Pam Sue

Frederick Pickering, CFA

6/18/2019

# Introduction

PamSue is a department store operating in the southeastern United States. The store primarily serves lower income residents. Recently PamSue’s forecasts have been missing the mark when it comes to selecting the placement for new stores. The management of PamSue is looking to add regression to the process of selecting the location for new stores. In addition to regression analysis they are looking at a new metric created by the planning department team where potential locations are assigned a *competitive type*. This analysis was conducted in R and the document was made using RMarkdown. All the exhibits are included in the exhibit section of the report. Readers can link to the exhibits by clicking links [colored like this.](#variable-correlation-matrix)

# Data Prep

## Correlation Analysis

Before we can start running regressions we need to familiarize ourselves with the underlying data. Here we will look at scatter plots of the various regressors (features) and how hey compare against the dependent variable (sales). We’ll also look at correlation matrices to check for possible multicollinearity in the variables.

The [correlation plot](#variable-correlation-matrix) of the various regressors to sales shows us that there is a strong correlation between the sales of a store and the percentage of Spanish speaking people in a neighborhood. The correlation also shows us that in neighborhoods where the percentage of residents earn between 10 and 14 thousand dollars has a strong relationship with sales in the sore. We also see strong negative correlations in the data set. Home ownership has a strong negative correlation with sales and the more in home utilizes. For example homes with dishwashers and dryers are more likely to see a strong negative relationship with sales.

## Question 1

*As a possible alternative to the subjective “competitive type” classification, how well can you forecast sales using the demographic variables (along with the store size and the percentage of hard goods)? What does your model reveal about the nature of location sites that are likely to have higher sales?*

Before we can begin to assess the accuracy of a model with all of these variables we nee to find the significant variables in the data set. Not all variables are going to be useful in predicting the sales of a store. In this model I start by making a linear regression model in R with all of the variables minus *Competitive Type*. The data has already been loaded into the R environment and the resulting line for a linear regression model is lm(sales ~ . -comType, data = regData) and assign the model to q1Reg. This provides a linear regression model with all of the variables minus compType. The next step is to assess the variables importance. Caret provides useful function for pulling variables importance from a model the function varImp(q1Reg) provides a ranking of the most important variables. This function looks ranks the variables by their contribution to lower the standard error of the model. I’ve taken the top 15 largest contributors to the model and passed them into a second linear regression model and I’ve come with with a linear regression model in the exhibits below.

[Question 1 Results](#question-1-results) [Question 1 Coefficients](#question-1-coefficients)

This model is able to explain approximately 68% of the variation in the data set and has an adjusted R^2 of 66%. This backwards method of modeling sales provides a slightly higher R^2 and adjusted R^2 values than the default step wise methodology provided by the RESESSIONS.XLSX file.

This regression model tells us several important things about the locations of PamSue stores. From the model I created population is the most important variable. This stands to reason as more populous locations are more likely to drive a higher level of foot traffic into the store and therefore more sales. The second and third most important variables are median home and selling square feet. These variables suggest that wealthier neighborhoods and larger stores are going to see a higher level of sales. Interestingly, demographic variables such as a home having an air conditioner, freezer, dryers, or a car are more likely to see lower level of sales. This suggests that PamSue does cater to a lower wealth demographic compared to the overall population.

## Question 2

*How good is the “competitive type” classification method (along with using the store size and percentage of hard goods) at predicting sales.*

Competitive type is a rather good predictor at estimating sales. Just using competitive type, percent hard goods, and selling square feet provides a model with an adjusted R^2 of 70%. This means the dummy variables, and selling square feet are better at predicting sales than the demographic data provided in the data set. This means that the real estate department has come up with a very useful methodology for categorizing store locations. Interestingly percent hard goods does not produce a statistically significant result when included in the model. This either means it’s explanatory analysis is dwarfed by the relationship between sales and the other variables or sales is not strongly related to percent hard goods.

[Question 2 Results](#question-2-results) [Question 2 Coefficients](#question-2-coefficients)

This is a strong model with all of the regressors having t-stats that are significantly above 2. The p-values also suggest that there is a s very small chance we are accepting a false Beta as true. Note these values show the estimated increase in sales over competitive type seven. Competitive type seven includes ***“stores located along the sides of major roads”*** meaning any other competitive type location will see higher sales than stores simply placed alongside major roads.

## Question 3

*Two sites, A and B, are currently under consideration for the next new store opening. Characteristics of the two sites are provided below in Table B. Which site would you recommend? What sales forecasting approach would you recommend.*

In order to approach this problem I created a hybrid model that looks at both the stores competitive type and demographic information. I used this approach because it’s able to get us a better model. This model is able to take in hard demographic data of the neighborhood and compare it to qualitative measurements from the real estate department.

[Question 3 Projections](#question-3-projections)

The model determined that there are 15 important regressors for determining the sales of a store. Four of these regressors are competitive types and ten are demographic. The model suggests that we should build a store in Site A. Site A was selected primarily due to it being classified as a type 1 location. Competitive Type 1 locations are expected to see $9952.5 more in sales than competitive type 7 locations. Comparatively, competitive type 5 locations can only expect $1805.104 more in sales. This benefit alone gives Site A a significant boost over Site B for planning purposes. Site A also benefits from having a higher population, higher home prices, and lower utilization of negatively important variables such as dryers, freezers, dishwashers, and air conditioners.

## Question 4

*Two of the variables in the data base are under managerial control: the size of the store and percent hard goods stocked in the store. Margins on hard goods are different from soft goods. What impact do these variables have on sales*

I’ve run a linear regression model with all of the regressors in the database. This model provides us with a list of all of the variables, their t-stat, and their contribution to the models performance. Due to the fact that the correlation matrix did not see any high levels of correlation, we can determine that there is little to worry about from the prospective of multicollinearity in the regressors. Looking at the model and the t-stats we can determine that the *Percent Hard Goods* is not a very strong predictor of the sales in the model we are using.

[Modl Variable Importance](#variable-importance)

Looking at the exhibit below we can see that the scatter plot and moving average for the *Percent Hard Goods* is relatively flat. This suggests that there isn’t a relationship between the *Percent Hard Goods* and the level of sales in a store. However, there is a small positive relationship between the *Selling Square Feet* and the sales in a store location suggesting that this is a more useful predictor of sales and something management should consider.

[Square Feet and Hard Goods Scatter](#scatterplots)

From the linear regression model we can see that the t-stat for the *Selling Square Feet* is significant variable with a t-stat of 2.211, in the comType and demographic inclusive model. This means that there is a strong relationship we can be confident in and management should look to control the size of a store to drive sales. However, management can’t be confident they will be able to impact sales through the use of the *Percent Hard Goods* regressor as it is to insignificant to appear in the model and we are not confident that it’s impact is any more statistically significant than zero.

## Question 5

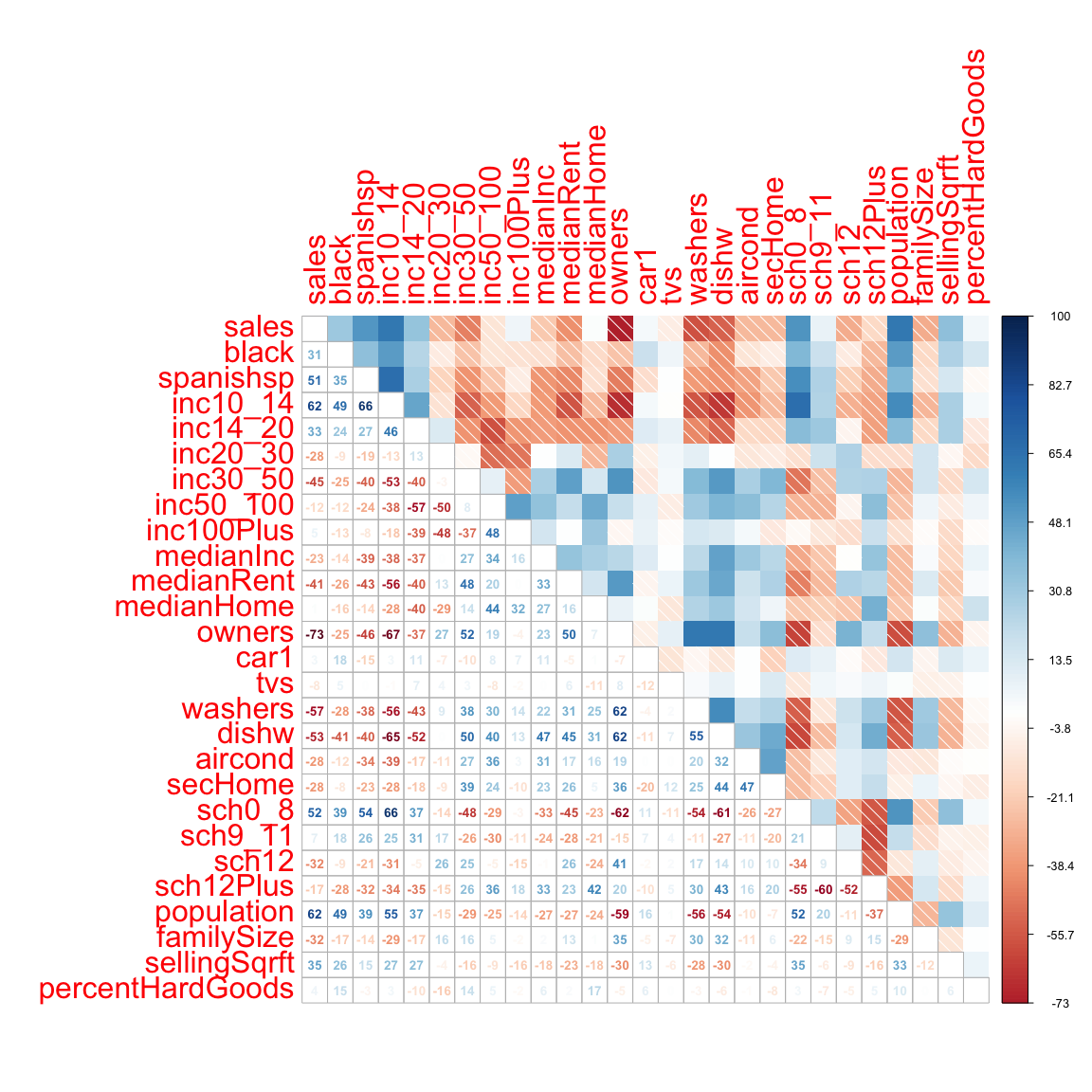
*TECHNICAL: For your recommended model, check to make sure the technical assumptions are satisfied. Comment on any points that would concern you based on the diagnostics.*

When looking at the linear regression model we need to diagnose any negative effective in the linear regression model such as, heteroskedasticity, multicollinearity, and nonlinear relationships in the residuals. Exhibit [Q5 Residuals](#q5-residuals) and exhibit [Q5 Residual Histogram](#q5-residual-histogram), show us the relationship between the the model’s predictions and the actual observed values in the data. We can see from [Q5 Residuals](#q5-residuals). that there is random noise in the residual. This is important because it means that there are no lurking non-linear relationships in the data we’re unable to account for. Also the standard deviation of the noise is constant meaning we don’t have an issue with heteroskedasticity.

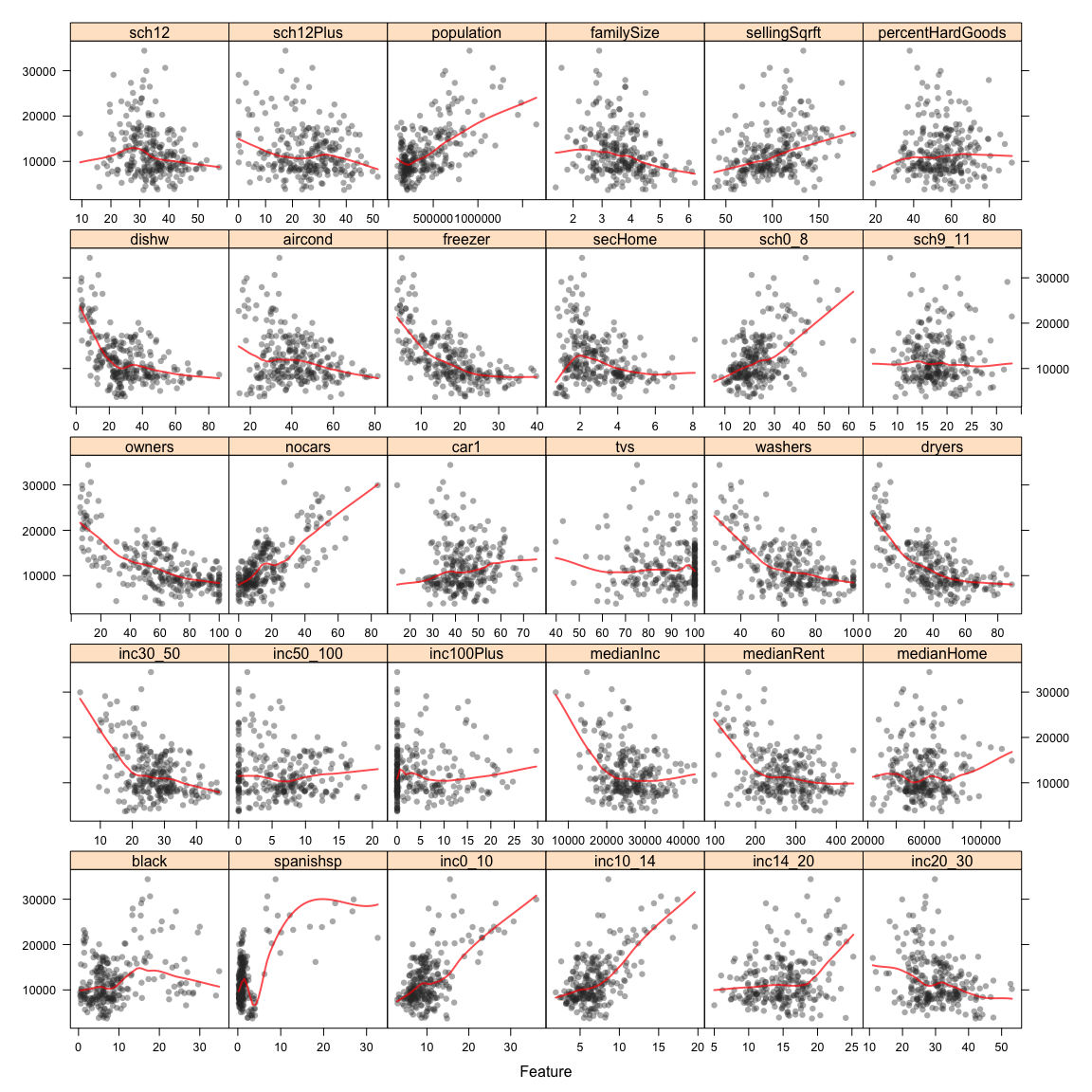
Exhibit [Q5 Projections](#q5-projections). shows the relationship between the predictions of the model and the actual values. This is another way of looking at the residuals and also visualizing the R and R^2 values of the model. The model has a strong 77% Adjusted R^2 which means roughly 77% of the variation in the underlying data is explained by the model. However there is some variation in the model. When looking at the residuals we can determine that this is random noise and challenging to model. If we were insistent on modeling with more variables we’d risk over fitting the model to the data set. This would mean that we have a model that does a better job explaining the underlying data set, but fails to correctly predict new out of sample data.

# Exhibits

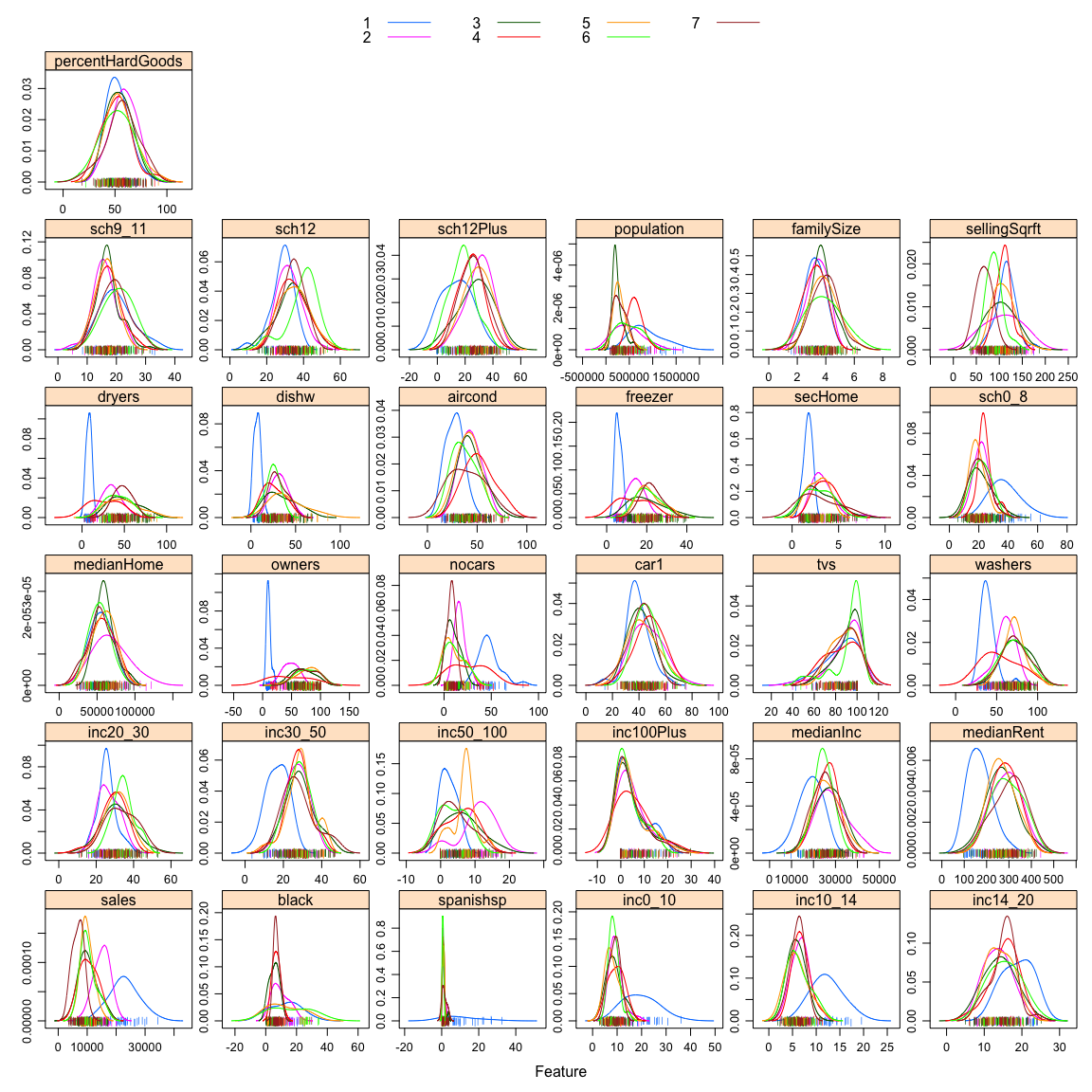
## Variable Correlation Matrix



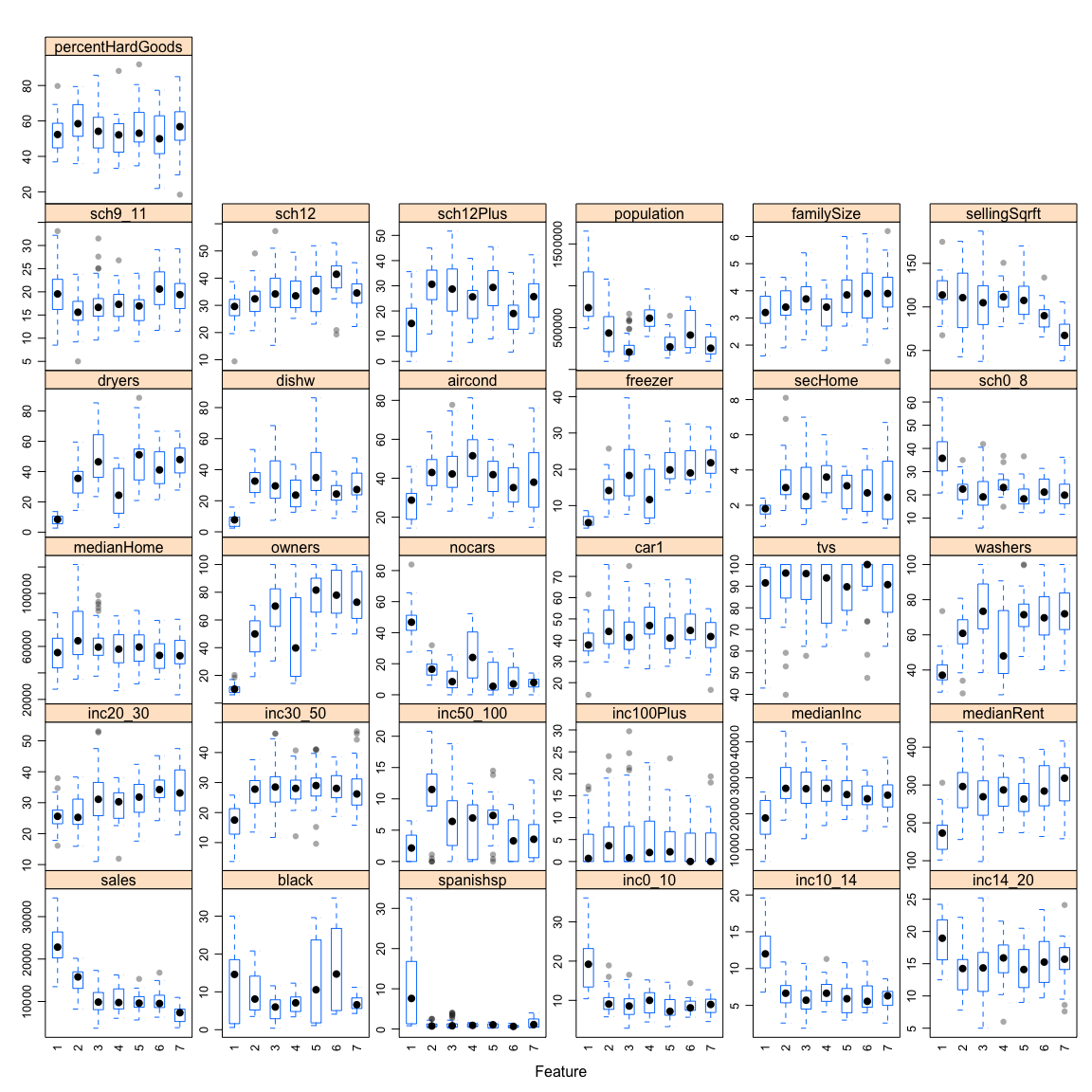
### Variable Scatter Plots



### Competitive Type Curves



### Competitive Type Box Plot



## Question 1 Exhibits

### Question 1 Results

| **Metric** | **Value** |
| --- | --- |
| r.squared | 0.683 |
| adj.r.squared | 0.662 |
| sigma | 3167.674 |
| statistic | 33.580 |
| p.value | 0.000 |
| df | 16.000 |
| logLik | -2361.655 |
| AIC | 4757.311 |
| BIC | 4817.176 |
| deviance | 2347993814.365 |
| df.residual | 234.000 |

### Question 1 Coefficients

| **Term** | **Beta** | **Standard Error** | **T-Stat** | **P-Value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 16916.622 | 2793.569 | 6.056 | 0.000 |
| population | 0.005 | 0.001 | 5.059 | 0.000 |
| medianHome | 0.057 | 0.015 | 3.883 | 0.000 |
| sellingSqrft | 21.864 | 7.609 | 2.873 | 0.004 |
| sch9\_11 | -142.762 | 49.355 | -2.893 | 0.004 |
| spanishsp | 266.625 | 70.510 | 3.781 | 0.000 |
| aircond | -52.976 | 20.193 | -2.623 | 0.009 |
| freezer | -118.013 | 48.807 | -2.418 | 0.016 |
| sch0\_8 | -68.139 | 34.728 | -1.962 | 0.051 |
| owners | -31.579 | 13.991 | -2.257 | 0.025 |
| car1 | -42.689 | 22.791 | -1.873 | 0.062 |
| secHome | -287.017 | 185.715 | -1.545 | 0.124 |
| dryers | -41.898 | 21.058 | -1.990 | 0.048 |
| black | -50.874 | 34.915 | -1.457 | 0.146 |
| inc10\_14 | 190.643 | 135.164 | 1.410 | 0.160 |
| inc50\_100 | 89.475 | 52.155 | 1.716 | 0.088 |

## Question 2 Exhibits

### Question 2 Results

| **Metric** | **Value** |
| --- | --- |
| r.squared | 0.714 |
| adj.r.squared | 0.705 |
| sigma | 2959.262 |
| statistic | 86.180 |
| p.value | 0.000 |
| df | 8.000 |
| logLik | -2348.843 |
| AIC | 4715.686 |
| BIC | 4747.379 |
| deviance | 2119250366.690 |
| df.residual | 242.000 |

### Question 2 Coefficients

| **Term** | **Beta** | **Standard Error** | **T-Stat** | **P-Value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 5219.535 | 725.403 | 7.195 | 0.000 |
| comType1 | 14895.319 | 837.641 | 17.782 | 0.000 |
| comType2 | 7192.957 | 774.730 | 9.284 | 0.000 |
| comType3 | 2544.779 | 685.767 | 3.711 | 0.000 |
| comType6 | 3030.626 | 868.942 | 3.488 | 0.001 |
| sellingSqrft | 24.234 | 7.192 | 3.370 | 0.001 |
| comType4 | 2677.296 | 907.221 | 2.951 | 0.003 |
| comType5 | 2076.237 | 822.162 | 2.525 | 0.012 |

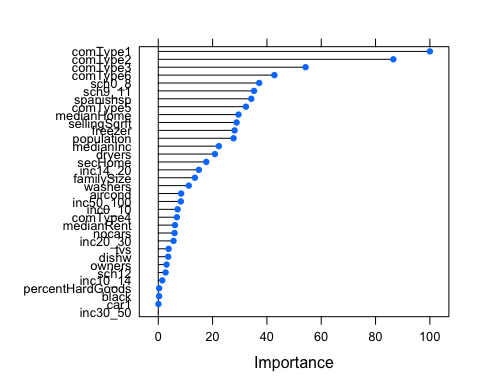
## Question 3 Exhibits

### Question 3 Projections

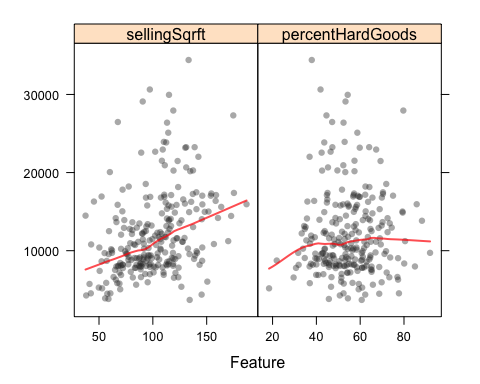
| **Location** | **Projection** |
| --- | --- |
| Site A | 23217.778 |
| Site B | 11056.596 |

## Question 4 Exhibits

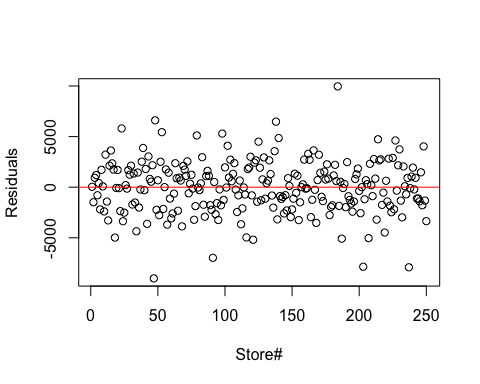
### Variable Importance



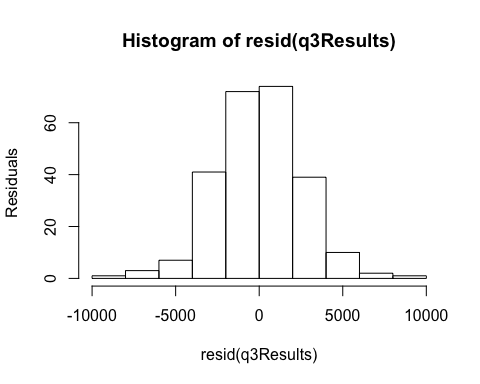
### ScatterPlots



### Q5 Residuals



### Q5 Residual Histogram



### Q5 Projections

