Project 6

By: Will Hallgren

February/3/2021

ECON 5645

University of North Texas

Why does Buster's sales matter?

Buster's Brewhaus is a popular sports bar with many locations throughout the United States. Buster's aims to provide the classic American, college-town sports bar experience for a reasonable price. Buster's offers traditional bar food, appealing mainly to sports-goers. They also offer a variety of pizzas including both New York-style thin crust pizza and Chicago-style deep dish pizza. Most importantly, Busters offers 110 different beers from around the world, 30 of which are on tap. They also offer standard cocktails and a modest selection of whiskey. The atmosphere is comfortable and informal.

Buster's has done well in 2019 and we would like to know more about the factors that influence the sales at Buster's Brewhaus so that they can continue to expand. We have taken demographic data from people who live close to each location in an attempt to find a link between certain demographics and increased sales. This information would give us an idea about which locations would be most profitable for future development, and who the target audience is for marketing.

The purpose of this paper is to build and estimate the best fitting models to explain sales at Buster's given the potential regressors found during pre-model analysis.

Variables:

In this regression, the dependent variable is Sales. Sales is the dollar value of sales at each store in 2019. The goal of this analysis is to determine the factors that influence sales. Store ID is a unique number from 1 to 78 identifying each store. Close date is simply the date that the store closed if applicable.

Pre-Model Analysis

Graph 1: Raw Sales Data

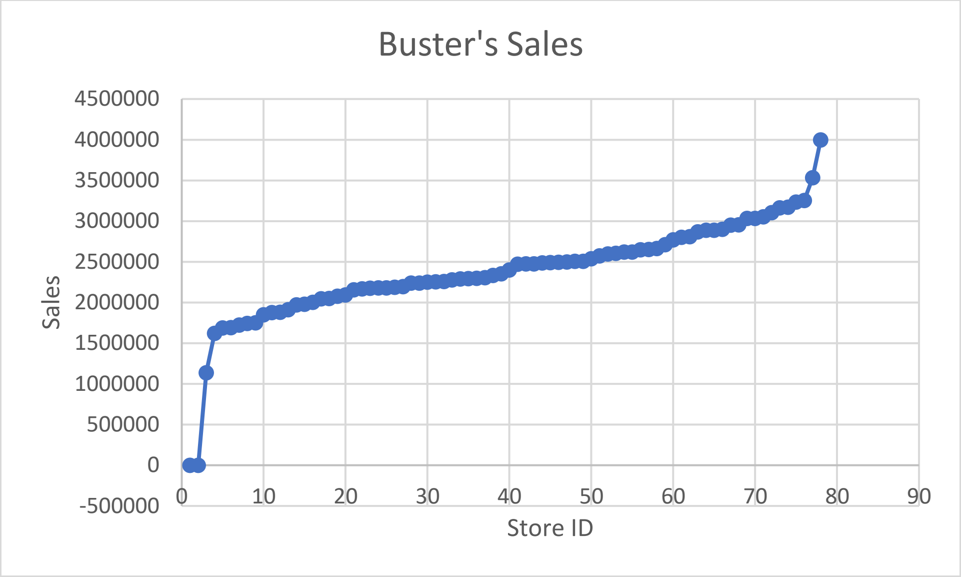

Table 1: Summary statistics for Sales:

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

Some values in this data set are unreasonable. Sales should be a number greater than zero. Stores 1 and 2 have values for sales that are -770.85 and 0 respectively. Because these values are unreasonable in this analysis and I have no reason to believe that the source of these unreasonable values was data entry error, I will strike them from the data set. These two observations were the only two observations that had an entry for close date, so close date data will be ignored.

Graph 2: Buster’s Sales data with Unreasonable Values Removed

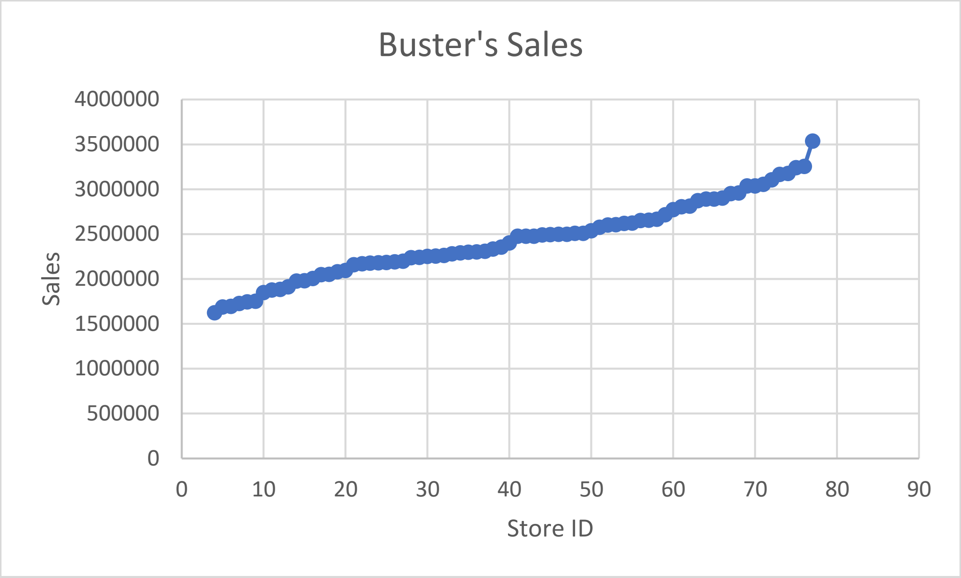

Table 2: Summary statistics for Sales with Unreasonable Values Removed:

# of Observations	Mean	Standard Deviation	Minimum Value	Maximum Value
76	2,436,925.39	487,021.26	1,136,846.32	3,997,991.13

Unreasonable values have been removed, however there are some outliers in this data as well. Outlier observations should be removed from the data as well since they do not represent the average trend of sales. They can cause parameter estimates to be inaccurate, which can produce misleading conclusions.

Mean Sales + 2.5(Standard deviation) = 2,436,925.39 + 2.5(487,021.26) = 3,654,478.85

Mean Sales - 2.5(Standard deviation) = 2,436,925.39 - 2.5(487,021.26) = 1,219,372.24

I will remove any observations in the data for sales larger than $3,654,478.85 and smaller than $1,219,372.24

Graph 3: Busters Sales with Outliers Removed

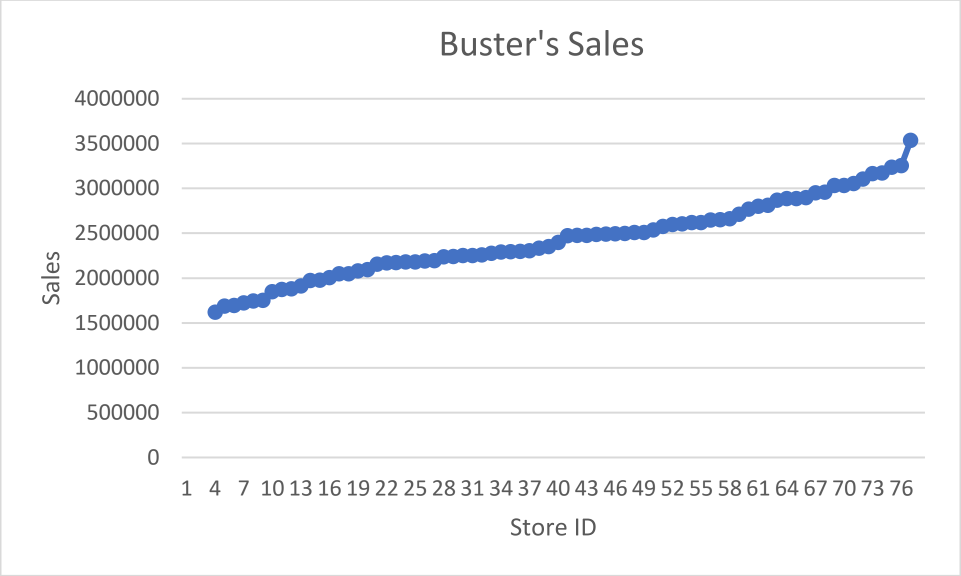

Table 3: Summary Statistics for Sales with Outliers Removed

# of Observations	Mean	Standard Deviation	Minimum Value	Maximum Value	Coefficient of Variation
74	2,433,398.55	432,596.11	1,620,139.23	3,535,412.28	17.78

All unreasonable observations and outliers have been removed. The dependent variable must also have sufficient variation in order to draw conclusions from the data. If the dependent variable does not have sufficient variation, then we do not have evidence that sales vary across stores and analyzing the differences between them would be frivolous. The coefficient of variation is the standard deviation divided by the mean, multiplied by 100. For continuous dependent variables, such as sales, the coefficient of variation should be greater than 2. The calculated Coefficient of Variation for sales is 17.78, well above our minimum value of 2, indicating that the dependent variable is varied enough for this regression.

Table 4: Total Observations Removed from Dependent variable

Observation Number	*Reason*
1	Sales were negative
2	Sales were zero
3	Sales were below the minimum outlier cutoff
78	Sales were above the maximum outlier cutoff

Potential Problems with Independent variable data

First, I check the summary statistics and Coefficient of variation of all variables. I am looking for missing values, unreasonable numbers, and outliers. I also check to see if the Coefficient of Variation is greater than 2 for all variables. We desire variation in the independent variable data as well as the dependent variable data because there will be no unique solution for a parameter estimate when we begin running regressions.

Table 5: Summary Statistics for Store ID and Population 45-50

Variable	Number of Observations	Mean	Standard Deviation	Minimum	Maximum	Coefficient of Variation
Store ID	78	39.5	22.66	1	78	57.37
Population 45-50	77	152.78	116.72	0	464	76.4

The only problem with this data is that, “Population 45-50” is missing an observation. Upon contacting the data collector, I determined that the best course of action is to replace the missing observation with the mean, rounding up to 153 since number of people is a discrete number.

A Brief explanation on Micronumerosity

Micronumerosity is a condition of the data in which there are too few degrees of freedom. When there are too few degrees of freedom, the accuracy of our regression will be low. We may accept a hypothesis that otherwise would be rejected if the data had more degrees of freedom.

In short, the degrees of freedom in statistics is the difference between the number of observations and the number of variables being tested. In this case, we are testing 78 observations against 32 variables.

78-32 = 46 degrees of freedom

In general, we need more than thirty degrees of freedom to prevent the micronumerosity issue. Since we have forty-six degrees of freedom, we will have no problems with micronumerosity in this data set.

Correlation

To see which variables would best fit in our regression, I will test the correlation between Sales and each variable tested to see which correlations are statistically significant. Some of our variables may be more correlated to sales than others and I want to narrow down a list of potential regressors so that we know which factors have the most influence on sales. I will use the hypothesis:

$$H_0: Correlation\ Coeff. = 0\ \ vs.\ \ H_a: Correlation\ Coeff. \neq\ 0$$

I will use the Pearson correlation coefficient to evaluate the correlation between independent variables and sales. It is the covariance between sales and one of the independent variables divided by the product of their standard deviations. The p-value given by a hypothesis test is the probability that we obtain test results at least as extreme as the results observed. Therefore, we want to minimize the p-value if we are looking for variables with a significant effect on sales. If the p-value in the hypothesis test is less than .10, then we reject the null hypothesis; we have sufficient statistical evidence to believe that our tested independent variable is correlated with sales. The null hypothesis states that our variable being tested is not correlated to sales at all. The correlation coefficient would equal 0. Below I have sorted the variables into six different categories where I give both the Correlation Coefficient to sales and P-value for each variable.

Correlation Tables with Sales:

Table 6: Age

	Population aged 45 to 50
Correlation Coefficient with sales	0.19729
P-value	0.0920

Table 7: Race

	Hispanic	**Caucasian**	**African American**	**Asian**
Correlation Coefficient with sales	0.01752	0.07077	0.03909	0.02501
P-value	0.8822	0.5491	0.7409	0.8325

Table 8: Education

	Bachelors	**Masters and above**
Correlation Coefficient with sales	0.06077	-0.01405
P-value	0.6070	0.9054

Table 9: Occupation

	Business	Financial	Computer	Engineer	Social Science	Repair
Correlation Coefficient with sales	-0.07231	-0.00856	-0.02501	0.23490	-0.04858	0.23950
P-value	0.5404	0.9423	0.8325	0.0440	0.6811	0.0399

Table 10: Recreational Interests

	Baseball	Basketball	Bowling	Football	Hockey	Volleyball	Yoga	Exercise Regularly
Correlation Coefficient with 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: Economic Activity

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

From this information, I can compile a list of nine potential regressors that are significantly correlated with sales. Level of education and race had very little correlation with sales while recreational preferences seem to have the highest correlation. Occupation had very little correlation with sales with the exception of engineering and repair.

Table 12: Potential Regressors

Potential Regressor	Correlation Coefficient	P-value
Engineering occupation in the area	0.23490	0.0440
Repair occupation in the area	0.23950	0.0399
Baseball players in the area	0.22251	0.0567
Basketball players in the area	0.25743	0.0268
Played Bowling players in the area	0.26449	0.0228
Played Football players in the area	0.24723	0.0337
Played Hockey players in the area	0.30242	0.0088
Restaurant Score	0.22524	0.0537
Nightlife Score	-0.22646	0.0524

Table 13: Potential Independent Variables for Regression

Variable:	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 13: Potential Independent Variables for Regression

Variable:	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 ore 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

Single-Trait Dummy Variables

Table 14: Variable Definitions for Existing Variables

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 1 radial mile of a store, and 0 if not.
University	=1 if there is University located within 1 radial mile of a store, and 0 if not.
CC	=1 if a store does not require a cover charge to enter the bar on Fridays and Saturdays, and 2 if it does.
DT	= “yes” if there the store has a drive-thru for take out food and beverages, and “no” if not.
BT	= “high” if a store is located in a city that has a bar tax greater than 9%, and “low” if not.
Champ	= “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.

Notice that cover charge, drive-thru, bar tax, and championship are all recorded in the data as “alpha variables” or variables that use letters or words to identify qualitative data. For the SAS to properly analyze the data, these variables must be edited so that they only include values of 0 and 1. If any numbers besides 0 and 1 are used, then values designated by higher numbers will show a higher effect than values designated by low numbers and the results will be misleading.

I have created new variables out of CC, DT, BT, and Champ so that they only contain values 0 and 1 and rewritten the names so that they are more recognizable at a glance.

Table 15: Variable Definitions for Newly Created Variables

Variable:	**Description:**
Cover_charge	=0 if a store does not require a cover charge to enter the bar on Fridays and Saturday (that is, if cc= 2, and 1 if it does (that is, if cc= 1).
Drive_thru	=0 if a store does not have a drive-thru (that is, if DT= no), and 1 if it does (that is, if DT= yes).
Bar_tax	=1 if a city where a store is located has a high tax rate (that is, if BT= high), and 0 if the city has a low tax rate (that is, if BT= low).
Sports_champ	=1 if a city where a store is located has won a professional sports championship within the last 4 years (that is, if champ= Y), 0 if not (that is, if champ= N).

Table 16: Summary Statistics for Single Trait Dummy Variables

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
Sports_champ	74	0.23	0.42	0	1

There are some issues with the raw data. The variable, "university" has too high of a mean. With dummy variables, if the mean is above .90, then the data is not varied enough. Too many Buster's locations are near a university, so a regression including this variable will give inaccurate estimates on its effect on sales. I think that it is a reasonable assumption to assume that locating a sports bar near a university will have a positive effect on sales, so I will remove, "University" from this regression. The variable, "Music" has too low of a mean. Too few Buster's locations provide music on the weekends for the regression to accurately model the effect of providing music on weekends on sales.

Table 17: Correlation of Single-Trait Dummy Variables with Sales

	Football	**Baseball**	**Basketball**	**Soccer**
Correlation to Sales	0.74228	0.48689	0.42647	-0.13607
P-value	<.0001	<.0001	0.0002	0.2477

	Cover_charge	**Drive_thru**	**Bar_tax**	**Sports_champ**
Correlation to Sales	0.78845	-0.02549	-0.36442	0.31965
P-value	<.0001	0.8293	0.0014	0.0055

Next, I will do a hypothesis test to find the variables with the strongest correlation to sales. The hypothesis being tested is that the true correlation between any of the variables and sales is 0; that is, they are insignificant. The p-value is the probability of obtaining test results that are at least as extreme as the results observed, so a small p-value is evidence that we should reject the hypothesis, meaning that we have found statistically significant correlation between sales and the variable. In this case, we are looking for p-values that are equal to or less than .10, a 90% confidence level. The variables Football, Baseball, Basketball, cover charge, bar tax, and sports champ all have p-values that are below .10. Football having a positive correlation coefficient and a P-value of <.0001 can be interpreted as, we have strong statistical evidence to suggest that on average, locating near a football stadium is expected to increase sales. The same goes for Baseball and Basketball stadiums, according to the data. Soccer has a high P-value, which means that we do not reject the null hypothesis. We do not have strong statistical evidence to conclude that the presence of a soccer stadium within one radial mile of a Buster's location is related to sales at all. Drive thru also has a high P-value, indicating that we do not have statistical evidence to conclude that the presence of a drive thru is correlated with sales. Cover_charge is positively correlated with sales and has an extremely significant P-value. On average, we expect bars that charge a cover charge to increase sales. Bar tax is negatively correlated with sales meaning, on average, we expect Buster's locations in cities with high taxes to lower sales. Sports Champ is also positive and significant. Locating in a city where a professional sports team has won a championship within the past four years is expected on average, to increase sales.

Table 18: Final Single-Trait Dummy Variables List

Variable:	Description:
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.
Cover_charge	=0 if a store does not require a cover charge to enter the bar on Fridays and Saturday (that is, if cc= 2, and 1 if it does (that is, if cc= 1).
Bar_tax	=1 if a city where a store is located has a high tax rate (that is, if BT= high), and 0 if the city has a low tax rate (that is, if BT= low).
Sports_champ	=1 if a city where a store is located has won a professional sports championship within the last 4 years (that is, if champ= Y), 0 if not (that is, if champ= N).

Multi-Trait Dummy Variables

Table 19: Multi-Trait Dummy Variable Descriptions

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.
Life_style	=1 if the store is located in a lifestyle mall, and 0 if not.
Pop_growth	Takes on a value of "1" if the rate of population growth is considered to be high, "2" if the rate of population growth is considered to be medium, "3" if the rate of population growth is considered to be low, and "4" if the rate of population growth is negative.
Region	Takes on a value of "W" if the store is located in the Western region of the United states, "MWest" if the store is located in the Midwest, "SWest" if the store is located in the Southwest, and "e" if the store is located in the east.

The population growth and region variables must be edited so that I am able to run a regression in SAS. The population growth variables are designated by numbers other than zero or one, which may skew the results. The region variables are designated by letters, which cannot be used by SAS. I have created new dummy variables out of the data that was provided by the data collector.

Table 20: Variable Definitions for Newly Created Multi-Trait Dummy Variables

Variable:	Description:
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 21: Summary Statistics for Type of Building

Variable	Number of Obs.	Mean	Std Dev	Minimum	Maximum
Stand Alone	74	0.3	0.46	0	1
Strip Mall	74	0.35	0.48	0	1
Lifestyle	74	0.35	0.48	0	1

Table 22: Summary Statistics for Population Growth

Variable	Number of Obs.	Mean	Std Dev	Minimum	Maximum
High Pop. Growth	74	0.05	0.23	0	1
Medium Pop. Growth	74	0.27	0.45	0	1
Low Pop. Growth	74	0.41	0.49	0	1
Negative Pop. Growth	74	0.27	0.45	0	1

Table 23: Summary Statistics for Regions

Variable	Number of Obs.	Mean	Std Dev	Minimum	Maximum
West Locations	74	0.03	0.16	0	1
Midwest Locations	74	0.27	0.45	0	1
Southwest Locations	74	0.32	0.47	0	1
East Locations	74	0.38	0.49	0	1

For pre-model analysis, I am looking for sufficient variation in the variables. The variables must be sufficiently varied because we need to be able to compare locations that have the trait to locations that do not. Insufficient variation means that too many observations possess the same trait and we cannot definitively say that this trait impacts sales because there is not enough data to compare it against. Dummy variables cannot be evaluated with the coefficient of variation because of the nature of the data. Dummy variables can only take on values of zero and one, otherwise the information is skewed in favor of the variables that take on higher numbers as an identity. Instead, we check to see if the mean of each variable is less than 0.1 or more than 0.9. Either case indicates that too many observations in the variable possess the same trait.

Table 21 and Table 22 both contain variables that do not have sufficient variation. The variables, “High Population Growth” and, “West” do not have a high enough mean to be considered sufficiently varied. Rather than throwing out the variable entirely, it is more appropriate to try to combine a variable with insufficient variation with another variable from the same qualitative set. Ideally, we would combine it with a variable that is logically, most likely to behave similarly. I will combine, “High Population Growth” with, “Medium Population Growth” to form the variable, “Medium-High Population Growth. I will combine, “West Locations” with, “Midwest Locations” to form the variable “West-Midwest Locations”. Below is a list of summary statistics for the adjusted dummy variables.

Table 24: Summary Statistics for Adjusted Dummy Variables

Variable	Number of Obs.	Mean	Std Dev	Minimum	Maximum
Medium-High Pop. Growth	74	0.32	0.47	0	1
West-Midwest Locations	74	0.3	0.46	0	1

Both means are above 0.1 and below 0.9, therefore sufficient variation exists in the adjusted variables.

Correlation Between Dummy Variables

Table 25: Correlation with Sales for Building Type

	Standalone	Strip Mall	Life Style
Correlation with Sales	0.80	-0.76	0.0009
P-Values	<.0001	<.0001	0.99

Table 26: Correlation with Sales for Population Growth

	Medium-High Pop. Growth	Low Pop. Growth	Negative Pop. Growth
Correlation with Sales	0.80	-0.11	-0.72
P-Values	<.0001	0.35	<.0001

Table 27: Correlation with Sales for Region

	West-Midwest Locations	Soutwest Locations	East Locations
Correlation with Sales	-0.74	0.80	-0.079
P-Values	<.0001	<.0001	0.5

I will use the Pearson correlation coefficient to evaluate the correlation between the multi-trait dummy variables and sales. It is the covariance between sales and one of the variables divided by the product of their standard deviations. A negative figure suggests that on average, the variable tends to move opposite with sales. A positive figure suggests that on average, the variable tends to move with sales. Then I test for significance using the null hypothesis that the true correlation coefficient is zero. The p-value given by a hypothesis test is the probability that we obtain test results at least as extreme as the ones observed. Therefore, we want the P-values to be small when looking for statistically significant variables. Usually the P-values should be no greater than 0.1, which gives us a 90% confidence level in our results.

Standalone and strip-mall-type buildings both had statistically significant P-values. Standalone-type buildings have a positive correlation coefficient, which suggests that on average standalone buildings have a positive influence on sales. Strip-mall-type buildings had a negative correlation coefficient, which suggests that on average, strip-mall locations tend to have a negative impact on sales. Lifestyle-type buildings did not have a statistically significant P-value, indicating that we do not have strong enough statistical evidence to suggest that Lifestyle buildings are correlated with sales. Medium-High population growth and negative population growth both had statistically significant P-values, however low population growth did not. There is not strong enough statistical evidence to suggest that low population growth is correlated to sales. Medium-High population growth had a positive correlation coefficient with sales, suggesting that on average, medium-high population growth tends to have a positive impact on sales. Negative population growth had a negative correlation coefficient, suggesting that on average, negative population growth tends to have a negative impact on sales. As for regions, West-Midwest and Southwest were both had statistically significant P-values but East did not. We do not have strong enough statistical evidence to suggest that East is correlated with Sales. West-Midwest had a negative correlation coefficient with sales suggesting that on average, West-Midwest locations tend to have a positive impact on sales. Southwest had a strong, positive correlation coefficient with sales. On average, Southwest locations tend to have a strong, positive impact on Sales.

Limited Integer Value (LIV) Variables

Limited integer value variables are variables that can only take on a very limited number of integer values, usually six or fewer. The LIV variable data gathered for this study is on the proximity and prevalence of competitors to Busters.

Table 28: List of 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.

To determine whether LIV variables have sufficient variation, each outcome in the variable must comprise at least 10% of the observations in the sample. There is a problem with sufficient variation with three out of the four LIV variables in this sample. The only variable with sufficient variation is the "Hooters" data. The following figures contain the frequency tables of the variables that violate the sufficient variation condition.

Table 29: Twin Peaks Frequency Chart

# of Twin Peaks	Frequency	Percent
0	69	93.24
1	3	4.05
2	2	2.7

Table 30: Buffalo Wild Wings Frequency Chart

# of Buffalo Wild Wings	Frequency	Percent
0	43	58.11
1	21	28.38
2	7	9.46
3	3	4.05

Table 31: Metrics Frequency Chart

# of Metrics	Frequency	Percent
0	7	9.46
1	8	10.81
2	8	10.81
3	7	9.46
4	7	9.46
5	8	10.81
6	6	8.11
7	5	6.76
8	7	9.46
9	5	6.76
10	6	8.11

LIV variables may be manipulated into dummy variables if they do not possess sufficient variation. We can combine the rest of the observations and make the variable equal "0" if there are no locations, or "1" if there are any locations, regardless of frequency. According to figure 8, the Twin Peaks variable cannot be changed to have sufficient variation. The combined share of Buster's locations having a Twin Peaks location within a 10-mile radius at all is 6.75%. Therefore, the Twin Peaks data must be removed from this regression. Metrics does not have the characteristics of an LIV variable. An

LIV variable generally takes on values of six or fewer. Metrics can take on a value of 10, making it an ineligible LIV variable. I can however use just Metrics as a normal, discrete random variable. It has a high enough correlation coefficient of 67.39 to be considered for regression. Buffalo Wild Wings can be corrected into a dummy variable.

Correlation with Sales

Now that I have adjusted the LIV variables, I will evaluate their correlation with sales. Again, I will use the Pearson Correlation Coefficient. Just to reiterate what I am looking for: a negative figure identifies a negative impact on sales. A positive figure identifies a positive impact on sales. A P-value of 0.1 or less identifies a variable that has statistical significance, giving us a 90% confidence level.

Table 32: Competitors' Correlation with Sales

	Hooters	Buffalo Wild Wings	Metrics
Correlation with Sales	-0.90458	-0.79144	-0.97438
P-Value	<.0001	<.0001	<.0001

Both Variables are highly statistically significant because the P-values are both very small. Both Hooters and Buffalo Wild Wings have a negative correlation coefficient, indicating that the presence of both competitors has a negative impact on sales. Both are suitable for regression.

Table 33: List of Potential Regressors from Multi-Trait Dummies and LIV Variables

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 categorized as having ether medium or high population growth, 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 Locations	=1 if the store is located in ether the West region or Midwest region of the US, and 0 if not.
Southwest Locations	=1 if the store is located in the Southwest region of the US, and 0 if not.
Hooters	=The number of “Hooters” locations within 10 radial miles of Busters.
Buffalo Wild Wings	=1 of there is a “Buffalo Wild Wings” location within 10 radial miles of Busters, =0 otherwise.
Metrics	=The number of “Metrics” locations within 10 radial miles of Busters.

Model Building and Inference in Regression

With the list of potential variables suitable for regression, now I will build and estimate the best fitting model to explain sales at Buster's given the potential regressors found during pre-model analysis. I will regress 22 different combinations of regressors and compare them against each other. The "best fitting" model is the model that has intuitively correct signs, P-values that are small enough to indicate statistical significance, a high R Squared, a high adjusted R squared, and low out-of-sample mean absolute percentage error (OOS MAPE).

Considerations on Theory and Logic

After going about pre-model analysis, it is important to consider demand theory and logical reasoning so that we do not miss out on any other potential regressors that may improve our understanding of the factors that influence sales. Demand theory suggests that sales would be related to: own price of goods, income, population, preferences, prices of substitutes and complements, expectations for the future. I believe that the current set of potential regressors sufficiently encompasses demand theory. Another variable may be added in regards to varying price level across locations; for example, a Buster's location in Los Angeles, CA may have a different price level than a Buster's location in San Angelo, TX. This discrepancy could potentially influence sales however I believe that the effect would be small considering that this is mainly due to cost-of-living differences and has little to do with differences in demand, therefore we do not need to expend more resources collecting data on this variable. This data would most likely be highly correlated with regional data as well, which we already have. Consumer preference data may also be useful. Buster's offers a wide variety of bar food but some consumers may prefer Buffalo Wild Wings or Hooters because they prefer a bar that specializes in chicken wings.

Figure 1: Potential Regressors List

1. Engineering Occupation
2. Repair Occupation
3. Played Baseball
4. Played Basketball
5. Played Bowling
6. Played Football
7. Played Hockey
8. Restaurant Score
9. Nightlife Score
10. Football Field
11. Baseball field
12. Cover Charge
13. Bar Tax
14. Sports Championship
15. Standalone Building
16. Strip mall Building
17. Med-High population Growth
18. West-Midwest Location
19. Southwest Location
20. Hooters
21. Buffalo Wild Wings
22. Metrics

Table 34: Top 5 Regression Models

Model:	Regressors:
H	Engineering occupation, repair occupation, restaurant score, nightlife score, cover charge, high tax, West-Midwest, Southwest
J	Played baseball, played basketball, played bowling, played football, played hockey, cover charge, high tax, Hooters, Buffalo Wild Wings
P	Football field, cover charge, high tax, championship, stand-alone, strip-mall, Hooters, Buffalo Wild Wings
T	Played baseball, played basketball, played hockey, championship, stand-alone, strip-mall, Hooters, Buffalo Wild Wings
U	Played baseball, played basketball, played hockey, football field, cover charge, high tax, Hooters, Buffalo Wild Wings

Table 35: Fit Statistics for the Top 5 Regression Models

Model	**Calc. General F Test (P-value)**	R^2	$\bar{R}^2$	**# of significant slopes with correct sign**	**# of significant slopes with incorrect sign**	**Value of OOS MAPE**	**Joint F test testing significance of region dummy variables (P-value)**
H	48.67 (<.0001)	0.8569	0.8393	3 out of 3	0	0.26	41.8 (<.0001)
J	58.61 (<.0001)	0.8918	0.8766	3 out of 3	0	0.20	NA
P	69.06 (<.0001)	0.8947	0.8818	4 out of 4	0	0.24	NA
T	58.00 (<.0001)	0.8771	0.8620	3 out of 3	0	0.18	NA
U	65.68 (<.0001)	0.8899	0.8764	3 out of 3	0	0.20	NA

Table 36: Summary and Assessment of the Top 5 Regresion Models

Model	**General F Test**	R^2	$\overline{R}^2$	**% Significant slopes with correct sign**	**Any significant regressors with the wrong sign**	**OOS MAPE**	**Overall Assessment**
H	Significant	Very Strong	Very Strong	100%	0	Fair	Good
J	Significant	Very Strong	Very Strong	100%	0	Good	Good
P	Significant	Very Strong	Very Strong	100%	0	Fair	Good
T	Significant	Very Strong	Very Strong	100%	0	Very Good	Good
U	Significant	Very Strong	Very Strong	100%	0	Good	Good

Multicollinearity

Next, I will evaluate the independent variables in the top five models for multicollinearity. Multicollinearity exists in the model if the independent variables are highly correlated with each other. If any of the variables are too correlated, then the variance of the estimated regression parameters will be over-inflated, test statistics will be artificially small, and the magnitude of the parameter estimates that we get from regression and their estimated signs may be counter-intuitive. We are more likely to fail to reject a null hypothesis that is false and we may get parameter estimates that suggest a counter-intuitive relationship between the independent variables and sales. This means that the accuracy of the model is overstated when multicollinearity is present. Moderate to severe multicollinearity exists if the absolute value of the correlation coefficient between regressors is greater than 0.5. A correlation coefficient of 1 indicates perfect multicollinearity.

Table 37: Variable Pairs that have Moderate to Severe Multicollinearity

Variable Pair:	Correlation Coefficient:
Played Football & Played Baseball	0.51
Played Football & Played Basketball	0.94
Played Football & Played Bowling	0.66
Played Football & Played Hockey	0.81
Cover Charge & Hooters	-0.73
Cover Charge & Buffalo Wild Wings	-0.52
Hooters and Buffalo Wild Wings	0.77
Football Field & Cover Charge	0.72
Football Field & Stand Alone	0.79
Football Field & Strip Mall	-0.53
Football Field and Hooters	-0.71
Football Field & Buffalo Wild Wings	-0.61
Hooters & Stand Alone	-0.78
Hooters & Strip Mall	0.78
Buffalo and Stand Alone	-0.55
Buffalo and Strip Mall	0.87
Played Hockey & Played Basketball	0.85

The variable, “Played Football” is correlated with the variables: “Played Baseball”, “Played Basketball”, “Played Bowling”, and “Played Hockey”. Intuitively, the data suggests that people who play at least one sport, particularly football, are likely to play many other sports as well. “Played Football” and “Played Basketball” had the highest correlation coefficient out of any other pair of regressors with a correlation coefficient of 0.94, indicating severe multicollinearity. “Played Hockey” and “Played Basketball” had a particularly high correlation coefficient as well, with a coefficient of 0.85. Models J, T, and U may be vulnerable to multicollinearity issues because of the combination “Played Football” and “Played Basketball”, and the combination “Played Hockey” and “Played Basketball”.

“Hooters” and “Buffalo Wild Wings” are highly correlated with each other with a correlation coefficient of 0.77. They are also both highly correlated with “Stand Alone” and “Strip Mall”. “Buffalo Wild Wings” had a correlation coefficient of -0.55 and 0.87 with “Stand Alone” and “Strip Mall” respectively. “Hooters” had a correlation coefficient of -.78 and .78 with “Stand Alone” and “Strip Mall” respectively. Busters’ competitors tend to locate near each other because industries tend to cluster in cities. The data also suggests that Hooters and Buffalo Wild wings tend to locate in strip malls, but tend to avoid stand alone buildings. Only model H is exempt from this vulnerability, containing nether competitor regressors nor building type regressors.

Best Model

Overall, I believe that model U is the best fitting model:

$$\begin{aligned}\widehat{Sales} = {} & 2{,}623{,}103 + 2{,}130.65407 * \text{played\_baseball} - 3{,}042.96820 * \text{played\_basketball} \\ & + 1{,}707.47677 * \text{played\_hockey} + 36{,}317 * \text{football} + 257{,}744 * \text{cover\_charge} \\ & - 39{,}515 * \text{high\_tax} - 267{,}143 * \text{Hooters} - 210{,}071 * \text{Buffalo}\end{aligned}$$

The model as it is contains large numbers, so I will measure sales in thousands of dollars, which will adjust the parameters by five decimal places to consolidate the numbers:

$$\begin{aligned}\widehat{Sales_t} = {} & 262.31 + 2.13 * \text{played\_baseball} - 3.04 * \text{played\_basketball} + 1.71 * \text{played\_hockey} \\ & + 3.63 * \text{football} + 257.74 * \text{cover\_charge} - 39.52 * \text{high\_tax} - 267.14 * \text{Hooters} \\ & - 210.07 * \text{Buffalo}\end{aligned}$$

Interpretations of coefficients:

Played_baseball: Estimated coefficient = 2.13

For each additional person who lives within one radial mile of a Buster's location, who said that they had played organized baseball within the last 12 months, sales are expected on average to increase by 2.13 thousand dollars.

Played_basketball: Estimated coefficient = -3.04

For each additional person who lives within one radial mile of a Buster's location, who said that they had played organized basketball within the last 12 months, sales are expected on average to decrease by 3.04 thousand dollars, holding all else constant.

Played_hockey: Estimated coefficient = 1.71

For each additional person who lives within one radial mile of a Buster's location, who said that they had played organized hockey within the last 12 months, sales are expected on average to increase by 1.71 thousand dollars, holding all else constant.

Football: Estimated coefficient = 3.63

If there is a National League Football stadium within one radial mile of a Buster's location, sales are expected on average to increase by 3.63 thousand dollars, holding all else constant.

Cover_charge: Estimated coefficient = 257.74

If a Buster’s location requires a cover charge to enter the bar on Fridays and Saturdays, then sales are expected on average to increase by 257.74 thousand dollars, holding all else constant.

High Tax: Estimated coefficient = -39.52

If a Buster’s location is located in a city with a high bar tax, then sales are expected on average to decrease by 39.52 thousand dollars, holding all else constant.

Hooters: Estimated coefficient = -267.14

Each additional Hooters location within 10 radial miles of a Buster’s location is expected on average to decrease sales by 267.14 thousand dollars, holding all else constant.

Buffalo: Estimated coefficient = -210.17

If there is a Buffalo Wild Wings location within 10 radial miles of a Buster’s location, then sales are expected on average to decrease by 210.17 thousand dollars, holding all else constant.

Model Scoring and Sequential Regression

After the results from my initial regression, I will be using sequential regression to try to estimate a better fitting model than I found from intuitively building a regression model. I will be using the forward-selection method, the stepwise selection method, the backward selection method, the maximum R-square improvement selection method, and the adjusted R-square selection method to find the best fitting models. The best models have the highest statistical significance while also having a high R-squared and adjusted R-squared. Then I will use model scoring to evaluate potential locations and predict the best location for Buster's Brewhaus. Potential locations will be organized into four different types: High sales potential, medium sales potential, above-average sales potential, and low sales potential.

The multi-trait dummy variables must be removed when doing sequential regression because all multi-trait dummies must be included the regression model, however the computer may try to add them one at a time. Therefore, multi-trait dummy variables must be added after sequential regression. There are two sets of multi-trait dummies to consider adding to this regression: the type of building regressors, and geographic location regressors. Adding these variables to the regression produces insignificant p-values so I will exclude the multi-trait dummy variables from these regressions and decide whether to add them after I have found the best models.

Figure 2: Potential Regressors List

1. Engineering Occupation
2. Repair Occupation
3. Played Baseball
4. Played Basketball
5. Played Bowling
6. Played Football
7. Played Hockey
8. Restaurant Score
9. Nightlife Score
10. Football Field
11. Baseball field
12. Cover Charge
13. Bar Tax
14. Sports Championship
15. Hooters
16. Buffalo Wild Wings
17. Metrics

Forward-Selection Sequential Regression

Forward-Selection sequential regression is a method of building a linear model where models are regressed with only one regressor, and the model with the smallest p-value is selected. The regressors with the smallest p-values are continuously added until the addition of another regressor produces a p-value above a threshold. In this case, the threshold is 0.2.

Table 38: Parameter Estimates for best Forward-Selection Model

Variable	Parameter Estimate	Standard Error	t Value	P-value
Intercept	2,674,041	49,635	53.87	<.0001
Baseball	-76,549	53,965	-1.42	0.1606
cover_charge	328,887	61,200	5.37	<.0001
champion	94,444	42,475	2.22	0.0295
Hooters	-259,224	44,732	-5.8	<.0001
Buffalo	-232,753	54,355	-4.28	<.0001

$$R^2 = 0.8927$$

$$\bar{R}^2 = 0.8848$$

I prefer model U to this model. Although it has a higher R-squared than model U, it has a statistically significant variable with an incorrect sign. Forward-Selection sequential regression produced the exact same results as stepwise sequential regression. Stepwise selection is a similar method of sequential regression only in order for a variable to remain in the model after each successive step, that variable must be statistically significant at the threshold.

Backward Elimination Sequential Regression

Backward elimination sequential regression is the opposite of forward-selection sequential regression. Instead, a model with all regressors is estimated and a joint F-test is conducted on each regressor and the one with the largest p-value is removed. This continues until all regressors have p-values that are beneath the threshold of 0.2.

After adding the multi-trait dummy variables, none produced significant p-values so I will leave them out of this model. Model U is still preferred because this model includes a statistically significant variable with a counter-intuitive sign.

Table 39: Parameter Estimates for the best Backward Elimination Model

Variable	Parameter Estimate	Standard Error	t Value	P-value
Intercept	2,780,399	141,121	19.70	<.0001
played_bowling	-3,447.25	1,917.85	-1.80	0.08
played_hockey	2,543.99	1,509.33	1.69	0.10
baseball	-84,797	53,584	-1.58	0.12
cover_charge	327,288	61,094	5.36	<.0001
champion	94,114	42,720	2.20	0.03
Hooters	-270,732	45,025	-6.01	<.0001
Buffalo	-221,748	54,088	-4.10	0.0002

$R^2 = 0.8981$

$\bar{R}^2 = 0.8873$

R-square Improvement Sequential Regression

R-square improvement sequential regression builds models based on the R-square value. All one-variable models will be estimated and the one with the highest R-squared is selected. Then, all two variable models are estimated and the model that produces the largest positive change in R-squared is selected. This continues until all possible model sizes are found. This method finds the best model for each number of variables.

The best five variable model is an exact match to the ones found in forward-selection and stepwise selection. The best seven variable model is an exact match to the one found in backwards elimination selection. I believe that the best model is from step 6 in the R-square sequential regression.

The best models with more regressors tend to cause the other regressors to become statistically insignificant. The seven variable model includes played_bowling, which had a negative sign.

The adjusted R-squared sequential regression method produced models containing parameter estimates with counterintuitive signs, so I will turn my attention to the models found from the other methods.

Table 40: Parameter Estimates for the Best R-square Improvement Model

Variable	Parameter	Standard	t Value	P-value
	Estimate	Error		
Intercept	2,516,917	140,473	17.92	<.0001
played_baseball	1,747.66	1,462	1.20	0.24
baseball	-78,364	53,817	-1.46	0.15
cover_charge	315,427	62,039	5.08	<.0001
champion	88,969	42,589	2.09	0.04
Hooters	-270,736	45,620	-5.93	<.0001
Buffalo	-217,394	55,688	-3.90	0.0002

$$R^2 = 0.8950$$

$$\bar{R}^2 = 0.8856$$

I will use model U to estimate sales for potential locations.

Model U:

$$\widehat{Sales}_t = 262.31 + 2.13 * \text{played\_baseball} - 3.04 * \text{played\_basketball} + 1.71 * \text{played\_hockey} + 3.63 * \text{football} + 257.74 * \text{cover\_charge} - 39.52 * \text{high\_tax} - 267.14 * \text{Hooters} - 210.07 * \text{Buffalo}$$

Threshold Values for Projected Sales

We can predict the performance of prospective locations by using the models we have found and compare them to the average sales across the current Buster's locations. Above average is between the average and one standard deviation. Medium Sales potential is between the average plus one standard deviation and the average plus two standard deviations. High sales potential is greater than the average plus 2 standard deviations.

Table 41: Threshold values for sales potential

Revenue Type:	Threshold Value
High Sales Potential	Projected Sales > $3,298,590.77
Medium Sales Potential	$2,865,994.66 < Projected Sales < $3,298,590.77
Above Average Sales Potential	$2,433,398.55 < Projected Sales < $2,865,994.66
Low Average Sales Potential	Projected Sales < $2,433,398.55

Table 42: Sales Projection Results

Revenue type:	Description:	Store number:
High Sales Potential	Projected sales ≥ (average sales plus two standard deviations)	None
Medium Sales Potential	Projected sales ≥ (average sales plus one standard deviation) but Projected sales < (average sales plus two standard deviations)	None
Above-Average Sales Potential	Projected sales ≥ average sales but Projected sales < (average sales plus one standard deviations)	87
Low Sales Potential	Projected sales < average sales	83, 84, 85, 86

Table 43: Explanation of Sales Levels

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

The only location worth pursuing for Buster's is store 87. The other locations all have below-average sales potential.

In addition to the model, I have some general advice for expansion strategies for Buster's. Buster's should locate in a city that has won a sports championship within the last 4 years, or a city that is projected to win a sports championship in the near future. Baseball was consistently estimated to have a negative parameter estimate, so Buster's should not locate near a baseball field as I believe that Buster's competes with the services provided at a baseball stadium. People who play one sport tend to play many sports, as indicated by the correlation coefficients in table 37, and most of them had positive parameter estimates when regressed. Buster's should aim to cater to sports enthusiasts rather than just to participants in one sport. It may also be a good idea to locate near a gym or recreation center for this reason. Many Buster's locations are near universities and I believe that a prime target audience for marketing would be college students participating in intramurals or other organized sports. Buster's could offer a waived cover charge to D1 college athletes to help drive demand and popularity with college students. Both engineering occupation and repair occupation were consistently positive and significant. It seems as though people who tinker with machines are interested in Buster's. Engineers, repairmen, and repairwomen would be another good target audience.