Project 11

Will Hallgren

4/21/2021

ECON 5645: Empirical Linear Modeling

University of North Texas

What is Timmy Tom's?

Timmy Tom's is a quick-service sandwich shop that provides subs, sandwiches, assorted beverages, and snacks. Many people need food fast and do not have time to cook for themselves, but many options in regards to fast-food are proven to be unhealthy. Timmy Tom's aims to be a healthier alternative to other established fast-food franchises by providing a wider variety of sandwiches with healthier options for ingredients. Timmy Tom's also provides nutritional information on the menu for customers wishing to keep track of their daily nutrition. Timmy Tom's features a sandwich of the month, which uses slightly more premium and unique ingredients than the rest of the sandwiches. Timmy Tom's offers house-made chips and a pickle spear with every sandwich meal package. The first Timmy Tom's location opened in 1974 in Champaign, Illinois and has been a hit among consumers ever since. They opened their first location in the downtown Champaign area. People greatly appreciated a new quick-service sandwich shop and Timmy Tom's was able to gain a footing financially. They decided to move into the Chicago area in an attempt to make the brand more lucrative. Timmy Toms was a hit in the bustling Windy City and was easily able to justify moving into such a high-profile, high-cost environment. The brand began franchising out more locations in the great lakes area, and soon enough began expanding into other regions of the United States. They expanded their reach to more cities including: Los Angeles, Seattle, Phoenix, Austin, Miami, Atlantic City, and New York. With the popularity of quick-service sandwich shops in the United States, Timmy Tom's was able to establish themselves as a competitive fast-food franchise with locations all across the country.

Timmy Tom's has experienced rapid growth up until March of 2020. They have done so well, that they believe that they have currently saturated the market. They have all of the high-revenue locations that they are aware of and are beginning to place more care and consideration on where they begin to open more new stores. Timmy Tom's has taken data from 307 of their stores in 2019 to contrive which characteristics of their current locations contribute to their high-revenue potential. Using this data, we can determine a clearer expansion path moving forward. We can gain insights into where Timmy Tom's should aim to locate and who the target audience is for marketing.

The Dependent Variable

The dependent variable in this regression is sales in 2019, measured in dollars. The intuition is that there are factors that contribute to the level of sales that a store generates. We are trying to find which of the factors which contribute to increased sales so that we can build a regression model that can accurately predict the performance of potential new locations in the form of predicted sales. "Store_ID" and, "year_open" are both descriptors to the dependent variable, sales. "Store_ID" is an identification number which corresponds to the age of the store. The oldest Timmy Tom's store opened has a store ID of one, and the newest store has a store ID of 306. "year_open" is simply the year in which the store was opened.

Graph 1: Raw Sales Data

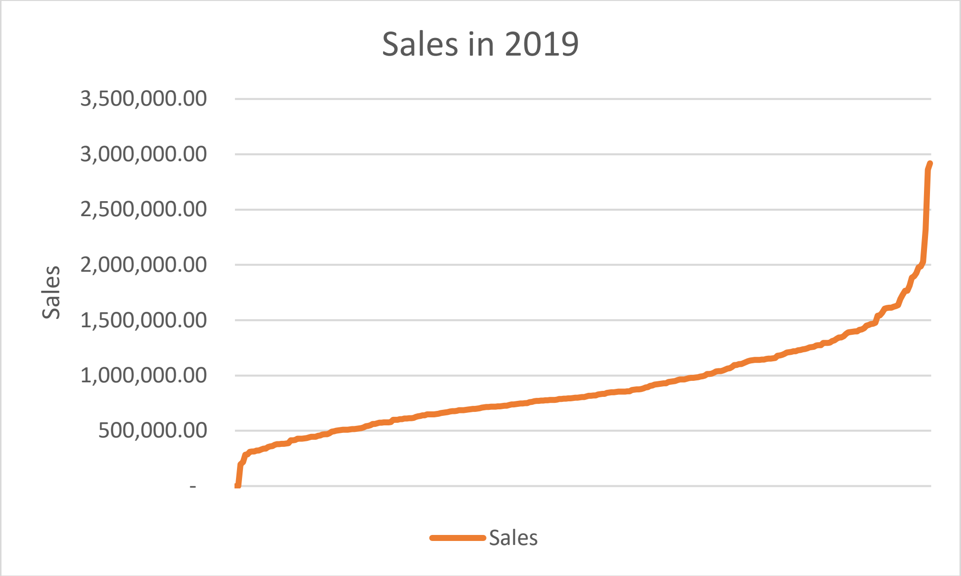

Chart 1: Raw Sales Summary Statistics, All Observations

Number of Obs.	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
306	895,285	420,174	0	2,918,588

There are some unreasonable values in this data set. There are two stores, the latest two stores to open, that have a value of 0 for sales. These stores had not been open long enough to generate sales when the data was collected. We must remove them from the data set as they will not help us understand the factors that lead to high-revenue potential.

Graph 2: Sales Data with Unreasonable Values Removed

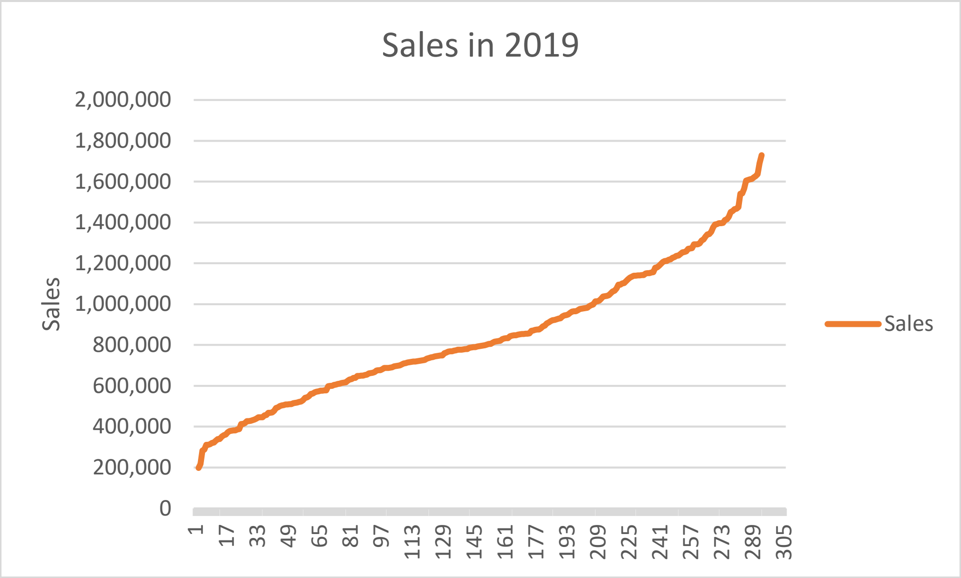

Chart 2: Summary Statistics with Unreasonable Values Removed

Number of Obs.	Mean	Standard Deviation	Minimum	Maximum
304	901,175	415,194	198,153	2,918,588

With unreasonable values removed, there are still a few observations that we need to remove from this data set. Outlier observations should be removed because they do not represent the general trend. They can cause parameter estimates to be inaccurate when performing OLS regression. Outliers are defined by any observation outside two and a half standard deviations from the mean.

Mean Sales + 2.5*(Standard Deviation) = 901,175 + 2.5*(415,194) = 1,939,160

Mean Sales – 2.5*(Standard Deviation) = 901,175 – 2.5*(415,194) = -136,810

There are 6 stores that produced sales above two and a half standard deviations from the mean. Observations 301 to 306 will be removed.

Graph 3: Sales Data with Unreasonable and Outlier Observations Removed

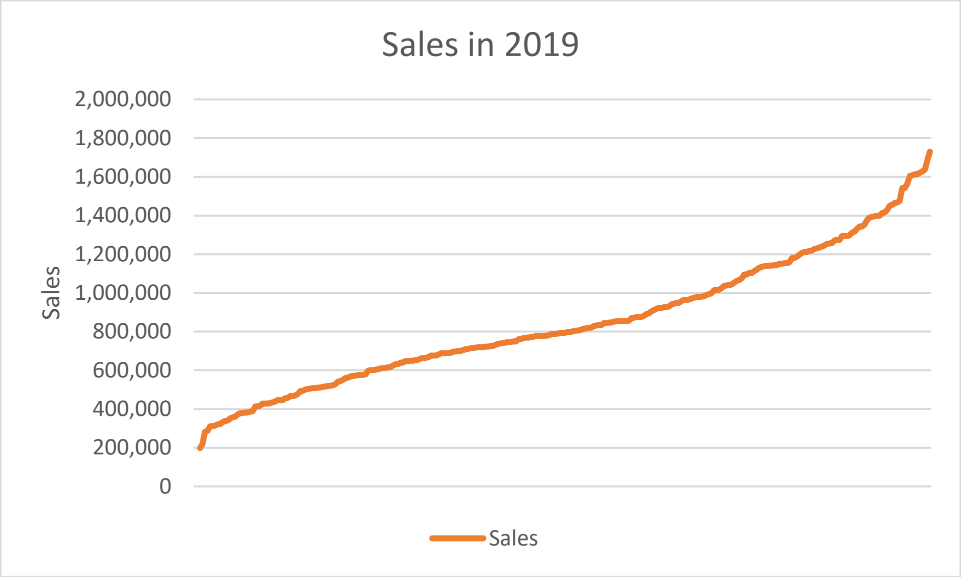

Chart 3: Summary Statistics for Sales with Unreasonable and Outlier Observations Removed

Number of Obs.	Mean	Standard Deviation	Minimum	Maximum	Coefficient of Variation
298.00	872,036.76	359,823.32	198,152.77	1,925,871.54	41.26

With unreasonable and outlier observations removed, now we must determine whether sufficient variation in the dependent variable exists. If the dependent variable does not have sufficient variation, then we do not have sufficient statistical evidence to suggest that sales vary across stores at all, therefore none of the data collected can explain sales at Timmy Tom's. The coefficient of variation is the following formula:

$$\frac{\sigma_Y}{\bar{Y}} * 100$$

For continuous dependent variables, such as sales, the coefficient of variation should be greater than 2. The coefficient of Variation for Sales in this data set is 39.32, so there is no problem with variation.

Table 4: Total Observations Removed from Dependent variable

OBSERVATION NUMBER:	REASON:
1	Sales were 0
2	Sales were 0
301	Sales were above 2 standard deviations from the mean
302	Sales were above 2 standard deviations from the mean
303	Sales were above 2 standard deviations from the mean
304	Sales were above 2 standard deviations from the mean
305	Sales were above 2 standard deviations from the mean
306	Sales were above 2 standard deviations from the mean

Table 5: Independent Variable Definitions

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.
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).

Table 5: Independent Variable Definitions

VARIABLE NAME:	DEFINITION:
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.
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.

Table 5: Independent Variable Definitions

VARIABLE NAME:	DEFINITION:
Pop_grades_9_12_3R	The number of people who live within 3 radial miles of a given store who are in grades 9 through 12.
Pop_grades_9_12_5T	The number of people who live within a 5-minute drive of a given store who are in grades 9 through 12.
Pop_grad_school_3R	The number of people who live within 3 radial miles of a given store who are in graduate school.
Pop_grad_school_5T	The number of people who live within a 5-minute drive of a given store who are in graduate school.
Pop_in_school_3R	The number of people who live within 3 radial miles of a given store who are in school (any school).
Pop_in_school_5T	The number of people who live within a 5-minute drive of a given store who are in school (any school).
Pop_undergrads_3R	The number of people who live within 3 radial miles of a given store who are undergraduates.
Pop_undergrads_5T	The number of people who live within a 5-minute drive of a given store who are undergraduates.
Pop_Masters_3R	The number of people living within 3 radial miles of a given store whose highest educational attainment is a Masters degree.
Pop_Masters_5T	The number of people living within a 5-minute drive of a given store whose highest educational attainment is a Masters degree.
Pop_some_college_3R	The number of people who live within 3 radial miles of a given store who have some college education, but no degree.
Pop_some_college_5T	The number of people who live within a 5-minute drive of a given store who have some college education, but no degree.
Tot_HH_Expend_3R	Total annual expenditure (in dollars) of households located within 3 radial miles of a given store.
Tot_HH_Expend_5T	Total annual expenditure (in dollars) of households located within a 5-minute drive of a given store.
Cust_value	A measure of the value of all residents in the Timmy Tom's network, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value).
Cust_value_per_cap	A measure of the value, per capita, of all residents in the Timmy Tom's , with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value).
Cust_value_region	A measure of the value of residents within the neighboring geographic region of a given store, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value).
Cust_value_per_cap_region	A measure of the value, per capita, of residents within the neighboring geographic region of a given store, with regard to how likely they are to purchase items from Timmy Tom's (a higher number implies a greater value).

Table 5: Independent Variable Definitions

VARIABLE NAME:	DEFINITION:
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.
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.

Table 5: Independent Variable Definitions

VARIABLE NAME:	DEFINITION:
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.
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.

Table 5: Independent Variable Definitions

VARIABLE NAME:	DEFINITION:
avg_LOR_3R	The average number of years that residents lived in their home (length of residence) for people who live within 3 radial miles of a given sortore.
Pop_married_3R	The number of married people who live within 3 radial miles of a given store.
Pop_married_5T	The number of married people who live within a 5-minute drive of a given store.
Distance_hwy	The distance, in miles, to the nearest highway.
Distance_hwy_interstate	The distance, in miles, to the nearest highway or interstate.
Distance_interstate	The distance, in miles, to the nearest interstate.
restaurants_3R	The number of restaurants (of all types) located within 3 radial miles of a given store.
retail_3R	The number of retail establishments (of all types) located within 3 radial miles of a given store.
restaurants_retail_3R	The number of restaurants and retail establishments (of all types) located within 3 radial miles of a given store.
Asian_HH_3R	The number of Asian households located within 3 radial miles of a given store.
Asian_HH_5T	The number of Asian households located within a 5-minute drive of a given store.
Asian_pop_3R	The Asian population (in people) living within 3 radial miles of a given store.
Asian_pop_5T	The Asian population (in people) living within a 5-minute drive of a given store.
Black_HH_3R	The number of black households located within 3 radial miles of a given store.
Black_HH_5T	The number of black households located within a 5-minute drive of a given store.
Black_pop_3R	The black population (in people) living within 3 radial miles of a given store.
Black_pop_5T	The black population (in people) living within a 5-minute drive of a given store.
Hispanic_HH_3R	The number of Hispanic households located within 3 radial miles of a given store.
Hispanic_HH_5T	The number of Hispanic households located within a 5-minute drive of a given store.
Hispanic_pop_3R	The Hispanic population (in people) living within 3 radial miles of a given store.
Hispanic_pop_5T	The Hispanic population (in people) living within a 5-minute drive of a given store.

Potential Issues with the Independent Variables

First, I check the summary statistics and Coefficient of variation of all variables. The coefficient of variation is defined by the following equation:

$$CV_{x_{ij}} = \left| \frac{\left(\sigma_{x_{ij}}\right)}{\bar{x}} * 100 \right|$$

I am looking for missing values, unreasonable numbers, and outliers. Standard deviation cannot be equal to zero. I also check to see if the Coefficient of Variation is greater than 2 for all variables. We desire variation in the independent variable data as well as the dependent variable data because without sufficient variation, there will be no unique solution for a parameter estimate when we begin running regressions.

Table 6: Summary statistics for Problematic Variables

Variable	**Number of Obs.**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**	**Coefficient of Variation**
Traffic_count	225	22,676.73	19,729.65	4,645.41	178,296.58	87.00
Fast_food_8T	298	25.14	14.70	-45.00	101.00	58.48
Big_box_1R	298	2,989.38	51,491.52	0.00	888,888.00	1722.48

The variable, "traffic count" is missing 73 observations. The variable, "Fast_food_8T" has negative values; the indicated minimum is -45. The variable, "Big_box_1R" has a suspicious maximum value of 888,888. After contacting the data collector, I was unable to recover the data for the traffic count observations. Since this variable is missing data for 24.5% of the observations, I will not include it in the list of potential regressors. The data collector did not have an explanation for the unusual observations in big_box_1R and fast_food_8T, so I will exclude them from the list of potential regressors as well.

A Brief explanation on Micronumerosity

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

In short, the degrees of freedom in statistics is the difference between the number of observations and the number of variables being tested. In this case, we are testing 292 observations against 113 variables.

292-113 = 179 degrees of freedom

In general, we need more than 30 degrees of freedom to prevent the micronumerosity issue. Since we have 179 degrees of freedom, we will have no problems with micronumerosity in this data set.

Correlation

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

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

I will use the Pearson correlation coefficient to evaluate the correlation between independent variables and sales. It is the covariance between sales and one of the independent variables divided by the product of their standard deviations. The p-value given by a hypothesis test is the probability that we obtain test results at least as extreme as the results observed. Therefore, we want to minimize the p-value, as opposed to maximize, if we are looking for variables with a significant effect on sales. If the p-value in the hypothesis test is less than .12, an 88% confidence level, then we reject the null hypothesis; we have sufficient statistical evidence to believe that our tested independent variable is correlated with sales. The null hypothesis states that our variable being tested is not correlated to sales at all. The correlation coefficient would equal 0.

Table 7: Correlation with Sales

Variable:	**Correlation Coefficient:**	**P-Value:**
Food_away_3R	0.12738	0.0279
Likely_customers_1R	-0.14658	0.0113
Pop_GE_18_3R	-0.17827	0.002
Pop_GE_18_5T	-0.10487	0.0707
Pop_21_39_3R	-0.15457	0.0075
Pop_21_39_5T	-0.09245	0.1112
Pop_40_49_3R	-0.16519	0.0042
Pop_40_49_5T	-0.10013	0.0844
Pop_50_69_3R	-0.19131	0.0009
Pop_50_69_5T	-0.10977	0.0584
Pop_70_85_3R	-0.18971	0.001
Pop_70_85_5T	-0.09091	0.1173
Mid_grocery_index_1R	-0.08884	0.126
Big_box_index_1R	0.11078	0.0561
Pop_Associates_3R	-0.15035	0.0093
Pop_Associates_5T	-0.08869	0.1266
Pop_Bachelors_3R	-0.09294	0.1094
Pop_Doctorate_3R	-0.09444	0.1037
Pop_grades_9_12_3R	-0.17034	0.0032
Pop_grades_9_12_5T	-0.10236	0.0777
Pop_grad_school_3R	-0.11565	0.0461
Pop_in_school_3R	-0.17088	0.0031
Pop_Masters_3R	-0.09268	0.1103
Pop_some_college_3R	-0.17537	0.0024
Pop_some_college_5T	-0.09548	0.0999
Tot_HH_Expend_3R	-0.15132	0.0089
Cust_value_per_cap_region	0.14837	0.0103
HHinc_LT_25K_3R	-0.18612	0.0012
HHinc_LT_25K_5T	-0.14236	0.0139

Table 7: Correlation with Sales

Variable:	Correlation Coefficient:	P-Value:
HHinc_25_49K_3R	-0.19072	0.0009
HHinc_25_49K_5T	-0.1235	0.0331
HHinc_50_74K_3R	-0.16376	0.0046
HHinc_75_99K_3R	-0.1131	0.0511
Med_HHinc_3R	0.11658	0.0443
HH_1person_3R	-0.14178	0.0143
HH_1person_5T	-0.09265	0.1105
HH_2person_3R	-0.1464	0.0114
HH_3person_3R	-0.17359	0.0026
HH_3person_5T	-0.09774	0.0921
HH_4person_3R	-0.16196	0.0051
HH_4person_5T	-0.096	0.0981
HH_5person_3R	-0.20113	0.0005
HH_5person_5T	-0.14302	0.0135
HH_6person_3R	-0.21447	0.0002
HH_6person_5T	-0.16166	0.0052
Brady_Bunch_3R	-0.21149	0.0002
Brady_Bunch_5T	-0.16844	0.0035
labor_blue_3R	-0.20053	0.0005
labor_blue_5T	-0.13468	0.02
labor_farm_3R	-0.16783	0.0037
labor_farm_5T	-0.10529	0.0695
labor_white_col_3R	-0.10673	0.0658
Pop_married_3R	-0.17544	0.0024
restaurants_3R	-0.11826	0.0413
retail_3R	-0.13693	0.018
restaurants_retail_3R	-0.13242	0.0222
Black_HH_3R	-0.20168	0.0005
Black_HH_5T	-0.16821	0.0036
Black_pop_3R	-0.19228	0.0008
Black_pop_5T	-0.16982	0.0033
Hispanic_HH_3R	-0.16298	0.0048
Hispanic_HH_5T	-0.12089	0.037

Table 7: Correlation with Sales

Variable:	Correlation Coefficient:	P-Value:
Hispanic_pop_3R	-0.16999	0.0032
Hispanic_pop_5T	-0.12942	0.0255

Potential Regressors:

I have compiled a list of potential regressors from the random variables for model building. All of these variables have no unreasonable values, sufficient variation, and a statistically significant P-value at the 88% confidence level from the correlation hypothesis test. If any of the variables above do not satisfy all of the conditions, I will exclude them from this list.

List 1: List of Potential Regressors

Potential Regressors:
1. Food_away_3R
2. Likely_customers_1R
3. Pop_GE_18_3R
4. Pop_GE_18_5T
5. Pop_21_39_3R
6. Pop_GE_18_5T
7. Pop_21_39_3R
8. Pop_21_39_5T
9. Pop_40_49_3R
10. Pop_40_49_5T
11. Pop_50_69_3R
12. Pop_50_69_5T
13. Pop_70_85_3R
14. Pop_70_85_5T
15. Mid_grocery_index_1R
16. Big_box_index_1R

List 1: List of Potential Regressors

Potential Regressors:
17. Pop_Associates_3R
18. Pop_Associates_5T
19. Pop_Bachelors_3R
20. Pop_Doctorate_3R
21. Pop_grades_9_12_3R
22. Pop_grades_9_12_5T
23. Pop_grad_school_3R
24. Pop_in_school_3R
25. Pop_Masters_3R
26. Pop_some_college_3R
27. Pop_some_college_5T
28. Tot_HH_Expend_3R
29. Cust_value_per_cap_region
30. HHinc_LT_25K_3R
31. HHinc_LT_25K_5T
32. HHinc_25_49K_3R
33. HHinc_25_49K_5T
34. HHinc_50_74K_3R
35. HHinc_75_99K_3R
36. Med_HHinc_3R
37. HH_1person_3R
38. HH_1person_5T
39. HH_2person_3R
40. HH_3person_3R
41. HH_3person_5T
42. HH_4person_3R
43. HH_4person_5T
44. HH_5person_3R
45. HH_5person_5T
46. HH_6person_3R
47. HH_6person_5T
48. Brady_Bunch_3R
49. Brady_Bunch_5T
50. labor_blue_3R
51. labor_blue_5T
52. labor_farm_3R
53. labor_farm_5T
54. labor_white_col_3R

List 1: List of Potential Regressors

of Potential Regressors:
55. Pop_married_3R
56. restaurants_3R
57. retail_3R
58. restaurants_retail_3R
59. Black_HH_3R
60. Black_HH_5T
61. Black_pop_3R
62. Black_pop_5T
63. Hispanic_HH_3R
64. Hispanic_HH_5T
65. Hispanic_pop_3R
66. Hispanic_pop_5T

Single Trait and Multi-Trait Dummy Variables

Table 8: Dummy Variable Names and Description

Variable Name:	Description:
Free_standing	=1 if the store is located in a free-standing building, =0 if otherwise.
Strip_mall	=1 if the store is located in a strip-mall building, =0 if otherwise.
Other_building	=1 if the store is located in any other building besides a free-standing building or a strip mall, =0 if otherwise. This is the base trait for building dummy variables.
HD	=1 if the store is located in a high-density area, =2 if not.
HV	=1 if the store is located in a "high visibility" area, =2 if not.
South	=1 if the store is located in the Southern region of the United States, =0 if not.
Central	=1 if the store is located in the central region of the United States, =0 if not.
West	=1 if the store is located in the Western region of the United States, =0 if not.
East	=1 if the store is located in the Eastern Region of the United States, =0 if not. This is the base trait for region dummy variables.

Table 9: Summary Statistics for Dummy Variables

Variable	Number of Obs.	Mean	Standard Deviation	Minimum	Maximum
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
HD	298	1.24	0.43	1	2
free_standing	298	0.61	0.49	0	1
strip_mall	298	0.33	0.47	0	1

"HD" and "HV" are invalid dummy variables. A dummy variable can only take on values of zero and one, otherwise results will be biased in favor of the traits that are designated by higher numbers. I will redefine these variables so that they are proper dummy variables. The minimum value should always be, "0" and the maximum should always be, "1" for dummy variables.

Table 19: Corrected Dummy Variables

Variable Name:	Description:
High_Density	=1 if the store is located in a high-density area, =0 if not.
High_Visibility	=1 if the store is located in a "high visibility" area, =0 if not.

Table 11: Summary Statistics for Corrected Dummy Variables

Variable	Number of Obs.	Mean	Standard Deviation	Minimum	Maximum
high_visibility	298	0.92	0.27	0	1
south	298	0.22	0.41	0	1
central	298	0.3	0.46	0	1
west	298	0.45	0.5	0	1
high_density	298	0.76	0.43	0	1
free_standing	298	0.61	0.49	0	1
strip_mall	298	0.33	0.47	0	1

For sufficient variation in dummy variables, the mean must be less than or equal to 0.9 or greater than or equal to 0.1. The variable, "high_visibility" has a mean of 0.92, indicating that roughly 92% of stores in the sample are located in high visibility areas. There are not enough stores that are not located in a high visibility area to compare the data with. This issue cannot be corrected so high_visibility will be excluded from further consideration and analysis. The group of region dummy variables does not have enough observations in the base trait. The group of building dummy variables does not have enough observations in the base trait ether. In order to fix this, I will merge one of the variables that is most similar to the base-trait with the base. For the region dummy variables, I will merge Central with East and for the building dummy variables, I will merge free_standing with other_building.

Table 12: Dummy Variables with Issues

Variable(s):	**Reason:**
High_visibility	Insufficient variation
West, Central, South	Insufficient variation in the base
Free_standing, strip_mall	Insufficient variation in the base

Correlation

Next, I will do a hypothesis test to determine whether the variables are significantly correlated with Sales to narrow down a list of regressors. Variables that do not have statistically significant correlation with sales should be excluded from the regression. I will use the Pearson correlation coefficient to evaluate the correlation between independent variables and sales. It is the covariance between sales and one of the independent variables divided by the product of their standard deviations. The hypothesis being tested is that the true correlation between any of the variables and sales is 0; that is, they are insignificant:

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

The p-value is the probability of obtaining test results that are at least as extreme or more than the results observed, so a small P-value is evidence that we should reject the hypothesis. If the P-value is sufficiently small, we have found statistically significant correlation between sales and the variable. In this case, we are looking for P-values that are less than or equal to 0.12, an 88% confidence level.

Table 13: Correlation between Sales and Dummy Variables

	south	**west**	**high_density**	**strip_mall**
Correlation Coefficient	-0.64	0.81	-0.03664	-0.25803
P-Value	<.0001	<.0001	0.5286	<.0001

The variable, “high_density” has a P-value greater than 0.12, which means that it is statistically insignificant at the 88% confidence level; we do not reject the null hypothesis. The correlation coefficient on “high_density” does not matter since we have sufficient statistical evidence to believe that the true correlation coefficient is zero. The rest of the variables have P-values less than .0001, so we will reject the null hypothesis, they have a statistically significant degree of linear association. “South” has a negative correlation coefficient which suggests that on average, sales are generally lower in the South than stores that are not in the South, holding all else constant. “West” has a positive correlation coefficient, which suggests that on average, sales are generally higher in the West than stores that are not in the west. “strip_mall” has a negative correlation coefficient, which suggests that on average, stores located in strip malls generally have lower sales than standalone- and other-type buildings.

Limited Integer Value (LIV) Variables

Table 14: Variable Definitions for LIV Variables

Variable Name:	Description:
All_malls_1R	The number of malls (of all types) 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.
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.
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.
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 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.
Mid_grocery_0_5R	The number of mid-level grocery stores that are located within a one-half radial mile of a given store.
Mid_grocery_1R	The number of mid-level grocery stores that are located within one radial mile of a given store.
Sandwich_shop_8T	The number of sandwich shops that are located within an 8-minute drive of a given store.

Table 14: Variable Definitions for LIV Variables

Variable Name:	Description:
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.

Limited integer value variables are variables that can only take on a very limited number of integer values, usually six or fewer. To determine whether LIV variables have sufficient variation, each outcome in the variable must comprise at least 10% of the observations in the sample or no more than 90% of the observations in the sample. To find sufficient variation, we will use frequency tables.

Frequency Tables

Table 15: Competitor_A_0_5R Frequency

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 value “2” comprises only .34% of the observations, which is not enough to be considered a LIV Variable. This Variable can be made into a dummy variable, which will have sufficient variation.

Table 16: Competitor_A_1R Frequency

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 value “2” comprises only 1.34% of the observations, which is not enough to be considered a LIV Variable. This variable can be made into a dummy variable, which will have sufficient variation.

Table 17: Competitor_B_0_5R Frequency

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 value “1” comprises only 3.36% of observations and the value “2” is only .34% of the observations, which is not enough to be considered a LIV Variable. This variable cannot be fixed and will be removed from further consideration.

Table 18: Competitor_B_1R Frequency

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 value “1” comprises only 5.37% of observations and the value “2” comprises only .34% of observations. This variable cannot be fixed and will be removed from further consideration.

Table 19: Competitor_C_0_5R Frequency

Competitor_C_0_5R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	276	92.62	276	92.62
1	22	7.38	298	100

There are not enough observations where competitor C is present and too many where they are not. This variable cannot be fixed and will be removed from further consideration.

Table 20: Competitor_C_1R Frequency

Competitor_C_1R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	248	83.22	248	83.22
1	50	16.78	298	100

There are no issues with this LIV variable.

Table 21: Competitor_D_0_5R Frequency

Competitor_D_0_5R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	277	92.95	277	92.95
1	21	7.05	298	100

There are not enough observations where competitor D is present and too many where they are not. This variable cannot be fixed and will be removed from further consideration.

Table 22: Competitor_D_1R Frequency

Competitor_D_1R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	240	80.54	240	80.54
1	58	19.46	298	100

There are no issues with this LIV variable.

Table 23: Bakeries_0_5R Frequency

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

Observations containing 2 or more bakeries do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 24: Bakeries_1R Frequency

Bakeries_1R	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

Observations containing 3 or more bakeries do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 25: Low_grocery_0_5R Frequency

Low_grocery_0_5R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	273	91.61	273	91.61
1	25	8.39	298	100

There are not enough observations where “Low-end” grocery stores are present and too many where they are not. This variable cannot be fixed and will be removed from further consideration.

Table 26: Low_grocery_1R Frequency

Low_grocery_1R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	246	82.55	246	82.55
1	51	17.11	297	99.66
2	1	0.34	298	100

Observations containing 2 “Low-end” grocery stores do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 27: Mid_grocery_0_5R Frequency

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

Observations containing 2 or more “medium-end” grocery stores do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 28: Mid_grocery_1R Frequency

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

Observations containing 3 or more “medium-end” grocery stores do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 29: Malls_300K_0_5R Frequency

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

Observations containing 2 or more malls do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 30: Malls_300K_1R Frequency

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

Observations containing 2 or more malls do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 31: Universities_0_5R Frequency

Universities_0_5R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	293	98.32	293	98.32
1	5	1.68	298	100

Not enough locations have a university within a half mile radius. This variable cannot be fixed and will be removed from further consideration.

Table 32: Universities_3R Frequency

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

Observations containing 2 or more universities within a 3-mile radius do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 33: Universities_5T Frequency

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

Not enough locations have one or more universities within a 5-minute travel time. This variable cannot be fixed and will be removed from further consideration.

Table 34: Universities_8T Frequency

Universities_8T	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

Observations containing 2 or more universities within a 8-minute travel time do not comprise enough of the data. This variable can be made into a dummy variable with sufficient variation.

Table 35: Big_box_0_5R Frequency

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

This variable is not a LIV variable. It takes on more than six integer values. This variable will be considered a continuous random variable.

Table 36: Sandwich_shop_8T Frequency

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

This variable is not a LIV variable. It takes on more than six integer values. This variable will be considered a continuous random variable.

Table 37: All_malls_1R Frequency

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
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

This variable is not a LIV variable. It takes on more than six integer values. This variable will be considered a continuous random variable.

Table 35: LIV Variables with no Issues

Variable Name:
Competitor_C_1R
Competitor_D_1R

Table 36: LIV Variables with Issues

Variable Name:	Description of Issue:
Competitor_A_0_5R	The value "2" is only .34% of the observations, which is not enough to be considered a LIV Variable. This Variable can be made into a dummy variable.
Competitor_A_1R	The value "2" is only 1.34% of the observations, which is not enough to be considered a LIV Variable. This variable can be made into a dummy variable.
Competitor_B_0_5R	The value "1" is only 3.36% of observations and the value "2" is only .34% of the observations, which is not enough to be considered a LIV Variable. This variable cannot be fixed and will be removed.
Competitor_B_1R	The value "1" is only 5.37% of observations and the value "2" is only .34% of observations. This variable cannot be fixed and will be removed.
Competitor_C_0_5R	There are not enough observations where competitor C is present and too many where they are not. This variable cannot be fixed and will be removed.
Competitor_D_0_5R	There are not enough observations where competitor D is present and too many where they are not. This variable cannot be fixed and will be removed.
Bakeries_0_5R	Observations containing 2 or more bakeries do not comprise enough of the data. This variable can be made into a Dummy Variable.
Bakeries_1R	Observations containing 3 or more bakeries do not comprise enough of the data. This variable can be made into a Dummy Variable.
Low_grocery_0_5R	There are not enough observations where "Low-end" grocery stores are present and too many where they are not. This variable cannot be fixed and will be removed.

Table 36: LIV Variables with Issues

Variable Name:	Description of Issue:
Low_grocery_1R	Observations containing 2 “Low-end” grocery stores do not comprise enough of the data. This variable can be made into a Dummy Variable.
Mid_grocery_0_5R	Observations containing 2 or more “Medium-end” grocery stores do not comprise enough of the data. This variable can be made into a Dummy Variable.
Mid_grocery_1R	Observations containing 3 or more “Medium-end” grocery stores do not comprise enough of the data. This variable can be made into a Dummy Variable.
Malls_300K_0_5R	Observations containing 2 or more malls do not comprise enough of the data. This variable can be made into a Dummy Variable.
Malls_300K_1R	Observations containing 2 or more malls do not comprise enough of the data. This variable can be made into a Dummy Variable.
Univeristies_0_5R	Not enough locations have a university within a half mile radius. This variable cannot be fixed and will be removed.
Universities_1R	Not enough locations have a university within a one-mile radius. This variable cannot be fixed and will be removed.
Universities_3R	Observations containing 2 or more universities within a 3-mile radius do not comprise enough of the data. This variable can be made into a Dummy Variable.
Universities_5T	Not enough locations have one or more universities within a 5-minute travel time. This variable cannot be fixed and will be removed.
Universities_8T	Observations containing 2 or more universities within an 8-minute travel time do not comprise enough of the data. This variable can be made into a Dummy Variable.

Table 10: Newly Created Variables

Variable Name:	**Description:**
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.
Sandwich_shop_8T	The number of sandwich shops that are located within an 8-minute drive of a given store.
All_malls_1R	The number of malls (of all types) that are located within one radial mile of a given store.
Competitor_A_0_5R_dum	=1 if Competitor A has a location within a half radial mile form Timmy Tom's, =0 if not.
Competitor_A_1R_dum	=1 if Competitor A has a location within a single radial mile from Timmy Tom's, =0 if not.
Bakeries_0_5R_dum	=1 if a Bakery is located within a half radial mile from Timmy Tom's, =0 if not.
Bakeries_1R_dum	=1 if a Bakery is located within a radial mile from Timmy Tom's, =0 if not.
Low_grocery_1R_dum	=1 if a "low-end" grocery store is located within a radial mile from Timmy Tom's, =0 if not.
Mid_grocery_0_5R_dum	=1 if a "medium-end" grocery store is located within a half-radial mile from Timmy Tom's, =0 if not.
Mid_grocery_1R	=1 if a "medium-end" grocery store is located within a radial mile from Timmy Tom's, =0 if not.
Malls_300K_0_5R_dum	=1 if a mall is located within a half-radial mile from Timmy Tom's, =0 if not.
Malls_300K_1R_dum	=1 if a mall is located within a radial mile from Timmy Tom's, =0 if not.
Universities_3R_dum	=1 if a university is located within 3 radial miles from Timmy Tom's, =0 if not.
Universities_8T	=1 if a university is located within an 8-minute travel time from Timmy Tom's, =0 if not.

Table 11: Summary Statistics for Newly Created Variables

Variable	Number of Obs.	Mean	Standard Deviation	Minimum	Maximum	Coefficient of Variation
Big_box_0_5R	298	3.29	4.24	0	26.00	128.83
Sandwich_shop_8T	298	3.94	3.49	0	19.00	88.59
All_malls_1R	298	3.74	3.27	0	18.00	87.33
Competitor_A_0_5R_dum	298	0.2	0.4	0	1	N/A
Competitor_A_1R_dum	298	0.38	0.49	0	1	N/A
Bakeries_0_5R_dum	298	0.28	0.45	0	1	N/A
Bakeries_1R_dum	298	0.44	0.5	0	1	N/A
Low_grocery_1R_dum	298	0.17	0.38	0	1	N/A
Mid_grocery_0_5R_dum	298	0.34	0.47	0	1	N/A
Malls_300K_0_5R_dum	298	0.26	0.44	0	1	N/A
Malls_300K_1R_dum	298	0.38	0.49	0	1	N/A
Universities_3R_dum	298	0.23	0.42	0	1	N/A
Universities_8T_dum	298	0.18	0.39	0	1	N/A

The summary statistics verify that the newly defined dummy variables are proper dummy variables for our regression. Each has a minimum value of 0 and a maximum value of 1, a non-zero standard deviation, and means that are less than or equal to 0.9 and greater than or equal to 0.1. The continuous random variables have a non-zero standard deviation and a sufficiently large coefficient of variation.

Correlation

I will do the same pre-model analysis procedure for the LIV variables and newly defined variables. I will find the Pearson Correlation Coefficient and do a hypothesis test to find out if the variables are correlated with Sales to narrow down a list of potential regressors. For the Correlation Coefficient, a negative figure identifies an average negative impact on sales. A positive figure identifies an average positive impact on sales. The hypothesis being tested is that the true correlation coefficient between any of the variables and sales is 0; that is, they are insignificant. The p-value is the probability of obtaining test results that are at least as extreme as the results observed, so a small P-value is evidence that we should reject the null hypothesis, meaning that we have found statistically significant correlation between sales and the independent variable. In this case, we are looking for P-values that are less than or equal to 0.12, an 88% confidence level.

Table 12: Correlation Coefficients and P-values for LIV Variables and Newly Defined Variables

Variable Name:	**Correlation Coefficient:**	**P-value:**
Big_box_0_5R	0.10083	0.0823
Sandwich_shop_8T	-0.03838	0.5092
All_malls_1R	-0.04306	0.4590
Competitor_C_1R	-0.00665	0.909
Competitor_D_1R	0.05836	0.3154
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
Mid_grocery_1R_dum	-0.12772	0.0275
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

Most of the variables here have insignificant P-values, suggesting that they are not strongly correlated with sales. “Big_box_0_5R” had a P-value of .08, which suggests that it is strongly correlated with sales. The sign of the correlation coefficient is positive, which suggests that the number of big box stores within a half-mile radius of a Timmy Tom’s location moves with sales. “Mid_grocery_1R_dum” had a statistically significant P-value of with a correlation coefficient of -0.13 and a negative sign on the correlation coefficient. The number of medium-end grocery stores within a one-mile radius moves against sales. "Malls_300K_0_5R_dum" also had a statistically significant P-value of .09, with a positive correlation coefficient. A Timmy Tom's location that is within a half radial-mile of a mall with more than 300,000 square-feet of leasable space is expected to have higher sales than a location that is not, holding all else constant.

The following is a list of potential regressors after pre-model analysis. Each regressor has sufficient variation and a statistically significant P-value from the correlation hypothesis test.

List 2: Potential Regressors from Variables Intended to be Dummies or LIV Variables

Potential Regressors:
1. South
2. West
3. Strip_mall
4. Mid_grocery_1R_dum
5. Malls_300K_0_5R_dum
6. Big_box_0_5R

List 3: Final List of Potential Regressors

Potential Regressors:
1. Food_away_3R
2. Likely_customers_1R
3. Pop_GE_18_3R
4. Pop_GE_18_5T
5. Pop_21_39_3R
6. Pop_21_39_5T
7. Pop_40_49_3R
8. Pop_40_49_5T
9. Pop_50_69_3R
10. Pop_50_69_5T
11. Pop_70_85_3R
12. Pop_70_85_5T
13. Mid_grocery_index_1R
14. Big_box_index_1R
15. Pop_some_college_3R
16. Pop_some_college_5T
17. Pop_Associates_3R
18. Pop_Associates_5T
19. Pop_Bachelors_3R
20. Pop_Masters_3R
21. Pop_Doctorate_3R
22. Pop_grades_9_12_3R
23. Pop_grades_9_12_5T
24. Pop_grad_school_3R
25. Pop_in_school_3R
26. Tot_HH_Expend_3R
27. Cust_value_per_cap_region
28. HHinc_LT_25K_3R
29. HHinc_LT_25K_5T
30. HHinc_25_49K_3R
31. HHinc_25_49K_5T
32. HHinc_50_74K_3R
33. HHinc_75_99K_3R
34. Med_HHinc_3R
35. HH_1person_3R
36. HH_1person_5T
37. HH_2person_3R
38. HH_3person_3R

List 3: Final List of Potential Regressors

Potential Regressors:
39. HH_3person_5T
40. HH_4person_3R
41. HH_4person_5T
42. HH_5person_3R
43. HH_5person_5T
44. HH_6person_3R
45. HH_6person_5T
46. Brady_Bunch_3R
47. Brady_Bunch_5T
48. labor_blue_3R
49. labor_blue_5T
50. labor_farm_3R
51. labor_farm_5T
52. labor_white_col_3R
53. Pop_married_3R
54. restaurants_3R
55. retail_3R
56. restaurants_retail_3R
57. Black_HH_3R
58. Black_HH_5T
59. Black_pop_3R
60. Black_pop_5T
61. Hispanic_HH_3R
62. Hispanic_HH_5T
63. Hispanic_pop_3R
64. Hispanic_pop_5T
65. South
66. West
67. Strip_mall
68. Mid_grocery_1R_dum
69. Malls_300K_0_5R_dum
70. Big_box_0_5R

Considerations on Theory and Logic

After going about pre-model analysis, it is important to consider demand theory and logical reasoning on the current set of variables so that we do not miss out on any other potential regressors that may improve our understanding of the factors that influence sales. We may want to add more variables to our data set upon consideration. Demand theory suggests that sales would be related to: own price of goods, income, population, preferences, prices of substitutes, prices of complements, and expectations for the future. Data could be taken to compare the price level of Timmy Tom's across locations, controlling for price level in the area. An index could be made to compare the price level of Timmy Tom's against the price level of competitors. Another index could be made to compare the price level at Timmy Tom's with the price level of retail stores, or price level at malls. We already have variables that examine the proximity of malls or retail stores such as, "Malls_300K_0_5R_dum" and, "retail_3R" however those variables do not take price level into account. We have "med_HH_inc_3R" to capture income. We have many population variables such as, "Black_HH_3R", "Hispanic_HH_3R", and "HH_3person_3R". The variable, "Likely_customers_1R" sufficiently captures expectations for the future. Gauging the amount of people who have indicated that they are likely customers can give Timmy Tom's a reasonable expectation for the kind of sales that they will see in the future. For preferences, it may be necessary to understand the nature of the health-food aspect of Timmy Tom's business model. Dummy variable data can be taken indicating whether or not Timmy Tom's customers prefer health-food. If the health-food variable improves sales, then we will know that the health food part of the business model should be emphasized and advertised more. If the health-food variable reduces sales, then Timmy Tom's may take steps toward becoming more of a fast-food franchise because the health-food aspect of the business model is an unnecessary cost. Logically speaking, I cannot think of an area that the current data set does not cover.

Table 1: Top 5 Models from Regression

Model:	**Regressors:**
S	Pop_GE_18_3R Pop_21_39_3R labor_white_col_3R restaurants_3R strip_mall Mid_grocery_1R_dum south west
V	Food_away_3R Pop_50_69_3R Big_box_index_1R Cust_value_per_cap_region HH_2person_3R restaurants_3R south west strip_mall
W	Pop_21_39_3R Pop_50_69_3R Big_box_index_1R HH_2person_3R Med_HHinc_3R south west strip_mall
Y	Pop_50_69_3R Mid_grocery_index_1R Big_box_index_1R HH_2person_3R Med_HHinc_3R south west strip_mall
AB	Pop_50_69_3R Big_box_index_1R HH_2person_3R Med_HHinc_3R Mid_grocery_1R_dum south west strip_mall

Table 2: Top 5 Models Fit Statistics

Model	Calc. General F-Test (P-value)	Value of R-Squared	Value of Adjusted R-Squared	Number of Significant Slopes with Correct Sign	Number of Significant Slopes with Incorrect Sign	Joint F-test testing significance of Region Dummy Variables[1] (P-value)	Value of OOS MAPE
S	103.54 (<.0001)	0.7413	0.7342	3 out of 3	0	357.44 (<.0001)	0.17
V	93.68 (<.0001)	0.7454	0.7374	6 out of 6	0	332.63 (<.0001)	0.198
W	106.81 (<.0001)	0.7473	0.7403	6 out of 6	0	358.99 (<.0001)	0.23
Y	108.50 (<.0001)	0.7502	0.7433	8 out of 8	0	362.49 (<.0001)	0.23
AB	107.70 (<.0001)	0.7488	0.7419	6 out of 6	0	355.23 (<.0001)	0.24

Table 3: Assessment of Fit Statistics

Model	General F Test	R-Squared	Adjusted R-Squared	% Significant slopes with Correct sign	Any Significant Regressors with Wrong Sign?	OOS MAPE	Overall Assessment
S	Significant	Fairly strong	Fairly Strong	100%	No	Good	Good
V	Significant	Fairly Strong	Fairly Strong	100%	No	Good	Good
W	Significant	Fairly Strong	Fairly Strong	100%	No	Good	Good
Y	Significant	Fairly Strong	Fairly Strong	100%	No	Good	Good
AB	Significant	Fairly Strong	Fairly Strong	100%	No	Good	Good

[1] $H_0 =$ The group of region dummy variables is insignificant.

Multicollinearity

Next, I will evaluate the independent variables in the top five models for multicollinearity. Multicollinearity exists in the model if the independent variables are highly correlated with each other. If any of the variables are too correlated, then the variance of the estimated regression parameters will be over-inflated and test statistics will be artificially small. We are more likely to fail to reject a null hypothesis that is false and we may get parameter estimates that suggest a counter-intuitive relationship between the independent variables and sales. In short, this means that the model is inaccurate when multicollinearity is present and we cannot rely on the predictions produced by the model. Moderate to severe multicollinearity exists if the absolute value of the correlation coefficient between regressors is greater than 0.5. A correlation coefficient of 1 indicates perfect multicollinearity.

Table 4: Variable Pairs with Moderate to Severe Multicollinearity

Variable Pair:	**Correlation Coefficient:**
Pop_50_69_3R & HH_2person_3R	0.94274
Pop_21_39_3R & Pop_50_69_3R	0.86768
HH_2person_3R & restaurants_3R	0.88799
Pop_GE_18_3R & Pop_21_39_3R	0.96855
Pop_GE_18_3R & labor_white_col_3R	0.92152
Pop_21_39_3R & labor_white_col_3R	0.88483
Pop_GE_18_3R & restaurants_3R	0.86788
Pop_21_39_3R & restaurants_3R	0.86744
labor_white_col_3R & restaurants_3R	0.82273

The variable pair with the highest correlation coefficient is “pop_ge_18_3R” and “pop_21_39_3R”. This makes intuitive sense because “pop_GE_18_3R” indicates the number of people in a 3-mile radius of a Timmy Tom’s location who are 18 years or older. “Pop_21_39_3R” indicates the number of people in a 3-mile radius of a Timmy Tom’s location who are aged between 21 and 39. Since people who are aged between 21 and 39 are indeed older than 18, it would make sense that this variable pair exhibits severe multicollinearity. Model S contains six variable pairs that have severe multicollinearity. Model S also exhibits a slightly smaller general F-test statistic and features far fewer significant regressor slopes than the other models. It also produced the worst MAPE of the group. Although model V contains only 2 variable pairs that exhibit severe multicollinearity, it has the lowest general F-statistic and the second lowest MAPE. Model W also contains two variable pairs that exhibit sever multicollinearity. Model W had a slightly smaller general F-statistic however its R-square and MAPE were only slightly smaller than models Y and AB. Models Y and AB both only contain one variable pair with severe multicollinearity. Model AB has the highest MAPE and model Y has the highest general F-statistic. Model Y produced the most significant regressor slopes and has the highest R-squared.

Best Fitting Model

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

$$\widehat{Sales} = 674593 + -8.78(\mathrm{Pop\_50\_69\_3R}) - 28007(\mathrm{Mid\_grocery\_index\_1R})$$

(<.0001) (0.0207) (0.0649)

$$+ 2625.93(\mathrm{Big\_box\_index\_1R}) + 15.68(\mathrm{HH\_2person\_3R}) + 1.16(\mathrm{Med\_HHinc\_3R})$$

(0.1711) (0.0125) (0.0489)

$$- 274990(\mathrm{south}) + 456691(\mathrm{west}) - 40552(\mathrm{strip\_mall})$$

(<.0001) (<.0001) (0.0889)

The model as it is contains large numbers, so I will measure sales in thousands of dollars, which will adjust the parameters by three decimal places to consolidate the numbers. I will measure pop_50_69_3R in thousands, HH_2Person_3R in thousands, and med_HHinc_3R in thousands so that these numbers do not appear too small:

$$\widehat{Sales}\_T = 674.59 + -8.78(\mathrm{Pop\_50\_69\_3R\_T}) - 28.01(\mathrm{Mid\_grocery\_index\_1R})$$

(<.0001) (0.0207) (0.0649)

$$+ 2.63(\mathrm{Big\_box\_index\_1R}) + 15.68(\mathrm{HH\_2person\_3R\_T})$$

(0.1711) (0.0125)

$$+ 1.16(\mathrm{Med\_HHinc\_3R\_T}) - 274.99(\mathrm{south}) + 456.69(\mathrm{west}) - 40.55(\mathrm{strip\_mall})$$

(0.0489) (<.0001) (<.0001) (0.0889)

Interpretations of Coefficients

Pop_50_69_3R_T: Estimated Coefficient = -8.78

Each additional thousand people who are aged between 50 and 69 who live within 3 radial miles of a Timmy Tom's location is expected to reduce sales on average by a factor of 8.78 thousand dollars, holding all else constant.

Mid_grocery_index_1R: Estimated Coefficient = -28.01

Each additional index point towards the value of mid-level grocery stores within one radial mile of a Timmy Tom's location is expected to reduce sales on average by a factor of 28.01 thousand dollars, holding all else constant.

Big_box_index_1R_H: Estimated Coefficient = 2.63

Each additional index point towards the value of "Big Box" stores within one radial mile of a Timmy Tom's location is expected to increase sales on average by a factor of 2.63 thousand dollars, holding all else constant.

HH_2person_3R_T: Estimated Coefficient = 15.68

Each additional thousand 2-person households within 3 radial miles of a Timmy Tom's location is expected to increase sales on average by a factor of 15.68 thousand dollars, holding all else constant.

Med_HHinc_3R_T: Estimated Coefficient =1.16

A one thousand dollar increase in median household income within 3 radial miles of a Timmy Tom's location is expected to increase sales on average by 1.16 thousand dollars, holding all else constant.

South: Estimated Coefficient = -274.99

A Timmy Tom's location that is located in the Southern region as opposed to Central and Eastern regions is expected on average to decrease sales by a factor of 274.99 thousand dollars, holding all else constant.

West: Estimated Coefficient = 456.69

A Timmy Tom's location that is located in the Western region as opposed to Central and Eastern regions is expected on average to increase sales by a factor of 45.67 thousand dollars, holding all else constant.

Strip_mall: Estimated Coefficient = -40.55

A Timmy Tom's location that is located in a strip mall as opposed to a standalone or other building is expected to decrease sales on average by a factor of 40.55 thousand dollars, holding all else constant.

Model Scoring and Sequential Regression

I will be using sequential regression to try to estimate a better fitting model than I found from intuitively building a regression model, model Y:

$$\begin{gathered}
\widehat{Sales}\_T = \underset{(<.0001)}{674.59} + \underset{(0.0207)}{-8.78}(\text{Pop\_50\_69\_3R\_T}) - \underset{(0.0649)}{28.01}(\text{Mid\_grocery\_index\_1R}) \\
+ \underset{(0.1711)}{2.63}(\text{Big\_box\_index\_1R}) + \underset{(0.0125)}{15.68}(\text{HH\_2person\_3R\_T}) \\
+ \underset{(0.0489)}{1.16}(\text{Med\_HHinc\_3R\_T}) - \underset{(<.0001)}{274.99}(\text{south}) + \underset{(<.0001)}{456.69}(\text{west}) - \underset{(0.0889)}{40.55}(\text{strip\_mall})
\end{gathered}$$

Sequential regression includes the forward-selection method, the stepwise selection method, the backward elimination method, the maximum R-square improvement selection method, and the adjusted R-square selection method to find the best fitting models. The best models have the highest statistical significance while also having a high R-squared and adjusted R-squared. I will then use model scoring to evaluate potential locations and predict sales potential for new Timmy Tom's locations. Potential locations will be organized into four different revenue types: High sales potential, medium sales potential, above-average sales potential, and low sales potential.

The multi-trait dummy variables must be removed when conducting sequential regression because all multi-trait dummies must be included the regression model, however the computer program may try to add the multi-trait dummies one at a time. This can redefine the base group in a way that does not make sense. Therefore, multi-trait dummy variables must be added after sequential regression. There is one multi-trait dummy to consider adding to this regression: the geographic location regressor. The building type regressor was intended to be a multi-trait dummy however I included free standing buildings in the base group to account for insufficient variation and therefore has only one trait.

Sequential Regression

Forward-Selection sequential regression is a method of building a linear model where models are regressed with only one regressor, and the model with the smallest p-value is selected. The regressors with the smallest p-values are continuously added until the addition of another regressor produces a p-value above a threshold. In this case, the threshold is 0.2. The forward-selection method did not produce a better model than the one found previously. It has an R-square of 0.17, which is far below the R-square of the model found previously. The forward-selection model also had one less significant variable and some of the coefficient signs were counter-intuitive. After adding the region multi-trait dummies, the R-square improved greatly but there were even more variables with counter intuitive signs. This model is not preferred to the previous model.

Stepwise selection is a similar method of sequential regression only in order for a variable to remain in the model after each successive step, that variable must be statistically significant at the threshold. Stepwise-selection used fewer variables but has the same problems as forward-selection. The R-square was 0.17 as well and produced variable coefficients that were counter-intuitive. Adding the region dummies to the model improved the R-square, however many of the variables ether became insignificant or their coefficients had a counter-intuitive sign. This model is not preferred to the previous model.

Backward elimination sequential regression is the opposite of forward-selection sequential regression. Instead, a model with all regressors is estimated and a joint F-test is conducted on each regressor and the one with the largest p-value is removed. This continues until all regressors have p-values that are beneath the threshold of 0.2. Backward elimination produced a model with far too many regressors. When building a regression model, we desire a parsimonious model; or a model with the fewest possible regressors while still providing the greatest possible explanation of the dependent variable. I am looking for models with 8 or more regressors, and 15 or fewer regressors. Backwards selection produced a model with 32 regressors. Although the R-square was a slightly higher than forward selection or stepwise selection with a value of 0.33, this model is far from parsimonious. The model found earlier had a much higher R-square value with far fewer regressors. Because of the length of the backwards elimination model, I will not add the region multi-trait dummy variables. This model is not preferred to the previous model.

R-square improvement sequential regression builds models based on the R-square value. All one-variable models will be estimated and the one with the highest R-squared is selected. Then, all two variable models are estimated and the model that produces the largest positive change in R-squared is selected. This continues until all possible model sizes are found. This method finds the best model for each number of variables. The adjusted R-square selection method is nearly the same, only using the adjusted R-square value instead of the normal R-square value. All one-variable models will be estimated and the one with the highest adjusted R-squared is selected. This process is repeated for all possible model sizes. Both R-square improvement and adjusted R-square selection provided similar results as backward elimination. The highest R-square produced was 0.33 and the highest adjusted R-square produced was 0.25. The best model had 33 regressors, which is too many to justify just a small change in R-square and adjusted R-square when the previous model had a much higher R-square and adjusted R-square with far fewer regressors. Because of the length of the R-square improvement and adjusted R-square selection models, I will not add the region multi-trait dummy variables. This model is not preferred to the previous model.

None of the sequential regression models were better than the previous model, model Y. I will use model Y to predict sales potential for the prospective new Timmy Tom's locations.

Model Scoring

Threshold Values for Projected Sales

We can estimate the performance of prospective locations by using the models we have found and compare our estimated sales potential to the average sales across current Timmy Tom's locations. Above average sales potential is defined as: between the average and one standard deviation from the average. Medium Sales potential is between the average plus one standard deviation and the average plus two standard deviations. High sales potential is greater than the average plus 2 standard deviations.

Table 1: Threshold Values for Sales Potential

Sales Level:	**Value of Sales:**
Average Sales:	$872,036.76
Average + 1 std. dev.	$1,231,860.08
Average + 2 std. dev.	$1,591,683.40

Table 2: Sales Projection Results

Store ID	**Projected Sales**
312	$509,604.78
313	$681,726.20
314	$1,138,532.86
315	$749,698.31
316	$790,570.24
317	$666,363.26
318	$417,419.99
319	$474,719.30
320	$1,143,668.81
321	$726,281.40

Table 3: Sales Projection Analysis

Revenue type:	**Description:**	**Store number:**
High Sales Potential	Projected sales ≥ (average sales plus two standard deviations)	None
Medium Sales Potential	Projected sales ≥ (average sales plus one standard deviation) but Projected sales < (average sales plus two standard deviations)	None
Above-Average Sales Potential	Projected sales ≥ average sales but Projected sales < (average sales plus one standard deviations)	314, 320
Low Sales Potential	Projected sales < average sales	312, 313, 315, 316, 317, 318, 319, 321

Locations 314 and 320 are the only two locations that have above average sales potential. The rest of the locations have low sales potential. There are no, "easy-win" high or medium sales potential locations of the locations provided.

General Advice

In addition to the model, I have some general advice for expansion strategies for Timmy Tom's. Both the mid-grocery index and the mid-grocery dummy consistently had negative coefficients when using them in regression, which suggests that Timmy Tom's competes with mid-level grocery stores. Many of the competitor variables, aside from mid-grocery stores, were insignificant which suggests that sales are rarely affected by locating near other sandwich competitors, so other sandwich competitors in the area should not be a major factor in deciding where to open future locations. Planners for future Timmy Tom's locations should avoid locating near grocery stores but other competitors are less important. Mall variables consistently had positive variable coefficients, which suggests that Timmy Tom's subs are compliments to mall shopping. Timmy Tom's should locate near malls, but not in strip malls. Number of white-collar workers in the area also had a consistently positive coefficient, which suggests that Timmy Tom's is popular among white-collar workers. Finally, adding the region dummy variables greatly improved the accuracy of all regressions and proved very significant in every joint F-test, which suggests that region is the most important factor to consider when searching for new locations. The two highest sales potential locations were both located in the West and they had the highest sales potential by far.