UNIVERSITY OF NORTH TEXAS

RETAIL LOCATION REGRESSION ANALYSIS

ADTA 5120

MARISSA MCKEE

Contents

Business Objectives 2

Data Description 2

General Data Construction 2

Data Exploration 3

Data Cleansing 5

Dummy Variables 5

Data Transformation 6

Variable Selection 7

Variable Selection 8

Model Technique 8

Multivariate Linear Regression Model Construction 9

Multivariate Linear Regression Model Evaluation 12

Mean Error Evaluation 12

R-Squared and Adjust R-Squared Evaluation 13

F Statistic Evaluation 14

Durbin Watson Evaluation 14

Final results .... **Error! Bookmark not defined.**

Mean Error Evaluation 16

R-Squared and Adjust R-Squared Evaluation 17

F Statistic Evaluation 17

Durbin Watson Evaluation 18

Final results .... **Error! Bookmark not defined.**

Conclusion 19

Resources 20

Appendix 20

Business Objectives

Slacks on Racks is a discount department store with 68 locations. Their missions is to price their products competitively low. They have experienced rapid growth in their niche market. They continue to grow their brand but are running out of *easy win* locations to open new stores. Slacks on Racks is pivoting to a more conservative growth strategy and is interested in areas that will provide strong revenue potential. The Sacks on Racks team has identified 5 potential locations to open 1 new store. The primary objective of this project is to identify which location of the 5 will have the highest revenue opportunities.

For this project I have built a sales forecasting model based on the 2016 sales data. In the subsequent sections, I will go through some initial EDA, identify outliers, calculate correlations, develop a simple linear regression for large and small stores, develop a multivariate linear regression for large stores, validate each regression model, and evaluate and compare the results.

Data Description

There were 447 variables provided for the 68 existing stores. Variables include subjects like:

- Sales for 2016
- Store characteristics (ex. opening year, region, state, square footage)
- Competitive prescence (ex. competitor counts and competitor distance scores)
- Demographic information (ex. marriage status, race, age, education)
- Estimated customer value (household income, home value)
- Retail and business density (housing units, labor)

General Data Construction

Cotenants are large retail stores that act as an area draw and bring in customer to nearby stores. Cotenants are assigned types like convenience stores, grocery stores, mall type stores, movie theaters, pharmacies, or big box stores like Target or Best Buy.

Spatially generated variables are used to show distance or time from the store. Variables denoted with a suffix of 16TO are read that the store is a 16-minute drive away. Variables denoted with a suffix of 1RO are read that the store is 1-mile radius away.

Distance scores are variables that measure the count and proximity of a given cotenant or competitor within 1 mile. The farther the business is from the location, the smaller the distance score.

Customer value are variables that measure the estimated value of household nearby. Residents and workplaces are evaluated in these variables. There are a few qualitative variables that measure the percentage rather than the count. These variables are designed to compare similar characteristics across both large and small stores.

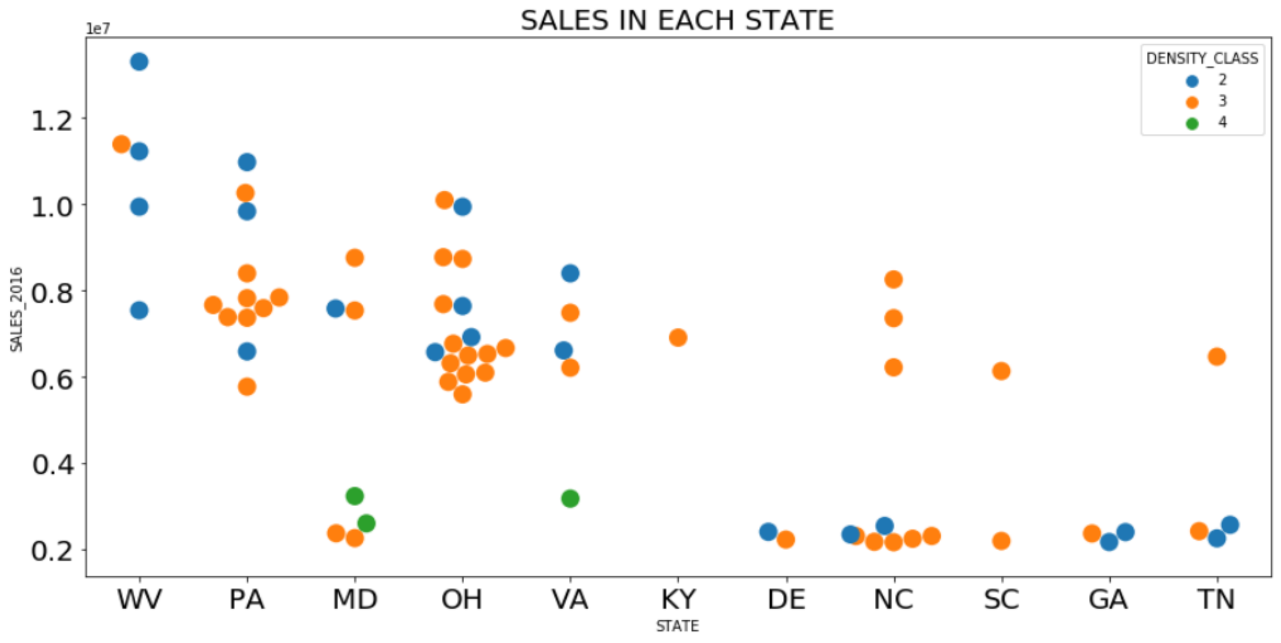
For this project I will be using the percentage variables as opposed to the actual values so I can compare small and large stores on the same scale. I will also use spatially generated variables suffixed with 8TO and 16TO.

Data Exploration

For more data exploration results please visit my GitHub repo for this project here.

Data exploration is the initial step in modeling. Please refer to Figure 1. 2016 sales are shown below by state and population density classes 2 (in town), 3 (suburban), and 4 (metro). The highest sales are seen in West Virginia in towns and the lowest sales are seen North Carolina, Georgia, Delaware, and Tennessee in suburban or metro areas. Ohio has the least deviation of sales and most are clustered hovering around $60,00 in sales on average.

Figure 1: Sales in each state shown by density class



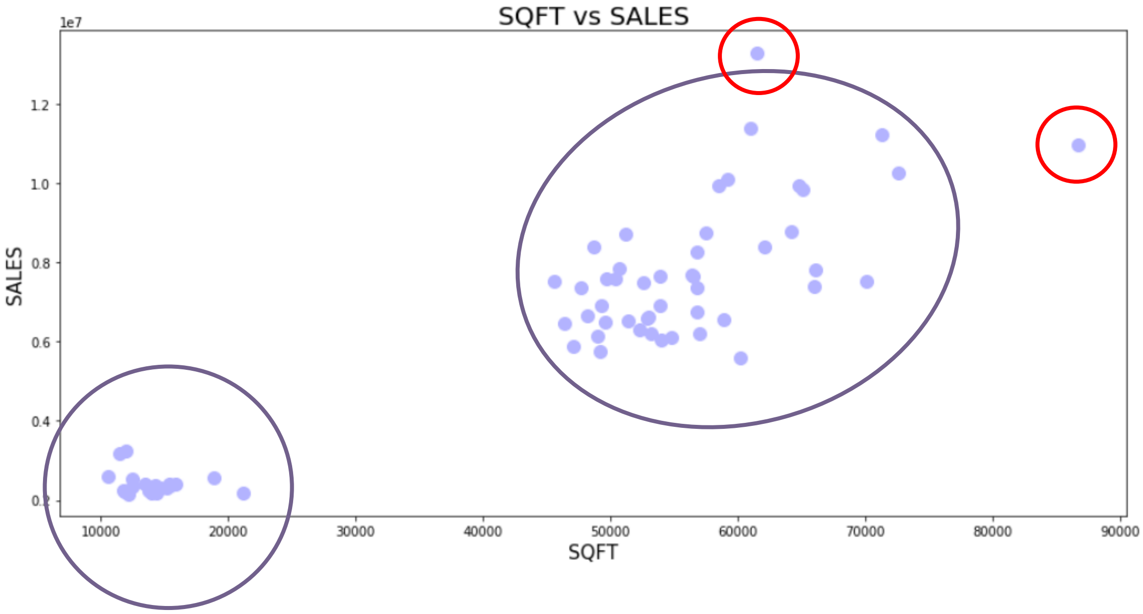
Data Exploration

Please refer to Figure 2. The scatter plot below shows sales by square footage. Linear regression models need for the relationship between the independent and dependent variables to be linear. It's also important to check for outliers since linear regression is sensitive to outlier effect.

There are two clouds of clustered sales that could potentially act differently from each other. Stores that have less than 23,000 square feet are clustered with smaller sales and stores having more than 23,000 square feet are clustered with higher sales. For this project I will examine both sets of large and small stores on their own. Large stores typically make higher sales and the goal of this project is to predict the store with the largest sales.

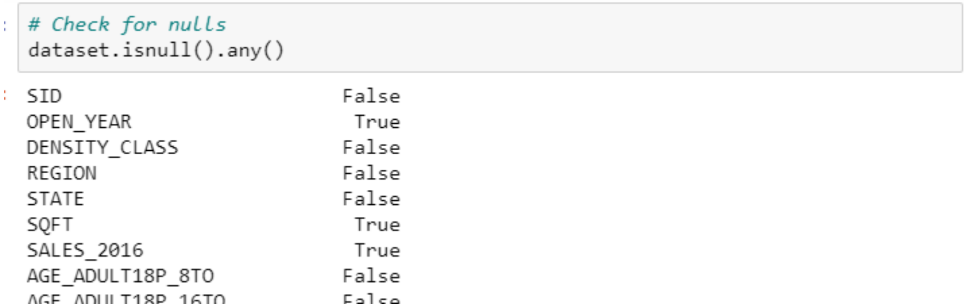
There are a couple notable outliers circled in red. The first outlier in which has the highest sales recorded is a store in West Virginia. From the previous EDA conducted, we know that West Virginia stores have on average higher sales, and it may not be an outlier for that region in the future if more stores are opened. The second outlier has the largest square footage and slightly higher than average sales for large stores. For this project, I will keep both outliers.

Figure 2: Scatter plot SQFT vs sales



Data Cleansing

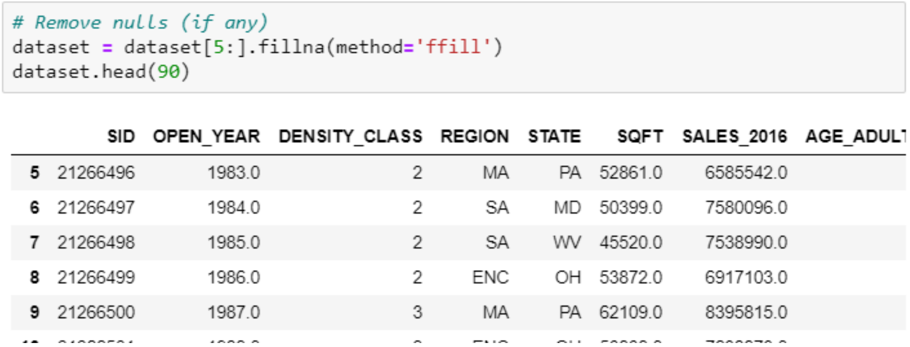
The data is fairly clean. However, there are some null values. The code below is used to check for null data. As you can see, we have null values in OPEN_YEAR, SQFT, and SALES_2016.



```
# Check for nulls
dataset.isnull().any()
```

```
SID                         False
OPEN_YEAR                    True
DENSITY_CLASS               False
REGION                      False
STATE                       False
SQFT                         True
SALES_2016                   True
AGE_ADULT18P_8TO            False
AGE_ADULT18P_16TO           False
```

The fillna method is used to propagate the last valid observation forward to the next valid value. The 5 potential new stores selected by the Slacks on Racks team are frontloaded in the data as the first five observations. We don't want to include those in our dataset and will remove those records for the time being.



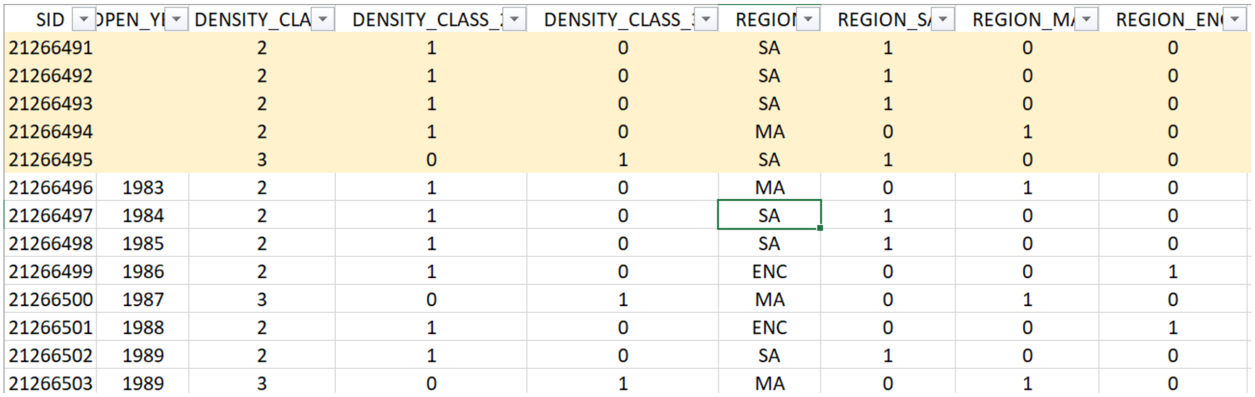
```
# Remove nulls (if any)
dataset = dataset[5:].fillna(method='ffill')
dataset.head(90)
```

	SID	OPEN_YEAR	DENSITY_CLASS	REGION	STATE	SQFT	SALES_2016	AGE_ADUL
5	21266496	1983.0	2	MA	PA	52861.0	6585542.0	
6	21266497	1984.0	2	SA	MD	50399.0	7580096.0	
7	21266498	1985.0	2	SA	WV	45520.0	7538990.0	
8	21266499	1986.0	2	ENC	OH	53872.0	6917103.0	
9	21266500	1987.0	3	MA	PA	62109.0	8395815.0	

Dummy Variables

Refer to Figure 13. Dummy variables are numerical variables used to analyze subgroups of categorical variables. In the figure below, I created 2 dummy variables for expressing DENSITY_CLASS subgroups and 3 dummy variables for expressing REGION subgroups.

Figure 13: Dummy variable creation



SID	OPEN_YI	DENSITY_CLA	DENSITY_CLASS_	DENSITY_CLASS_	REGIOI	REGION_S/	REGION_M/	REGION_EN(
21266491		2	1	0	SA	1	0	0
21266492		2	1	0	SA	1	0	0
21266493		2	1	0	SA	1	0	0
21266494		2	1	0	MA	0	1	0
21266495		3	0	1	SA	1	0	0
21266496	1983	2	1	0	MA	0	1	0
21266497	1984	2	1	0	SA	1	0	0
21266498	1985	2	1	0	SA	1	0	0
21266499	1986	2	1	0	ENC	0	0	1
21266500	1987	3	0	1	MA	0	1	0
21266501	1988	2	1	0	ENC	0	0	1
21266502	1989	2	1	0	SA	1	0	0
21266503	1989	3	0	1	MA	0	1	0

Data Transformation

The dependent variable sales when split into small and large stores is heavily skewed to the right for large stores. Please refer to Figure 6 and Figure 7 below. In order to normalize the dependent variable, the lognormal function is applied to the sales variable. The new dependent variable will contain the log of the sales values.

Figure 6: Sales histogram for large stores

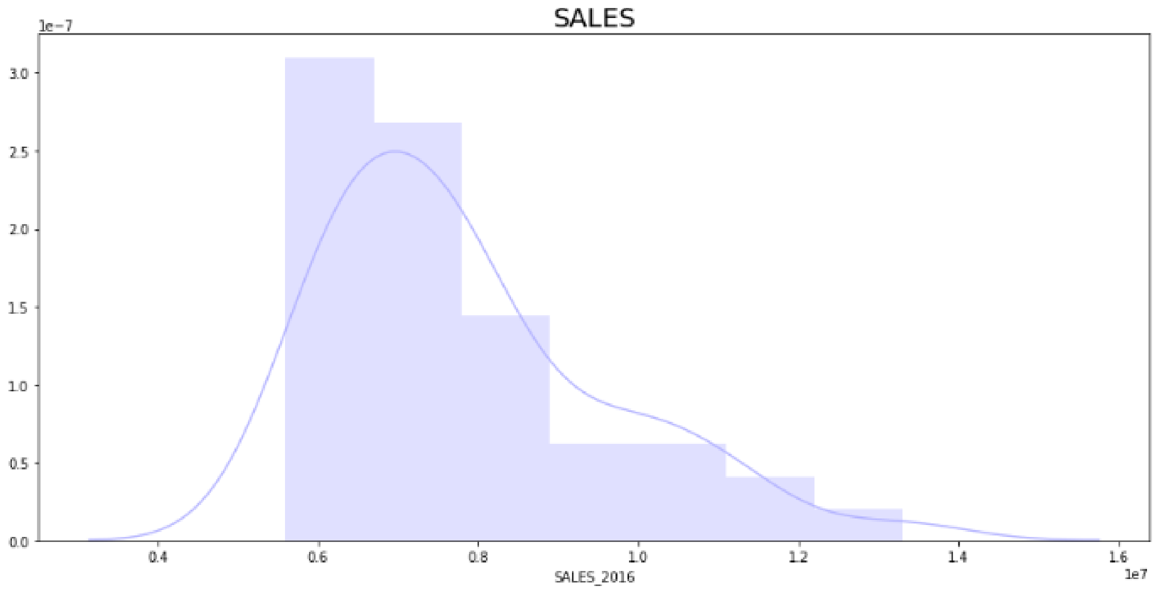
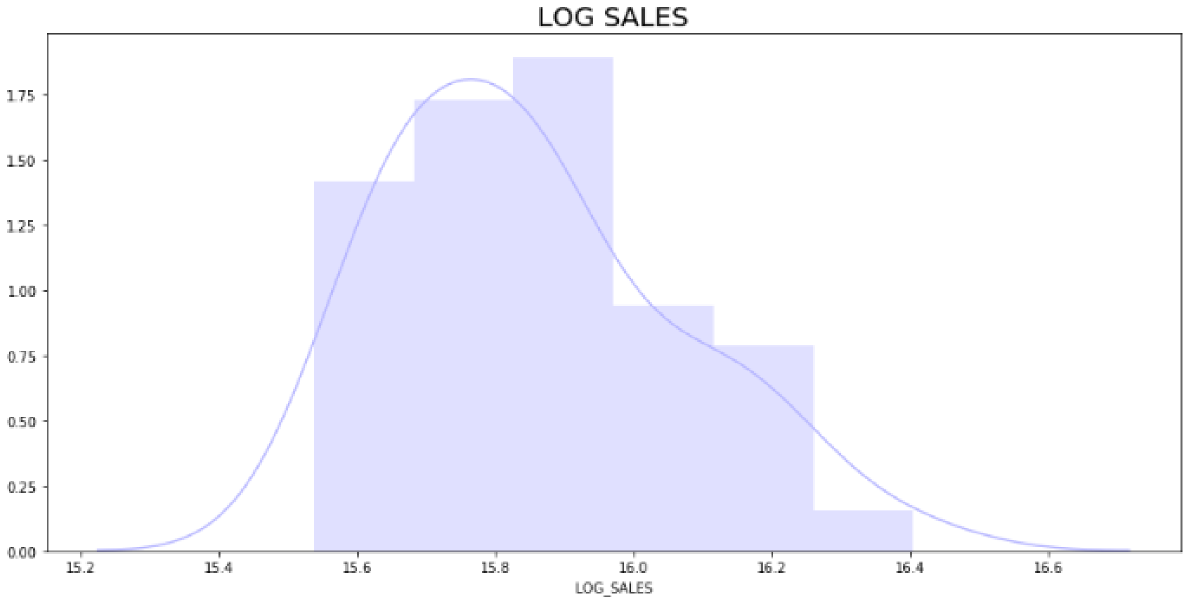


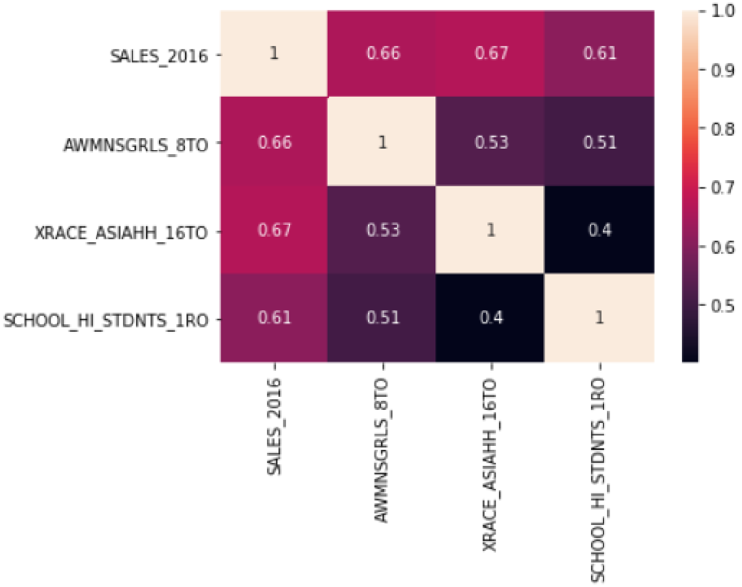
Figure 7: Lognormal sales histogram for large stores



Variable Selection

Please refer to Figure 3. Below is a correlation heatmap of several highly correlated variables with the dependent variable sales. Variables with a positively calculated .20 correlation or higher were analyzed in the final multivariate regression model. Higher correlations indicate strong linear relationships between dependent and independent variables. Low correlations don't indicate a lack of significant relationship between independent and dependent variables. The relationship could be nonlinear or masked by other correlations with other independent variables.

Figure 3: Correlation heatmap of small store model variables



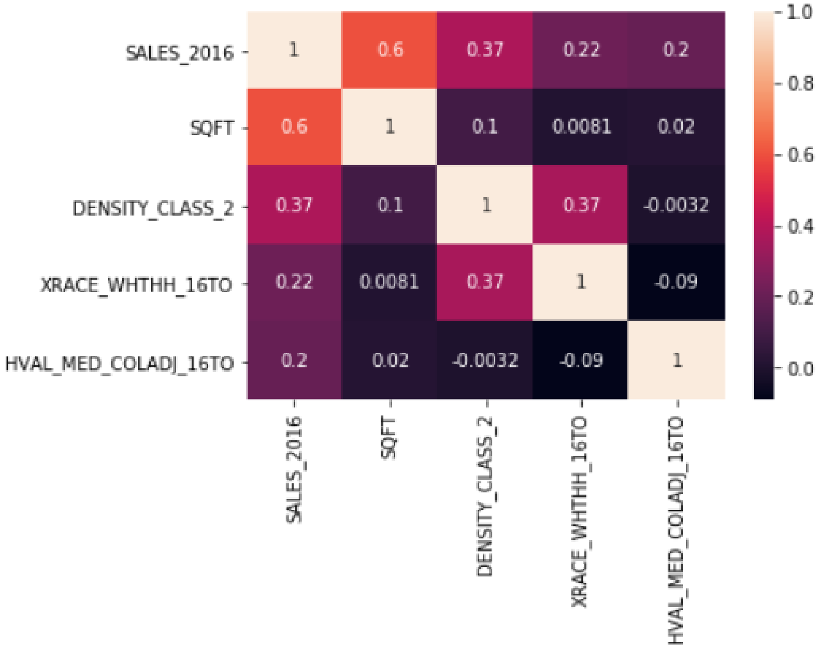
The correlation matrix above does not show strong collinearity among the groupings of variables. When predictor variables in the same regression model are correlated, individually they will not be able to predict the value of the dependent variable as accurately. Autocorrelated variables used in the model would explain some of the same variance in the dependent variable, which in turn will reduce their statistical significance in predicting the dependent variable.

Multicollinearity occurs when more than two predictors are correlated amongst each other. Multicollinearity in a regression model can cause higher standard errors. It can make it harder to explain the variation caused by the dependent variable and the confidence intervals will become wider. The simplest way to avoid multicollinearity is to remove the independent variables with high correlation values among other independent variables. It's important in the variable selection to avoid variables that are highly correlated with other predictors..

Variable Selection

Refer to Figure 4. Shown below are the variables chosen for the large store multivariate regression model. SQFT, DENSITY_CLASS_2, XRACE_WHTHH_16TO, and XHVAL_MED_COLADJ_16TO had high positive correlations with the dependent variable SALES_2016 and moderate to low correlations with each other.

Figure 4: Correlation heatmap of large store model variables



Model Technique

Regression is a type of supervised learning that is used to predict the dependent variable (y) based on historical data of the independent variables (X). Linear regression has 5 main assumptions:

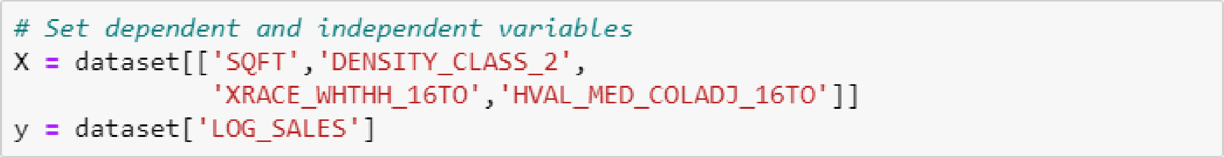
1. There's a linear relationship between the independent and dependent variables.
2. All independent variables should be multivariate normal.
3. There is no or little multicollinearity.
4. There is no auto correlation where the residuals are not independent from each other.
5. There is homoscedasticity meaning the residuals are equal across the regression line.

Multivariate Linear Regression Model Construction

To create my model, I used python in a Jupyter Notebook. You can find the code for this project here. I will also share snippets of the code in this report.

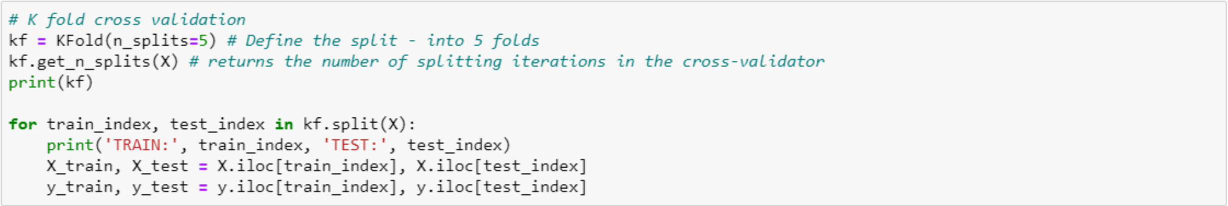
I decided to look at large stores first. Large stores will have a better chance at driving higher sales. They also have more data points and comprise majority of the 5 stores selected by Slacks on Racks, which will make predicting the dependent variable easier.

First, I set my dependent (LOG_SALES) and independent variables for large stores (SQFT, DENSITY_CLASS_2, XRACE_WHTHH_16TO, and HVAL_MED_COLAD_16TO).



```
# Set dependent and independent variables
X = dataset[['SQFT','DENSITY_CLASS_2',
             'XRACE_WHTHH_16TO','HVAL_MED_COLADJ_16TO']]
y = dataset['LOG_SALES']
```

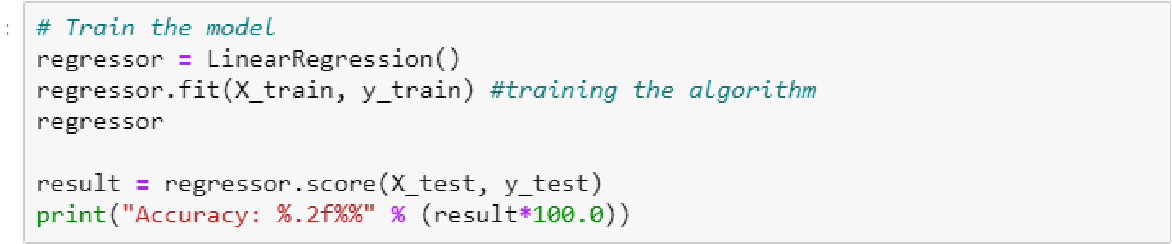
In order to train my model, I defined my validation up front. K fold cross validation is a validation process used for limited datasets. K refers to the number of groups the data will be split into. The dataset is shuffled randomly and split into 5 groups. For each group a portion of the data will be held for training and testing. Once I fit the model on the training data, the model will be evaluated with the test data.



```
# K fold cross validation
kf = KFold(n_splits=5) # Define the split - into 5 folds
kf.get_n_splits(X) # returns the number of splitting iterations in the cross-validator
print(kf)

for train_index, test_index in kf.split(X):
    print('TRAIN:', train_index, 'TEST:', test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
```

I trained my model with the LinearRegression class using the fit method. The LinearRegression class fits a linear model to minimize the residual sum of squares between the observed and predicted targets. The fit method fits the linear model based on the training data selected.



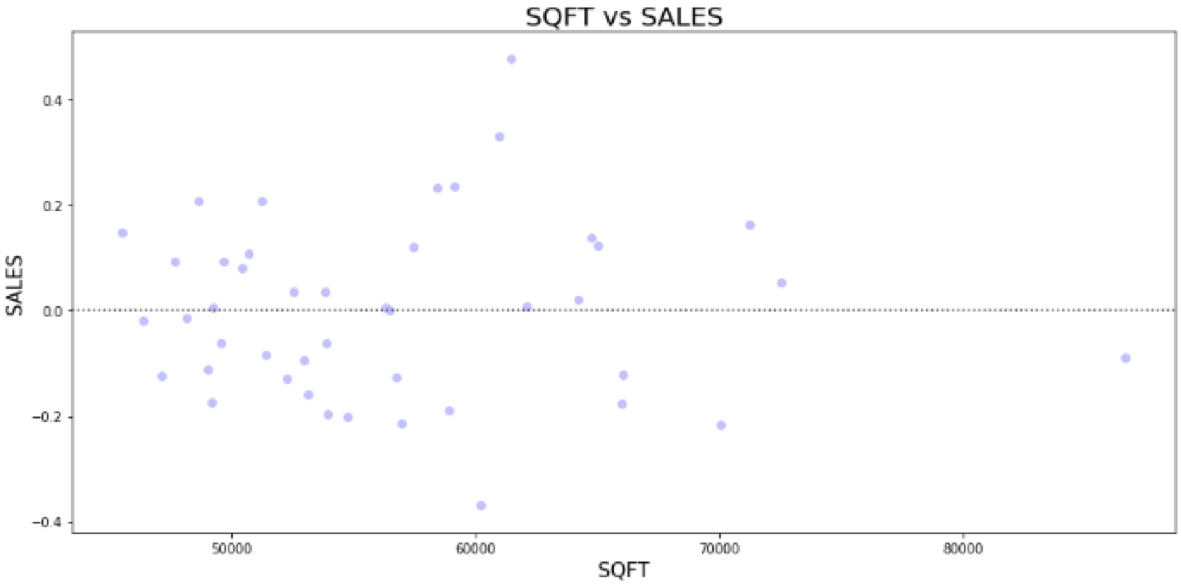
```
# Train the model
regressor = LinearRegression()
regressor.fit(X_train, y_train) #training the algorithm
regressor

result = regressor.score(X_test, y_test)
print("Accuracy: %.2f%%" % (result*100.0))
```

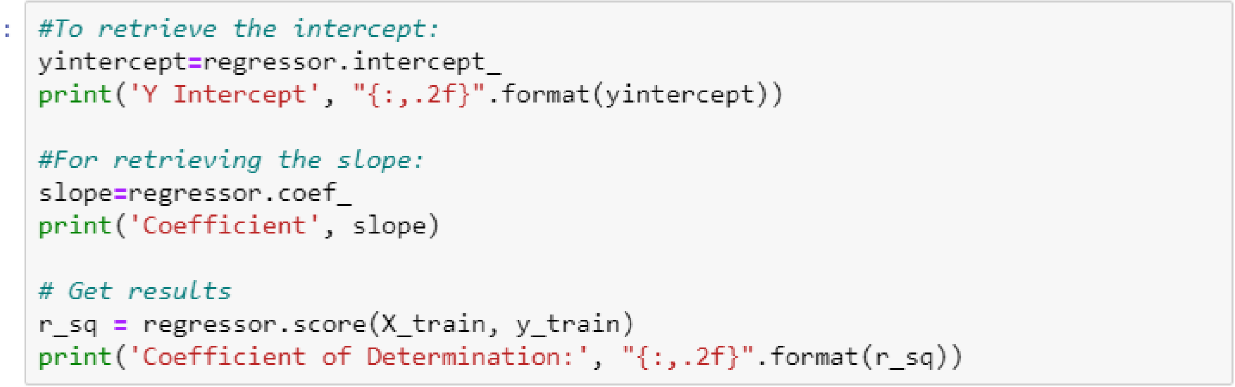
Multivariate Linear Regression Model Construction

Residuals are the difference between the observed value of the dependent variable and the predicted value. Refer to Figure 8. Below is the residual plot for the SQFT independent variable. The points are located randomly around the horizontal axis. This suggests the relationship between sales and the square foot variable is linear.

Figure 8: Residual plot for SQFT



I used the coefficient method to estimate the coefficients for the linear regression and the intercept method to find the y intercept for the regression equation.



```
#To retrieve the intercept:
yintercept=regressor.intercept_
print('Y Intercept', "{:,.2f}".format(yintercept))

#For retrieving the slope:
slope=regressor.coef_
print('Coefficient', slope)

# Get results
r_sq = regressor.score(X_train, y_train)
print('Coefficient of Determination:', "{:,.2f}".format(r_sq))
```

Multivariate Linear Regression Model Construction

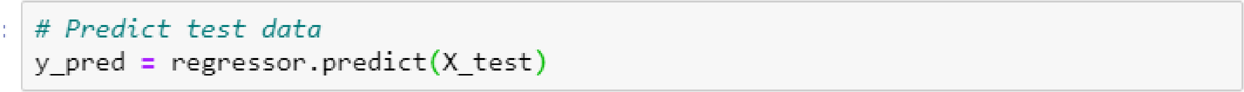
The regression equation for large stores is denoted as follows:

$$Y = 15.03 + 0.0000127(\text{SQFT}) + 0.1118375(\text{DENSITY\_CLASS\_2}) - 0.0002764(\text{XRACE\_WHTHH\_16TO}) + 0.0000007(\text{HVAL\_MED\_COLADJ})$$

The regression equation for small stores is denoted as follows:

$$Y = 14.27 + 0.00064183(\text{AWMNSGRLS\_8TO}) + 0.01251566(\text{XRACE\_ASIAHH\_16TO}) - 0.00406766(\text{XAGE\_MIDLFE3544\_16TO})$$

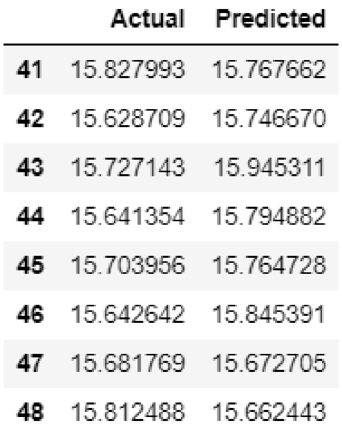
I used the predict method to predict the test data based on the trained model.



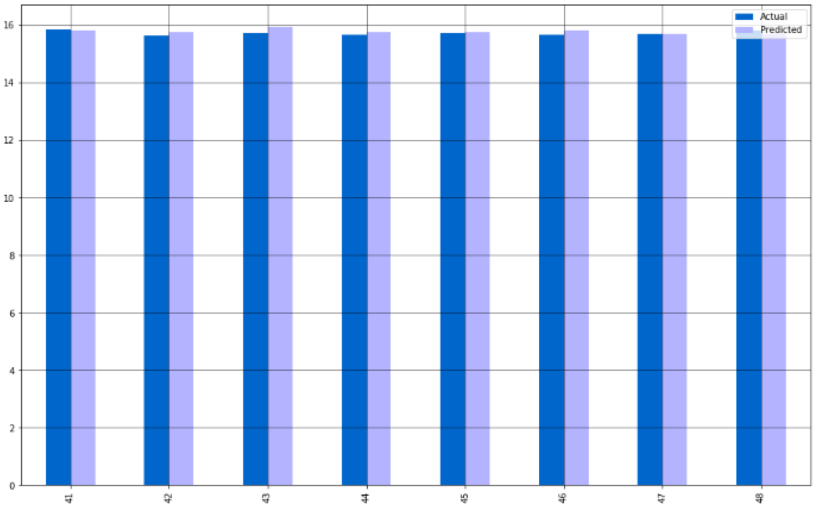
```
# Predict test data
y_pred = regressor.predict(X_test)
```

Refer to Figure 11 below. For this set of testing data, the actual and predicted values were very close.

Figure 11: Actual vs predicted values



	Actual	Predicted
41	15.827993	15.767662
42	15.628709	15.746670
43	15.727143	15.945311
44	15.641354	15.794882
45	15.703956	15.764728
46	15.642642	15.845391
47	15.681769	15.672705
48	15.812488	15.662443



The same application of the model built for large stores was applied to small stores but had differing independent variables.

Multivariate Linear Regression Model Evaluation

Large Stores

The final step is to evaluate the performance of the algorithm. In evaluating the results of the model I originally started out with 8 variables SQFT, DENSITY_CLASS_2, SCHOOL_MID_STDNTS_0_5RO, XLABOR_SRV_16TO, REGION_SA, XHVAL_L49K_16TO, XRACE_WHTHH_16TO, AND HVAL_MED_COLADJ_16TO.

The variables were chosen because they had high correlations with large store sales and moderate to low correlation amongst themselves. Upon further analysis, several of these variables had high p values and were deemed insignificant to the model. The final set of variables used for the large store model are SQFT, DENSITY_CLASS_2, XRACE_WHTHH_16TO, HVAL_MED_COLADJ_16TO.

The set of final variables suggests Slacks on Racks markets to consumers in predominately white households in towns that span across all regions. The household income adjusted for the cost of living is a factor that helps predict the sales of Slacks on Racks.

Mean Error Evaluation

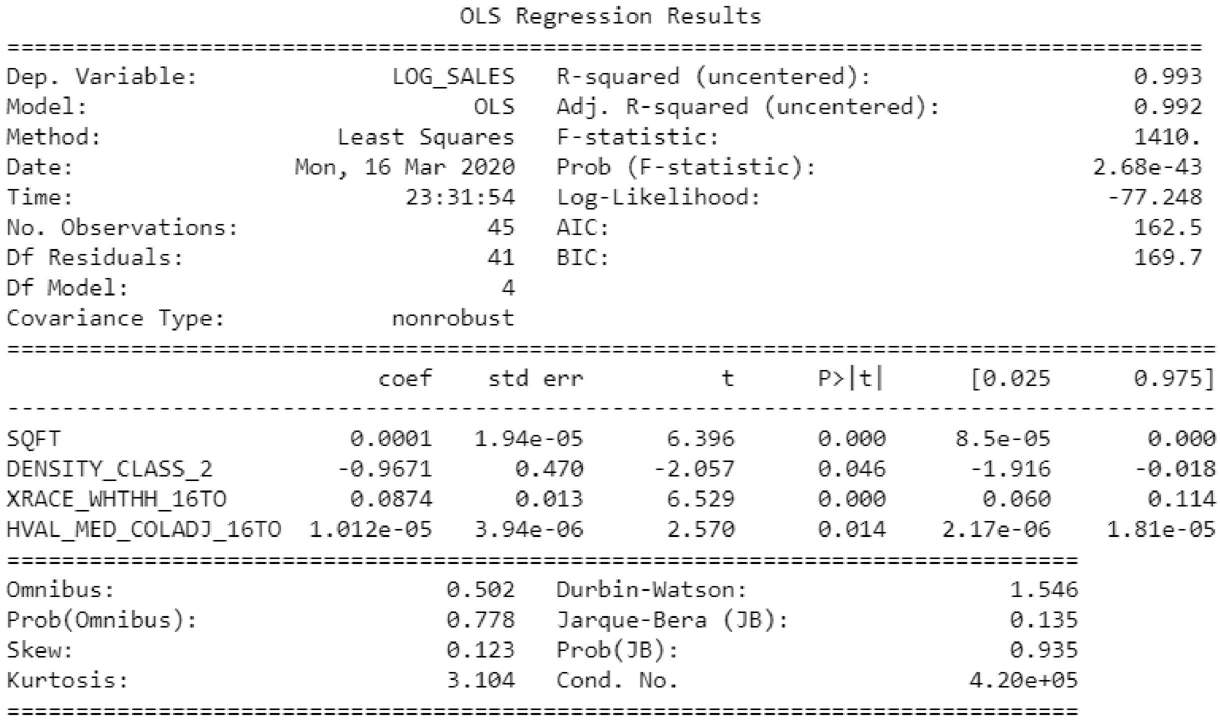
```
Mean Absolute Error: 0.12
Mean Squared Error: 0.02
Root Mean Squared Error: 0.14
```

The mean absolute error (MAE) is the absolute value of the errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. The mean squared error (MSE) tells you how close a regression line is to a set of points. The lower both these statistics measures are, the better. The results are listed below. The MAE is low as well as the MSE, meaning difference of the absolute error between the forecasted and actual values are small. There are many factors that could contribute to inaccuracy regarding high MAE and MSE values. Bad assumptions made in error, mediocre features that may've not had a high enough correlation to the dependent variable, having possible multicollinearity, or having too little data could be a few reasons for inaccuracy.

Multivariate Linear Regression Model Evaluation

For the subsequent sections refer to Figure 9. Below are the summary results of the ordinary least squares method for large stores.

Figure 9: Regression results for large stores



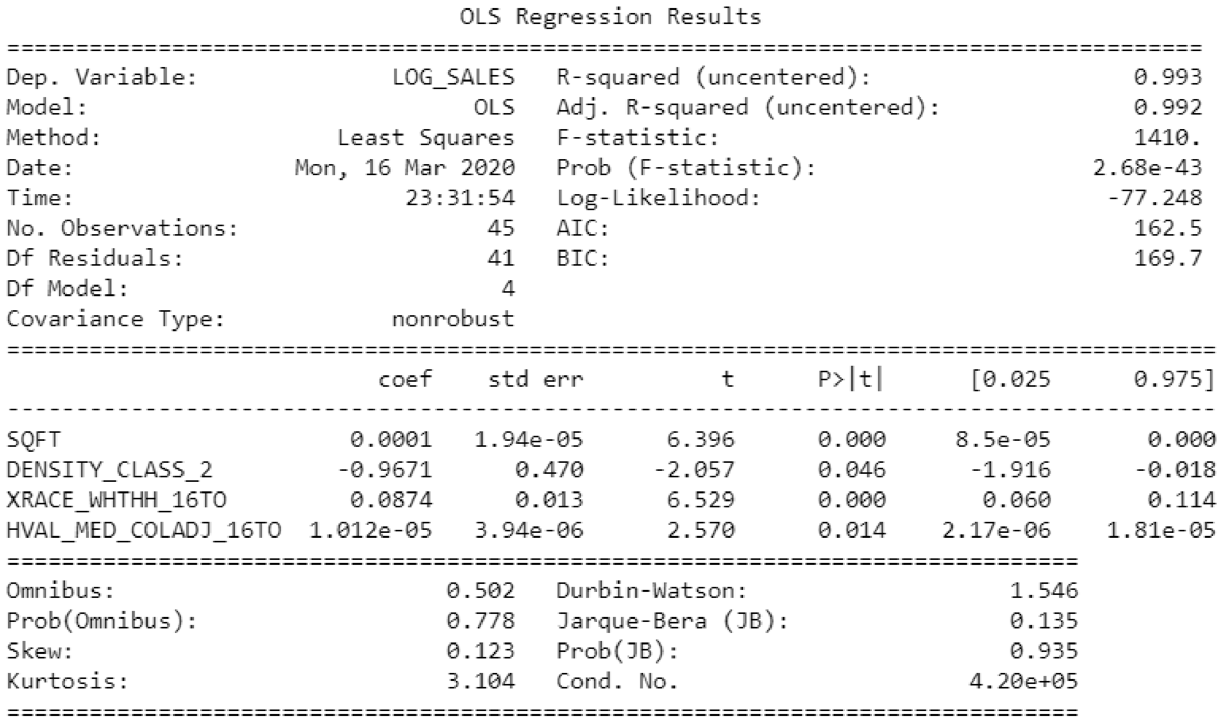
```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              LOG_SALES   R-squared (uncentered):                   0.993
Model:                            OLS   Adj. R-squared (uncentered):              0.992
Method:                 Least Squares   F-statistic:                              1410.
Date:                Mon, 16 Mar 2020   Prob (F-statistic):                    2.68e-43
Time:                        23:31:54   Log-Likelihood:                         -77.248
No. Observations:                  45   AIC:                                      162.5
Df Residuals:                      41   BIC:                                      169.7
Df Model:                           4
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
SQFT                     0.0001   1.94e-05      6.396      0.000     8.5e-05       0.000
DENSITY_CLASS_2         -0.9671      0.470     -2.057      0.046      -1.916      -0.018
XRACE_WHTHH_16TO         0.0874      0.013      6.529      0.000       0.060       0.114
HVAL_MED_COLADJ_16TO  1.012e-05   3.94e-06      2.570      0.014    2.17e-06    1.81e-05
==============================================================================
Omnibus:                        0.502   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.778   Jarque-Bera (JB):                0.135
Skew:                           0.123   Prob(JB):                        0.935
Kurtosis:                       3.104   Cond. No.                     4.20e+05
==============================================================================
```

R-Squared and Adjust R-Squared Evaluation

R-squared, also known as the coefficient of determination, is the measure of how well the regression line fits the data. Higher R-squared values represent smaller differences between the observed data and the fitted values. The R-squared value is always between 0% and 100%. 0% represent a model that does not explain any of the variation in the dependent variable around it's mean. 100% represents a model that explains all the variation in the dependent variable. The R-squared value is very high, .993. The difference between the fitted values and the observed values is very small and almost always landing on the actual regression line. R-squared should not be solely used to determine whether the coefficient estimates are biased. The R-squared values doesn't indicate if a regression model has an adequate fit to the data. There are many different circumstances that can inflate the R-squared value like overfitting. An overfit model is when the model fits all the scenarios of the sample trained. The adjusted R-squared adjusts the R-squared statistics based on the number of independent variables used. In this case there were 4 independent variables used. Again, the adjusted R-squared value is very high meaning the difference between the observed values and fitted values is small.

Multivariate Linear Regression Model Evaluation

Figure 9: Regression results for large stores



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                LOG_SALES   R-squared (uncentered):                  0.993
Model:                              OLS   Adj. R-squared (uncentered):             0.992
Method:                   Least Squares   F-statistic:                             1410.
Date:                  Mon, 16 Mar 2020   Prob (F-statistic):                   2.68e-43
Time:                          23:31:54   Log-Likelihood:                        -77.248
No. Observations:                    45   AIC:                                     162.5
Df Residuals:                        41   BIC:                                     169.7
Df Model:                             4
Covariance Type:              nonrobust
========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
SQFT                    0.0001   1.94e-05      6.396      0.000     8.5e-05       0.000
DENSITY_CLASS_2        -0.9671      0.470     -2.057      0.046      -1.916      -0.018
XRACE_WHTHH_16TO        0.0874      0.013      6.529      0.000       0.060       0.114
HVAL_MED_COLADJ_16TO 1.012e-05   3.94e-06      2.570      0.014    2.17e-06    1.81e-05
==============================================================================
Omnibus:                        0.502   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.778   Jarque-Bera (JB):                0.135
Skew:                           0.123   Prob(JB):                        0.935
Kurtosis:                       3.104   Cond. No.                     4.20e+05
==============================================================================
```

F Statistic Evaluation

The F statistic is the calculated ratio of the mean squared error of the model and the mean squared error of residuals. It is used in combination with the p value. The combination of the F statistic and p value can determine if the results are significant. A low p value, less than alpha, indicates that the coefficient is a meaningful addition to the regression model. A larger p value suggests that changes in the predictor are not associated with changes in the dependent variable.

Durbin Watson Evaluation

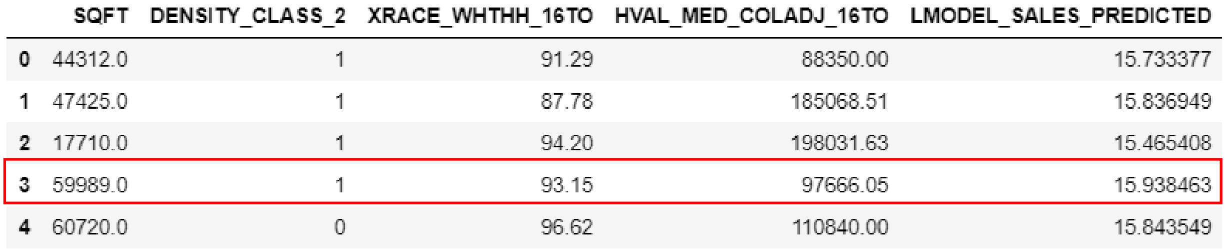
The Durbin Watson statistic tests if the residuals are linearly auto correlated. Durbin Watson values from 1.5 to 2.5 indicate little to no autocorrelation. Linear regression requires that there is little to no autocorrelations. Autocorrelations appear when the residuals are not independent from one another. The Durban Watson test for the regression model is on the upper limit of the spread. There could be potential autocorrelation but is still within the limits.

Multivariate Linear Regression Model Evaluation

Results

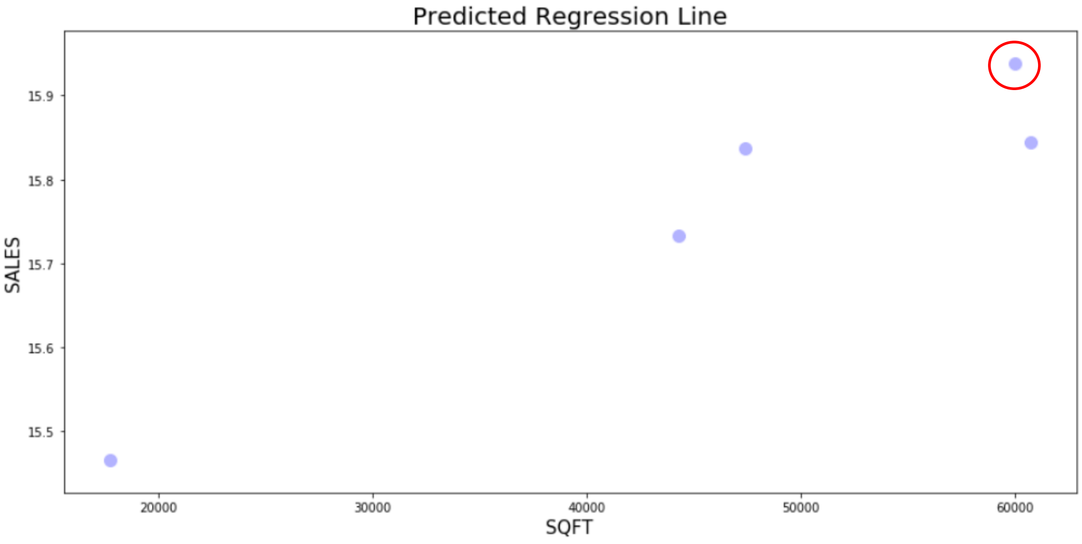
Refer to Figure 10 and Figure 12. The 5 potential stores and their predictions are below. The store with the highest sales is store number 4 with 59,989 square feet. Store 4 is located in a town and has a high percentage of white households in the area, within a 16-minute drive. While although this store looks like a winner in these graphs, this store compared to their previous locations is middle of the pack. It will be a good investment into the company brand, but it will not become their money maker.

Figure 10: Coefficients and predicted values for the 5 potential stores - large store model



	SQFT	DENSITY_CLASS_2	XRACE_WHTHH_16TO	HVAL_MED_COLADJ_16TO	LMODEL_SALES_PREDICTED
0	44312.0	1	91.29	88350.00	15.733377
1	47425.0	1	87.78	185068.51	15.836949
2	17710.0	1	94.20	198031.63	15.465408
3	59989.0	1	93.15	97666.05	15.938463
4	60720.0	0	96.62	110840.00	15.843549

Figure 12: Predicted sales for the 5 potential stores plotted - large store model



Multivariate Linear Regression Model Evaluation

Small Stores

The final step is to evaluate the performance of the algorithm. In evaluating the results of the model, I originally started out with 44 variables. I ended up using the AWMNSGRLS_8TO, XRACE_ASIAHH_16TO, and XAGE_MIDLFE3544_16TO in the final model. The omitted variables were chosen originally because they had high correlations with small store sales. However, there were many variables that are autocorrelated. Upon further analysis, all these variables had high p values and were deemed insignificant to the model.

The set of final variables suggests Slacks on Racks markets to consumers who live in predominately Asian households and spends money on apparel and services for women and girls. Their consumers may be men in their 30's to 40's.

Mean Error Evaluation

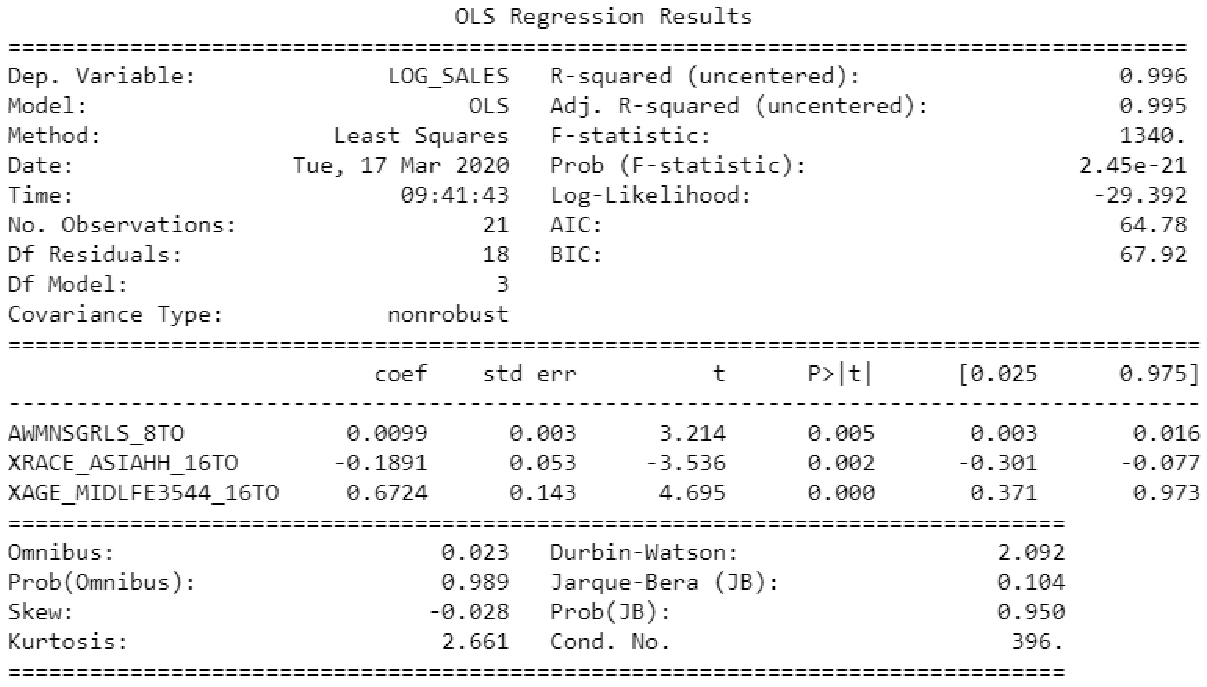
```
Mean Absolute Error: 0.06
Mean Squared Error: 0.00
Root Mean Squared Error: 0.07
```

The mean absolute error (MAE) is the absolute value of the errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. The mean squared error (MSE) tells you how close a regression line is to a set of points. The lower both these statistics measures are, the better. The results are listed above. The MAE and MSE are both low, meaning the difference of the absolute error between the forecasted and actual values is minimal. In comparison to the large store MAE and MSE results, the small store model is better at reducing the error between actual and predicted values.

Multivariate Linear Regression Model Evaluation

For the subsequent sections refer to Figure 14. Below are the summary results of the ordinary least squares method for large stores.

Figure 14: Regression results for small stores



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              LOG_SALES   R-squared (uncentered):                   0.996
Model:                            OLS   Adj. R-squared (uncentered):              0.995
Method:                 Least Squares   F-statistic:                              1340.
Date:                Tue, 17 Mar 2020   Prob (F-statistic):                    2.45e-21
Time:                        09:41:43   Log-Likelihood:                         -29.392
No. Observations:                  21   AIC:                                      64.78
Df Residuals:                      18   BIC:                                      67.92
Df Model:                           3
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
AWMNSGRLS_8TO            0.0099      0.003      3.214      0.005       0.003       0.016
XRACE_ASIAHH_16TO       -0.1891      0.053     -3.536      0.002      -0.301      -0.077
XAGE_MIDLFE3544_16TO     0.6724      0.143      4.695      0.000       0.371       0.973
==============================================================================
Omnibus:                        0.023   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.989   Jarque-Bera (JB):                0.104
Skew:                          -0.028   Prob(JB):                        0.950
Kurtosis:                       2.661   Cond. No.                         396.
==============================================================================
```

R-Squared and Adjust R-Squared Evaluation

The R-squared value is very high, .993. This tells us the differences between the fitted values and the observed values is small. As mentioned previously, the R-squared value should not be solely used to determine whether the coefficient estimates are biased. There are many different circumstances that can inflate the R-squared value like overfitting. The adjusted R-squared value is very high meaning the difference between the observed values and fitted values is small.

F Statistic Evaluation

The F statistic is the calculated ratio of the mean squared error of the model and the mean squared error of residuals. It is used in combination with the p value. The combination of the F statistic and p value can determine if the results are significant. A low p value, less than alpha, indicates that the coefficient is a meaningful addition to the regression model. A larger p value suggests that changes in the predictor are not associated with changes in the dependent variable. The p values for AWMNSGRLS_8TO, XRACE_ASIAHH_16TO, and XAGE_MIDLFE3544_16TO are significant coupled with a large F statistic.

Multivariate Linear Regression Model Evaluation

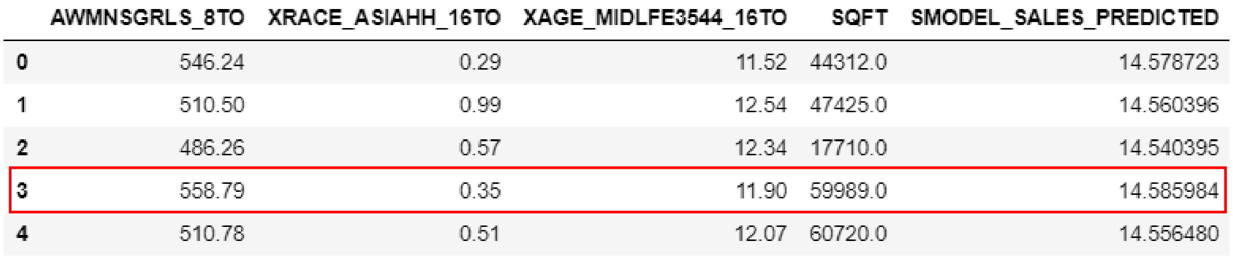
Durbin Watson Evaluation

The Durbin Watson statistic tests if the residuals are linearly auto correlated. Durbin Watson values from 1.5 to 2.5 indicate little to no autocorrelation. The Durban Watson test for the regression model is midway of the acceptable range.

Results

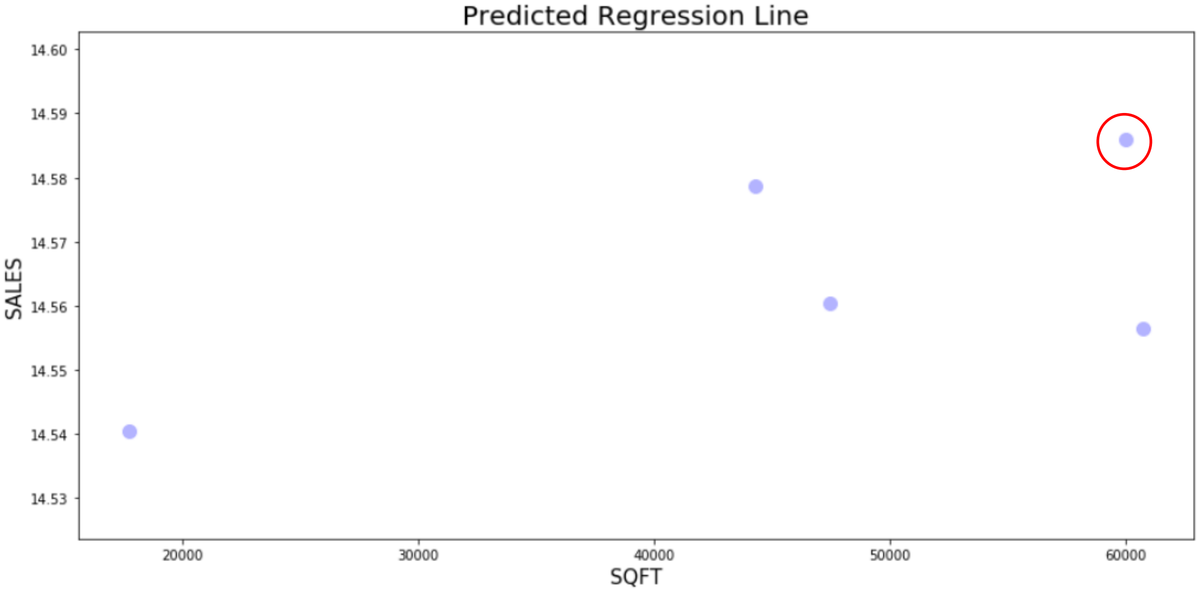
Refer to Figure 15 and Figure 16. The 5 potential stores and their predictions are below. The store with the highest sales is store number 4 with 59,989 square feet. Store 4 has the highest average of consumer expenditures of apparel and services for women and girls. Store 4 also has a decent percentage of predominately Asian households and middle-aged men in the area. Store 4 was also chosen as having the highest predicted sales of the large store model. While although this store looks like a winner in these graphs, this store compared to their previous locations is middle of the pack. It will be a good investment into the company brand, but it will not become their money maker.

Figure 15: Coefficients and predicted values for the 5 potential stores – small store model



	AWMNSGRLS_8TO	XRACE_ASIAHH_16TO	XAGE_MIDLFE3544_16TO	SQFT	SMODEL_SALES_PREDICTED
0	546.24	0.29	11.52	44312.0	14.578723
1	510.50	0.99	12.54	47425.0	14.560396
2	486.26	0.57	12.34	17710.0	14.540395
3	558.79	0.35	11.90	59989.0	14.585984
4	510.78	0.51	12.07	60720.0	14.556480

Figure 16: Predicted sales for the 5 potential stores plotted – small store model



Conclusion

In conclusion, the small and large stores behaved much differently from one another. Small stores were more affected by their communities characteristics. Age, race, income, value of homes in the area, and buying habits were highly correlated variables pertaining to the sales in small stores. Large stores were more affected by the region and square footage of the establishment. The larger the store the better they performed. The large store cluster had a more linear plot, while the small store cluster had more of a blob shape. The blob factor may've attributed to the complexity in variable selection for the small store model. To add to that complexity, there was a lack of small store data points and plenty of autocorrelation.

From the results, the small store model had a harder time predicting the sales value of the large stores. The large store model had the same issue. It had a hard time predicting the sales value of the small stores. The large store model should only be used to predict values of large store. Likewise, the small store model should only be used to predict the sales of small stores. With that in mind, store 4 with 59,989 square feet in Pennsylvania is the store I would suggest Slacks on Racks to open. While although it's not the largest store, it does perform the best. As stated previously, store 4 will not be the best or worst performer for Slack on Racks. It sits in the middle of the pack among their other locations and sales for 2016.

Resources

https://github.com/marissamckee/Retail_Location_Regression

Appendix

Figure 1: Sales in each state shown by density class

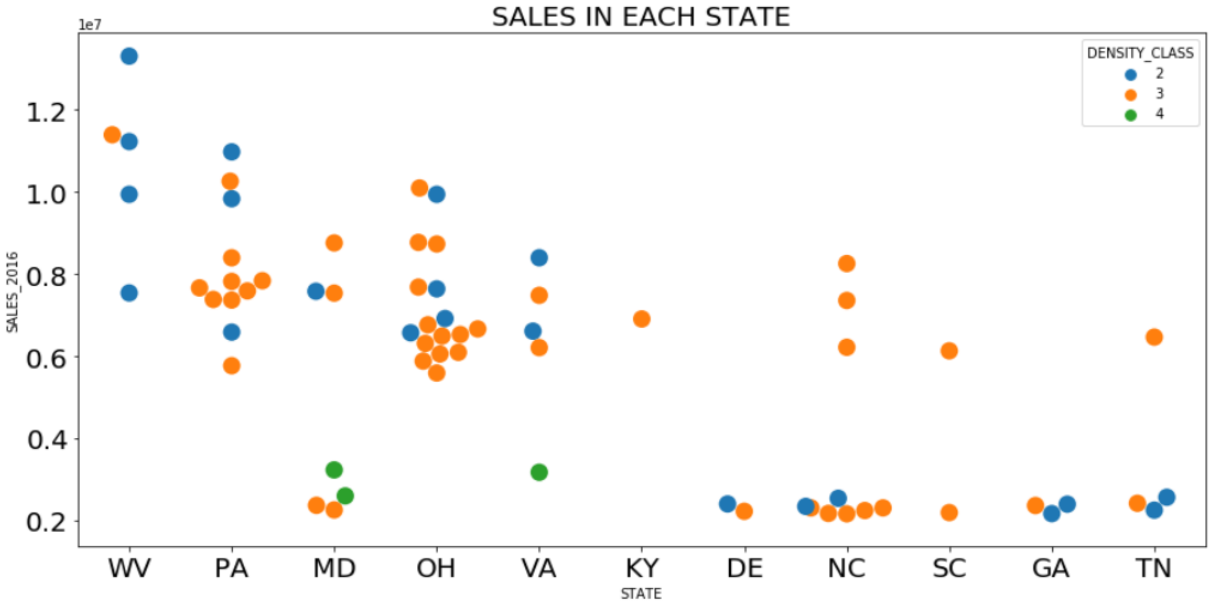


Figure 2: Scatter plot SQFT vs sales

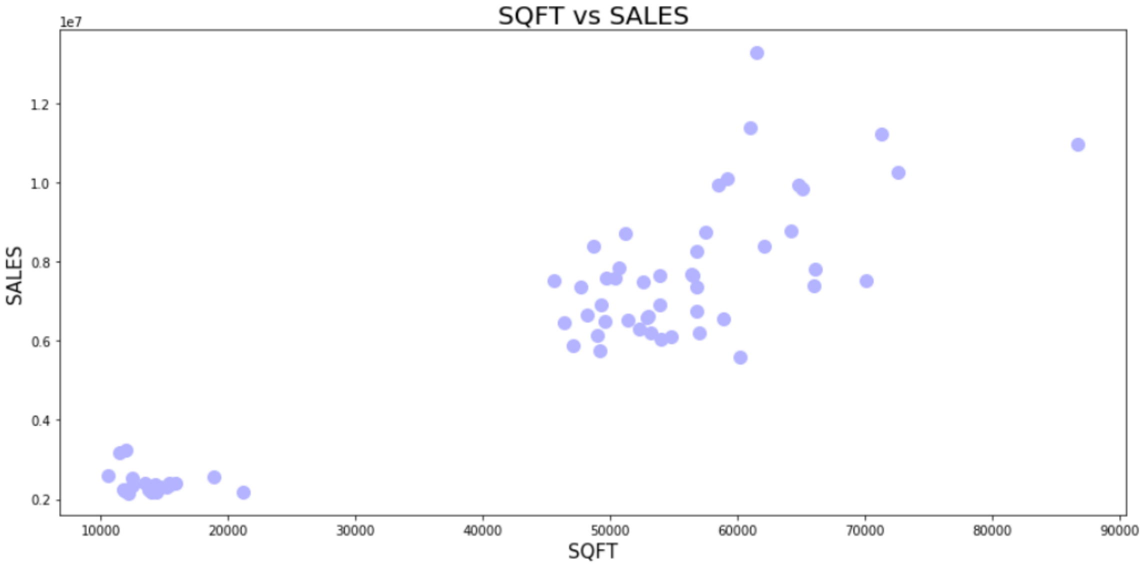


Figure 3: Correlation heatmap of small store model variables

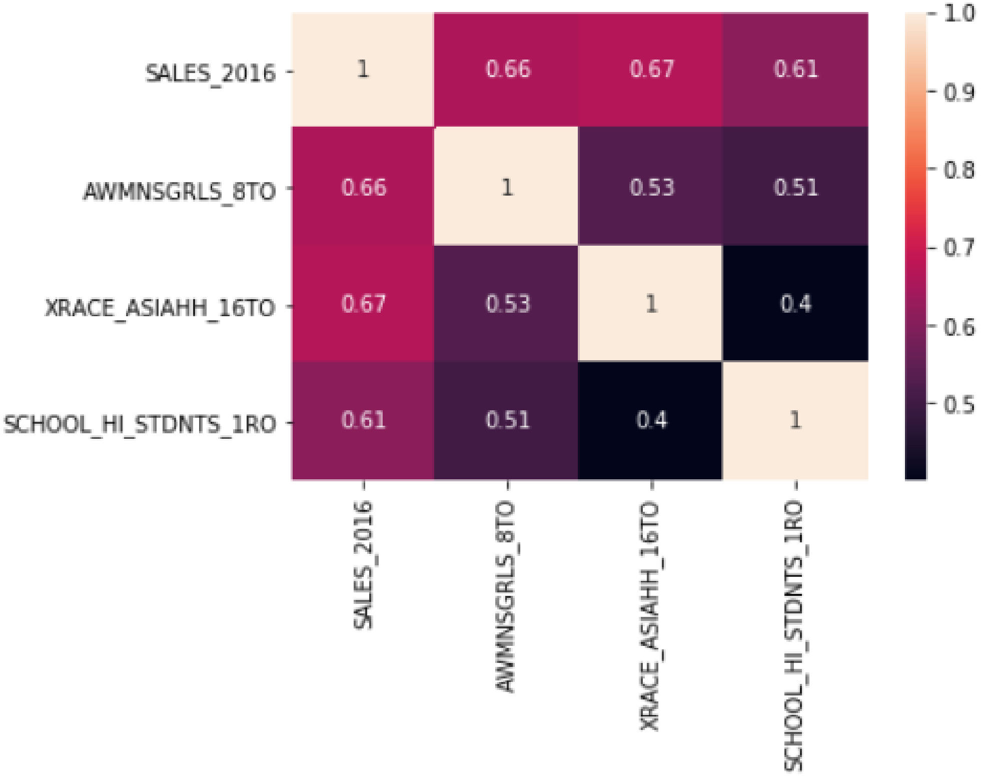


Figure 4: Correlation heatmap of large store model variables

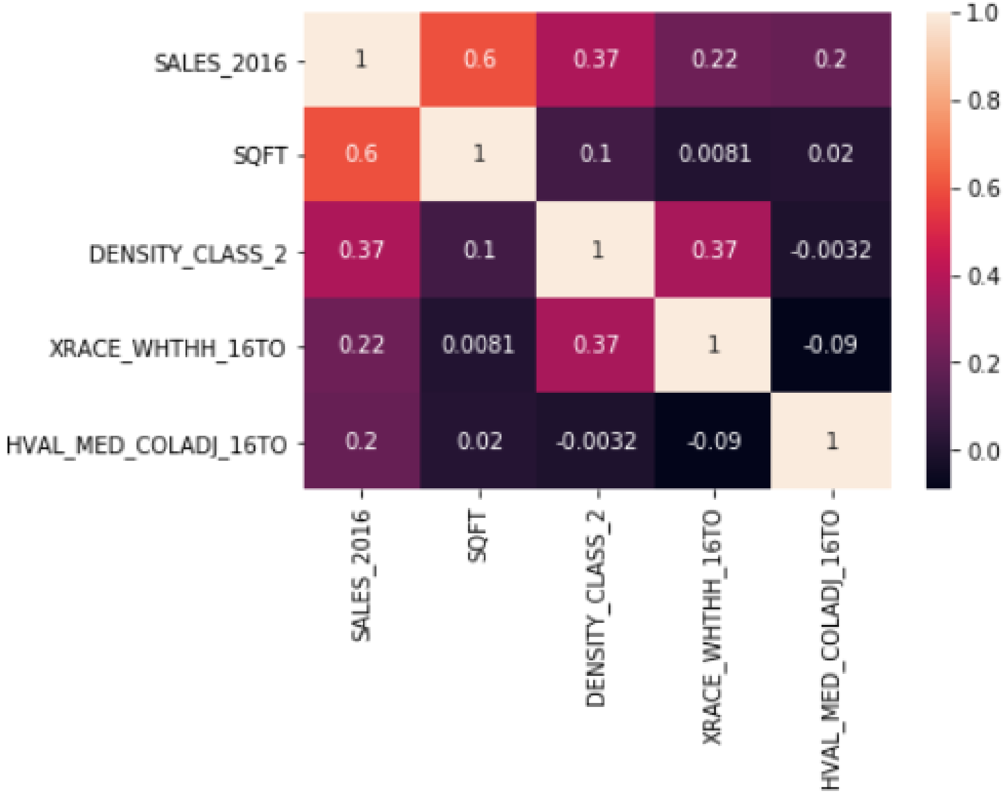


Figure 5: Scatter plot of sales vs square footage for large stores

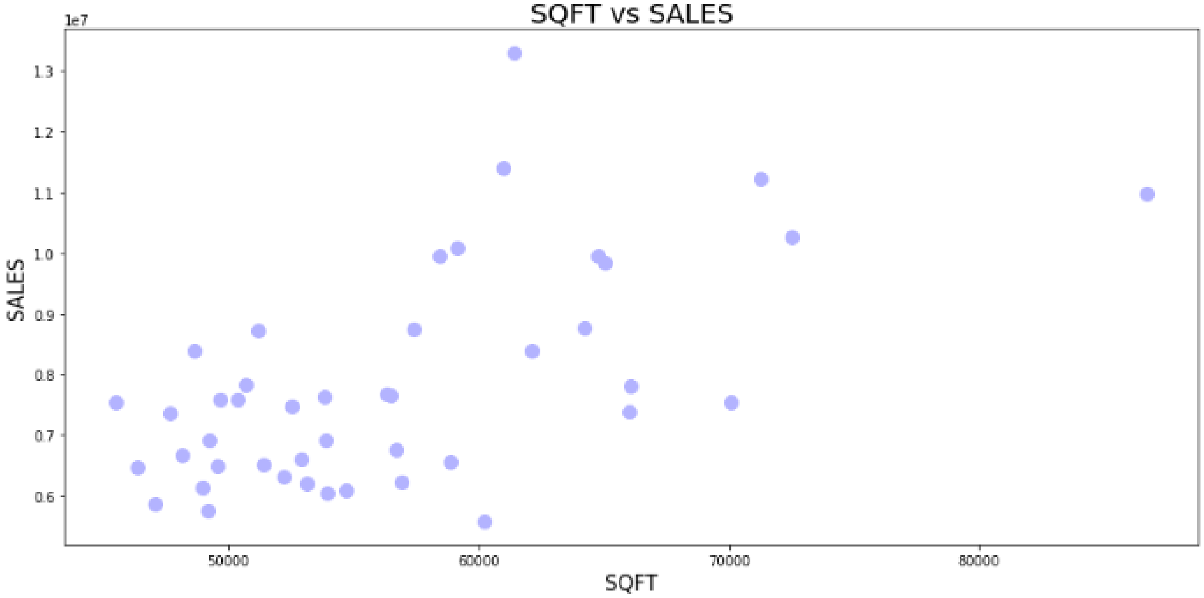


Figure 6: Sales histogram for large stores

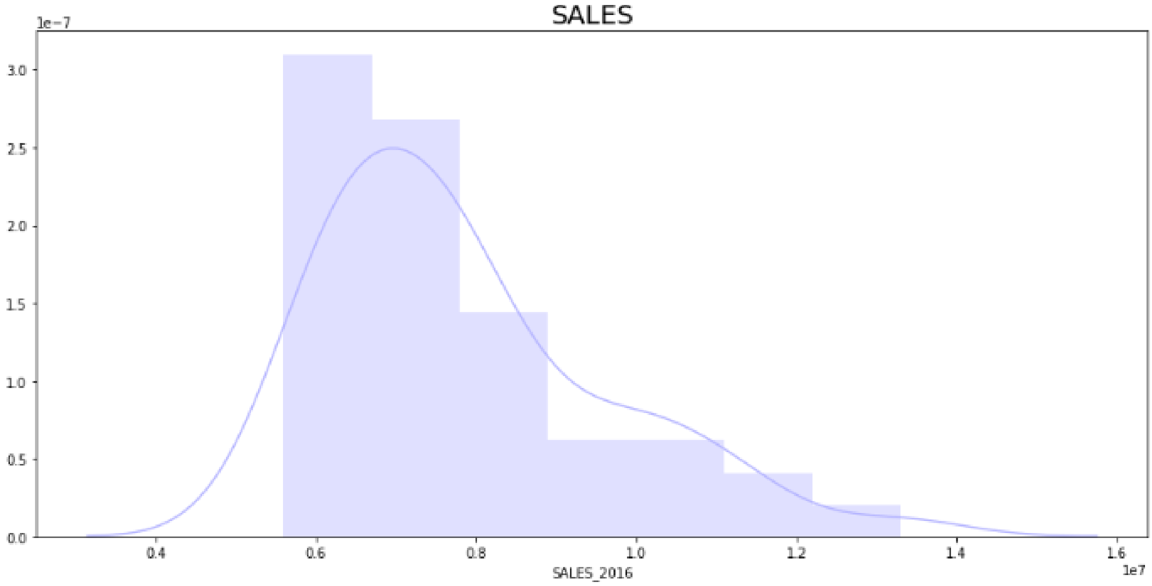


Figure 7: Lognormal sales histogram for large stores

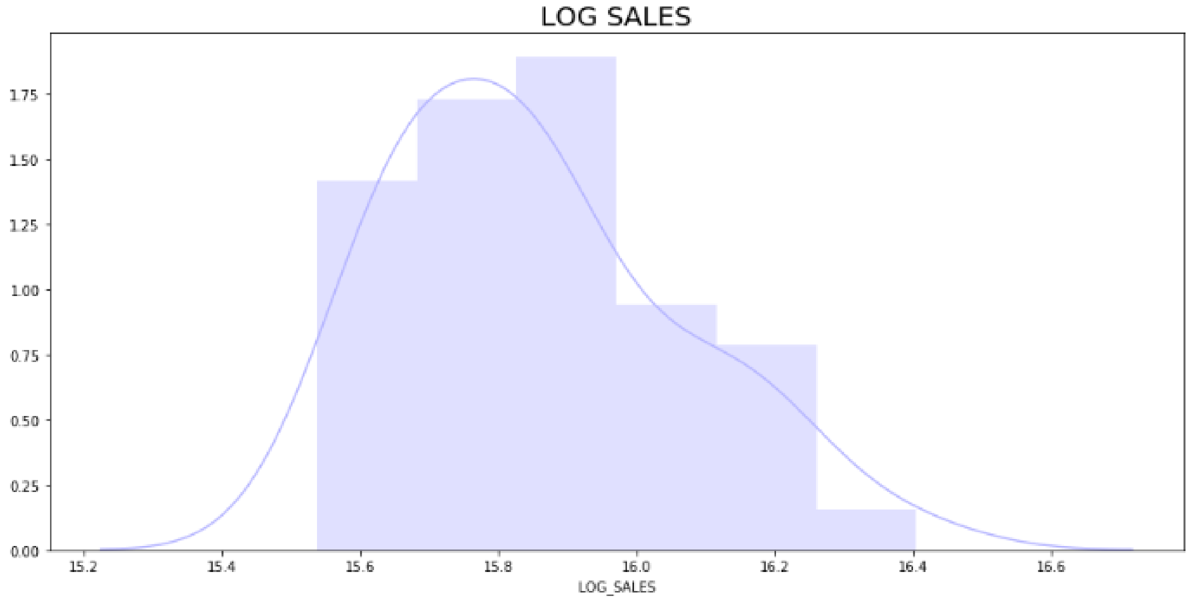


Figure 8: Residual plot for SQFT

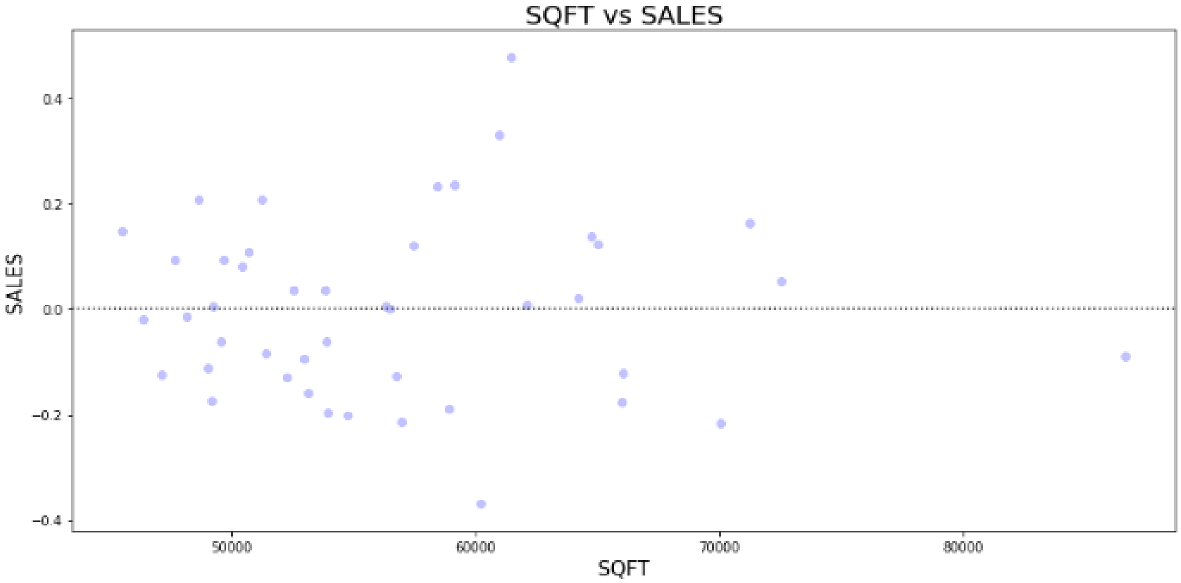
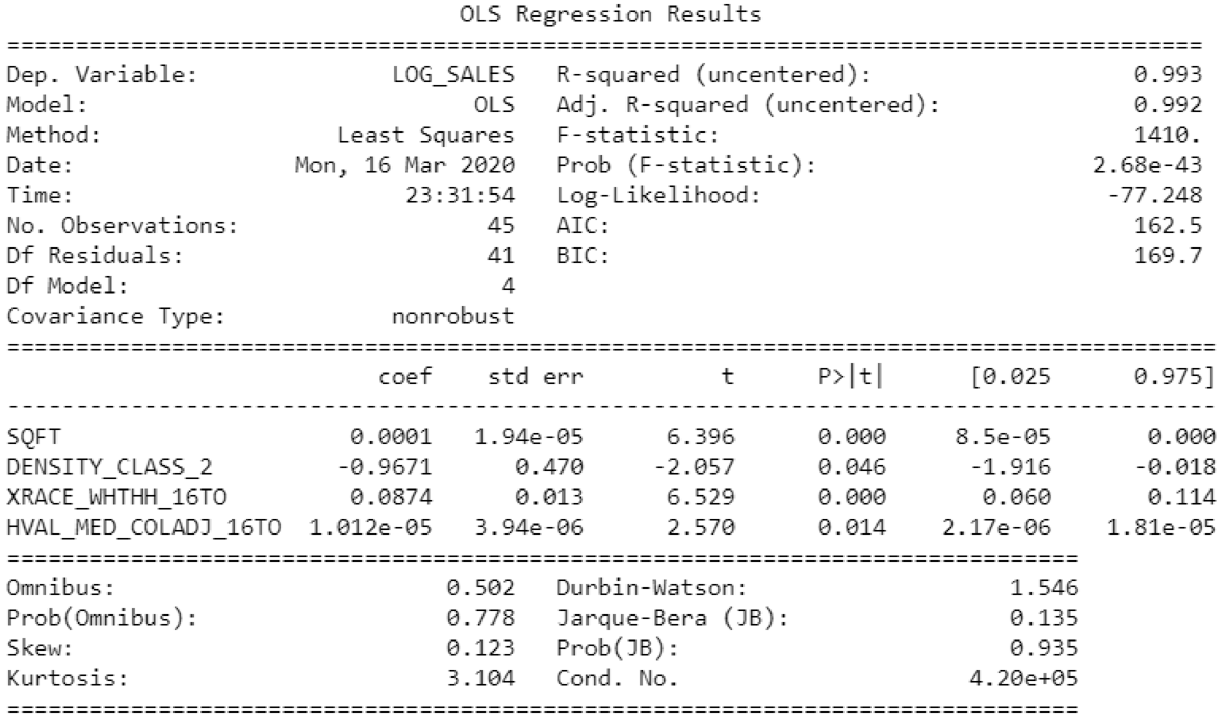
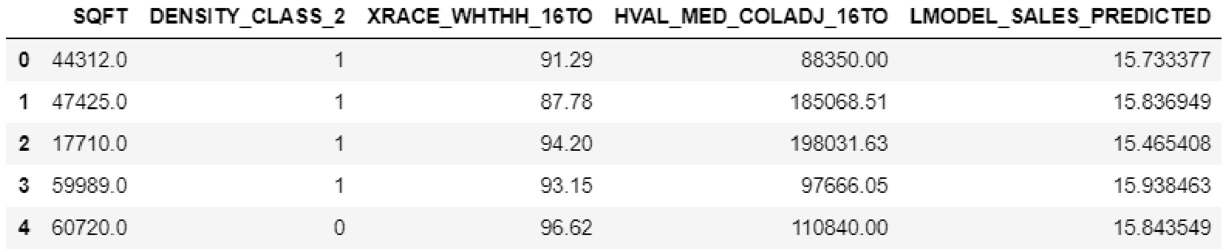


Figure 9: Regression results for large stores



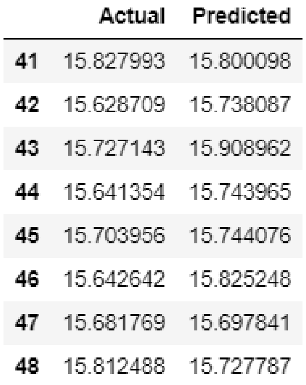
```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              LOG_SALES   R-squared (uncentered):                   0.993
Model:                            OLS   Adj. R-squared (uncentered):              0.992
Method:                 Least Squares   F-statistic:                              1410.
Date:                Mon, 16 Mar 2020   Prob (F-statistic):                    2.68e-43
Time:                        23:31:54   Log-Likelihood:                         -77.248
No. Observations:                  45   AIC:                                      162.5
Df Residuals:                      41   BIC:                                      169.7
Df Model:                           4
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
SQFT                     0.0001   1.94e-05      6.396      0.000     8.5e-05       0.000
DENSITY_CLASS_2         -0.9671      0.470     -2.057      0.046      -1.916      -0.018
XRACE_WHTHH_16TO         0.0874      0.013      6.529      0.000       0.060       0.114
HVAL_MED_COLADJ_16TO  1.012e-05   3.94e-06      2.570      0.014    2.17e-06    1.81e-05
==============================================================================
Omnibus:                        0.502   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.778   Jarque-Bera (JB):                0.135
Skew:                           0.123   Prob(JB):                        0.935
Kurtosis:                       3.104   Cond. No.                     4.20e+05
==============================================================================
```

Figure 10: Coefficients and predicted values for the 5 potential stores - large store model



	SQFT	DENSITY_CLASS_2	XRACE_WHTHH_16TO	HVAL_MED_COLADJ_16TO	LMODEL_SALES_PREDICTED
0	44312.0	1	91.29	88350.00	15.733377
1	47425.0	1	87.78	185068.51	15.836949
2	17710.0	1	94.20	198031.63	15.465408
3	59989.0	1	93.15	97666.05	15.938463
4	60720.0	0	96.62	110840.00	15.843549

Figure 11: Actual vs predicted values



	Actual	Predicted
41	15.827993	15.800098
42	15.628709	15.738087
43	15.727143	15.908962
44	15.641354	15.743965
45	15.703956	15.744076
46	15.642642	15.825248
47	15.681769	15.697841
48	15.812488	15.727787

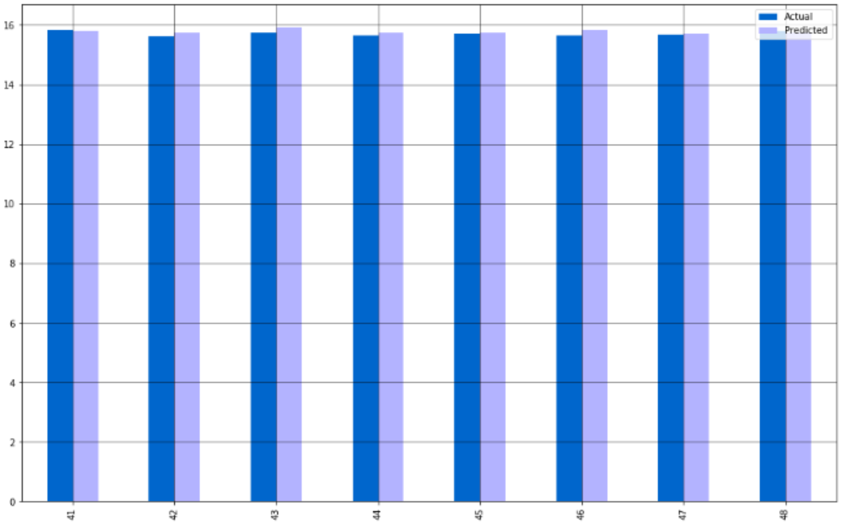


Figure 12: Predicted sales for the 5 potential stores plotted - large store model

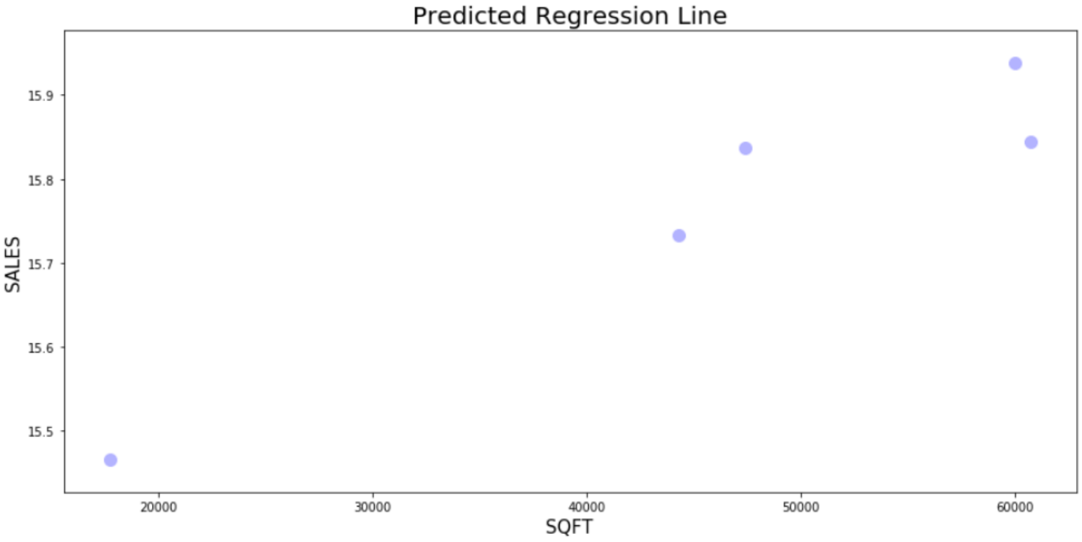
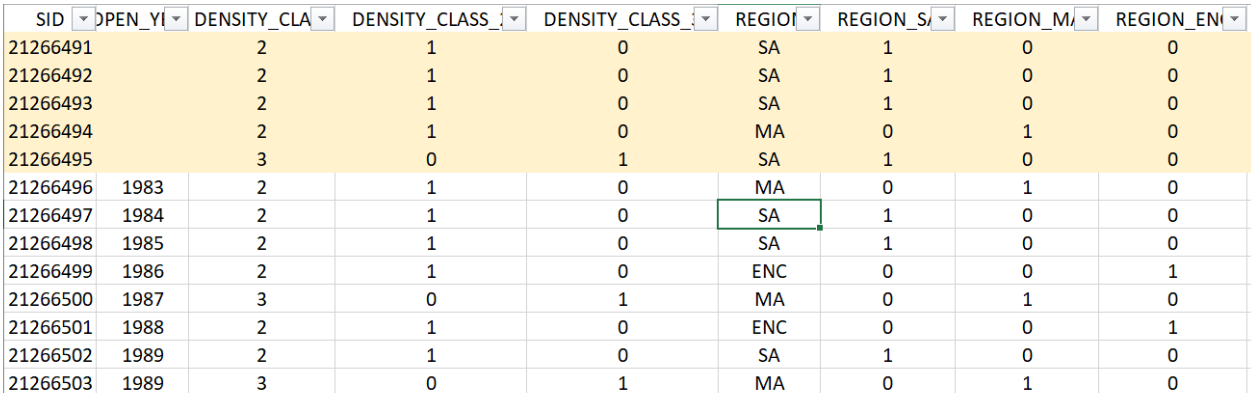
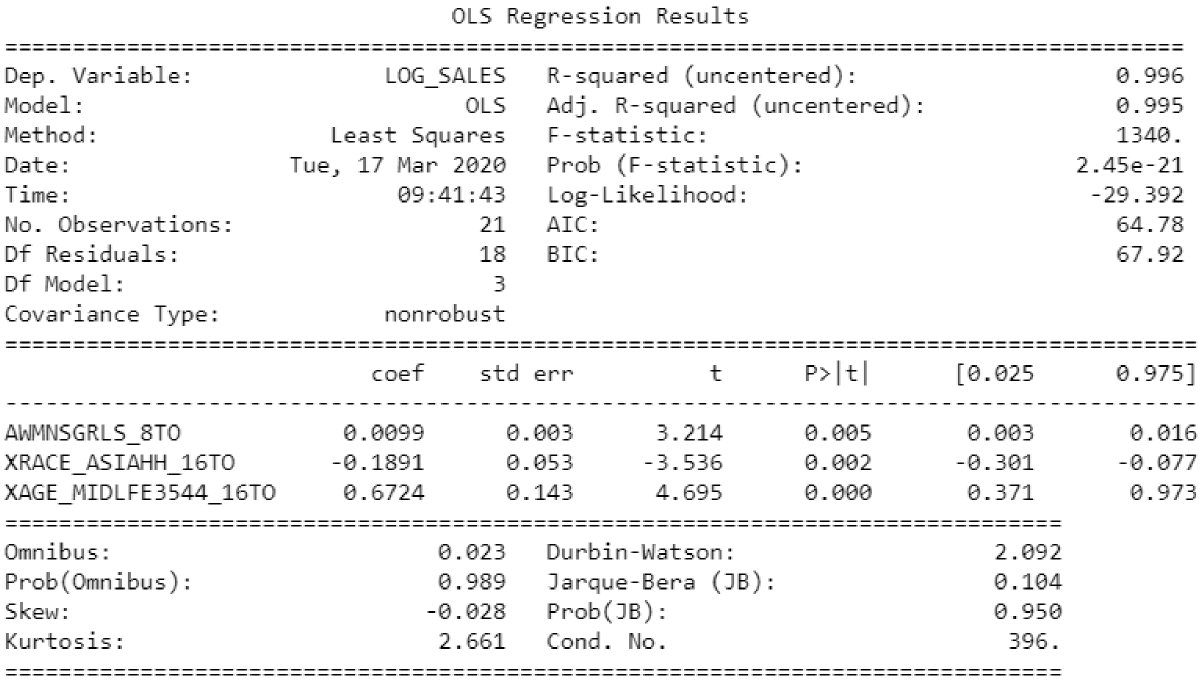


Figure 13: Dummy variable creation



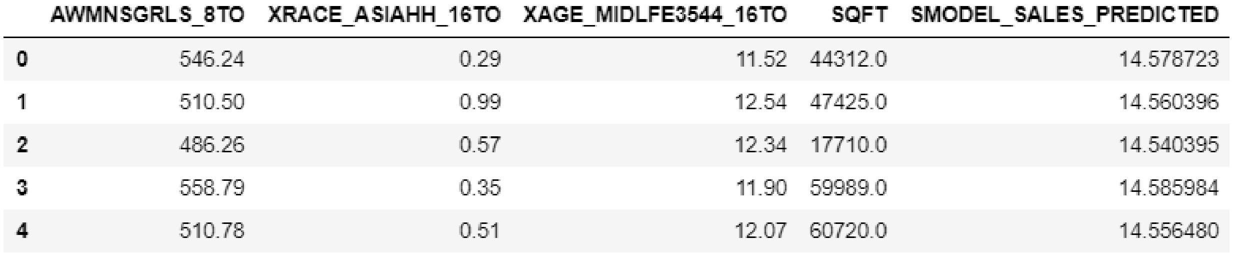
SID	)PEN_YI	DENSITY_CLA	DENSITY_CLASS_:	DENSITY_CLASS_:	REGIOI	REGION_S/	REGION_M/	REGION_EN(
21266491		2	1	0	SA	1	0	0
21266492		2	1	0	SA	1	0	0
21266493		2	1	0	SA	1	0	0
21266494		2	1	0	MA	0	1	0
21266495		3	0	1	SA	1	0	0
21266496	1983	2	1	0	MA	0	1	0
21266497	1984	2	1	0	SA	1	0	0
21266498	1985	2	1	0	SA	1	0	0
21266499	1986	2	1	0	ENC	0	0	1
21266500	1987	3	0	1	MA	0	1	0
21266501	1988	2	1	0	ENC	0	0	1
21266502	1989	2	1	0	SA	1	0	0
21266503	1989	3	0	1	MA	0	1	0

Figure 14: Regression results for small stores



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              LOG_SALES   R-squared (uncentered):                   0.996
Model:                            OLS   Adj. R-squared (uncentered):              0.995
Method:                 Least Squares   F-statistic:                              1340.
Date:                Tue, 17 Mar 2020   Prob (F-statistic):                    2.45e-21
Time:                        09:41:43   Log-Likelihood:                         -29.392
No. Observations:                  21   AIC:                                      64.78
Df Residuals:                      18   BIC:                                      67.92
Df Model:                           3
Covariance Type:            nonrobust
========================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
AWMNSGRLS_8TO         0.0099      0.003      3.214      0.005       0.003       0.016
XRACE_ASIAHH_16TO    -0.1891      0.053     -3.536      0.002      -0.301      -0.077
XAGE_MIDLFE3544_16TO  0.6724      0.143      4.695      0.000       0.371       0.973
==============================================================================
Omnibus:                        0.023   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.989   Jarque-Bera (JB):                0.104
Skew:                          -0.028   Prob(JB):                        0.950
Kurtosis:                       2.661   Cond. No.                         396.
==============================================================================
```

Figure 15: Coefficients and predicted values for the 5 potential stores – small store model



	AWMNSGRLS_8TO	XRACE_ASIAHH_16TO	XAGE_MIDLFE3544_16TO	SQFT	SMODEL_SALES_PREDICTED
0	546.24	0.29	11.52	44312.0	14.578723
1	510.50	0.99	12.54	47425.0	14.560396
2	486.26	0.57	12.34	17710.0	14.540395
3	558.79	0.35	11.90	59989.0	14.585984
4	510.78	0.51	12.07	60720.0	14.556480

Figure 16: Predicted sales for the 5 potential stores plotted – small store model

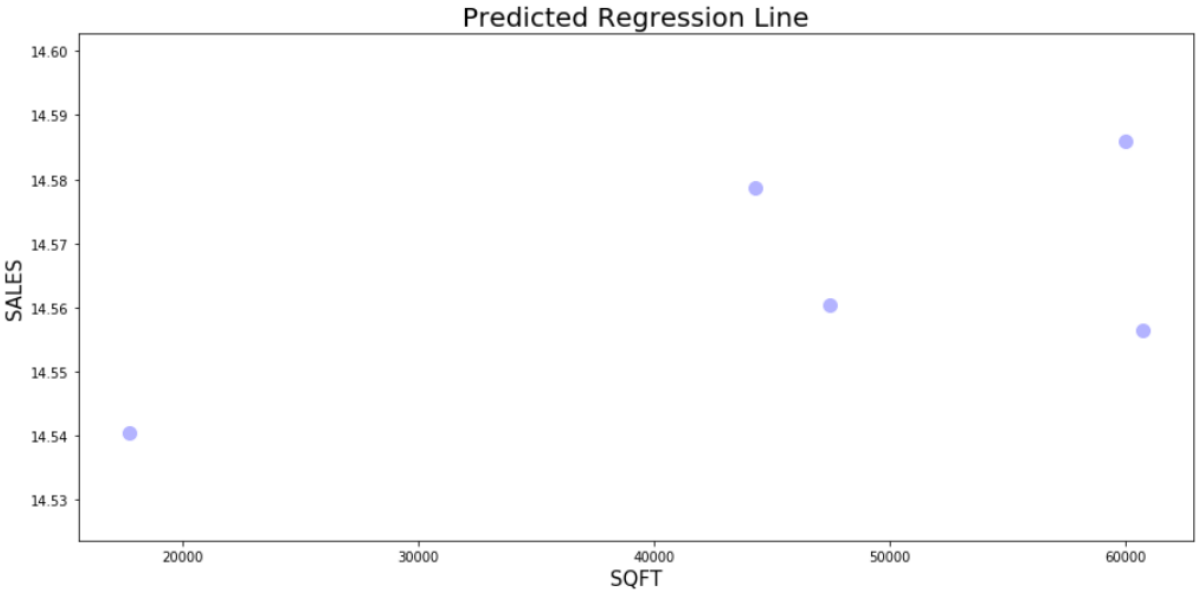
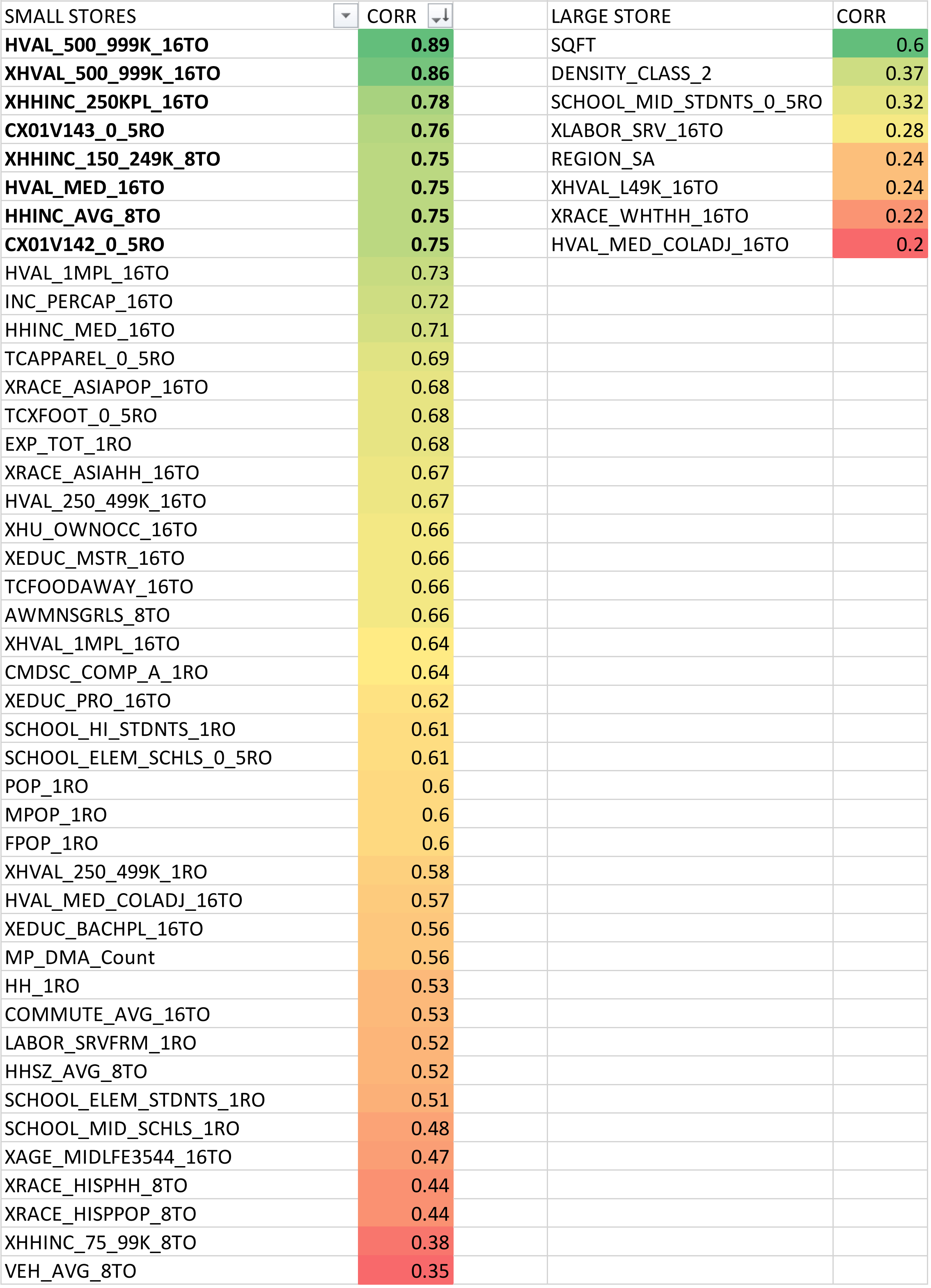


Figure 17: Variables highly correlated with the independent variable



SMALL STORES	CORR	LARGE STORE	CORR
HVAL_500_999K_16TO	**0.89**	SQFT	0.6
XHVAL_500_999K_16TO	**0.86**	DENSITY_CLASS_2	0.37
XHHINC_250KPL_16TO	**0.78**	SCHOOL_MID_STDNTS_0_5RO	0.32
CX01V143_0_5RO	**0.76**	XLABOR_SRV_16TO	0.28
XHHINC_150_249K_8TO	**0.75**	REGION_SA	0.24
HVAL_MED_16TO	**0.75**	XHVAL_L49K_16TO	0.24
HHINC_AVG_8TO	**0.75**	XRACE_WHTHH_16TO	0.22
CX01V142_0_5RO	**0.75**	HVAL_MED_COLADJ_16TO	0.2
HVAL_1MPL_16TO	0.73		
INC_PERCAP_16TO	0.72		
HHINC_MED_16TO	0.71		
TCAPPAREL_0_5RO	0.69		
XRACE_ASIAPOP_16TO	0.68		
TCXFOOT_0_5RO	0.68		
EXP_TOT_1RO	0.68		
XRACE_ASIAHH_16TO	0.67		
HVAL_250_499K_16TO	0.67		
XHU_OWNOCC_16TO	0.66		
XEDUC_MSTR_16TO	0.66		
TCFOODAWAY_16TO	0.66		
AWMNSGRLS_8TO	0.66		
XHVAL_1MPL_16TO	0.64		
CMDSC_COMP_A_1RO	0.64		
XEDUC_PRO_16TO	0.62		
SCHOOL_HI_STDNTS_1RO	0.61		
SCHOOL_ELEM_SCHLS_0_5RO	0.61		
POP_1RO	0.6		
MPOP_1RO	0.6		
FPOP_1RO	0.6		
XHVAL_250_499K_1RO	0.58		
HVAL_MED_COLADJ_16TO	0.57		
XEDUC_BACHPL_16TO	0.56		
MP_DMA_Count	0.56		
HH_1RO	0.53		
COMMUTE_AVG_16TO	0.53		
LABOR_SRVFRM_1RO	0.52		
HHSZ_AVG_8TO	0.52		
SCHOOL_ELEM_STDNTS_1RO	0.51		
SCHOOL_MID_SCHLS_1RO	0.48		
XAGE_MIDLFE3544_16TO	0.47		
XRACE_HISPHH_8TO	0.44		
XRACE_HISPPOP_8TO	0.44		
XHHINC_75_99K_8TO	0.38		
VEH_AVG_8TO	0.35		