Exercise 11 - Housing Data analysis and Prediction

Ninad Patkhedkar

2020-10-18

Work individually on this assignment. You are encouraged to collaborate on ideas and strategies pertinent to this assignment. Data for this assignment is focused on real estate transactions recorded from 1964 to 2016 and can be found in Week 6 Housing.xlsx. Using your skills in statistical correlation, multiple regression and R programming, you are interested in the following variables: Sale Price and several other possible predictors.

```
## # A tibble: 6 x 24
##   `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning
##   <dttm>                     <dbl>       <dbl>           <dbl> <chr>
## 1 2006-01-03 00:00:00       698000           1               3 <NA>
## 2 2006-01-03 00:00:00       649990           1               3 <NA>
## 3 2006-01-03 00:00:00       572500           1               3 <NA>
## 4 2006-01-03 00:00:00       420000           1               3 <NA>
## 5 2006-01-03 00:00:00       369900           1               3 15
## 6 2006-01-03 00:00:00       184667           1              15 18 51
## # ... with 19 more variables: sitetype <chr>, addr_full <chr>, zip5 <dbl>,
## #   ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
## #   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
## #   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>,
## #   year_built <dbl>, year_renovated <dbl>, current_zoning <chr>,
## #   sq_ft_lot <dbl>, prop_type <chr>, present_use <dbl>
```

a) Explain why you chose to remove data points from your 'clean' dataset.

Below variables/columns doesn't seems to affect Sale price in my opinion. Hence thsese won't be considered.
sale_instrument sitetype addr_full postalcyn lon lat prop_type present_use

```
clean_housing_df = subset(housing_df, select = -c(sale_reason,sale_instrument,sitetype,addr_full,postalc
head(clean_housing_df)
```

```
## # A tibble: 6 x 15
##   `Sale Date`         `Sale Price` sale_warning  zip5 ctyname building_grade
##   <dttm>                     <dbl> <chr>        <dbl> <chr>            <dbl>
## 1 2006-01-03 00:00:00       698000 <NA>         98052 REDMOND              9
## 2 2006-01-03 00:00:00       649990 <NA>         98052 REDMOND              9
## 3 2006-01-03 00:00:00       572500 <NA>         98052 <NA>                 8
## 4 2006-01-03 00:00:00       420000 <NA>         98052 REDMOND              8
## 5 2006-01-03 00:00:00       369900 15           98052 REDMOND              7
## 6 2006-01-03 00:00:00       184667 18 51        98053 <NA>                 7
## # ... with 9 more variables: square_feet_total_living <dbl>, bedrooms <dbl>,
## #   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>,
## #   year_built <dbl>, year_renovated <dbl>, current_zoning <chr>,
## #   sq_ft_lot <dbl>
```

b) Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price and several additional predictors of your choice. Explain the basis for your additional predictor selections.

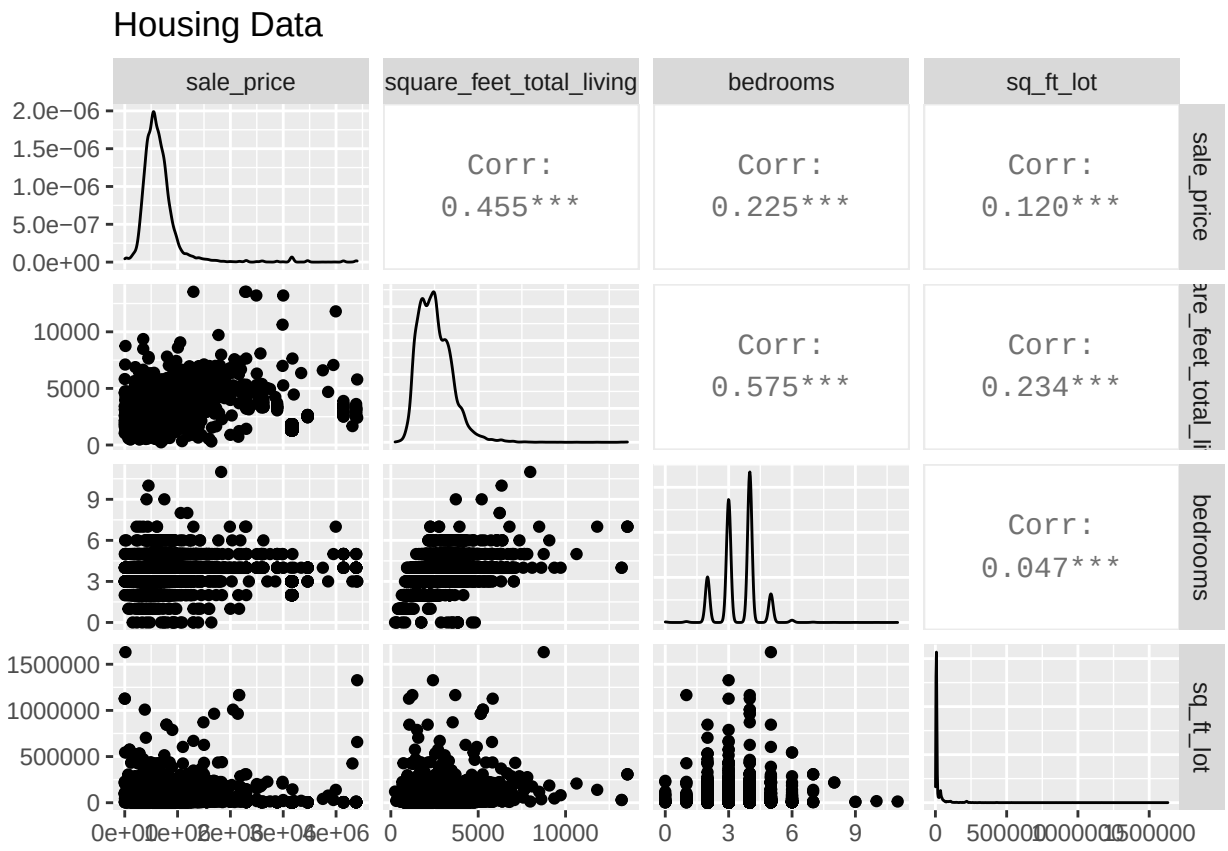
There are multiple variables like year, renovated year etc. which I am not aware of how to account for. Also there are some categorical variables like zipcodes, current zoning, warning those would also have small effect but not significant. There are also some ordinal variables like sale-reason etc.

I am considering simple numeric fields for Model which include square_feet_total_living, number of bedrooms, sq_ft_lot. For bathrooms I just sum up full, half and 3qtr bathroom and considered it as single numenric value.

I used ggpairs to check the correlation.

```
clean_housing_df = subset(housing_df, select = -c(1,sale_reason,sale_instrument,sale_warning,zip5,ctynar
clean_housing_df <- rename(clean_housing_df,sale_price = "Sale Price")

ggpairs(data=clean_housing_df,columns =c(1,2,3,7) , title="Housing Data")
```




```
housing_lm <- lm(sale_price ~ square_feet_total_living + bedrooms + sq_ft_lot,data=clean_housing_df)
summary(housing_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     sq_ft_lot, data = clean_housing_df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1940694  -118436   -40283    43781  3783524
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.431e+05  1.304e+04  18.642  < 2e-16 ***
## square_feet_total_living  1.972e+02  4.050e+00  48.694  < 2e-16 ***
## bedrooms                 -2.433e+04  4.454e+03  -5.463 4.76e-08 ***
## sq_ft_lot                 6.596e-02  5.766e-02   1.144    0.253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 359800 on 12861 degrees of freedom
## Multiple R-squared:  0.2087, Adjusted R-squared:  0.2085
## F-statistic:  1131 on 3 and 12861 DF,  p-value: < 2.2e-16
```

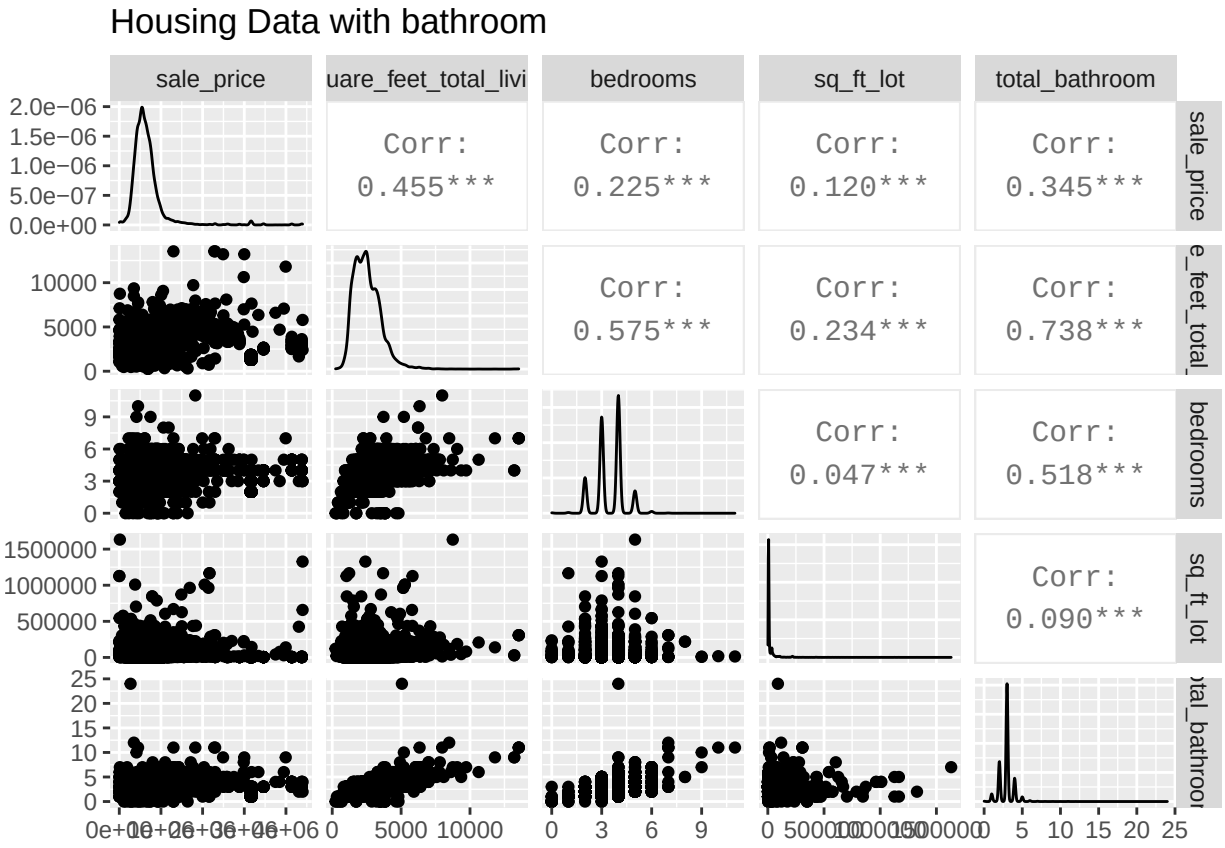
Adding number of bathroom counts in Linear Model

```
clean_housing_df$total_bathroom <-clean_housing_df$bath_full_count + clean_housing_df$bath_half_count +

clean_housing_df <- subset(clean_housing_df,select = -c(bath_full_count,bath_half_count,bath_3qtr_count)
head(clean_housing_df)
```

```
## # A tibble: 6 x 5
##   sale_price square_feet_total_living bedrooms sq_ft_lot total_bathroom
##        <dbl>                    <dbl>    <dbl>     <dbl>          <dbl>
## 1     698000                     2810        4      6635              3
## 2     649990                     2880        4      5570              3
## 3     572500                     2770        4      8444              3
## 4     420000                     1620        3      9600              2
## 5     369900                     1440        3      7526              2
## 6     184667                     4160        4      7280              4
```

```
ggpairs(data=clean_housing_df, title="Housing Data with bathroom")
```




```
housing_lm <- lm(sale_price ~ square_feet_total_living + bedrooms + sq_ft_lot + total_bathroom ,data=cl
summary(housing_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     sq_ft_lot + total_bathroom, data = clean_housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1974696  -118280   -40135    43904  3781461
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.268e+05  1.416e+04  16.009  < 2e-16 ***
## square_feet_total_living  1.875e+02  5.224e+00  35.892  < 2e-16 ***
## bedrooms                 -2.643e+04  4.509e+03  -5.862 4.69e-09 ***
## sq_ft_lot                 8.466e-02  5.799e-02   1.460  0.14434
## total_bathroom            1.648e+04  5.597e+03   2.944  0.00324 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 359700 on 12860 degrees of freedom
## Multiple R-squared:  0.2092, Adjusted R-squared:  0.209
## F-statistic: 850.6 on 4 and 12860 DF,  p-value: < 2.2e-16
```

c) Execute a summary() function on two variables defined in the previous step to compare the model results. What are the R2 and Adjusted R2 statistics? Explain what these results tell you about the overall model. Did the inclusion of the additional predictors help explain any large variations found in Sale Price?

R2 and Adjusted R2 values before adding extra predictor variable i.e. Bathrooms Multiple R-squared: 0.2087, Adjusted R-squared: 0.2085

R2 and Adjusted R2 values after adding extra predictor variable i.e. Bathrooms Multiple R-squared: 0.2092, Adjusted R-squared: 0.209

So adding extra predictor variable increased R2 square values a bit but not significantly. Though additional variable selected improves Model , it doesn't very significantly.

d) Considering the parameters of the multiple regression model you have created. What are the standardized betas for each parameter and what do the values indicate?

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     sq_ft_lot + total_bathroom, data = clean_housing_df)
##
## Standardized Coefficients::
##               (Intercept) square_feet_total_living                bedrooms
##                0.00000000               0.45895498             -0.05726393
##                 sq_ft_lot           total_bathroom
##                0.01191913               0.03493640
```

square_feet_total_living is the most significant deciding factor with value 0.45. Coefficient for bedroom is showing negative. Hence I suspect something is wrong with my assumption or model.

e) Calculate the confidence intervals for the parameters in your model and explain what the results indicate.

```
##                                   5 %          95 %
## (Intercept)               2.034549e+05  2.500537e+05
## square_feet_total_living  1.789085e+02  1.960953e+02
## bedrooms                 -3.384756e+04 -1.901337e+04
## sq_ft_lot                -1.073104e-02  1.800475e-01
## total_bathroom            7.271208e+03  2.568502e+04
```

f) Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.

Not sure what to do here. Its already done in #b above and Model didn't improve significantly as R square remains same.

g) Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     698 1300000 1540000 1904783 2583000 4400000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5050    5305    5770    6153    6355   13540
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   6.000   4.362   6.000  11.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23618   35302   45738   85673   93845 1631322
```

There are some big outliers for sale_price, bedroom size and sq_ft_lot ## h) Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create. housing.res = resid(housing_lm) summary(housing.res)

i) Use the appropriate function to show the sum of large residuals.

TO-DO

j) Which specific variables have large residuals (only cases that evaluate as TRUE)?

TO-DO ## k) Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics. TO-DO ## l) Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not. TO-DO ## m) Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not. TO-DO ## n) Visually check the assumptions related to the residuals using the plot() and hist() functions. Summarize what each graph is informing you of and if any anomalies are present.

o) Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

TO-DO