Pene Seloti, Shawn Anderson, Ian Morris, Jafar Hosseinzadeh

IST 687 | Applied Data Science

R Project

Flinders Home Analysis

Table of Contents

[Introduction- background and scope 2](#_Toc20160582)

[Business Questions 3](#_Toc20160583)

[Table of Figures 4](#_Toc20160584)

[Data acquisition, cleanse, and munging: 5](#_Toc20160585)

[Descriptive: histograms, maps, SCATTER AND BOX PLOTS 6](#_Toc20160586)

[Descriptive Continues: SCATTER AND BOX PLOTS 14](#_Toc20160587)

[PREDICTIVE: LINEAR MODELS 16](#_Toc20160588)

[DATA MINING: LINEAR, KSVM, SVM MODELS 19](#_Toc20160589)

[OVERALL INTERPRETATION OF RESULTS/ACTIONABLE INSIGHTS 22](#_Toc20160590)

[Appendix 23](#_Toc20160591)

[Data Dictionary 23](#_Toc20160592)

# Introduction- background and scope

John and Pam Flinders entered our Sacramento office today, June 21st 2009, with questions about building or purchasing a single family home. They are also exploring the possibility of investing in a second dwelling – preferably a condo – for their college age children. Their search criteria centers around the City of Sacramento and its immediately surrounding communities (hereafter called Region).

The Region Tax & Assessments office releases a daily real estate sales report through the Sacramento Bee website 30 days in arears. We pulled the most recent 5-day period of sales activity; which spans a weekend, May 15 – 21. With nearly 1,000 real estate closings over this period and the larger two-home spending budget at hand, we determined an in-depth analysis was needed.

This report answers some initial questions we received from their visit, coupled with the means and measures of a more thorough quantitative assessment for our proceeding appointment. It concludes with our chief recommendations for the Flinders family, based on the initial intake information we received.

# Business Questions

The Flinders presented our team with the four following questions:

1. Q – *“What is the overall price range for properties in this market?*
2. Q – *“We prefer another 5-bedroom home. What is the average price in this region?”*
3. Q - *“Having a potential two-home budget, we need to be mindful of cost. Which property feature is the best measure for reducing our final price (cost-driver)?*
4. Q – *“How about condos? What are the current pricing conditions?”*
5. Q - *We have a $450,000 budget but wonder if this provides enough room for us to purchase a smaller investment home which our college kids can use. What do you think?”*

# Table of Figures

|  |  |  |  |
| --- | --- | --- | --- |
| Page Number | Figure Number | Title | Figure Description |
| 7 | **1** | **Structure of "re1"** | **Diagnostic of the internal structure of our dataset** |
| 7 | **2** | **Summary of "re1"** | **Summary of the results of our dataset** |
| 8 | **3** | **Vector Information** | **Mean/median/SD/skewness/range/quantile of 4 variables of our dataset** |
| 8 | **4** | **Descriptive: Bedrooms** | **Dataframe of mean/median/SD/Kurtosis by number of bedrooms in property** |
| 9 | **5** | **Histogram: Housing Sales by Price** | **Histogram of housing sales by price with fill set to "count." Darkblue = higher count** |
| 10 | **6** | **Barplot: Sales Price Filled** | **Barplot of housing sales price with fill set to “type of property”** |
| 10 | **7** | **Density plot: Type of Property** | **Density plot of housing sales price with fill set to "type of property”** |
| 11 | **8** | **Scatterplot: Price by Square Feet + Regression** | **Scatterplot of price by square feet with color set to “number of bedrooms” and a regression line denoting "best fit"** |
| 12 | **9** | **Boxplot: Price by Number of Bedrooms** | **Boxplot of price arranged by the categorical variable “bedrooms” (1-8)** |
| 13 | **10** | **Boxplot: Price by Property Type** | **Boxplot of sales price and type of property faceted by sales date** |
| 14 | **11** | **Geographic Map: Points of Home Sales** | **Map of Sacramento area with points representing home sales, scaled by color to reflect price** |
| 15 | **12** | **Scatterplot: Beds and Price** | **Plot illustrating number of beds and price, informing modeling process** |
| 15 | **13** | **Scatterplot: Baths and Price** | **Plot illustrating number of bathrooms and price** |
| 16 | **14** | **Scatterplot: Square Feet and Price** | **Plot illustrating square footage and price** |
| 16 | **15** | **Scatterplot: Property Type and Price** | **Plot illustrating property type and price** |

# Data acquisition, cleanse, and munging:

We obtained the latest data from the following Sacramento Bee website spanning May 15-21:

http://samplecsvs.s3.amazonaws.com/Sacramentorealestatetransactions.csv

From this we received a list of 985 real estate transactions in the region reported over a five-day period, as described by the [Sacramento Bee](http://www.sacbee.com/). Our intent was to analyze a dataset that had enough categorical variables from which we could determine correlation to housing sale price. We chose to “read-in” the entire dataset and cleanse the data to meet present needs of our clients.

The dataset which we imported as “rawrealestate” is a data frame with 985 observations and 12 variables. While we believed that several of the variables were classed properly, we assessed that we would need to adjust the class of beds, baths, and date variables. These variables provide further analytical opportunities when represented as factors.

Our initial assessment of the dataset necessitated that we perform minor adjustments to the “date” and “zip” variables. Additionally, we assumed that square footage would be a major indicator of housing price and thus decided to omit entries that were missing this key data point. We developed the following code to accomplish this task:

# cleanse the data

> rawRealEstate <-read.csv(url("http://samplecsvs.s3.amazonaws.com/Sacramentorealestatetransactions.csv"))

> re1 <- data.frame(rawRealEstate)

> re1$sale\_date <- gsub("00:00:00 EDT ","",re1$sale\_date)

> re1 %>% separate(sale\_date,c("Weekday","Month","Monthday",NA,NA,"Year"))

> re1$zip <- as.factor(re1$zip)

> re1$beds <- as.numeric(re1$beds)

> re1$sq\_\_ft[re1$sq\_\_ft==0] <- NA

> is.na(re1$sq\_\_ft) <- mean(re1$sq\_\_ft)

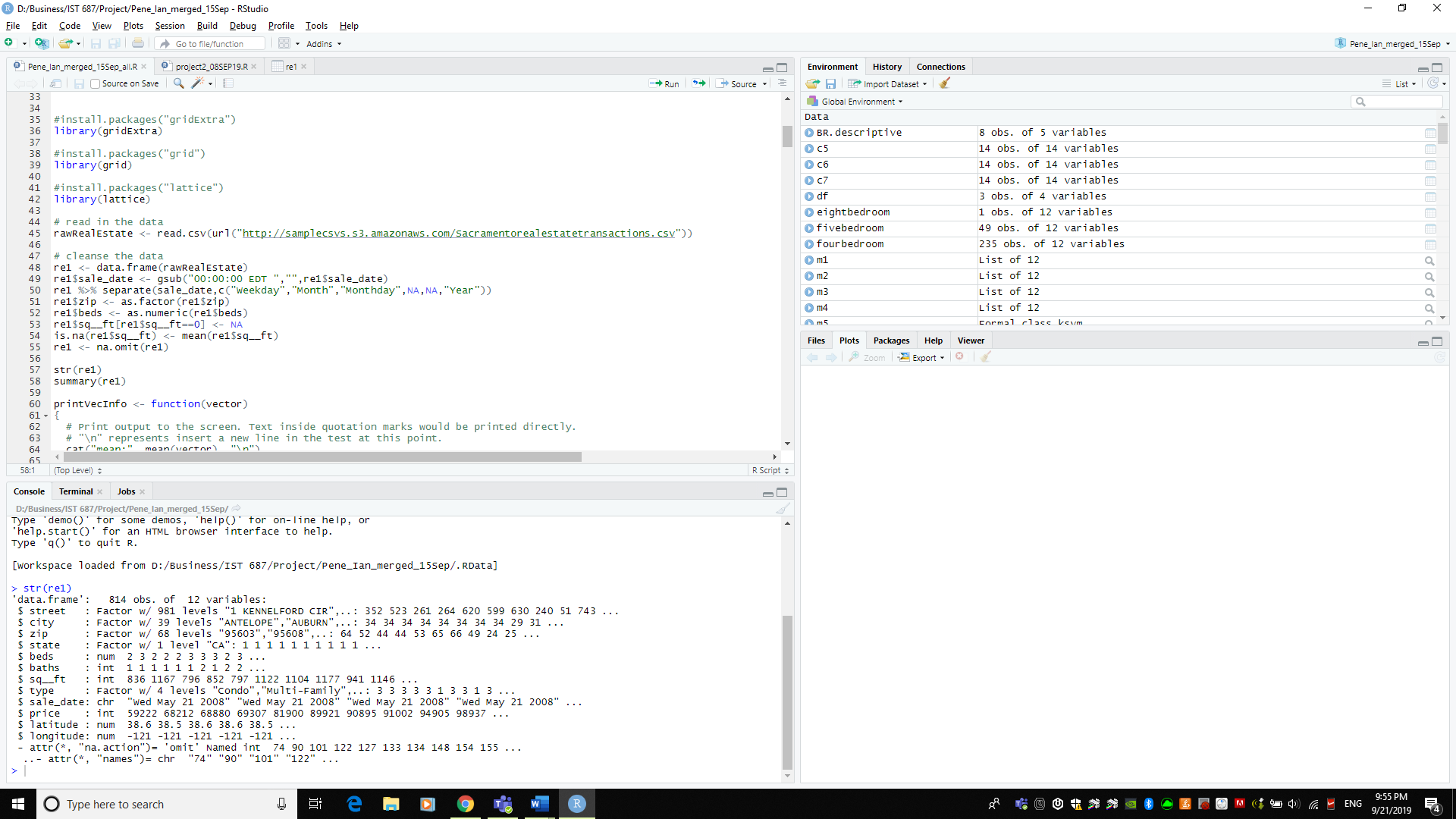
> re1 <- na.omit(re1)

Considering these minor corrections, we deemed the resulting dataset sufficient to begin our analysis.

# Descriptive: histograms, maps, SCATTER AND BOX PLOTS

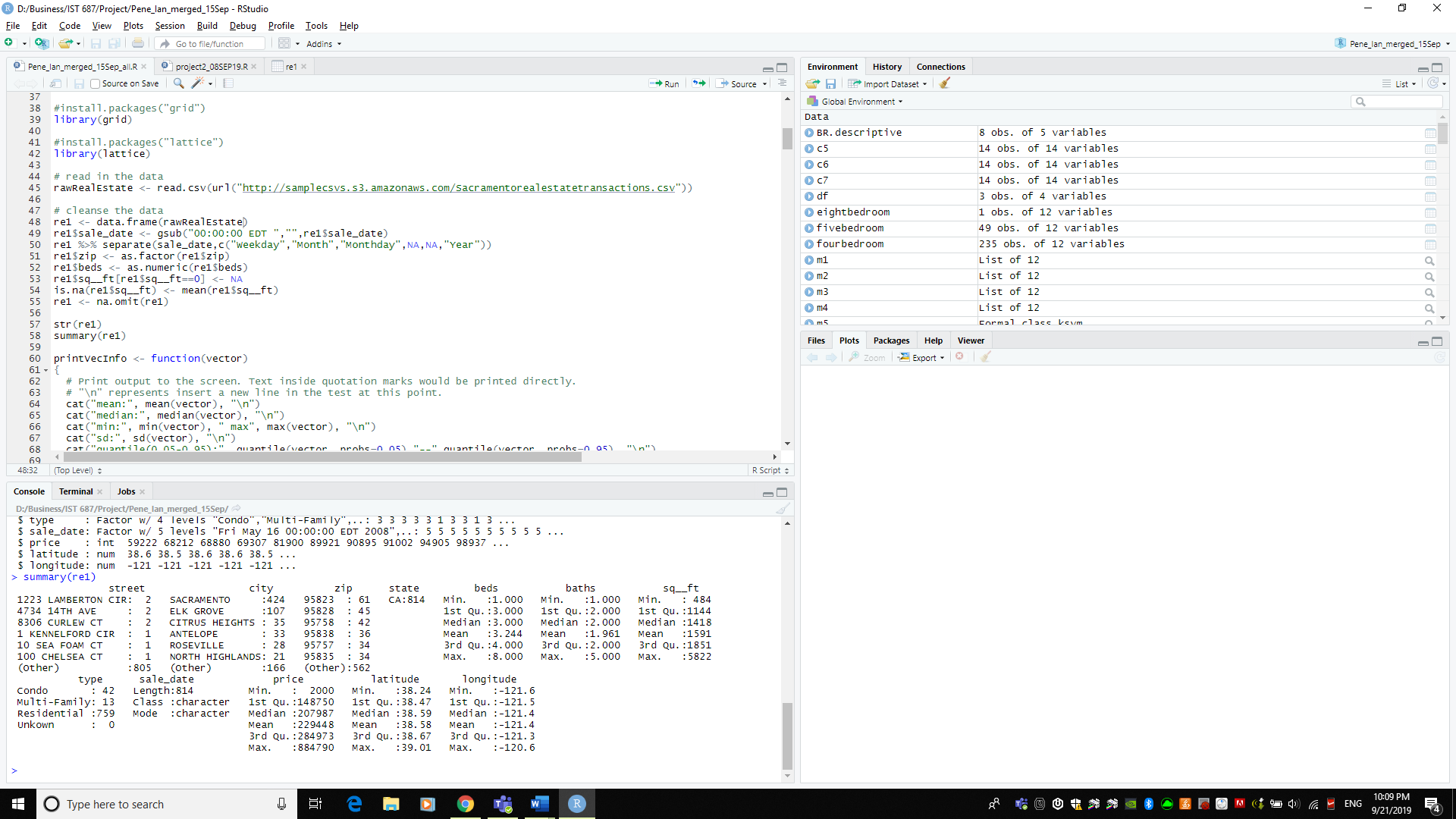
Running str() on the dataset yields the following data structure:

**Figure 1**



There are several key variables which inform our analysis. Most notably, zip code, beds, baths, square feet, type, sale price, latitude and longitude. We utilized the summary() function to better familiarize ourselves with the data which populated:

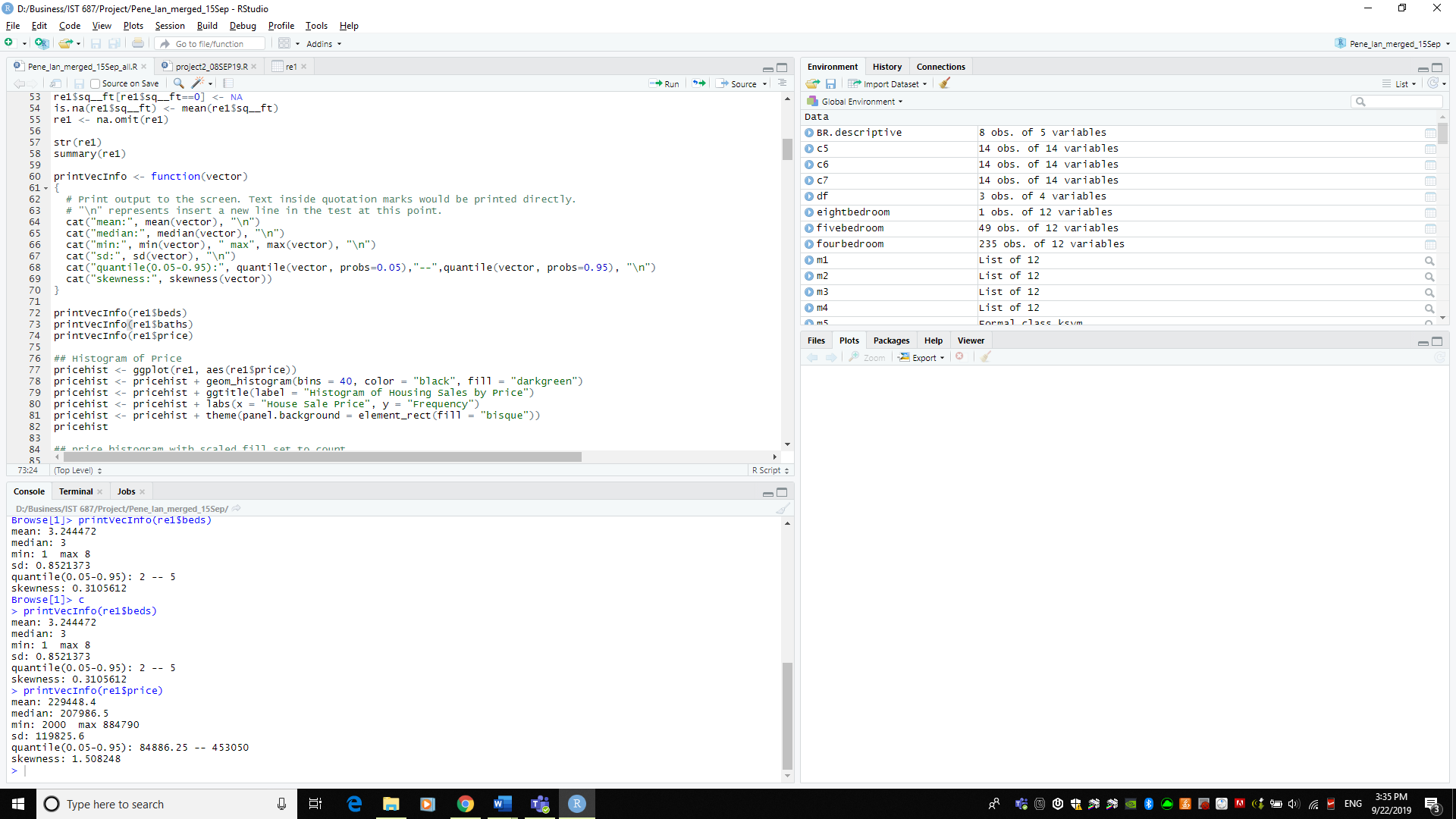
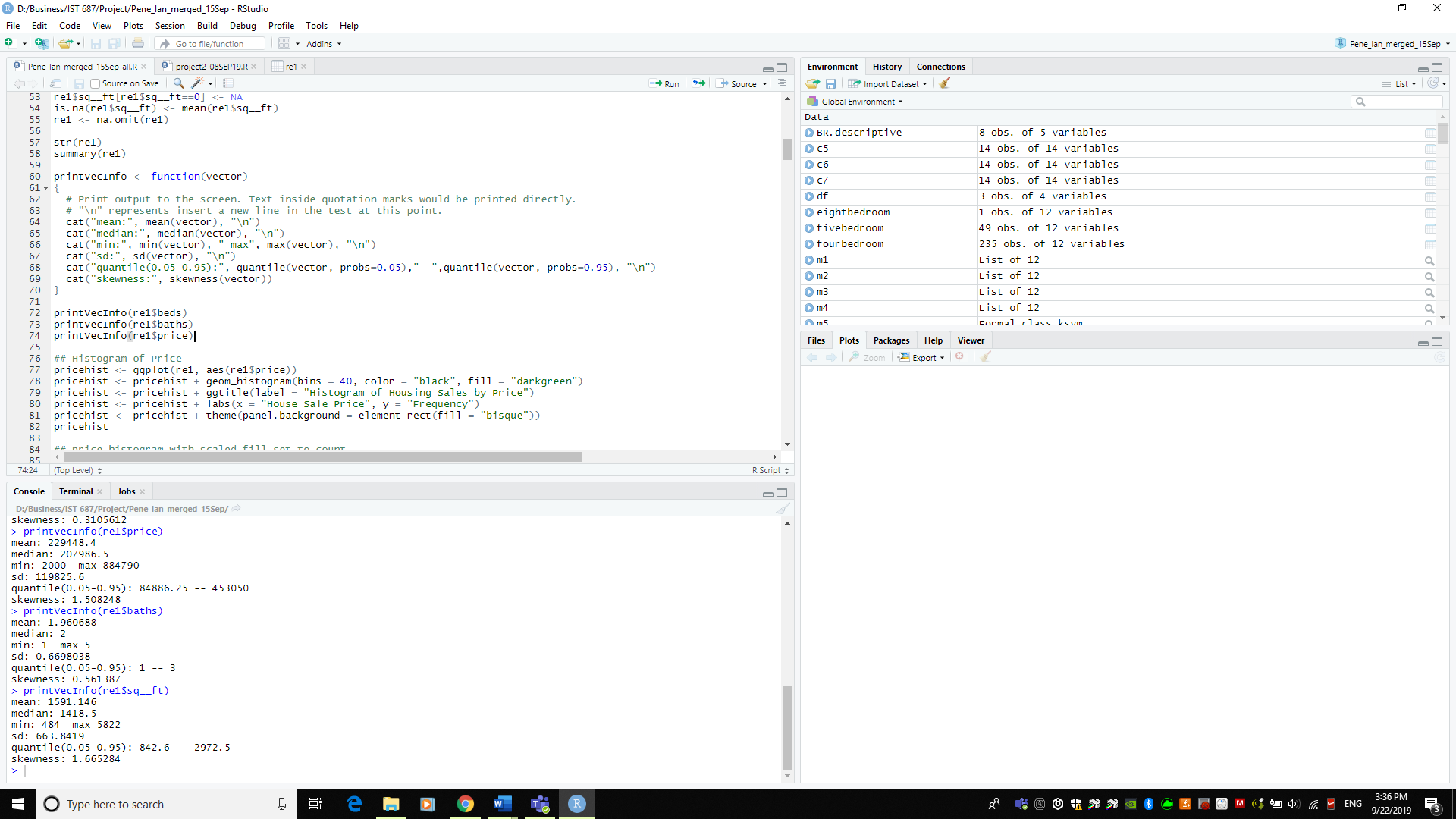
**Figure 2**



At first glance, we noted that approximately 90% of our observations were of type “residential,” with a fairly normal distribution of both beds and baths with a slight right skew. This provided us with enough information to run preliminary descriptive analysis.

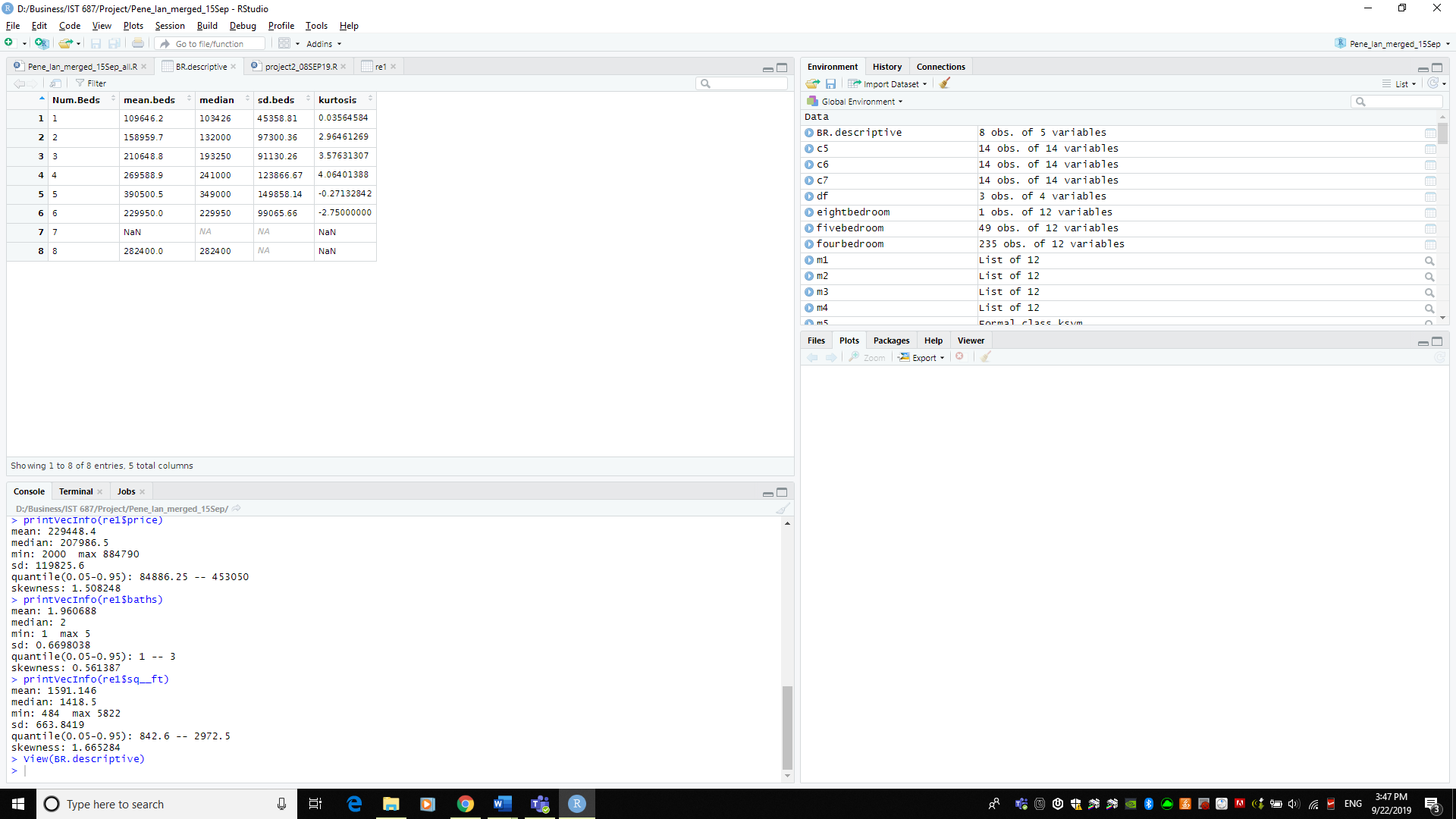
We utilized a function which prints the mean, median, minimum/maximum, standard deviation, quantile and skewness. Of greatest interest, these functions expressed that nearly all of our variables had a right skew, indicating that the majority of our data fell on the lower end of the overall range and that there were several outliers altering the data.

**Figure 3**

Next, we created a data frame which organized the data into 8 observations of 5 variables. Each row of the data represented the number of beds while the variables denoted the (price) mean, median, standard deviation and kurtosis for properties with the corresponding number of bedrooms.

**FIGURE 4**



It was interesting to note that the number of bedrooms did not directly correlate to a higher sales price. This is a result of multifamily, condos and residential homes all being grouped together under the “type” of property. A multifamily house with six bedrooms could cost less and did cost less on average than a single family “residential” home with four or five bedrooms due to other factors such as square feet or variables not accounted for in our dataset, such as lot size and neighborhood characteristics. Once we reach a higher number of bedrooms, e.g. six or seven, the limited amount of observations with this quantity of bedrooms created errors in the data frame, which we accounted for in our modeling process.

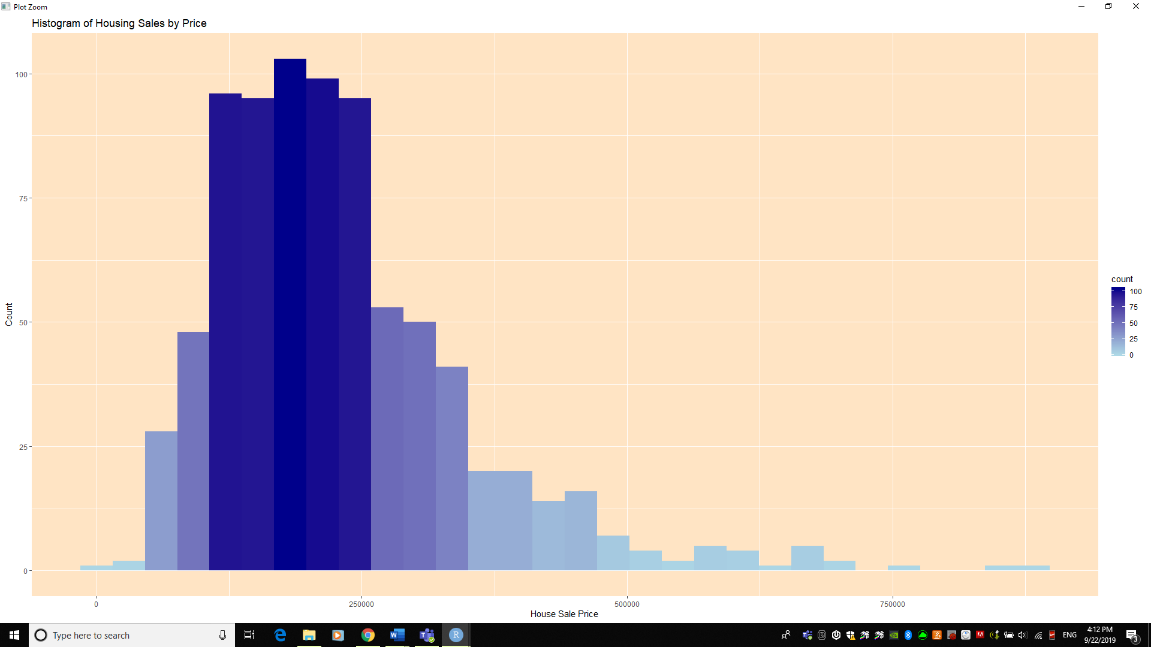
We utilized a series of graphing techniques to gain a visual representation of the data. These included:

* Histogram: housing sales by price
* Bar plot: sales price filled
* Density plot: type of property
* Scatterplot: price by square feet
* Boxplot: price by number of bedrooms
* Boxplot: price by property type
* Geographic map: points of home sales

Histogram: housing sales by price

In the graph below, fill is set to “count” which visually displays the corresponding bins closest to the median as a darker color. As was indicated earlier, this histogram of sales price showed a right skew, informing us that there were several outliers pulling the mean approximately $20k higher than the median. To gain greater understanding of what might be affecting this graph, we set fill to “type” to determine if the type of property was what caused this change.

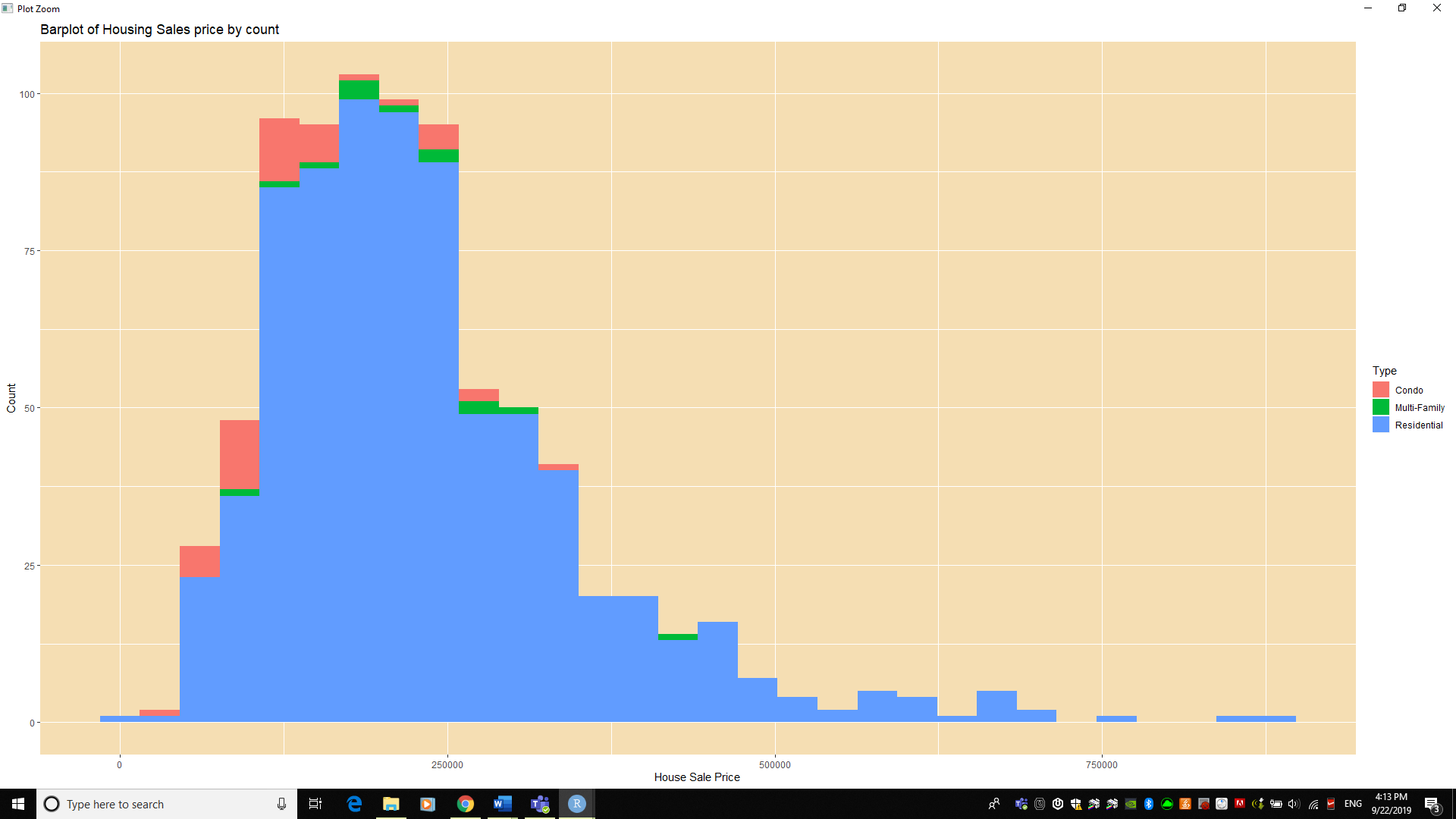
**Figure 5**



Bar plot: sales price filled

The graph below, with fill set to property type, we can discern that condos are disproportionately lower in price than residential and multi-family. These lower prices pull the overall mean distribution lower.

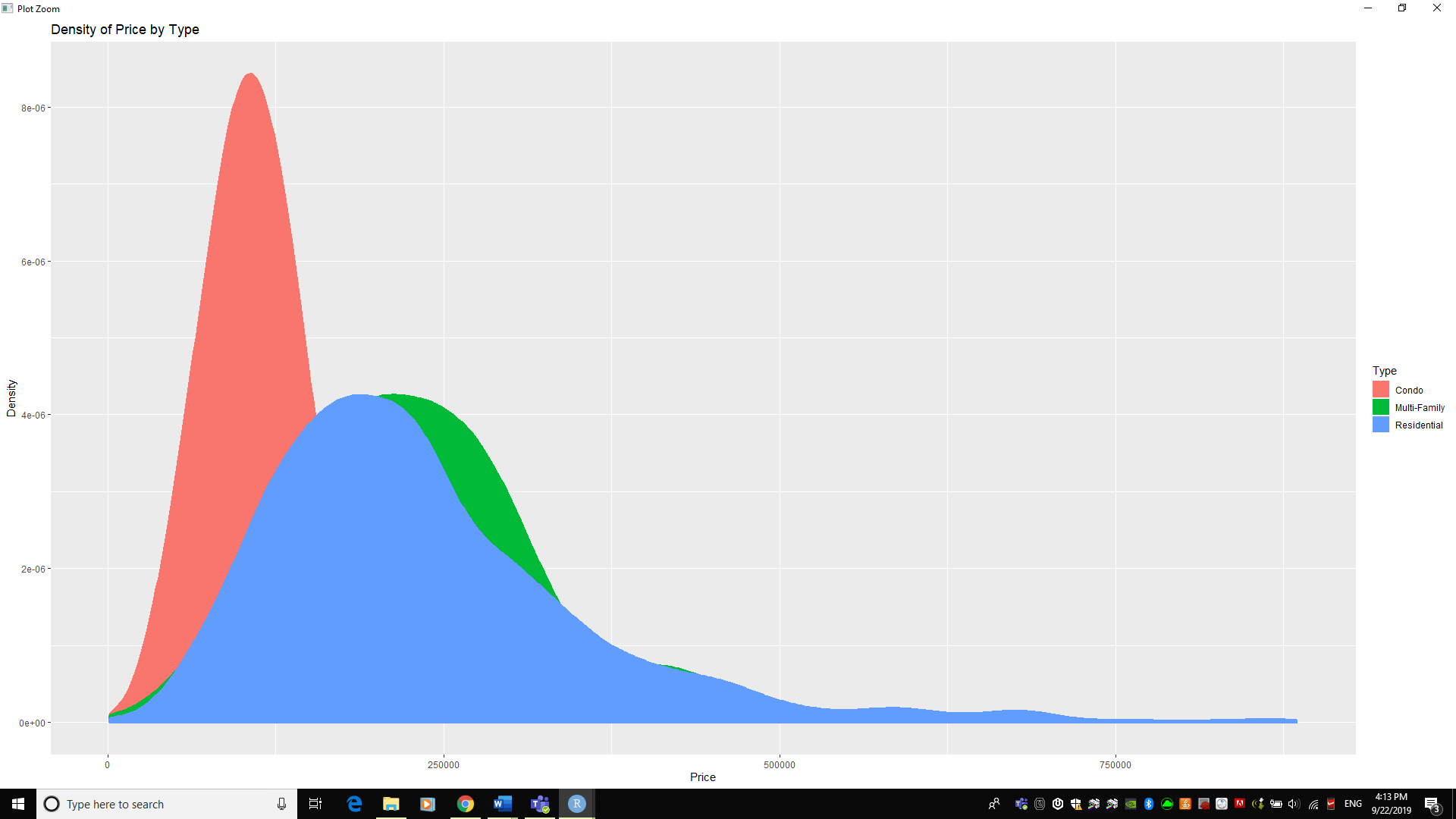
**Figure 6**



Density plot: type of property

This density graph below confirms our previous assumption of condos lowering the overall mean price as condos represent a much greater percentage of the lower sales price spectrum.

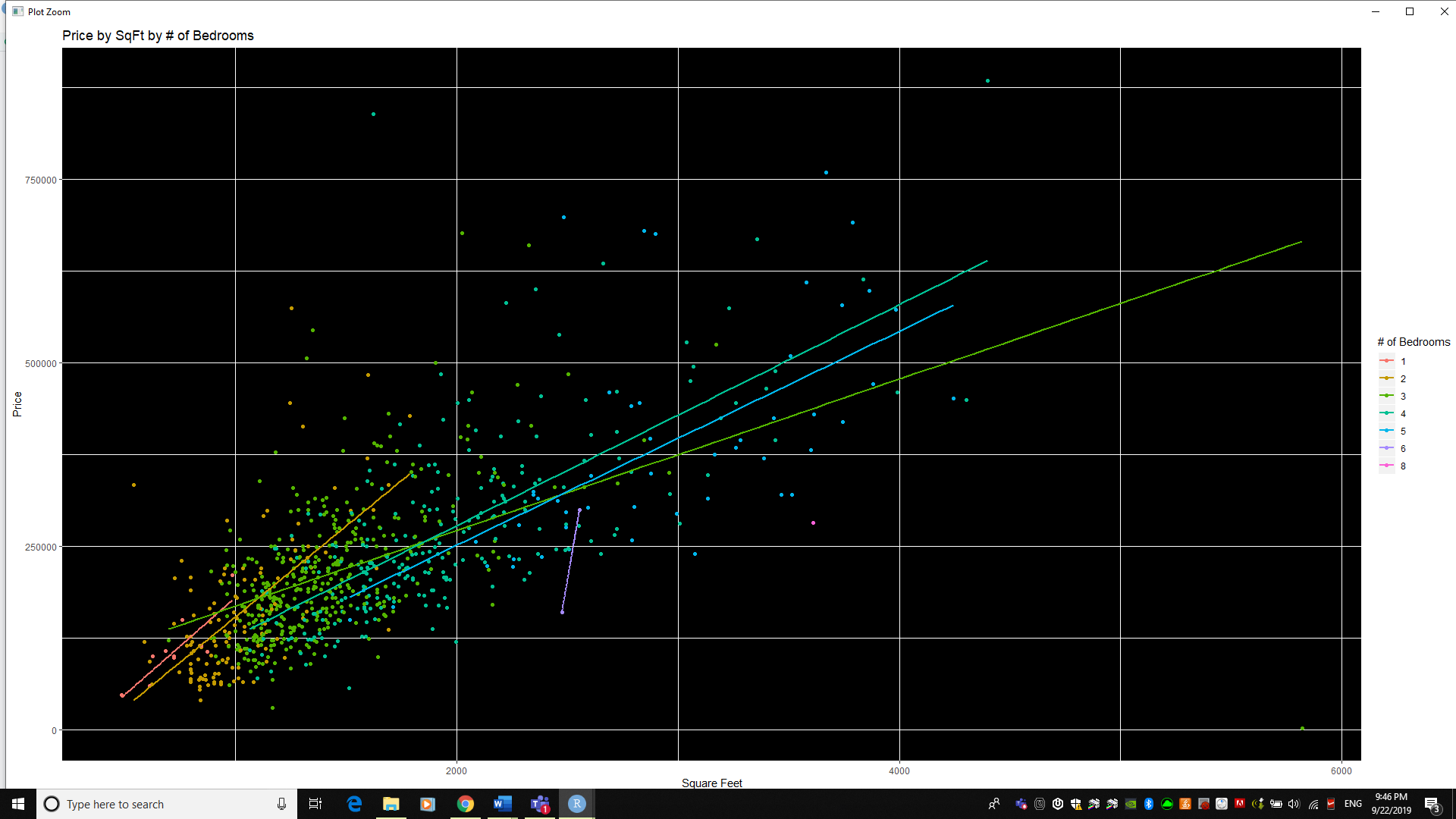
**Figure 7**



Scatterplot: price by square feet

Furthermore, the graph below shows the price of each property increases as square footage increases. The color of each point is set to the number of bedrooms in the property. While it is intuitive that both continuous variables of price and square footage share a positive correlation, the point color set to number of bedrooms provides additional insights. The data shows that there is a large concentration of one- to two-bedroom properties with a square footage less than 1900 square feet and price less than $300,000. There are relatively fewer one- to three-bedroom properties with square footage above 1,900, as those are reserved for four or more-bedroom properties. As square footage increases from this point onward, the density of datapoints begins to decrease and it becomes much more difficult to understand average sales price. Presumably, there are other factors affecting the overall price that are not represented in this data.

**Figure 8**

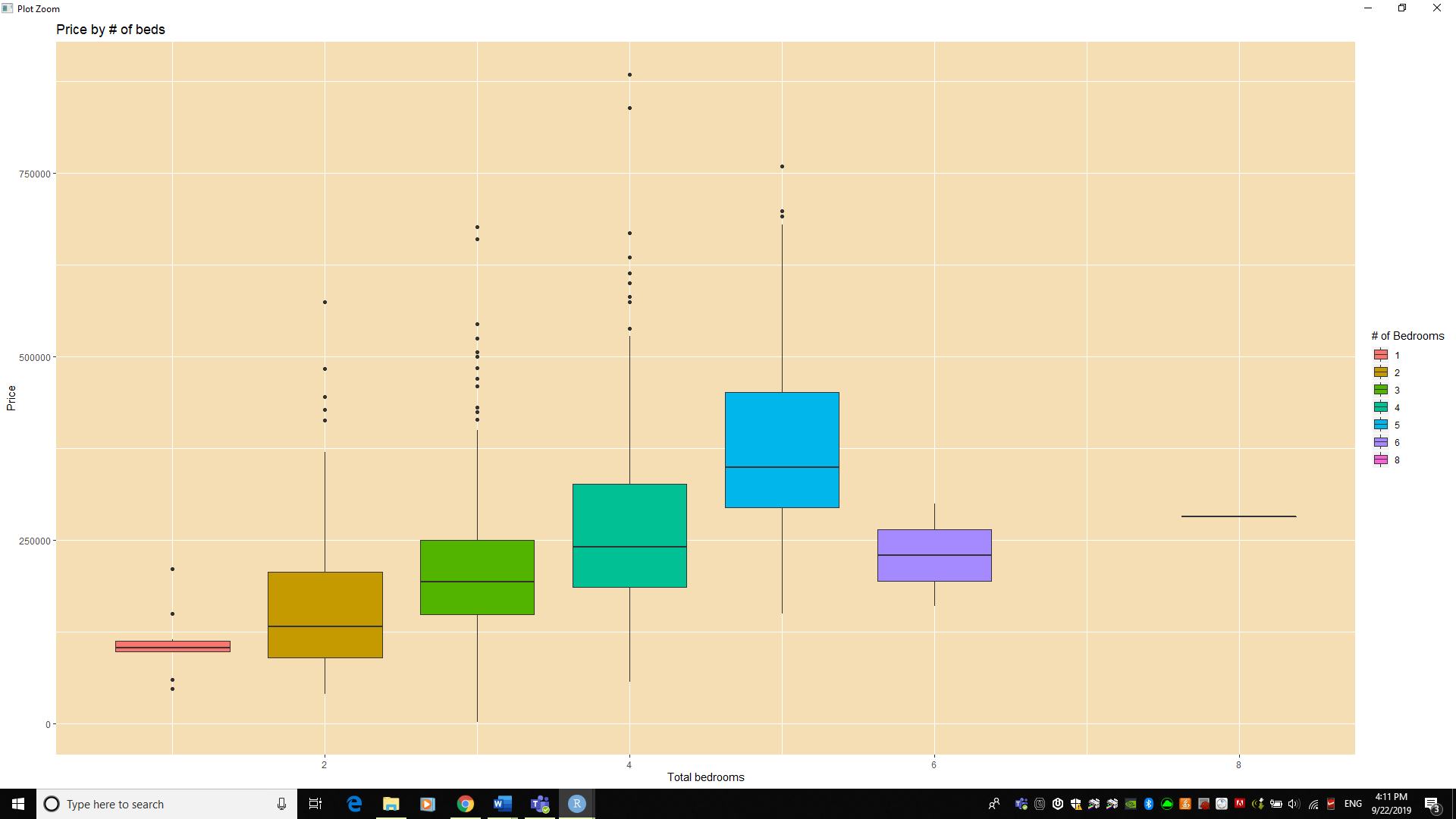


**POINT OF INTEREST:** The one- and two-bedroom properties show a steeper regression line of price to square feet. This means that people are willing to pay for more square feet even though the number of bedrooms remain the same. This is particularly true for condo properties. Additionally, these properties have a lower y-axis value, which is to be expected. Furthermore, the y-axis represents the base cost prior to a property being developed. This would include costs such as land, sewer, roads, utilities, etc. Conversely, for three- to five-bedroom properties the y-axis value is higher as the base costs would be higher. Lastly, for six- and eight- bedroom properties, there are only three datapoints available, and more data is needed to conclude any meaningful / reliable insight.

Boxplot: price by number of bedrooms

The graph below represents the sale price of the property as expressed by the categorical variable of “number of bedrooms.” As discussed earlier, four-bedroom properties have a higher average cost than six-bedroom properties. Six-bedroom properties have a near-identical sales price as three-bedroom properties, which brings into question the “type” of property the six-bedroom home belongs to.

**Figure 9**



Boxplot: price by property type

Additionally, we plotted a graph of sale price as expressed by the type of property, faceted by sale date. The intent was to show if sale date influenced the sale price for each type of property. We determined that it did not, due to the relatively constant results in the data. The only property with significant variation with sale price were condominiums.

**Figure 10**

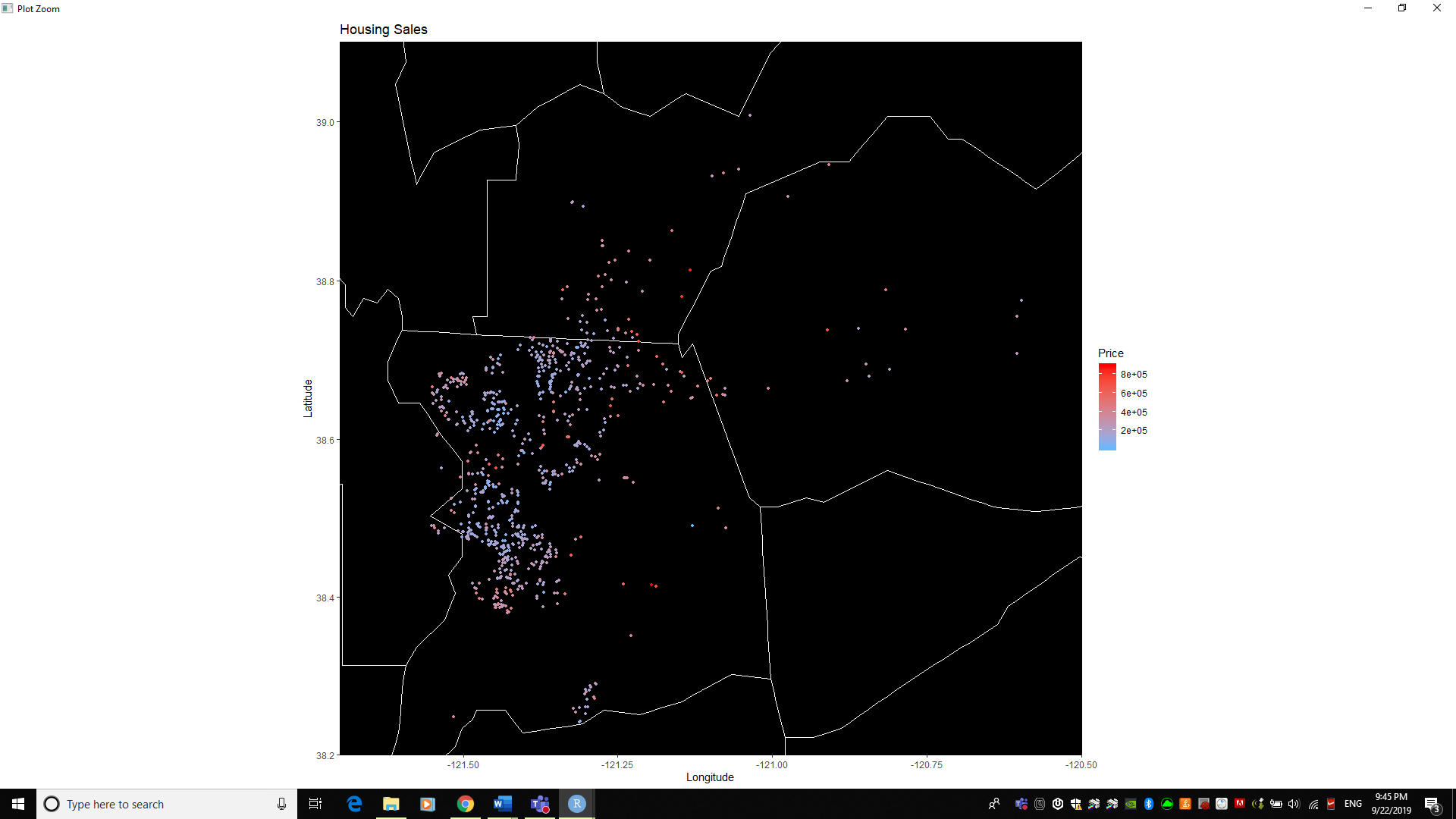


Geographic map: points of home sales

The graph below is a geographic representation of our dataset, plotted over a county map of California and enhanced to encompass all data points. The x-axis represents longitude and the y-axis represents latitude. Each point is signified by a point with alpha set to 0.5 to prevent over-plotting. The color of each point is set to price with blue expressing the lower values and red expressing the higher values. The intent with this graph was to determine if there were “pockets” of wealth or higher amenities which might inform our understanding of sales price.

While the majority of the datapoints rest within the city of Sacramento, CA; there are datapoints located to the northeast in Placerville, CA; and Elk Grove, CA. We could not determine if geographic circumstances influenced sales price due to the inability to identify pockets of higher sales price. To better understand geographic and socio-economic circumstances, we would need to understand livability index, crime rates, education levels and other factors outside of the scope of this project.

**Figure 11**



# Descriptive Continues: SCATTER AND BOX PLOTS

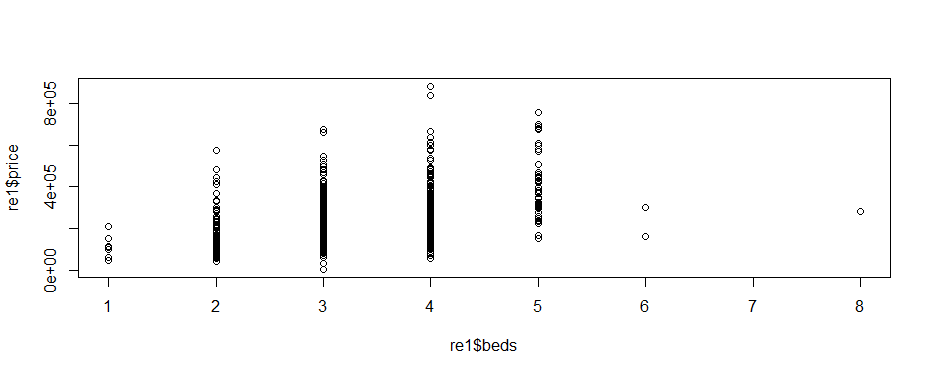
Additionally, we obtained the following visualizations to help us decide on the best modeling techniques to use. We present several scatter plots where “price” is the dependent variable (y-axis) and the number of beds, baths, and square feet (sqft) are independent variables (x-axis). Also, we present a box plot of different home type: condo, multi-family, and residential.

Intuitively, one expects that as the number of beds, baths, and sqft increase, so does price (i.e. positive correlation). The visualization presented below validates this assumption. However, there appears to be a celling for the number of beds and baths. For example, price increases from 1 to 4 beds, but it appears to flatten from 4 to 5 beds. The number of beds also shows a similar pattern, where price increases from 1 to 3 baths, but it flattens from 3-5 baths. The insight from this data is the number of beds and baths matters to price, but only to a certain point.

**Figure 12**

> # scatter plot beds and price

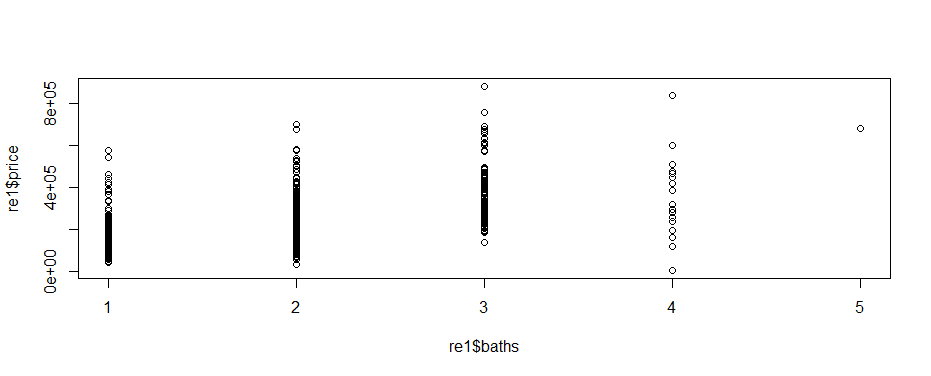
> plot (re1$beds, re1$price)



**Figure 13**

> #scatter plot baths and price

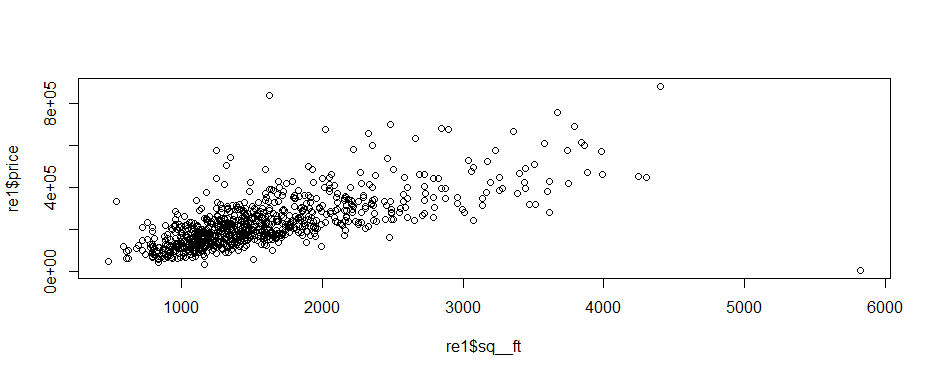
> plot (re1$baths, re1$price)



**Figure 14**

> #scatter plot sqft and price

> plot (re1$sq\_\_ft, re1$price)

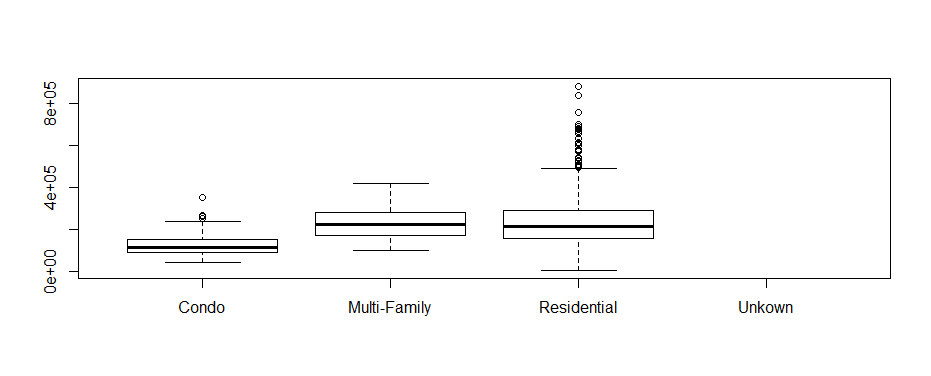


The box plot for condo, multi-family, and residential clearly shows condo as its own category, the minimum, median, and maximum values are materially different from the others. Multi-Family and Residential could arguably be treated in the same category, since the first quartile, median, and third quartile are similar. However, Residential does show several outliers and therefore, it would make sense to treat them separately. The insight from this data is the selected model will need to be performed for each home type separately.

**Figure 15**

> #scatter plot type and price

> plot (re1$type, re1$price)



# PREDICTIVE: LINEAR MODELS

For model techniques, we present linear model which is the basis of most predictive models. Linear models helps us understand the response of a dependent variable (y-axis) as a function of one or more independent variables (x-axis). For example, one may want to purchase a residential home that has 3 beds, 2 baths, and 1200 square feet (x-axis). A basic question would be, what is the price? Using the linear model technique, the price can be calculated as follow.

# re2 = residential

> m8 <- lm(formula=price ~ beds+baths+sq\_\_ft, data=re2)

> summary(m8)

Call:

lm(formula = price ~ beds + baths + sq\_\_ft, data = re2)

Residuals:

Min 1Q Median 3Q Max

-796568 -54082 -12986 37207 592876

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 47562.607 14210.704 3.347 0.000857 \*\*\*

beds -11920.799 5763.819 -2.068 0.038960 \*

baths 9285.505 7408.466 1.253 0.210461

sq\_\_ft 128.758 7.548 17.058 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87050 on 755 degrees of freedom

Multiple R-squared: 0.4795, Adjusted R-squared: 0.4774

F-statistic: 231.8 on 3 and 755 DF, p-value: < 2.2e-16

> test4=data.frame(beds=3,baths=2,sq\_\_ft=1200)

> p4 <- predict(m8,test4,type="response")

> p4

1

184880.2

The linear model above shows that in this example, one should expect the price to be $184,880.

There other important statistical output that is worth noting. For instance, the p-value is 2.2e-16, which is a small number (close to zero) and less than 0.05. As a rule of thumb, when one observes a p-value that is less than 0.05, one can call that “statistically significant”, and can take away a sense that the results were not due to randomness (a problem known as “sampling error”).

Additionally, the Adjusted R-squared value is 0.4774. This means that about 48% of the variation in price can be explained by the number of beds, baths, and square feet (sqft). The R-squared value is also known as the coefficient of determination. The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables. The closer to 1.0, the greater the influence the independent variable has on predicting the value of the dependent variable.

Assuming one can tolerate that only 48% of the variation in price can be explained by beds, baths, and square feet, then it’s safe to assume that this model can be used.

In contrast, a linear model for multi-family homes could not be used and for the following reasons. For example, for a multi-family home with 5 beds, 3 baths, and 2300 square feet, one would expect the price to be $235,943. However, the p-value is 0.8489 (greater than 0.05), which means the independent variables are “not” statistically significant. Also, the Adjusted R-squared value is -0.2251. This means price has a negative correlation to beds, baths, and square feet, which is not intuitive. Given this insight, one can assume that this model should not be used, and that more data needs to be collected.

# re3 = multi-family

> m9 <- lm(formula=price ~ beds+baths+sq\_\_ft, data=re3)

> summary(m9)

Call:

lm(formula = price ~ beds + baths + sq\_\_ft, data = re3)

Residuals:

Min 1Q Median 3Q Max

-95147 -75864 -14071 54162 185511

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 225489 99060 2.276 0.0489 \*

beds -41942 57898 -0.724 0.4872

baths -23996 37188 -0.645 0.5349

sq\_\_ft 127 147 0.864 0.4100

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94150 on 9 degrees of freedom

Multiple R-squared: 0.08121, Adjusted R-squared: -0.2251

F-statistic: 0.2652 on 3 and 9 DF, p-value: 0.8489

> test4=data.frame(beds=5,baths=3,sq\_\_ft=2300)

> p4 <- predict(m9,test4,type="response")

> p4

1

235942.7

The linear model for condos fares well than the earlier models. The p-value is less than 0.05 and Adjusted R-squared is 0.6343. This means the data is statistically significant and 63% of price variation is influenced by beds, baths, and square feet. Assuming one can tolerate a 63% of price variation due to known independent variables, then it is safe to assume the model can be used. For example, a condo with 2 beds, 1 bath, and 900 square feet, one can expect the price to be $110,829.

# re4 = condo

> m10 <- lm(formula=price ~ beds+baths+sq\_\_ft, data=re4)

> summary(m10)

Call:

lm(formula = price ~ beds + baths + sq\_\_ft, data = re4)

Residuals:

Min 1Q Median 3Q Max

-57357 -27542 -7026 25567 152026

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -10032.08 24193.73 -0.415 0.6807

beds -42899.27 16282.30 -2.635 0.0121 \*

baths 4581.03 17635.70 0.260 0.7965

sq\_\_ft 224.53 36.31 6.183 3.19e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41280 on 38 degrees of freedom

Multiple R-squared: 0.6611, Adjusted R-squared: 0.6343

F-statistic: 24.7 on 3 and 38 DF, p-value: 4.877e-09

> test4=data.frame(beds=2,baths=1,sq\_\_ft=900)

> p4 <- predict(m10,test4,type="response")

> p4

1

110829

# DATA MINING: LINEAR, KSVM, SVM MODELS

Another modeling technique that can be applied is Supervised Machine Learning, where two thirds of the data is established for training purposes and the remaining one third is set aside for testing. For example, the 759 records for residential homes can be divided such that 506 records are used for training purposes, and the remaining 253 records are used for testing purposes.

# Create train and test data sets for residential homes

> nrows <- nrow(re2)

> nrows

[1] 759

> cutPoint <- floor(nrows/3\*2)

> cutPoint

[1] 506

> rand <- sample(1:nrows)

> head(rand)

[1] 63 43 339 328 155 470

> re2.train <- re2[rand[1:cutPoint],]

> re2.test <- re2[rand[(cutPoint+1):nrows],]

> nrow(re2.train)

[1] 506

> nrow(re2.test)

[1] 253

Subsequently, different modeling techniques such as Support Vector Machine models (SVM and KSVM) and Linear Model (LM) can then be created, trained, and tested. The output from this computation is the RMSE value or Root Mean Squared Error. RMSE is the standard deviation of the prediction errors (residuals), and prediction errors are simply the differences between the observed values and predicted values. RMSE is a measure of how spread out these errors are. In other words, it shows how concentrated the data is around the “line of best fit”. Therefore, the smaller the RMSE value is the more optimal the model is in predicting observed values.

As an example, for residential homes the RMSE values for SVM, KSVM, and LM models are $81,498, $95,008, and $78,423, respectively. Therefore, the result suggests that in this case, the Liner Model (LM) is the optimal model to use for residential homes.

> # Build svm model

> m6 <- svm(price ~ beds+baths+sq\_\_ft, data=re2.train)

> p6 <- predict(m6,re2.test)

> c6 <- data.frame(re2.test,p6)

> c6$error <- c6$price-c6$p6

> re2rmse2 <- sqrt(mean((c6$error)^2))

> re2rmse2

[1] 81498.01

> # Build ksvm model

> m5 <- ksvm(price ~ beds+baths+sq\_\_ft, data=re2.train)

> p5 <- predict(m5,re2.test)

> c5 <- data.frame(re2.test,p5)

> c5$error <- c5$price-c5$p5

> re2rmse1 <- sqrt(mean((c5$error)^2))

> re2rmse1

[1] 95007.59

# Build linear (lm) model

> m7 <- lm(price ~ beds+baths+sq\_\_ft, data=re2.train)

> p7 <- predict(m7,re2.test)

> c7 <- data.frame(re2.test,p7)

> c7$error <- c7$price-c7$p7

> re2rmse3 <- sqrt(mean((c7$error)^2))

> re2rmse3

[1] 78423.38

Similarly for condos, the RMSE for SVM, KSVM, and LM models are $56,885, $58,031, and $38,237, respectively. Again, the result suggests that in this case, the Liner Model (LM) is the optimal model to use for condo homes, since the RMSE value for LM is smaller than the RSMSE values for SVM, and KSVM.

> # Create train and test data sets

> nrows <- nrow(re4)

> nrows

[1] 42

> cutPoint <- floor(nrows/3\*2)

> cutPoint

[1] 28

> rand <- sample(1:nrows)

> head(rand)

[1] 28 12 38 7 34 18

> re4.train <- re4[rand[1:cutPoint],]

> re4.test <- re4[rand[(cutPoint+1):nrows],]

> nrow(re4.train)

[1] 28

> nrow(re4.test)

[1] 14

> # Build ksvm model

> m5 <- ksvm(price ~ beds+baths+sq\_\_ft, data=re4.train)

> p5 <- predict(m5,re4.test)

> c5 <- data.frame(re4.test,p5)

> c5$error <- c5$price-c5$p5

> re4rmse1 <- sqrt(mean((c5$error)^2))

> re4rmse1

[1] 56885.19

>

> # Build svm model

> m6 <- svm(price ~ beds+baths+sq\_\_ft, data=re4.train)

> p6 <- predict(m6,re4.test)

> c6 <- data.frame(re4.test,p6)

> c6$error <- c6$price-c6$p6

> re4rmse2 <- sqrt(mean((c6$error)^2))

> re4rmse2

[1] 58031.45

>

> # Build lm model

> m7 <- lm(price ~ beds+baths+sq\_\_ft, data=re4.train)

> p7 <- predict(m7,re4.test)

> c7 <- data.frame(re4.test,p7)

> c7$error <- c7$price-c7$p7

> re4rmse3 <- sqrt(mean((c7$error)^2))

> re4rmse3

[1] 38236.85

For multi-family homes, more data needs to be collected as noted earlier. Applying Supervised Machine Learning in this case, using the current available data will not yield reliable results.

# OVERALL INTERPRETATION OF RESULTS/ACTIONABLE INSIGHTS

The Flinders presented our team with the four question topics. Their questions and our answers follow:

1. Q – *“What is the overall price range for properties in this market?”*
   * A - $2,000 – $884,790 - [**Figure 2**]
2. Q – *“We prefer another 5-bedroom home. What is the average price in this region?”*
   * A - The average for 5-bedroom homes is $390,501 [**Figure 4**]. However, in that price variance and standard deviation have a positive correlation; the standard deviation in **Figure 4** shows that these homes have the broadest range in pricing with 4-bedroom homes close behind. This condition is further evidenced in **Figure 8**.
3. Q - *“Having a potential two-home budget, we need to be mindful of cost. Which property feature is the best measure for reducing our final price (cost-driver)?*
   * Of the most impacting cost contributors, square footage provides the strongest correlation in price rises/falls within each respective property type. [**Figure 14**]
4. Q – *“How about condos? What are the current pricing conditions?”*
   * As shown in **Figures 7**, condo prices are among the most affordable; averaging ~$110,000 and with a clear majority falling relatively close to that range,~$90,000 to ~$150,000. **Figure 7** also highlights that condos offer the most affordable and least variable prices within all property types.
5. *“Thank you for your time and interest. We have one last question, after having taken all of this in, we realize that we have a $450,000 budget, but wonder if there is enough room for us to purchase the 5-bedroom home and a smaller investment property which our college kids can use. What do you think?”*
   * As discussed in Q-2, 5-bedroom homes have the greatest price variability around a $390,000 average and condos a rather narrow variability near $110,000. This equates to a an average $500,000 spend or $50,000 shortage in your case.
   * With your future home having the much greater price variation and the condo relatively little, I would recommend that you first obtain a 5-bedroom home ~$330,000 which is of course livable, but underpriced due to needed improvements which you feel you can reasonably complete over the course of your occupancy. This will provide you the 5-bedroom space you need in addition to another investment opportunity as you increase your equity via home improvements.
   * After we have a sales contract in place on your next home, we can then gain a firmer idea on your remaining $170,000 budget. If your primary (5-bedroom) home requires immediate improvements, you will have ample reserve in the budget to meet the need for a typically priced condo or even two value-priced units.

# Appendix

## Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Variable Name** | **Variable Type** | **Variable Description** | **Example** |
| **1** | **Street** | **Text/Numeric** | **Home street address of the property** | **3526 HIGH ST** |
| **2** | **City** | **Text** | **City of the property** | **SACRAMENTO** |
| **3** | **Zip** | **Numeric** | **Zip Code** | **95838** |
| **4** | **State** | **Text** | **State of the property** | **CA** |
| **5** | **Beds** | **Numeric** | **Number of the bedroom in the house** | **2** |
| **6** | **Baths** | **Numeric** | **Number of the bathroom in the house** | **1** |
| **7** | **sq.\_\_ft** | **Numeric** | **Home Square feet** | **836** |
| **8** | **Type** | **Text** | **Type of home, code, single family, multifamily** | **Residential** |
| **9** | **Sale\_date** | **Date** | **Date of the sales** | **Wed May 21 00:00:00 EDT 2008** |
| **10** | **Price** | **Numeric** | **Price of the property** | **59222** |
| **11** | **Latitude** | **Numeric** | **Latitude of the property** | **38.631913** |
| **12** | **Longitude** | **Numeric** | **Longitude of the property** | **121.434879** |