**Miriam Garcia**

**March 24, 2021**

**University of North Texas**

**ECON 5645**

**Model Building & Inference in**

**Regression for Timmy Tom’s**

**I. Introduction**

Timmy Tom’s Gourmet Sandwiches is a rapid service restaurant that serves sandwiches, subs, beverages, and other side dishes such as chips and desserts. Timmy Tom’s mission is to provide their customers appetizing sandwiches at a feasible price. Timmy Tom’s was on an expansion path and rapid growth in their key markets up until March of 2020. By that point of time, Timmy Tom’s had infiltrated the market with locations that were considered high-revenue and easy-win over their competitors. Around the same time, a global pandemic, COVID-19, hindered any plans for further growth or expansion. This global crisis forced Timmy Tom’s to switch the direction of their expansion strategy into a more conservative approach of extreme caution of any new stores being opened. After a year, Timmy Tom’s is confident that it can continue its expansion strategy as long as we are able to identify locations with really strong revenue potential. Therefore, we have been provided with sales data from all of their locations in the United States from 2019. This data shows previous sales in locations that were operating in 2019 and also gives us several variables that could potentially explain sales.

As a data analyst, we must construct econometric models to forecast revenue potential of retail sites for Timmy Tom’s. With the sales data from 2019, our long-term goal is to be able to estimate potential sales in order to determine what future locations are worth pursuing as part of our expansion plan. To do this, we will perform full and comprehensive pre-model analysis on our dependent variable, Sales, and all potential independent variables. We then must build the best possible model to explain sales at Timmy Tom’s during the 2019 year. We will then select the best model to score several potential locations in order to determine which ones will produce high revenue. Our current short-run objective is to provide full and comprehensive pre-model analysis on the dependent variable and the continuous potential independent variables. 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 "year\_open". “Year\_openi” is the year in which store “i” opened, which ranges from 1974 up until 2015. “Store\_IDi” refers to a unique number (from 1 to 306) that identifies each store, “i,” in the data set. We must ensure that all observations on our dependent variable, Yi, make sense. To begin our process, we have 306 observations and therefore 306 stores to pull data from.

Our first 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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Variable: Sales | | | | |
| N | Mean | Std Dev | Minimum | Maximum |
| 306 | 895,284.54 | 420,174.16 | 0 | 2,918,588.07 |

As our next step, we will plot a graph that shows us the original observations and their respective Sales.

Figure 1: Average Sales for all Timmy Tom’s 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. We can observe that the minimum for Sales is zero, which indicates that one or more observations had no data for Sales. From there, we identify what stores have these sales and remove the observations that are unreasonable using SAS. In this case store 1 and store 2 have zero sales. These two stores have 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 SUBS2 to keep track now that we have removal of store 1 and store 2. We can now create new summary statistics after the removal of these unreasonable observations.

Table 2: Summary Statistics (Minus Unreasonable Observations)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Analysis Variable: Sales** | | | | |
| **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| 304 | 901,174.6 | 415,194.2 | 198,153 | 2,918,588.1 |

After the removal of these two stores, we can now create an adjusted graph.

Figure 2: Average Sales for all Timmy Tom’s Locations (without unreasonable observations)

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: 901,174.6. And the standard deviation is: 415,194.2

Low outlier -$136,810.90

High outlier $1,939,160.10

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 Subs3. Since we have already eliminated any unreasonable observation with sales equal to zero and had no sales below our low outlier threshold, we do not need to worry about low outliers.

Table 3: Stores with Sales Higher than our High Outlier Threshold

|  |  |
| --- | --- |
| STORE\_ID | Sales |
| 301 | $1,978,412.03 |
| 302 | $1,986,173.10 |
| 303 | $ 2,027,240.52 |
| 304 | $2,318,110.56 |
| 305 | $2,861,590.24 |
| 306 | $2,918,588.07 |

.

These six stores have sales figures that our higher than our high outlier threshold. The removal of these stores gives us the following summary statistics and new graph.

Table 4: Summary Statistics (Adj. for Unreasonable Observations & Outliers)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis Variable: Sales** | | | | | |
| **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** | **Coeff of Variation** |
| 298 | 872,036.76 | 359,823.32 | 198,152.77 | 1,925,871.54 | 41.26 |

Figure 3: Avg Sales for all TT’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. 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 5: Summary of removed observations and the reasoning

|  |  |
| --- | --- |
| **# of observation/STORE\_ID** | **Reason for removal** |
| STORE\_ID = 1 | Unreasonable observation since sales were zero. |
| STORE\_ID = 2 | Unreasonable observation since sales were zero. |
| STORE\_ID = 301 | Has sales of $1,978,412.03 which is above our high outlier threshold. |
| STORE\_ID = 302 | Has sales of $1,986,173.10 which is above our high outlier threshold. |
| STORE\_ID = 303 | Has sales of $2,027,240.52 which is above our high outlier threshold. |
| STORE\_ID = 304 | Has sales of $2,318,110.56 which is above our high outlier threshold. |
| STORE\_ID = 305 | Has sales of $2,861,590.24 which is above our high outlier threshold. |
| STORE\_ID = 306 | Has sales of $2,918,588.07 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 6: Potential Independent Variables

|  |  |
| --- | --- |
| **VARIABLE NAME** | **DEFINITION** |
| Store\_ID | A number from 1 to 305, that identifies each store in the sample. |
| Sales | The dollar value of total sales at each store from January 1, 2019 through December 31, 2019. |
| Year\_open | The year in which the given store opened. |
| Traffic\_count | The average number of vehicles (per day) that travel on the road near a given store. |
| Food\_away\_3R | Average expenditure on food (away from home) by people who live within 3 radial miles of a given store. |
| Food\_away\_5T | Average expenditure on food (away from home) by people who live within a 5-minute drive of a given store. |
| Pop\_GE\_18\_3R | The number of people who are are 18 years old or older, who live within 3 radial miles of a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| Pop\_GE\_18\_5T | The number of people who are are 18 years old or older, who live within a 5-minute drive of a given store. |
| Pop\_18\_21\_3R | The number of people who are are 18 to 21 years old who live within 3 radial miles of a given store. |
| Pop\_18\_21\_5T | The number of people who are are 18 to 21 years old who live within a 5-minute drive of a given store. |
| Pop\_21\_39\_3R | The number of people who are are 21 to 39 years old who live within 3 radial miles of a given store. |
| Pop\_21\_39\_5T | The number of people who are are 21 to 39 years old who live within a 5-minute drive of a given store. |
| Pop\_40\_49\_3R | The number of people who are are 40 to 49 years old who live within 3 radial miles of a given store. |
| Pop\_40\_49\_5T | The number of people who are are 40 to 49 years old who live within a 5-minute drive of a given store. |
| Pop\_50\_69\_3R | The number of people who are are 50 to 69 years old who live within 3 radial miles of a given store. |
| Pop\_50\_69\_5T | The number of people who are are 50 to 69 years old who live within a 5-minute drive of a given store. |
| Pop\_70\_85\_3R | The number of people who are are 70 to 85 years old who live within 3 radial miles of a given store. |
| Pop\_70\_85\_5T | The number of people who are are 70 to 85 years old who live within a 5-minute drive of a given store. |
| Likely\_customers\_1R | The number of people who live within one radial mile of a given store who are likely to be customers. |
| Likely\_customers\_5T | The number of people who live within a 5-minute drive of a given store who are likely to be customers. |
| Competitor\_A\_index | An index of how valuable competitior A is to a given store (the more valuable, the larger the index). |
| Competitor\_B\_index | An index of how valuable competitior B is to a given store (the more valuable, the larger the index). |
| Competitor\_C\_index | An index of how valuable competitior C is to a given store (the more valuable, the larger the index). |
| Competitor\_D\_index | An index of how valuable competitior D is to a given store (the more valuable, the larger the index). |
| Bakeries\_index\_1R | An index of how valuable bakery-type restaurants (such as Panera, etc.) that are located within 1 radial mile are to a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| Casual\_dining\_index\_1R | An index of how valuable casual-dining-type restaurants (such as Applebee's, Chili's, BJ's, etc.) that are located within 1 radial mile are to a given store. |
| Fast\_food\_index\_1R | An index of how valuable fast-food-type restaurants (such as McDonald's, Burger King, etc.) that are located within 1 radial mile are to a given store. |
| Low\_grocery\_index\_1R | An index of how valuable low-end grocery stores (such as Aldi, Sack N Save, etc.) that are located within 1 radial mile are to a given store. |
| Mid\_grocery\_index\_1R | An index of how valuable mid-level grocery stores (such as Kroger, Safeway, etc.) that are located within 1 radial mile are to a given store. |
| Big\_box\_index\_1R | An index of how valuable big box stores (such as Best Buy, Target, etc.) that are located within 1 radial mile are to a given store. |
| Sandwich\_shop\_index\_1R | An index of how valuable sandwich shops (other than Timmy Tom's, such as Subway, Jersey Mike's, etc.) that are located within 1 radial mile are to a given store. |
| Fast\_food\_8T | The number of fast-food-type restaurants (such as McDonald's, Burger King, etc.) that are located within an 8-minute drive of a given store. |
| Big\_box\_1R | The number of big-box stores (such as Best Buy, Target, etc.) that are located within one radial mile of a given store. |
| Pop\_Associates\_3R | The number of people living within 3 radial miles of a given store whose highest educational attainment is an Associates degree. |
| Pop\_Associates\_5T | The number of people living within a 5-minute drive of a given store whose highest educational attainment is an Associates degree. |
| Pop\_Bachelors\_3R | The number of people living within 3 radial miles of a given store whose highest educational attainment is a Bachelors degree. |
| Pop\_Bachelors\_5T | The number of people living within a 5-minute drive of a given store whose highest educational attainment is a Bachelors degree. |
| Pop\_Doctorate\_3R | The number of people living within 3 radial miles of a given store whose highest educational attainment is a doctorate degree. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| Pop\_Doctorate\_5T | The number of people living within a 5-minute drive of a given store whose highest educational attainment is a doctorate degree. |
| Pop\_grades\_9\_12\_3R | The number of people who live within 3 radial miles of a given store who are in grades 9 through 12. |
| Pop\_grades\_9\_12\_5T | The number of people who live within a 5-minute drive of a given store who are in grades 9 through 12. |
| Pop\_grad\_school\_3R | The number of people who live within 3 radial miles of a given store who are in graduate school. |
| Pop\_grad\_school\_5T | The number of people who live within a 5-minute drive of a given store who are in graduate school. |
| Pop\_in\_school\_3R | The number of people who live within 3 radial miles of a given store who are in school (any school). |
| Pop\_in\_school\_5T | The number of people who live within a 5-minute drive of a given store who are in school (any school). |
| Pop\_undergrads\_3R | The number of people who live within 3 radial miles of a given store who are undergraduates. |
| Pop\_undergrads\_5T | The number of people who live within a 5-minute drive of a given store who are undergraduates. |
| Pop\_Masters\_3R | The number of people living within 3 radial miles of a given store whose highest educational attainment is a Masters degree. |
| Pop\_Masters\_5T | The number of people living within a 5-minute drive of a given store whose highest educational attainment is a Masters degree. |
| Pop\_some\_college\_3R | The number of people who live within 3 radial miles of a given store who have some college education, but no degree. |
| Pop\_some\_college\_5T | The number of people who live within a 5-minute drive of a given store who have some college education, but no degree. |
| Tot\_HH\_Expend\_3R | Total annual expenditure (in dollars) of households located within 3 radial miles of a given store. |
| Tot\_HH\_Expend\_5T | Total annual expenditure (in dollars) of households located within a 5-minute drive of a given store. |
| Cust\_value | A measure of the value of all residents in the Timmy Tom's network, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value). |
| Cust\_value\_per\_cap | A measure of the value, per capita, of all residents in the Timmy Tom's , with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value). |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| Cust\_value\_region | A measure of the value of residents within the neighboring geographic region of a given store, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value). |
| Cust\_value\_per\_cap\_region | A measure of the value, per capita, of residents within the neighboring geographic region of a given store, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value). |
| HHinc\_LT\_25K\_3R | The number of households within 3 radial miles of a given store, with annual income less than $25,000. |
| HHinc\_LT\_25K\_5T | The number of households within a 5-minute drive of a given store, with annual income less than $25,000. |
| HHinc\_25\_49K\_3R | The number of households within 3 radial miles of a given store, with annual income between $25,000 and $49,000. |
| HHinc\_25\_49K\_5T | The number of households within a 5-minute drive of a given store, with annual income between $25,000 and $49,000. |
| HHinc\_50\_74K\_3R | The number of households within 3 radial miles of a given store, with annual income between $50,000 and $74,999. |
| HHinc\_50\_74K\_5T | The number of households within a 5-minute drive of a given store, with annual income between $50,000 and $74,999. |
| HHinc\_75\_99K\_3R | The number of households within 3 radial miles of a given store, with annual income between $75,000 and $99,999. |
| HHinc\_75\_99K\_5T | The number of households within a 5-minute drive of a given store, with annual income between $75,000 and $99,999. |
| HHinc\_GE\_100K\_3R | The number of households within 3 radial miles of a given store, with annual income greater than or equal to $100,000. |
| HHinc\_GE\_100K\_5T | The number of households within a 5-minute drive of a given store, with annual income greater than or equal to $100,000. |
| Avg\_HHinc\_3R | Average annual household income (in dollars) of households within 3 radial miles of a given store. |
| Avg\_HHinc\_5T | Average annual household income (in dollars) of households within a 5-minute drive of a given store. |
| Med\_HHinc\_3R | Median annual household income (in dollars) of households within 3 radial miles of a given store. |
| Med\_HHinc\_5T | Median annual household income (in dollars) of households within a 5-minute drive of a given store. |
| HH\_1person\_3R | The number of 1-person households located within 3 radial miles of a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| HH\_1person\_5T | The number of 1-person households located within a 5-minute drive of a given store. |
| HH\_2person\_3R | The number of 2-person households located within 3 radial miles of a given store. |
| HH\_2person\_5T | The number of 2-person households located within a 5-minute drive of a given store. |
| HH\_3person\_3R | The number of 3-person households located within 3 radial miles of a given store. |
| HH\_3person\_5T | The number of 3-person households located within a 5-minute drive of a given store. |
| HH\_4person\_3R | The number of 4-person households located within 3 radial miles of a given store. |
| HH\_4person\_5T | The number of 4-person households located within a 5-minute drive of a given store. |
| HH\_5person\_3R | The number of 5-person households located within 3 radial miles of a given store. |
| HH\_5person\_5T | The number of 5-person households located within a 5-minute drive of a given store. |
| HH\_6person\_3R | The number of 6-person households located within 3 radial miles of a given store. |
| HH\_6person\_5T | The number of 6-person households located within a 5-minute drive of a given store. |
| Brady\_Bunch\_3R | The number of households with 7 or more people, located within 3 radial miles of a given store. |
| Brady\_Bunch\_5T | The number of households with 7 or more people, located within a 5-minute drive of a given store. |
| med\_home\_value\_3R | The median value (in dollars) of homes located within 3 radial miles of a given store. |
| med\_home\_value\_5T | The median value (in dollars) of homes located within a 5-minute drive of a given store. |
| med\_home\_value\_adj\_3R | The median value (in dollars, and adjusted for the cost of living) of homes located within 3 radial miles of a given store. |
| med\_home\_value\_adj\_5T | The median value (in dollars, and adjusted for the cost of living) of homes located within a 5-minute drive of a given store. |
| per\_cap\_inc\_3R | Per capita income (in dollars) of people living with 3 radial miles of a given store. |
| per\_cap\_inc\_5T | Per capita income (in dollars) of people living with a 5-minute drive of a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| labor\_blue\_3R | The number of people who live within 3 radial miles of a given store, who work in blue collar occupations. |
| labor\_blue\_5T | The number of people who live within a 5-minute drive of a given store, who work in blue collar occupations. |
| labor\_farm\_3R | The number of people who live within 3 radial miles of a given store, who work in service or farm occupations. |
| labor\_farm\_5T | The number of people who live within a 5-minute drive of a given store, who work in service or farm occupations. |
| labor\_white\_col\_3R | The number of people who live within 3 radial miles of a given store, who work in white collar occupations. |
| labor\_white\_col\_5T | The number of people who live within a 5-minute drive of a given store, who work in white collar occupations. |
| avg\_LOR\_3R | The average number of years that residents lived in their home (length of residence) for people who live within 3 radial miles of a given sortore. |
| Pop\_married\_3R | The number of married people who live within 3 radial miles of a given store. |
| Pop\_married\_5T | The number of married people who live within a 5-minute drive of a given store. |
| Distance\_hwy | The distance, in miles, to the nearest highway. |
| Distance\_hwy\_interstate | The distance, in miles, to the nearest highway or interstate. |
| Distance\_interstate | The distance, in miles, to the nearest interstate. |
| restaurants\_3R | The number of restaurants (of all types) located within 3 radial miles of a given store. |
| retail\_3R | The number of retail establishments (of all types) located within 3 radial miles of a given store. |
| restaurants\_retail\_3R | The number of restaurants and retail establishments (of all types) located within 3 radial miles of a given store. |
| Asian\_HH\_3R | The number of Asian households located within 3 radial miles of a given store. |
| Asian\_HH\_5T | The number of Asian households located within a 5-minute drive of a given store. |
| Asian\_pop\_3R | The Asian population (in people) living within 3 radial miles of a given store. |
| Asian\_pop\_5T | The Asian population (in people) living within a 5-minute drive of a given store. |
| Black\_HH\_3R | The number of black households located within 3 radial miles of a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| Black\_HH\_5T | The number of black households located within a 5-minute drive of a given store. |
| Black\_pop\_3R | The black population (in people) living within 3 radial miles of a given store. |
| Black\_pop\_5T | The black population (in people) living within a 5-minute drive of a given store. |
| Hispanic\_HH\_3R | The number of Hispanic households located within 3 radial miles of a given store. |
| Hispanic\_HH\_5T | The number of Hispanic households located within a 5-minute drive of a given store. |
| Hispanic\_pop\_3R | The Hispanic population (in people) living within 3 radial miles of a given store. |
| Hispanic\_pop\_5T | The Hispanic population (in people) living within a 5-minute drive of a given store. |

**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):

Graphical user interface, text, application, Word

Description automatically generated

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 7: Summary Statistics for Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** | **Coeff of Variation** |
| Store\_ID | 298 | 152.35 | 88.64 | 1 | 304 | 58.18 |
| Year\_open | 298 | 1998.21 | 10.89 | 1974 | 2015 | 0.55 |
| Traffic\_count | 225 | 22676.73 | 19729.65 | 4645.41 | 178296.58 | 87 |

.

Upon analysis, we observe two potential variables with issues. One that exhibits a low coefficient of variation below our threshold and another variable that has a high percentage of missing observations. The variable Year\_open has a coefficient of variation of 0.55, which is below our required threshold of a value of 2. We want all of our variables to have sufficient variation in order to be relevant to our analysis. In this case, it seems to be that the year that stores in the data were opened do not really vary. Upon observation, it seems to be that many stores were opened in “bulk” every so years. For example, in 2013 thirteen stores were opened. Therefore, if there is little variation in the year the stores opened, the variable Year\_open is irrelevant to our analysis. Traffic\_count has 225 observations, which means that are 73 missing observations now that our adjusted dataset has a total of 298 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 we cannot figure out the missing values in Traffic\_count. Since there is 24.5% missing data for Traffic\_count, it is in our best interest to completely drop the variable Traffic\_count from our analysis. Therefore, both Year\_open and Traffic\_count will be dropped from further analysis.

**V. Micronumerosity**

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 298 observations against 113 variables.

298-113 = 185 degrees of freedom

**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 an 88% confidence level, which gives us a p-value of 0.12. If the p-value in our hypothesis test is a value less than 0.12, 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 0.12, 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 8: Impact of Potential Variables – Food Away

|  |  |  |
| --- | --- | --- |
| **Variable** | Food\_away\_3R | Food\_away\_5T |
| **Correlation Coefficient to Sales** | 0.12738 | 0.04871 |
| **P-value** | 0.0279 | 0.4021 |

Table 9: Impact of Potential Variables – Likely Customers

|  |  |  |
| --- | --- | --- |
| **Variable** | Likely\_customers\_1R | Likely\_customers\_5T |
| **Correlation Coefficient to Sales** | -0.14658 | -0.04610 |
| **P-value** | 0.0113 | 0.4279 |

.

Table 10: Impact of Potential Variables – Population

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Pop\_GE\_18\_3R | Pop\_GE\_18\_5T | Pop\_18\_21\_3R |
| **Correlation Coefficient to Sales** | -0.17827 | -0.10487 | -0.08028 |
| **P-value** | 0.0020 | 0.0707 | 0.1669 |

Table 11: Impact of Potential Variables – Population (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Pop\_18\_21\_5T | Pop\_21\_39\_3R | Pop\_21\_39\_5T |
| **Correlation Coefficient to Sales** | -0.03465 | -0.15457 | -0.09245 |
| **P-value** | 0.5513 | 0.0075 | 0.1112 |

Table 12: Impact of Potential Variables – Population (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Pop\_40\_49\_3R | Pop\_40\_49\_5T | Pop\_50\_69\_3R |
| **Correlation Coefficient to Sales** | -0.16519 | -0.10013 | -0.19131 |
| **P-value** | 0.0042 | 0.0844 | 0.0009 |

Table 13: Impact of Potential Variables – Population (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Pop\_50\_69\_5T | Pop\_70\_85\_3R | Pop\_70\_85\_5T |
| **Correlation Coefficient to Sales** | -0.10977 | -0.18971 | -0.09091 |
| **P-value** | 0.0584 | 0.0010 | 0.1173 |

Table 14: Impact of Potential Variables – Competitor Index

|  |  |  |
| --- | --- | --- |
| **Variable** | Competitor\_A\_index | Competitor\_B\_index |
| **Correlation Coefficient to Sales** | -0.02003 | -0.01345 |
| **P-value** | 0.7305 | 0.8171 |

Table 15: Impact of Potential Variables – Competitor Index (cont.)

|  |  |  |
| --- | --- | --- |
| **Variable** | Competitor\_C\_index | Competitor\_D\_index |
| **Correlation Coefficient to Sales** | 0.02873 | 0.08109 |
| **P-value** | 0.6214 | 0.1626 |

Table 16: Impact of Potential Variables – Dining Index

|  |  |  |
| --- | --- | --- |
| **Variable** | Casual\_dining\_index\_1R | Fast\_food\_index\_1R |
| **Correlation Coefficient to Sales** | 0.02462 | 0.04603 |
| **P-value** | 0.6721 | 0.4285 |

Table 17: Impact of Potential Variables – Grocery Index

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Bakeries\_index\_1R | Low\_grocery\_index\_1R | Mid\_grocery\_index\_1R |
| **Correlation Coefficient to Sales** | 0.01846 | 0.01159 | -0.08884 |
| **P-value** | 0.7509 | 0.8421 | 0.1260 |

Table 18: Impact of Potential Variables – Other Indexes

|  |  |  |
| --- | --- | --- |
| **Variable** | Big\_box\_index\_1R | Sandwich\_shop\_index\_1R |
| **Correlation Coefficient to Sales** | 0.11078 | -0.00006 |
| **P-value** | 0.0561 | 0.9991 |

Table 19: Impact of Potential Variables – Population with Associates Degree

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_Associates\_3R | Pop\_Associates\_5T |
| **Correlation Coefficient to Sales** | -0.15035 | -0.08869 |
| **P-value** | 0.0093 | 0.1266 |

Table 20: Impact of Potential Variables – Population with Bachelor’s Degree

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_Bachelors\_3R | Pop\_Bachelors\_5T |
| **Correlation Coefficient to Sales** | -0.09294 | -0.01885 |
| **P-value** | 0.1094 | 0.7459 |

Table 21: Impact of Potential Variables – Population with Doctorate Degree

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_Doctorate\_3R | Pop\_Doctorate\_5T |
| **Correlation Coefficient to Sales** | -0.09444 | -0.05163 |
| **P-value** | 0.1037 | 0.3745 |

Table 22: Impact of Potential Variables – Population in High School

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_grades\_9\_12\_3R | Pop\_grades\_9\_12\_5T |
| **Correlation Coefficient to Sales** | -0.17034 | -0.10236 |
| **P-value** | 0.0032 | 0.0777 |

Table 23: Impact of Potential Variables – Population in Graduate School

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_grad\_school\_3R | Pop\_grad\_school\_5T |
| **Correlation Coefficient to Sales** | -0.11565 | -0.03064 |
| **P-value** | 0.0461 | 0.5983 |

Table 24: Impact of Potential Variables – Population in School

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_in\_school\_3R | Pop\_in\_school\_5T |
| **Correlation Coefficient to Sales** | -0.17088 | -0.08855 |
| **P-value** | 0.0031 | 0.1272 |

Table 25: Impact of Potential Variables – Population in College (Undergraduates)

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_undergrads\_3R | Pop\_undergrads\_5T |
| **Correlation Coefficient to Sales** | -0.08153 | -0.03814 |
| **P-value** | 0.1604 | 0.5119 |

Table 26: Impact of Potential Variables – Population with Master’s Degree

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_Masters\_3R | Pop\_Masters\_5T |
| **Correlation Coefficient to Sales** | -0.09268 | -0.02842 |
| **P-value** | 0.1103 | 0.6251 |

Table 27: Impact of Potential Variables – Population with Some College Level Education

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_some\_college\_3R | Pop\_some\_college\_5T |
| **Correlation Coefficient to Sales** | -0.17537 | -0.09548 |
| **P-value** | 0.0024 | 0.0999 |

Table 28: Impact of Potential Variables – Total Annual Expenditure of Households

|  |  |  |
| --- | --- | --- |
| **Variable** | Tot\_HH\_Expend\_3R | Tot\_HH\_Expend\_5T |
| **Correlation Coefficient to Sales** | -0.15132 | -0.08212 |
| **P-value** | 0.0089 | 0.1573 |

Table 29: Impact of Potential Variables – Customer Value in Network

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Cust\_value | Cust\_value\_per\_cap | Cust\_value\_region |
| **Correlation Coefficient to Sales** | -0.02328 | 0.06624 | 0.01033 |
| **P-value** | 0.6890 | 0.2543 | 0.8591 |

Table 30: Impact of Potential Variables – Customer Value in Network & Number of Households with Specific Income Level

|  |  |  |
| --- | --- | --- |
| **Variable** | Cust\_value\_per\_cap\_region | HHinc\_LT\_25K\_3R |
| **Correlation Coefficient to Sales** | 0.14837 | -0.18612 |
| **P-value** | 0.0103 | 0.0012 |

Table 31: Impact of Potential Variables – Number of Households with Specific Income Level

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | HHinc\_LT\_25K\_5T | HHinc\_25\_49K\_3R | HHinc\_25\_49K\_5T |
| **Correlation Coefficient to Sales** | -0.14236 | -0.19072 | -0.12350 |
| **P-value** | 0.0139 | 0.0009 | 0.0331 |

Table 32: Impact of Potential Variables – # of Households with Specific Income Level (cont.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | HHinc\_50\_74K\_3R | HHinc\_50\_74K\_5T | HHinc\_75\_99K\_3R | HHinc\_75\_99K\_5T |
| **Correlation Coefficient to Sales** | -0.16376 | -0.08348 | -0.11310 | -0.04123 |
| **P-value** | 0.0046 | 0.1506 | 0.0511 | 0.4783 |

Table 33: Impact of Potential Variables – Number of Households with Specific Income Level

& Distance/Driving Time to Location

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | HHinc\_GE\_100K\_3R | HHinc\_GE\_100K\_5T | Avg\_HHinc\_3R | Avg\_HHinc\_5T |
| **Correlation Coefficient to Sales** | -0.08341 | -0.01553 | 0.08205 | 0.03184 |
| **P-value** | 0.1509 | 0.7895 | 0.1577 | 0.5841 |

Table 34: Impact of Potential Variables – Median Value of Home & Median Value of Home Adjusted for Cost of Living

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Med\_HHinc\_3R | Med\_HHinc\_5T | med\_home\_value\_adj\_3R | med\_home\_value\_adj\_5T |
| **Correlation Coefficient to Sales** | 0.11658 | 0.04299 | 0.05843 | 0.03780 |
| **P-value** | 0.0443 | 0.4597 | 0.3147 | 0.5157 |

Table 35: Impact of Potential Variables – Per Capita Income

|  |  |  |
| --- | --- | --- |
| **Variable** | per\_cap\_inc\_3R | per\_cap\_inc\_5T |
| **Correlation Coefficient to Sales** | 0.10664 | 0.02717 |
| **P-value** | 0.0660 | 0.6404 |

Table 36: Impact of Potential Variables – Households Headcount

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | HH\_1person\_3R | HH\_1person\_5T | HH\_2person\_3R | HH\_2person\_5T |
| **Correlation Coefficient to Sales** | -0.14178 | -0.09265 | -0.14640 | -0.07015 |
| **P-value** | 0.0143 | 0.1105 | 0.0114 | 0.2273 |

Table 37: Impact of Potential Variables – Households Headcount (cont.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | HH\_3person\_3R | HH\_3person\_5T | HH\_4person\_3R | HH\_4person\_5T |
| **Correlation Coefficient to Sales** | -0.17359 | -0.09774 | -0.16196 | -0.09600 |
| **P-value** | 0.0026 | 0.0921 | 0.0051 | 0.0981 |

Table 38: Impact of Potential Variables – Households Headcount (cont.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | HH\_5person\_3R | HH\_5person\_5T | HH\_6person\_3R | HH\_6person\_5T |
| **Correlation Coefficient to Sales** | -0.20113 | -0.14302 | -0.21447 | -0.16166 |
| **P-value** | 0.0005 | 0.0135 | 0.0002 | 0.0052 |

Table 39: Impact of Potential Variables – Households Headcount (cont.) & Median Home Value

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Brady\_Bunch\_3R | Brady\_Bunch\_5T | med\_home\_value\_3R |
| **Correlation Coefficient to Sales** | -0.21149 | -0.16844 | -0.02172 |
| **P-value** | 0.0002 | 0.0035 | 0.7088 |

Table 40: Impact of Potential Variables – Median Home Value & Adjusted Median Home Value

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | med\_home\_value\_5T | med\_home\_value\_adj\_3R | med\_home\_value\_adj\_5T |
| **Correlation Coefficient to Sales** | -0.03971 | 0.05843 | 0.03780 |
| **P-value** | 0.4946 | 0.3147 | 0.5157 |

Table 41: Impact of Potential Variables – Blue Collar & Service/Farm Occupation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | labor\_blue\_3R | labor\_blue\_5T | labor\_farm\_3R | labor\_farm\_5T |
| **Correlation Coefficient to Sales** | -0.20053 | -0.13468 | -0.16783 | -0.10529 |
| **P-value** | 0.0005 | 0.0200 | 0.0037 | 0.0695 |

Table 42: Impact of Potential Variables – White Collar Occupation & Length of Residence

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | labor\_white\_col\_3R | labor\_white\_col\_5T | avg\_LOR\_3R |
| **Correlation Coefficient to Sales** | -0.10673 | -0.03801 | -0.08367 |
| **P-value** | 0.0658 | 0.5133 | 0.1496 |

Table 43: Impact of Potential Variables – Married Population & Distance from Highway

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Pop\_married\_3R | Pop\_married\_5T | Distance\_hwy |
| **Correlation Coefficient to Sales** | -0.17544 | -0.08726 | -0.05912 |
| **P-value** | 0.0024 | 0.1329 | 0.3091 |

Table 44: Impact of Potential Variables – Distance

|  |  |  |
| --- | --- | --- |
| **Variable** | Distance\_hwy\_interstate | Distance\_interstate |
| **Correlation Coefficient to Sales** | -0.05912 | 0.00610 |
| **P-value** | 0.3091 | 0.9164 |

Table 45: Impact of Potential Variables – Number of restaurants/ retail establishments

|  |
| --- |
| within 3 radial miles of a TT store |

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | restaurants\_3R | retail\_3R | restaurants\_retail\_3R |
| **Correlation Coefficient to Sales** | -0.11826 | -0.13693 | -0.13242 |
| **P-value** | 0.0413 | 0.0180 | 0.0222 |

Table 46: Impact of Potential Variables – Demographics - Asian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Asian\_HH\_3R | Asian\_HH\_5T | Asian\_pop\_3R | Asian\_pop\_5T |
| **Correlation Coefficient to Sales** | -0.08766 | -0.02739 | -0.08199 | -0.02772 |
| **P-value** | 0.1311 | 0.6378 | 0.1580 | 0.6336 |

Table 47: Impact of Potential Variables – Demographics - Black

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Black\_HH\_3R | Black\_HH\_5T | Black\_pop\_3R | Black\_pop\_5T |
| **Correlation Coefficient to Sales** | -0.20168 | -0.16821 | -0.19228 | -0.16982 |
| **P-value** | 0.0005 | 0.0036 | 0.0008 | 0.0033 |

Table 48: Impact of Potential Variables – Demographics - Hispanic

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Hispanic\_HH\_3R | Hispanic\_HH\_5T | Hispanic\_pop\_3R | Hispanic\_pop\_5T |
| **Correlation Coefficient to Sales** | -0.16298 | -0.12089 | -0.16999 | -0.12942 |
| **P-value** | 0.0048 | 0.0370 | 0.0032 | 0.0255 |

Now we can create a list of potential regressors that are significantly correlated to sales. With an 88% level of confidence, we will set our p-value to 0.12. Which indicates that any variable with a p-value less than 0.12 will be statistically significant to our analysis and will affect Sales.

Table 49: Regressors Correlated with Sales & Statistical Significance\*

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Correlation Coefficient** | **P-value** |
| HH\_6person\_3R | -0.21447 | 0.0002 |
| Brady\_Bunch\_3R | -0.21149 | 0.0002 |
| HH\_5person\_3R | -0.20113 | 0.0005 |
| labor\_blue\_3R | -0.20053 | 0.0005 |
| Black\_HH\_3R | -0.20168 | 0.0005 |
| Black\_pop\_3R | -0.19228 | 0.0008 |
| Pop\_50\_69\_3R | -0.19131 | 0.0009 |
| HHinc\_25\_49K\_3R | -0.19072 | 0.0009 |
| Pop\_70\_85\_3R | -0.18971 | 0.001 |
| HHinc\_LT\_25K\_3R | -0.18612 | 0.0012 |
| Pop\_GE\_18\_3R | -0.17827 | 0.002 |
| Pop\_some\_college\_3R | -0.17537 | 0.0024 |
| Pop\_married\_3R | -0.17544 | 0.0024 |
| HH\_3person\_3R | -0.17359 | 0.0026 |
| Pop\_in\_school\_3R | -0.17088 | 0.0031 |
| Pop\_grades\_9\_12\_3R | -0.17034 | 0.0032 |
| Hispanic\_pop\_3R | -0.16999 | 0.0032 |
| Black\_pop\_5T | -0.16982 | 0.0033 |
| Brady\_Bunch\_5T | -0.16844 | 0.0035 |
| Black\_HH\_5T | -0.16821 | 0.0036 |
| labor\_farm\_3R | -0.16783 | 0.0037 |
| Pop\_40\_49\_3R | -0.16519 | 0.0042 |
| HHinc\_50\_74K\_3R | -0.16376 | 0.0046 |
| Hispanic\_HH\_3R | -0.16298 | 0.0048 |
| HH\_4person\_3R | -0.16196 | 0.0051 |
| HH\_6person\_5T | -0.16166 | 0.0052 |
| Pop\_21\_39\_3R | -0.15457 | 0.0075 |
| Tot\_HH\_Expend\_3R | -0.15132 | 0.0089 |
| Pop\_Associates\_3R | -0.15035 | 0.0093 |
| Cust\_value\_per\_cap\_region | 0.14837 | 0.0103 |
| Likely\_customers\_1R | -0.14658 | 0.0113 |
| HH\_2person\_3R | -0.1464 | 0.0114 |
| HH\_5person\_5T | -0.14302 | 0.0135 |
| HHinc\_LT\_25K\_5T | -0.14236 | 0.0139 |
| HH\_1person\_3R | -0.14178 | 0.0143 |
| retail\_3R | -0.13693 | 0.018 |
| labor\_blue\_5T | -0.13468 | 0.02 |

Table 49: Regressors Correlated with Sales & Statistical Significance (cont.)

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Correlation Coefficient** | **P-value** |
| restaurants\_retail\_3R | -0.13242 | 0.0222 |
| Hispanic\_pop\_5T | -0.12942 | 0.0255 |
| Food\_away\_3R | 0.12738 | 0.0279 |
| HHinc\_25\_49K\_5T | -0.1235 | 0.0331 |
| Hispanic\_HH\_5T | -0.12089 | 0.037 |
| restaurants\_3R | -0.11826 | 0.0413 |
| Med\_HHinc\_3R | 0.11658 | 0.0443 |
| Pop\_grad\_school\_3R | -0.11565 | 0.0461 |
| HHinc\_75\_99K\_3R | -0.1131 | 0.0511 |
| Big\_box\_index\_1R | 0.11078 | 0.0561 |
| Pop\_50\_69\_5T | -0.10977 | 0.0584 |
| labor\_white\_col\_3R | -0.10673 | 0.0658 |
| per\_cap\_inc\_3R | 0.10664 | 0.066 |
| labor\_farm\_5T | -0.10529 | 0.0695 |
| Pop\_GE\_18\_5T | -0.10487 | 0.0707 |
| Pop\_grades\_9\_12\_5T | -0.10236 | 0.0777 |
| Pop\_40\_49\_5T | -0.10013 | 0.0844 |
| HH\_3person\_5T | -0.09774 | 0.0921 |
| HH\_4person\_5T | -0.096 | 0.0981 |
| Pop\_some\_college\_5T | -0.09548 | 0.0999 |
| Pop\_Doctorate\_3R | -0.09444 | 0.1037 |
| Pop\_Bachelors\_3R | -0.09294 | 0.1094 |
| Pop\_Masters\_3R | -0.09268 | 0.1103 |
| HH\_1person\_5T | -0.09265 | 0.1105 |
| Pop\_21\_39\_5T | -0.09245 | 0.1112 |
| Pop\_70\_85\_5T | -0.09091 | 0.1173 |

\*Note: Table 49 is sorted so that variables with the smallest p-values appear first.