Kalaivani Kalyanasundaram – DSC Week 8
Github: https://github.com/kalaikalyan/hello-world

Housing Data

1. 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 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.
    1. If you worked with the Housing dataset in previous week – you are in luck, you likely have already found any issues in the dataset and made the necessary transformations. If not, you will want to take some time looking at the data with all your new skills and identifying if you have any clean up that needs to happen.
2. Complete the following:
    1. Explain any transformations or modifications you made to the dataset

I split the dataset using the plyr function so that only important factors of our analysis are seen through the data. I did some aggregation and groupby to see any trends in the data.

    2. 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.

```
predict_lm <- lm(housing_df$`Sale Price` ~ housing_df$square_feet_total_living, housing_df)
predict_lm2 <- lm(housing_df$`Sale Price` ~ housing_df$square_feet_total_living + housing_df$bedrooms + housing_df$bath_full_count)
```

Usually the number of bedrooms and bathrooms influence the sale price. The higher the count, the more expensive houses are in my view.

    3. 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?

```
> summary(predict_lm)

Call:
lm(formula = housing_df$`Sale Price` ~ housing_df$square_feet_total_living,
   data = housing_df)

Residuals:
   Min     1Q  Median     3Q    Max
-1800136  -120257   -41547    44028  3811745

Coefficients:
                                Estimate Std. Error t value
(Intercept)                     1.891e+05  8.745e+03   21.62
housing_df$square_feet_total_living 1.857e+02  3.208e+00   57.88
                                Pr(>|t|)
```

```
(Intercept)                    <2e-16 ***
housing_df$square_feet_total_living  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 360200 on 12863 degrees of freedom
Multiple R-squared:  0.2066,        Adjusted R-squared:  0.2066
F-statistic:  3351 on 1 and 12863 DF,  p-value: < 2.2e-16

> summary(predict_lm2)

Call:
lm(formula = housing_df$`Sale Price` ~ housing_df$square_feet_total_living +
    housing_df$bedrooms + housing_df$bath_full_count)

Residuals:
    Min      1Q  Median      3Q     Max
-1760583  -117559   -41529    43918  3832099

Coefficients:
                              Estimate Std. Error t value
(Intercept)                 204679.013  14013.468  14.606
housing_df$square_feet_total_living    184.150     4.353  42.302
housing_df$bedrooms          -25206.328   4417.404  -5.706
housing_df$bath_full_count     42309.808   5685.497   7.442
                             Pr(>|t|)
(Intercept)                   < 2e-16 ***
housing_df$square_feet_total_living  < 2e-16 ***
housing_df$bedrooms           1.18e-08 ***
housing_df$bath_full_count    1.06e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 359000 on 12861 degrees of freedom
Multiple R-squared:  0.212,  Adjusted R-squared:  0.2118
F-statistic:  1153 on 3 and 12861 DF,  p-value: < 2.2e-16
```

4. 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?

```
library(lm.beta)
lm.beta(predict_lm)
lm.beta(predict_lm2)
```

The beta values clearly indicate a positive relationship between sale price and square feet and full bathrooms, but a negative correlation between sale price and bedrooms.

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

```
> confint(predict_lm)
                                2.5 %      97.5 %
```

```
(Intercept)                   171965.2516 206247.8664
housing_df$square_feet_total_living   179.4286   192.0067
> confint(predict_lm2)
                                2.5 %     97.5 %
(Intercept)                   177210.5359 232147.4899
housing_df$square_feet_total_living   175.6175   192.6834
housing_df$bedrooms            -33865.0966 -16547.5597
housing_df$bath_full_count      31165.3899  53454.2268
```

Based on the confidence interval, I am 95% sure that there is a positive relationship between the sale price nd square footage and number of full bathrooms. Also, a negative relationship between sale price and bedrooms.

6. 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.

```
> aov(predict_lm)
Call:
  aov(formula = predict_lm)

Terms:
                housing_df$square_feet_total_living    Residuals
Sum of Squares                  4.347032e+14 1.668870e+15
Deg. of Freedom                        1       12863

Residual standard error: 360197.1
Estimated effects may be unbalanced
> aov(predict_lm2)
Call:
  aov(formula = predict_lm2)

Terms:
                housing_df$square_feet_total_living housing_df$bedrooms
Sum of Squares                  4.347032e+14       4.092943e+12
Deg. of Freedom                        1                  1
                housing_df$bath_full_count    Residuals
Sum of Squares             7.137738e+12 1.657640e+15
Deg. of Freedom                   1       12861

Residual standard error: 359010.9
Estimated effects may be unbalanced
```

Residuals decreased so first prediction is better.

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

```
library(car)
outliers <- outlierTest(predict_lm)
outliers2 <- outlierTest(predict_lm2)
resid <- resid(predict_lm)
resid2 <-resid(predict_lm2)
```

8. Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create

```
standard_res <- rstandard(predict_lm)
standard_res2 <- rstandard(predict_lm2)
```

9. Use the appropriate function to show the sum of large residuals.

```
sum(resid, na.rm = TRUE)
sum(resid2, na.rm= TRUE)
```

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

```
resid <- resid(predict_lm)
resid2 <-resid(predict_lm2)
```

11. Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.

```
leverage <- hatvalues(predict_lm)
leverage2 <- hatvalues(predict_lm2)
cookD <- cooks.distance(predict_lm)
cookD2 <- cooks.distance(predict_lm2)
cov(housing_df$`Sale Price`, housing_df$sq_ft_lot)
cov(housing_df$`Sale Price`,housing_df$bath_full_count)
```

The problematic cases take away from the data and prove to be tricky to deal with. However, for the most part, the data values reinforce my assumptions.

12. Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

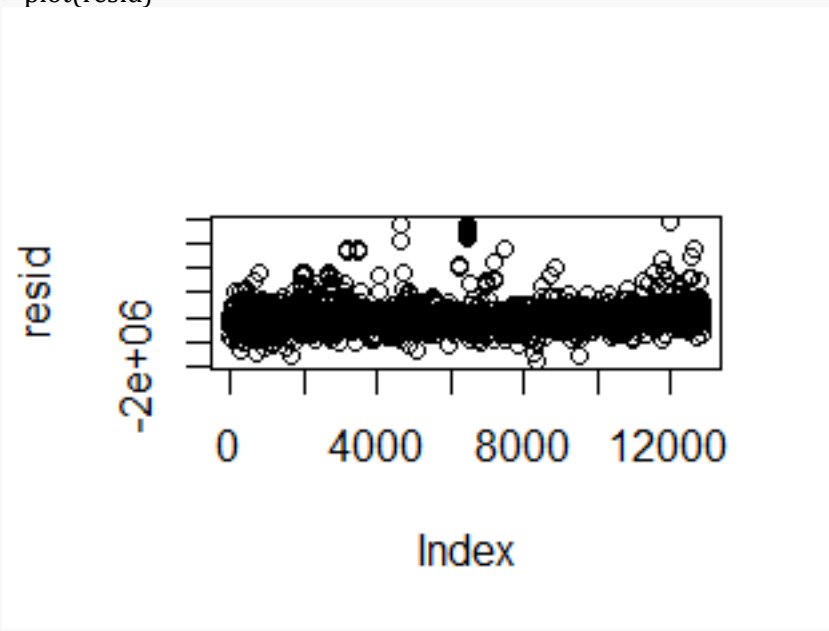
The condition is met because one probability doesn't affect the other.

13. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.

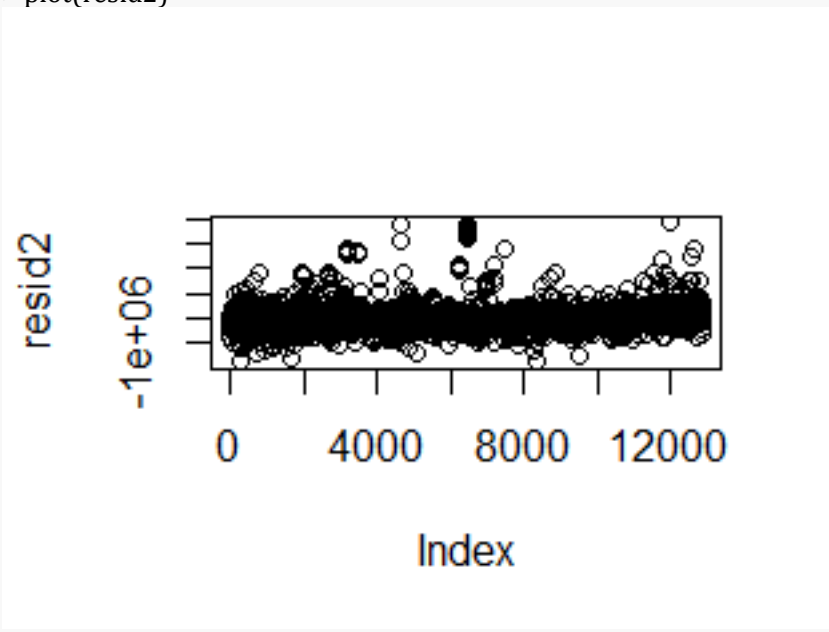
```
> vif(predict_lm2)
housing_df$square_feet_total_living          housing_df$bedrooms
                  1.853062                     1.494959
    housing_df$bath_full_count
                  1.366432
> vif(predict_lm)
```

14. 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.

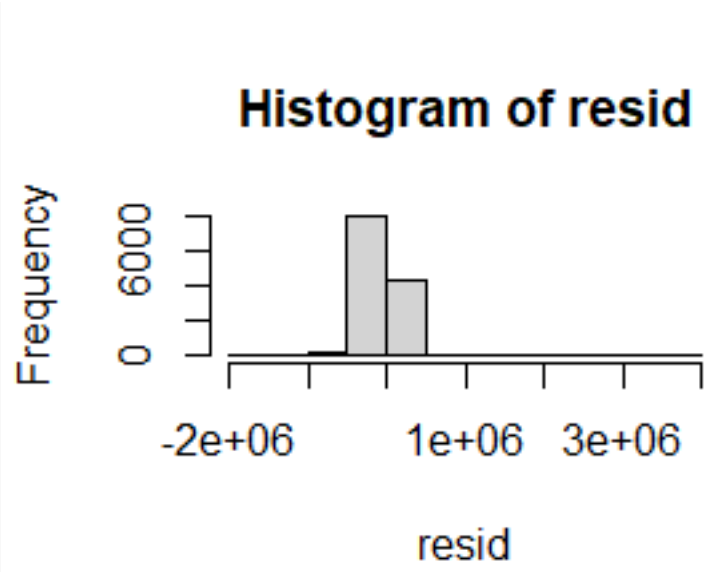
> plot(resid)



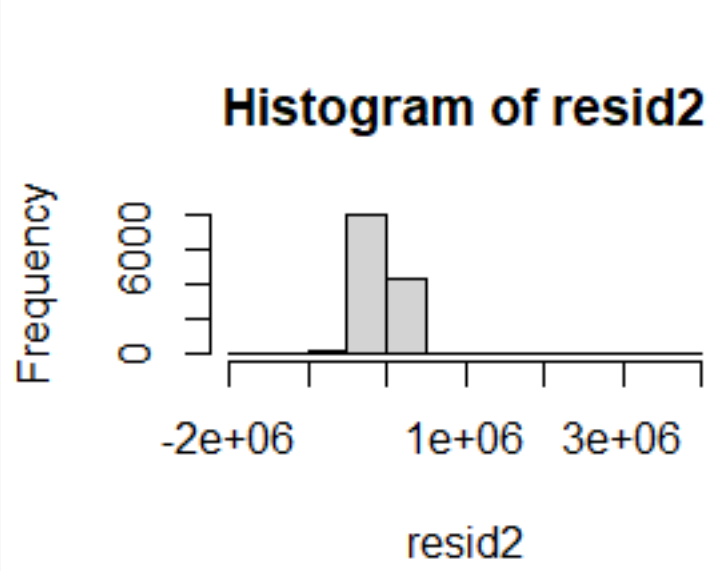
> plot(resid2)



> hist(resid)




```
> hist(resid2)
```



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

I think the model is unbiased based on the multicollinearity and the independence test. It shows that a sample approximation can be made about a population but it cannot be the true population itself.