**Miriam Garcia**

**April 21, 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 competitive edge rests on its nutritional value as a healthier alternative to other established fast-food franchises by providing a wider variety of sandwiches with fresh and locally grown ingredients. Timmy Tom’s first ever store opened in 1974 in San Francisco, California. The following three stores were opened in 1976 in the Bay Area in California as Timmy Tom’s sought to expand in the Western region of the United States. Two more stores were also opened in late 1976, one located in the Southern region in Dallas, Texas and the other in the Central region in Chicago, Illinois. As the popularity of quick-service sandwich shops in the United States grew, Timmy Tom’s was able to expand and establish themselves as a competitive healthy fast-food franchise. Although Timmy Tom’s expansion focused mainly on the Western Region—specifically in California, Colorado, Oregon, and Washington—it soon began to infiltrate different regions of the United States.

Timmy Tom’s was on an expansion path and rapid growth in their key markets up until March of 2020. Timmy Tom’s had about 67 locations in the Southern region, 88 in the Central-Eastern region, and 139 locations in the Western Region as of 2019. 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 objectives include to provide full and comprehensive pre-model analysis on the dependent variable, the continuous potential independent variables, the dummy variables, and the LIV variables in the data file. We will also use the 8 steps of model building to construct and estimate the best fitting model possible to explain sales at Timmy Tom’s, given the potential regressors by performing model validation through the calculation of the Out-of-Sample Mean Absolute Percentage Error (OOS MAPE). Our overall purpose is to use our previous regression results to give advice to Timmy Tom’s regarding where to build future locations. 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. 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.

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 Observations

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 a value of zero 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 indicates we are missing the value of Sales for these stores. 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 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 are 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 |

**Coefficient of Variation**

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean (Hayes, 2020). The coefficient of variation represents the ratio of the standard deviation to the mean. Now that we have narrowed down our observations, we get a coefficient of variation of 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

|  |  |
| --- | --- |
| **STORE\_ID/ Number of observation** | **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 defines 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. |
| 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. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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. |
| 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. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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. |
| 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. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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). |
| 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). |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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. |
| 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. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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. |
| 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. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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 store. |
| 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. |
| Black\_HH\_5T | The number of black households located within a 5-minute drive of a given store. |

Table 6: Potential Independent Variables (cont.)

|  |  |
| --- | --- |
| **Variable Name** | **Definition** |
| 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, 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** |
| Traffic\_count | 225 | 22676.73 | 19729.65 | 4645.41 | 178296.58 | 87 |
| Fast\_food\_8T | 298 | 25.14 | 14.7 | -45 | 101 | 58.48 |
| Big\_box\_1R | 298 | 2989.38 | 51491.52 | 0 | 888,888 | 1722.48 |

.

Upon analysis, we observe three potential variables with issues. 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. 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, Traffic\_count will be dropped from further analysis.

Fast\_food\_8T has a negative minimum, which would be indicating that there is a negative

number of fast-food-type restaurants that are located within an 8-minute drive of a given store. This number is misleading as a minimum as there is no such thing as negative fast-food-type restaurants in an area. Therefore, we can drop this variable. The variable Big\_box\_1R refers to the number of big-box stores that are located within one radial mile of a given store. In our summary statistics, we see that the maximum value in our observations is 888,888, which would be indicating that there are 888,888 big-box stores within one radial mile which seems extremely unlikely. Therefore, we can drop this variable.

**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 explain sales, 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. Where Rh0 is the hypothesis that the regressor we are testing for has zero impact on our dependent variable. 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 potential regressor against the null hypothesis. Our null hypothesis states that a regressor’s impact on sales will be equal to zero. This in turn indicates that the regressor is not statistically correlated with sales.

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: Potential Variables – Food Away & Likely Customers

|  |  |  |
| --- | --- | --- |
| **Variable** | Food\_away\_3R | Likely\_customers\_1R |
| **Correlation Coefficient to Sales** | 0.12738 | -0.14658 |
| **P-value** | 0.0279 | 0.0113 |

.

Table 9: Potential Variables – Population

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Pop\_GE\_18\_3R | Pop\_GE\_18\_5T | Pop\_21\_39\_3R | Pop\_21\_39\_5T |
| **Correlation Coefficient to Sales** | -0.17827 | -0.10487 | -0.15457 | -0.09245 |
| **P-value** | 0.0020 | 0.0707 | 0.0075 | 0.1112 |

Table 10: 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 11: 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 12: Potential Variables – Other Indexes

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

Table 13: Potential Variables – Population with Specific Degree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Pop\_Associates\_3R | Pop\_Bachelors\_3R | Pop\_Masters\_3R | Pop\_Doctorate\_3R |
| **Correlation Coefficient to Sales** | -0.15035 | -0.09294 | -0.09268 | -0.09444 |
| **P-value** | 0.0093 | 0.1094 | 0.1103 | 0.1037 |

Table 14: 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 15: Potential Variables – Population in school and Graduate School

|  |  |  |
| --- | --- | --- |
| **Variable** | Pop\_in\_school\_3R | Pop\_grad\_school\_3R |
| **Correlation Coefficient to Sales** | -0.17088 | -0.11565 |
| **P-value** | 0.0031 | 0.0461 |

Table 16: 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 17: Potential Variables – Total Annual Expenditure of Households and Customer Value

|  |  |  |
| --- | --- | --- |
| **Variable** | Tot\_HH\_Expend\_3R | Cust\_value\_per\_cap\_region |
| **Correlation Coefficient to Sales** | -0.15132 | 0.14837 |
| **P-value** | 0.0089 | 0.0103 |

Table 18: Potential Variables – Number of Households with Specific Income Level

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | HHinc\_LT\_25K\_3R | HHinc\_LT\_25K\_5T | HHinc\_25\_49K\_3R |
| **Correlation Coefficient to Sales** | -0.18612 | -0.14236 | -0.19072 |
| **P-value** | 0.0012 | 0.0139 | 0.0009 |

Table 19: Potential Variables – Number of Households with Specific Income Level (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | HHinc\_25\_49K\_5T | HHinc\_50\_74K\_3R | HHinc\_75\_99K\_3R |
| **Correlation Coefficient to Sales** | -0.12350 | -0.16376 | -0.11310 |
| **P-value** | 0.0331 | 0.0046 | 0.0511 |

Table 20: Potential Variables – Median Value of Home & Per Capita Income

|  |  |  |
| --- | --- | --- |
| **Variable** | Med\_HHinc\_3R | per\_cap\_inc\_3R |
| **Correlation Coefficient to Sales** | 0.11658 | 0.10664 |
| **P-value** | 0.0443 | 0.0660 |

Table 21: Potential Variables – Households Headcount

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

Table 22: 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 23: 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 24: Potential Variables – Households Headcount (cont.)

|  |  |  |
| --- | --- | --- |
| **Variable** | Brady\_Bunch\_3R | Brady\_Bunch\_5T |
| **Correlation Coefficient to Sales** | -0.21149 | -0.16844 |
| **P-value** | 0.0002 | 0.0035 |

Table 25: 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 26: Impact of Potential Variables – White Collar Occupation & Married Population

|  |  |  |
| --- | --- | --- |
| **Variable** | labor\_white\_col\_3R | Pop\_married\_3R |
| **Correlation Coefficient to Sales** | -0.10673 | -0.17544 |
| **P-value** | 0.0658 | 0.0024 |

|  |
| --- |
| Table 27: Potential Variables –Restaurants/Retail Establishments Within 3 Radial Miles |

|  |  |  |  |
| --- | --- | --- | --- |
| **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 28: 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 29: 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 30: 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 |

Table 30: Regressors Correlated with Sales & Statistical Significance (cont.)

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Correlation Coefficient** | **P-value** |
| 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 |
| 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 |

Table 30: Regressors Correlated with Sales & Statistical Significance

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Correlation Coefficient** | **P-value** |
| 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 30 is sorted so that variables with the smallest p-values appear first.

**VII. Dummy and Multi-Trait Dummy Variables**

Table 31: New Independent Variables for Analysis – Dummy and Multi-trait Dummy Variables

|  |  |
| --- | --- |
| **Variable** | **Description** |
| HD | This variable indicates whether or not a store is located in a relatively densely populated (high density) geographic region. This variable takes on a value of 1 if a store is located in a high-density area, and it takes on a value of 2 if not. |
| HV | This variable takes on a value of 1 if a store is located in a "high visibility" area, and it takes on a value of 2 if not. |
| free\_standing, strip\_mall | These variables represent a multi-trait dummy variable that indicates the three types of buildings in which a Timmy Tom's store is located: a free-standing building, a strip mall, or any other type of building.  The variable "free\_standing" takes on a value of 1 if a store is located in a free-standing building, and it takes on a value of 0 if not.  The variable "strip\_mall" takes on a value of 1 if a store is located in a strip mall, and it takes on a value of 0 if not.  The base trait is "other\_building." That is, the base trait is any other type of building other than free standing or strip mall.  These three building types are mutually exclusive and exhaustive. |
| south, central, west | These variables represent a multi-trait dummy variable that measures the four regions of the US in which a store is located.  The variable "south" takes on a value of 1 if a store is located in the southern region of the US, and 0 if not.  The variable "central" takes on a value of 1 if a store is located in the central region of the US, and 0 if not.  The variable "west" takes on a value of 1if a store is located in the western region of the US, and 0 if not.  The base trait is "east."  These 4 traits are mutually exclusive and exhaustive. |

.

Table 32: Summary Statistics for Dummy and Multi-Trait Dummy Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| HD | 298 | 1.24 | 0.43 | 1 | 2 |
| HV | 298 | 1.08 | 0.27 | 1 | 2 |
| south | 298 | 0.22 | 0.41 | 0 | 1 |
| central | 298 | 0.3 | 0.46 | 0 | 1 |
| west | 298 | 0.45 | 0.5 | 0 | 1 |
| free\_standing | 298 | 0.61 | 0.49 | 0 | 1 |
| strip\_mall | 298 | 0.33 | 0.47 | 0 | 1 |

When we performed our summary statistics for our dummy and multi-trait dummy variables, we notice that values of minimum and maximum are not our typical 0 and 1 for dummy variables for both HD and HV. This is because they have different thresholds set for if the characteristic is present or absent. We can then transform our dummy variables in order to have the appropriate values of zero and one. Both HD and HV have a different threshold for if the characteristic is present. We will adjust accordingly and convert these into our traditional dummy variables and give them a new name.

If we have east as the base group for geographical locations and calculate its mean, we notice it is 0.03. Which means that the probability of a store being in the east is 3%. This means that east as a base is not properly weighted and cannot stand by itself. The best alternative is to select a new base and combine central with east. The reason for this is that it would make sense since both the central USA and eastern USA are close geographical region wise.

If we have other\_buildings as the base group for location type and calculate its mean, we notice it is 0.06. Which signifies that only 6% of stores are located in other\_buildings. Since this is the base group, it is not properly weighted and cannot stand by itself as the base. Our best alternative is again to select a new base, which we can do first by combining other\_buildings with strip\_mall. This move would make sense since the alternative would then be that a store is either free standing or “attached” to other businesses.

Table 33: Corrected Dummy and Multi-Trait Dummy Variables

|  |  |
| --- | --- |
| **Variable** | **Description** |
| High\_density | This variable indicates whether or not a store is located in a relatively densely populated (high density) geographic region.  This variable takes on a value of 1 if a store is located in a high-density area, and it takes on a value of 0 if not. |
| High\_visibility | This variable takes on a value of 1 if a store is located in a "high visibility" area, and it takes on a value of 0 if not. |
| Central\_east | The variable "central\_east" takes on a value of 1 if a store is located in the central or eastern region of the US, and 0 if not.  The base trait is now "central\_east." |
| Strip\_mall\_other | The variable "strip\_mall\_other" takes on a value of 1 if a store is located in a strip mall or any other type of building, and it takes on a value of 0 if not.  The base trait is now “strip\_mall\_other”. |

.

Now, we can compute our summary statistics for our adjusted dummy variables and our multi-trait dummy variables.

Table 34: Summary Statistics for Corrected Dummy and Multi-trait Dummy Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| high\_visibility | 298 | 0.92 | 0.27 | 0 | 1 |
| high\_density | 298 | 0.76 | 0.43 | 0 | 1 |
| south | 298 | 0.22 | 0.41 | 0 | 1 |
| central\_east | 298 | 0.33 | 0.46 | 0 | 1 |
| west | 298 | 0.45 | 0.5 | 0 | 1 |
| free\_standing | 298 | 0.61 | 0.49 | 0 | 1 |
| strip\_mall\_other | 298 | 0.39 | 0.47 | 0 | 1 |

.

We require sufficient variation in order for a dummy variable to be relevant in our analysis. For example, a very high mean close to 1 indicates that almost all observations possess this trait. And a very low mean close to 0, indicates that almost no observations possess this trait. We set our threshold of sufficient variation in a dummy variable as a high mean of 0.9 or higher, indicates this variable is significantly present in all observations and is therefore not displaying sufficient variation since the characteristic is highly present in most of our observations. And for that trait to not be present in most of our observations, we set our threshold to be 0.1 or lower to indicate that since there is not sufficient variation, this trait is lacking in most of our observations. In our case, the variable high\_visibility has a mean of 0.92, which is higher than our 0.9 threshold for a trait to be considered predominantly present in all observations. This translates to the fact that almost all of our stores—92% of our observations—are located in a "high visibility" area and this variable can be omitted in our analysis.

Table 35: Dummy Variables Without Sufficient Variation

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Deviation** |
| high\_visibility | 298 | 0.92 | 0.27 |

.

Now that we have corrected our dummy variables and taken in account sufficient variation, we can now compute our correlation coefficients and the associated p-values for each dummy variable in order to determine which ones if any are significantly correlated to Sales.

Table 36: Dummy and Multi-Trait Dummy Variables Correlation Coefficients

|  |  |  |
| --- | --- | --- |
| **Variable** | **Correlation to Sales** | **P-value** |
| high\_density | -0.03664 | 0.5286 |
| south | -0.63594 | <.0001 |
| central\_east | -0.23884 | <.0001 |
| west | 0.80827 | <.0001 |
| free\_standing | 0.32271 | <.0001 |
| strip\_mall\_other | -0.25803 | <.0001 |

.

With these correlation statistics, we can now perform hypothesis testing in order to determine what variables have the strongest correlation to our dependent variable, Sales. Our hypothesis being tested is that if there is any correlation between these variables to sales, and if there is not any correlation then that variable is insignificant. If so, our null hypothesis establishes that correlation does not exist and is zero. We use our p-values from our correlation table in order to determine this. If we have a small p-value, this serves as evidence that we should reject the null hypothesis. If we reject the null hypothesis, this indicates that there is significant correlation between sales and that variable. Since we are testing at a 88% confidence level, we are looking for p-values that are less than or equal to 0.12. We will utilize the Pearson correlation coefficient to evaluate the correlation between these multi-trait dummy variables and Sales. There is a negative relationship between sales and a trait being present if there is a negative correlation coefficient. And the opposite is true for positive correlation coefficient. The P-value given can be evaluated under the null hypothesis to see if a variable will have impact on the dependent variable.

All of these variables except for high\_density have a p-value less than 0.12 which means that they are statistically significant to our dependent variable, Sales, south, central\_east, and strip\_mall\_other all have negative correlation coefficients. Which indicates that if these traits are present in our observations, sales will be negatively impacted. For south, this indicates that if a Timmy Tom’s location is located in the south, sales will decrease. For central, this tells us that if a location is in the west region, sales will decrease. If a location is in a strip mall or other type of building, sales will decrease. On the other hand, west and free standing have positive correlation coefficients. Which indicates that if these traits are present in our observations, sales will be positively impacted. If a location is free standing, sales will increase. And if a location is in the west region, sales will increase.

**VIII. Limited Integer Value (LIV) Variables**

Limited-integer-value (LIV) variables are variables that take on only integer values and there are a limited number of values that they can take on, in our case it will be limited to six values. In the case of LIV variables, we must create frequency tables since our usual summary statistics will not be enough. The reason being that a frequency table shows all of the values that a variable can take on, and how many times it takes on each value, both in absolute terms and in percentage term. In our data, the LIV variables are related to the proximity and presence of the following to a Timmy Tom’s location: malls, bakeries, big-box stores, competitors, low-end grocery stores, mid-level grocery stores, other sandwich shops, and universities. We will consider the following potential LIV Variables.

Table 37: New Independent Variables for Analysis – Potential LIV Variables

|  |  |
| --- | --- |
| **Variable** | **Description** |
| All\_malls\_1R | The number of malls (of all types) that are located within one radial mile of a given store. |
| Malls\_300K\_0\_5R | The number of malls (with more than 300,000 square feet of gross leasable area) that are located within a one-half radial mile of a given store. |
| Malls\_300K\_1R | The number of malls (with more than 300,000 square feet of gross leasable area) that are located within one radial mile of a given store. |
| Bakeries\_0\_5R | The number of bakery-type restaurants (such as Panera, etc.) that are located within a one-half radial mile of a given store. |
| Bakeries\_1R | The number of bakery-type restaurants (such as Panera, etc.) that are located within one radial mile of a given store. |
| Competitor\_A\_0\_5R | The number of competitor A stores that are located within a one-half radial mile of a given store. |
| Competitor\_A\_1R | The number of competitor A stores that are located within one radial mile of a given store. |
| Competitor\_B\_0\_5R | The number of competitor B stores that are located within a one-half radial mile of a given store. |
| Competitor\_B\_1R | The number of competitor B stores that are located within one radial mile of a given store. |
| Competitor\_C\_0\_5R | The number of competitor C stores that are located within a one-half radial mile of a given store. |
| Competitor\_C\_1R | The number of competitor C stores that are located within one radial mile of a given store. |
| Competitor\_D\_0\_5R | The number of competitor D stores that are located within a one-half radial mile of a given store. |
| Competitor\_D\_1R | The number of competitor D stores that are located within one radial mile of a given store. |
| Big\_box\_0\_5R | The number of big-box stores (such as Best Buy, Target, etc.) that are located within a one-half radial mile of a given store. |
| Low\_grocery\_0\_5R | The number of low-end grocery stores (such as Aldi, Sack N Save, etc.) that are located within a one-half radial mile of a given store. |
| Low\_grocery\_1R | The number of low-end grocery stores (such as Aldi, Sack N Save, etc.) that are located within one radial mile of a given store. |

Table 37: New Independent Variables for Analysis – Potential LIV Variables (cont.)

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Mid\_grocery\_0\_5R | The number of mid-level grocery stores (such as Kroger, Safeway, etc.) that are located within a one-half radial mile of a given store. |
| Mid\_grocery\_1R | The number of mid-level grocery stores (such as Kroger, Safeway, etc.) that are located within one radial mile of a given store. |
| Sandwich\_shop\_8T | The number of sandwich shops (other than Timmy Tom's, such as Subway, Jersey Mike's, etc.) that are located within an 8-minute drive of a given store. |
| Universities\_0\_5R | The number of 4-year universities located within a one-half radial mile of a given store. |
| Universities\_1R | The number of 4-year universities located within one radial mile of a given store. |
| Universities\_3R | The number of 4-year universities located within 3 radial miles of a given store. |
| Universities\_5T | The number of 4-year universities located within a 5-minute drive of a given store. |
| Universities\_8T | The number of 4-year universities located within an 8-minute drive of a given store. |

.

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

**Malls**

Table 38: All Malls Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **All\_malls\_1R** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| 0 | 43 | 14.43 | 43 | 14.43 |
| 1 | 39 | 13.09 | 82 | 27.52 |
| 2 | 44 | 14.77 | 126 | 42.28 |
| 3 | 45 | 15.1 | 171 | 57.38 |
| 4 | 29 | 9.73 | 200 | 67.11 |
| 5 | 27 | 9.06 | 227 | 76.17 |
| 6 | 19 | 6.38 | 246 | 82.55 |
| 7 | 11 | 3.69 | 257 | 86.24 |
| 8 | 13 | 4.36 | 270 | 90.6 |
| 9 | 9 | 3.02 | 279 | 93.62 |
| 10 | 7 | 2.35 | 286 | 95.97 |

Table 38: All Malls Within One Radial Mile Frequency Table (cont.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **All\_malls\_1R** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| 11 | 6 | 2.01 | 292 | 97.99 |
| 12 | 1 | 0.34 | 293 | 98.32 |
| 13 | 1 | 0.34 | 294 | 98.66 |
| 15 | 1 | 0.34 | 295 | 98.99 |
| 16 | 2 | 0.67 | 297 | 99.66 |
| 18 | 1 | 0.34 | 298 | 100 |

.

The frequency table for All\_malls\_1R shows us that this variable can take on more than six values, and is therefore not an LIV variable but a typical regressor.

Table 39: Malls (300k sqft) Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Malls\_300K\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 221 | 74.16 | 221 | 74.16 |
| 1 | 62 | 20.81 | 283 | 94.97 |
| 2 | 14 | 4.7 | 297 | 99.66 |
| 3 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Malls\_300K\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 74.16. Our criteria requires all of them to be 10% or higher.

Table 40: Malls (300k sqft) Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Malls\_300K\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 184 | 61.74 | 184 | 61.74 |
| 1 | 73 | 24.5 | 257 | 86.24 |
| 2 | 27 | 9.06 | 284 | 95.3 |
| 3 | 8 | 2.68 | 292 | 97.99 |
| 4 | 6 | 2.01 | 298 | 100 |

.

The frequency table for Malls\_300K\_1R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 2.01 up to 61.74. Our criteria requires all of them to be 10% or higher.

**Bakeries**

Table 41: Bakeries Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bakeries\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 216 | 72.48 | 216 | 72.48 |
| 1 | 64 | 21.48 | 280 | 93.96 |
| 2 | 15 | 5.03 | 295 | 98.99 |
| 3 | 2 | 0.67 | 297 | 99.66 |
| 4 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Bakeries\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 72.48. Our criteria requires all of them to be 10% or higher.

Table 42: Bakeries Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bakeries\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 167 | 56.04 | 167 | 56.04 |
| 1 | 78 | 26.17 | 245 | 82.21 |
| 2 | 38 | 12.75 | 283 | 94.97 |
| 3 | 12 | 4.03 | 295 | 98.99 |
| 4 | 3 | 1.01 | 298 | 100 |

.

The frequency table for Bakeries\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 56.04. Our criteria requires all of them to be 10% or higher.

**Competitors**

Table 43: Competitor A Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_A\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 238 | 79.87 | 238 | 79.87 |
| 1 | 59 | 19.8 | 297 | 99.66 |
| 2 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Competitor\_A\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 79.87. Our criteria requires all of them to be 10% or higher.

Table 44: Competitor A Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_A\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 184 | 61.74 | 184 | 61.74 |
| 1 | 110 | 36.91 | 294 | 98.66 |
| 2 | 4 | 1.34 | 298 | 100 |

.

The frequency table for Competitor\_A\_1R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 1.34 up to 61.74. Our criteria requires all of them to be 10% or higher.

Table 45: Competitor B Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_B\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 287 | 96.31 | 287 | 96.31 |
| 1 | 10 | 3.36 | 297 | 99.66 |
| 2 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Competitor\_B\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 96.31. Our criteria requires all of them to be 10% or higher. Here we notice a high percent of zero stores within one-half radial mile, which indicates that the frequency of Competitor B’s presence within one-half radial mile to a Timmy Tom’s within that threshold is almost non-existing.

Table 46: Competitor B Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_B\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 281 | 94.3 | 281 | 94.3 |
| 1 | 16 | 5.37 | 297 | 99.66 |
| 2 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Competitor\_B\_1R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 94.3. Our criteria requires all of them to be 10% or higher. Here we notice a high percent of zero stores within one-half radial mile, which indicates that the frequency of Competitor B’s presence within one radial mile to a Timmy Tom’s within that threshold is almost non-existent.

Table 47: Competitor C Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_C\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 276 | 92.62 | 276 | 92.62 |
| 1 | 22 | 7.38 | 298 | 100 |

.

Now that we have computed the frequency table for Competitor\_C\_0\_5R, we see that it is in fact a traditional dummy variable that takes a value of 1 if the trait is present and a value of 0 if the trait is absent. Here we notice a high percent of zero stores within one-half radial mile, which indicates that the frequency of Competitor C’s presence within one-half radial mile to a Timmy Tom’s within that threshold is almost non-existent.

Table 48: Competitor C Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_C\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 248 | 83.22 | 248 | 83.22 |
| 1 | 50 | 16.78 | 298 | 100 |

.

After computing the frequency table for Competitor\_C\_1R, we see that it is in fact a traditional dummy variable.

Table 49: Competitor D Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_D\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 277 | 92.95 | 277 | 92.95 |
| 1 | 21 | 7.05 | 298 | 100 |

.

After computing the frequency table for Competitor\_D\_0\_5R, we see that it is in fact a traditional dummy variable. Here we notice a high percent of zero stores within one-half radial mile, which indicates that the frequency of Competitor D’s presence within one-half radial mile to a Timmy Tom’s within that threshold is almost non-existent.

Table 50: Competitor D Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competitor\_D\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 240 | 80.54 | 240 | 80.54 |
| 1 | 58 | 19.46 | 298 | 100 |

After computing the frequency table for Competitor\_D\_1R, we see that it is in fact a traditional dummy variable.

**Big-Box and Grocery Stores**

Table 51: Big-box Stores Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Big\_box\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 86 | 28.86 | 86 | 28.86 |
| 1 | 65 | 21.81 | 151 | 50.67 |
| 2 | 28 | 9.4 | 179 | 60.07 |
| 3 | 20 | 6.71 | 199 | 66.78 |
| 4 | 18 | 6.04 | 217 | 72.82 |
| 5 | 10 | 3.36 | 227 | 76.17 |
| 6 | 14 | 4.7 | 241 | 80.87 |
| 7 | 15 | 5.03 | 256 | 85.91 |
| 8 | 10 | 3.36 | 266 | 89.26 |
| 9 | 7 | 2.35 | 273 | 91.61 |
| 10 | 5 | 1.68 | 278 | 93.29 |
| 11 | 4 | 1.34 | 282 | 94.63 |
| 12 | 3 | 1.01 | 285 | 95.64 |
| 13 | 5 | 1.68 | 290 | 97.32 |
| 14 | 2 | 0.67 | 292 | 97.99 |
| 15 | 1 | 0.34 | 293 | 98.32 |
| 19 | 3 | 1.01 | 296 | 99.33 |
| 24 | 1 | 0.34 | 297 | 99.66 |
| 26 | 1 | 0.34 | 298 | 100 |

The frequency table for Big\_box\_0\_5R shows us that this variable can take on more than six values, and is therefore not an LIV variable but a typical regressor.

Table 52: Low-Grocery Store Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Low\_grocery\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 273 | 91.61 | 273 | 91.61 |
| 1 | 25 | 8.39 | 298 | 100 |

.

After computing the frequency table for Low\_grocery\_0\_5R, we see that it is in fact a traditional dummy variable. Here we notice a high percent of zero low grocery stores within one-half radial mile, which indicates that the frequency of Low grocery stores’ presence within one-half radial mile to a Timmy Tom’s within that threshold is almost non-existent.

Table 53: Low-Grocery Store Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Low\_grocery\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 184 | 61.74 | 184 | 61.74 |
| 1 | 110 | 36.91 | 294 | 98.66 |
| 2 | 4 | 1.34 | 298 | 100 |

The frequency table for Low\_grocery\_1R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 1.34 up to 61.74. Our criteria requires all of them to be 10% or higher.

Table 54: Mid-Grocery Store Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mid\_grocery\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 197 | 66.11 | 197 | 66.11 |
| 1 | 89 | 29.87 | 286 | 95.97 |
| 2 | 9 | 3.02 | 295 | 98.99 |
| 3 | 3 | 1.01 | 298 | 100 |

.

The frequency table for Mid\_grocery\_0\_5R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 1.01 up to 66.11. Our criteria requires all of them to be 10% or higher. Although it might not be an LIV variable, it can still be a variable relevant to our analysis.

Table 55: Mid-Grocery Store Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mid\_grocery\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 143 | 47.99 | 143 | 47.99 |
| 1 | 121 | 40.6 | 264 | 88.59 |
| 2 | 30 | 10.07 | 294 | 98.66 |
| 3 | 3 | 1.01 | 297 | 99.66 |
| 4 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Mid\_grocery\_1R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 47.99. Our criteria requires all of them to be 10% or higher. Three of these percentages meet the criteria, which indicates we might be able to adjust this variable in order to use it for our analysis.

**Sandwich Shops**

Table 56: Sandwich Shops Within 8 Minute Drive Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sandwich\_shop\_8T | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 40 | 13.42 | 40 | 13.42 |
| 1 | 45 | 15.1 | 85 | 28.52 |
| 2 | 45 | 15.1 | 130 | 43.62 |
| 3 | 28 | 9.4 | 158 | 53.02 |
| 4 | 40 | 13.42 | 198 | 66.44 |
| 5 | 21 | 7.05 | 219 | 73.49 |
| 6 | 13 | 4.36 | 232 | 77.85 |
| 7 | 22 | 7.38 | 254 | 85.23 |
| 8 | 10 | 3.36 | 264 | 88.59 |
| 9 | 12 | 4.03 | 276 | 92.62 |
| 10 | 6 | 2.01 | 282 | 94.63 |
| 11 | 6 | 2.01 | 288 | 96.64 |
| 12 | 3 | 1.01 | 291 | 97.65 |
| 13 | 3 | 1.01 | 294 | 98.66 |
| 15 | 1 | 0.34 | 295 | 98.99 |
| 17 | 2 | 0.67 | 297 | 99.66 |
| 19 | 1 | 0.34 | 298 | 100 |

.

The frequency table for Sandwich\_Shop\_8T shows us that this variable can take on more than six values, and is therefore not an LIV variable but a typical regressor.

**Universities**

Table 57: Universities Within One-Half Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Universities\_0\_5R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 293 | 98.32 | 293 | 98.32 |
| 1 | 5 | 1.68 | 298 | 100 |

.

After computing the frequency table for Universities\_0\_5R, we see that it is in fact a traditional dummy variable. Here we notice a high percent of zero universities within one-half radial mile, which indicates that the frequency of universities presence within one-half radial mile to a Timmy Tom’s is almost non-existent.

Table 58: Universities Within One Radial Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Universities\_1R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 278 | 93.29 | 278 | 93.29 |
| 1 | 20 | 6.71 | 298 | 100 |

After computing the frequency table for Universities\_1R, we see that it is in fact a traditional dummy variable. Here we notice a high percent of zero universities within one radial mile, which might be indicating that the frequency of universities presence within one radial mile to a Timmy Tom’s is almost non-existent.

Table 59: Universities Within Three Radial Miles Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Universities\_3R | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 228 | 76.51 | 228 | 76.51 |
| 1 | 56 | 18.79 | 284 | 95.3 |
| 2 | 11 | 3.69 | 295 | 98.99 |
| 3 | 1 | 0.34 | 296 | 99.33 |
| 4 | 2 | 0.67 | 298 | 100 |

.

The frequency table for Universities\_3R, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 76.51. Our criteria requires all of them to be 10% or higher.

Table 60: Universities Within 5 Minute Drive Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Universities\_5T | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 279 | 93.62 | 279 | 93.62 |
| 1 | 18 | 6.04 | 297 | 99.66 |
| 2 | 1 | 0.34 | 298 | 100 |

The frequency table for Universities\_5T, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 93.62. Our criteria requires all of them to be 10% or higher. In this case, we see that in all of our observations, we frequently see zero universities within a five-minute drive close to a Timmy Tom’s (93.62% of the time), which indicates that this variable would not be useful in our analysis.

Table 61: Universities Within 8 Minute Drive Mile Frequency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Universities\_5T | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 0 | 244 | 81.88 | 244 | 81.88 |
| 1 | 47 | 15.77 | 291 | 97.65 |
| 2 | 6 | 2.01 | 297 | 99.66 |
| 3 | 1 | 0.34 | 298 | 100 |

The frequency table for Universities\_8T, does not meet our criteria since this LIV variable does not have sufficient variation. Percentages range from 0.34 up to 81.88. Our criteria requires all of them to be 10% or higher.

Table 62: Summary Statistics for New Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| Big\_box\_0\_5R | 298 | 3.29 | 4.24 | 0 | 26 |
| Sandwich\_shop\_8T | 298 | 3.94 | 3.49 | 0 | 19 |
| All\_malls\_1R | 298 | 3.74 | 3.27 | 0 | 18 |
| Competitor\_A\_0\_5R | 298 | 0.2 | 0.41 | 0 | 2 |
| Competitor\_A\_1R | 298 | 0.4 | 0.52 | 0 | 2 |
| Competitor\_B\_0\_5R | 298 | 0.04 | 0.21 | 0 | 2 |
| Competitor\_B\_1R | 298 | 0.06 | 0.25 | 0 | 2 |
| Competitor\_C\_0\_5R | 298 | 0.07 | 0.26 | 0 | 1 |
| Competitor\_C\_1R | 298 | 0.17 | 0.37 | 0 | 1 |
| Competitor\_D\_0\_5R | 298 | 0.07 | 0.26 | 0 | 1 |
| Competitor\_D\_1R | 298 | 0.19 | 0.4 | 0 | 1 |
| Bakeries\_0\_5R | 298 | 0.35 | 0.64 | 0 | 4 |
| Bakeries\_1R | 298 | 0.68 | 0.92 | 0 | 4 |
| Low\_grocery\_0\_5R | 298 | 0.08 | 0.28 | 0 | 1 |
| Low\_grocery\_1R | 298 | 0.18 | 0.39 | 0 | 2 |
| Mid\_grocery\_0\_5R | 298 | 0.39 | 0.6 | 0 | 3 |
| Mid\_grocery\_1R | 298 | 0.65 | 0.73 | 0 | 4 |
| Malls\_300K\_0\_5R | 298 | 0.31 | 0.57 | 0 | 3 |
| Malls\_300K\_1R | 298 | 0.59 | 0.91 | 0 | 4 |
| Universities\_0\_5R | 298 | 0.02 | 0.13 | 0 | 1 |
| Universities\_1R | 298 | 0.07 | 0.25 | 0 | 1 |
| Universities\_3R | 298 | 0.3 | 0.62 | 0 | 4 |
| Universities\_5T | 298 | 0.07 | 0.26 | 0 | 2 |
| Universities\_8T | 298 | 0.21 | 0.48 | 0 | 3 |

In our analysis, we have found proper and improper LIV variables. Given our potential LIV variables, the given frequency tables, and the summary statistics of all these potential LIV variables, we can determine that the only two variables that meets our criteria of having sufficient variation and being a proper LIV variables are Competitor\_C\_1R and Competitor\_D\_1R but even so they are actually traditional dummy variables that take on a value of zero or one. We also can determine that although some variables were intended to be LIV variables, they can function as regular variables in our analysis. Those include All\_malls\_1R, Big\_box\_0\_5R, and Sandwich\_shop\_8T.

Table 63: Variables Without Issues

|  |  |
| --- | --- |
| **Variable Name** | **Details** |
| Competitor\_C\_1R | LIV proper variable |
| Competitor\_D\_1R | LIV proper variable |
| All\_malls\_1R | Regular proper variable |
| Big\_box\_0\_5R | Regular proper variable |
| Sandwich\_shop\_8T | Regular proper variable |

.

Of our potential LIV variables, there are some that we have determined that given their frequency tables, we can toss out or corrected. For some of these variables, we can perform corrections. These potential improper LIV variables can be corrected to ensure they have sufficient variation. To do so, we can manipulate those LIV variables without sufficient variation and turn them into dummy variables. For that, we must combine the rest of the observations and have the variable be equal to “zero” if there are no locations or “one” if there are any locations regardless of frequency.

Table 64: Potential Variables and Solutions

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Issue** | **Solution (Convert/Toss)** |
| Malls\_300K\_0\_5R | Improper LIV variable, convert to dummy variable. | Convert |
| Malls\_300K\_1R | Improper LIV variable, convert to dummy variable. | Convert |
| Bakeries\_0\_5R | Improper LIV variable, convert to dummy variable. | Convert |
| Bakeries\_1R | Improper LIV variable, convert to dummy variable. | Convert |
| Competitor\_A\_0\_5R | Improper LIV variable, convert to dummy variable. | Convert |
| Competitor\_A\_1R | Improper LIV variable, convert to dummy variable. | Convert |
| Competitor\_B\_0\_5R | In 96.31% of our observations, there are zero Competitor B’s stores within one-half radial mile of a Timmy Tom’s. | Toss |
| Competitor\_B\_1R | In 94.3% of our observations, there are zero Competitor B’s stores within one radial mile of a Timmy Tom’s. | Toss |
| Competitor\_C\_0\_5R | In 92.62% of our observations, there are zero Competitor C’s stores within one-half radial mile of a Timmy Tom’s. | Toss |

Table 64: Potential Variables and Solutions (cont.)

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Issue** | **Solution (Convert/Toss)** |
| Competitor\_D\_0\_5R | In 92.95% of our observations, there are zero Competitor D’s stores within one-half radial mile of a Timmy Tom’s. | Toss |
| Low\_grocery\_0\_5R | In 91.61% of our observations, there are zero Competitor D’s stores within one-half radial mile of a Timmy Tom’s. | Toss |
| Low\_grocery\_1R | Improper LIV variable, convert to dummy variable. | Convert |
| Mid\_grocery\_0\_5R | Improper LIV variable, convert to dummy variable. | Convert |
| Universities\_0\_5R | In 98.32% of our observations, there are zero universities within one-half radial mile of a Timmy Tom’s. | Toss |
| Universities\_1R | In 93.29% of our observations, there are zero universities within one radial mile of a Timmy Tom’s. | Toss |
| Universities\_3R | Improper LIV variable, convert to dummy variable. | Convert |
| Universities\_5T | Frequently see zero universities within a five-minute drive close to a Timmy Tom’s in 93.62% of our observations. | Toss |
| Universities\_8T | Improper LIV variable, convert to dummy variable. | Convert |

.

Of those variables that we will toss since they do not display sufficient variation, we must remove from our regression analysis since the frequency of there being “zero” locations with that criteria near our stores indicates that adding it to our analysis would be irrelevant since they are not predominant near our stores.

The reasoning behind converting the found improper variables into dummy variables is under the assumption that these variables have sufficient variation if instead we set the parameter to be having zero of this variable or one or more of these variables in the data. In our frequency tables, we can see that Competitor\_A\_0\_5R can be saved if we convert it into a traditional dummy variable since in all our observations Competitor\_A\_0\_5R has a 79.87% of frequency of having zero stores near Timmy Tom’s, and combining the alternative of at least one store or more that is 20.14%. For the variable Competitor\_A\_1R, there is a 61.74% chance that there are zero stores near Timmy Tom’s and a 38.25% chance of at least one or more store being present. For the variable Bakeries\_0\_5R, there is a 72.48% chance that there are zero bakeries near Timmy Tom’s within one-half radial mile and a 27.52% chance of at least one or more bakeries being present within one-half radial mile.

For the variable Bakeries\_1R, there is a 56.04% chance that there are zero bakeries near Timmy Tom’s within one radial mile and a 43.96% chance of at least one or more bakeries being present within one radial mile. For the variable Low\_grocery\_1R, there is a 82.55% chance that there are zero low end grocery stores within one radial mile of a Timmy Tom’s and a 17.45% chance that there is at least one or more. For the variable Mid\_grocery\_0\_5R, there is a 66.11% chance that there are zero mid-level grocery stores within one-half radial mile of a Timmy Tom’s and a 33.9% chance that there is at least one or more.

For the variable Malls\_300K\_0\_5R, 74.16% of our observations have zero of these malls and there is at least one or more of these malls in 25.85% of our observations. For the variable Malls\_300K\_1R, 61.74% of our observations have zero of these malls and there is at least one or more of these malls in 38.25% of our observations.

For the variable Universities\_3R, 76.51% of our observations have zero universities within three radial miles and at least one or more of these universities 23.49% of the time. For the variable Universities\_8T, 81.88% of our observations have zero universities within an eight-minute drive and at least one or more of these universities 18.12% of the time.

**V. Corrected LIV variables**

Of the improper LIV variables we have identified, we can convert them into dummy variables and renamed as such.

Table 65: Previous and Corrected LIV Variables

|  |  |
| --- | --- |
| **Previous Variable Name** | **Corrected Variable Name** |
| Competitor\_A\_0\_5R | Competitor\_A\_0\_5R\_dum |
| Competitor\_A\_1R | Competitor\_A\_1R\_dum |
| Bakeries\_0\_5R | Bakeries\_0\_5R\_dum |
| Bakeries\_1R | Bakeries\_1R\_dum |
| Low\_grocery\_1R | Low\_grocery\_1R\_dum |
| Mid\_grocery\_0\_5R | Mid\_grocery\_0\_5R\_dum |
| Malls\_300K\_0\_5R | Malls\_300K\_0\_5R\_dum |
| Malls\_300K\_1R | Malls\_300K\_1R\_dum |
| Universities\_3R | Universities\_3R\_dum |
| Universities\_8T | Universities\_8T\_dum |

.

Table 66: Newly Defined Dummy Variables

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Competitor\_A\_0\_5R\_dum | The variable "Competitor\_A\_0\_5R\_dum" takes on a value of 1 if a store is located near a competitor A store within a one-half radial mile of a given store, and it takes on a value of 0 if not. |
| Competitor\_A\_1R\_dum | The variable "Competitor\_A\_1R\_dum" takes on a value of 1 if a store is located near a competitor A store within a one radial mile of a given store, and it takes on a value of 0 if not. |
| Bakeries\_0\_5R\_dum | The variable "Bakeries\_0\_5R\_dum" takes on a value of 1 if a store is located near a bakery-type restaurants located within one-half radial mile of a given store, and it takes on a value of 0 if not. |
| Bakeries\_1R\_dum | The variable "Bakeries\_0\_5R\_dum" takes on a value of 1 if a store is located near a bakery-type restaurants located within one radial mile of a given store, and it takes on a value of 0 if not. |
| Low\_grocery\_1R\_dum | The variable " Low\_grocery\_1R\_dum " takes on a value of 1 if a store is located near a low-end grocery store located within one radial mile of a given store, and it takes on a value of 0 if not. |
| Mid\_grocery\_0\_5R\_dum | The variable “Mid\_grocery\_0\_5R\_dum" takes on a value of 1 if a store is located near a mid-end grocery store located within one-half radial mile of a given store, and it takes on a value of 0 if not. |
| Malls\_300K\_0\_5R\_dum | The variable “Malls\_300K\_0\_5R\_dum" takes on a value of 1 if a store is located near a 300 square feet mall within one-half radial mile of a given store, and it takes on a value of 0 if not. |

Table 66: Newly Defined Dummy Variables (cont.)

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Malls\_300K\_1R\_dum | The variable “Malls\_300K\_1R\_dum" takes on a value of 1 if a store is located near a 300 square feet mall within one radial mile of a given store, and it takes on a value of 0 if not. |
| Universities\_3R\_dum | The variable “Universities\_3R\_dum" takes on a value of 1 if a store is located within three radial miles of a university, and it takes on a value of 0 if not. |
| Universities\_8T\_dum | The variable “Universities\_8T\_dum" takes on a value of 1 if a store is located within an eight minute drive of a university, and it takes on a value of 0 if not. |

Table 67: Summary Statistics of Corrected Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| Competitor\_A\_0\_5R\_dum | 298 | 0.2 | 0.4 | 0 | 1 |
| Competitor\_A\_1R\_dum | 298 | 0.38 | 0.49 | 0 | 1 |
| Bakeries\_0\_5R\_dum | 298 | 0.28 | 0.45 | 0 | 1 |
| Bakeries\_1R\_dum | 298 | 0.44 | 0.5 | 0 | 1 |
| Low\_grocery\_1R\_dum | 298 | 0.17 | 0.38 | 0 | 1 |
| Mid\_grocery\_0\_5R\_dum | 298 | 0.34 | 0.47 | 0 | 1 |
| Malls\_300K\_0\_5R\_dum | 298 | 0.26 | 0.44 | 0 | 1 |
| Malls\_300K\_1R\_dum | 298 | 0.38 | 0.49 | 0 | 1 |
| Universities\_3R\_dum | 298 | 0.23 | 0.42 | 0 | 1 |
| Universities\_8T\_dum | 298 | 0.18 | 0.39 | 0 | 1 |

.

We can also now calculate the correlation to Sales of these adjusted LIV variables and for any non-problematic variables. Using the Pearson Correlation Coefficient, we will evaluate at an 88% confidence level where P-values should be no greater than 0.12. Under the null hypothesis, we state that if a coefficient has a P-value smaller than 0.12, then we fail to reject the null hypothesis that states that that coefficient is statistically significant.

Table 68: Summary Statistics for LIV and Other Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| Sales | 298 | 872037 | 359823 | 2.6E+08 | 198153 | 1925872 |
| Competitor\_C\_1R | 298 | 0.16779 | 0.3743 | 50 | 0 | 1 |
| Competitor\_D\_1R | 298 | 0.19463 | 0.39658 | 58 | 0 | 1 |
| All\_malls\_1R | 298 | 3.74161 | 3.26759 | 1115 | 0 | 18 |
| Big\_box\_0\_5R | 298 | 3.28859 | 4.23676 | 980 | 0 | 26 |
| Sandwich\_shop\_8T | 298 | 3.93624 | 3.48725 | 1173 | 0 | 19 |

.

Table 69: Correlation Coefficient Statistics for LIV and Other Variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Correlation Coefficient | P-value |
| Competitor\_C\_1R | -0.00665 | 0.909 |
| Competitor\_D\_1R | 0.05836 | 0.3154 |
| All\_malls\_1R | -0.04306 | 0.459 |
| Big\_box\_0\_5R | 0.10083 | 0.0823 |
| Sandwich\_shop\_8T | -0.03838 | 0.5092 |

.

Given our null hypothesis, we state that if a coefficient has a P-value smaller than 0.12, then we fail to reject the null hypothesis that states that that coefficient is statistically significant. In our variables above, the only variable that has p-values smaller than 0.12 is Big\_box\_0\_5R. This indicates that Big\_box\_0\_5R is statistically significant in the explanation of sales.

Table 70: Summary Statistics for Corrected LIV Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| Competitor\_A\_0\_5R\_dum | 298 | 0.20134 | 0.40168 | 60 | 0 | 1 |
| Competitor\_A\_1R\_dum | 298 | 0.38255 | 0.48683 | 114 | 0 | 1 |
| Bakeries\_0\_5R\_dum | 298 | 0.27517 | 0.44735 | 82 | 0 | 1 |
| Bakeries\_1R\_dum | 298 | 0.4396 | 0.49717 | 131 | 0 | 1 |
| Low\_grocery\_1R\_dum | 298 | 0.1745 | 0.38017 | 52 | 0 | 1 |
| Mid\_grocery\_0\_5R\_dum | 298 | 0.33893 | 0.47414 | 101 | 0 | 1 |
| Malls\_300K\_0\_5R\_dum | 298 | 0.25839 | 0.43849 | 77 | 0 | 1 |
| Malls\_300K\_1R\_dum | 298 | 0.38255 | 0.48683 | 114 | 0 | 1 |
| Universities\_3R\_dum | 298 | 0.2349 | 0.42465 | 70 | 0 | 1 |
| Universities\_8T\_dum | 298 | 0.18121 | 0.38584 | 54 | 0 | 1 |

Table 71: Correlation Coefficient Statistics for Corrected LIV Variables

|  |  |  |
| --- | --- | --- |
| **Variable** | Correlation Coefficient | P-value |
| Competitor\_A\_0\_5R\_dum | 0.05983 | 0.3033 |
| Competitor\_A\_1R\_dum | -0.0651 | 0.2626 |
| Bakeries\_0\_5R\_dum | 0.01996 | 0.7315 |
| Bakeries\_1R\_dum | 0.03456 | 0.5523 |
| Low\_grocery\_1R\_dum | -0.00145 | 0.9801 |
| Mid\_grocery\_0\_5R\_dum | -0.05408 | 0.3522 |
| Malls\_300K\_0\_5R\_dum | 0.09808 | 0.091 |
| Malls\_300K\_1R\_dum | 0.08093 | 0.1635 |
| Universities\_3R\_dum | 0.00358 | 0.951 |
| Universities\_8T\_dum | 0.03906 | 0.5018 |

.

If we run our correlation coefficients statistics for all of our corrected LIV variables, we won’t see any significant variables except for Malls\_300K\_0\_5R\_dum at an 88% confidence level.

Table 72: List of Potential Regressors

|  |
| --- |
| **Variable Name** |
| South |
| West |
| Free\_standing |
| Strip\_mall |
| Big\_box\_0\_5R |
| Malls\_300K\_0\_5R\_dum |

**IX. Model Building**

After identifying potential regressors and selecting which fit our regression model best based on our set parameters, we can now move on to the section of model building in our analysis. We can estimate and select the best fitting model that can more accurately explain sales at Timmy Tom’s by using our eight steps of model building. Our purpose is now transitioning into finding the best fitting regression model and performing model validation through the calculation of the Out-of-Sample Mean Absolute Percentage Error (OOS MAPE). Out of our forty-six potential regressors, we can narrow our list down in order to create distinct models that can get us closer to our best fitted model that allows us to explain sales at Timmy Tom’s.

Table 73: List of Potential Regressors

|  |
| --- |
| **Variable Name** |
| HH\_6person\_3R |
| HH\_5person\_3R |
| labor\_blue\_3R |
| Black\_HH\_3R |
| Black\_pop\_3R |
| Pop\_50\_69\_3R |
| HHinc\_25\_49K\_3R |
| Pop\_70\_85\_3R |
| HHinc\_LT\_25K\_3R |
| Pop\_GE\_18\_3R |
| Pop\_some\_college\_3R |
| Pop\_married\_3R |
| HH\_3person\_3R |
| Pop\_in\_school\_3R |
| Pop\_grades\_9\_12\_3R |
| Hispanic\_pop\_3R |
| Black\_pop\_5T |
| Brady\_Bunch\_5T |
| Black\_HH\_5T |
| labor\_farm\_3R |
| Pop\_40\_49\_3R |
| HHinc\_50\_74K\_3R |
| Hispanic\_HH\_3R |
| HH\_4person\_3R |
| HH\_6person\_5T |
| Pop\_21\_39\_3R |
| Tot\_HH\_Expend\_3R |
| Pop\_Associates\_3R |
| Cust\_value\_per\_cap\_region |
| Likely\_customers\_1R |
| HH\_2person\_3R |
| HH\_5person\_5T |
| HHinc\_LT\_25K\_5T |
| HH\_1person\_3R |
| retail\_3R |
| restaurants\_3R |
| per\_cap\_inc\_3R |
| labor\_blue\_5T |
| Big\_box\_0\_5R |
| Mid\_grocery\_1R\_dum |
| Malls\_300K\_0\_5R\_dum |
| South |
| West |

Demand theory and logical reasoning carries us into our next step for our analysis. Our demand theory indicates that income, population, preferences, prices of our own goods, prices of substitutes and complements will be correlated to our sales. Considering theory, we can assume that we are building a “demand” equation for our sales. Where our sales are related to how much demand there is for our services at Timmy Tom’s. We are building a model that helps us predict sales and the demand of customers for our services given significant regressors.

We have relevant income variables that we can incorporate into the model such as HHinc\_LT\_25K\_3R, HHinc\_25\_49K\_3R, and HHinc\_50\_74K\_3R since we want to analyze what levels of income are more likely to visit our stores. Higher levels of income indicate a higher disposable income that can be spent at Timmy Tom’s. We can also include Pop\_Associates\_3R and Pop\_Bachelors\_3R to determine our sales based on highest educational attainment. We will also incorporate per\_cap\_inc\_3R which is the per capita income (in dollars) of people living with 3 radial miles of a given store on the foundation that per capita income indicates levels of income within a certain area.

We can also include Pop\_married\_3R as a regressor since the percentage of people that are married will have an influx on sales if the population is more likely to be married since they could have a higher combined income. Pop\_GE\_18\_3R refers to the number of people who are 18 years old or older, who live within 3 radial miles of a given store. We can attest that those who are 18 and older are more likely to have a source of income by being active participants in the labor force. We also must add regional variables like south and west that indicate the location of a Timmy Tom’s in the United States. We will use central\_east as a base group. This in turn tells us where our business is more popular in the country and how sales are affected from it.

We also have to consider that our customers have preferences and that they will impact how they perceive our stores. We can include the variable Cust\_value\_per\_cap\_region which is 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. This variable can indicate what the likelihood of customers in the area is to visit one of our stores. Pop\_grades\_9\_12\_3R refers to the number of people who live within 3 radial miles of a given store who are in grades 9 through 12. This variable is relevant since we can look at the preferences of high schoolers to visit our store since it is affordable and provides a good variety of options for that specific population.

Considering logic, we must look at our competition nearby in terms of what other options are available for customers to eat such as restaurants\_3R. We must also look at retail\_3R as other retail establishments surrounding Timmy Tom’s can impact sales depending on the type of retail store. We can also include the combined variable restaurants\_retail\_3R to analyze as a whole how the number of restaurants and retail establishments of all types located within 3 radial miles of a Timmy Tom’s affect sales.

We could look at the variables Big\_box\_0\_5R and Big\_box\_1R as potential positive drivers to our sales. These variables indicate the number of big-box stores (such as Best Buy, Target, etc.) that are located within a one-half or one radial mile of a given store. If our customers are going on a shopping spree to a big box store, they might be hungry in between the shopping session and decide to have lunch or dinner at one of our locations. We can follow the same logic with Malls\_300k\_0\_5R\_dum.

Table 74: Final List of Potential Regressors

|  |
| --- |
| **Variable Name** |
| HHinc\_LT\_25K\_3R |
| HHinc\_25\_49K\_3R |
| HHinc\_50\_74K\_3R |
| Pop\_married\_3R |
| Pop\_Associates\_3R |
| Pop\_Bachelors\_3R |
| Pop\_GE\_18\_3R |
| Pop\_grades\_9\_12\_3R |
| South |
| West |
| restaurants\_3R |
| restaurants\_retail\_3R |
| retail\_3R |
| cust\_value\_per\_cap\_region |
| per\_cap\_inc\_3R |
| Big\_box\_index\_1R |
| Big\_box\_0\_5R |
| Malls\_300K\_0\_5R\_dum |

**.**

**X. Fit Statistics for the Selected Models**

With our final list of regressors, we proceed to create 20 different versions of regression models with subsets of our chosen variables. These models have at least 8 regressors and no more than 15. From all of the models that that were estimated, we can now choose the top five "contender models" and display their fit statistics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 646578 | 38892 | 16.63 | <.0001 |
| Likely\_customers\_1R | 1 | -19.69767 | 11.47539 | -1.72 | 0.0871 |
| Pop\_GE\_18\_3R | 1 | 1.18645 | 0.71447 | 1.66 | 0.0979 |
| Pop\_Associates\_3R | 1 | -9.94348 | 10.3285 | -0.96 | 0.3365 |
| Big\_box\_index\_1R | 1 | 1754.40639 | 2385.6002 | 0.74 | 0.4627 |
| Malls\_300K\_0\_5R\_dum | 1 | 7307.61258 | 30695 | 0.24 | 0.812 |
| per\_cap\_inc\_3R | 1 | 2.97948 | 1.04543 | 2.85 | 0.0047 |
| south | 1 | -276969 | 29942 | -9.25 | <.0001 |
| west | 1 | 470024 | 24542 | 19.15 | <.0001 |
| Mid\_grocery\_1R\_dum | 1 | -26464 | 22027 | -1.2 | 0.2306 |

. Table 75: Model P Summary Statistics

Table 76: Model Q Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 647666 | 38838 | 16.68 | <.0001 |
| Likely\_customers\_1R | 1 | -20.10207 | 11.34955 | -1.77 | 0.0776 |
| Pop\_GE\_18\_3R | 1 | 2.12046 | 1.22319 | 1.73 | 0.0841 |
| Pop\_grades\_9\_12\_3R | 1 | -17.66553 | 11.26943 | -1.57 | 0.1181 |
| restaurants\_retail\_3R | 1 | -44.74451 | 64.76304 | -0.69 | 0.4902 |
| Big\_box\_0\_5R | 1 | 2392.23921 | 2587.92411 | 0.92 | 0.3561 |
| per\_cap\_inc\_3R | 1 | 3.21366 | 1.05598 | 3.04 | 0.0026 |
| south | 1 | -277682 | 29835 | -9.31 | <.0001 |
| west | 1 | 469512 | 24557 | 19.12 | <.0001 |
| Mid\_grocery\_1R\_dum | 1 | -27445 | 22027 | -1.25 | 0.2138 |

Table 77: Model R Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 650928 | 38585 | 16.87 | <.0001 |
| Likely\_customers\_1R | 1 | -18.91138 | 11.41893 | -1.66 | 0.0988 |
| Pop\_GE\_18\_3R | 1 | 1.36273 | 0.71072 | 1.92 | 0.0562 |
| Pop\_grades\_9\_12\_3R | 1 | -13.16816 | 9.50439 | -1.39 | 0.167 |
| Malls\_300K\_0\_5R\_dum | 1 | 21374 | 24712 | 0.86 | 0.3878 |
| per\_cap\_inc\_3R | 1 | 3.08125 | 1.03444 | 2.98 | 0.0031 |
| south | 1 | -277522 | 29845 | -9.3 | <.0001 |
| west | 1 | 469632 | 24478 | 19.19 | <.0001 |
| Mid\_grocery\_1R\_dum | 1 | -25661 | 21843 | -1.17 | 0.241 |

Table 78: Model S Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 632843 | 39245 | 16.13 | <.0001 |
| Likely\_customers\_1R | 1 | -16.60747 | 11.71385 | -1.42 | 0.1573 |
| Pop\_married\_3R | 1 | -3.73639 | 2.4426 | -1.53 | 0.1272 |
| Pop\_GE\_18\_3R | 1 | 2.41738 | 1.39759 | 1.73 | 0.0848 |
| Big\_box\_0\_5R | 1 | 2281.07479 | 2580.00002 | 0.88 | 0.3774 |
| restaurants\_retail\_3R | 1 | -16.88537 | 57.49364 | -0.29 | 0.7692 |
| per\_cap\_inc\_3R | 1 | 3.71677 | 1.12759 | 3.3 | 0.0011 |
| south | 1 | -275543 | 29835 | -9.24 | <.0001 |
| west | 1 | 469280 | 24566 | 19.1 | <.0001 |
| Mid\_grocery\_1R\_dum | 1 | -27818 | 22040 | -1.26 | 0.2079 |

Table 79: Model T Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 604405 | 43472 | 13.9 | <.0001 |
| Likely\_customers\_1R | 1 | -19.66765 | 11.77267 | -1.67 | 0.0959 |
| Pop\_married\_3R | 1 | -4.88585 | 2.56287 | -1.91 | 0.0576 |
| Pop\_GE\_18\_3R | 1 | 3.46893 | 1.55685 | 2.23 | 0.0266 |
| restaurants\_3R | 1 | -597.42643 | 356.35719 | -1.68 | 0.0947 |
| retail\_3R | 1 | 190.64065 | 137.61975 | 1.39 | 0.167 |
| per\_cap\_inc\_3R | 1 | 4.7117 | 1.255 | 3.75 | 0.0002 |
| south | 1 | -276253 | 29712 | -9.3 | <.0001 |
| west | 1 | 473732 | 24515 | 19.32 | <.0001 |
| Mid\_grocery\_1R\_dum | 1 | -24397 | 21761 | -1.12 | 0.2632 |

Table 80: Values of Fit Statistics for the Selected Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Calc. General F-Test (p-value)\* | Value of R^2 | Value of Adj R^2 | # of significant \*\* slopes with correct sign | # of significant slopes with incorrect sign | Value of OOSMAPE |
| P | 95.03 (<0.0001) | 0.7481 | 0.7402 | 4 of 9 | 1 of 9 | 0.7619 |
| Q | 95.48 (<0.0001) | 0.7490 | 0.7411 | 4 of 9 | 2 of 9 | 0.2399 |
| R | 107.53 (<0.0001) | 0.7485 | 0.7416 | 4 of 9 | 1 of 9 | 0.2368 |
| S | 95.43 (<0.0001) | 0.7489 | 0.7410 | 4 of 9 | 0 of 9 | 0.2409 |
| T | 96.29 (<0.0001) | 0.7506 | 0.7428 | 6 of 9 | 1 of 9 | 0.2465 |

\*Where the General F-test has the null hypothesis (H0) that states that the variables in that model together as a group do not explain the dependent variable. For our five selected models the p-value is less than 0.12, which means we fail to reject the null hypothesis. That indicates that the variables as a group cannot explain sales.

\*\* At a 90% confidence level

We expected Likely\_customers\_1R to have a positive coefficient but in all of these models we see that it is in fact negative, which indicates that this variable drives sales down. This might be due to the fact that if customers are likely to come to our store, they are as likely to go to a competitor instead. Even so, we still should not see this negative coefficient in any of our models since it is an index of customers that are likely to come to our stores, and the higher the index the higher our sales should be. We also expected Pop\_married\_3R to have a positive coefficient but in all of these models we see that it is in fact negative, which indicates that this variable drives sales down. But we can reevaluate our theory and say that a higher married population indicates lower sales since married couples are less likely to eat out since they might have higher expenditures due to having children or higher expenses that in turn translate to lower disposable income. This new theory would then justify Pop\_married\_3R having a negative coefficient.

Table 81: Regressors in Selected Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model P | Model Q | Model R | Model S | Model T |
| Likely\_customers\_1R | Likely\_customers\_1R | Likely\_customers\_1R | Likely\_customers\_1R | Likely\_customers\_1R |
| Pop\_GE\_18\_3R | Pop\_GE\_18\_3R | Pop\_GE\_18\_3R | Pop\_GE\_18\_3R | Pop\_GE\_18\_3R |
| Pop\_Associates\_3R | Pop\_grades\_9\_12\_3R | Pop\_grades\_9\_12\_3R | Pop\_married\_3R | Pop\_married\_3R |
| Big\_box\_index\_1R | Big\_box\_0\_5R | Cust\_value\_per\_cap\_region | Big\_box\_0\_5R | restaurants\_3R |
| Malls\_300k\_0\_5R\_dum | restaurants\_retail\_3R | Malls\_300K\_0\_5R\_dum | restaurants\_retail\_3R | retail\_3R |
| per\_cap\_inc\_3R | per\_cap\_inc\_3R | per\_cap\_inc\_3R | per\_cap\_inc\_3R | per\_cap\_inc\_3R |
| south | south | south | south | south |
| west | west | west | west | west |
| Mid\_grocery\_1R\_dum | Mid\_grocery\_1R\_dum | Mid\_grocery\_1R\_dum | Mid\_grocery\_1R\_dum | Mid\_grocery\_1R\_dum |

The table below summarizes and assesses the fit statistics of the top five contender models.

Table 82: Assessment of Fit Statistics for Selected Models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | General F-Test | R^2 | Adj R^2 | % of significant slopes with correct sign | Any significant slopes with incorrect sign? | Value of OOS MAPE | Overall Assessment |
| P | significant | strong | strong | 44% | yes | bad | bad |
| Q | significant | strong | strong | 44% | yes | good | good |
| R | significant | strong | strong | 44% | yes | good | good |
| S | significant | strong | strong | 44% | yes | good | good |
| T | significant | strong | strong | 66.66% | yes | good | very good |

**XI. Multicollinearity**

Text

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

| r | = 1 → perfect multicollinearity

0.9 ≤ | r | < 1 → severe multicollinearity

0.5 ≤ | r | < 0.9 → moderate to strong multicollinearity

0 < | r | < 0.5 → weak multicollinearity

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

Table 83: Variable Pairs that have Moderate to Severe Multicollinearity in Model P

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| Likely\_customers\_1R and Pop\_Associates\_3R | 0.67342 |
| Likely\_customers\_1R and Pop\_GE\_18\_3R | 0.76751 |
| Pop\_GE\_18\_3R and Pop\_Associates\_3R | 0.84254 |
| Big\_box\_index\_1R and Malls\_300K\_0\_5R\_dum | 0.59896 |

.

Model P displays severe multicollinearity in many variables. Here we see that Likely\_customers\_1R and Pop\_Associates\_3R display multicollinearity, which could indicate that these two variables are correlated since it is very likely that those who hold at least an Associate’s degree have a higher income level and therefore likely customers. Pop\_GE\_18\_3R having severe multicollinearity with Likely\_customers\_1R indicates that that population of age greater than 18 is part of our likely customer group. Pop\_Associates\_3R and Pop\_GE\_18\_3R being correlated makes sense since most individuals that hold Associate degrees are of legal age or above. Big\_box\_index\_1R and Malls\_300K\_0\_5R\_dum being correlated indicate that it is likely that a big-box store is close to a big mall.

Table 84: Variable Pairs that have Moderate to Severe Multicollinearity in Model Q

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| Likely\_customers\_1R and Pop\_grades\_9\_12\_3R | 0.67238 |
| Likely\_customers\_1R and Pop\_GE\_18\_3R | 0.76751 |
| Likely\_customers\_1R and restaurants\_retail\_3R | 0.65986 |
| Pop\_GE\_18\_3R and Pop\_grades\_9\_12\_3R | 0.84326 |
| Pop\_GE\_18\_3R and restaurants\_retail\_3R | 0.86153 |
| Pop\_grades\_9\_12\_3R and restaurants\_retail\_3R | 0.58345 |

.

Model Q displays severe multicollinearity in many variables. Here we see that Likely\_customers\_1R and Pop\_grades\_9\_12\_3R display multicollinearity, which could indicate that these two variables are correlated since it is very likely that those in high school are likely customers. Pop\_GE\_18\_3R having severe multicollinearity with Likely\_customers\_1R indicates that that population of age greater than 18 is part of our likely customer group. Likely\_customers\_1R and restaurants\_retail\_3R being correlated indicates that those that are our likely customers are also likely to be likely customers for other restaurant and retail establishments. The same logic follows for Pop\_GE\_18\_3R and Pop\_grades\_9\_12\_3R displaying multicollinearity with restaurants\_retail\_3R since those populations are also potential customers for other restaurant and retail establishments. The high correlation between Pop\_GE\_18\_3R and Pop\_grades\_9\_12\_3R is a bit problematic since we know that usually those in high school are between the ages of 14 to 18. This would indicate that a very high number of high school students in the data are 18 or older, which could indicate that there are more last year high school students in the data.

Table 85: Variable Pairs that have Moderate to Severe Multicollinearity in Model R

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| Likely\_customers\_1R and Pop\_grades\_9\_12\_3R | 0.67238 |
| Likely\_customers\_1R and Pop\_GE\_18\_3R | 0.76751 |
| Pop\_GE\_18\_3R and Pop\_grades\_9\_12\_3R | 0.84326 |

.

In Model R, we again see “Likely\_customers\_1R and Pop\_grades\_9\_12\_3R”, “Likely\_customers\_1R and Pop\_GE\_18\_3R”, and “Pop\_GE\_18\_3R and Pop\_grades\_9\_12\_3R”.

Table 86: Variable Pairs that have Moderate to Severe Multicollinearity in Model S

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| Likely\_customers\_1R and Pop\_married\_3R | 0.79719 |
| Likely\_customers\_1R and Pop\_GE\_18\_3R | 0.76751 |
| Likely\_customers\_1R and restaurants\_retail\_3R | 0.65986 |
| Pop\_married\_3R and Pop\_GE\_18\_3R | 0.94514 |
| Pop\_married\_3R and restaurants\_retail\_3R | 0.77998 |
| Pop\_GE\_18\_3R and restaurants\_retail\_3R | 0.86153 |

.

Model S displays severe multicollinearity in many variables. Here we see that Likely\_customers\_1R and Pop\_married\_3R display multicollinearity, which could indicate that these two variables are correlated since it is very likely that those who are married have a higher income level and therefore likely customers. We again see “Likely\_customers\_1R and Pop\_GE\_18\_3R” and “Likely\_customers\_1R and restaurants\_retail\_3R” again. Pop\_married\_3R and Pop\_GE\_18\_3R being correlated makes perfect sense since most of the married individuals are of legal age or above. Likely\_customers\_1R also has severe multicollinearity with restaurants\_retail\_3R which indicates that our likely customers are likely to attend our stores given there are restaurants and retail stores located within 3 radial miles of a Timmy Tom’s. Pop\_married\_3R being correlated to restaurants\_retail\_3R is a telling sign that those that are married are likely to attend our stores given there are restaurants and retail stores located within 3 radial miles of a Timmy Tom’s. The same logic follows for Pop\_GE\_18\_3R with restaurants\_retail\_3R since those individuals older than 18 are likely to attend our stores if restaurants and retail stores are present in the area.

Table 87: Variable Pairs that have Moderate to Severe Multicollinearity in Model T

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| Likely\_customers\_1R and Pop\_married\_3R | 0.79719 |
| Likely\_customers\_1R and Pop\_GE\_18\_3R | 0.76751 |
| Likely\_customers\_1R and restaurants\_3R | 0.65266 |
| Likely\_customers\_1R and retail\_3R | 0.65316 |
| Pop\_married\_3R and Pop\_GE\_18\_3R | 0.94514 |
| Pop\_married\_3R and restaurants\_3R | 0.77780 |
| Pop\_married\_3R and retail\_3R | 0.77919 |
| Pop\_GE\_18\_3R and restaurants\_3R | 0.86788 |
| Pop\_GE\_18\_3R and retail\_3R | 0.84553 |

.

Model T displays severe multicollinearity in many variables. Here we see that Likely\_customers\_1R and Pop\_married\_3R display multicollinearity, which could indicate that these two variables are correlated since it is very likely that those who are married have a higher income level and therefore likely customers. We actually see that Likely\_customers\_1R is highly correlated with most of the variables in our model. Pop\_GE\_18\_3R having severe multicollinearity with Likely\_customers\_1R indicates that that population of age greater than 18 is part of our likely customer group. Likely\_customers\_1R also has severe multicollinearity with restaurants\_3R and retail\_3R, which indicates that our likely customers are likely to attend our stores given there are restaurants and retail stores located within 3 radial miles of a Timmy Tom’s. Pop\_married\_3R and Pop\_GE\_18\_3R being correlated makes perfect sense since most of the married individuals are of legal age or above. Pop\_married\_3R being correlated to restaurants\_3R and retail\_3R is a telling sign that those that are married are likely to attend our stores given there are restaurants and retail stores located within 3 radial miles of a Timmy Tom’s. The same logic follows for Pop\_GE\_18\_3R with restaurants\_3R and retail\_3R since those individuals older than 18 are likely to attend our stores if restaurants and retail stores are present in the area.

The presence of multicollinearity in these models could be one of the factors contributing to the slopes with incorrect signs. The presence of multicollinearity in many of these models is tied to the variable Likely\_customers\_1R, which is one of the major problems in our models since it has a negative slope coefficient.

**XII. Chosen Model**

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

This leaves us with the selection between two models:

Table 88: Model R vs Model T

|  |  |  |
| --- | --- | --- |
| Model | R | T |
| General F-test | Significant | Significant |
| R^2 | Strong | Strong |
| Adjusted R^2 | Strong | Strong |
| % of significant slopes with correct sign | 44% | 66% |
| Any significant slopes with incorrect sign? | Yes | Yes |
| Value of OOS MAPE | Good | Good |
| Multicollinearity presence | Low  (3 occurrences) | Very high  (9 occurrences) |
| Overall Assessment | Very good | Good |

.

Model R is the best out of our contender models. Its OOS MAPE value is 23.68% and there are only 3 occurrences of multicollinearity, it’s high R^ 2 and high percentage of significant slopes with correct signs indicates that we have many positive indicators that lead us to select this model. This now allows us to write our estimated form of our equation for Model R with their respective p-values underneath each parameter estimate.

= 650,928 –18.91Likely\_customers\_1Ri + 1.36Pop\_GE\_18\_3Ri

(<.0001) (0.0988) (0.0562)

–13.17Pop\_grades\_9\_12\_3Ri +21,374 Malls\_300K\_0\_5R\_dumi

(0.167) (0.378)

+3.08per\_cap\_inc\_3Ri –277,522southi + 469,632westi )–25,661Mid\_grocery\_1R\_dumi

(0.0031) (<.0001) (<.0001) (0.241)

Finally, we can evaluate the "attractiveness" of our estimated coefficients. We want to scale our regressors in order to have a smoother interpretation of the dollar figure values. Starting from the top, we consider Likely\_customer\_1R as an attractive regressors since it helps us estimate how our sales perform under the likelihood that someone could be a likely customer. This negative coefficient, however, tells us that we must be careful on how our likely customers are perceived since it can have a negative impact on sales perhaps on the foundation that our likely customers might also approach our competitors or are our competitors fan base. This variable is not attractive since it is hard to justify the negative coefficient on sales, so we could possibly toss it out in our final modeling. Pop\_GE\_18\_3R is a very attractive regressor since it illustrates the point of how most of our customers are 18 or older. This could indicate that our customer base is those that most likely completed high school, are in the labor force, and have more disposable income. It has a positive coefficient which indicates we want to ensure our future placements in areas were people of the age of 18 and above are located. We see the variable per\_cap\_inc\_3R having a positive impact, which indicates that the higher the per capita income is within a three-mile radius, the higher our sales.

Pop\_grades\_9\_12\_3R has a negative coefficient, which indicates that if there is a high school population near our stores, sales tend to go down. This might be because these high school students do not have a lot of disposable income to spare, they eat at their school cafeteria, or it might drive those that are not in high school far from the high school areas to avoid teenagers during lunch rush hours. Malls\_300K\_0\_5R\_dum has a positive high impact on sales. This variable indicates that if there is a big mall very close by, our sales will increase. This could be due to the fact that individuals might get hungry during or after a big shopping spree. We also see that if we place our new locations in the West we will have a positive impact on sales. And if in the South we will have a decrease in sales. This indicates that we should steer away from placing any new locations in the Southern regions and instead opting for example on placing new locations in the West. We must also steer away from locations that have a mid-level grocery store within one mile as this has a significant negative impact on sales.

From this analysis, we can determine that it is best to put new locations nearby big malls, far from high schools, in areas with higher populations of adults (over the age of 18), in areas with higher per capita income, in the Western region of the United States, and far from mid-level grocery stores.

**XV. Interpretations of the estimated coefficients:**

• Likely\_customers\_1R: estimated coefficient = –18.91

For each additional individual that lives within one radial mile of a store and is likely to be a customer, sales at Buster’s are expected to decrease by about 18.91 dollars, ceteris paribus.

• Pop\_GE\_18\_3R: estimated coefficient = 1.36

For every additional person who is 18 years old or older and is within 3 radial miles of a given store, sales at Timmy Tom’s are expected to increase by about 1.36 dollars, ceteris paribus.

• Pop\_grades\_9\_12\_3R : estimated coefficient = –13.17

For every additional person who is in high school and is within 3 radial miles of a given store, sales at Timmy Tom’s are expected to decrease by about 13.17 dollars, ceteris paribus.

• Malls\_300K\_0\_5R\_dum: estimated coefficient = 21,374

If there is a 300+ square foot mall within one-half radial mile, the average sales at Timmy Tom’s are expected to be higher by 21,374 dollars, ceteris paribus.

• Per\_cap\_inc\_3R: estimated coefficient = 3.08

For every additional dollar increase of the per capita income within three radial miles, sales at Timmy Tom’s are expected to increase by about 3.08 dollars, ceteris paribus.

•South: estimated coefficient = -277,522

If a store is located in the South, the average sales at Timmy Tom’s are expected to be lower by about $277,522, ceteris paribus.

•West: estimated coefficient = 469,632

If a store is located in the West, the average sales at Timmy Tom’s are expected to be higher by about $469,632, ceteris paribus.

•Mid\_grocery\_1R\_dum: estimated coefficient = –25,661

If there is a mid-end grocery store present within one radial mile, the average sales at Timmy Tom’s are expected to be lower by about $25,661, ceteris paribus.

**XVI. Model Building through Sequential Regression**

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

Table 89: Temporarily Removed Multi-Characteristic Dummy Variables

|  |
| --- |
| **Variable Name** |
| South |
| West |

..

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

Table 90: Potential Regressors (Omission of Multi-Characteristic Dummy Variable)

|  |
| --- |
| **Variable Name** |
| HHinc\_LT\_25K\_3R |
| HHinc\_25\_49K\_3R |
| HHinc\_50\_74K\_3R |
| Pop\_married\_3R |
| Pop\_Associates\_3R |
| Pop\_Bachelors\_3R |
| Pop\_GE\_18\_3R |
| Pop\_grades\_9\_12\_3R |
| restaurants\_3R |
| restaurants\_retail\_3R |
| retail\_3R |
| cust\_value\_per\_cap\_region |
| per\_cap\_inc\_3R |
| Big\_box\_index\_1R |
| Big\_box\_0\_5R |
| Malls\_300K\_0\_5R\_dum |

Although the best approach to create models is through our previous steps of model building, there is a distinct method in which the computer estimates every possible model given the current collection of potential regressors. From there, the computer selects the best model based on a given statistical criterion. Sequential regression uses algorithm-based decisions. It builds a regression model using the adjusted R^2 selection model by producing the best one-variable model, then the best two-variable model, and so on up to the best (K-1) variable model in which K is the number of potential regressors, and finally a model that contains all potential regressors. From all these models estimated, the “best” model is the one with the largest adjusted R2. There are five different methods for sequential regression: forward-selection, stepwise selection, backward-elimination selection, "Maximum R^2 Improvement" selection, and the "Adjusted R^2" selection method.

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

**C(p) Statistic**

Text

Description automatically generatedSince we want to choose the model that provides the best prediction using the sample estimates, we must be aware of the possibility of estimating more parameters than can be reliably estimated with the given sample size by using a moderate significance level in the range of 10 percent to 25 percent (Daniel & Wood, 1980). This is where the C(p) statistic comes into play. In our model selection methods, a C(p) statistic is displayed for each model. The C(p) statistic was proposed by C.L. Mallows in 1973—hence why it is also known as the Mallows’ *Cp* statistic—and serves as a criterion for selecting a model. It is a measure of total squared error defined as:

where is the MSE for the full model, and SSEp is the sum-of-squares error for a model with p parameters including the intercept, if any. If C(p) is plotted against p, we select the model where C(p) first approaches p (Mallows, 1973). When the right model is chosen, the parameter estimates are unbiased, and this is reflected in C(p) near p.

**Partial R-squared**

If we set up a general linear *F*-test, we can determine what percent of the variation in the response cannot be explained by the predictors in the reduced model but can be explained by the rest of the predictors in the full model. If the result is a large percentage, then it is likely we would want to specify some or all of the remaining predictors to be in the final model since they explain a lot of the variation in the model. This percentage is called partial R-square or the coefficient of partial determination. It refers to how much of the Corrected Total Sums of Squares can be attributed to the Sums of Squares for this particular effect. In some of our sequential regression methods, we partially add variables after each step.

**Type II Error**

A Type II error is a term that is used within the context of hypothesis testing. It describes the error that occurs when one accepts a null hypothesis that is actually false. A type II error produces a false negative, also known as an error of omission (Hayes, 2021). It is also known as a false negative since the researcher wrongly concludes that there is not a statistically significant effect when in reality there is.

**XVII. Forward Selection Sequential Regression Method**

When using the "forward-selection" method, we must set an entry tolerance level, where our tolerance level SLENTRY is referring to our tolerance to a Type I error. This method first estimates all possible regression models that contain a single regressor. It then conducts an F test of H0: βj = 0 for the slope variable for every single model. After, it discards any model for which the p-value for that F test is greater than the tolerance level. If there are any models remaining the one with the smallest p-value is selected and named “Model 1”. The method then estimates all possible regression models that contain two regressors, follows the same steps as above, and then gives us a “Model 2” if any models remain. This process continues until we end up adding the total number of regressors. In our case we have 16 regressors once we exclude multi-characteristic dummy variables and we want to set the tolerance level to 0.20. Therefore, we will be using a value of 0.20 for SLENTRY.

Table 91: Forward Selection

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Forward Selection** | | | | | | | |
| **Step** | **Variable Entered** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | HHinc\_25\_49K\_3R | 1 | 0.0364 | 0.0364 | 8.8247 | 11.17 | 0.0009 |
| **2** | per\_cap\_inc\_3R | 2 | 0.0110 | 0.0474 | 7.3746 | 3.40 | 0.0662 |
| **3** | Pop\_married\_3R | 3 | 0.0105 | 0.0578 | 6.0802 | 3.27 | 0.0715 |
| **4** | Cust\_value\_per\_cap\_region | 4 | 0.0172 | 0.0750 | 2.6837 | 5.44 | 0.0204 |
| **5** | Big\_box\_index\_1R | 5 | 0.0114 | 0.0864 | 1.1038 | 3.64 | 0.0574 |
| **6** | HHinc\_50\_74K\_3R | 6 | 0.0084 | 0.0948 | 0.4489 | 2.72 | 0.1004 |

Going back to the topic of partial R-square, in forward selection we get a specific column for partial R-Square that refers to how much of the Corrected Total Sums of Squares can be attributed to the Sums of Squares for each variable added in each step. According to forward selection, the best model This now points us to the best model according to forward selection that has an R-Square of 0.0948. Using forward selection’s given regressors, we will create this model and name it Model U.

Table 92: Analysis of Variance for Best Model Using Forward Selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 3.647287E12 | 6.078812E11 | 5.08 | <.0001 |
| **Error** | 291 | 3.480614E13 | 1.196087E11 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

Table 93: Best Model Using Forward Selection Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| Intercept | 502,382 | 147,997 | 1.378233E12 | 11.52 | 0.0008 |
| HHinc\_25\_49K\_3R | -4.45 | 14.95 | 10599870067 | 0.09 | 0.7661 |
| HHinc\_50\_74K\_3R | 46.48 | 28.20 | 3.248541E11 | 2.72 | 0.1004 |
| Pop\_married\_3R | -12.19 | 3.87 | 1.184857E12 | 9.91 | 0.0018 |
| Cust\_value\_per\_cap\_region | 3,260.56 | 1,295.32 | 7.578675E11 | 6.34 | 0.0124 |
| per\_cap\_inc\_3R | 6.34 | 2.28 | 9.768503E11 | 8.17 | 0.0046 |
| Big\_box\_index\_1R | 6,604.22 | 3,606.81 | 4.010148E11 | 3.35 | 0.0681 |

.

After the computer selected this as the “best” model, we must consider if we want to add any multi-characteristic dummy variables into the final model. In SAS, we created Model U that has the six regressors that were selected via forward selection.

Table 94: Model U – Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 3.65E+12 | 6.08E+11 | 5.08 | <.0001 |
| **Error** | 291 | 3.48E+13 | 1.20E+11 |  |  |
| **Corrected Total** | 297 | 3.85E+13 |  |  |  |

Table 95: Model U – Analysis of Variance (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 345845 | **R-Square** | 0.0948 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.0762 |
| **Coeff Var** | 39.65945 |  |  |

Table 96: Model U Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 502382 | 147997 | 3.39 | 0.0008 |
| HHinc\_25\_49K\_3R | 1 | -4.44934 | 14.94607 | -0.3 | 0.7661 |
| HHinc\_50\_74K\_3R | 1 | 46.48207 | 28.20478 | 1.65 | 0.1004 |
| Pop\_married\_3R | 1 | -12.19286 | 3.87395 | -3.15 | 0.0018 |
| Cust\_value\_per\_cap\_region | 1 | 3260.55903 | 1295.31942 | 2.52 | 0.0124 |
| per\_cap\_inc\_3R | 1 | 6.33585 | 2.21703 | 2.86 | 0.0046 |
| Big\_box\_index\_1R | 1 | 6604.21938 | 3606.80567 | 1.83 | 0.0681 |

We can then create a Model V that reincorporates the multi-characteristic dummy variables. When creating models that reincorporate the multi-characteristic dummy variables (South and West) we see that the new Model V brings a significant higher R-Square when adding them back in. Model U has an R-Square of 0.0948 versus Model V that has an R-Square of 0.7459. It must be noted that three additional variables were insignificant in Model V compared to Model R.

Table 97: Model V – Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 8 | 2.86811E13 | 3.585137E12 | 106.02 | <.0001 |
| **Error** | 289 | 9.77233E12 | 33814292096 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

Table 98: Model V – Analysis of Variance (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 183887 | **R-Square** | 0.7459 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.7388 |
| **Coeff Var** | 21.08703 |  |  |

.

Table 99: Model V Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Intercept | 1 | 566,663 | 80,420 | 7.05 | <.0001 |
| HHinc\_25\_49K\_3R | 1 | 6.54 | 7.96 | 0.82 | 0.4116 |
| HHinc\_50\_74K\_3R | 1 | 8.23 | 15.07 | 0.55 | 0.5853 |
| Pop\_married\_3R | 1 | -3.69 | 2.084 | -1.77 | 0.0771 |
| Cust\_value\_per\_cap\_region | 1 | 622.78 | 697.58 | 0.89 | 0.3727 |
| per\_cap\_inc\_3R | 1 | 3.64 | 1.18 | 3.08 | 0.0023 |
| Big\_box\_index\_1R | 1 | 2,054.79 | 1,925.25 | 1.07 | 0.2867 |
| south | 1 | -281,332 | 29,951 | -9.39 | <.0001 |
| west | 1 | 463,894 | 24,603 | 18.86 | <.0001 |

.

Now we can compare this “best” forward selection Model V that reincorporates the multi characteristic dummy variables. Compared to our previous five candidate models, Model V has the lowest R-Square score of 0.7459. It has a 50% of significant slopes with the correct sign, which is higher than 4 of our contender models, with Model T having 66.6%. Our human selected Model R has a higher R-Square of 0.7485.

However, Model V has a lower Adjusted R-Square of 0.7388 while Model R has one of 0.7416.

Model V and Model R both have Cust\_value\_per\_cap\_region, per\_cap\_inc\_3R, south, and west as regressors. Model V incorporates HHinc\_25\_49K\_3R, HHinc\_50\_74K\_3R, Pop\_married\_3R, and Big\_box\_index\_1R. And Model R incorporates Likely\_customers\_1R, Pop\_GE\_18\_3R, Pop\_grades\_9\_12\_3R, Malls\_300K\_0\_5R\_dum, and Mid\_grocery\_1R\_dum. Given all this, we still decide to go with “human selected” Model R.

**XVIII. Stepwise Selection Sequential Regression Method**

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

Table 100: Stepwise Selection

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Stepwise Selection** | | | | | | | | |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars in** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | HHinc\_25\_49K\_3R |  | 1 | 0.0364 | 0.0364 | 8.8247 | 11.17 | 0.0009 |
| **2** | per\_cap\_inc\_3R |  | 2 | 0.0110 | 0.0474 | 7.3746 | 3.40 | 0.0662 |
| **3** | Pop\_married\_3R |  | 3 | 0.0105 | 0.0578 | 6.0802 | 3.27 | 0.0715 |
| **4** |  | HHinc\_25\_49K\_3R | 2 | 0.0001 | 0.0578 | 4.0996 | 0.02 | 0.8899 |
| **5** | Cust\_value\_per\_cap\_region |  | 3 | 0.0136 | 0.0714 | 1.8216 | 4.31 | 0.0388 |
| **6** | HHinc\_50\_74K\_3R |  | 4 | 0.0122 | 0.0836 | -0.0188 | 3.91 | 0.0490 |
| **7** | Malls\_300K\_0\_5R\_dum |  | 5 | 0.0111 | 0.0947 | -1.4991 | 3.57 | 0.0597 |

.

We actually end up getting the exact same model as with forward selection, Model U. Which means that the stepwise selection regression method produced the exact same model. Again, we end up choosing “human selected Model R”.

**XIX. Backward-Elimination Selection Sequential Regression Method**

The Backward-Elimination Selection is a stepwise regression approach that begins with a full model then at each step it gradually eliminates variables from the regression model to find a reduced model. When building a regression model using the backward-elimination selection method, regressors are selected in a way that is the reverse of the forward-selection method.

Table 101: Backward Selection

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Backward Elimination** | | | | | | | | |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** |  | restaurants\_retail\_3R | 14 | 0.0001 | 0.1026 | 14.0169 | 0.02 | 0.8965 |
| **2** | retail\_3R |  | 15 | 0.0001 | 0.1026 | 16.0000 | 0.02 | 0.8965 |
| **3** |  | retail\_3R | 14 | 0.0001 | 0.1026 | 14.0169 | 0.02 | 0.8965 |
| **4** |  | HHinc\_LT\_25K\_3R | 13 | 0.0002 | 0.1024 | 12.0861 | 0.07 | 0.7923 |
| **5** |  | Big\_box\_0\_5R | 12 | 0.0004 | 0.1020 | 10.2148 | 0.13 | 0.7192 |
| **6** |  | Pop\_Bachelors\_3R | 11 | 0.0004 | 0.1016 | 8.3432 | 0.13 | 0.7190 |
| **7** |  | Pop\_GE\_18\_3R | 10 | 0.0003 | 0.1013 | 6.4247 | 0.08 | 0.7741 |
| **8** |  | HHinc\_25\_49K\_3R | 9 | 0.0006 | 0.1007 | 4.6203 | 0.20 | 0.6561 |
| **9** |  | Pop\_Associates\_3R | 8 | 0.0009 | 0.0998 | 2.9084 | 0.29 | 0.5884 |
| **10** |  | restaurants\_3R | 7 | 0.0010 | 0.0987 | 1.2381 | 0.34 | 0.5621 |
| **11** |  | Pop\_grades\_9\_12\_3R | 6 | 0.0013 | 0.0974 | -0.3639 | 0.41 | 0.5238 |
| **12** |  | Big\_box\_index\_1R | 5 | 0.0028 | 0.0947 | -1.4991 | 0.89 | 0.3470 |

.

The backward elimination method actually ends up giving us a very different model. Just by sight we can tell that all of the P-values are insignificant. Therefore, we can omit the result for this model and move on to the next step.

**XX. Maximum R-Squared Improvement Sequential Regression Method**

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

Once we run our code in SAS, we find that in this case without the incorporation of multi trait dummy variables, our maximum R-Squared given model has a very low R-Square of 0.1026. It also has only three statistically significant regressors. We can refer to this as model Z.

Table 102: Model Z – Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 15 | 3.946949E12 | 2.631299E11 | 2.15 | 0.0082 |
| **Error** | 282 | 3.450648E13 | 1.223634E11 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 349197 | **R-Square** | 0.1026 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.0582 |
| **Coeff Var** | 40.04381 |  |  |

Table 103: Model Z – Analysis of Variance (cont.)

Table 104: Model Z - Maximum R-Squared Improvement Sequential Regression Method

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| Intercept | 447,457 | 175,729 | 7.933514E11 | 6.48 | 0.0114 |
| HHinc\_LT\_25K\_3R | -5.88 | 21.21 | 9420551963 | 0.08 | 0.7816 |
| HHinc\_25\_49K\_3R | -9.14 | 25.38 | 15873162898 | 0.13 | 0.7190 |
| HHinc\_50\_74K\_3R | 41.53 | 37.37 | 1.510762E11 | 1.23 | 0.2675 |
| Pop\_married\_3R | -17.99 | 8.88 | 5.025068E11 | 4.11 | 0.0437 |
| Pop\_Associates\_3R | 12.03 | 27.51 | 23394555292 | 0.19 | 0.6623 |
| Pop\_Bachelors\_3R | -4.11 | 9.75 | 21714181321 | 0.18 | 0.6739 |
| Pop\_GE\_18\_3R | 3.15 | 6.51 | 28628505156 | 0.23 | 0.6290 |

Table 104: Model Z - Maximum R-Squared Improvement Sequential Regression Method (cont.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| Pop\_grades\_9\_12\_3R | 12.42 | 33.74 | 16583121006 | 0.14 | 0.7130 |
| restaurants\_3R | 188.47 | 747.36 | 7781703767 | 0.06 | 0.8011 |
| retail\_3R | 41.28 | 317.21 | 2072679351 | 0.02 | 0.8965 |
| Cust\_value\_per\_cap\_region | 3541.15 | 1,335.99 | 8.596667E11 | 7.03 | 0.0085 |
| per\_cap\_inc\_3R | 7.11 | 3.15 | 6.245936E11 | 5.10 | 0.0246 |
| Big\_box\_index\_1R | 7834.33 | 11,888 | 53139212895 | 0.43 | 0.5104 |
| Big\_box\_0\_5R | -5138.10 | 15,406 | 13609848572 | 0.11 | 0.7390 |
| Malls\_300K\_0\_5R\_dum | 50,079 | 59,303 | 87258744859 | 0.71 | 0.3991 |

,

If we create a Model ZZ that incorporates the multi-trait dummy variables, we get the following.

Table 105: Model ZZ – Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 17 | 2.883153E13 | 1.695973E12 | 49.35 | <.0001 |
| **Error** | 280 | 9.621894E12 | 34363907856 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

Table 106: Model ZZ – Analysis of Variance

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 185375 | **R-Square** | 0.7498 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.7346 |
| **Coeff Var** | 21.25771 |  |  |

Table 107: Model ZZ – Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Intercept | 1 | 525,592 | 94,775 | 5.55 | <.0001 |
| HHinc\_LT\_25K\_3R | 1 | 5.73 | 11.29 | 0.51 | 0.6119 |
| HHinc\_25\_49K\_3R | 1 | -6.73 | 13.46 | -0.50 | 0.6177 |
| HHinc\_50\_74K\_3R | 1 | 16.55 | 19.85 | 0.83 | 0.4053 |
| Pop\_married\_3R | 1 | -5.06 | 4.78 | -1.06 | 0.2900 |
| Pop\_Associates\_3R | 1 | -5.26 | 14.61 | -0.36 | 0.7189 |
| Pop\_Bachelors\_3R | 1 | -0.99 | 5.17 | -0.19 | 0.8476 |
| Pop\_GE\_18\_3R | 1 | 2.23 | 3.46 | 0.64 | 0.5204 |
| Pop\_grades\_9\_12\_3R | 1 | -2.89 | 17.91 | -0.16 | 0.8717 |
| restaurants\_3R | 1 | -441.29 | 397.69 | -1.11 | 0.2681 |
| retail\_3R | 1 | 106.99 | 168.37 | 0.64 | 0.5257 |

Table 107: Model ZZ – Summary Statistics (cont.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Cust\_value\_per\_cap\_region | 1 | 716.23158 | 719.004 | 1.00 | 0.3200 |
| per\_cap\_inc\_3R | 1 | 4.49 | 1.67 | 2.69 | 0.0077 |
| Big\_box\_index\_1R | 1 | 6,202.47 | 6304.90 | 0.98 | 0.3261 |
| Big\_box\_0\_5R | 1 | -6,592.81 | 8175.98 | -0.81 | 0.4207 |
| Malls\_300K\_0\_5R\_dum | 1 | 1,3026 | 31,487 | 0.41 | 0.6794 |
| south | 1 | -27,9905 | 30,746 | -9.10 | <.0001 |
| west | 1 | 467,492 | 24,994 | 18.70 | <.0001 |

Model ZZ has an R-Square of 0.7498 and Model R has an R-Square of 0.7485. Model ZZ, however, have less statistically significant regressors with the correct sign, 3 out of 17 which is 17.65 %. Compared to Model R, that has 44. Again, we continue with human selected Model R.

**XXI. Adjusted R-Squared Sequential Regression Method**

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

Table 108: Adjusted R-Squared Sequential Regression Method

|  |  |  |  |
| --- | --- | --- | --- |
| **Number in Model** | **Adjusted R-Square** | **R-Square** | **Variables in Model** |
| 5 | 0.0792 | 0.0947 | HHinc\_50\_74K\_3R Pop\_married\_3R Cust\_value\_per\_cap\_region per\_cap\_inc\_3R Malls\_300K\_0\_5R\_dum |

.

If we create a model with these variables and name it Model AA, we will get both a low adjusted R-Square and a low R-Square. With the given results in SAS, we create a Model AB that incorporates the same variables as Model AA along with the multi-trait dummy variables. Model AB then has the following.

Table 109: Model AB - Analysis of Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 7 | 2.865965E13 | 4.094235E12 | 121.23 | <.0001 |
| **Error** | 290 | 9.793781E12 | 33771659178 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

Table 110: Model AB - Analysis of Variance (cont.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 183771 | **R-Square** | 0.7453 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.7392 |
| **Coeff Var** | 21.07373 |  |  |

.

Table 111: Model AB – Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Intercept | 1 | 592,411 | 72,983 | 8.12 | <.0001 |
| HHinc\_50\_74K\_3R | 1 | 17.32 | 10.40 | 1.66 | 0.0970 |
| Pop\_married\_3R | 1 | -3.80 | 2.08 | -1.83 | 0.0686 |
| Cust\_value\_per\_cap\_region | 1 | 471.996 | 662.14 | 0.71 | 0.4765 |
| per\_cap\_inc\_3R | 1 | 3.34 | 1.09 | 3.07 | 0.0024 |
| Malls\_300K\_0\_5R\_dum | 1 | 24,782 | 24,657 | 1.01 | 0.3157 |
| south | 1 | -279,064 | 29,994 | -9.30 | <.0001 |
| west | 1 | 464,295 | 24,556 | 18.91 | <.0001 |

.

Model AB has an Adjusted R-Square of 0.7392, an R-Square of 0.7453, and five regressors out of seven being statistically significant which is 62.5%. Model R has a higher R-Square of 0.7485 and 44% of statistically significant regressors with the correct sign. Model AB and Model R have Malls\_300K\_0\_5R\_dum, per\_cap\_inc\_3R, south, and west as common regressors. If we check for multicollinearity in Model AB, it might become our optimal model if it has less multicollinearity than Model R.

Table 112: Variable Pairs that have Moderate to Severe Multicollinearity in Model AB

|  |  |
| --- | --- |
| **Variable Pair** | **Correlation Coefficient** |
| HHinc\_50\_74K\_3R and Pop\_married\_3R | 0.92430 |

There is only one instance of Multicollinearity in Model AB. which means that given its 62.5% statistically significant regressors with correct sign, its R-Square of 0.7453, and single instance of multicollinearity we will select machine generated Model AB that was generated via Adjusted R-Squared Sequential Regression Method with the incorporation of multi-trait dummy variables.

**XXII. Model Scoring**

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

Table 113: Sales Potential Definition

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

.

Using our corrected data set with the 298 observations, we got that our average sales were equal to $872,036.76 and our standard deviation being equal to $359,823.32.

Table 114: Sales Potential Figures

|  |  |
| --- | --- |
| **Sales Level** | **Value of Sales** |
| Average Sales: | $872,036.76 |
| Average + 1 standard deviation: | $1,231,860.08 |
| Average + 2 standard deviations: | $1,591,683.40 |

Table 115: Sales Potential and Threshold Values

|  |  |
| --- | --- |
| **Sales Potential** | **Threshold Value** |
| Below Average Sales Potential | Projected Sales < $872,036.76 |
| Above Average Sales Potential | $872,036.76< Projected Sales < $1,231,860.08 |
| Medium Sales Potential | $1,231,860.08< Projected Sales < $1,591,683.40 |
| High Sales Potential | Projected Sales > $1,591,683.40 |

.

We have five potential stores where we could place a location. The potential stores have a Storeid ranging from Store 312 to Store 321. Using Model AB, we calculated the projected sales for these ten potential locations and sorted them by highest to lowest projected sales figures.

Table 116: Potential Timmy’s Toms Locations and Sales Potential

|  |  |  |
| --- | --- | --- |
| **Storeid** | **Projected Sales** | **Sales Potential** |
| Store 320 | $1,181,496.02 | Above Average |
| Store 314 | $1,122,621.91 | Above Average |
| Store 321 | $747,458.31 | Below Average |
| Store 315 | $746,906.54 | Below Average |
| Store 316 | $745,717.33 | Below Average |
| Store 317 | $ 719,743.12 | Below Average |
| Store 313 | $660,224.36 | Below Average |
| Store 312 | $557,153.01 | Below Average |
| Store 319 | $484,838.28 | Below Average |
| Store 318 | $467,508.67 | Below Average |

.

We have now scored each location based on their sales potential using Model AB. Store 320 and Store 314 are the only two stores that have an above average sales potential. The other eight stores have a below average sales potential, and therefore should not be considered as potential future locations. Store 320 and Store 314 fall under the above average sales potential category and should therefore we can advise that these locations are candidates for our new locations. These above average figures indicate positive profits and will give us above average sales and increase our overall revenue. Store 320 has the highest projected sales figure of all the potential locations.

We can consider how the regressors will impact our sales figures. We consider those factors that are positively related and increase sales. The best locations for future Timmy Tom’s stores are those that possess positive factors that according to our Model AB will increase sales, we must also avoid those potential stores with the characteristics that drive sales down such as a high married population and placing potential locations in the South. Therefore, we want to consider potential locations in places that have a population with averages of $50,000 to $74,999 in income, that have a high measure of customer value per capita geographical region, a high number of per capita income of people living with 3 radial miles of a given store, that has one or more malls with more than 300,000 square feet of gross leasable area that is located within a one-half radial mile and is located in the Western region of the United States. Among those factors that will have negative impact on sales we have if the potential location is located in the Southern region of the United States and if the married population within three radial miles of the store is high.

Given this, we know that potential Store 320 would be impacted positively by the facts that there is a good amount of people that earn an average income between $50,000 and $74,999, has the lowest population married within three radial miles of all of our potential stores, has the highest measure of customer value per capita geographical region of all of our potential stores, has a good number of per capita income of people living with 3 radial miles of a given store, and it is located in the Western region of the United States. Although Store 320 has no mall with more than 300,000 square feet of gross leasable area within a one-half radial mile of the potential location, there might be an opportunity for a mall to be constructed in the future and drive its sales further up since Store 320 is located in a developing area.

Due to the Covid-19 pandemic’s impact, Timmy Tom’s has adopted a conservative approach and seeks to only consider potential locations that have projected “high revenue” figures. Given this approach, Timmy Tom’s will not seek to open any of these potential locations, even store 320.

In general, we draft the following recommendations for Timmy Tom’s future selection of potential locations. Current Timmy Tom’s locations that are located within three radial miles of having a population with average income of $50,000 to $74,999 have higher sales on average. So, we want to seek to expand in locations with this average level of income in the area. Current stores with higher married population have lower sales on average, so we want to strategically select new locations with lower married populations in the three-radial mile of a given store area. Stores with a higher measure of customer value per capita geographical region have higher sales on average. So, we must seek high measures of customer value per capita geographical region.

We observe that current stores that have a higher per capita income of people living with 3 radial miles of a given store have higher sales on average. Therefore, we must seek to place potential locations in areas that have higher per capita income. We can also observe that current stores that are within one-half radial mile of a mall that has over 300,000 square foot have higher sales on average. Then, we observe the phenomena of our current stores having higher sales figure in part due to this characteristic being present. Which leads us to believe that if we place future locations near to these types of malls, our sales will be higher on average. Stores located in the southern region have lower sales on average, so we want to steer clear of placing new locations in the South. Stores located in the western region have higher sales on average, so we want to consider placing new locations in the West.

**REFERENCES**

5.8 - PARTIAL R-SQUARED. (n.d.). Retrieved April 07, 2021, from https://online.stat.psu.edu/stat462/node/138/

Daniel, C. and Wood, F. (1980), Fitting Equations to Data, Revised Edition, New York: John Wiley & Sons, Inc.

Hayes, A. (2021, April 09). Type ii error definition. Retrieved April 14, 2021, from https://www.investopedia.com/terms/t/type-ii-error.asp

Hayes, A. (2020, September 13). Coefficient of variation (cv) definition. Retrieved April 15, 2021, from https://www.investopedia.com/terms/c/coefficientofvariation.asp

Mallows, C.L. (1973), “Some Comments on Cp,” Technometrics, 15, 661–675

**Appendix**

MODEL R

The REG Procedure

Model: R

Dependent Variable: Sales

|  |  |
| --- | --- |
| **Number of Observations Read** | 298 |
| **Number of Observations Used** | 298 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 8 | 2.878377E13 | 3.597971E12 | 107.53 | <.0001 |
| **Error** | 289 | 9.669657E12 | 33459019356 |  |  |
| **Corrected Total** | 297 | 3.845343E13 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 182918 | **R-Square** | 0.7485 |
| **Dependent Mean** | 872037 | **Adj R-Sq** | 0.7416 |
| **Coeff Var** | 20.97596 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 650928 | 38585 | 16.87 | <.0001 |
| **Likely\_customers\_1R** | **1** | -18.91138 | 11.41893 | -1.66 | 0.0988 |
| **Pop\_GE\_18\_3R** | **1** | 1.36273 | 0.71072 | 1.92 | 0.0562 |
| **Pop\_grades\_9\_12\_3R** | **1** | -13.16816 | 9.50439 | -1.39 | 0.1670 |
| **Malls\_300K\_0\_5R\_dum** | **1** | 21374 | 24712 | 0.86 | 0.3878 |
| **per\_cap\_inc\_3R** | **1** | 3.08125 | 1.03444 | 2.98 | 0.0031 |
| **south** | **1** | -277522 | 29845 | -9.30 | <.0001 |
| **west** | **1** | 469632 | 24478 | 19.19 | <.0001 |
| **Mid\_grocery\_1R\_dum** | **1** | -25661 | 21843 | -1.17 | 0.2410 |