James Eun

Roman Sielewicz

Jacob Lee

Alek Carlson

Dr. Sharif

DSO545 Statistical Computation

[REAL ESTATE FIRM] Analytics Project

**Problem Statement**

Our team has been tasked with assisting [Real Estate Firm], a real estate asset management firm based in the Southern United States. Their current investment strategy is focused on the acquisition and management of apartment communities and shopping centers throughout the southeast United States and Texas. [Real Estate Firm]’s goals are to bring quantitative finance into the real estate industry as a first mover and develop algorithms to automate the asset valuation process. Ideally, [Real Estate Firm] would be able to identify undervalued properties using statistical methods and make returns on these properties from both rent and capital gains. Currently, [Real Estate Firm] has gathered and purchased data on approximately 25,000 commercial properties in the Southeast US, including building-specific details and sales transaction data. Our goal is to analyze the data and develop insights into which factors are most important in determining the value of these real estate assets.

**Key Business Insights**

We believe the most important variables (in order of significance) to consider in predicting sales price are:

1. Building Age
   1. As the age of a building increases, its sale price tends to decrease
2. Number of Units
   1. As Number of units per building increases, its sale price tends to increase
3. Building Square Foot
   1. As building square foot increases, its sale price tends to increase
4. Fitness Center (Amenity)
   1. If a property has a fitness center on site, its sale price tends to be higher
5. 2019 Avg Age: The average Age of the area in which the property is located
   1. A higher average market age tends to indicate higher prices
6. Star ratings
   1. The Star Rating of a building is positively associated with its sale price
7. Affordable\_Housing
   1. The existence of an affordable housing program within the unit tends to correlate with lower sales prices
8. Playground (Amenity)
   1. Properties with playgrounds on site tend to have lower sale prices
9. Construction Material
   1. Structures built of Masonry and structures built of reinforced concrete tend to have higher sale prices
10. Out of State: Indicates the owner is from another state than the property
    1. Properties with out of state owners tend to have lower sale prices
11. Grill (Amenity)
    1. Apartment complexes with grills on site tends to have higher sales prices
12. Gated (Amenity)
    1. Gated apartment complexes tend to have lower sales prices

**Data Cleaning**

We each split up the data set to figure out what variables should be under consideration.

We split it up as follows:

* Roman: Building Age, Building Materials, Features and Amenities, Star Rating, Type of Owner (out of town, larger owner)
* Alek: Floor to Area Ratio, Number of Floors, Land Area, Market Time, Net Income
* Jacob: Unit Mix, City, County, Other Location Identifiers, Date Sold
* James: Vacancy, Market Rent/Sales Price, Closest Transit, Concessions, Demographic Data, Supply

In cases (variables, columns) that were missing over 80% of the records, except for factor variables like ‘Affordable Type’ where the missing data were at a reference level, we removed them from consideration.

Here are some manipulations our team made to the data set:

* Roman:
  + Out of State
    - Separated Owners into two categories, in-state
  + Building Age
    - Created a new column for building age based on year built
  + Years Since Renovation
    - New Column assessing the years since the most recent renovation
  + Masonry, Reinforced Concrete, Wood Frame
    - Created Dummy variables for each Construction Material type
  + Amenities - Split them up into discrete dummy variables
  + Rental Revenue
    - Expert Variable: the sum of the rent per type of unit time the number of units of that kind
  + Star rating dummy variables
    - Created dummy variables from the star rating variable
* Alek:
  + Floor to area ratio
    - Created a new column by dividing the building square feet column by land area square feet column.
* Jacob:
  + States
    - Created dummy variables
  + Secondary Type
    - Created dummy variables
  + Sales price
    - Deleted rows with nonsensical prices (<$100,000)
  + Price per sqft
    - Deleted rows with nonsensical prices (<$20/sqft and >$10,000/sqft)
  + ln(Price)
    - Natural log of the sales price
    - Natural log of sales price per sqft
  + 2019 Average Age
    - Deleted zeroes
* James:
  + Vacancy %
    - Substituted missing data with median value
  + Avg Concessions
    - Since concessions are essentially coupons, missing data were replaced with 0’s
  + Affordable.TypeNone
    - Since most values were missing data, this column was binarized as a factor column with ‘None’ and ‘Distinction.’ Meaning that if there was a value, there was some sort of distinction for affordability action.
  + Closest Transit Stop
    - Missing values were replaced with median value

**Exploratory Data Analysis**

To define the relationship between the given variables, we set out to create a linear model to predict sales and rental prices.

Since there were many instances of missing data in the records related to rental prices, we decided to focus on sales price.

We went through the columns in the data and picked variables that were likely to have correlations with the sales price. We also plotted each variable against the dependent variable to ensure a linear relation was apparent. We then added these variables to a multivariate linear regression.

When estimating Sales Price, we noticed that the residuals plot was cone-shaped, indicating that the model was heteroscedastic. Thus, we decided to estimate ln(Sales Price) instead, which removed the issue of heteroscedasticity.

Using the t-statistic for each independent variable and an alpha of 0.05, we removed variables with an insignificant slope estimate (p-value > 0.05) and kept variables that had significance (p-value < 0.05). Using the VIF statistic, we made sure to avoid collinearity between variables by ensuring that all values were below a threshold of 5.

The variables in our final model were able to account for about 68% of the variation in Sales Price. We applied the same method for predicting sales price per sqft and obtained an r-squared of 46%.

Next we wanted to explore the data visually to gain a better understanding of the markets with the highest number of sales. After determining which markets had the highest number of sales we imported our data into Tableau to visualize how these sales were dispersed throughout the top performing markets.

We used the latitude and longitude to plot a point for each property sold. After that we filtered the data to only show points in the markets we deemed were top performing. Finally, we added color to each point based off of the Sale Price variable so we could distinguish the properties sold at the highest prices. With these visualizations we are able to better distinguish patterns throughout the top performing markets.

**Future Work**

If we were given the opportunities and the resources, we would aim to accomplish the following:

1. Create combinations of variables to combine highly correlated variables to produce better modeling solutions
2. Explore and integrate third party data to increase the prediction value of the models
3. Reach out to property owner constituents to fill in missing data. For example, some of the properties might be missing rental price data, which could be useful for prediction.
4. Lastly, we would like to come up with more creative ideas for using our variables. We believe all of the data could be useful for prediction.

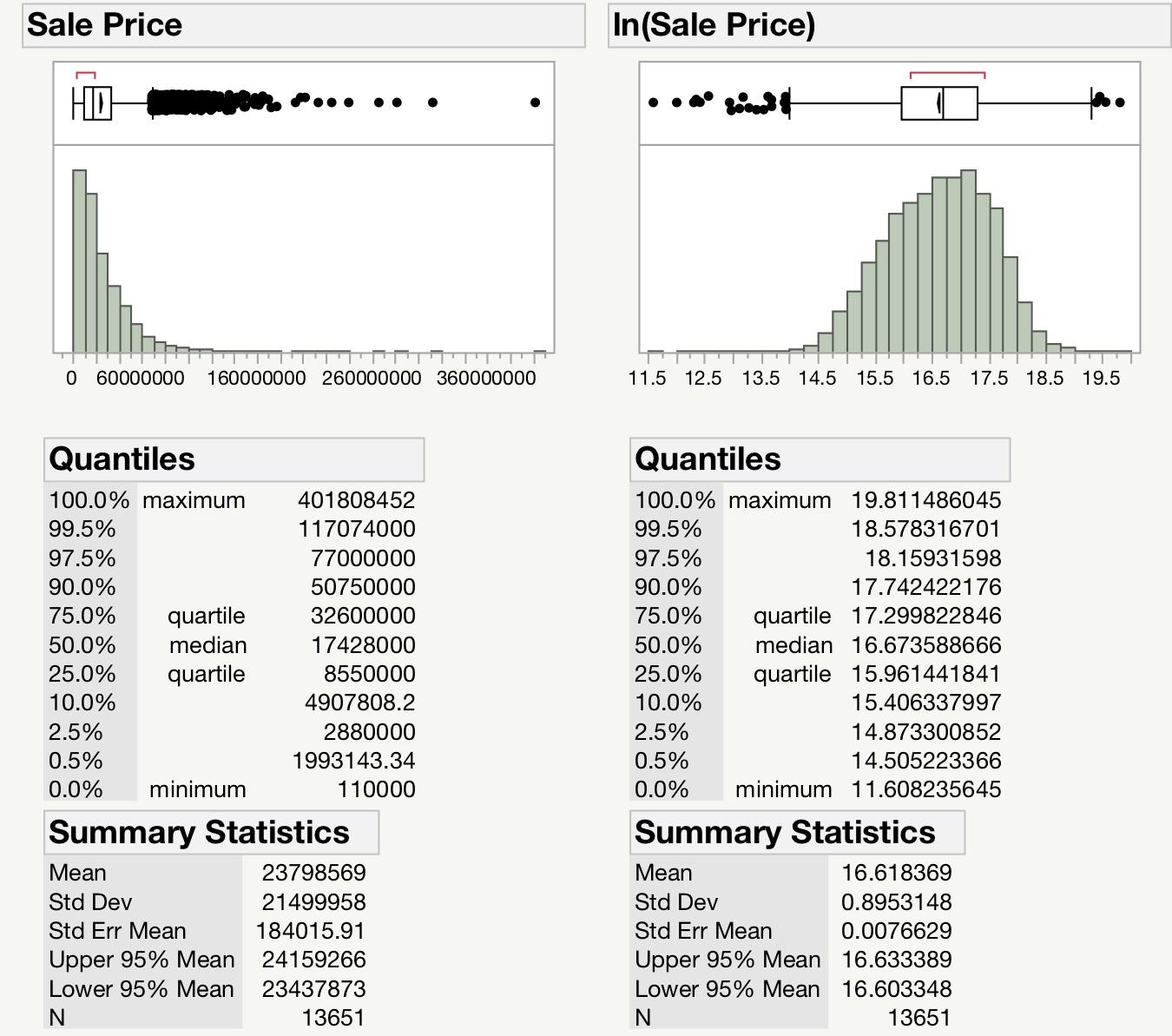
**Team Contribution**

Our team split up data cleaning and analysis, taking responsibility for different variable categories. After making attempts at linear models vs sales price, we combined significant variables for one large regression model.

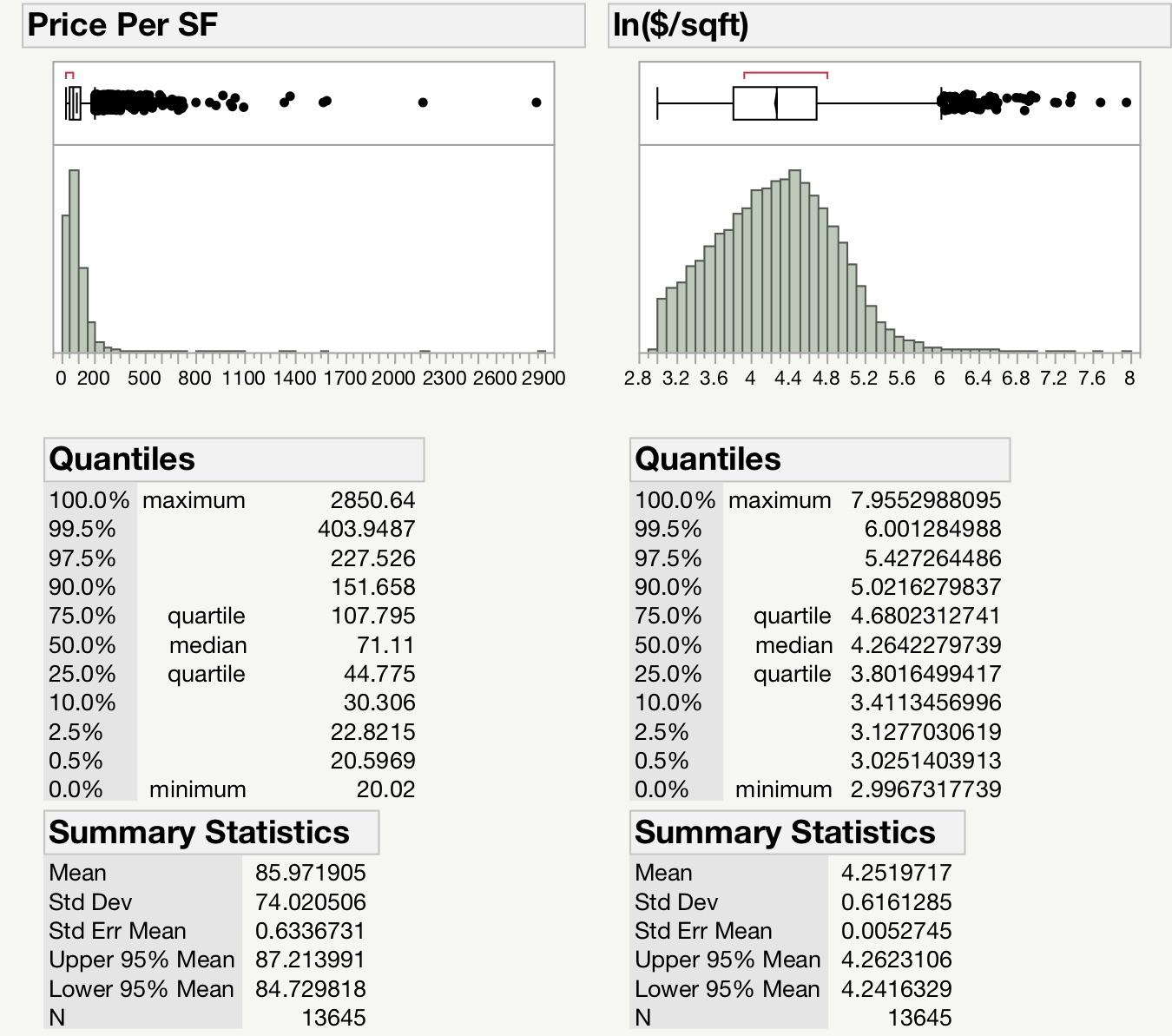
|  |  |  |
| --- | --- | --- |
| **Member Name** | **Variable Interests** | **Deliverables** |
| James Eun | Vacancy, Market Rent/Sales Price,  Closest Transit, Concessions, Demographic Data, Supply | Analysis/visualizations in R, report and presentation |
| Roman Sielewicz | Building Age, Building Materials, Features and Amenities, Star Rating, Type of Owner (out of town, larger owner) | Cleaning/analysis in python, report and presentation |
| Jacob Lee | Unit Mix, City, County, Other Location Identifiers, Date Sold | Analysis/visualizations on JMP, final regression models, report and presentation |
| Alek Carlson | Floor to Area Ratio, Number of Floors, Land Area, Market Time, Net Income | Tableau analysis/visuals, report and presentation |

**Appendix:**

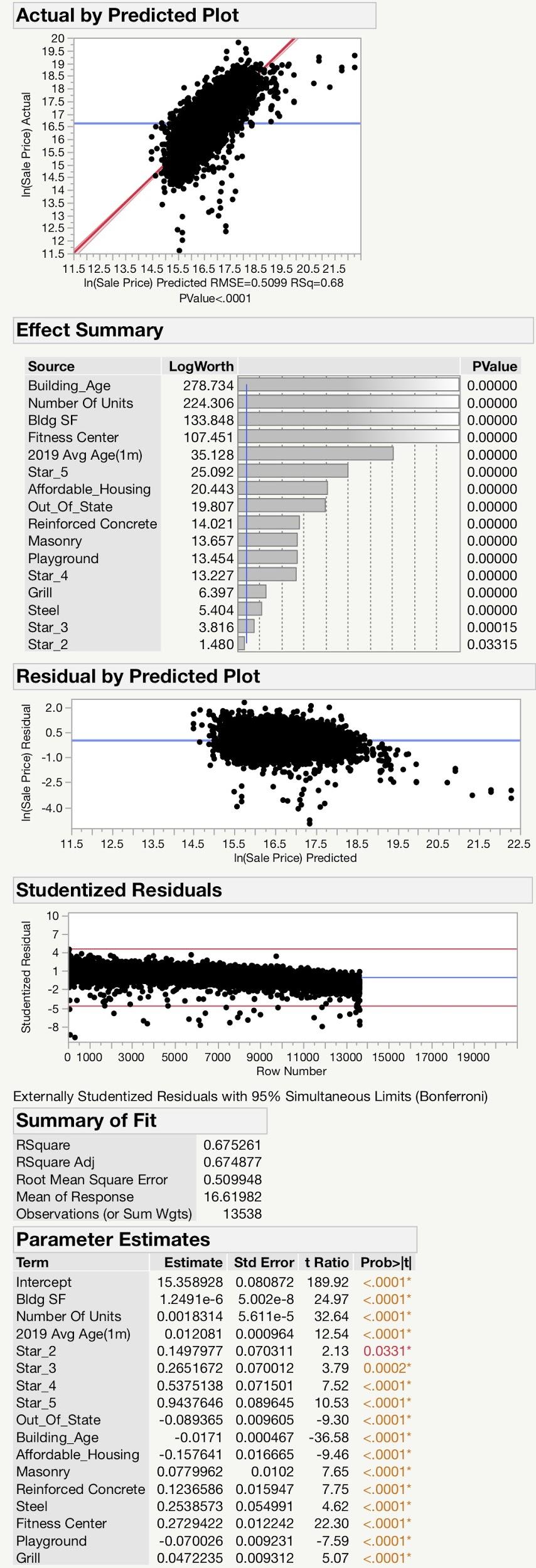
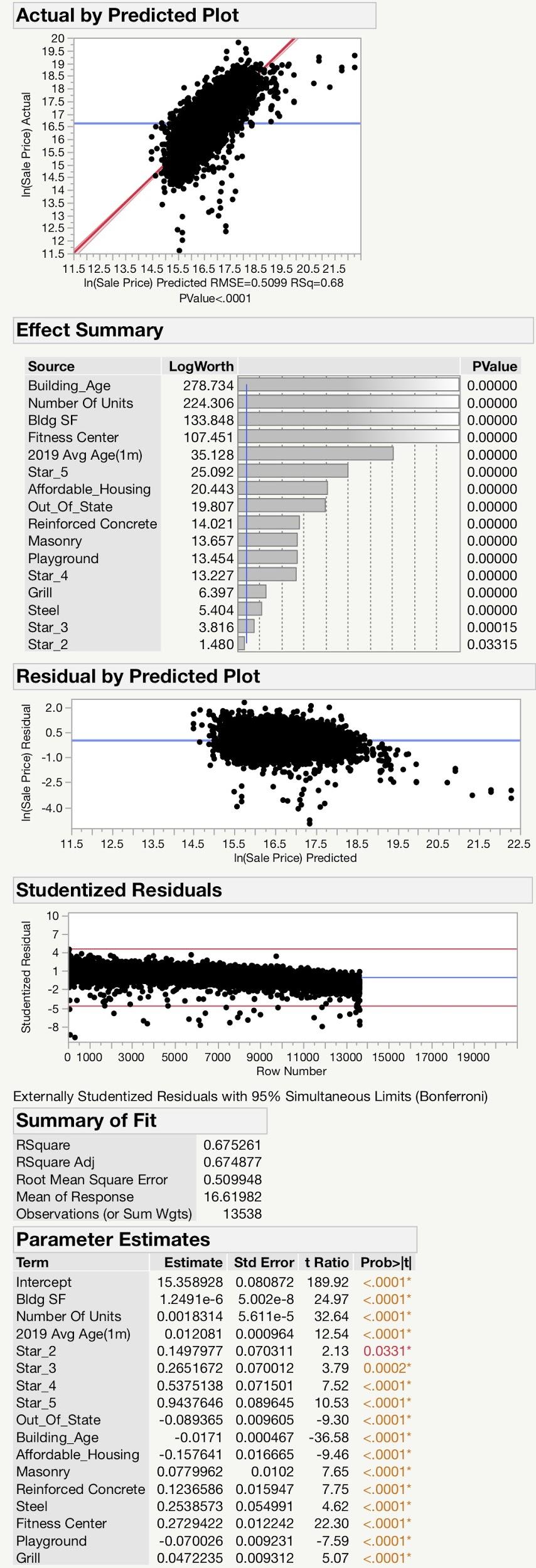
Histograms of Sales Price and its log transform.



Histograms of Price per Sqft and its log transform.



Regression model after variable transformations and log based transform to Sales Price. Residual plot also in log form. Model statistics produced in JMP software.



Regression model after variable transformations and log based transform to Sales Price per square foot. Residual plot also in log form. Model statistics produced in JMP software.

