# HousingSurvey_YoungStephen.R

young

2022-06-03

```
# Assignment: ASSIGNMENT 7
# Name: Young, Stephen
# Date: 2022-05-08

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/young/Desktop/Classes/DSC520/GIT")

## Load the `housing data
library(readxl)
housing_df <- read_excel("data/week-7-housing.xlsx")

#i. Explain any transformations or modifications you made to the dataset
#I honestly do not remember if I did any transformations. I beleive there was a blank space that has be

#ii.Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same v
#and one that will contain Sale Price and several additional predictors of your choice. Explain the bas
names(housing_df)[names(housing_df)=='Sale Price'] <- 'sale_price'
sale_vs_lot_lm <- lm(sale_price ~ sq_ft_lot, data = housing_df)
sale_vs_zip_lm <- lm(sale_price ~ zip5+square_feet_total_living, data = housing_df)

#iii. Execute a summary() function on two variables defined in the previous step to compare the model r
#Explain what these results tell you about the overall model. Did the inclusion of the additional predi
summary(sale_vs_lot_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
#The Rsquared is 0.01435 and adjusted is 0.01428. This would be considered a very weak to no correlatio
#model itself does not explain the variation in response between the variables and does not explain the

summary(sale_vs_zip_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ zip5 + square_feet_total_living, data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1797898  -120283   -41451    44176  3813341
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.932e+08  1.845e+08  -2.131   0.0331 *
## zip5                      4.012e+03  1.882e+03   2.132   0.0330 *
## square_feet_total_living  1.851e+02  3.223e+00  57.422   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360100 on 12862 degrees of freedom
## Multiple R-squared:  0.2069, Adjusted R-squared:  0.2068
## F-statistic:  1678 on 2 and 12862 DF,  p-value: < 2.2e-16
```

```
#The R-squared is 0.003618 and adjusted is 0.00354. This adjusted r square is even less than the ones f
#correlation between the variables and also does not explain the large variable size.


#iv. Considering the parameters of the multiple regression model you have created. What are the standar
library(lm.beta)
lm.beta(sale_vs_lot_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_df)
##
## Standardized Coefficients::
## (Intercept)   sq_ft_lot
##          NA   0.1198122
```

```
#The .1198122 indicates that as the sales price increases, so does the lot size of the property
lm.beta(sale_vs_zip_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ zip5 + square_feet_total_living, data = housing_df)
##
## Standardized Coefficients::
##              (Intercept)                      zip5 square_feet_total_living
##                       NA               0.01681726               0.45297887
```

```
#With a .45297887 indicator, the square feet total living has a much greater impact than the .01681726

#v. Calculate the confidence intervals for the parameters in your model and explain what the results in
confint(sale_vs_lot_lm)
```

```
##                    2.5 %       97.5 %
## (Intercept) 6.343730e+05 6.492698e+05
## sq_ft_lot   7.291208e-01 9.728641e-01
```

```
#There is a small number for teh sq_ft_lot which indicates it is not a very good indicator

confint(sale_vs_zip_lm)
```

```
##                                 2.5 %       97.5 %
## (Intercept)              -7.548758e+08 -3.149451e+07
## zip5                      3.231303e+02  7.700648e+03
## square_feet_total_living  1.787433e+02  1.913776e+02
```

```
#Both zip and square_feet_total_living have small numbers indicating that they would be very good indic

#vi. Assess the improvement of the new model compared to your original model (simple regression model)
aov(housing_df$sq_ft_lot ~ square_feet_total_living, data = housing_df)
```

```
## Call:
##    aov(formula = housing_df$sq_ft_lot ~ square_feet_total_living,
##     data = housing_df)
##
## Terms:
##                 square_feet_total_living    Residuals
## Sum of Squares              2.285221e+12 3.941214e+13
## Deg. of Freedom                        1        12863
##
## Residual standard error: 55353.35
## Estimated effects may be unbalanced
```

```
#vii. Perform casewise diagnostics to identify outliers and/or influential cases, storing each function


#viii. Calculate the standardized residuals using the appropriate command, specifying those that are +-

sales_lot_resid <- rstandard(sale_vs_lot_lm)
summary(sales_lot_resid)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -5.185311 -0.485326 -0.157656 -0.000013  0.228076  9.303661
```

```
#ix Use the appropriate function to show the sum of large residuals.
sum(sales_lot_resid^2)
```

```
## [1] 12871.11
```

```
#x Which specific variables have large residuals (only cases that evaluate as TRUE)?


#xi Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on

#sales_lot_resid <- housing_df[housing_df$large_residuals,
#                                      c("standardized_residuals", "hat_values",
#                                        "cooks_distance", "covariance_ratio")]

#leverage_threshold <- (3 / nrow(housing_df)) * 3
#sales_lot_resid$large_leverage <- sales_lot_resid$hat_values > leverage_threshold
#sum(sales_lot_resid$large_leverage)

#xii Perform the necessary calculations to assess the assumption of independence and state if the condi


#xiii Perform the necessary calculations to assess the assumption of no multicollinearity and state if


#xiv Visually check the assumptions related to the residuals using the plot() and hist() functions. Sum


#xv Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us
```