Miriam Garcia

March 17, 2021

University of North Texas

ECON 5645

Model Building & Inference in Regression for Buster's Brewhaus

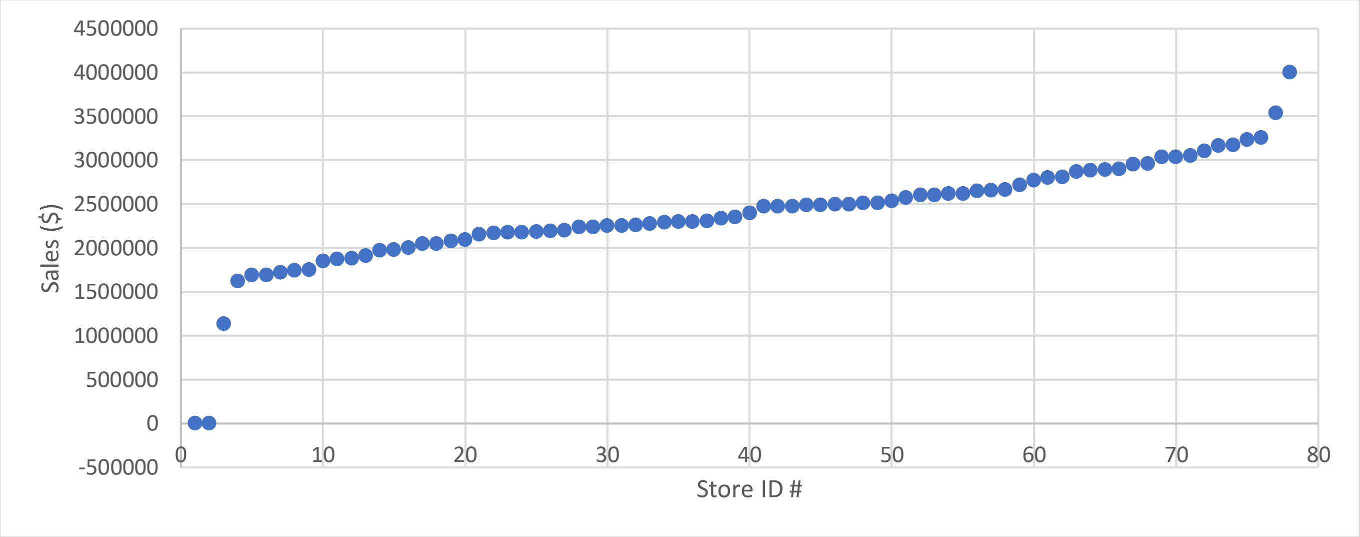
I. Introduction

Buster’s Brewhaus is a trending sports bar with various locations throughout the United States. Buster’s Brewhaus is known for its affordable pricing for lunch and dinner. Along with takeout options and reputable customer service. As a data analyst, I was hired to analyze sales from the 2019 in order to build predictive models for projected sales. The purpose of this project is to perform pre-model analysis on Buster’s Brewhaus’ sales and the relevant analysis for continuous regressors that affect said sales. Buster’s Brewhaus has done well in 2019 and we would like to know more about the factors that influence the sales. We have taken demographics data to find relationships between certain demographics and how to increase sales. We have also developed multiple dummy variables in order to determine how sales are affected if a trait is present or absent. We must perform pre-model analysis on the multi-trait dummy variables and limited integer value (LIV) variables that are contained in the data set. Our long run goal is to determine what factors affect sales for Buster’s and how to tailor marketing and our efforts to increase sales. Being able to predict where it is best to expand and build future locations for Buster’s Brewhaus is our primary goal. For statistical analysis, we will use the analytics software SAS and for graphs we will use Excel.

II. Dependent Variable

Our dependent variable—also known as Yi—is the main factor that we are trying to understand. In our model, we will analyze ***sales*** as our dependent variable for the model. The units of measure for sales are in U.S. dollars (dollar value of sales at each store). The time period covered in sales is 2019. Other important variables that provide information about the data are "Store_ID" and "close_date". “Store_IDi” refers to a unique number (from 1 to 78) that identifies each store, “i,” in the data set. “Close_datei” is the date on which store “i” closed, if it closed within the year as all of the stores in the sample opened before 2019. The value will be blank if the store did not close. Dates are written using the first three letters of the month, then an underscore, then the 4 digits that identify the year (for example MAR_2019 means March 2019). We must ensure that all observations on our dependent variable, Yi, make sense. As our first step, we will plot a graph that shows us the original observations and their respective Sales.

Figure 1: Average Sales for all Buster's Brewhaus Locations in our sample



By observation of Figure 1, there are some unreasonable observations towards the bottom with sales being equal to zero, below the trend, or above the trend. Our next step in order is then to use SAS in order to write a program that will generate the summary statistics of the original data set.

Table 1: Summary Statistics

Number of Observations	Mean	Standard Deviation	Minimum Value	Maximum Value
78	2,374,430.24	617,552.32	-770.85	3,997,991.13

From the results, we can see that the minimum for Sales is a negative value, which indicates that one or more observations had negative sales. We can also see that one observation has zero sales. From there, we identify what stores have these sales and remove the observations that are unreasonable using SAS. In this case store 1 has negative sales, and store 2 has zero sales. Store 1 might have negative sales if it is a store that is not performing its best and in further analysis might indicate that it must be shutdown. Store 2 has zero sales, which probably indicate that we do not have a given observation number for sales. After removing both observations, we rename our data set to BUSTERS2 to keep track now that we have removal of store 1 and store 2. After the removal of these two stores, we can now create an adjusted graph.

Figure 2: Average Sales for all Buster's Brewhaus Locations (without unreasonable observations)

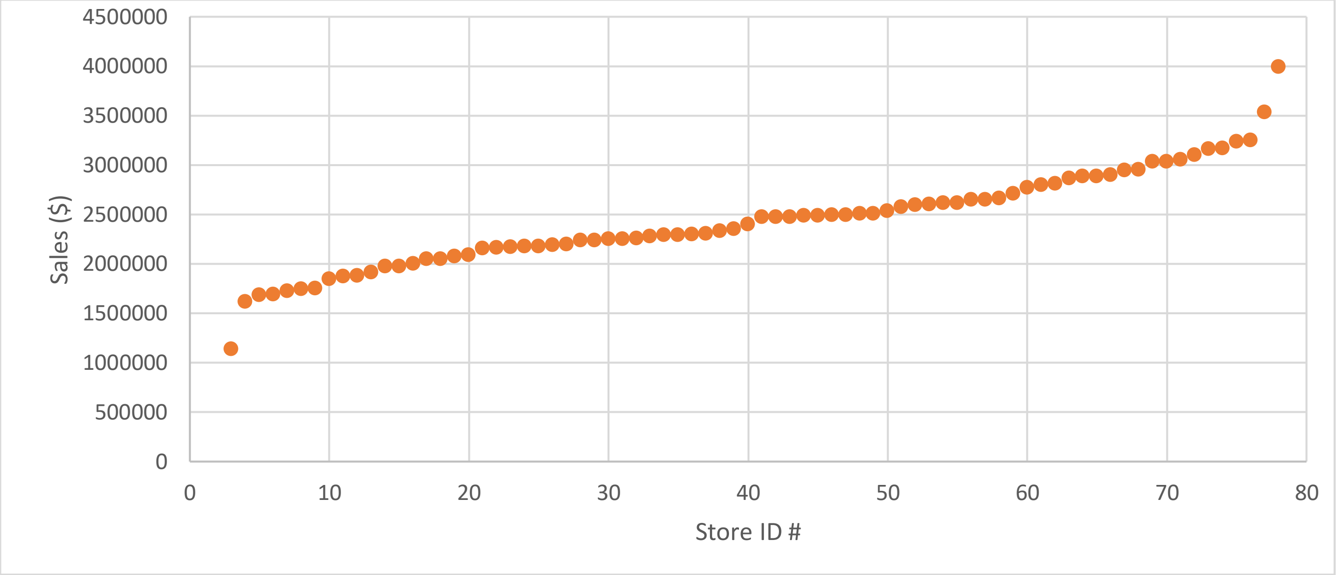


Table 2: Summary Statistics (adjusted for unreasonable observations)

Number of Observations	Mean	Std Dev	Minimum Value	Maximum Value
76	2,436,925.39	487,021.26	1,136,846.32	3,997,991.13

Furthermore, we can continue to analyze any abnormalities in our graph. We can compute the threshold values a "low" outlier and a "high" outlier using Excel and the given summary statistics above. The purpose of this is to identify observations that do not align to the overall population. This process will help us identify the mean of the population and how far variables can differ in order to stay in our analysis. This will set a threshold to have our values as no lower than the mean minus two and a half standard deviations and no higher than the mean plus two and a half standard deviations. In this case, the threshold formulas are:

Low outlier: mean – 2.5std

High outlier: mean + 2.5std

Where the mean is: 2436925.39. And the standard deviation is: 487021.26

Low outlier	1,219,372.24
High outlier	3,654,478.54

We now have these parameters of what a low and a high outlier is, we utilize SAS to remove any outliers that do not fit with our parameters and rename the dataset as BUSTERS3. STORE3 has sales of 1,136,846.32 which is below our low outlier threshold. STORE78 has sales of 3,997,991.13 which is above our high outlier threshold.

Figure 3: Avg Sales for all Buster's Locations (adj. for unreasonable observations & outliers)

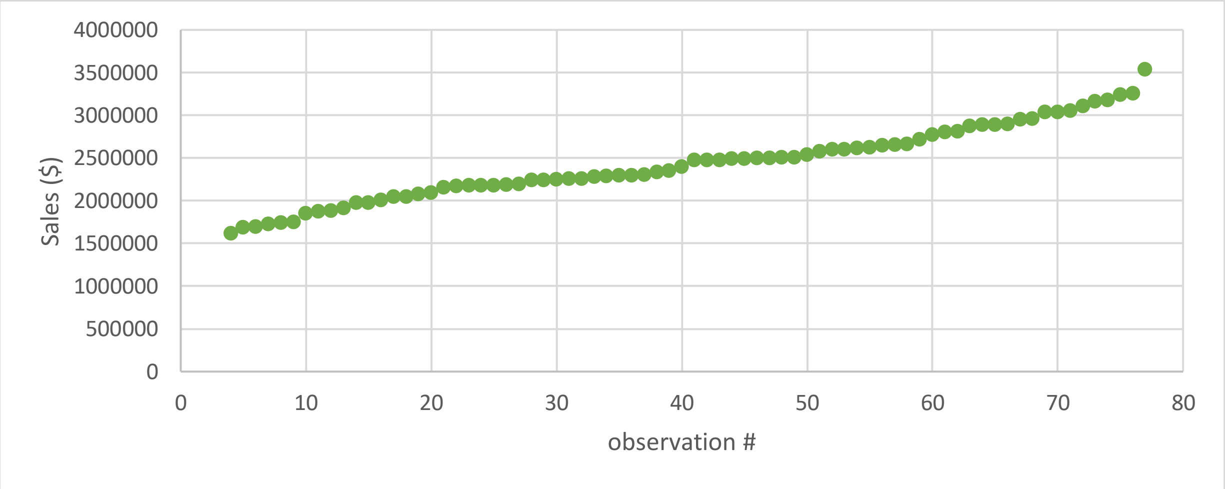


Figure 3 is now a graph that is more uniform as we have eliminated unreasonable observations and outliers. Now that these two unreasonable outliers have been dropped, we can now get our new summary statistics.

Table 3: Summary Statistics (adjusted for unreasonable observations and outliers)

Number of Observations	Mean	Std Dev	Minimum	Maximum	Coefficient of Variation
74	2433398.55	432596.11	1620139.23	3535412.28	17.78

Our dependent variable has now been cleaned from unreasonable observations and outliers adjusted to our threshold, the following table summarizes what observations were removed from the dataset and the reasoning.

Table 4: Summary of removed observations and the reasoning

Number of observation/STORE_ID	Reason for removal
STORE_ID = 1	Unreasonable observation since sales were negative.
STORE_ID = 2	Unreasonable observation since sales were zero.
STORE_ID = 3	Has sales of 1,136,846.32 which is below our low outlier threshold
STORE_ID = 78	Has sales of 3,997,991.13 which is above our high outlier threshold

III. Continuous Potential Independent Variables

The table below illustrates the continuous potential independent variables that are relevant to our analysis.

Table 5: Potential Independent Variables

Independent Variables	**Description**
Population 45-50	The number of people aged 45-50 who live within a one-half mile radius of a store.
Hispanic Population	Number of Hispanic individuals who live within one radial mile of a store.
Caucasian Population	Number of Caucasian individuals who live within one radial mile of a store.
African American Population	Number of African American individuals who live within one radial mile of a store.
Asian Population	Number of Asian individuals who live within one radial mile of a store.
Single Population	The number of single individuals who live within one radial mile of a store.
Married Population	The number of married individuals who live within one radial mile of a store.
Income $40k-100k	The number of households within one radial mile of a store that earn a household income between $40,000 and $50,000.
Income Greater than $100,000	The number of households within one radial mile of a store that earn a household income greater than $100,000.
Per Capita Income	The per capita income, in dollars, of households located within one radial mile of a store.
Average Income	Average income, in dollars, of households located within one radial mile of a store.
Bachelor's Degree	The number of people living within one radial mile of a store whose highest level of education is a bachelor's degree.
Master's Degree or higher	The number of people living within one radial mile of a store who hold a master's degree or higher.
Business	The number of people living within one radial mile of a store whose occupation is business-related.
Financial	The number of people living within one radial mile of a store whose occupation is finance-related.
Engineer	The number of people living within one radial mile of a store whose occupation is Engineering-related
Computer Science	The number of people living within one radial mile of a store whose occupation is computer science-related.

Table 5: Potential Independent Variables (cont.)

Independent Variables	Description
Social Science	The number of people living within one radial mile of a store whose occupation is social science-related.
Repair	The number of people living within a one-half mile radius of a store whose occupation is repair-related.
Played Baseball	An index of the number of people who live within one radial mile who played organized baseball within the last 12 months.
Played Basketball	An index of the number of people who live within one radial mile who played organized basketball within the last 12 months.
Played Bowling	An index of the number of people who live within one radial mile who played organized bowling within the last 12 months.
Played Football	An index of the number of people who live within one radial mile who played organized football within the last 12 months.
Played Hockey	An index of the number of people who live within one radial mile who played organized hockey within the last 12 months.
Played Volleyball	An index of the number of people who live within one radial mile who played organized volleyball within the last 12 months.
Played Yoga	An index of the number of people who live within one radial mile who participated in yoga within the last 12 months.
Exercise Regularly	An index of the number of people who live within one radial mile of a store who indicated that they exercise regularly.
Restaurant Score	An index of the number of people who live within one radial mile of a store who indicated that they had eaten out 10 or more times at a restaurant within the last 30 days
Night Life Score	An index of the number of people who live within one radial mile of a store who indicated that at least one household member went to a bar or nightclub within the past 12 months

IV. Potential Regressors Analysis

With our adjusted data set, we can observe the summary statistics along with the coefficient of variation for all variables. Doing analysis on our potential regressors, we look for unreasonable numbers, outliers, null or missing values, and their respective coefficient of variation. When estimating regression models, all independent variables must have "sufficient variation”. As long as the model contains an intercept, if any regressor does not vary then it will be perfectly collinear with the constant. That is, we have perfect multicollinearity; the OLS estimates do not exist in unique form. For continuous regressors, we can use the coefficient of variation (CVX) to measure variation. The CV of X is the standard deviation of X standardized by the mean of X (times 100, in absolute value):

$$\text{(coefficient of variation for } X_{ji}\text{)} = CV_{X_{ji}} = \left| \left(\frac{\hat{\sigma}_{X_{ji}}}{\overline{X}_{ji}} \right) \times 100 \right|$$

where: $\overline{X}_{ji}$ = the mean of X_{ji}, and $\hat{\sigma}_{X_{ji}}$ = the standard deviation of X_{ji}.

As a rule of thumb for continuous independent variables, if the CV is greater than two, there is sufficient variation in X. Therefore, if a variable does not have sufficient variation then it must be excluded from further consideration. The coefficient of variation we are looking for is one that is greater than 2 for all variables. With these requirements set, we create a table of the basic summary statistics and the coefficient of variation of variables that have issues.

Table 6: Summary Statistics for Analysis

Variable	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**	**Coeff of Variation**
Store_ID	74	40.5	21.51	4	77	53.1
Close_date	0	.	.	.	.	.
Pop_45to50	73	156.51	118.13	0	464	75.48

Pop_45to50 has 73 observations, which means that there is one missing observation now that our adjusted dataset has a total of 74 stores. In order to see why there is an observation missing we contact the data collector and figure out if we must replace it or drop it. After contacting the data collector, we have determined that the missing value is closely related to the mean of this variable. So, we can therefore replace the missing value with the mean of Pop_45to50 which is 156.51. STORE4 is the one with the missing value, so we will use SAS to insert the mean. Another issue is that Close_date has zero observations. So, it is irrelevant and can be dropped from our potential regressors analysis.

V. Micronumerosity, Rescaling and Adjustments for Potential Continuous Regressors

Micronumerosity is a condition of too few observations, or very small degrees of freedom (df < 30). The smaller the degrees of freedom, the less accurate your estimates are. In our adjusted dataset, our degrees of freedom are higher than 30. We do not have to worry about micronumerosity since we are testing 74 observations against 32 variables.

74-32 = 42 degrees of freedom

Since SALES is in millions, I would rescale it to a smaller number using:

SALES/100 = SALES_T

Which would convert the figure as "sales in thousands". For example, the mean of $2,433,398.55 would become 24,333.9855 "thousands". This gives us figures that align closer with other means in our data.

VI. Correlation

In order to find what independent variables best fit our model, we will determine the correlation between sales and every possible regressor. To do this we will use hypothesis testing to see what variables are statistically significant. This will help us determine what variables have the highest impact on our dependent variable. We use Pearson Correlation Coefficients where Prob > |r| under H0: Rho = 0. Our null hypothesis H0, tells us that we are testing statistical significance for a regressor in our overall model. By stating H0 as Rh0 = 0 we are indicating that we will test a predicted regressor against the null hypothesis that its impact on the overall model and in our case against sales, will be equal to zero. This in turn indicates that the regressor is not statistically significant in our analysis.

We will analyze using our p-value for every independent variable. The p-value given by a hypothesis test represents the likelihood that our null hypothesis is true. In our case we will state that our null hypothesis is that that variable has no impact on sales. Our alternative hypothesis will be that it does. We will use a 90% confidence level, which gives us a p-value of 0.1. If the p-value in our hypothesis test is a value less than 0.1, we then reject the null hypothesis since we have evidence to believe that the independent variable is statistically significant and has correlation to sales. If the p-value is greater than 0.1, we fail to reject the null hypothesis and are led to believe that this variable has no impact on the dependent variable. We can now break down different potential regressors in order to observe their impact.

Table 7: Demographics with correlation coefficient and p-value

	Age	Hispanic	Caucasian	African American	Asian
Correlation Coefficient to Sales	0.19729	0.01752	0.07077	0.03909	0.02501
P-value	0.092	0.8822	0.5491	0.7409	0.8325

Table 8: Education with correlation coefficient and p-value

	Bachelors	Masters and above
Correlation Coefficient to Sales	0.06077	-0.01405
P-value	0.607	0.9054

Table 9: Workforce with correlation coefficient and p-value

	Business	**Financial**	**Computer**	**Engineer**	**Social Science**	**Repair**
Correlation Coefficient to sales	-0.07231	-0.00856	-0.02501	0.2349	-0.04858	0.2395
P-value	0.5404	0.9423	0.8325	0.044	0.6811	0.0399

Table 10: Sports with correlation coefficient and p-value

	Baseball	**Basketball**	**Bowling**	**Football**	**Hockey**	**Volleyball**	**Yoga**	**Exercise Regularly**
Correlation Coefficient to sales	0.22251	0.25743	0.26449	0.24723	0.30242	0.12675	0.15532	-0.10507
P-value	**0.0567**	**0.0268**	**0.0228**	**0.0337**	**0.0088**	0.2819	0.1864	0.3729

Table 11: Other variables with correlation coefficient and p-value

	Restaurant Score	**Nightlife Score**
Correlation Coefficient to sales	0.22524	-0.22646
P-value	0.0537	0.0524

Now we can create a list of potential regressors that are significantly correlated to sales.

Table 12: Demographics with correlation coefficient and p-value

Potential Regressors	Correlation Coefficient	P-value
Age	0.19729	0.092
Engineer (occupation)	0.2349	0.044
Repair (occupation)	0.2395	0.0399
Baseball players (sports)	0.22251	0.0567
Basketball players (sports)	0.25743	0.0268
Bowling players (sports)	0.26449	0.0228
Football players (sports)	0.24723	0.0337
Hockey players (sports)	0.30242	0.0088
Restaurant Score	0.22524	0.0537
Nightlife Score	-0.22646	0.0524

VII. Alpha Variables

We now have additional variables to analyze, including some that are “alpha variables” and variables that are not legitimate binary dummy variables. Alpha variables are variables that utilize letters or words to identify qualitative data. We must correct some of these variables in order to have proper and unbiased analysis.

For example, CC (cover charge) can either be a value of 1 or 2, in order to perform proper analysis using dummy variables we require a value of 0 (if the property is not present) or 1. This is required since if the value is a higher number it will translate as a higher effect and not if the property is present or absent. DT (drive-thru) has a similar issue, but here it is related to being formatted as words instead of a number value to see if the property is present or absent. We must convert the value of “Yes” and “No” into values of 0 and 1 to establish if the property is there. Champ (Championship) has an almost identical issue to DT, but the keywords here are “Y” and “N”. BT (bar tax) has a similar issue as DT but the keywords are “high” and “low”. In order to perform proper analysis, we must create new variables to correct these errors.

Table 13: Definitions of New Independent Variables for Analysis

Variable Name	**Description**
Music	= 1 if the store offers live music on the weekends, and 0 if not.
Football	= 1 if there is an NFL stadium located within 1 radial mile of a store, and 0 if not.
Baseball	= 1 if there is an MLB stadium located within 1 radial mile of a store, and 0 if not.
Basketball	= 1 if there is an NBA stadium located within 1 radial mile of a store, and 0 if not.
Soccer	= 1 if there is a Major League Soccer stadium located within one radial mile of a store, and 0 if not.
University	= 1 if there is a university located within 1 radial mile of a store, and 0 if not.
cover_charge	=1 if a store does not require a cover charge to enter the bar on Fridays and Saturdays, and 2 if it does.
drive_thru	= "yes" if there the store has a drive-thru for take out food and beverages, and "no" if not.
bar tax	= "high" if a store is located in a city that has a bar tax greater than 9%, and "low" if not.
champion	= "Y" if the city in which a store is located has won a professional sports championship within the last 4 years, and "N" if not.

Now we create new variables for cover charge, drive-thru, bar tax, and championship. We will transform then so they only contain the value of 0 or 1.

Table 14: Previous Variable and Adjusted/New Independent Variables

Previous Variable & Adjusted/New Variable	**Description**
cover_charge	= 0 if a store does not require a cover charge to enter the bar on Fridays and Saturday, and 1 if it does.
drive_thru	= 0 if the store lacks a drive-thru, and 1 if it has one.
bar_tax	= 0 if a city where a store is located has a high tax rate, and 1 if the city has a low tax rate.
champion	= 0 if a city where a store is located has not won a professional sports championship within the last 4 years, 1 if it has won.

Now we can get our five basic summary statistics for the adjusted proper 2-trait dummy variables and for our other independent variables.

Table 15: Summary Statistics

Variable	**N**	**Mean**	**Std Deviation**	**Minimum**	**Maximum**
music	74	0.05	0.23	0	1
football	74	0.34	0.48	0	1
baseball	74	0.2	0.4	0	1
basketball	74	0.2	0.4	0	1
soccer	74	0.14	0.34	0	1
university	74	0.97	0.16	0	1
Cover_charge	74	0.27	0.45	0	1
Drive_thru	74	0.2	0.4	0	1
Bar_tax	74	0.62	0.49	0	1
champion	74	0.23	0.42	0	1

From observation, we can see that the independent variable music has a very low mean value of 0.05, which is extremely close to 0. Since this is a dummy variable, we know that a value of 1 indicates if the store offers live music on the weekends, and 0 if not. This dummy variable does not display sufficient variation. We require sufficient variation in order for a dummy variable to be relevant in our analysis. For example, a very high mean close to 1 indicates that almost all observations possess this trait. And a very low mean close to 0, indicates that almost no observations possess this trait. In our data, we have two dummy variables that have means that will indicate that they do not have sufficient variation, music and university.

This low mean tells us that almost all of our 74 observations do not offer live music on the weekends. Now we turn to analyze the independent variable university. We can see that the mean value for university is 0.97, which is extremely close to 1. Since this is a dummy variable, we know that a value of 1 indicates that there is a university located within one radial mile of a store, and 0 if not. This dummy variable does not display sufficient variation. This high mean tells us that virtually all of our 74 observations are located within one radial mile of a university.

Table 16: Dummy Variables Without Sufficient Variation

Variable	**N**	**Mean**	**Std Deviation**
music	74	0.05	0.23
university	74	0.97	0.16

Now that we have finalized our dummy variables that are relevant for our analysis and display sufficient variation, we can compute correlation statistics.

Table 17: Independent Variables & Correlation to Sales

Independent Variable Name	**Correlation to Sales**	**P-value**
Football	0.74228	< 0.0001
Baseball	0.48689	< 0.0001
Basketball	0.42647	0.0002
Soccer	-0.13607	0.2477
Cover_charge	0.78845	< 0.0001
Drive_thru	-0.02549	0.8293
Bar_tax	0.36442	0.0014
champion	0.31965	0.0055

With these correlation statistics, we can now perform hypothesis testing in order to determine what variables have the strongest correlation to our dependent variable, Sales. Our hypothesis being tested is that if there is any correlation between these variables to sales, and if there is not any correlation then that variable is insignificant. If so, our null hypothesis establishes that correlation does not exist and is zero. We use our p-values from our correlation table in order to determine this. If we have a small p-value, this serves as evidence that we should reject the null hypothesis. If we reject the null hypothesis, this indicates that there is significant correlation between sales and that variable. Since we are testing at a 90% confidence level, we are looking for p-values that are less than or equal to 0.10.

We now observe that the variables Football, Baseball, Basketball, Cover_charge, Bar_tax, and Sports_champ all have p-values that are less than 0.10. This tells us that we reject the null hypothesis that these variables have no impact on sales. Which means that these variables do have an impact on sales and must be used for further analysis. In comparison, the variables Soccer and Drive_thru have high p-values, which indicate we fail to reject the null hypothesis. This means that these variables have no statistical significance.

Football has a positive high correlation to sales, which indicates that if there is an NFL stadium located within 1 radial mile of a store, sales will be higher for the store. Baseball and Basketball also have a positive correlation to sales, in this case lower than Football, which indicates if there is an MLB stadium located within one radial mile of a store for baseball and if there is an NBA stadium located within one radial mile of a store for basketball sales will be higher. Cover_charge is another variable with a high positive correlation to sales, which indicates that sales are affected if a store does not require a cover charge to enter the bar on Fridays and Saturday. Bar_tax has a positive correlation to sales, lower than other variables but still significant. This indicates that if a city where a store is located has a low tax rate, sales will increase. And finally, Sports_champ also has a positive correlation to sales, lower than other variables but still significant. This indicates that if a city where a store is located has won a professional sports championship within the last 4 years, sales will increase.

VIII. Multi-characteristic intercept dummy variables.

Before continuing with any further analysis, we must convert alpha variables and qualitative variables that are not proper intercept dummy variables. To do so, we wrote the required block of code on SAS. Afterwards, we can begin analysis with the variables that are intended to be multi-characteristic intercept dummy variables.

Table 18: Multi-trait Dummy Variables

Name of Variable	**Description**
Stand Alone	=1 if the store is a stand-alone store, and 0 if not.
Strip Mall	=1 if the store is located in a strip mall, and 0 if not.
Lifestyle	=1 if the store is located in a lifestyle mall, and 0 if not.
High Pop. Growth	=1 if the store is located in a high rate of population growth area, and 0 if not.
Medium Pop. Growth	=1 if the store is located in a medium rate of population growth area, and 0 if not.
Low Pop. Growth	=1 if the store is located in a low rate of population growth area, and 0 if not.
Negative Pop. Growth	=1 if the store is located in a negative rate of population growth area, and 0 if not.
West	=1 if the store is located in the West region of the US, and 0 if not.
Midwest	=1 if the store is located in the Midwest region of the US, and 0 if not.
Southwest	=1 if the store is located in the Southwest region of the US, and 0 if not.
East	=1 if the store is located in the East region of the US, and 0 if not.

Table 19: Summary Statistics for New Qualitative Variables

Variable	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
stand_alone	74	0.3	0.46	0	1
strip_mall	74	0.35	0.48	0	1
life_style	74	0.35	0.48	0	1
pop_high	74	0.05	0.23	0	1
pop_medium	74	0.27	0.45	0	1
pop_low	74	0.41	0.49	0	1
pop_negative	74	0.27	0.45	0	1
West	74	0.03	0.16	0	1
MW	74	0.27	0.45	0	1
SW	74	0.32	0.47	0	1
East	74	0.38	0.49	0	1

Upon analysis, we find just one abnormality with these qualitative variables. Since they are dummy variables, we require a minimum of 0 and a maximum of 1 which they all possess. We also must have a standard deviation greater than zero. The mean for all variables here except for West and pop_high are greater than 0.1 and less than 0.9. Pop_high has a mean of 0.05 and it also has a low standard deviation of 0.23. This low mean indicates that pop_high has a mean of 0.05, which tells us that this dummy variable is closing in to have a mean of zero and therefore indicates that almost no Buster's stores are located in an area where the rate of growth of the population is considered to be high (as determined by the definition set forth by the US Census). This violates our assumption of sufficient variation existing for this variable. In order to correct this problem, we could take the drastic measure of dropping the variable but alternatively we will converge it. Since pop_medium has sufficient variation and is relatively close to the standard for pop_high, we will combine these two variables as pop_high_medium, which is the sum of the pop_medium and pop_high variables.

West has a mean of 0.03 and it also has a low standard deviation of 0.16. This low mean indicates that West has a mean of 0.03, which tells us that this dummy variable is closing in to have a mean of zero and therefore indicates that almost no Buster's stores are located in the western region of the United States. This violates our assumption of sufficient variation existing for this variable. In order to correct this problem, we will combine this region with another one. Since the Midwest region has sufficient variation and the western region of the United states is relatively close, we will combine these two variables as W_MW, which is the sum of the West and the Midwest. We then generate a new set of summary statistics for these corrected variables and see a much uniform mean.

Table 20: Summary Statistics for Corrected Variables

Variable	N	Mean	Std Dev	Minimum	Maximum
pop_high_medium	74	0.32	0.47	0	1
W_MW	74	0.3	0.46	0	1

With the proper intercept dummy variables that have sufficient variation, we can now generate the correlation coefficient between them and Sales. We will utilize the Pearson correlation coefficient to evaluate the correlation between these multi-trait dummy variables and Sales. Those with a negative coefficient indicate that they have a negative impact on sales while those with a positive coefficient indicate a positive impact on sales. This is because there is a negative relationship between sales and a trait being present if there is a negative correlation coefficient. And the opposite is true for positive correlation coefficient. The P-value given can be evaluated under the null hypothesis to see if a variable will have impact on the dependent variable. We require small P-values when looking for statistically significant variables. For a 90% confidence level, P-values should be no greater than 0.1. Under the null hypothesis, we state that if a coefficient has a P-value smaller than 0.1, then we fail to reject the null hypothesis that states that that coefficient is statistically significant.

Table 21: Correlation Coefficients for Sales – Building/Infrastructure Type

	Standalone	Strip Mall	Life Style
Correlation Coefficient (correlation with Sales)	0.80	-0.76	0.0009
P-Values	<0.0001	<0.0001	0.99

Under a 90% confidence interval, we have a P-value of 0.1. In this case both Standalone and Strip Mall have a small P-value, that indicates that both of these variables are statistically significant. Both of these buildings are statistically significant to our analysis, and we can see that if a building is standalone it has a positive effect on Sales while being located in a Strip Mall has a negative impact on Sales. This means that a Buster's Brewhaus located in a Strip Mall will have less sales than one that is a stand alone. Since Life Style has a P-value drastically larger than 0.1, we reject the null hypothesis. That is, Life Style has no statistically significant impact.

Table 22: Correlation Coefficients for Sales – Population Growth

	Medium-High Pop. Growth	Low Pop. Growth	Negative Pop. Growth
Correlation Coefficient (correlation with Sales)	0.80	-0.11	-0.72
P-Values	<0.0001	0.35	<0.0001

Under a 90% confidence interval, we have a P-value of 0.1. In this case both Medium-High Population Growth and Negative Population Growth have a small P-value, that indicates that both of these variables are statistically significant. Both of these levels of population growth are statistically significant to our analysis, and we can see that if there is a medium to high population growth then we have a positive effect on Sales. This indicates that if a city has medium to high population growth, the sales of that Buster's location will be higher. While having negative population growth has a negative impact on Sales. Which indicates lower sales for those Buster's in cities with low population growth. However, since Low Population Growth has a p-value larger than 0.1, we reject the null hypothesis. That is, that areas with low population growth have no statistically significant impact.

Table 23: Correlation Coefficients for Sales – Regional

	West-Midwest Locations	Southwest Locations	East Locations
Correlation with Sales	-0.74	0.80	-0.079

P-Values	<0.0001	<0.0001	0.5

For the regional correlation coefficients, we find that West-Midwest and Southwest locations are statistically significant. In this case both West-Midwest and Southwest locations have a small P-value, that indicates that both of these variables are statistically significant. Both of these regions are statistically significant to our analysis. We can see that if the location is in the West-Midwest region, then we have a negative effect on Sales. Which in turn indicates that if a Buster's location is in that region, sales will be lower. While locations in the Southwest have a positive impact on Sales, which indicates that any Buster's location in the Southwest will have higher sales based on being in this region, ceteris paribus. Since East has a P-value larger than 0.1, we reject the null hypothesis. That is, that locations in the East of the United States have no statistically significant impact.

IX. Limited Integer Value (LIV) Variables

Limited-integer-value (LIV) variables are variables that take on only integer values and there are a limited number of values that they can take on. For our analysis, we will assume that LIV variables take on six or fewer integer values. In the case of LIV variables, must create frequency tables since our usual summary statistics will not be enough. The reason being that a frequency table shows all of the values that a variable can take on, and how many times it takes on each value, both in absolute terms and in percentage term. In our data, the LIV variables are related to the proximity and presence of Buster's Brewhaus' competitors.

Table 24: Potential LIV Variables

LIV Variable	Description
Hooters	The number of “Hooters” locations within 10 radial miles of Busters
Twin Peaks	The number of “Twin Peaks” locations within 10 radial miles of Busters
Buffalo Wild Wings	The number of “Buffalo Wild Wings” locations within 10 radial miles of Busters
Metrics	The number of “Metrics” locations within 10 radial miles of Busters

In order to determine if a LIV variable has sufficient variation, each outcome in the variable must comprise of at least 10% of the observations in the sample. For that we then turn to look at the frequency tables.

Table 25: Hooters Frequency Table

Hooters Locations within a ten-mile radius	Frequency in our 74 observations	Percent	Cumulative Frequency	Cumulative Percent
0	25	33.78	25	33.78
1	30	40.54	55	74.32
2	19	25.68	74	100.00

The frequency table for our competitor, Hooters, meets our criteria since this LIV variable has sufficient variation because each outcome in the variable comprises of at least 10% of the observations in the sample.

Table 26: Twin Peaks Frequency Table

Twin Peaks Locations within a ten-mile radius	Frequency in our 74 observations	Percent	Cumulative Frequency	Cumulative Percent
0	69	93.24	69	93.24
1	3	4.05	72	97.30
2	2	2.70	74	100.00

The frequency table for Twin Peaks, does not meet our criteria since this LIV variable does not have sufficient variation. Although one outcome is higher than 10%, we find that two of them are lower (4.05 and 2.70 respectively).

Table 27: Buffalo Wild Wings Frequency Table

BWW Locations within a ten-mile radius	**Frequency in our 74 observations**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	43	58.11	43	58.11
1	21	28.38	64	86.49
2	7	9.46	71	95.95
3	3	4.05	74	100.00

The frequency table for Buffalo Wild Wings, does not meet our criteria since this LIV variable does not have sufficient variation. Although two outcomes are higher than 10%, we find that two of them are lower (9.46 and 4.05 respectively).

Table 28: Metrics Frequency Table

Metrics Locations within a ten-mile radius	**Frequency in our 74 observations**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	7	9.46	7	9.46
1	8	10.81	15	20.27
2	8	10.81	23	31.08
3	7	9.46	30	40.54
4	7	9.46	37	50.00
5	8	10.81	45	60.81
6	6	8.11	51	68.92
7	5	6.76	56	75.68
8	7	9.46	63	85.14
9	5	6.76	68	91.89
10	6	8.11	74	100.00

Finally, the frequency table for Metrics does not meet our criteria since this LIV variable does not have sufficient variation. Some of the percentages are higher than 10% but in order to meet our criteria of a proper LIV variable we require that all of them are 10% or higher.

Therefore, given these four potential LIV variables, we can determine that the only variable that meets our criteria of having sufficient variation is Hooters. This means that we must correct Twin Peaks, Buffalo Wild Wings, and Metrics to ensure they have sufficient variation. To do so, we can manipulate those LIV variables without sufficient variation and turn them into dummy variables. For that, we must combine the rest of the observations and have the variable be equal to "zero" if there are no locations or "one" if there are any locations regardless of frequency. If we look at Table 9, we can observe that the Twin Peaks variable cannot be transformed to have sufficient variation since the combined share of Buster's locations that have no Twin Peaks locations within a ten-mile radius is 93.24%. This indicates that we must remove the Twin Peaks variable from our regression analysis since the frequency of there being "zero" Twin Peaks location to our stores indicates that adding it to our analysis would be irrelevant since they are not predominant near our stores.

Table 29: Variables Removed

Variable	Reason for removal
TP	Cannot be transformed into a dummy variable since the frequency of there being "zero" Twin Peaks location to our stores indicates that adding it to our analysis would be irrelevant since they are not predominant near our stores.
Metrics	Does not meet our criteria since the variable does not have sufficient variation. Some of the percentages are higher than 10% but in order to meet our criteria of a proper LIV variable we require that all of them are 10% or higher.

In comparison, we can see that Buffalo Wild Wings can be saved if we convert it into dummy variables. Buffalo Wild Wings has a high percentage since the combined share of Buster's locations that have no BWW locations within a ten-mile radius is 58.11%. This indicates that there is a 58% percent chance of there not being any BBW near our stores, but in turn it tells us that there is a 42% chance that there is one or more BWW locations near us.

X. Corrected LIV variables

Now that the LIV variables have been adjusted, we can evaluate our summary statistics with these final variables.

Table 30: Summary Statistics for Corrected LIV variables

Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
SALES	74	2433399.00	432596.00	180071492.00	1620139.00	3535412.00
Hooters	74	0.92	0.77	68.00	0.00	2.00
Buffalo	74	0.42	0.50	31.00	0.00	1.00

We can also now calculate the correlation to Sales of these adjusted LIV variables. Again, we will use Pearson Correlation Coefficient. We will evaluate at a 90% confidence level where P-values should be no greater than 0.1. Under the null hypothesis, we state that if a coefficient has a P-value smaller than 0.1, then we fail to reject the null hypothesis that states that that coefficient is statistically significant.

Table 31: Correlation Coefficients for Sales – Corrected LIV variables

	Hooters	Buffalo
Correlation Coefficient (correlation with Sales)	-0.90	-0.79
P-Values	<0.0001	<0.0001

Here, we can observe that Hooters and Buffalo are statistically significant since their P-value is smaller than 0.1. The interpretation is that all of these variables do indeed have an impact on Buster's Sales. These three LIV variables have a negative impact on Buster's Sales, which indicates that if there are at least one store of any of these three competitors our sales are expected to go down. Finally, we conclude this section of our research with a table of our potential regressors.

Table 32: List of Potential Regressors

Name of Variable	Description
Stand Alone	=1 if the store is a stand-alone store, and 0 if not.
Strip Mall	=1 if the store is located in a strip mall, and 0 if not.
Medium-High Pop. Growth	=1 if the store is located in a high or medium rate of population growth area, and 0 if not.
Negative Pop. Growth	=1 if the store is located in a negative rate of population growth area, and 0 if not.
West-Midwest	=1 if the store is located in the West or Midwest region of the US, and 0 if not.
Southwest	=1 if the store is located in the Southwest region of the US, and 0 if not.
Hooters	= 1 if there is one or more Hooters location within 10 radial miles of Busters, and 0 if not.
Buffalo Wild Wings	= 1 if there is one or more BWW location within 10 radial miles of Busters, and 0 if not.

After identifying potential regressors and selecting which fit our regression model best based on our set parameters, we can now move on to the section of model building in our analysis. We can estimate and select the best fitting model that can more accurately explain sales at Buster's Brewhaus by using our eight steps of model building. Our purpose is now transitioning into finding the best fitting regression model and performing model validation through the calculation of the Out-of-Sample Mean Absolute Percentage Error (OOS MAPE).

XI. Final Potential Regressors

Demand theory and logical reasoning carries us into our next step for our analysis. Our demand theory indicates that income, population, preferences, prices of our own goods, prices of substitutes and complements will be correlated to our sales. Considering theory, we can assume that we are building a "demand" equation for our sales. Where our sales are related to how much demand there is for our services at Buster's Brewhaus. We are building a model that aids us predict sales and the demand of customers for our services given significant regressors. We have relevant income variables that we can incorporate into the model "Inc_40kto100k" is one that we should include since a high level of income indicates disposable income that can be spent at Buster's Brewhaus.

We also have to consider that our customers have preferences and that they will impact how they perceive our stores. We will include variables that indicate if customers play a sport such as football, baseball, basketball, and hockey. The reason being is that as a sports bar we can expect an influx of customers that are into sports. We can also include Married_pop as a regressor since the percentage of people that are married will have an influx on sales if the population is more likely to be married since they could have more combined income. We also must add regional variables like W_MW and SW that indicate the location of a Buster's Brewhaus. This in turn tells us where our business is more popular in the United States and how sales are affected from it.

Considering logic, we can assume that our biggest competitors are Hooters and Buffalo Wild Wings. Therefore, it is logical to think that sales at Buster's Brewhaus would be strongly (negatively) affected by the presence of any of these competitors in the vicinity. Logic indicates that we should include the variables Hooters and Buffalo as potential regressors. We can also consider that by logic if there are any football or baseball stadiums close, we could see an impact on sales. For example, if there is a football stadium nearby, fans that were unable to get tickets but want to be close to the action will visit Buster's.

Table 33: Final List of Potential Regressors

Variable Name
Married_pop
Inc_40Kto100K
Occ_repair
Played_football
Played_baseball
Played_basketball
Played_hockey
restaurant_score
football
baseball
cover_charge
champion
W_MW
SW
stand_alone
strip_mall
Hooters
Buffalo

XII. Fit Statistics for the Selected Models

Out of our twenty-four potential regressors, we narrowed out list down in order to create distinct models that can get us closer to our best fitted model that allows us to explain sales at Busters. We proceeded to create 20 different versions of regression models with subsets of our chosen variables. These models have at least 8 regressors and no more than 15. From all of the models that that were estimated, we can now choose the top five "contender models" and display their fit statistics.

Table 34: Values of Fit Statistics for the Selected Models

Model	Calc. General F Test (p-value)*	Value of R^2	Value of Adj R^2	# of significant slopes with correct sign**	# of significant slopes with incorrect sign	Value of OOSMAPE

P	66.65 (<0.0001)	0.8913	0.878	3 of 8	0 of 8	1.19425009
Q	70.27 (<0.0001)	0.8964	0.8836	4 of 8	0 of 8	0.240411581
R	65.58 (<0.0001)	0.8898	0.8762	3 of 8	0 of 8	0.204559113
S	67.56 (<0.0001)	0.8926	0.8794	3 of 8	0 of 8	0.272305925
T	72.96 (<0.0001)	0.8998	0.8875	5 of 8	0 of 8	0.307245734

Table 35: Regressors in Selected Models

Model P	Model Q	Model R	Model S	Model T
Married_pop	Married_pop	Married_pop	Married_pop	Married_pop
inc_40kto100k	Restaurant_score	Played_baseball	occ_repair	inc_40kto100k
cover_charge	champion	Played_basketball	football	champion
baseball	cover_charge	Played_hockey	cover_charge	cover_charge
Strip_mall	W_MW	football	W_MW	W_MW
Stand_alone	SW	Cover_charge	SW	SW
Hooters	Hooters	Hooters	Hooters	Hooters
Buffalo	Buffalo	Buffalo	Buffalo	Buffalo

*Where the General F-test has the null hypothesis (H0) that states that the variables in that model together as a group do not explain the dependent variable. For our five selected models the p-value is less than 0.1, which means we fail to reject the null hypothesis. That indicates that the variables as a group cannot explain sales.

** At a 90% confidence level

The table below summarizes and assesses the fit statistics of the top five contender models. We can categorize our models and their fit statistics in order to make a final decision of what model possesses the best fit statistics and will give us the best model in order to predict Buster’s sales.

Table 36: Assessment of Fit Statistics for Selected Models

Model	General F-Test	R^2	Adj R^2	% of significant slopes with correct sign	Any significant slopes with incorrect sign?	Value of OOS MAPE	Overall Assessment
P	significant	very strong	very strong	37.50%	no	bad	good
Q	significant	very strong	very strong	50%	no	good	good
R	significant	very strong	very strong	37.50%	no	very good	very good
S	significant	very strong	very strong	37.50%	no	good	good
T	significant	very strong	very strong	87.50%	no	good	very good

XIII. Multicollinearity

When it comes to detecting multicollinearity in our models, we can detect it by using the Sample Correlation Coefficient, which measures the degree of linear association that exists between two regressors. Having moderate or strong multicollinearity makes the variance of our OLS estimators overinflated, our t -tests than they should be, and the magnitudes of the parameter estimates and their estimated signs can be counter-intuitive. If we are trying to measure the degree of linear association between regressor Xij and Xsi we then have the following expression and thresholds:

$$\mathbf{r}_{\mathrm{Xj\,Xs}} = \left[\frac{\mathrm{cov}(X_{ji}, X_{si})}{\sqrt{\mathrm{var}(X_{ji}) \times \mathrm{var}(X_{si})}} \right], \qquad -1 \le r \le 1$$

$|r| = 1 \rightarrow$ perfect multicollinearity

$0.9 \le |r| < 1 \rightarrow$ severe multicollinearity

$0.5 \le |r| < 0.9 \rightarrow$ moderate to strong multicollinearity

$0 < |r| < 0.5 \rightarrow$ weak multicollinearity

In order to identify if multicollinearity exists in our analysis, we set the threshold to be that: 0.5 < | correlation coefficient | < 1, then moderate to severe multicollinearity exists.

Table 37: Variable Pairs that have Moderate to Severe Multicollinearity in Model P

Variable Pair	**Correlation Coefficient**
Inc_40kto100k and Married_pop	0.77242
Baseball and cover_charge	0.60144
Stand_alone and cover_charge	0.93564
Stand_alone and baseball	0.62811
Hooters and cover_charge	-0.72929
Hooters and baseball	-0.51656
Hooters and strip_mall	0.77919
Hooters and stand_alone	-0.77945
Hooters and Buffalo	0.76842
Buffalo and cover_charge	-0.51673
Buffalo and strip_mall	0.8668
Buffalo and stand_alone	-0.55228

Model P displays severe multicollinearity in many variables. Here we see that inc_pop40kto100k and married_pop display multicollinearity, which could indicate that these two variables are correlated since it is very likely that those who are married make a combined income between 40 to 100k. We also see that cover_charge is very correlated to most of the variables in our model. We see different numbers here that indicate either a negative or positive multicollinearity relationship, which indicates that some variables are negatively correlated, and others are positively. We also see a correlation between Hooters and Buffalo, which are our two main competitors, this could be indicating that these two competitors have a very similar impact on Buster's and the distinction between them could be irrelevant. Which would mean that whether there is a Buffalo or Hooters near us, the title of the competitor is irrelevant as both affect our sales in a negative way (they drive our sales down regardless). We can see that if our location is on a strip_mall or a stand_alone we will see high correlation to other variables, which could indicate for example with stand_alone and cover_charge that we usually do charge a cover_charge at our stand_alone locations.

Table 38: Variable Pairs that have Moderate to Severe Multicollinearity in Model Q

Variable Pair	**Correlation Coefficient**
Cover_charge and SW	0.87841
Cover_charge and Hooters	-0.72929
Cover_charge and Buffalo	-0.51673
Hooters and W_MW	0.80134
Hooters and SW	-0.83024
Hooters and Buffalo	0.76842

Model Q has less variables that display multicollinearity. Again, here we see that Hooters and Buffalo display high multicollinearity. We see an impact again from cover_charge and its correlation to our competitors—which here could indicate that if we charge a cover charge then our sales will go down given that our competitors are in the area. Model Q has regional regressors, W_MW and SW. Both of which have correlation to our competitor Hooters. It seems to be that if our competitor Hooters is in the W_MW region, then positive multicollinearity is displayed—which could indicate that those Hooters in the W_MW region make higher sales when compared to Busters. And the opposite is true for Hooters and SW. We also see that cover_charge and the SW region have high positive multicollinearity, which could be a telling sign on how most of our locations in the SW region charge a cover_charge.

Table 39: Variable Pairs that have Moderate to Severe Multicollinearity in Model R

Variable Pair	**Correlation Coefficient**
Played_hockey and played_basketball	0.85075
Cover_charge and football	0.72334
Cover_charge and Hooters	-0.72929
Cover_charge and Buffalo	-0.60648
Hooters and football	-0.70692
Hooters and Buffalo	0.76842
Buffalo and football	-0.60648

Model R has less variables that display multicollinearity. Again, Hooters and Buffalo display high multicollinearity. Charging a cover_charge has a correlation to our competitors. Here we introduce the regressors of football stadium presence. We see that both of our competitors, Hooters and Buffalo, share negative multicollinearity with football. This could indicate that if there is a football stadium nearby, there is a negative relationship between its presence and our competitors. Which would be favorable for us. We also see a high correlation coefficient of played_hockey and played_basketball, which indicates that a very high number of hockey players also play basketball.

Table 40: Variable Pairs that have Moderate to Severe Multicollinearity in Model S

Variable Pair	**Correlation Coefficient**
Married_pop and occ_repair	0.7036
Cover_charge and football	0.72334
Cover_charge and SW	0.87841
Cover_charge and Hooters	-0.72929
Cover_charge and Buffalo	-0.51673
Football and SW	0.78685
Football and Hooters	-0.70692
Football and Buffalo	-0.60648

We again see previous regressors having multicollinearity like previous models. Model S introduces occ_repair. Here we see high positive multicollinearity between the married_pop and occ_repair, this could indicate that a high number of the married_pop includes at least one spouse that has the trait of working in repairs.

Finally, Model T displays most of our previous findings in all other models regarding multicollinearity.

Table 41: Variable Pairs that have Moderate to Severe Multicollinearity in Model T

Variable Pair	**Correlation Coefficient**
Married_pop and inc_40kto100k	0.77242
Cover_charge and SW	0.87841
Cover_charge and Hooters	-0.72929
Cover_charge and Buffalo	-0.51673
W_MW and Hooters	0.80134
W_MW and Buffalo	0.76606
Hooters and SW	-0.83024
Hooters and Buffalo	0.76842
Buffalo and SW	-0.58826

In most of our models we see that football, W_MW, SW, baseball, Married_pop, Inc_40kto100k are insignificant. The constant presence of multicollinearity of these variables related to others could be an indicator on why they are insignificant to our model. We also see repeatedly that our competitors have high correlation coefficient to most of these regressors. Which could be indicating that they are common shared occurrences.

XIV. Chosen Model

Considering different factors such as R^2, adjusted R^2, percentage of significant slopes with correct signs, presence of multicollinearity, general F-test, and the value of OOS MAPE, we can determine an overall assessment of which is the best model out of our five contender models. This leaves us with the selection between two models:

Table 42: Model R vs Model T

Model	R	T
General F-test	Significant	Significant
R^2	Very strong	Very strong
Adjusted R^2	Very strong	Very strong
% of significant slopes with correct sign	37.50%	87.50%
Any significant slopes with incorrect sign?	No	No
Value of OOS MAPE	Very good	okay
Multicollinearity presence	High (7 occurrences)	Very high (9 occurrences)
Overall Assessment	Very good	Very good

We will select Model T. Although it has an okay value of OOS MAPE of about 30.7% and 9 occurrences of multicollinearity, it’s high R^ 2 and high percentage of significant slopes with correct signs indicates that we have many positive indicators that lead us to select this model. This now allows us to write our estimated form of our equation for Model T.

$$\widehat{SALES\imath} = \underset{(<.0001)}{2{,}685{,}493} - \underset{(0.0938)}{24.251}\text{Married\_pop}i + \underset{(0.1306)}{119.968}\text{Inc\_40kto100k}i + \underset{(0.0265)}{100{,}752}\text{champion}i$$

$$+ \underset{(<.0001)}{415{,}959}\text{cover\_charge}i - \underset{(0.2944)}{96{,}334}\text{W\_MW}i - \underset{(0.3668)}{104{,}457}\text{SW}i - \underset{(0.0034)}{241{,}114}\text{Hooters}i - \underset{(0.0051)}{177{,}975}\text{Buffalo}i + \varepsilon i$$

Where Ɛi is our error term, which means it is a residual variable that accounts for whatever our Model T is missing in fully represent the actual relationship between the independent variables and Sales.

Finally, we can evaluate the "attractiveness" of our estimated coefficients. Starting from the top, we consider married_pop as an attractive regressors since it helps us estimate how our sales perform under the condition that a city's population is married. This positive coefficient tells us that we want to consider new Buster's Brewhaus locations in cities that have a higher population that is married. We also consider Inc_40kto100k to be an attractive regressor since it points that placing our new locations in cities that have a population that has averages of 40k-100k in income will positively increase sales. We also see that if a store is located in a city that has won a professional sports championship within the last four years, sales at our Buster's locations are expected to increase. Which means that we should consider placing new locations in cities that have high records of winning professional sports championships within the last four years. Here, we see an interesting phenomenon with cover_charge. We see that if a store charges a cover charge to enter the bar on Fridays and Saturdays, sales at Buster's are expected to increase. This could be indicating that we get additional revenue from the cover charge, or that customers consider Buster's Brewhaus to be more exclusive and are willing to pay a higher premium for access.

We also see that if we place our new locations on the West, Midwest, or Southwest we will have a negative impact on sales. This indicates that we should steer away from placing any new locations in those regions and instead opting for example on placing new locations on the East, Northeast, Northwest regions. We also see a negative impact on sales if we have our competitors within 10 radial miles. Both the presence of one or more Hooters and Buffalo Wild Wings drive our current sales down. So, for future locations, we want to ensure to place our stores at least 10 miles or more further from our competitors. This will give us access to higher sales if our customers do not have any other sports bars alternatives in the area.

XV. Interpretations of the estimated coefficients:

•Married_pop: estimated coefficient = 24.251

For each additional married individual that lives within one radial mile of a store, sales at Buster's are expected to increase by about 24.251 dollars, ceteris paribus.

•Inc_40kto100k: estimated coefficient = 119.968

For every additional household within one radial mile of a store that earns a household income between $40,000 and $100,000, sales at Buster's are expected to increase by about 119.968 dollars, ceteris paribus.

•champion: estimated coefficient = 100,752

If a store is located in a city that has won a professional sports championship within the last 4 years, sales at Buster's are expected to increase by about $100,752, ceteris paribus.

•cover_charge: estimated coefficient = 415,959

If a store charges a cover charge to enter the bar on Fridays and Saturdays, sales at Buster's are expected to increase by about $415,959, ceteris paribus.

•W_MW: estimated coefficient = 96334

If a store is located in the West or Midwest, sales at Buster's are expected to decrease by about $96,334 , ceteris paribus.

•SW: estimated coefficient = 104,457

If a store is located in the Southwest, sales at Buster's are expected to decrease by about $104,457 , ceteris paribus.

•Hooters: estimated coefficient = 241,114

If there is one or more Hooters location within 10 radial miles of Busters, sales at Buster’s are expected to decrease by about $241,114, ceteris paribus.

•Buffalo: estimated coefficient = 177,975

If there is one or more Buffalo Wild Wings location within 10 radial miles of Busters, sales at Buster’s are expected to decrease by about $177,975, ceteris paribus.

XVI. Model Building through Sequential Regression

After identifying potential regressors and selecting which fit our regression model best based on our set parameters, we can now move on to the section of model building in our analysis. We can estimate and select the best fitting model that can more accurately explain sales at Buster’s Brewhaus by using our eight steps of model building. We have previously found the best fitting regression model and performed model validation through the calculation of OOS MAPE. We can now have a final list of potential regressors that were previously used to estimate the 20 alternative models and proceed to remove any multi-characteristic dummy variables.

Table 43: Temporarily Removed Multi-Characteristic Dummy Variables

Variable Name
W_MW
SW
stand_alone
strip_mall

Then, we can use collection of regressors to estimate the best possible model using the forward selection sequential regression method with our final list of potential regressors.

Table 44: Potential Regressors (Omission of Multi-Characteristic Dummy Variable)

Variable Name
Married_pop
Inc_40Kto100K
Occ_repair
Played_football
Played_baseball
Played_basketball
Played_hockey
restaurant_score
football
baseball
cover_charge
champion
Hooters
Buffalo

Although the best approach to create models is through our previous steps of model building, there is a distinct method in which the computer estimates every possible model given the current collection of potential regressors. From there, the computer selects the best model based on a given statistical criterion. Sequential regression uses algorithm-based decisions. It builds a regression model using the adjusted R2 selection model by producing the best one-variable model, then the best two-variable model, and so on up to the best (K-1) variable model in which K is the number of potential regressors, and finally a model that contains all potential regressors. From all these models estimated, the “best” model is the one with the largest adjusted R2. There are five different methods for sequential regression: forward-selection, step-wise selection, backward-elimination selection, "Maximum R2 Improvement" selection, and the "Adjusted R2" selection method.

Some of the disadvantages of using sequential regression is that there is no theory or logic in the selection of regressors, no consideration of correct signs, reasonable magnitudes, parsimony, and most importantly of multicollinearity. It also uses only one single fit statistic rather than a collection of criteria to select the best model. Sequential regression also has issues with including multi-characteristic dummy variables as potential regressors. It produces errors since the computer can re-define the base group and shift things around.

XVII. Forward Selection Sequential Regression Method

When using the "forward-selection" method, we must set an entry tolerance level, where our tolerance level SLENTRY is referring to a Type I error. This method first estimates all possible regression models that contain a single regressor. It then conducts an F test of H0: βj = 0 for the slope variable for every single model. After, it discards any model for which the p-value for that F test is greater than the tolerance level. If there are any models remaining the one with the smallest p-value is selected and named “Model 1”. The method then estimates all possible regression models that contain two regressors, follows the same steps as above, and then gives us a “Model 2” if any models remain. This process continues until we end up adding the total number of regressors. In our case we have 14 regressors and we want to set the tolerance level to 0.20. We will be using a value of 0.20 for SLENTRY.

Table 45: Forward Selection

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Hooters	1	0.8183	0.8183	39.7692	324.19	<.0001
2	cover_charge	2	0.0354	0.8537	20.3802	17.18	<.0001
3	Buffalo	3	0.0290	0.8826	4.8912	17.27	<.0001
4	champion	4	0.0069	0.8896	2.7081	4.33	0.0412
5	baseball	5	0.0032	0.8927	2.7910	2.01	0.1606

This now points us to the best model according to forward selection that has an R-Square of 0.8927 and C(p) of 2.791 with bounds on condition number being 4.0403 to 58.795.

Table 46: Analysis of Variance for Best Model Using Forward Selection

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.219578E13	2.439157E12	113.19	<.0001
Error	68	1.465391E12	21549868741		
Corrected Total	73	1.366118E13			

Table 47: Analysis of Variance for Best Model Using Forward Selection Part 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2674041	49635	6.254637E13	2902.40	<.0001
baseball	-76549	53965	43361070677	2.01	0.1606
cover_charge	328887	61200	6.223587E11	28.88	<.0001
champion	94444	42475	1.065444E11	4.94	0.0295
Hooters	-259224	44732	7.237001E11	33.58	<.0001
Buffalo	-232753	54355	3.951476E11	18.34	<.0001

After the computer selected this as the “best” model, we must consider if we want to add add any multi-characteristic dummy variables into the final model. When creating models that reincorporate these multi-characteristic dummy variables (W_MW, SW, stand_alone, and strip_mall), we see that those new models give us no significant change when adding them. In SAS, we created models that have these five regressors that were selected via forward selection. Additionally, these models then add W_MW and SW, stand_alone and strip_mall, and finally one with W_MW, SW, stand_alone, and strip_mall. The P-values given from each model indicate that these multi-characteristic dummy variables are insignificant and therefore can be tossed out of the final forward selection "best" model. R-Square scores also do not differ by much with the integration of these multi-characteristic dummy variables.

This "best" forward selection model will be called Model U. Model U has an R-Square of 0.8927. Compared to our previous five candidate models, Model U has a higher R-Square than just two out of five models. It has an 80% of significant slopes with the correct sign, which is higher than 4 of our contender models, with Model T having 87.5%. Model T has a higher R-Square of 0.8998. However, Model U has a higher Adjusted R-Square of 0.8848 while Model T has one of 0.8875. Model U and Model T both have champion, cover_charge, Hooters, and Buffalo as regressors. While Model T also incorporates Married_pop, inc_40kto100k, and the multi-characteristic dummy variables W_MW and SW (albeit W_MW and SW are insignificant in Model T). Model U in contrast has baseball as a regressor. Given all this, we still decide to go with "human selected" Model T.

XVIII. Stepwise Selection Sequential Regression Method

When using the "step-wise" method, regressors are selected in the same way as they are in the forward-selection method, but in order for a variable to remain in the model at each successive step, it must be statistically significant at the "stay tolerance level". This term refers to our tolerance for a Type I error for variables that end up staying in the model. Again, we have 14 regressors and we want to set the tolerance level to 0.20. But we also want to set our stay tolerance level, SLSTAY, at 0.20.

Table 48: Stepwise Selection

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Hooters		1	0.8183	0.8183	39.7692	324.19	<.0001

2	cover_charge		2	0.0354	0.8537	20.3802	17.18	<.0001
3	Buffalo		3	0.0290	0.8826	4.8912	17.27	<.0001
4	champion		4	0.0069	0.8896	2.7081	4.33	0.0412
5	baseball		5	0.0032	0.8927	2.7910	2.01	0.1606

We actually end up getting the exact same model as with forward selection, Model U. Which means that the step-wise selection regression method produced the exact same model. Again, we end up selecting "human selected Model T".

XIX. Backward-Elimination Selection Sequential Regression Method

When building a regression model using the backward-elimination selection method, regressors are selected in a way that is the reverse of the forward-selection method.

Table 49: Backward Selection

Summary of Backward Elimination							
Step	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	played_football	13	0.0001	0.9023	13.0373	0.04	0.8475
2	played_basketball	12	0.0001	0.9021	11.1048	0.07	0.7943
3	Occ_repair	11	0.0002	0.9019	9.2282	0.13	0.7224
4	played_baseball	10	0.0004	0.9015	7.4871	0.27	0.6045
5	football	9	0.0014	0.9001	6.3177	0.88	0.3519
6	restaurant_score	8	0.0018	0.8983	5.4173	1.17	0.2841
7	played_hockey	7	0.0016	0.8968	4.3567	0.99	0.3224

The backward elimination method actually ends up giving us a very different model. Just by sight we can tell that all of the P-values are insignificant. Even if we create a Model Y with these regressors, our R-Square is 0.5910 which is much lower than our five candidate models and Model U that was given by forward selection and step-wise selection methods. Therefore, we can omit the result for this model and move on to the next step.

XX. Maximum R-Squared Improvement Sequential Regression Method*

When building a regression model using the maximum R-Square improvement selection method, we are not given a “best” model as our previous methods did. With this method, the best one-variable model is produced, the best two-variable model is produced, and so on up to the best (K – 1) variable model (where K is the number of potential regressors), and a model that contains all potential regressors is estimated. Under this method, the "best" model is the one with the largest R-Square.

Once we run our code in SAS, we find that in this case there are two models that share the same R-Squared and are the highest of all models. The R-Square is 0.9023 for both of these models. The models are the best 13-variable model found and the best 14-variable model found. Both models also have five significant regressors that in this case are the same five: baseball, cover_charge, champion, Hooters, and Buffalo. The only difference is that the 14-variable model adds played_football, which we found to be insignificant in both the 13-variable and 14-variable model. Hence, we can turn our attention to the 13-variable model which we will refer to as Model Z.

If we were to select Model Z purely out of having the highest R-Square score of 0.9023 out of all of our models, we would be ignoring statistical significance in our regressors and its effect on our overall model. Model Z gives us only five out of thirteen regressors that are statistically significant with the correct sign, which is just 38.46%, compared to “human selected” Model T which has 87.5% statistically significant regressors with the correct sign. Although Model T has a lower R-Square of 0.8998 which is lower than Model Z’s 0.9023, we would still move forward with the selection of Model T due to the higher percentage of statistically significant regressors in the model.

XXI. Adjusted R-Squared Sequential Regression Method

Finally, we turn our attention to the adjusted R-Squared selection method. In this method, there is no single model that is selected as the "best." It follows the same process as the maximum R-Square improvement selection method with the difference that the "best" model is the one with the largest adjusted R-Square.

With the given results in SAS, we create a Model AA which has the highest Adjusted R-Square of 0.8870, an R-Square of 0.8994, and eight regressors (five out of eight being statistically significant). Both Model T and Model AA have Married_pop, Inc_40Kto100K, cover_charge, champion, Hooters, and Buffalo. However, Model AA has played_baseball and baseball. While Model T has the statistically insignificant variables W_MW and SW. Model T has a 0.8875 R-Square and 87.5% of statistically significant variables compared to Model AA's 62.5%. Even if we incorporate W_MW and SW to Model AA and call it Model AB, the Adjusted R-Square of this Model AB is 0.8880 which isn't that much higher than Model AA's and the percentage of statistically significant variables actually drops from 62.65% to 50%.

XXII. Sales Volume Analysis

In order to analyze potential sales volume, we can compute the three threshold values of sales that divide Sales into four categories: low-potential sales, above-average potential sales, medium potential sales, and high potential sales. Now that we have selected our "best" regression model, we can now consider what potential locations that Buster's Brewhaus is considering opening are projected to produce higher profits. We can define our benchmarks for what is considered low, medium, above average, and high potential sales.

Table 50: Sales Potential Definition

Sales Potential	**Threshold**
Low Sales Potential	Projected Sales < Average Sales
Above Average Sales Potential	Average Sales < Projected Sales < (Average Sales + 1 standard deviation)
Medium Sales Potential	(Average Sales + 1 std. dev.) ≤ Projected Sales < (Average Sales + 2 standard deviations)
High Sales Potential	Projected Sales ≥ (average Sales + 2 standard deviations)

Using our corrected data set with the 74 observations, we got that our average sales were equal to $2,433,398.55 and our standard deviation being equal to $432,596.11.

Table 51: Sales Potential Figures

Sales Level	Value of Sales
Average Sales:	$2,433,398.55
Average + 1 standard deviation:	$2,865,994.66
Average + 2 standard deviations:	$3,298,590.77

Table 52: Sales Potential and Threshold Values

Sales Potential	**Threshold Value**
Below Average Sales Potential	Projected Sales < $2,433,398.55
Above Average Sales Potential	$2,433,398.55 < Projected Sales < $2,865,994.66
Medium Sales Potential	$2,865,994.66 < Projected Sales < $3,298,590.77
High Sales Potential	Projected Sales > $3,298,590.77

We have five potential stores where we could place a location, Storeid being from 83 to 87. Using Model T, we calculated the projected sales for these five potential locations.

Table 53: Potential Buster's Brewhaus Locations and Sales Potential

Storeid	**Projected Sales**	**Sales Potential**
Store 85	$2,412,491.57	Below Average

Store 86	$2,454,692.93	Above Average
Store 84	$2,490,775.20	Above Average
Store 83	$2,623,764.32	Above Average
Store 87	$2,682,100.02	Above Average

We have now scored each location based on their sales potential using Model T. Store 85 gives us the lowest projected sales and falls under the below average sales potential category. Hence, Buster's should not consider this location. Store 86, 84, 83, and 87 all fall under the above average sales potential category and should therefore we can advise that these locations are candidates for our new locations. These above average figures indicate positive profits and will give us above average sales and increase our overall revenue. Store 87 has the highest projected sales figure of all the potential locations.

We can consider how the regressors will impact our sales figures. We consider those factors that are positively related and increase sales. Therefore, the best locations for Buster's are locations that are in cities that have a population that have averages of 40k-100k in income, locations that are popular enough to charge a cover charge on the weekend and are located in a city where a local team has won a sports championship within the last 4 years. Among those factors that will have negative impact on sales we have if the potential location is located in the West-Midwest or Southwest region, if married population is high, and if we have our competitors—Hooters and Buffalo Wild Wings—within 10 radial miles. For future locations, we want to ensure to place our stores at least 10 miles or more further from our competitors, Hooters and Buffalo Wild Wings. This will give us access to higher sales if our customers do not have any other sports bars alternatives in the area.

Given this, we know that potential Store 87 would be impacted positively by the facts that there is a good amount of people that earn an income between 40k-100k, charges a cover charge on the weekends, has won a sports championship within the last 4 years, and has no Buffalo Wild Wings within 10 radial miles. Sales would be impacted negatively since Store 87 has a high married population, it is located in the West-Midwest region, and has two Hooters within 10 radial miles.

Appendix

Model T

Dependent Variable: SALES	
Number of Observations Read	74
Number of Observations Used	74

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	1.23E+13	1.54E+12	72.96	<.0001
Error	65	1.37E+12	21060385559		
Corrected Total	73	1.37E+13			

Root MSE	145122	**R-Square**	0.8998
Dependent Mean	2433399	**Adj R-Sq**	0.8875
Coeff Var	5.96376		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	2685493	81166	33.09	<.0001
Married_pop	**1**	-24.251	14.2606	-1.7	0.0938
Inc_40Kto100K	**1**	119.968	78.3624	1.53	0.1306
champion	**1**	100752	44367	2.27	0.0265
cover_charge	**1**	415959	82582	5.04	<.0001

W_MW	**1**	-96334	91140	-1.06	0.2944
SW	**1**	-104457	114941	-0.91	0.3668
Hooters	**1**	-241114	79288	-3.04	0.0034
Buffalo	**1**	-177975	61464	-2.9	0.0051

Model U

Dependent Variable: SALES	
Number of Observations Read	74
Number of Observations Used	74

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	1.22E+13	2.44E+12	113.19	<.0001
Error	68	1.47E+12	21549868741		
Corrected Total	73	1.37E+13			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	2674041	49635	53.87	<.0001
baseball	**1**	-76549	53965	-1.42	0.1606
cover_charge	**1**	328887	61200	5.37	<.0001

champion	**1**	94444	42475	2.22	0.0295
Hooters	**1**	-259224	44732	-5.8	<.0001
Buffalo	**1**	-232753	54355	-4.28	<.0001