Exercise 8.2

Luke Syverson

May 15, 2023

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.

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.

###Complete the following:

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

Explain any transformations or modifications you made to the dataset The ctyname variable is redundant by postalctyn, and contains null values. Being of higher granularity and lesser quality, it should be removed. Sale_warning contains many nulls, and denotes an exception type that implies different sale behavior than the remaining observations. Subsetting records with sale warnings will lead to more meaningful analysis. Year_renovated imputes 0 values for non-renovated houses. It may also be subsetted, as renovated houses will implicate another exceptional behavior.

Much of the qualitative data is rendered trivial by the unknown distinction between codes and values. Methods integrating the variables may create confounding influences per the potentially codified entries. Analysis of the categorical fields will yield results lacking in final interpretation. Gross lot size vs. living lot size will also introduce variance per the disproportionate value of housing vs. land. Filtering to lot sizes that are arbitrarily some amount larger than the living size could induce selection bias, so the resulting noise should be accommodated for with additional analysis. It will also be useful to sort observations by cities to bring meaning to the lat & lon fields.

The below instructions seem to warrant a more 'blind' analysis for Sale Price that takes into account the totality of the population's behavior. Following this exercise with mirrored analysis for a few subsets will yield additional insight into the housing market behavior across demographics, but this does not appear to be in the scope of the instruction.

Prop_type is entirely 'R', which is meaningless other than presumably indicating residential, and will be removed. Lastly, spaces need to be removed from column names to perform modeling functions.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- housing %>% select(-'sale_warning', -'year_renovated', -'ctyname', -'prop_type')
colnames(df) <- gsub(" ", "_", colnames(df))
```

```
price_lot <- df %>% select(Sale_Price, sq_ft_lot)
predict_sale <- df %>% select(Sale_Price, year_built, bedrooms, sq_ft_lot, sitetype, building_grade)

pl_model <- lm(Sale_Price ~ sq_ft_lot, data = price_lot)
sale_model <- lm(Sale_Price ~ year_built + bedrooms + sq_ft_lot + building_grade + sitetype, data = pre
```

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. Living space, property usage/zoning & all geographic qualifiers except for zip5 are excluded per assumed multicollinearity with total lot size, sitetype and geography. Sale date is excluded per the granularity, despite potential to quantify potential seasonality. Future analysis might aggregate the dates to the month level to increase generalization potential. Bathrooms are excluded per the complexity of relationships between each possible combination. A weighed total of each type might be suited for future analysis. — #### 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(pl_model)
```

```
##
## Call:
## lm(formula = Sale_Price ~ sq_ft_lot, data = price_lot)
##
## 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
```

```
summary(sale_model)
```

```
##
## Call:
## lm(formula = Sale_Price ~ year_built + bedrooms + sq_ft_lot +
##     building_grade + sitetype, data = predict_sale)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2208337  -133010   -48407    49473  3713419
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -7.333e+06  4.218e+05 -17.382  < 2e-16 ***
## year_built      3.650e+03  2.063e+02  17.691  < 2e-16 ***
## bedrooms        5.410e+04  3.965e+03  13.644  < 2e-16 ***
## sq_ft_lot       6.699e-01  5.858e-02  11.436  < 2e-16 ***
## building_grade  1.030e+05  3.464e+03  29.724  < 2e-16 ***
## sitetypeC1     -6.648e+04  2.468e+05  -0.269 0.787635
## sitetypeDV     -4.782e+05  1.970e+05  -2.427 0.015224 *
## sitetypeR1     -3.292e+05  1.292e+05  -2.549 0.010829 *
## sitetypeR2     -4.429e+05  1.304e+05  -3.396 0.000686 ***
## sitetypeR3     -5.465e+05  1.617e+05  -3.378 0.000731 ***
## sitetypeR4     -5.705e+05  3.866e+05  -1.476 0.139994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 364300 on 12854 degrees of freedom
## Multiple R-squared:  0.1891, Adjusted R-squared:  0.1885
## F-statistic: 299.8 on 10 and 12854 DF,  p-value: < 2.2e-16
```

pl_model: Multiple R-squared: 0.01435, Adjusted R-squared: 0.01428 The low R-statistics indicate little predictability of Sale Price given lot size. ##### sale_model Multiple R-squared: 0.1891, Adjusted R-squared: 0.1885 The additional predictors fared slightly better, but lacked substantial improvement.

```
library("lm.beta")
lm.beta(sale_model)
```

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 ~ year_built + bedrooms + sq_ft_lot +
##     building_grade + sitetype, data = predict_sale)
##
## Standardized Coefficients::
##    (Intercept)     year_built       bedrooms      sq_ft_lot building_grade
##             NA    0.155430490    0.117218425    0.094314059    0.278167861
##     sitetypeC1     sitetypeDV     sitetypeR1     sitetypeR2     sitetypeR3
##   -0.002510141   -0.025535719   -0.142824813   -0.184732747   -0.044555330
##     sitetypeR4
##   -0.012438933
```

It appears that building grade has the largest influences on the standard deviation of Sale Price, with building site types, year built, bedrooms and lot size following in descending order of magnitude. Building grade is the best predictor by far.

```
confint(sale_model)
```

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

```
##                        2.5 %        97.5 %
## (Intercept)    -8.159417e+06 -6.505670e+06
## year_built      3.245564e+03  4.054383e+03
## bedrooms        4.633032e+04  6.187522e+04
## sq_ft_lot       5.550627e-01  7.847095e-01
## building_grade  9.616106e+04  1.097393e+05
## sitetypeC1     -5.501770e+05  4.172244e+05
## sitetypeDV     -8.644465e+05 -9.204831e+04
## sitetypeR1     -5.823665e+05 -7.599807e+04
## sitetypeR2     -6.985466e+05 -1.872633e+05
## sitetypeR3     -8.634974e+05 -2.294039e+05
## sitetypeR4     -1.328252e+06  1.871915e+05
```

Most variables could be significant, and the range of the intervals other than site types and bedrooms ar fairly small. Site type seems to be confounding the significance of the model.

```
anova(pl_model, sale_model)
```

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.

```
## Analysis of Variance Table
##
## Model 1: Sale_Price ~ sq_ft_lot
## Model 2: Sale_Price ~ year_built + bedrooms + sq_ft_lot + building_grade +
##     sitetype
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12854 1.7058e+15  9 3.6759e+14 307.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value indicates a significant improvement from the original model.

```
library(broom)
cwise <- augment(sale_model) %>% select(everything(), .fitted, .std.resid, .cooksd, .hat)
```

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

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

```
cwise$large.residual <- abs(cwise$.std.resid) > 2
sum(cwise$large.residual, na.rm = TRUE)
```

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

```
## [1] 322
```

```
cwise_lr <- subset(cwise, cwise$large.residual == TRUE)
names(cwise_lr)
```

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

```
##  [1] "Sale_Price"     "year_built"     "bedrooms"       "sq_ft_lot"
##  [5] "building_grade" "sitetype"       ".fitted"        ".resid"
##  [9] ".hat"           ".sigma"         ".cooksd"        ".std.resid"
## [13] "large.residual"
```

All the variables contain large residuals, but none contain exclusively TRUE values. It's possible that I'm misunderstanding the prompt here.

```
cwise$large.cd <- cwise$.cooksd > 1
sum(cwise$large.cd, na.rm = TRUE)
```

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

```
## [1] 0
```

```
cwise$large.lev <- cwise$.hat > 2*(mean(cwise$.hat))
sum(cwise$large.lev)
```

```
## [1] 577
```

No records have a Cook's Distance > 1, which implies stability; however, 577 records have high leverage, indicating possible outliers with exaggerated influence on the model.

I wasn't able to get a covariance matrix per a non-logical/numeric error that I failed to resolve.

```
library(car)
```

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

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
dwt(sale_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7009978     0.5979958       0
##  Alternative hypothesis: rho != 0
```

It appears there is certainly autocorrelation within this model (p-value of 0), which I presume is within bedrooms, lot size and building grade.

```
cor.test(x = predict_sale$Sale_Price, y = predict_sale$sq_ft_lot, method = 'spearman')
```

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

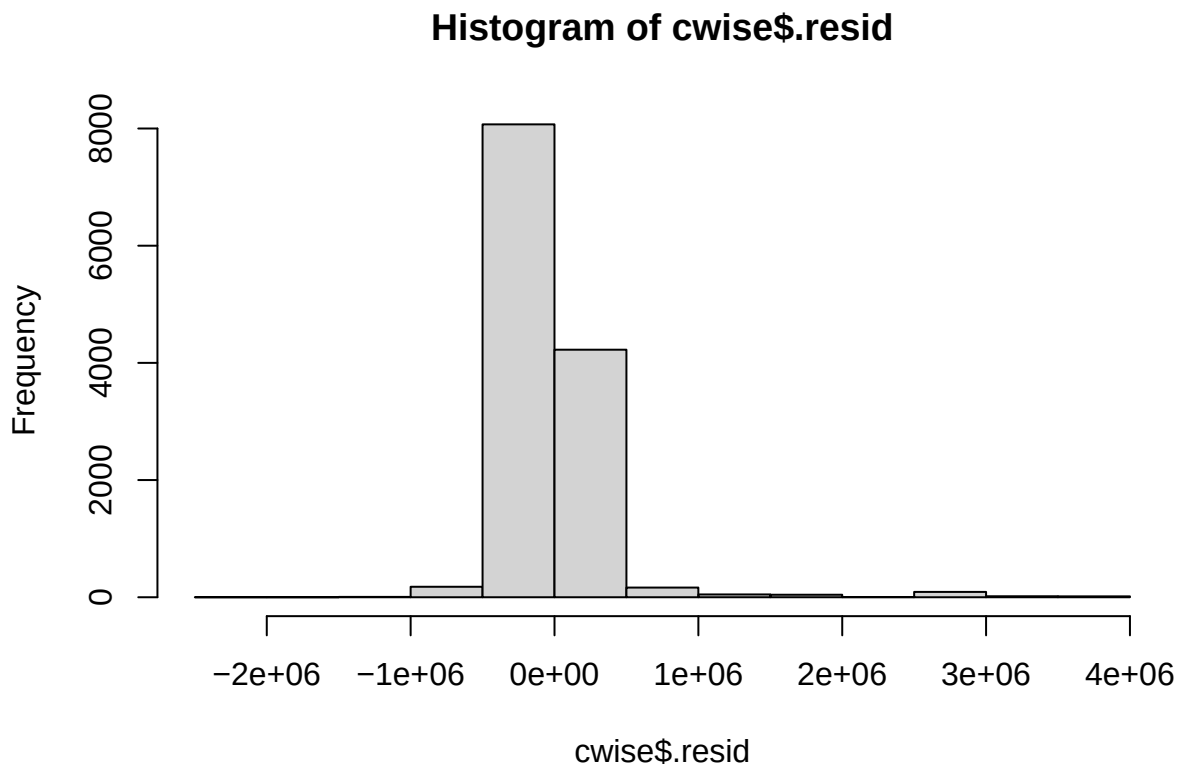
```
## Warning in cor.test.default(x = predict_sale$Sale_Price, y =
## predict_sale$sq_ft_lot, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  predict_sale$Sale_Price and predict_sale$sq_ft_lot
## S = 2.9773e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1610402
```

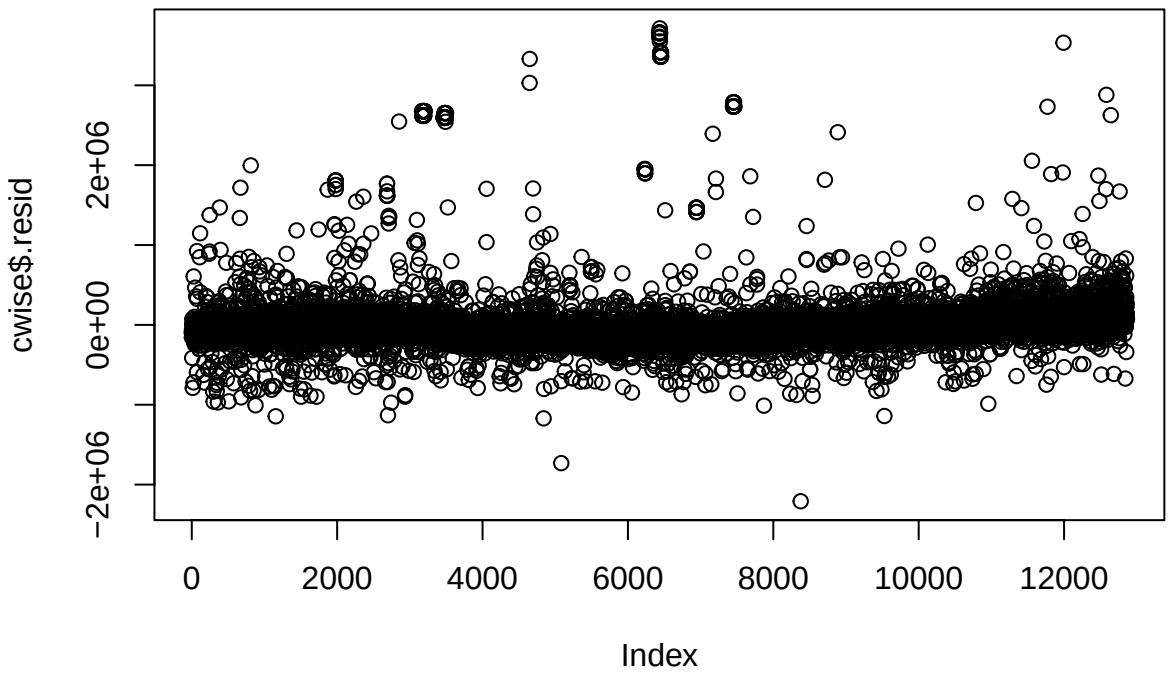
I'm not entirely certain of how to assess the entire dataframe with cor.test, but regardless, it appears Sale Price and lot size are substantially correlated.

```
hist(cwise$.resid)
```

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(cwise$.resid)
```



The residual histogram implies right-skewedness and very large outliers.The plot also implicates right-skewedness and outliers. Heteroskedasticity is present as well. The relationship is otherwise linear.

Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model? The model is a poor generalization of the population due to the presence of outliers, skewedness & heteroskedasticity. This indicates that many extreme exceptions to the normal housing market behavior are present within the sample dataset, indicating large variance within the population or possible selection bias.

If the model were unbiased, it would implicate a slight right-skewedness across a normal, leptokurtic distribution.