title: “WEEK 7 - EXERCISE 12” author: “Jena, Binay” date: ‘2020-10-18’ output: pdf\_document: default html\_document: default word\_document: default bibliography: bibliography.bib —

**a. Explain why you chose to remove data points from your ‘clean’ dataset.** **Answer** maybe its due to bad data or specific sale record being erroneous

summary(housing\_data)

## Sale Date Sale Price sale\_reason   
## Min. :2006-01-03 00:00:00 Min. : 2500 Min. : 0.000   
## 1st Qu.:2008-05-27 00:00:00 1st Qu.: 485075 1st Qu.: 1.000   
## Median :2012-01-24 00:00:00 Median : 605000 Median : 1.000   
## Mean :2011-08-17 23:50:44 Mean : 645051 Mean : 1.107   
## 3rd Qu.:2014-07-29 00:00:00 3rd Qu.: 749950 3rd Qu.: 1.000   
## Max. :2016-12-16 00:00:00 Max. :4311000 Max. :18.000   
## sale\_instrument sale\_warning sitetype addr\_full   
## Min. : 0.000 Length:10568 Length:10568 Length:10568   
## 1st Qu.: 3.000 Class :character Class :character Class :character   
## Median : 3.000 Mode :character Mode :character Mode :character   
## Mean : 3.147   
## 3rd Qu.: 3.000   
## Max. :26.000   
## zip5 ctyname postalctyn lon   
## Min. :98052 Length:10568 Length:10568 Min. :-122.2   
## 1st Qu.:98052 Class :character Class :character 1st Qu.:-122.1   
## Median :98052 Mode :character Mode :character Median :-122.1   
## Mean :98053 Mean :-122.1   
## 3rd Qu.:98053 3rd Qu.:-122.0   
## Max. :98074 Max. :-121.9   
## lat building\_grade square\_feet\_total\_living bedrooms   
## Min. :47.46 Min. : 2.000 Min. : 240 Min. : 0.000   
## 1st Qu.:47.67 1st Qu.: 8.000 1st Qu.: 1870 1st Qu.: 3.000   
## Median :47.69 Median : 8.000 Median : 2450 Median : 4.000   
## Mean :47.68 Mean : 8.273 Mean : 2545 Mean : 3.482   
## 3rd Qu.:47.71 3rd Qu.: 9.000 3rd Qu.: 3110 3rd Qu.: 4.000   
## Max. :47.73 Max. :13.000 Max. :13540 Max. :11.000   
## bath\_full\_count bath\_half\_count bath\_3qtr\_count year\_built   
## Min. : 0.000 Min. :0.0000 Min. :0.0000 Min. :1900   
## 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1980   
## Median : 2.000 Median :1.0000 Median :0.0000 Median :1999   
## Mean : 1.803 Mean :0.6175 Mean :0.5006 Mean :1993   
## 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:2007   
## Max. :23.000 Max. :6.0000 Max. :8.0000 Max. :2016   
## year\_renovated current\_zoning sq\_ft\_lot prop\_type   
## Min. : 0.00 Length:10568 Min. : 785 Length:10568   
## 1st Qu.: 0.00 Class :character 1st Qu.: 5400 Class :character   
## Median : 0.00 Mode :character Median : 7850 Mode :character   
## Mean : 21.93 Mean : 19921   
## 3rd Qu.: 0.00 3rd Qu.: 12037   
## Max. :2016.00 Max. :1631322   
## present\_use   
## Min. : 0.000   
## 1st Qu.: 2.000   
## Median : 2.000   
## Mean : 6.546   
## 3rd Qu.: 2.000   
## Max. :300.000

**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.**

#sales\_price\_with\_sq\_ft\_lot <- housing\_data[,c(“Sale Price”,“sq\_ft\_lot”)]

cor(housing\_data$'Sale Price', housing\_data$square\_feet\_total\_living)^2 \* 100

## [1] 50.02422

cor.test(housing\_data$'Sale Price', housing\_data$square\_feet\_total\_living)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$square\_feet\_total\_living  
## t = 102.84, df = 10566, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6976194 0.7166796  
## sample estimates:  
## cor   
## 0.707278

cor(housing\_data$'Sale Price', housing\_data$building\_grade)^2 \* 100

## [1] 42.06535

cor.test(housing\_data$'Sale Price', housing\_data$building\_grade)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$building\_grade  
## t = 87.589, df = 10566, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6373939 0.6594890  
## sample estimates:  
## cor   
## 0.6485781

cor(housing\_data$'Sale Price', housing\_data$year\_built)^2 \* 100

## [1] 6.737224

cor.test(housing\_data$'Sale Price', housing\_data$year\_built)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$year\_built  
## t = 27.628, df = 10566, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2416917 0.2772556  
## sample estimates:  
## cor   
## 0.2595616

cor(housing\_data$'Sale Price', housing\_data$bedrooms)^2 \* 100

## [1] 10.88933

cor.test(housing\_data$'Sale Price', housing\_data$bedrooms)

##   
## Pearson's product-moment correlation  
##   
## data: housing\_data$"Sale Price" and housing\_data$bedrooms  
## t = 35.933, df = 10566, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3128924 0.3468735  
## sample estimates:  
## cor   
## 0.3299898

**Answer:** I see all variables shown above are positively correlated. building grade and square feet of living room share 20% and 15% variation in determining sales price. so i chose building grade and square feet of living as predictors for the model over year\_built and bedrooms.

sales\_price\_with\_sq\_ft\_lot <- lm(housing\_data$'Sale Price' ~ housing\_data$sq\_ft\_lot, data = housing\_data)  
sales\_price\_with\_others <- lm(housing\_data$'Sale Price' ~ housing\_data$sq\_ft\_lot + housing\_data$square\_feet\_total\_living + housing\_data$building\_grade, data = housing\_data)

**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?**

summary(sales\_price\_with\_sq\_ft\_lot)

##   
## Call:  
## lm(formula = housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot,   
## data = housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2615922 -151493 -35572 106230 3293158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.205e+05 2.598e+03 238.9 <2e-16 \*\*\*  
## housing\_data$sq\_ft\_lot 1.232e+00 4.830e-02 25.5 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 248100 on 10566 degrees of freedom  
## Multiple R-squared: 0.05799, Adjusted R-squared: 0.0579   
## F-statistic: 650.5 on 1 and 10566 DF, p-value: < 2.2e-16

summary(sales\_price\_with\_others)

##   
## Call:  
## lm(formula = housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot +   
## housing\_data$square\_feet\_total\_living + housing\_data$building\_grade,   
## data = housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2259688 -80526 -14898 59973 3661458   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.671e+05 1.545e+04 -17.296 <2e-16  
## housing\_data$sq\_ft\_lot 3.427e-01 3.489e-02 9.823 <2e-16  
## housing\_data$square\_feet\_total\_living 1.321e+02 2.703e+00 48.886 <2e-16  
## housing\_data$building\_grade 6.879e+04 2.371e+03 29.013 <2e-16  
##   
## (Intercept) \*\*\*  
## housing\_data$sq\_ft\_lot \*\*\*  
## housing\_data$square\_feet\_total\_living \*\*\*  
## housing\_data$building\_grade \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 173200 on 10564 degrees of freedom  
## Multiple R-squared: 0.5407, Adjusted R-squared: 0.5406   
## F-statistic: 4145 on 3 and 10564 DF, p-value: < 2.2e-16

**Answer:**

R-Squared indicates how good the model is predicted. Higher the R2 value, means better the correlation coefficient. First model with value 0.05799, implies square foot of the lot contributes to 5.8% to the sales price In the other model , other attributes contribute 54% towards sale price.

Adjusted R2 indicates how well the model generalizes, and we don’t expect much deviation off R2. So cross validity of models is reasonable in our case.

**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?**

library("QuantPsyc")

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':  
##   
## norm

lm.beta(sales\_price\_with\_others)

## housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living   
## 0.06700386 0.48325016   
## housing\_data$building\_grade   
## 0.28180588

**Answer:**

The standardized beta estimates tell us the number of standard deviations by which the outcome will change as a result of one standard deviation change in the predictor.

In this case, 1 standard deviation of change in Sq\_ft\_lot causes sales price to change by 0.019 standard deviation. 1 standard deviation change in square\_feet\_total\_living causes sales price to change by 0.361 standard deviation and 1 standard deviation change in building\_grade can cause 0.120 standard deviation change in sale price.

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

confint(sales\_price\_with\_others, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -2.974257e+05 -2.368716e+05  
## housing\_data$sq\_ft\_lot 2.743323e-01 4.111075e-01  
## housing\_data$square\_feet\_total\_living 1.268507e+02 1.374482e+02  
## housing\_data$building\_grade 6.413951e+04 7.343430e+04

**Answer:**

The confidence interval shows that there is positive relation between all predictors and outcome. Also the 95% confidence interval range is not very big which means this sample if close to the beta of population.

**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.**

anova(sales\_price\_with\_sq\_ft\_lot, sales\_price\_with\_others)

## Analysis of Variance Table  
##   
## Model 1: housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot  
## Model 2: housing\_data$"Sale Price" ~ housing\_data$sq\_ft\_lot + housing\_data$square\_feet\_total\_living +   
## housing\_data$building\_grade  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 10566 6.5023e+14   
## 2 10564 3.1704e+14 2 3.3319e+14 5551.2 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Answer:** The value in column Pr(>F) is 2.2e−16 (i.e.~10^-22); sales\_price\_with\_others significantly improved the fit of the model to the data compared to sales\_price\_with\_sq\_ft\_lot.

**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.**

housing\_data$standardized\_residuals<- rstandard(sales\_price\_with\_others)  
head(housing\_data$standardized\_residuals)

## 1 2 3 4 5 6   
## -0.1590285 -0.4874806 -0.4594890 -0.4648446 0.4187673 -0.2760891

housing\_data$studentized\_residuals<-rstudent(sales\_price\_with\_others)  
head(housing\_data$studentized\_residuals)

## 1 2 3 4 5 6   
## -0.1590212 -0.4874630 -0.4594718 -0.4648273 0.4187509 -0.2760770

housing\_data$cooks\_distance<-cooks.distance(sales\_price\_with\_others)  
head(housing\_data$cooks\_distance)

## 1 2 3 4 5 6   
## 1.023147e-06 9.442883e-06 7.834462e-06 1.255278e-05 2.243542e-05 6.738877e-06

housing\_data$dfbeta<-dfbeta(sales\_price\_with\_others)  
head(housing\_data$dfbeta)

## (Intercept) housing\_data$sq\_ft\_lot housing\_data$square\_feet\_total\_living  
## 1 15.81733 1.835679e-05 0.001106875  
## 2 44.55331 6.337927e-05 0.001900835  
## 3 -44.30328 4.836378e-05 -0.008248112  
## 4 17.26970 -9.067978e-06 0.014261197  
## 5 -67.20274 1.732024e-04 0.002322996  
## 6 53.17910 8.944630e-06 -0.000610211  
## housing\_data$building\_grade  
## 1 -2.611925  
## 2 -7.089177  
## 3 6.866005  
## 4 -7.374515  
## 5 7.821870  
## 6 -6.809395

housing\_data$dffit<-dffits(sales\_price\_with\_others)  
head(housing\_data$dffit)

## 1 2 3 4 5 6   
## -0.002022922 -0.006145633 -0.005597813 -0.007085718 0.009472838 -0.005191643

housing\_data$leverage<-hatvalues(sales\_price\_with\_others)  
head(housing\_data$leverage)

## 1 2 3 4 5 6   
## 0.0001617998 0.0001589209 0.0001484070 0.0002323184 0.0005114775 0.0003535049

housing\_data$covariance\_ratios<-covratio(sales\_price\_with\_others)  
head(housing\_data$covariance\_ratios)

## 1 2 3 4 5 6   
## 1.000531 1.000448 1.000447 1.000529 1.000824 1.000704

**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\_data$large\_residual <- housing\_data$studentized\_residuals > 2 | housing\_data$studentized\_residuals < -2

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

sum(housing\_data$large\_residual)

## [1] 347