MSDS601 Final Report: Restaurant Revenue Regression Analysis

By: Katelyn Vuong, Serigne Diaw, Lukas Amare, Ricky Miura

I. Data Introduction

For this regression analysis, we used the Restaurant Revenue dataset from Kaggle. This dataset includes information from 8,368 different restaurants that are distinguished by 16 different attributes (categorical and numerical) such as Cuisine, Rating, Seating Capacity, Service Quality Score, etc. Our main objective behind looking at restaurant data is to find what features are the most impactful and are the best predictors of restaurant revenue for our predictive model. The goal with analyzing this data is to gain a deeper understanding of how the restaurant market works and to potentially find the key factors that drive revenue growth, which could help restaurant owners make data-informed decisions to optimize their profit and overall success. The objectives we focused on addressing with different data analysis methods are:

- What are the strongest factors of restaurant revenue?
- How do the effects of these strong predictors compare to one another?
- How does the model's predictive power change when including/excluding potential key features?
- Can we identify any non-linear relationships between the predictor variables and restaurant revenue?

II. Methodologies

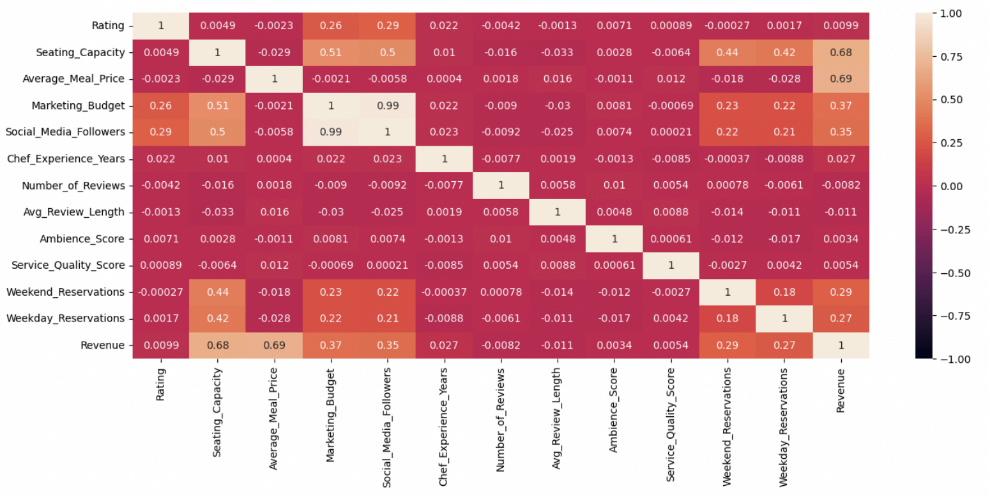
In our analysis, we employed several methods within the different stages of our analysis. During EDA, we visualized key features to identify potential predictors, observed pairwise correlations, and checked for any data imbalances. In the SLR phase, we explored single predictors and addressed issues such as heteroskedasticity through log transformations. For MLR, we also assessed heteroskedasticity through log transformations, carefully selected a subset of predictors and validated models based on various criteria such as AIC, BIC, and PRESS.

Throughout our analysis, we performed diagnostic tests to verify that our models met linear regression assumptions with a focus on normality of residuals, multicollinearity, heteroskedasticity, and the impact of outliers and influential points. We performed any necessary transformations, evaluated VIF to assess multicollinearity, and analyzed influential points to determine their effects on our model. We ultimately wanted to select a model that we thought best limited these violations and balanced model simplicity with predictive accuracy.

III. Exploratory Data Analysis

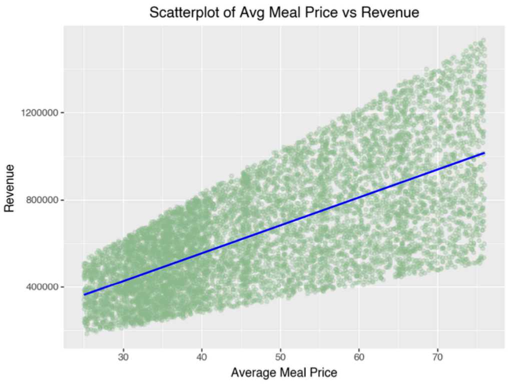
We looked at the distribution of potential predictors to see obvious signs of outliers, explored categorical features to see if distributions were balanced, and checked for nulls. From our findings, categorical features were evenly distributed, there were no obvious outliers across predictors, and there were no nulls in the dataset.

We first created a heatmap for a large scaled correlation matrix to identify any strong correlations to our target variable and any potential multicollinearity between features. The strongest correlations we found with revenue were with **average meal price** (0.69), **seating capacity** (0.68), **marketing budget** (0.37), and **social media followers** (0.35). The largest coefficients that could allude to multicollinearity is between seating capacity and marketing budget (0.51) and seating capacity with weekend reservations (0.44).



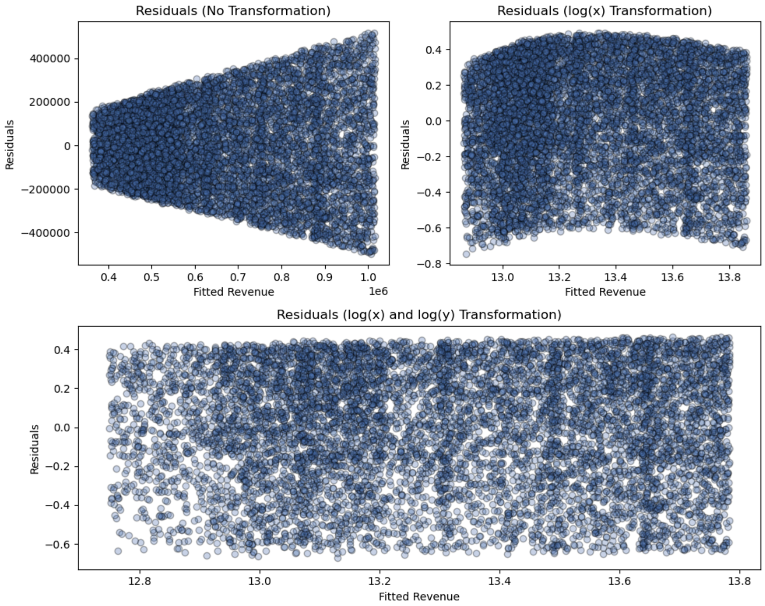
	Rating	Seating_Capacity	Average_Meal_Price	Marketing_Budget	Social_Media_Followers	Chef_Experience_Years	Number_of_Reviews	Avg_Review_Length	Ambience_Score	Service_Quality_Score	Weekend_Reservations	Weekday_Reservations	Revenue
Rating	1	0.0049	-0.0023	0.26	0.29	0.022	-0.0042	-0.0013	0.0071	0.00089	-0.00027	0.0017	0.0099
Seating_Capacity	0.0049	1	-0.029	0.51	0.5	0.01	-0.016	-0.033	0.0028	-0.0064	0.44	0.42	0.68
Average_Meal_Price	-0.0023	-0.029	1	-0.0021	-0.0058	0.0004	0.0018	0.016	-0.0011	0.012	-0.018	-0.028	0.69
Marketing_Budget	0.26	0.51	-0.0021	1	0.99	0.022	-0.009	-0.03	0.0081	-0.00069	0.23	0.22	0.37
Social_Media_Followers	0.29	0.5	-0.0058	0.99	1	0.023	-0.0092	-0.025	0.0074	0.00021	0.22	0.21	0.35
Chef_Experience_Years	0.022	0.01	0.0004	0.022	0.023	1	-0.0077	0.0019	-0.0013	-0.0085	-0.00037	-0.0088	0.027
Number_of_Reviews	-0.0042	-0.016	0.0018	-0.009	-0.0092	-0.0077	1	0.0058	0.01	0.0054	0.00078	-0.0061	-0.0082
Avg_Review_Length	-0.0013	-0.033	0.016	-0.03	-0.025	0.0019	0.0058	1	0.0048	0.0088	-0.014	-0.011	-0.011
Ambience_Score	0.0071	0.0028	-0.0011	0.0081	0.0074	-0.0013	0.01	0.0048	1	0.00061	-0.012	-0.017	0.0034
Service_Quality_Score	0.00089	-0.0064	0.012	-0.00069	0.00021	-0.0085	0.0054	0.0088	0.00061	1	-0.0027	0.0042	0.0054
Weekend_Reservations	-0.00027	0.44	-0.018	0.23	0.22	-0.00037	0.00078	-0.014	-0.012	-0.0027	1	0.18	0.29
Weekday_Reservations	0.0017	0.42	-0.028	0.22	0.21	-0.0088	-0.0061	-0.011	-0.017	0.0042	0.18	1	0.27
Revenue	0.0099	0.68	0.69	0.37	0.35	0.027	-0.0082	-0.011	0.0034	0.0054	0.29	0.27	1

Since average meal price had the highest correlation with revenue, we decided to explore this relationship further. We checked to see if average meal price and revenue had a linear relationship to conclude whether we should even be using linear regression as a model. The scatter plot shows a pretty linear relationship so we fit a SLR model using average meal price as a predictor. Before we proceed any further, we needed to make sure we were not violating any model assumptions before deciding on a final model.


Scatterplot of Avg Meal Price vs Revenue

IV. SLR Diagnostics & Assessment

We checked for heteroskedasticity with a residual plot against our fitted revenue values and saw obvious heteroskedasticity with the points spreading out as revenue gets higher. This showed that our variance was not constant. We attempted to fix this by performing a log transformation onto our predictor value and visualized our residual scatterplot again, where we saw a much better looking residual plot with more even spread. We then log transformed the Y and log transformed the X and saw a very good linearly regressed plot and graphed the residual scatter plot with both axes transformed. We found a much more homoscedastic residual plot, with points scattered above and below evenly.

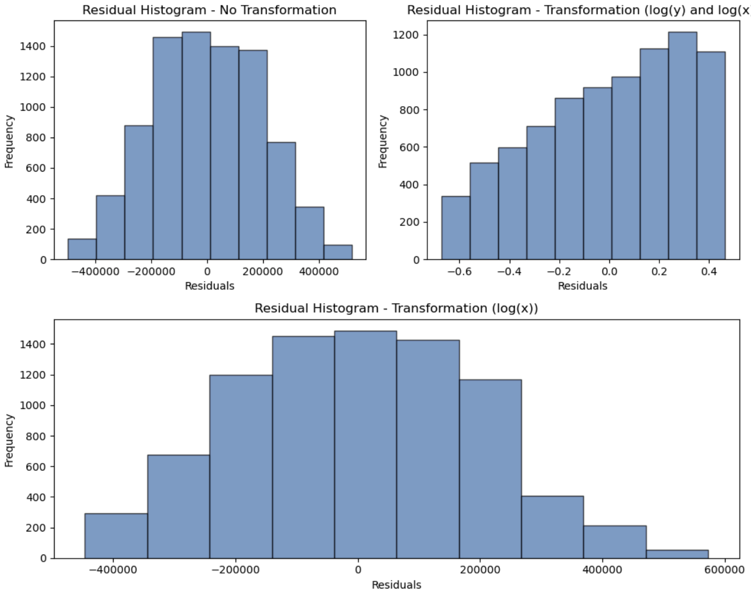

Residuals (No Transformation)

Residuals (log(x) Transformation)

Residuals (log(x) and log(y) Transformation)

Before we made our transformations we observed normality in our residual histogram plot, however, after making our transformations to fix heteroskedasticity we saw our residuals on a histogram show a skewed distribution. This was disheartening because in our attempt to fix heteroskedasticity, we created a new problem with normality.

We then graphed 4 different residual histogram plots, Y~X, Y~log(X), log(Y)~X, and log(Y)~log(X) and realized that logging Y was causing the change in normality in our function and making the residual distribution right-skewed. To compromise, we kept Y untouched and log transformed only the X to still remove most of the heteroskedasticity but maintain normality in the model.



In regards to our variables being independent, after we transformed X we did not see much of a pattern, which can allude to independence between observations. Therefore we will assume that the observations are independent.

The t-test for the individual predictor had a p-value of 0, meaning we can reject the null hypothesis for this predictor and conclude that log(X) average meal price is a strong predictor individually of Y (revenue). When looking at the R^2 adjusted for our SLR, we can see that it is 0.465 which is a solid R^2 value for one predictor variable. Now we want to see if we can get a better model from this with multiple predictors.

V. Multiple Linear Regression

It’s more realistic that we perform multiple regression analysis since a combination of predictors is most likely to reveal the truest relationship with revenue. We chose the following predictors to consider for our multiple linear regression based on our EDA: seating capacity, average meal price, marketing budget, social media followers, number of reviews, parking availability, and location.

To explore a little further, we looked to see whether or not the inclusion of some of our categorical variables would make any significant difference. We used location and parking availability and found that when we include these categorical variables and the numerical variables we mentioned right above, the Adjusted R^2 went from 0.9575 to 0.9576 when comparing that to the model without the categorical variables. While there is an increase, we believe this marginal difference does not outweigh the cost of having to add more dummy variables potentially exposing us to the curse of dimensionality. After further thought, we decided to leave out these categorical variables and stick with the predictors mentioned earlier.

VI. MLR Model Selection

Selecting the appropriate model is an important step in the process of fitting a multiple linear regression model, as it directly influences both the accuracy of predictions and the interpretability of results. It is important to find a balance between choosing an underspecified model that is missing important predictors, and a model with extraneous or redundant variables. In this section, we will split our data into training and validation sets, and perform best subsets regression to determine the model that meets an objective criterion of choice.

First, we split our dataset into training and validation sets. The purpose of this is to prevent overfitting, and ensure that our model generalizes to data that it hasn't yet seen. We use the training set to fit our models, and the validation set to select a model that meets our criterion. We chose to have 80% of our data to fall into the training set, while 20% is used for validation.

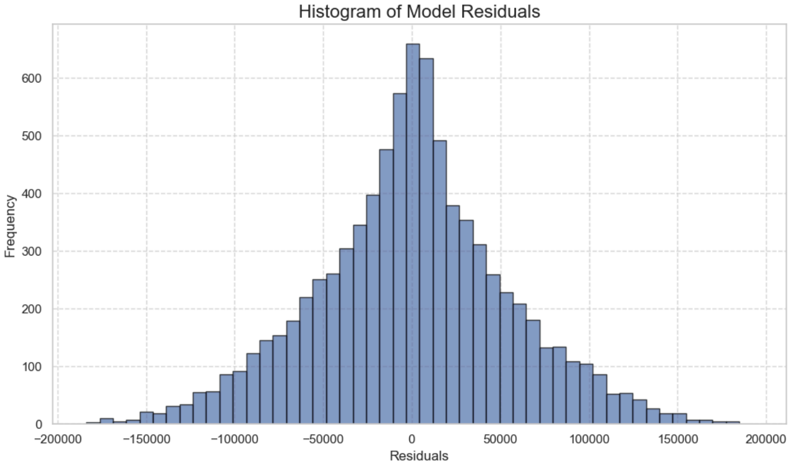
Next, we performed best subsets regression to select a subset of our predictors that meet our criterion. We fit a model for every combination of our seven predictors on our training set, resulting in 127 $(2^7 - 1)$ total models to validate. For each of these models, we calculated the following metrics: Mallow's C_p , Akaike's Information Criterion (AIC), Schwarz's Bayesian information criterion (BIC), Prediction Sum of Squared Error (PRESS), R^2, and Adjusted R^2.

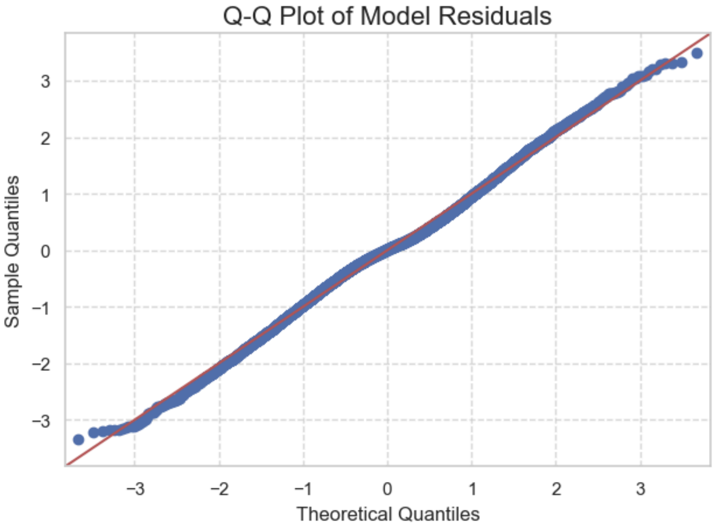
These metrics were calculated with our validation set to ensure that we are choosing a model that generalizes to data it was not trained on. As we are interested in the effects that specific predictors have on restaurant revenue, we chose the model that was selected by BIC because BIC helps with the inference and interpretability of our model. Due to the fact that BIC penalizes model complexity, even more so than AIC, we were hopeful that this would limit multicollinearity between our predictors. It is likely that social media followers and marketing budget are correlated, but marketing budget was not selected by BIC. This is preferred since multicollinearity would adversely affect our ability to look at the effects of specific predictors on revenue. The model we selected includes the predictors: seating capacity, average meal price, and social media followers. Now that we have selected a model, we can proceed to diagnostic measures designed to test it further.

VII. MLR Diagnostics

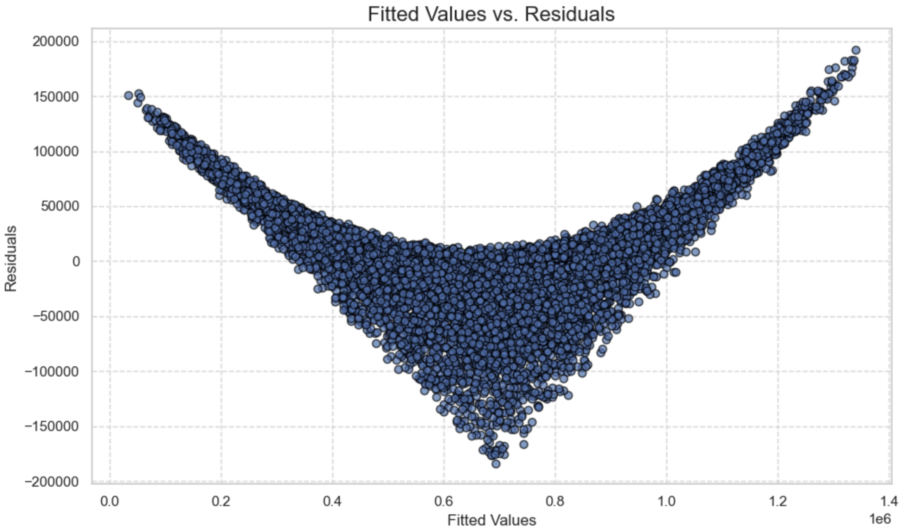
In our next phase, it is essential to conduct a diagnostic analysis to ensure that our model and its assumptions are valid. This ensures that we identify issues that could lead to unreliable results and inference. Like what we did in our SLR, we will focus on several key tests, including those for heteroskedasticity, outliers, leverage, and influential points.

Our first step is to ensure that our model residuals are normally distributed, a key classical assumption for linear regression. This will allow us to perform subsequent inference, including hypothesis tests and constructing confidence intervals. First, we plot a histogram and QQ plot of the residuals.



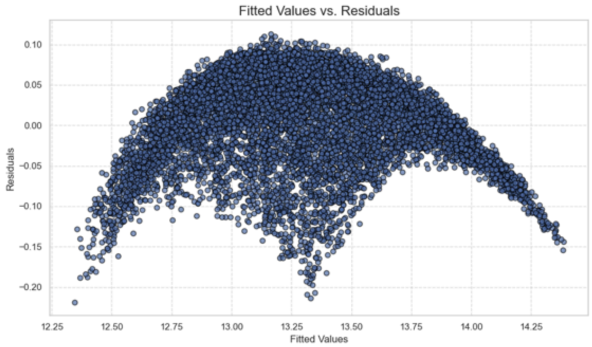


It appears that the residuals are normally distributed. Our histogram follows the normal distribution, and the QQ plot is relatively linear. We see slight heavy tails which is not a major concern, but may suggest that the significance of our subsequent tests may not be completely reliable. Next, we will plot our fitted values against our residuals, to infer whether our model suffers from heteroskedasticity.

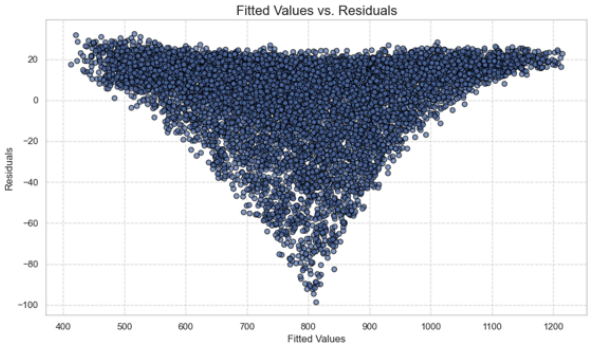


Unfortunately, we see that our residuals show heteroskedasticity. The variance of residuals in our model is not constant. It doesn't initially appear that this plot follows any known shape, such as a logarithmic or quadratic one. However, we will still try to transform some of our variables to see if we can fix this issue.

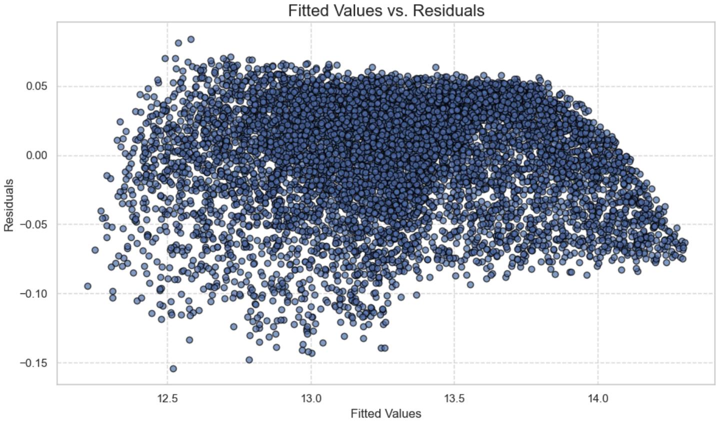
We will start by transforming our response variable, revenue, to stabilize the variance. We take the log and square root of revenue respectively, and re-fit our model. These two transformations still do not fix our heteroskedasticity. Therefore, we'll also try to transform a selection of our predictors. We choose to take the log of both the seating capacity and social media followers variables, in addition to revenue.



log(revenue) ~ seating capacity + social media followers + average meal price

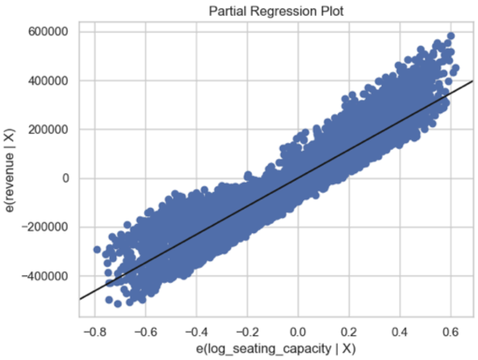


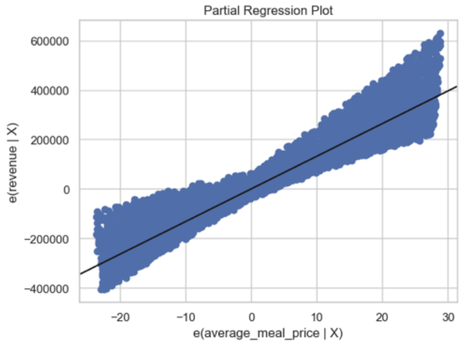
sqrt(revenue) ~ seating capacity + social media followers + average meal price

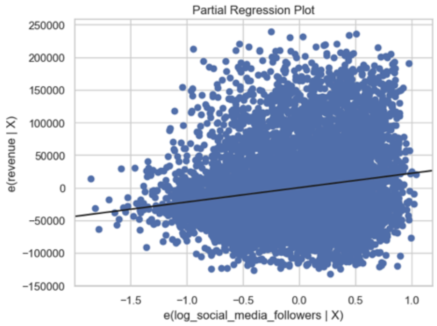


log(revenue) ~ log(seating capacity) + log(social media followers) + average meal price

Thankfully, this transformation of our predictors results in a much less variable residuals plot. Our residuals are still not completely constant, but are much improved. Finally, we will plot the relationship between each predictor and revenue, to test if their relationship is linear. We will visualize partial regression plots for each predictor, showing their relationship with the response variable while keeping the other predictors constant.







After plotting the relationship between our different predictors and revenue, we see that both seating capacity and average meal price show a clear linear relationship. However, social media followers do not. The points stray much farther above the line than below. However, we attempted to address this by applying several different transformations to social media followers, to no avail. Therefore, we will leave the model as it is and hope that it does not affect our tests too much.

It is also important to check if any of our predictors are correlated. In order to test this, we will calculate a variance inflation factor (VIF) for each predictor. The VIF for average meal price is extremely close to 1, indicating almost no multicollinearity. The other two predictors, seating capacity and social media followers both have a VIF of about 1.33.

Features	VIF Factor
Seating Capacity	1.329080
Average Meal Price	1.000928
Social Media Followers	1.328021

Finally, we want to look at leverage, outliers, and influential points. We will try to determine if we have any problematic data points and then make a decision about whether to remove them. To start with leverage, our measure of outliers with respect to our predictors, we calculate a statistic for each point and note if it is above $2p/n$. We find that 1197 of our 8368 data points have high leverage. Next, we will threshold our externally studentized residuals, noting if they are larger than the critical value of the t-distribution with alpha level $.05$ and $n - p - 1$ degrees of freedom. We find that 355 points are outliers with respect to revenue. Finally, we calculate Cook's distance to find influential points, thresholding at a level of $4/n$. We find that there are 621 influential points.

There are quite a few of each, so we wanted to try dropping all of these points and refitting our model to see if there is a big change in our coefficients and statistics. After fitting the new model, we find that our coefficients change very minimally. Our BIC measure even increased, indicating that this model is not preferred. Since we do not have any objective reason to delete these data points, and our summary analyses did not change much, we will choose to stick with our original model.

VIII. Conclusion

In conclusion, our regression analysis of the Restaurant Revenue dataset provided valuable insights into the factors influencing restaurant revenue and allowed us to find a strong, predictive model with practical applications. By examining both single and multiple regression models, we identified key predictors, such as average meal price, seating capacity, and social media followers, that may have significant impacts on revenue.

After performing model selection using best subsets regression and comparing several metrics like AIC, BIC, PRESS, R^2, and Adjusted R^2 ,we found our strongest model to be with the predictors **seating capacity**, **average meal price**, and **social media followers.** We looked into their VIF factors and found seating capacity and social media followers to be slightly higher, indicating they are very slightly collinear, which makes sense since larger businesses will likely have greater capacity and be more popular on social media. However, this VIF is on the lower end of the moderate correlation range, so it does not worry us too much.

We also performed diagnostics in both the SLR and MLR case to address issues of heteroskedasticity, multicollinearity, dependence of observations, as well as analyze the effects of leverage, outliers, and influential points. We tested various log transformations to meet linear regression assumptions and minimize violations, ultimately leading to a model that balances simplicity with accuracy.

A small issue we ran into assessing non-linear relationships arose between revenue and social media followers when we looked at their scatterplot and did not find too much success with transforming the data. However, we don’t believe it affected our model selection process by a concerning amount.

Overall, our findings were carefully considered and trialed with multiple practical explanations for factors that could influence revenue. In the end we were left with the best model that we think could be helpful for restaurant owners to understand for optimizing profitability and to support strategic planning within the competitive restaurant industry.