August 1, 2023

The results below are generated from an R script.

```
Johnson_Franchesca

Cleaning df =   Looked for unique values in all columns (lat/Lon), removed any duplicates.
Omitted all NAs, removed several columns that wouldn't help or hurt the model.  Sorted the columns to ve

7.  Housing_T_sqft <- lm(housingdata2$`Sale Price`~ housingdata2$square_feet_total_living)
    housinhousingdata2$studentized.residuals <- rstandard(housingdata2)gmodel <- lm(housingdata2$`Sale P
    housingdata2$square_feet_total_living + housingdata2$bath_full_count + housingdata2$bedrooms + housi

8.  Housing_T_sqft <- lm(housingdata2$`Sale Price`~ housingdata2$square_feet_total_living)
    Housing_T_sqft

    Call:
      lm(formula = housingdata2$`Sale Price` ~ housingdata2$square_feet_total_living)

    Coefficients:
      (Intercept)  housingdata2$square_feet_total_living
        190236.6                                 185.3

summary(Housing_T_sqft)

        Call:
          lm(formula = housingdata2$`Sale Price` ~ housingdata2$square_feet_total_living)

        Residuals:
          Min       1Q   Median       3Q      Max
        -1797527  -120336   -41637    43858  3811329

        Coefficients:
          Estimate Std. Error t value Pr(>|t|)
        (Intercept)                           190236.608   8780.272   21.67   <2e-16 ***
          housingdata2$square_feet_total_living    185.290      3.224   57.48   <2e-16 ***
          ---
          Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Residual standard error: 360800 on 12804 degrees of freedom
        Multiple R-squared:  0.2051,Adjusted R-squared:  0.205
        F-statistic:  3304 on 1 and 12804 DF,  p-value: < 2.2e-16


summary(housingmodel)

Call:
```

```
  lm(formula = housingdata2$`Sale Price` ~ housingdata2$square_feet_total_living +
       housingdata2$bath_full_count + housingdata2$bedrooms + housingdata2$year_built,
     data = housingdata2)

Residuals:
  Min       1Q   Median       3Q      Max
-1716509  -120674   -42542    45647  3905691

Coefficients:
  Estimate   Std. Error t value Pr(>|t|)
(Intercept)                            -4470679.262   420767.971 -10.625  < 2e-16 ***
  housingdata2$square_feet_total_living      173.859        4.443  39.129  < 2e-16 ***
  housingdata2$bath_full_count             16753.605     6113.930   2.740  0.00615 **
  housingdata2$bedrooms                   -13436.194     4535.156  -2.963  0.00306 **
  housingdata2$year_built                   2361.521      212.370  11.120  < 2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 357900 on 12801 degrees of freedom
Multiple R-squared:  0.2179,Adjusted R-squared:  0.2177
F-statistic: 891.7 on 4 and 12801 DF,  p-value: < 2.2e-16

8a.  Multiple R-squared:  0.2051,Adjusted R-squared:  0.205
8b.  The F-stat is a lrg number and a p-value  less than .001.
     The results say the sq ft of a lot can predict/affect the sale price of the home.
8c.  Yes, square feet of total lot, bathroom, bedrooms and year built all affected the Sale price of hom

9.   Below is the model summary of multiple linear Reg. model.  The standardized betas for each paramete
    All have a positive relationship except for the bedroom count. So, as the number of bathrooms, total
    year built increase the Sale price could also increase, as long as the other predictors are held con
    I also found that an increase in bedrooms is associated with a decrease in sale prices.  So if the t
    of the house increase by .425 then the sale of the house will increase by 0.425 sd. Only if the the

    housingdata2$square_feet_total_living housingdata2$bath_full_count          housingdata2$bedrooms
    0.42494143                            0.02693749                           -0.02911215

    housingdata2$year_built
    0.10057206

10. Square ft total, year built and bathroom count don't cross zero thus saying that 95% of the populati
    would have a true beta value.
    Even though bathrooms has a large C.I. and the other two (sq. total and year built) have a small
    C.I., it's still significant.  However,
    bedrooms do cross zero, this is telling me that some samples in the population will have a positive
    I can't say that 95% of the population will have a true b value.

    confint(housingmodel)
                                              2.5 %        97.5 %
    (Intercept)                         -5295447.3152 -3645911.2079
    housingdata2$square_feet_total_living      165.1495      182.5684
    housingdata2$bath_full_count              4769.3887    28737.8208
    housingdata2$bedrooms                    -2225.7774    -4546.6107
    housingdata2$year_built                   1945.2444     2777.7976
```

```
11. I believe it's signifigant by 69.9%.  hou

    anova(Housing_T_sqft, housingmodel)
    Analysis of Variance Table

    Model 1: housingdata2$`Sale Price` ~ housingdata2$square_feet_total_living
    Model 2: housingdata2$`Sale Price` ~ housingdata2$square_feet_total_living +
    housingdata2$bath_full_count + housingdata2$bedrooms + housingdata2$year_built
    Res.Df        RSS Df      Sum of Sq      F    Pr(>F)
  1  12804 1.6666e+15
  2  12801 1.6398e+15  3 26849432422523 69.866 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

12. > housingdata2$studentized.residuals <- rstandard(housingmodel)
  > housingdata2$studentized.residuals <- rstudent(housingmodel)
  > housingdata2$standardized.residuals <- rstandard(housingmodel)
  > housingdata2$residuals <- resid(housingmodel)
  > housingdata2$cooks.distance <- cooks.distance(housingmodel)
  > housingdata2$dfbeta <- dfbeta(housingmodel)
  > housingdata2$dffit <- dffits(housingmodel)
  > housingdata2$leverage <- hatvalues(housingmodel)
  > housingdata2$covariance.ratios <- covratio(housingmodel)
  > housingdata2
  # A tibble: 12,806 × 14
  `Sale Price` square_feet_total_living bath_full_count bedrooms year_built residuals standarized.residu
  <dbl>                  <dbl>           <dbl>    <dbl>      <dbl>     <dbl>                  <dbl>
    1       897990                   3830              2        5       2013   -17278.
  2       569990                   2370              2        3       1988   -59279.               -0
  3       731000                   2370              2        3       1988   101731.                0
  4       519000                   2690              3        5       1985  -148700.               -0
  5       515000                   2670              3        5       1981  -139777.               -0
  6       785000                   1850              2        4       2010   207621.                0
  7       357886                   1850              2        4       2010  -219493.               -0
  8       510000                   1880              2        4       1987   -18280.               -0
  9       550000                   2530              2        4       1986   -88927.               -0
  10       550000                   3150              2        4       2003  -236865.                -


13.  housingdata2$large.residual <- housingdata2$standarized.residuals>2|housingdata2$standarized.residu
14.  sum(housingdata2$large.residual)
15.  [1] 327

16.   The cooks.distance are greater than 1, none have an influence. Leverage is low having no influence
      for CVR and I don't believe any value is outside of the CVR.

    cooks.distance leverage covariance.ratios
    <dbl>    <dbl>             <dbl>
      1        0.00285 0.00153             0.998
      2        0.00332 0.000279            0.978
      3        0.00303 0.000258            0.978
```

```
      4        0.00344 0.000288             0.977
      5        0.00332 0.000279             0.978
      6        0.00246 0.000208             0.978
      7        0.00291 0.000247             0.978
      8        0.00349 0.000291             0.977
      9        0.00289 0.000246             0.978
     10       0.00246 0.000208            0.978

17.  The condition was met, the dwt was close to 2 and greater than 1 but less than 3.

      durbinWatsonTest(housingmodel)
     lag Autocorrelation D-W Statistic p-value
       1      0.01038419      1.979231   0.246
     Alternative hypothesis: rho != 0

18. The condition has been met but the average is slightly greater than one, their could be a small amou

vif(housingmodel)
housingdata2$square_feet_total_living          housingdata2$bath_full_count                 housingdata2
                             1.930416                              1.581702
 1/vif(housingmodel)
housingdata2$square_feet_total_living          housingdata2$bath_full_count                 housingdata2
                            0.5180232                             0.6322304
 mean(vif(housingmodel))
[1] 1.607849

19.  Each plot isn't linear.  Even the histogram is isn't a nice bell shape, it's slightly skewed.


plot(housingmodel)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
hist(housingdata2$studentized.residuals)

20.  I think overall our model is slightly unbiased and does represent the general population.
     I do feel there was some discretion with the data but would need more information concerning the dat
     Such as several homes having 0 bedrooms (studio apt?) and homes having 0 bathrooms didn't see accura
     However, without that being clear, I don't think it caused too much of an issue with the model.


## Error:  <text>:3:10:  unexpected symbol
## 2:
## 3:  Cleaning df
##             ^
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.3.0 (2023-04-21 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22621)
##
```

```
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] splines   stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] QuantPsyc_1.6        MASS_7.3-58.4        purrr_1.0.1          boot_1.3-28.1
##  [5] GGally_2.1.2         ggplot2_3.4.2        Rcmdr_2.8-0          effects_4.2-2
##  [9] RcmdrMisc_2.7-2      sandwich_3.0-2       car_3.1-2            carData_3.0-5
## [13] knitr_1.43           rmarkdown_2.23       scales_1.2.1         reshape2_1.4.4
## [17] tidyr_1.3.0          DataExplorer_0.8.2   janitor_2.2.0        tibble_3.2.1
## [21] dplyr_1.1.2          data.validator_0.2.0 data.table_1.14.8
##
## loaded via a namespace (and not attached):
##  [1] DBI_1.1.3          gridExtra_2.3      tcltk_4.3.0        readxl_1.4.2
##  [5] rlang_1.1.1        magrittr_2.0.3     snakecase_0.11.0   e1071_1.7-13
##  [9] compiler_4.3.0     vctrs_0.6.3        stringr_1.5.0      crayon_1.5.2
## [13] pkgconfig_2.0.3    fastmap_1.1.1      backports_1.4.1    ellipsis_0.3.2
## [17] labeling_0.4.2     utf8_1.2.3         haven_2.5.2        nloptr_2.0.3
## [21] tinytex_0.45       xfun_0.39          cachem_1.0.8       jsonlite_1.8.5
## [25] progress_1.2.2     highr_0.10         reshape_0.8.9      prettyunits_1.1.1
## [29] parallel_4.3.0     cluster_2.1.4      R6_2.5.1           RColorBrewer_1.1-3
## [33] bslib_0.5.0        stringi_1.7.12     pkgload_1.3.2      rpart_4.1.19
## [37] lubridate_1.9.2    jquerylib_0.1.4    cellranger_1.1.0   Rcpp_1.0.10
## [41] zoo_1.8-12         base64enc_0.1-3    Matrix_1.5-4       nnet_7.3-18
## [45] igraph_1.5.0       timechange_0.2.0   tidyselect_1.2.0   rstudioapi_0.14
## [49] abind_1.4-5        yaml_2.3.7         lattice_0.21-8     plyr_1.8.8
## [53] withr_2.5.0        evaluate_0.21      foreign_0.8-84     survival_3.5-5
## [57] proxy_0.4-27       survey_4.2-1       pillar_1.9.0       checkmate_2.2.0
## [61] nortest_1.0-4      insight_0.19.3     generics_0.1.3     hms_1.1.3
## [65] munsell_0.5.0      minqa_1.2.5        class_7.3-21       glue_1.6.2
## [69] Hmisc_5.1-0        tools_4.3.0        lme4_1.1-33        forcats_1.0.0
## [73] grid_4.3.0         mitools_2.4        colorspace_2.1-0   nlme_3.1-162
## [77] networkD3_0.4      htmlTable_2.4.1    Formula_1.2-5      cli_3.6.1
## [81] fansi_1.0.4        tcltk2_1.2-11      gtable_0.3.3       relimp_1.0-5
## [85] sass_0.4.6         digest_0.6.31      htmlwidgets_1.6.2  farver_2.1.1
## [89] htmltools_0.5.5    lifecycle_1.0.3

Sys.time()

## [1] "2023-08-01 23:00:14 EDT"
```