDistance, PromotionsCount, EconomicIndicator have p-values greater than 0.05, indicating they do not contribute significantly to the model, which did fit our preliminary exploration.



```
                             OLS Regression Results
===============================================================================
Dep. Variable:     MonthlySalesRevenue   R-squared:                       0.809
Model:                             OLS   Adj. R-squared:                  0.807
Method:                  Least Squares   F-statistic:                     395.0
Date:                 Tue, 17 Dec 2024   Prob (F-statistic):               0.00
Time:                         01:56:00   Log-Likelihood:                -6291.6
No. Observations:                 1320   AIC:                         1.261e+04
Df Residuals:                     1305   BIC:                         1.269e+04
Df Model:                           14
Covariance Type:             nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          23.5901      6.327      3.728      0.000      11.177      36.003
ProductVariety                  0.2931      0.005     54.107      0.000       0.282       0.304
MarketingSpend                 -0.0585      0.164     -0.357      0.721      -0.380       0.263
CustomerFootfall               -0.0015      0.002     -0.892      0.373      -0.005       0.002
StoreSize                       0.2938      0.006     48.830      0.000       0.282       0.306
EmployeeEfficiency              0.3467      0.061      5.641      0.000       0.226       0.467
StoreAge                        0.1839      0.091      2.012      0.044       0.005       0.363
CompetitorDistance              0.0687      0.254      0.271      0.787      -0.429       0.566
PromotionsCount                 1.0352      1.535      0.674      0.500      -1.977       4.047
EconomicIndicator               0.0575      0.054      1.069      0.285      -0.048       0.163
StoreLocation_Los Angeles       5.3379      2.071      2.578      0.010       1.276       9.400
StoreLocation_Palo Alto         5.8882      2.100      2.804      0.005       1.769      10.008
StoreLocation_Sacramento        5.7786      2.094      2.759      0.006       1.670       9.888
StoreLocation_San Francisco     6.5854      2.097      3.141      0.002       2.472      10.698
StoreCategory_Clothing          8.4661      2.397      3.531      0.000       3.763      13.169
StoreCategory_Electronics       7.6614      2.365      3.239      0.001       3.021      12.302
StoreCategory_Grocery           7.4625      2.396      3.115      0.002       2.763      12.162
==============================================================================
Omnibus:                      918.548   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               77.794
Skew:                           0.027   Prob(JB):                     1.28e-17
Kurtosis:                       1.812   Cond. No.                     1.41e+19
==============================================================================
```

Some key regression metrics include R^2, F-stat and condition number.

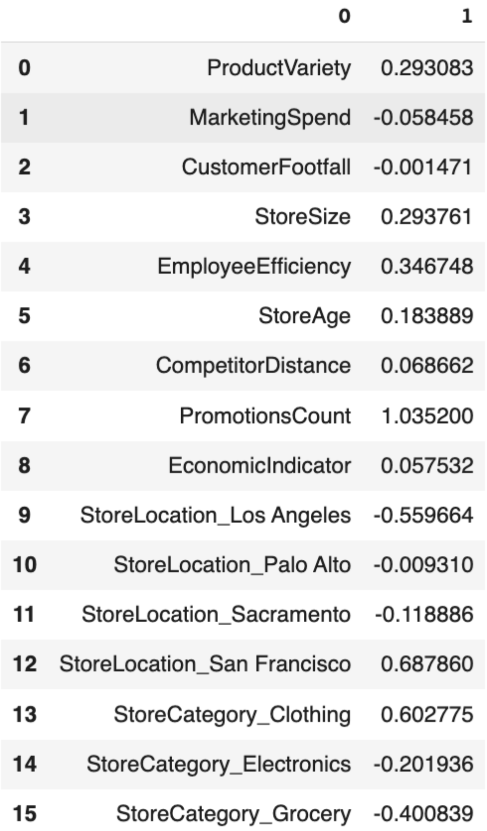
- For the R^2, the model explains 80.9% of the variance in Monthly Sales Revenue, indicating a strong fit to the data.
- The F-statistic shows that overall model is statistically significant.
- The condition number of 1.41e+19 is quite high, signaling multicollinearity among predictors. This can make some coefficients unstable and less reliable.

In terms of model validity, the strengths are its high R^2 and adjusted R^2, they show the model fits the data well. However, the big weaknesses is its multicollinearity. This means there is a high correlation among predictors, that reduces confidence in individual coefficient estimates. Some predictors like MarketingSpend, PromotionsCount) are unexpectedly insignificant, which can be caused by either poor feature quality or multicollinearity masking their impact.

Linear Regression Model with sklearn

First, I built a baseline model to predict all the sales based on average sales. Based on this, the MSE is 4293. This will be the base line that the other models will be compare to.

Then, I ran the multivariate linear regression. From the coefficients, we can see: PromotionsCount, EmployeeEfficiency, ProductVariety, StoreSize, StoreAge has relatively higher numbers. Also, from store location's standpoint, San Francisco is better as it has a higher and positive coefficient. This might be due to how fierce the competition level is in the cities. And, from store type perspective, clothing stores have higher coefficients, while grocery stores have lower. These findings are somehow different from what we observed purely from the scatter plots and bar plots. This regression is giving us more complete and accurate information. For example, from scatter plot, we did not see big difference or strong correlation between promotions count with monthly sales revenue, but this regression model tells us that the coefficient is relatively high. Same with employee efficiency and store age.

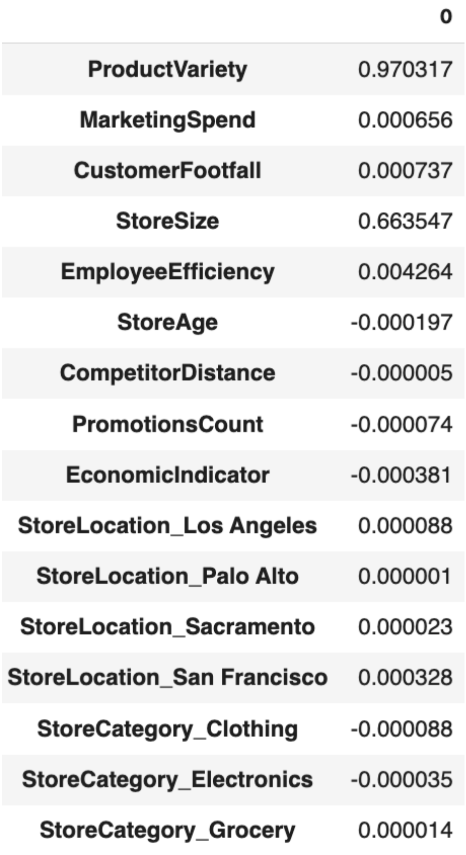


	0	1
0	ProductVariety	0.293083
1	MarketingSpend	-0.058458
2	CustomerFootfall	-0.001471
3	StoreSize	0.293761
4	EmployeeEfficiency	0.346748
5	StoreAge	0.183889
6	CompetitorDistance	0.068662
7	PromotionsCount	1.035200
8	EconomicIndicator	0.057532
9	StoreLocation_Los Angeles	-0.559664
10	StoreLocation_Palo Alto	-0.009310
11	StoreLocation_Sacramento	-0.118886
12	StoreLocation_San Francisco	0.687860
13	StoreCategory_Clothing	0.602775
14	StoreCategory_Electronics	-0.201936
15	StoreCategory_Grocery	-0.400839

I also evaluated the model score and MSE. The MSE is 808 on training dataset and 848 on test data set. This is very much improved vs. our baseline model, which had a MSE of 4293. These test and training MSE numbers are quite close, which means the model performed almost equally well on training dataset and test dataset. There is not much over-fitting or under-fitting issue. The R^2 score of 0.813 shows the model captures most of the variability in the data, making it a reliable predictive model for Monthly Sales Revenue. To further improve, one might consider adding interaction terms, non-linear transformations, or more features, to capture potential non-linear relationships.

To summarize, I looked at the importance of each factor:

- Dominance of ProductVariety and StoreSize: The results highlight that these two features are crucial drivers of sales, as they strongly influence the model's accuracy.
- Low Impact of External Factors: Features like CompetitorDistance, EconomicIndicator, and even MarketingSpend contribute minimally, suggesting internal store operations (e.g., variety and size) are more critical than external factors.
- Actionable Focus: Businesses should focus on increasing product variety and optimizing store size to improve sales. Less emphasis might be placed on external competition or promotional efforts unless coupled with other strategies.



	0
ProductVariety	0.970317
MarketingSpend	0.000656
CustomerFootfall	0.000737
StoreSize	0.663547
EmployeeEfficiency	0.004264
StoreAge	-0.000197
CompetitorDistance	-0.000005
PromotionsCount	-0.000074
EconomicIndicator	-0.000381
StoreLocation_Los Angeles	0.000088
StoreLocation_Palo Alto	0.000001
StoreLocation_Sacramento	0.000023
StoreLocation_San Francisco	0.000328
StoreCategory_Clothing	-0.000088
StoreCategory_Electronics	-0.000035
StoreCategory_Grocery	0.000014

KNN Regressor

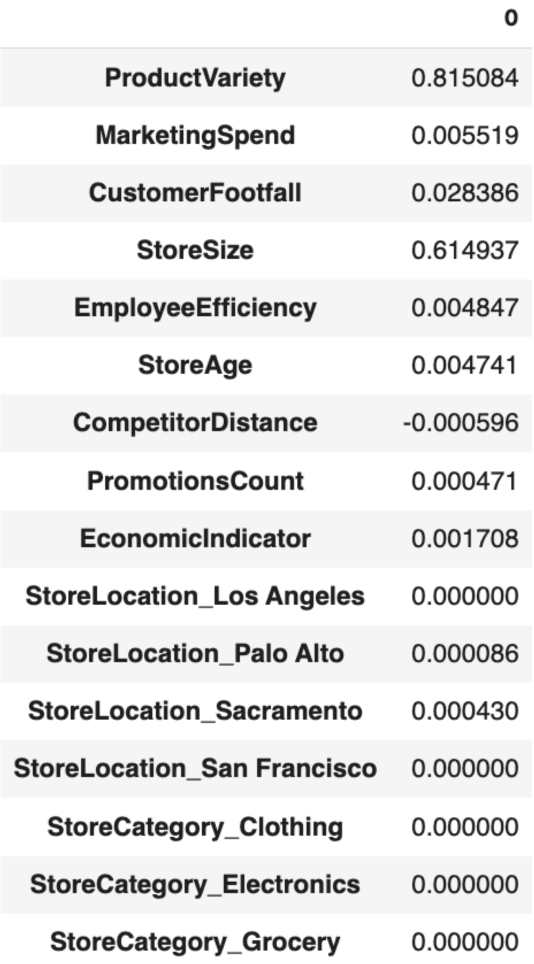
I also used KNN regressor to see how well the model is compared to the linear one.

The training MSE is 697.35 and testing MSE is 1074.44. The training MSE is lower than the linear regressor in sklearn. However, the testing dataset's MSE is a bit high. This might mean the regressor built a model that "over-fit" model based on the training set, when it applies to test set, it does not fit very well. As a result, not surprisingly, the model score is 0.76, the KNN model explains 76.29% of the variance in the target variable for the test set. This is not a bad score, but it is relatively lower than the sklearn linear regression model. It indicates that the model is reasonably effective but less robust than linear regression for this dataset.

I tried to improve the parameters in KNN model to improve the performance. I used the GridSearchCV process and optimized the KNN model. The final R^2 score of 0.761 on the test set shows that the optimized KNN model captures approximately 76.1% of the variance in the target variable. While this is a solid result, the performance is slightly lower than the linear regression model, suggesting that linear relationships dominate the dataset.

To summarize, I looked at the importance of the factors again.

- Dominant Features: The two dominant predictors are ProductVariety and StoreSize, which collectively explain most of the variation in sales revenue. This is in line with what we found out earlier.
- Minimal Impact Features: Factors such as Marketing Spend, Promotions Count, and Employee Efficiency appear to have minimal influence on sales. This suggests these features may not directly drive revenue, or their impact could be masked by other variables.
- Irrelevant Features: Features like Competitor Distance, Store Locations, and Store Categories were deemed unimportant, indicating that sales are not significantly affected by store geography or category in this dataset.



	0
ProductVariety	0.815084
MarketingSpend	0.005519
CustomerFootfall	0.028386
StoreSize	0.614937
EmployeeEfficiency	0.004847
StoreAge	0.004741
CompetitorDistance	-0.000596
PromotionsCount	0.000471
EconomicIndicator	0.001708
StoreLocation_Los Angeles	0.000000
StoreLocation_Palo Alto	0.000086
StoreLocation_Sacramento	0.000430
StoreLocation_San Francisco	0.000000
StoreCategory_Clothing	0.000000
StoreCategory_Electronics	0.000000
StoreCategory_Grocery	0.000000

Results and Model Comparison

Above, I have done a study using different regression models: statsmodel OLS model, sklearn linear regression, KNN model to develop a prediction model to understand the relationship between the different factors and monthly sales revenue. The results slightly differ, but mostly agree that store size and product variety are the most important factors.

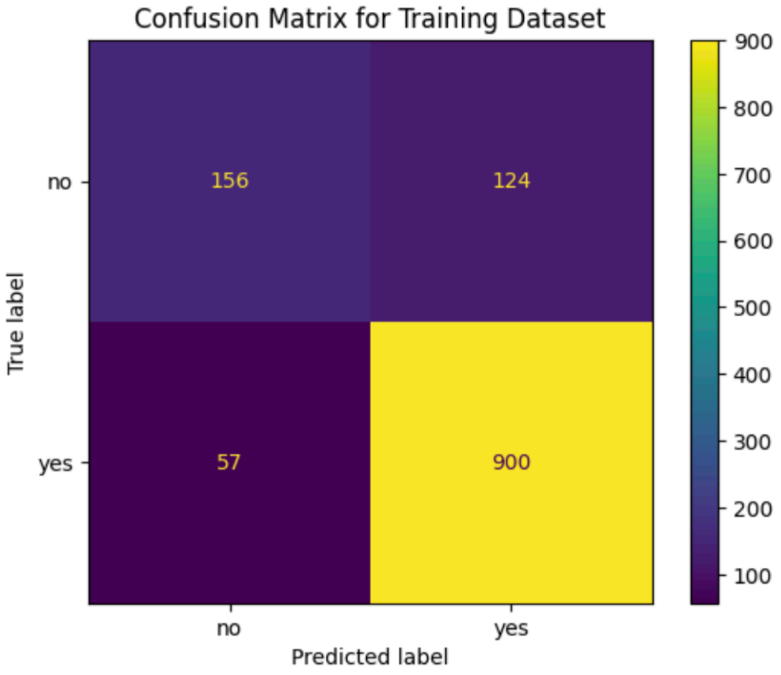
In comparison, in terms of performance, linear models (StatsModels OLS and Sklearn Linear Regression) outperform KNN on this dataset, achieving higher R^2and lower MSE. In terms of interpretability, stats model OLS excels in interpretability due to statistical summaries, helping identify significant predictors. In terms of handling complexity, KNN is more flexible for capturing non-linear relationships. However, it underperforms here because the data shows strong linear trends.

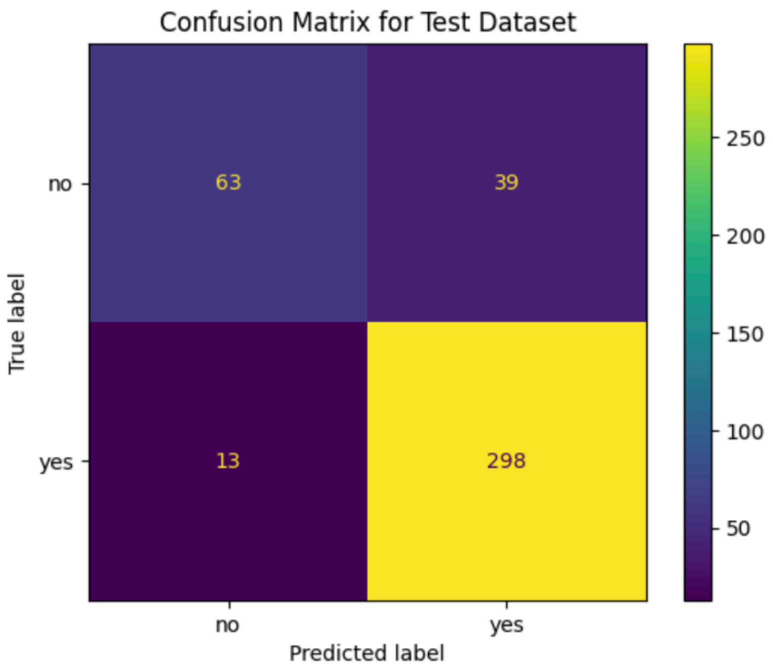
Extension: Make It a Logistic Regression Problem

Assume in our survey of store data, we can only know whether the store is making profit or not, and store owners don't want to share the exact numbers of their monthly sales revenue. For practice reason, I will try make it a logistic regression problem. The problem is: if I invest a new store (or buy a existing store), will I make profit or not?

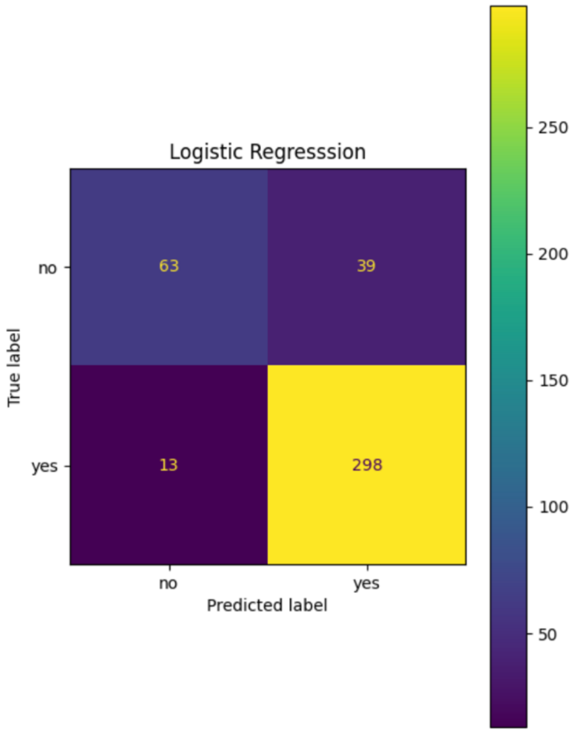
By checking the range and average of the data, I assumed that the breakeven line for monthly sales is 250. I then made the 'MonthlySalesRevenue' a binary data. Following, I trained logistic regression model on the training and evaluated using accuracy scores for both training and test sets.

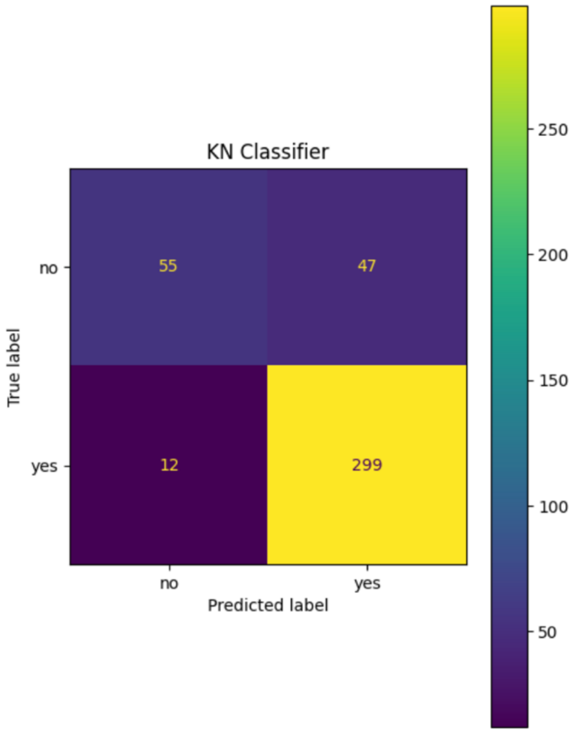
The prediction model works well with a score of 0.85. And, both training data score and testing data set score are very close to each other, the test data score is even slightly higher. This shows there is probably no "over-fitting" concern. I also graphed the confusion matrices of training and test datasets. We can also tell the confusion matrix of the model for both training dataset and test dataset. From the data, we can see again, the prediction accuracy is pretty good. The model correctly predicts a high proportion of positive cases (profit stores) while making fewer errors on negative cases. The accuracy is robust across training and test datasets.





I did another KN Classifier model on the same dataset to compare the performance of the two different models. The model score of test and train are both 0.86, which were pretty good. The KNN model's performance is very close to logistic regression, with a slightly lower test score. The KNN model slightly underperforms compared to logistic regression. It has a slightly higher false positive rate (47 vs. 40). The difference is not big, though.





Results and Model Comparison

Logistic regression achieves slightly better accuracy and fewer false positives, making it preferable for this scenario due to its simplicity and interpretability. KNN performs comparably but requires hyperparameter tuning to optimize performance.

Conclusion and Next Step

Conclusion

In this analysis, I examined factors influencing store sales to help small business owners, such as grocery store owners or supermarket managers, make informed decisions to increase revenue. By using predictive models including Linear Regression, KNN, and Logistic Regression, the key findings are as follows:

1. Product Variety and Store Size emerged as the most significant predictors of higher monthly sales revenue. This highlights the importance of offering a diverse product range and operating larger stores to attract customers.
2. Employee Efficiency and Promotions Count also contributed positively to sales performance, suggesting that improving staff productivity and running effective promotional campaigns can drive additional revenue.
3. Competitor Distance and Marketing Spend showed relatively minor impacts on sales. This implies that while competitor proximity matters, it does not dominate sales outcomes, and marketing spend needs more strategic allocation to ensure effectiveness.

Key Recommendations

- Focus on increasing **product variety** and expanding **store size** where feasible to boost customer attraction and retention.
- Improve **employee efficiency** through training programs and incentives.

- Design and implement well-planned **promotional strategies** to maximize impact on revenue.
- Analyze marketing spend further to ensure it delivers a positive return on investment.

Next Steps

To make the analysis more useful and practical, we can consider:

1. More detailed sales revenue data. The current monthly sales revenue data is most likely an average data. If we could get more about each month, and we can try to understand the dynamics of each factor influencing each other, this could also help us to make more thorough thought decision.
2. Consider other factors. In our analysis, we only found 2 factors were quite influential. One way to make this analysis more accurate is to collect more data, so we can have a more unbiased dataset. We can also consider more factors, such as: hospitality of store owner, opening hours, and price of products in the store etc.
3. Incorporate external data. We can include regional economic indicators like GDP, inflation rate, and unemployment rate to understand how macroeconomic conditions impact store sales. Use data on population density, household income, age distribution, and customer preferences to refine location-based recommendations. A more detailed analysis with time series and data throughout time can be done to draw more useful conclusions.