Multicollinearity: Question

STA 506 2.0 Linear Regression Analysis

Thiyanga S. Talagala

Data

```
library(tidyverse)
realestate <- read.csv("real-estate.csv")
head(realestate)
```

```
  ID  Price Sqft Bedroom Bathroom Airconditioning Garage Pool YearBuild Quality
1  1 360000 3032       4        4               1      2    0      1972       2
2  2 340000 2058       4        2               1      2    0      1976       2
3  3 250000 1780       4        3               1      2    0      1980       2
4  4 205500 1638       4        2               1      2    0      1963       2
5  5 275500 2196       4        3               1      2    0      1968       2
6  6 248000 1966       4        3               1      5    1      1972       2
    Lot AdjHighway
1 22221          0
2 22912          0
3 21345          0
4 17342          0
5 21786          0
6 18902          0
```

```
glimpse(realestate)
```

```
Rows: 522
Columns: 12
$ ID              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
$ Price           <int> 360000, 340000, 250000, 205500, 275500, 248000, 229900~
$ Sqft            <int> 3032, 2058, 1780, 1638, 2196, 1966, 2216, 1597, 1622, ~
$ Bedroom         <int> 4, 4, 4, 4, 4, 4, 3, 2, 3, 3, 7, 3, 5, 5, 3, 5, 2, 3, ~
$ Bathroom        <int> 4, 2, 3, 2, 3, 3, 2, 1, 2, 3, 5, 4, 4, 4, 3, 5, 2, 4, ~
$ Airconditioning <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, ~
$ Garage          <int> 2, 2, 2, 2, 2, 5, 2, 1, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, ~
$ Pool            <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
$ YearBuild       <int> 1972, 1976, 1980, 1963, 1968, 1972, 1972, 1955, 1975, ~
$ Quality         <int> 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 1, 1, 1, 2, 2, 2, 2, ~
$ Lot             <int> 22221, 22912, 21345, 17342, 21786, 18902, 18639, 22112~
$ AdjHighway      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Q1: Identify qualitative and quantitative variables.

```
      ID              Price             Sqft          Bedroom
 Min.   :  1.0   Min.   : 84000   Min.   : 980   Min.   :0.000
 1st Qu.:131.2   1st Qu.:180000   1st Qu.:1701   1st Qu.:3.000
 Median :261.5   Median :229900   Median :2061   Median :3.000
 Mean   :261.5   Mean   :277894   Mean   :2261   Mean   :3.471
 3rd Qu.:391.8   3rd Qu.:335000   3rd Qu.:2636   3rd Qu.:4.000
 Max.   :522.0   Max.   :920000   Max.   :5032   Max.   :7.000
    Bathroom     Airconditioning     Garage    Pool      YearBuild    Quality
 Min.   :0.000   0: 88           Min.   :0.0   0:486   Min.   :1885   1: 68
 1st Qu.:2.000   1:434           1st Qu.:2.0   1: 36   1st Qu.:1956   2:290
 Median :3.000                   Median :2.0           Median :1966   3:164
 Mean   :2.642                   Mean   :2.1           Mean   :1967
 3rd Qu.:3.000                   3rd Qu.:2.0           3rd Qu.:1981
 Max.   :7.000                   Max.   :7.0           Max.   :1998
      Lot         AdjHighway
 Min.   : 4560   0:511
 1st Qu.:17205   1: 11
 Median :22200
 Mean   :24370
 3rd Qu.:26787
 Max.   :86830
```

Q2: What is wrong with the following graph?

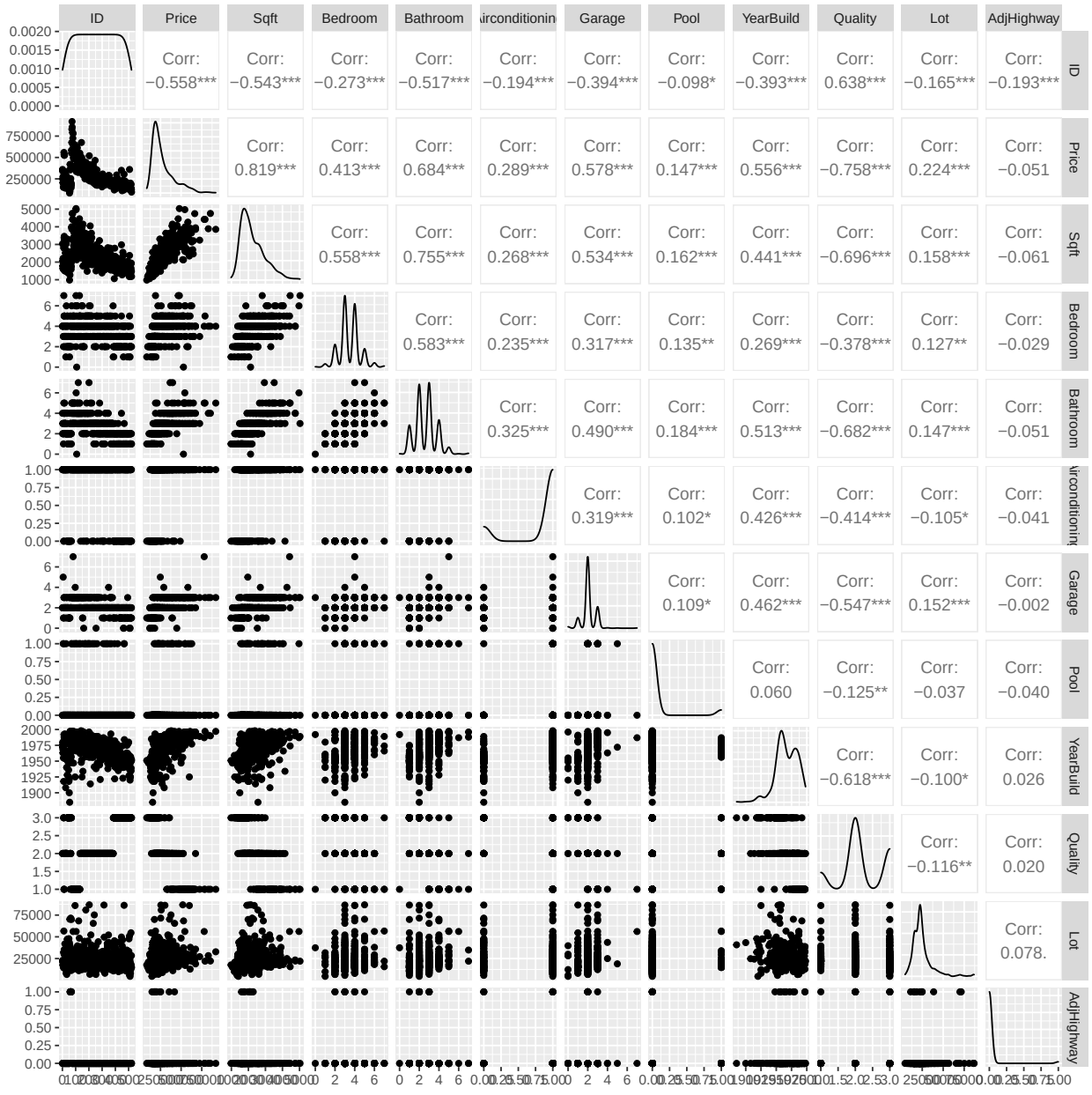

Figure 1: Pairwise correlation plot

Q3: Figure 1 is modified as follows. Now what can you say about the graph?

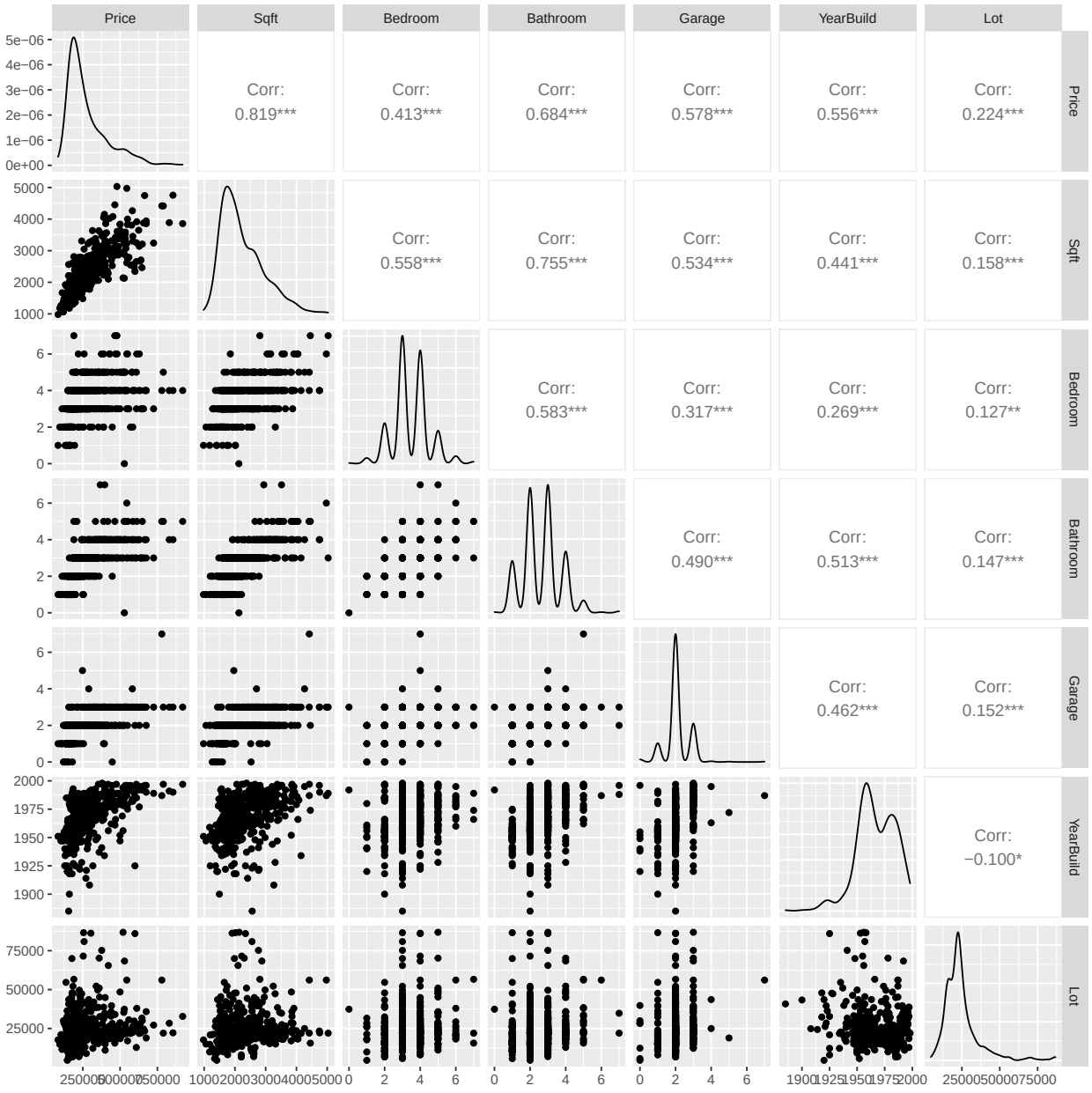

Figure 2: Pairwise correlation plot

Q4: What can you say about the source of multicollinearity in these data?

```
realestate$Airconditioning <- factor(realestate$Airconditioning)
realestate$Pool <- factor(realestate$Pool)
realestate$AdjHighway <- factor(realestate$AdjHighway)
realestate$Quality <- factor(realestate$Quality)
realestate <- realestate[, -1]
summary(realestate)
```

```
     Price             Sqft         Bedroom         Bathroom
 Min.   : 84000   Min.   : 980   Min.   :0.000   Min.   :0.000
 1st Qu.:180000   1st Qu.:1701   1st Qu.:3.000   1st Qu.:2.000
 Median :229900   Median :2061   Median :3.000   Median :3.000
 Mean   :277894   Mean   :2261   Mean   :3.471   Mean   :2.642
 3rd Qu.:335000   3rd Qu.:2636   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :920000   Max.   :5032   Max.   :7.000   Max.   :7.000
 Airconditioning     Garage    Pool      YearBuild    Quality      Lot
 0: 88           Min.   :0.0   0:486   Min.   :1885   1: 68   Min.   : 4560
 1:434           1st Qu.:2.0   1: 36   1st Qu.:1956   2:290   1st Qu.:17205
                 Median :2.0           Median :1966   3:164   Median :22200
                 Mean   :2.1           Mean   :1967           Mean   :24370
                 3rd Qu.:2.0           3rd Qu.:1981           3rd Qu.:26787
                 Max.   :7.0           Max.   :1998           Max.   :86830
 AdjHighway
 0:511
 1: 11
```

```
model <- lm(Price ~ . , data=realestate)
summary(model)
```

```
Call:
lm(formula = Price ~ ., data = realestate)

Residuals:
    Min      1Q  Median      3Q     Max
-204865  -28010   -4973   21315  298892

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.358e+06  3.991e+05  -5.909 6.29e-09 ***
Sqft             8.700e+01  6.570e+00  13.242  < 2e-16 ***
Bedroom         -5.125e+03  3.275e+03  -1.565   0.1182
Bathroom         8.127e+03  4.288e+03   1.895   0.0586 .
Airconditioning1 4.851e+03  8.086e+03   0.600   0.5488
Garage           1.089e+04  5.060e+03   2.152   0.0319 *
Pool1            1.014e+04  1.040e+04   0.975   0.3303
YearBuild        1.269e+03  2.024e+02   6.272 7.60e-10 ***
Quality2        -1.430e+05  1.021e+04 -14.007  < 2e-16 ***
```

```
Quality3         -1.484e+05  1.404e+04 -10.564  < 2e-16 ***
Lot               1.556e+00  2.363e-01   6.587 1.12e-10 ***
AdjHighway1      -2.737e+04  1.810e+04  -1.512   0.1311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58770 on 510 degrees of freedom
Multiple R-squared:  0.8223,    Adjusted R-squared:  0.8184
F-statistic: 214.5 on 11 and 510 DF,  p-value: < 2.2e-16
```

```
library(car)
car::vif(model)
```

```
                   GVIF Df GVIF^(1/(2*Df))
Sqft           3.292569  1        1.814544
Bedroom        1.664845  1        1.290289
Bathroom       3.141563  1        1.772445
Airconditioning 1.385038  1        1.176876
Garage         1.651938  1        1.285277
Pool           1.050442  1        1.024911
YearBuild      1.922344  1        1.386486
Quality        3.322305  2        1.350081
Lot            1.150133  1        1.072443
AdjHighway     1.021444  1        1.010665
```

Note:

1.1: note

Multicollinearity occurs if we do not treat this appropriately.

1.2 Model with only quantitative variables: VIF

```
summary(Duncan)
```

```
   type        income          education         prestige
 bc  :21   Min.   : 7.00   Min.   :  7.00   Min.   : 3.00
 prof:18   1st Qu.:21.00   1st Qu.: 26.00   1st Qu.:16.00
 wc  : 6   Median :42.00   Median : 45.00   Median :41.00
           Mean   :41.87   Mean   : 52.56   Mean   :47.69
           3rd Qu.:64.00   3rd Qu.: 84.00   3rd Qu.:81.00
           Max.   :81.00   Max.   :100.00   Max.   :97.00
```

```
m1 <- lm(prestige ~ income + education, data=Duncan)
vif(m1)
```

```
   income education
   2.1049    2.1049
```

1.3 Model with only quantitative variables and qualitative variables: Generalized variance-inflation factors

```
m2 <- lm(prestige ~ income + education + type, data=Duncan)
vif(m2)
```

```
              GVIF Df GVIF^(1/(2*Df))
income    2.209178  1        1.486330
education 5.297584  1        2.301648
type      5.098592  2        1.502666
```

Q5: Write down the estimated model.

```
summary(model)
```

```
Call:
lm(formula = Price ~ ., data = realestate)

Residuals:
    Min      1Q  Median      3Q     Max
-204865  -28010   -4973   21315  298892

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.358e+06  3.991e+05  -5.909 6.29e-09 ***
Sqft              8.700e+01  6.570e+00  13.242  < 2e-16 ***
Bedroom          -5.125e+03  3.275e+03  -1.565   0.1182
Bathroom          8.127e+03  4.288e+03   1.895   0.0586 .
Airconditioning1  4.851e+03  8.086e+03   0.600   0.5488
Garage            1.089e+04  5.060e+03   2.152   0.0319 *
Pool1             1.014e+04  1.040e+04   0.975   0.3303
YearBuild         1.269e+03  2.024e+02   6.272 7.60e-10 ***
Quality2         -1.430e+05  1.021e+04 -14.007  < 2e-16 ***
Quality3         -1.484e+05  1.404e+04 -10.564  < 2e-16 ***
Lot               1.556e+00  2.363e-01   6.587 1.12e-10 ***
AdjHighway1      -2.737e+04  1.810e+04  -1.512   0.1311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58770 on 510 degrees of freedom
Multiple R-squared:  0.8223,    Adjusted R-squared:  0.8184
F-statistic: 214.5 on 11 and 510 DF,  p-value: < 2.2e-16
```

cont.