Robert Hall, 12/04/2025
STA 4713, Dr. Dae Young Ham

Analyzing Predictors of Coffee Shop Daily Revenue Using Multiple Linear Regression

Kaggle Dataset Link | GitHub Code Link | Report Google Doc Link

Methods learned in class: multiple linear regression, residuals analysis for normality check
Methods from outside class: homoskedasticity check (Breusch-Pagan), k-Fold Cross Validation

1. Introduction

This project seeks to use multiple linear regression to determine the strength and significance of multiple predictors for a coffee shop's daily revenue. Data is publicly available on Kaggle.

The data set contains 2000 rows of data from coffee shops, and is designed for predictive analytics and business optimization, according to the Kaggle description.

2. Building the Multiple Linear Regression Model

All six predictors given in the data set were used in the multiple linear regression model:

- Number of customers visiting the coffee shop per day
- Average order value ($USD)
- Operating hours per day
- Number of employees
- Marketing and promotions spending per day
- Location foot traffic

OLS Regression Results

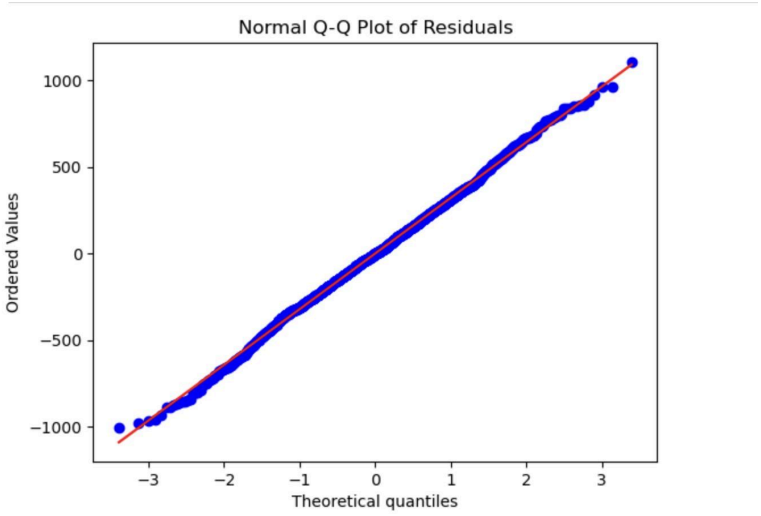
Dep. Variable:	Daily_Revenue	R-squared:	0.892
Model:	OLS	Adj. R-squared:	0.891
Method:	Least Squares	F-statistic:	2732.
Date:	Thu, 04 Dec 2025	Prob (F-statistic):	0.00
Time:	21:58:22	Log-Likelihood:	-14383.
No. Observations:	2000	AIC:	2.878e+04
Df Residuals:	1993	BIC:	2.882e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1517.9327	43.768	-34.681	0.000	-1603.768	-1432.097
Number_of_Customers_Per_Day	5.5680	0.056	100.051	0.000	5.459	5.677
Average_Order_Value	243.2842	3.311	73.469	0.000	236.790	249.778
Operating_Hours_Per_Day	-0.2338	2.096	-0.112	0.911	-4.344	3.876
Number_of_Employees	-2.3213	1.927	-1.204	0.229	-6.101	1.459
Marketing_Spend_Per_Day	1.5555	0.051	30.455	0.000	1.455	1.656
Location_Foot_Traffic	0.0242	0.027	0.913	0.361	-0.028	0.076

Omnibus:	1.534	Durbin-Watson:	1.983
Prob(Omnibus):	0.464	Jarque-Bera (JB):	1.503
Skew:	-0.008	Prob(JB):	0.472
Kurtosis:	3.133	Cond. No.	4.23e+03

3. Checking Normality Assumptions using Q-Q Plot

The Normal Q-Q plot of residuals demonstrates strong and clear normality:



4. Checking for Heteroskedasticity using Breusch-Pagan

Breusch-Pagan was chosen over other common methods for homoskedasticity checking because, since the predictors scale upward alongside business activity, heteroskedasticity would be systematically tied to the regressor variables.

Breusch-Pagan tests whether error variance is a function of the predictor variables. This proves an advantage over White's test, which is an omnibus test for detecting general forms of homoskedasticity, including non-linear effects.

The hypotheses of the Breusch-Pagan test are structured as such:

H_0: $Var(\varepsilon_i) = \sigma^2$

H_a: $Var(\varepsilon_i) = \sigma^2 \bullet h(X_i)$ for unknown variance function $h(X_i)$ linked to observation i

The results of the Breusch-Pagan test indicate that we fail to reject H_0 at the 5% level of significance, and conclude that the assumption of homoskedasticity holds:

	Statistic	Test Statistic	p-value
0	LM Test	4.349572	0.629482
1	F Test	0.723966	0.630309

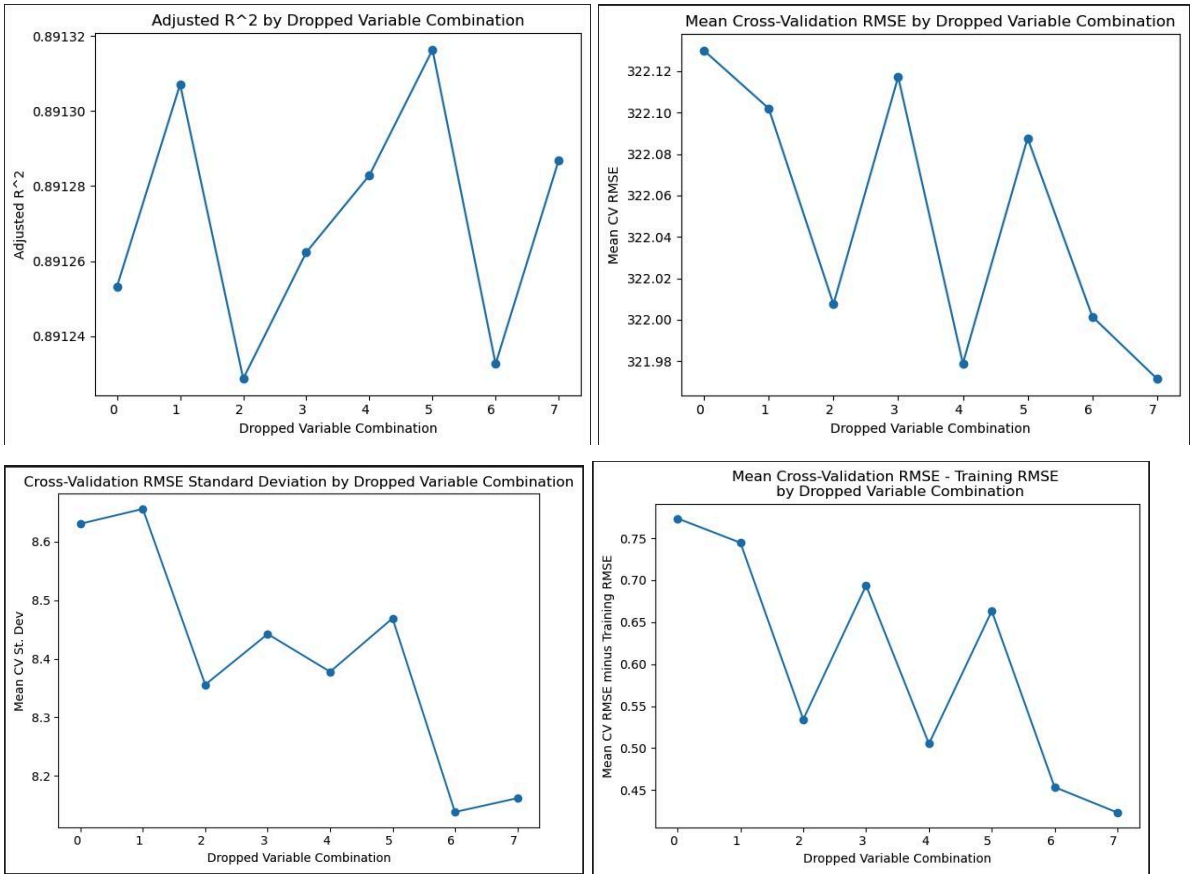
5. Testing Whether Omitting Non-Significant Features Greatly Affects Model Accuracy

Three features were found to be non-significant – Location foot traffic, Number of employees, and Operating hours per day. The model was tested iteratively by removing one of each non-significant variable, then combinations of two variables, then removing all three, to see which combination of variable omission produced the best results, if any combination at all.

5-Fold Cross Validation was used to evaluate model predictions, generalization, and reliability through measuring (respectively) the mean CV root mean-squared error (RMSE), the gap between mean CV RMSE and training RMSE (as a measure of overfitting), and CV RMSE standard deviation. Adjusted R^2 was also taken into consideration, as a measurement of model accuracy after complexity penalization applied.

Ideally, the mean CV RMSE, CV RMSE standard deviation, and the gap between mean CV RMSE and training RMSE, should all be low, while the Adjusted-R^2 should be relatively high.

The following results were yielded from cross-validation:



It appears from the results that mean CV RMSE, CV RMSE standard deviation, and the gap between mean CV RMSE and training RMSE, all dropped to their lowest after the three non-significant variables were omitted. Variability in Adjusted-R^2 was marginal, but was reasonably high on a relative basis with the other combinations, such that removing all three non-significant variables seemed acceptable under all four conditions.

It is also likely prudent to use a simpler model rather than a more complex one, if it does not gravely sacrifice accuracy and predictive authority, so that small coffee shop owners with little or no quantitative backgrounds can easily interpret the results. It does not appear that removing the three non-significant features of interest will do substantial damage to the integrity of the model.

Dropped variable combinations are matched with ordinal numerical keys as such, with 7 indicating the model when omitting all three non-significant terms:

	dropped_variables
0	None (full model)
1	Operating_Hours_Per_Day
2	Number_of_Employees
3	Location_Foot_Traffic
4	Operating_Hours_Per_Day, Number_of_Employees
5	Operating_Hours_Per_Day, Location_Foot_Traffic
6	Number_of_Employees, Location_Foot_Traffic
7	Operating_Hours_Per_Day, Number_of_Employees, ...

6. Simplified Model

The simplified model was constructed after removing the three non-significant features:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Daily_Revenue   R-squared:                       0.891
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     5464.
Date:                Thu, 04 Dec 2025   Prob (F-statistic):               0.00
Time:                        21:58:23   Log-Likelihood:                -14384.
No. Observations:                2000   AIC:                         2.878e+04
Df Residuals:                    1996   BIC:                         2.880e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    -1525.6781     29.477    -51.758      0.000   -1583.487   -1467.869
Number_of_Customers_Per_Day      5.5680      0.056    100.066      0.000       5.459       5.677
Average_Order_Value            243.3025      3.310     73.514      0.000     236.812     249.793
Marketing_Spend_Per_Day          1.5532      0.051     30.433      0.000       1.453       1.653
==============================================================================
Omnibus:                        1.578   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.454   Jarque-Bera (JB):                1.553
Skew:                          -0.006   Prob(JB):                        0.460
Kurtosis:                       3.136   Cond. No.                     1.63e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

7. TLDR (Findings & Conclusions)

The best predictors of daily revenue are, in descending order:

- How many customers the coffee shop is getting per day
- How much, on average, people are paying for their order ($$)
- The shop's daily marketing and promotions spending

An increase in 1…	**…brings an average of ___ in daily revenue!**
Customer	+$5.57
Dollar of Average Order Payment ($$)	+$243.30
Dollar of Marketing & Promotions Spending ($$)	+$1.55

8. Sources

Data:

https://www.kaggle.com/datasets/himelsarder/coffee-shop-daily-revenue-prediction-dataset?utm_source=chatgpt.com