Data Science II (P8106) Homework #1

Zachary Katz (UNI: zak2132)

2/22/2022

Contents

Set-Up and Pre-Processing

In this exercise, we'll predict the sale price of a home based on a set of feature variables. We begin by loading the appropriate libraries and setting our defaults. Then, we'll import our data and eliminate any rows with NAs before providing summary statistics.

```
# Load data
training_data = read_csv("./Data/housing_training.csv") %>%
  janitor::clean_names()
```

```
## Rows: 1440 Columns: 26
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Overall_Qual, Kitchen_Qual, Fireplace_Qu, Exter_Qual
## dbl (22): Gr_Liv_Area, First_Flr_SF, Second_Flr_SF, Total_Bsmt_SF, Low_Qual_...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
test_data = read_csv("./Data/housing_test.csv") %>%
  janitor::clean_names()
```

```
## Rows: 959 Columns: 26
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Overall_Qual, Kitchen_Qual, Fireplace_Qu, Exter_Qual
## dbl (22): Gr_Liv_Area, First_Flr_SF, Second_Flr_SF, Total_Bsmt_SF, Low_Qual_...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Eliminate rows containing NA entries
training_data = na.omit(training_data)
test_data = na.omit(test_data)
```

```
# Summary statistics of training data
summary(training_data)
```

```
##   gr_liv_area    first_flr_sf    second_flr_sf    total_bsmt_sf
##  Min.   : 492   Min.   : 372.0   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:1093   1st Qu.: 870.5   1st Qu.:   0.0   1st Qu.: 792.5
##  Median :1432   Median :1055.0   Median :   0.0   Median : 975.0
##  Mean   :1478   Mean   :1134.2   Mean   : 338.6   Mean   :1035.1
##  3rd Qu.:1721   3rd Qu.:1346.2   3rd Qu.: 704.8   3rd Qu.:1262.2
##  Max.   :4316   Max.   :3228.0   Max.   :1872.0   Max.   :3206.0
##  low_qual_fin_sf    wood_deck_sf     open_porch_sf     bsmt_unf_sf
##  Min.   :  0.000   Min.   :   0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.: 193.8
##  Median :  0.000   Median :   0.00   Median : 26.00   Median : 431.5
##  Mean   :  4.681   Mean   :  98.43   Mean   : 45.14   Mean   : 520.6
##  3rd Qu.:  0.000   3rd Qu.: 169.00   3rd Qu.: 68.00   3rd Qu.: 768.2
##  Max.   :697.000   Max.   :1424.00   Max.   :570.00   Max.   :2336.0
##   mas_vnr_area      garage_cars     garage_area       year_built
##  Min.   :   0.00   Min.   :0.000   Min.   :   0.0   Min.   :1872
##  1st Qu.:   0.00   1st Qu.:1.000   1st Qu.: 336.0   1st Qu.:1954
##  Median :   0.00   Median :2.000   Median : 480.0   Median :1972
##  Mean   :  95.38   Mean   :1.764   Mean   : 471.9   Mean   :1970
##  3rd Qu.: 149.25   3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.:1998
##  Max.   :1600.00   Max.   :5.000   Max.   :1356.0   Max.   :2009
##  tot_rms_abv_grd   full_bath    overall_qual       kitchen_qual
##  Min.   : 3.00   Min.   :0.00   Length:1440        Length:1440
##  1st Qu.: 5.00   1st Qu.:1.00   Class :character   Class :character
##  Median : 6.00   Median :2.00   Mode  :character   Mode  :character
##  Mean   : 6.36   Mean   :1.54
##  3rd Qu.: 7.00   3rd Qu.:2.00
##  Max.   :12.00   Max.   :4.00
##    fireplaces     fireplace_qu        exter_qual         lot_frontage
##  Min.   :0.0000   Length:1440        Length:1440        Min.   :  0.00
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 37.00
```

```
##  Median :1.0000   Mode  :character   Mode  :character   Median : 60.00
##  Mean   :0.5972                                        Mean   : 54.98
##  3rd Qu.:1.0000                                        3rd Qu.: 76.00
##  Max.   :3.0000                                        Max.   :174.00
##     lot_area        longitude         latitude        misc_val
##  Min.   :  1470   Min.   :-93.69   Min.   :41.99   Min.   :    0.00
##  1st Qu.:  7238   1st Qu.:-93.66   1st Qu.:42.02   1st Qu.:    0.00
##  Median :  9306   Median :-93.64   Median :42.03   Median :    0.00
##  Mean   : 10101   Mean   :-93.64   Mean   :42.03   Mean   :   54.42
##  3rd Qu.: 11425   3rd Qu.:-93.62   3rd Qu.:42.05   3rd Qu.:    0.00
##  Max.   :164660   Max.   :-93.58   Max.   :42.06   Max.   :15500.00
##    year_sold      sale_price
##  Min.   :2006   Min.   : 52000
##  1st Qu.:2007   1st Qu.:130000
##  Median :2008   Median :159000
##  Mean   :2008   Mean   :177568
##  3rd Qu.:2009   3rd Qu.:207000
##  Max.   :2010   Max.   :755000
```

```
skimr::skim(training_data)
```

Table 1: Data summary

Name	training_data
Number of rows	1440
Number of columns	26
Column type frequency:	
character	4
numeric	22
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
overall_qual	0	1	4	14	0	8	0
kitchen_qual	0	1	4	9	0	4	0
fireplace_qu	0	1	4	12	0	6	0
exter_qual	0	1	4	9	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
gr_liv_area	0	1	1477.55	484.87	492.00	1092.75	1432.50	1721.00	4316.00	
first_flr_sf	0	1	1134.25	367.92	372.00	870.50	1055.00	1346.25	3228.00	
second_flr_sf	0	1	338.62	422.77	0.00	0.00	0.00	704.75	1872.00	
total_bsmt_sf	0	1	1035.09	413.85	0.00	792.50	975.00	1262.25	3206.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
low_qual_fin_sf	0	1	4.68	44.33	0.00	0.00	0.00	0.00	697.00	
wood_deck_sf	0	1	98.43	133.83	0.00	0.00	0.00	169.00	1424.00	
open_porch_sf	0	1	45.14	63.49	0.00	0.00	26.00	68.00	570.00	
bsmt_unf_sf	0	1	520.57	415.02	0.00	193.75	431.50	768.25	2336.00	
mas_vnr_area	0	1	95.38	168.00	0.00	0.00	0.00	149.25	1600.00	
garage_cars	0	1	1.76	0.72	0.00	1.00	2.00	2.00	5.00	
garage_area	0	1	471.94	201.58	0.00	336.00	480.00	576.00	1356.00	
year_built	0	1	1970.22	29.36	1872.00	1954.00	1972.00	1998.00	2009.00	
tot_rms_abv_grd	0	1	6.36	1.53	3.00	5.00	6.00	7.00	12.00	
full_bath	0	1	1.54	0.54	0.00	1.00	2.00	2.00	4.00	
fireplaces	0	1	0.60	0.65	0.00	0.00	1.00	1.00	3.00	
lot_frontage	0	1	54.98	32.50	0.00	37.00	60.00	76.00	174.00	
lot_area	0	1	10101.03	8302.15	1470.00	7238.25	9306.50	11425.25	164660.00	
longitude	0	1	-93.64	0.03	-93.69	-93.66	-93.64	-93.62	-93.58	
latitude	0	1	42.03	0.02	41.99	42.02	42.03	42.05	42.06	
misc_val	0	1	54.42	590.37	0.00	0.00	0.00	0.00	15500.00	
year_sold	0	1	2007.88	1.29	2006.00	2007.00	2008.00	2009.00	2010.00	
sale_price	0	1	177568.50	73659.38	52000.00	130000.00	159000.00	207000.00	755000.00	

```
# Summary statistics of test data
summary(test_data)
```

```
##   gr_liv_area    first_flr_sf    second_flr_sf    total_bsmt_sf
##  Min.   : 520   Min.   : 453.0   Min.   :   0.0   Min.   :   0
##  1st Qu.:1120   1st Qu.: 866.5   1st Qu.:   0.0   1st Qu.: 784
##  Median :1432   Median :1077.0   Median :   0.0   Median : 970
##  Mean   :1487   Mean   :1139.5   Mean   : 344.2   Mean   :1015
##  3rd Qu.:1730   3rd Qu.:1358.0   3rd Qu.: 706.0   3rd Qu.:1225
##  Max.   :3820   Max.   :3820.0   Max.   :1796.0   Max.   :2461
##  low_qual_fin_sf     wood_deck_sf    open_porch_sf     bsmt_unf_sf
##  Min.   :   0.000   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:   0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 230.0
##  Median :   0.000   Median :  0.00   Median : 24.00   Median : 460.0
##  Mean   :   3.603   Mean   : 93.26   Mean   : 44.46   Mean   : 545.2
##  3rd Qu.:   0.000   3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.: 783.5
##  Max.   :1064.000   Max.   :870.00   Max.   :547.00   Max.   :2153.0
##   mas_vnr_area      garage_cars     garage_area       year_built
##  Min.   :   0.00   Min.   :0.000   Min.   :   0.0   Min.   :1875
##  1st Qu.:   0.00   1st Qu.:1.000   1st Qu.: 311.0   1st Qu.:1952
##  Median :   0.00   Median :2.000   Median : 460.0   Median :1970
##  Mean   :  97.43   Mean   :1.699   Mean   : 449.5   Mean   :1969
##  3rd Qu.: 148.00   3rd Qu.:2.000   3rd Qu.: 564.0   3rd Qu.:1998
##  Max.   :1378.00   Max.   :4.000   Max.   :1488.0   Max.   :2010
##  tot_rms_abv_grd   full_bath     overall_qual       kitchen_qual
##  Min.   : 3.00   Min.   :0.000   Length:959         Length:959
##  1st Qu.: 5.00   1st Qu.:1.000   Class :character   Class :character
##  Median : 6.00   Median :2.000   Mode  :character   Mode  :character
##  Mean   : 6.41   Mean   :1.546
##  3rd Qu.: 7.00   3rd Qu.:2.000
##  Max.   :13.00   Max.   :4.000
##    fireplaces     fireplace_qu        exter_qual         lot_frontage
```

```
##  Min.   :0.0000   Length:959         Length:959         Min.   :  0.00
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 36.00
##  Median :1.0000   Mode  :character   Mode  :character   Median : 62.00
##  Mean   :0.6194                                         Mean   : 56.12
##  3rd Qu.:1.0000                                         3rd Qu.: 78.00
##  Max.   :4.0000                                         Max.   :313.00
##     lot_area        longitude         latitude        misc_val
##  Min.   :  1300   Min.   :-93.69   Min.   :41.99   Min.   :   0.00
##  1st Qu.:  7500   1st Qu.:-93.66   1st Qu.:42.02   1st Qu.:   0.00
##  Median :  9480   Median :-93.64   Median :42.03   Median :   0.00
##  Mean   :  9987   Mean   :-93.64   Mean   :42.03   Mean   :  47.81
##  3rd Qu.: 11384   3rd Qu.:-93.62   3rd Qu.:42.05   3rd Qu.:   0.00
##  Max.   :215245   Max.   :-93.60   Max.   :42.06   Max.   :8300.00
##    year_sold      sale_price
##  Min.   :2006   Min.   : 40000
##  1st Qu.:2007   1st Qu.:130000
##  Median :2008   Median :159000
##  Mean   :2008   Mean   :174214
##  3rd Qu.:2009   3rd Qu.:206450
##  Max.   :2010   Max.   :625000
```

```
skimr::skim(test_data)
```

Table 4: Data summary

Name	test_data
Number of rows	959
Number of columns	26
Column type frequency:	
character	4
numeric	22
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
overall_qual	0	1	4	14	0	8	0
kitchen_qual	0	1	4	9	0	4	0
fireplace_qu	0	1	4	12	0	6	0
exter_qual	0	1	4	9	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
gr_liv_area	0	1	1487.35	475.01	520.00	1120.50	1432.00	1730.00	3820.00	
first_flr_sf	0	1	1139.52	361.70	453.00	866.50	1077.00	1358.00	3820.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
second_flr_sf	0	1	344.22	425.38	0.00	0.00	0.00	706.00	1796.00	
total_bsmt_sf	0	1	1015.13	393.13	0.00	784.00	970.00	1225.00	2461.00	
low_qual_fin_sf	0	1	3.60	45.86	0.00	0.00	0.00	0.00	1064.00	
wood_deck_sf	0	1	93.26	125.19	0.00	0.00	0.00	168.00	870.00	
open_porch_sf	0	1	44.46	62.37	0.00	0.00	24.00	68.00	547.00	
bsmt_unf_sf	0	1	545.16	420.85	0.00	230.00	460.00	783.50	2153.00	
mas_vnr_area	0	1	97.43	179.33	0.00	0.00	0.00	148.00	1378.00	
garage_cars	0	1	1.70	0.73	0.00	1.00	2.00	2.00	4.00	
garage_area	0	1	449.48	198.84	0.00	311.00	460.00	564.00	1488.00	
year_built	0	1	1968.80	29.50	1875.00	1952.50	1970.00	1998.00	2010.00	
tot_rms_abv_grd	0	1	6.41	1.48	3.00	5.00	6.00	7.00	13.00	
full_bath	0	1	1.55	0.55	0.00	1.00	2.00	2.00	4.00	
fireplaces	0	1	0.62	0.65	0.00	0.00	1.00	1.00	4.00	
lot_frontage	0	1	56.12	35.06	0.00	36.00	62.00	78.00	313.00	
lot_area	0	1	9987.10	8123.03	1300.00	7500.00	9480.00	11384.00	215245.00	
longitude	0	1	-93.64	0.03	-93.69	-93.66	-93.64	-93.62	-93.60	
latitude	0	1	42.03	0.02	41.99	42.02	42.03	42.05	42.06	
misc_val	0	1	47.81	378.58	0.00	0.00	0.00	0.00	8300.00	
year_sold	0	1	2007.84	1.34	2006.00	2007.00	2008.00	2009.00	2010.00	
sale_price	0	1	174213.92	65843.62	40000.00	130000.00	159000.00	206450.00	625000.00	

```
table(sapply(training_data, class))
```

```
##
## character   numeric
##         4        22
```

```
# NOTE TO SELF LATER: use summarytools package?
```

To predict our single continuous, numeric outcome (`sale_price`), we have 25 predictors, including 21 numeric variables (such as `lot_area` and `open_porch_sf`) and 4 factor variables (`overall_qual`, `kitchen_qual`, `fireplace_qu`, and `external_qual`). Our training data has 1440observations, while our testing data has 959 hold-out observations.

Before developing any models, we create appropriate vectors and matrices for future use:

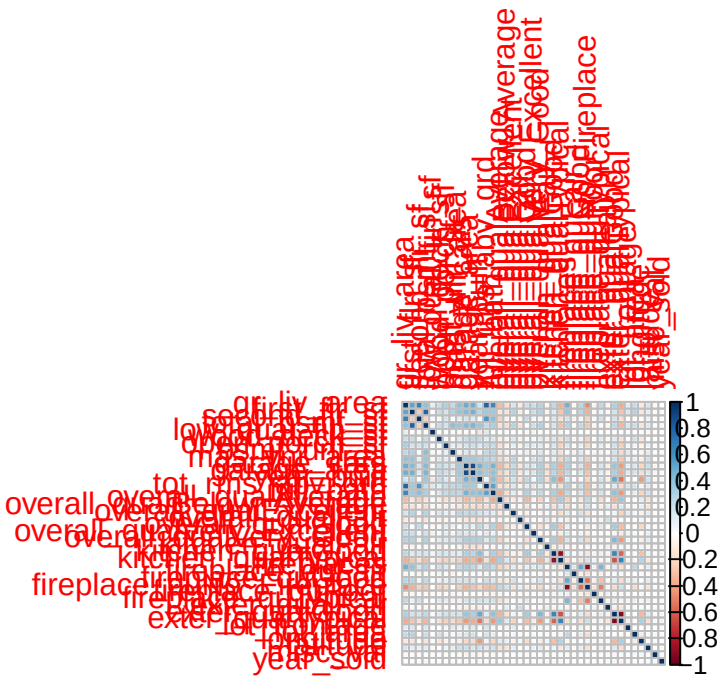
```
training_predictors_matrix = model.matrix(sale_price ~ ., training_data)[, -1]
training_outcomes_vector = training_data$sale_price

testing_predictors_matrix = model.matrix(sale_price ~ ., test_data)[, -1]
testing_outcomes_matrix = test_data$sale_price
testing_matrix_all = model.matrix(sale_price ~ ., test_data)
```

Linear regression

Correlation Plots and Predictor Subsetting

For completeness, let's check for potential collinearities between predictors in our training data.

```
# Correlation plot for all predictors
corrplot(cor(training_predictors_matrix), method = "circle", type = "full")
```




```
# Correlation matrix
round(rcorr(training_predictors_matrix)$r, 2)
```

```
##                           gr_liv_area first_flr_sf second_flr_sf total_bsmt_sf
## gr_liv_area                      1.00         0.53          0.67          0.40
## first_flr_sf                     0.53         1.00         -0.26          0.81
## second_flr_sf                    0.67        -0.26          1.00         -0.23
## total_bsmt_sf                    0.40         0.81         -0.23          1.00
## low_qual_fin_sf                  0.12         0.00          0.04         -0.03
## wood_deck_sf                     0.24         0.25          0.05          0.25
## open_porch_sf                    0.32         0.19          0.20          0.20
## bsmt_unf_sf                      0.26         0.31          0.02          0.41
## mas_vnr_area                     0.40         0.40          0.12          0.39
## garage_cars                      0.50         0.44          0.19          0.43
## garage_area                      0.48         0.47          0.15          0.45
## year_built                       0.20         0.28          0.00          0.37
## tot_rms_abv_grd                  0.80         0.36          0.59          0.24
## full_bath                        0.65         0.37          0.42          0.32
## overall_qualAverage             -0.32        -0.14         -0.24         -0.15
## overall_qualBelow_Average       -0.20        -0.16         -0.09         -0.19
## overall_qualExcellent            0.19         0.29         -0.04          0.30
## overall_qualFair                -0.13        -0.09         -0.06         -0.15
## overall_qualGood                 0.22         0.01          0.24          0.03
## overall_qualVery_Excellent       0.23         0.22          0.06          0.21
## overall_qualVery_Good            0.30         0.31          0.07          0.31
## kitchen_qualFair                -0.07        -0.07         -0.02         -0.05
## kitchen_qualGood                 0.26         0.15          0.17          0.19
## kitchen_qualTypical             -0.35        -0.27         -0.17         -0.31
```

```
## fireplaces                        0.47         0.40          0.19          0.30
## fireplace_quFair                 -0.05         0.03         -0.08          0.03
## fireplace_quGood                  0.27         0.29          0.06          0.25
## fireplace_quNo_Fireplace         -0.46        -0.35         -0.22         -0.27
## fireplace_quPoor                 -0.07        -0.04         -0.05         -0.04
## fireplace_quTypical               0.27         0.10          0.22          0.03
## exter_qualFair                   -0.06        -0.05         -0.03         -0.06
## exter_qualGood                    0.32         0.20          0.19          0.25
## exter_qualTypical                -0.37        -0.29         -0.18         -0.34
## lot_frontage                      0.14         0.17          0.01          0.17
## lot_area                          0.25         0.29          0.03          0.22
## longitude                        -0.13        -0.14         -0.03         -0.16
## latitude                          0.18         0.12          0.11          0.15
## misc_val                         -0.04        -0.01         -0.03         -0.01
## year_sold                         0.01         0.04         -0.02          0.07
##                          low_qual_fin_sf wood_deck_sf open_porch_sf
## gr_liv_area                         0.12         0.24          0.32
## first_flr_sf                        0.00         0.25          0.19
## second_flr_sf                       0.04         0.05          0.20
## total_bsmt_sf                      -0.03         0.25          0.20
## low_qual_fin_sf                     1.00         0.00          0.02
## wood_deck_sf                        0.00         1.00          0.04
## open_porch_sf                       0.02         0.04          1.00
## bsmt_unf_sf                         0.04        -0.04          0.11
## mas_vnr_area                       -0.06         0.15          0.14
## garage_cars                        -0.04         0.24          0.19
## garage_area                        -0.02         0.25          0.21
## year_built                         -0.15         0.25          0.17
## tot_rms_abv_grd                     0.15         0.13          0.23
## full_bath                           0.01         0.17          0.25
## overall_qualAverage                 0.01        -0.11         -0.17
## overall_qualBelow_Average           0.04        -0.07         -0.10
## overall_qualExcellent              -0.02         0.16          0.11
## overall_qualFair                   -0.01        -0.08         -0.07
## overall_qualGood                   -0.03         0.05          0.14
## overall_qualVery_Excellent          0.08         0.05          0.13
## overall_qualVery_Good              -0.03         0.13          0.14
## kitchen_qualFair                   -0.01        -0.09         -0.04
## kitchen_qualGood                   -0.07         0.17          0.18
## kitchen_qualTypical                 0.05        -0.20         -0.22
## fireplaces                          0.01         0.22          0.17
## fireplace_quFair                   -0.02         0.08         -0.02
## fireplace_quGood                    0.01         0.09          0.11
## fireplace_quNo_Fireplace            0.00        -0.23         -0.20
## fireplace_quPoor                   -0.01        -0.03         -0.01
## fireplace_quTypical                 0.01         0.15          0.12
## exter_qualFair                     -0.01        -0.06         -0.04
## exter_qualGood                     -0.02         0.15          0.21
## exter_qualTypical                   0.03        -0.17         -0.24
## lot_frontage                        0.03         0.00          0.07
## lot_area                            0.01         0.21          0.08
## longitude                           0.03        -0.15         -0.07
## latitude                           -0.05         0.00          0.09
## misc_val                           -0.01         0.02          0.01
```

```
## year_sold                          0.00         0.03         -0.02
##                           bsmt_unf_sf mas_vnr_area garage_cars garage_area
## gr_liv_area                      0.26         0.40        0.50        0.48
## first_flr_sf                     0.31         0.40        0.44        0.47
## second_flr_sf                    0.02         0.12        0.19        0.15
## total_bsmt_sf                    0.41         0.39        0.43        0.45
## low_qual_fin_sf                  0.04        -0.06       -0.04       -0.02
## wood_deck_sf                    -0.04         0.15        0.24        0.25
## open_porch_sf                    0.11         0.14        0.19        0.21
## bsmt_unf_sf                      1.00         0.08        0.17        0.15
## mas_vnr_area                     0.08         1.00        0.36        0.35
## garage_cars                      0.17         0.36        1.00        0.88
## garage_area                      0.15         0.35        0.88        1.00
## year_built                       0.07         0.32        0.48        0.41
## tot_rms_abv_grd                  0.25         0.27        0.37        0.34
## full_bath                        0.28         0.27        0.47        0.40
## overall_qualAverage             -0.14        -0.21       -0.32       -0.24
## overall_qualBelow_Average       -0.08        -0.14       -0.21       -0.17
## overall_qualExcellent            0.08         0.27        0.25        0.25
## overall_qualFair                -0.04        -0.07       -0.13       -0.12
## overall_qualGood                 0.09         0.02        0.16        0.11
## overall_qualVery_Excellent       0.06         0.18        0.15        0.18
## overall_qualVery_Good            0.16         0.22        0.32        0.32
## kitchen_qualFair                 0.01        -0.06       -0.18       -0.16
## kitchen_qualGood                 0.14         0.13        0.32        0.27
## kitchen_qualTypical             -0.17        -0.23       -0.37       -0.34
## fireplaces                      -0.02         0.25        0.31        0.27
## fireplace_quFair                -0.09         0.00        0.03        0.03
## fireplace_quGood                 0.15         0.17        0.19        0.17
## fireplace_quNo_Fireplace        -0.02        -0.25       -0.35       -0.29
## fireplace_quPoor                -0.07        -0.03       -0.02       -0.03
## fireplace_quTypical             -0.09         0.09        0.18        0.14
## exter_qualFair                   0.00        -0.06       -0.11       -0.09
## exter_qualGood                   0.19         0.19        0.38        0.32
## exter_qualTypical               -0.20        -0.28       -0.43       -0.38
## lot_frontage                     0.16         0.09        0.12        0.16
## lot_area                         0.03         0.10        0.17        0.19
## longitude                       -0.04        -0.09       -0.21       -0.18
## latitude                         0.13         0.22        0.23        0.17
## misc_val                        -0.03        -0.02       -0.06       -0.05
## year_sold                        0.03         0.01        0.04        0.03
##                           year_built tot_rms_abv_grd full_bath
## gr_liv_area                     0.20            0.80      0.65
## first_flr_sf                    0.28            0.36      0.37
## second_flr_sf                   0.00            0.59      0.42
## total_bsmt_sf                   0.37            0.24      0.32
## low_qual_fin_sf                -0.15            0.15      0.01
## wood_deck_sf                    0.25            0.13      0.17
## open_porch_sf                   0.17            0.23      0.25
## bsmt_unf_sf                     0.07            0.25      0.28
## mas_vnr_area                    0.32            0.27      0.27
## garage_cars                     0.48            0.37      0.47
## garage_area                     0.41            0.34      0.40
## year_built                      1.00            0.07      0.43
```

```
## tot_rms_abv_grd                 0.07           1.00      0.54
## full_bath                       0.43           0.54      1.00
## overall_qualAverage            -0.31          -0.16     -0.37
## overall_qualBelow_Average      -0.20          -0.17     -0.18
## overall_qualExcellent           0.20           0.15      0.13
## overall_qualFair               -0.16          -0.06     -0.05
## overall_qualGood                0.26           0.14      0.30
## overall_qualVery_Excellent      0.05           0.15      0.09
## overall_qualVery_Good           0.30           0.15      0.25
## kitchen_qualFair               -0.17          -0.04     -0.05
## kitchen_qualGood                0.42           0.11      0.33
## kitchen_qualTypical            -0.45          -0.18     -0.39
## fireplaces                      0.15           0.31      0.26
## fireplace_quFair                0.03          -0.06     -0.02
## fireplace_quGood                0.04           0.21      0.16
## fireplace_quNo_Fireplace       -0.21          -0.32     -0.30
## fireplace_quPoor               -0.02          -0.06     -0.08
## fireplace_quTypical             0.19           0.17      0.20
## exter_qualFair                 -0.11          -0.02     -0.06
## exter_qualGood                  0.51           0.15      0.41
## exter_qualTypical              -0.52          -0.20     -0.43
## lot_frontage                   -0.02           0.18      0.07
## lot_area                        0.00           0.18      0.12
## longitude                      -0.42          -0.07     -0.21
## latitude                        0.20           0.15      0.20
## misc_val                       -0.04          -0.02     -0.04
## year_sold                       0.05          -0.01      0.05
##                          overall_qualAverage overall_qualBelow_Average
## gr_liv_area                            -0.32                     -0.20
## first_flr_sf                           -0.14                     -0.16
## second_flr_sf                          -0.24                     -0.09
## total_bsmt_sf                          -0.15                     -0.19
## low_qual_fin_sf                         0.01                      0.04
## wood_deck_sf                           -0.11                     -0.07
## open_porch_sf                          -0.17                     -0.10
## bsmt_unf_sf                            -0.14                     -0.08
## mas_vnr_area                           -0.21                     -0.14
## garage_cars                            -0.32                     -0.21
## garage_area                            -0.24                     -0.17
## year_built                             -0.31                     -0.20
## tot_rms_abv_grd                        -0.16                     -0.17
## full_bath                              -0.37                     -0.18
## overall_qualAverage                     1.00                     -0.18
## overall_qualBelow_Average              -0.18                      1.00
## overall_qualExcellent                  -0.11                     -0.05
## overall_qualFair                       -0.08                     -0.03
## overall_qualGood                       -0.33                     -0.14
## overall_qualVery_Excellent             -0.06                     -0.03
## overall_qualVery_Good                  -0.22                     -0.10
## kitchen_qualFair                        0.05                      0.09
## kitchen_qualGood                       -0.30                     -0.16
## kitchen_qualTypical                     0.33                      0.16
## fireplaces                             -0.23                     -0.18
## fireplace_quFair                        0.01                     -0.02
```

```
## fireplace_quGood                          -0.14                    -0.12
## fireplace_quNo_Fireplace                   0.26                     0.20
## fireplace_quPoor                           0.05                     0.00
## fireplace_quTypical                       -0.18                    -0.11
## exter_qualFair                            -0.02                     0.13
## exter_qualGood                            -0.34                    -0.17
## exter_qualTypical                          0.37                     0.15
## lot_frontage                               0.02                    -0.06
## lot_area                                   0.00                    -0.05
## longitude                                  0.13                     0.10
## latitude                                  -0.16                    -0.16
## misc_val                                  -0.03                     0.01
## year_sold                                 -0.02                     0.01
##                           overall_qualExcellent overall_qualFair
## gr_liv_area                                0.19            -0.13
## first_flr_sf                               0.29            -0.09
## second_flr_sf                             -0.04            -0.06
## total_bsmt_sf                              0.30            -0.15
## low_qual_fin_sf                           -0.02            -0.01
## wood_deck_sf                               0.16            -0.08
## open_porch_sf                              0.11            -0.07
## bsmt_unf_sf                                0.08            -0.04
## mas_vnr_area                               0.27            -0.07
## garage_cars                                0.25            -0.13
## garage_area                                0.25            -0.12
## year_built                                 0.20            -0.16
## tot_rms_abv_grd                            0.15            -0.06
## full_bath                                  0.13            -0.05
## overall_qualAverage                       -0.11            -0.08
## overall_qualBelow_Average                 -0.05            -0.03
## overall_qualExcellent                      1.00            -0.02
## overall_qualFair                          -0.02             1.00
## overall_qualGood                          -0.09            -0.06
## overall_qualVery_Excellent                -0.02            -0.01
## overall_qualVery_Good                     -0.06            -0.04
## kitchen_qualFair                          -0.02             0.12
## kitchen_qualGood                          -0.06            -0.08
## kitchen_qualTypical                       -0.19             0.06
## fireplaces                                 0.14            -0.10
## fireplace_quFair                          -0.03            -0.02
## fireplace_quGood                           0.19            -0.04
## fireplace_quNo_Fireplace                  -0.15             0.10
## fireplace_quPoor                          -0.02            -0.02
## fireplace_quTypical                       -0.02            -0.06
## exter_qualFair                            -0.02             0.21
## exter_qualGood                             0.07            -0.06
## exter_qualTypical                         -0.23             0.02
## lot_frontage                               0.13             0.01
## lot_area                                   0.05            -0.02
## longitude                                 -0.07             0.06
## latitude                                   0.13            -0.02
## misc_val                                  -0.02             0.02
## year_sold                                  0.02            -0.02
##                           overall_qualGood overall_qualVery_Excellent
```

```
## gr_liv_area                          0.22                    0.23
## first_flr_sf                         0.01                    0.22
## second_flr_sf                        0.24                    0.06
## total_bsmt_sf                        0.03                    0.21
## low_qual_fin_sf                     -0.03                    0.08
## wood_deck_sf                         0.05                    0.05
## open_porch_sf                        0.14                    0.13
## bsmt_unf_sf                          0.09                    0.06
## mas_vnr_area                         0.02                    0.18
## garage_cars                          0.16                    0.15
## garage_area                          0.11                    0.18
## year_built                           0.26                    0.05
## tot_rms_abv_grd                      0.14                    0.15
## full_bath                            0.30                    0.09
## overall_qualAverage                 -0.33                   -0.06
## overall_qualBelow_Average           -0.14                   -0.03
## overall_qualExcellent               -0.09                   -0.02
## overall_qualFair                    -0.06                   -0.01
## overall_qualGood                     1.00                   -0.05
## overall_qualVery_Excellent          -0.05                    1.00
## overall_qualVery_Good               -0.18                   -0.03
## kitchen_qualFair                    -0.07                   -0.01
## kitchen_qualGood                     0.35                   -0.06
## kitchen_qualTypical                 -0.29                   -0.10
## fireplaces                           0.11                    0.14
## fireplace_quFair                    -0.01                   -0.02
## fireplace_quGood                     0.01                    0.07
## fireplace_quNo_Fireplace            -0.11                   -0.09
## fireplace_quPoor                    -0.03                   -0.01
## fireplace_quTypical                  0.14                   -0.01
## exter_qualFair                      -0.04                   -0.01
## exter_qualGood                       0.34                   -0.05
## exter_qualTypical                   -0.31                   -0.13
## lot_frontage                        -0.04                    0.09
## lot_area                             0.01                    0.04
## longitude                           -0.21                   -0.02
## latitude                             0.05                    0.10
## misc_val                            -0.02                   -0.01
## year_sold                            0.00                    0.01
##                          overall_qualVery_Good kitchen_qualFair
## gr_liv_area                               0.30            -0.07
## first_flr_sf                              0.31            -0.07
## second_flr_sf                             0.07            -0.02
## total_bsmt_sf                             0.31            -0.05
## low_qual_fin_sf                          -0.03            -0.01
## wood_deck_sf                              0.13            -0.09
## open_porch_sf                             0.14            -0.04
## bsmt_unf_sf                               0.16             0.01
## mas_vnr_area                              0.22            -0.06
## garage_cars                               0.32            -0.18
## garage_area                               0.32            -0.16
## year_built                                0.30            -0.17
## tot_rms_abv_grd                           0.15            -0.04
## full_bath                                 0.25            -0.05
```

```
## overall_qualAverage                      -0.22             0.05
## overall_qualBelow_Average                -0.10             0.09
## overall_qualExcellent                    -0.06            -0.02
## overall_qualFair                         -0.04             0.12
## overall_qualGood                         -0.18            -0.07
## overall_qualVery_Excellent               -0.03            -0.01
## overall_qualVery_Good                     1.00            -0.04
## kitchen_qualFair                         -0.04             1.00
## kitchen_qualGood                          0.34            -0.10
## kitchen_qualTypical                      -0.35            -0.14
## fireplaces                                0.18            -0.05
## fireplace_quFair                         -0.03            -0.02
## fireplace_quGood                          0.19             0.00
## fireplace_quNo_Fireplace                 -0.21             0.06
## fireplace_quPoor                         -0.05            -0.02
## fireplace_quTypical                       0.07            -0.05
## exter_qualFair                           -0.03             0.09
## exter_qualGood                            0.44            -0.08
## exter_qualTypical                        -0.42             0.06
## lot_frontage                              0.08             0.05
## lot_area                                  0.02            -0.02
## longitude                                -0.11             0.07
## latitude                                  0.13            -0.05
## misc_val                                 -0.02             0.04
## year_sold                                 0.04            -0.05
##                          kitchen_qualGood kitchen_qualTypical fireplaces
## gr_liv_area                          0.26               -0.35       0.47
## first_flr_sf                         0.15               -0.27       0.40
## second_flr_sf                        0.17               -0.17       0.19
## total_bsmt_sf                        0.19               -0.31       0.30
## low_qual_fin_sf                     -0.07                0.05       0.01
## wood_deck_sf                         0.17               -0.20       0.22
## open_porch_sf                        0.18               -0.22       0.17
## bsmt_unf_sf                          0.14               -0.17      -0.02
## mas_vnr_area                         0.13               -0.23       0.25
## garage_cars                          0.32               -0.37       0.31
## garage_area                          0.27               -0.34       0.27
## year_built                           0.42               -0.45       0.15
## tot_rms_abv_grd                      0.11               -0.18       0.31
## full_bath                            0.33               -0.39       0.26
## overall_qualAverage                 -0.30                0.33      -0.23
## overall_qualBelow_Average           -0.16                0.16      -0.18
## overall_qualExcellent               -0.06               -0.19       0.14
## overall_qualFair                    -0.08                0.06      -0.10
## overall_qualGood                     0.35               -0.29       0.11
## overall_qualVery_Excellent          -0.06               -0.10       0.14
## overall_qualVery_Good                0.34               -0.35       0.18
## kitchen_qualFair                    -0.10               -0.14      -0.05
## kitchen_qualGood                     1.00               -0.87       0.10
## kitchen_qualTypical                 -0.87                1.00      -0.16
## fireplaces                           0.10               -0.16       1.00
## fireplace_quFair                     0.01                0.01       0.15
## fireplace_quGood                     0.08               -0.17       0.51
## fireplace_quNo_Fireplace            -0.14                0.19      -0.90
```

```
## fireplace_quPoor                       -0.06                0.08       0.08
## fireplace_quTypical                     0.10               -0.06       0.46
## exter_qualFair                         -0.07                0.05      -0.06
## exter_qualGood                          0.65               -0.63       0.13
## exter_qualTypical                      -0.58                0.66      -0.17
## lot_frontage                            0.02               -0.09       0.03
## lot_area                                0.03               -0.04       0.26
## longitude                              -0.30                0.30      -0.05
## latitude                                0.09               -0.14       0.12
## misc_val                               -0.01                0.01      -0.02
## year_sold                               0.05               -0.05       0.01
##                           fireplace_quFair fireplace_quGood
## gr_liv_area                           -0.05             0.27
## first_flr_sf                           0.03             0.29
## second_flr_sf                         -0.08             0.06
## total_bsmt_sf                          0.03             0.25
## low_qual_fin_sf                       -0.02             0.01
## wood_deck_sf                           0.08             0.09
## open_porch_sf                         -0.02             0.11
## bsmt_unf_sf                           -0.09             0.15
## mas_vnr_area                           0.00             0.17
## garage_cars                            0.03             0.19
## garage_area                            0.03             0.17
## year_built                             0.03             0.04
## tot_rms_abv_grd                       -0.06             0.21
## full_bath                             -0.02             0.16
## overall_qualAverage                    0.01            -0.14
## overall_qualBelow_Average             -0.02            -0.12
## overall_qualExcellent                 -0.03             0.19
## overall_qualFair                      -0.02            -0.04
## overall_qualGood                      -0.01             0.01
## overall_qualVery_Excellent            -0.02             0.07
## overall_qualVery_Good                 -0.03             0.19
## kitchen_qualFair                      -0.02             0.00
## kitchen_qualGood                       0.01             0.08
## kitchen_qualTypical                    0.01            -0.17
## fireplaces                             0.15             0.51
## fireplace_quFair                       1.00            -0.10
## fireplace_quGood                      -0.10             1.00
## fireplace_quNo_Fireplace              -0.17            -0.55
## fireplace_quPoor                      -0.02            -0.07
## fireplace_quTypical                   -0.09            -0.29
## exter_qualFair                        -0.02            -0.04
## exter_qualGood                        -0.03             0.13
## exter_qualTypical                      0.04            -0.19
## lot_frontage                          -0.01             0.10
## lot_area                               0.00             0.11
## longitude                             -0.01             0.01
## latitude                              -0.03             0.12
## misc_val                              -0.01            -0.02
## year_sold                              0.01             0.02
##                           fireplace_quNo_Fireplace fireplace_quPoor
## gr_liv_area                                -0.46           -0.07
## first_flr_sf                               -0.35           -0.04
```

```
## second_flr_sf                                -0.22           -0.05
## total_bsmt_sf                                -0.27           -0.04
## low_qual_fin_sf                               0.00           -0.01
## wood_deck_sf                                 -0.23           -0.03
## open_porch_sf                                -0.20           -0.01
## bsmt_unf_sf                                  -0.02           -0.07
## mas_vnr_area                                 -0.25           -0.03
## garage_cars                                  -0.35           -0.02
## garage_area                                  -0.29           -0.03
## year_built                                   -0.21           -0.02
## tot_rms_abv_grd                              -0.32           -0.06
## full_bath                                    -0.30           -0.08
## overall_qualAverage                           0.26            0.05
## overall_qualBelow_Average                     0.20            0.00
## overall_qualExcellent                        -0.15           -0.02
## overall_qualFair                              0.10           -0.02
## overall_qualGood                             -0.11           -0.03
## overall_qualVery_Excellent                   -0.09           -0.01
## overall_qualVery_Good                        -0.21           -0.05
## kitchen_qualFair                              0.06           -0.02
## kitchen_qualGood                             -0.14           -0.06
## kitchen_qualTypical                           0.19            0.08
## fireplaces                                   -0.90            0.08
## fireplace_quFair                             -0.17           -0.02
## fireplace_quGood                             -0.55           -0.07
## fireplace_quNo_Fireplace                      1.00           -0.13
## fireplace_quPoor                             -0.13            1.00
## fireplace_quTypical                          -0.51           -0.07
## exter_qualFair                                0.06            0.09
## exter_qualGood                               -0.19           -0.06
## exter_qualTypical                             0.22            0.05
## lot_frontage                                 -0.04           -0.01
## lot_area                                     -0.19           -0.01
## longitude                                     0.05            0.03
## latitude                                     -0.17           -0.02
## misc_val                                      0.03            0.00
## year_sold                                    -0.01           -0.01
##                          fireplace_quTypical exter_qualFair exter_qualGood
## gr_liv_area                             0.27         -0.06           0.32
## first_flr_sf                            0.10         -0.05           0.20
## second_flr_sf                           0.22         -0.03           0.19
## total_bsmt_sf                           0.03         -0.06           0.25
## low_qual_fin_sf                         0.01         -0.01          -0.02
## wood_deck_sf                            0.15         -0.06           0.15
## open_porch_sf                           0.12         -0.04           0.21
## bsmt_unf_sf                            -0.09          0.00           0.19
## mas_vnr_area                            0.09         -0.06           0.19
## garage_cars                             0.18         -0.11           0.38
## garage_area                             0.14         -0.09           0.32
## year_built                              0.19         -0.11           0.51
## tot_rms_abv_grd                         0.17         -0.02           0.15
## full_bath                               0.20         -0.06           0.41
## overall_qualAverage                    -0.18         -0.02          -0.34
## overall_qualBelow_Average              -0.11          0.13          -0.17
```

```
## overall_qualExcellent                    -0.02          -0.02           0.07
## overall_qualFair                         -0.06           0.21          -0.06
## overall_qualGood                          0.14          -0.04           0.34
## overall_qualVery_Excellent               -0.01          -0.01          -0.05
## overall_qualVery_Good                     0.07          -0.03           0.44
## kitchen_qualFair                         -0.05           0.09          -0.08
## kitchen_qualGood                          0.10          -0.07           0.65
## kitchen_qualTypical                      -0.06           0.05          -0.63
## fireplaces                                0.46          -0.06           0.13
## fireplace_quFair                         -0.09          -0.02          -0.03
## fireplace_quGood                         -0.29          -0.04           0.13
## fireplace_quNo_Fireplace                 -0.51           0.06          -0.19
## fireplace_quPoor                         -0.07           0.09          -0.06
## fireplace_quTypical                       1.00          -0.05           0.11
## exter_qualFair                           -0.05           1.00          -0.07
## exter_qualGood                            0.11          -0.07           1.00
## exter_qualTypical                        -0.07          -0.14          -0.92
## lot_frontage                             -0.08           0.04           0.04
## lot_area                                  0.11          -0.02          -0.01
## longitude                                -0.06           0.04          -0.33
## latitude                                  0.10          -0.03           0.15
## misc_val                                 -0.01           0.04          -0.05
## year_sold                                -0.02           0.05           0.05
##                          exter_qualTypical lot_frontage lot_area longitude
## gr_liv_area                          -0.37         0.14     0.25     -0.13
## first_flr_sf                         -0.29         0.17     0.29     -0.14
## second_flr_sf                        -0.18         0.01     0.03     -0.03
## total_bsmt_sf                        -0.34         0.17     0.22     -0.16
## low_qual_fin_sf                       0.03         0.03     0.01      0.03
## wood_deck_sf                         -0.17         0.00     0.21     -0.15
## open_porch_sf                        -0.24         0.07     0.08     -0.07
## bsmt_unf_sf                          -0.20         0.16     0.03     -0.04
## mas_vnr_area                         -0.28         0.09     0.10     -0.09
## garage_cars                          -0.43         0.12     0.17     -0.21
## garage_area                          -0.38         0.16     0.19     -0.18
## year_built                           -0.52        -0.02     0.00     -0.42
## tot_rms_abv_grd                      -0.20         0.18     0.18     -0.07
## full_bath                            -0.43         0.07     0.12     -0.21
## overall_qualAverage                   0.37         0.02     0.00      0.13
## overall_qualBelow_Average             0.15        -0.06    -0.05      0.10
## overall_qualExcellent                -0.23         0.13     0.05     -0.07
## overall_qualFair                      0.02         0.01    -0.02      0.06
## overall_qualGood                     -0.31        -0.04     0.01     -0.21
## overall_qualVery_Excellent           -0.13         0.09     0.04     -0.02
## overall_qualVery_Good                -0.42         0.08     0.02     -0.11
## kitchen_qualFair                      0.06         0.05    -0.02      0.07
## kitchen_qualGood                     -0.58         0.02     0.03     -0.30
## kitchen_qualTypical                   0.66        -0.09    -0.04      0.30
## fireplaces                           -0.17         0.03     0.26     -0.05
## fireplace_quFair                      0.04        -0.01     0.00     -0.01
## fireplace_quGood                     -0.19         0.10     0.11      0.01
## fireplace_quNo_Fireplace              0.22        -0.04    -0.19      0.05
## fireplace_quPoor                      0.05        -0.01    -0.01      0.03
## fireplace_quTypical                  -0.07        -0.08     0.11     -0.06
```

```
## exter_qualFair                          -0.14         0.04    -0.02      0.04
## exter_qualGood                          -0.92         0.04    -0.01     -0.33
## exter_qualTypical                        1.00        -0.09    -0.01      0.33
## lot_frontage                            -0.09         1.00     0.02      0.00
## lot_area                                -0.01         0.02     1.00     -0.11
## longitude                                0.33         0.00    -0.11      1.00
## latitude                                -0.19         0.04    -0.03      0.02
## misc_val                                 0.04        -0.02     0.04      0.05
## year_sold                               -0.08         0.04    -0.02     -0.02
##                              latitude misc_val year_sold
## gr_liv_area                      0.18    -0.04      0.01
## first_flr_sf                     0.12    -0.01      0.04
## second_flr_sf                    0.11    -0.03     -0.02
## total_bsmt_sf                    0.15    -0.01      0.07
## low_qual_fin_sf                 -0.05    -0.01      0.00
## wood_deck_sf                     0.00     0.02      0.03
## open_porch_sf                    0.09     0.01     -0.02
## bsmt_unf_sf                      0.13    -0.03      0.03
## mas_vnr_area                     0.22    -0.02      0.01
## garage_cars                      0.23    -0.06      0.04
## garage_area                      0.17    -0.05      0.03
## year_built                       0.20    -0.04      0.05
## tot_rms_abv_grd                  0.15    -0.02     -0.01
## full_bath                        0.20    -0.04      0.05
## overall_qualAverage             -0.16    -0.03     -0.02
## overall_qualBelow_Average       -0.16     0.01      0.01
## overall_qualExcellent            0.13    -0.02      0.02
## overall_qualFair                -0.02     0.02     -0.02
## overall_qualGood                 0.05    -0.02      0.00
## overall_qualVery_Excellent       0.10    -0.01      0.01
## overall_qualVery_Good            0.13    -0.02      0.04
## kitchen_qualFair                -0.05     0.04     -0.05
## kitchen_qualGood                 0.09    -0.01      0.05
## kitchen_qualTypical             -0.14     0.01     -0.05
## fireplaces                       0.12    -0.02      0.01
## fireplace_quFair                -0.03    -0.01      0.01
## fireplace_quGood                 0.12    -0.02      0.02
## fireplace_quNo_Fireplace        -0.17     0.03     -0.01
## fireplace_quPoor                -0.02     0.00     -0.01
## fireplace_quTypical              0.10    -0.01     -0.02
## exter_qualFair                  -0.03     0.04      0.05
## exter_qualGood                   0.15    -0.05      0.05
## exter_qualTypical               -0.19     0.04     -0.08
## lot_frontage                     0.04    -0.02      0.04
## lot_area                        -0.03     0.04     -0.02
## longitude                        0.02     0.05     -0.02
## latitude                         1.00     0.01      0.03
## misc_val                         0.01     1.00      0.03
## year_sold                        0.03     0.03      1.00
```

```
# Check numerically
round(rcorr(training_predictors_matrix)$r, 2) %>%
  as.data.frame() %>%
  pivot_longer(cols = gr_liv_area:year_sold, names_to = "predictor", values_to = "corr") %>%
```

```
  filter(corr < 1) %>%
  arrange(desc(corr)) %>%
  head()
```

```
## # A tibble: 6 x 2
##   predictor        corr
##   <chr>           <dbl>
## 1 garage_area      0.88
## 2 garage_cars      0.88
## 3 total_bsmt_sf    0.81
## 4 first_flr_sf     0.81
## 5 tot_rms_abv_grd  0.8
## 6 gr_liv_area      0.8
```

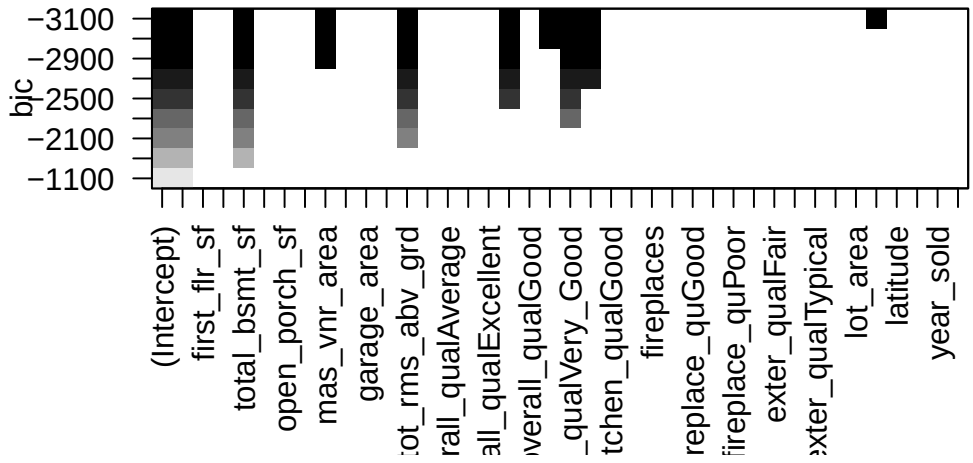
We have a few collinearities greater than 0.8. We may decide to reduce the number of predictors using a `regsubsets` (best subsets) procedure.

```
regsubsets_obj = regsubsets(sale_price ~ ., data = training_data, method = "exhaustive", nbest = 1)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
plot(regsubsets_obj, scale = "bic")
```




```
summary(regsubsets_obj)
```

```
## Subset selection object
## Call: regsubsets.formula(sale_price ~ ., data = training_data, method = "exhaustive",
##     nbest = 1)
## 39 Variables  (and intercept)
##                              Forced in Forced out
## gr_liv_area                      FALSE      FALSE
## first_flr_sf                     FALSE      FALSE
## second_flr_sf                    FALSE      FALSE
## total_bsmt_sf                    FALSE      FALSE
## wood_deck_sf                     FALSE      FALSE
## open_porch_sf                    FALSE      FALSE
## bsmt_unf_sf                      FALSE      FALSE
## mas_vnr_area                     FALSE      FALSE
## garage_cars                      FALSE      FALSE
## garage_area                      FALSE      FALSE
## year_built                       FALSE      FALSE
## tot_rms_abv_grd                  FALSE      FALSE
## full_bath                        FALSE      FALSE
## overall_qualAverage              FALSE      FALSE
## overall_qualBelow_Average        FALSE      FALSE
## overall_qualExcellent            FALSE      FALSE
## overall_qualFair                 FALSE      FALSE
## overall_qualGood                 FALSE      FALSE
## overall_qualVery_Excellent       FALSE      FALSE
## overall_qualVery_Good            FALSE      FALSE
## kitchen_qualFair                 FALSE      FALSE
## kitchen_qualGood                 FALSE      FALSE
## kitchen_qualTypical              FALSE      FALSE
## fireplaces                       FALSE      FALSE
## fireplace_quFair                 FALSE      FALSE
## fireplace_quGood                 FALSE      FALSE
## fireplace_quNo_Fireplace         FALSE      FALSE
## fireplace_quPoor                 FALSE      FALSE
## fireplace_quTypical              FALSE      FALSE
## exter_qualFair                   FALSE      FALSE
## exter_qualGood                   FALSE      FALSE
## exter_qualTypical                FALSE      FALSE
## lot_frontage                     FALSE      FALSE
## lot_area                         FALSE      FALSE
## longitude                        FALSE      FALSE
## latitude                         FALSE      FALSE
## misc_val                         FALSE      FALSE
## year_sold                        FALSE      FALSE
## low_qual_fin_sf                  FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          gr_liv_area first_flr_sf second_flr_sf total_bsmt_sf low_qual_fin_sf
## 1  ( 1 ) "*"         " "          " "           " "           " "
## 2  ( 1 ) "*"         " "          " "           "*"           " "
## 3  ( 1 ) "*"         " "          " "           "*"           " "
## 4  ( 1 ) "*"         " "          " "           "*"           " "
## 5  ( 1 ) "*"         " "          " "           "*"           " "
## 6  ( 1 ) "*"         " "          " "           "*"           " "
## 7  ( 1 ) "*"         " "          " "           "*"           " "
```

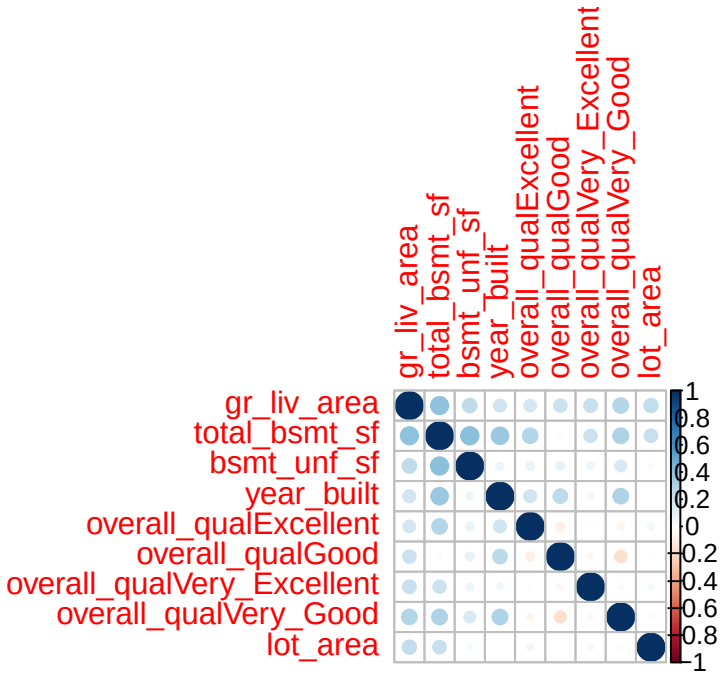
```
## 8  ( 1 ) "*"          " "           " "            "*"            " "
## 9  ( 1 ) "*"          " "           " "            "*"            " "
##          wood_deck_sf open_porch_sf bsmt_unf_sf mas_vnr_area garage_cars
## 1  ( 1 ) " "           " "            " "          " "           " "
## 2  ( 1 ) " "           " "            " "          " "           " "
## 3  ( 1 ) " "           " "            " "          " "           " "
## 4  ( 1 ) " "           " "            " "          " "           " "
## 5  ( 1 ) " "           " "            " "          " "           " "
## 6  ( 1 ) " "           " "            " "          " "           " "
## 7  ( 1 ) " "           " "            "*"          " "           " "
## 8  ( 1 ) " "           " "            "*"          " "           " "
## 9  ( 1 ) " "           " "            "*"          " "           " "
##          garage_area year_built tot_rms_abv_grd full_bath overall_qualAverage
## 1  ( 1 ) " "          " "        " "              " "       " "
## 2  ( 1 ) " "          " "        " "              " "       " "
## 3  ( 1 ) " "          "*"        " "              " "       " "
## 4  ( 1 ) " "          "*"        " "              " "       " "
## 5  ( 1 ) " "          "*"        " "              " "       " "
## 6  ( 1 ) " "          "*"        " "              " "       " "
## 7  ( 1 ) " "          "*"        " "              " "       " "
## 8  ( 1 ) " "          "*"        " "              " "       " "
## 9  ( 1 ) " "          "*"        " "              " "       " "
##          overall_qualBelow_Average overall_qualExcellent overall_qualFair
## 1  ( 1 ) " "                       " "                   " "
## 2  ( 1 ) " "                       " "                   " "
## 3  ( 1 ) " "                       " "                   " "
## 4  ( 1 ) " "                       " "                   " "
## 5  ( 1 ) " "                       "*"                   " "
## 6  ( 1 ) " "                       "*"                   " "
## 7  ( 1 ) " "                       "*"                   " "
## 8  ( 1 ) " "                       "*"                   " "
## 9  ( 1 ) " "                       "*"                   " "
##          overall_qualGood overall_qualVery_Excellent overall_qualVery_Good
## 1  ( 1 ) " "              " "                        " "
## 2  ( 1 ) " "              " "                        " "
## 3  ( 1 ) " "              " "                        " "
## 4  ( 1 ) " "              "*"                        " "
## 5  ( 1 ) " "              "*"                        " "
## 6  ( 1 ) " "              "*"                        "*"
## 7  ( 1 ) " "              "*"                        "*"
## 8  ( 1 ) "*"              "*"                        "*"
## 9  ( 1 ) "*"              "*"                        "*"
##          kitchen_qualFair kitchen_qualGood kitchen_qualTypical fireplaces
## 1  ( 1 ) " "              " "              " "                 " "
## 2  ( 1 ) " "              " "              " "                 " "
## 3  ( 1 ) " "              " "              " "                 " "
## 4  ( 1 ) " "              " "              " "                 " "
## 5  ( 1 ) " "              " "              " "                 " "
## 6  ( 1 ) " "              " "              " "                 " "
## 7  ( 1 ) " "              " "              " "                 " "
## 8  ( 1 ) " "              " "              " "                 " "
## 9  ( 1 ) " "              " "              " "                 " "
##          fireplace_quFair fireplace_quGood fireplace_quNo_Fireplace
## 1  ( 1 ) " "              " "              " "
```

```
## 2  ( 1 ) " "               " "               " "
## 3  ( 1 ) " "               " "               " "
## 4  ( 1 ) " "               " "               " "
## 5  ( 1 ) " "               " "               " "
## 6  ( 1 ) " "               " "               " "
## 7  ( 1 ) " "               " "               " "
## 8  ( 1 ) " "               " "               " "
## 9  ( 1 ) " "               " "               " "
##          fireplace_quPoor fireplace_quTypical exter_qualFair exter_qualGood
## 1  ( 1 ) " "               " "                 " "            " "
## 2  ( 1 ) " "               " "                 " "            " "
## 3  ( 1 ) " "               " "                 " "            " "
## 4  ( 1 ) " "               " "                 " "            " "
## 5  ( 1 ) " "               " "                 " "            " "
## 6  ( 1 ) " "               " "                 " "            " "
## 7  ( 1 ) " "               " "                 " "            " "
## 8  ( 1 ) " "               " "                 " "            " "
## 9  ( 1 ) " "               " "                 " "            " "
##          exter_qualTypical lot_frontage lot_area longitude latitude misc_val
## 1  ( 1 ) " "                " "          " "      " "       " "      " "
## 2  ( 1 ) " "                " "          " "      " "       " "      " "
## 3  ( 1 ) " "                " "          " "      " "       " "      " "
## 4  ( 1 ) " "                " "          " "      " "       " "      " "
## 5  ( 1 ) " "                " "          " "      " "       " "      " "
## 6  ( 1 ) " "                " "          " "      " "       " "      " "
## 7  ( 1 ) " "                " "          " "      " "       " "      " "
## 8  ( 1 ) " "                " "          " "      " "       " "      " "
## 9  ( 1 ) " "                " "          "*"      " "       " "      " "
##          year_sold
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
## 9  ( 1 ) " "
```

As expected, one primary collinearity was found (between `garage_area` and `garage_cars`). Were we to choose only several predictors to use, they would be, according to our algorithm: `gr_liv_area`, `total_bsmt_sf`, `bsmt_unf_sf`, `year_built`, `overall_qualExcellent`, `overall_qualGood`, `overall_qualVery_Excellent`, `overall_qualVery_Good`, and `lot_area`.

We can check out what the correlation matrix looks like with only these predictors.

```
# Correlation plot for only best subset predictors
x_best = training_predictors_matrix[, c("gr_liv_area", "total_bsmt_sf", "bsmt_unf_sf", "year_built", "o

corrplot(cor(x_best), method = "circle", type = "full")
```




```
# Check numerically
round(rcorr(x_best)$r, 2) %>%
  as.data.frame()
```

```
##                           gr_liv_area total_bsmt_sf bsmt_unf_sf year_built
## gr_liv_area                      1.00          0.40        0.26       0.20
## total_bsmt_sf                    0.40          1.00        0.41       0.37
## bsmt_unf_sf                      0.26          0.41        1.00       0.07
## year_built                       0.20          0.37        0.07       1.00
## overall_qualExcellent            0.19          0.30        0.08       0.20
## overall_qualGood                 0.22          0.03        0.09       0.26
## overall_qualVery_Excellent       0.23          0.21        0.06       0.05
## overall_qualVery_Good            0.30          0.31        0.16       0.30
## lot_area                         0.25          0.22        0.03       0.00
##                           overall_qualExcellent overall_qualGood
## gr_liv_area                                0.19             0.22
## total_bsmt_sf                              0.30             0.03
## bsmt_unf_sf                                0.08             0.09
## year_built                                 0.20             0.26
## overall_qualExcellent                      1.00            -0.09
## overall_qualGood                          -0.09             1.00
## overall_qualVery_Excellent                -0.02            -0.05
## overall_qualVery_Good                     -0.06            -0.18
## lot_area                                   0.05             0.01
##                           overall_qualVery_Excellent overall_qualVery_Good
## gr_liv_area                                     0.23                  0.30
## total_bsmt_sf                                   0.21                  0.31
## bsmt_unf_sf                                     0.06                  0.16
## year_built                                      0.05                  0.30
## overall_qualExcellent                          -0.02                 -0.06
## overall_qualGood                               -0.05                 -0.18
## overall_qualVery_Excellent                      1.00                 -0.03
```

```
## overall_qualVery_Good                          -0.03                  1.00
## lot_area                                        0.04                  0.02
##                             lot_area
## gr_liv_area                     0.25
## total_bsmt_sf                   0.22
## bsmt_unf_sf                     0.03
## year_built                      0.00
## overall_qualExcellent           0.05
## overall_qualGood                0.01
## overall_qualVery_Excellent      0.04
## overall_qualVery_Good           0.02
## lot_area                        1.00
```

We no longer have any major collinearities.

Linear regression modeling

We can train our model on our training data (including all predictors) using cross-validation). Here are two possible computational approaches.

Using `fit_lm` First, let's fit our model using `fit_lm`.

```
set.seed(2132)

# Set cross-validation parameters
control_cv = trainControl(method = "repeatedcv", number = 20, repeats = 5)

# Fit linear model on training data using cross-validation
fit_lm = train(sale_price ~ .,
               preProcess = "scale",
               data = training_data,
               method = "lm",
               trControl = control_cv)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
# Coefficients of final model
fit_lm$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
##                (Intercept)               gr_liv_area
##                  -4984627.9                   11918.1
##               first_flr_sf              second_flr_sf
##                    15645.6                   17658.1
##              total_bsmt_sf             low_qual_fin_sf
##                    14564.2                        NA
##               wood_deck_sf              open_porch_sf
##                     1609.2                    1027.2
##                bsmt_unf_sf               mas_vnr_area
##                    -8661.3                    1756.9
##                garage_cars               garage_area
##                     3056.3                    1566.1
##                 year_built             tot_rms_abv_grd
##                     9546.4                   -5883.7
##                  full_bath         overall_qualAverage
##                    -2344.0                   -2287.2
##  overall_qualBelow_Average       overall_qualExcellent
##                    -3314.3                   12221.9
##           overall_qualFair            overall_qualGood
##                    -1367.7                    4994.2
## overall_qualVery_Excellent       overall_qualVery_Good
##                    12336.0                   11604.6
##           kitchen_qualFair            kitchen_qualGood
##                    -3410.1                   -9158.7
##        kitchen_qualTypical                  fireplaces
##                   -13332.5                    7400.0
##           fireplace_quFair            fireplace_quGood
##                    -1199.0                     258.9
##   fireplace_quNo_Fireplace            fireplace_quPoor
##                     1697.4                    -677.4
##        fireplace_quTypical              exter_qualFair
##                    -2624.3                   -3914.4
##             exter_qualGood           exter_qualTypical
##                    -9346.0                  -11719.8
```

```
##               lot_frontage                   lot_area
##                     3328.0                     5015.9
##                  longitude                   latitude
##                     -923.3                     1071.9
##                   misc_val                  year_sold
##                      541.4                     -832.0
```

Then, we apply the model to the test data and obtain the MSE as a measure of accuracy.

```
predict_lm = predict(fit_lm, newdata = test_data)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
rmse_lm = RMSE(predict_lm, test_data$sale_price)

rmse_lm
```

```
## [1] 21149.18
```

Using `caret` Notably, we obtain the same linear model coefficients using the `caret` package:

```
set.seed(2132)

# Re-do linear model using glmnet
fit_lm_caret = train(x = training_predictors_matrix,
                     y = training_outcomes_vector,
                     method = "lm",
                     preProcess = c("center", "scale"),
                     trControl = control_cv)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
round(coef(fit_lm_caret$finalModel), 1)
```

```
##                (Intercept)               gr_liv_area
##                   177568.5                   11918.1
##               first_flr_sf              second_flr_sf
##                    15645.6                   17658.1
##              total_bsmt_sf             low_qual_fin_sf
##                    14564.2                        NA
##               wood_deck_sf              open_porch_sf
##                     1609.2                    1027.2
##                bsmt_unf_sf               mas_vnr_area
##                    -8661.3                    1756.9
##                garage_cars               garage_area
##                     3056.3                    1566.1
##                 year_built            tot_rms_abv_grd
##                     9546.4                   -5883.7
##                  full_bath        overall_qualAverage
##                    -2344.0                   -2287.2
##  overall_qualBelow_Average      overall_qualExcellent
##                    -3314.3                   12221.9
##           overall_qualFair           overall_qualGood
##                    -1367.7                    4994.2
## overall_qualVery_Excellent      overall_qualVery_Good
##                    12336.0                   11604.6
##           kitchen_qualFair           kitchen_qualGood
##                    -3410.1                   -9158.7
##        kitchen_qualTypical                 fireplaces
##                   -13332.5                    7400.0
##           fireplace_quFair           fireplace_quGood
##                    -1199.0                     258.9
##   fireplace_quNo_Fireplace           fireplace_quPoor
##                     1697.4                    -677.4
##        fireplace_quTypical             exter_qualFair
##                    -2624.3                   -3914.4
##             exter_qualGood          exter_qualTypical
##                    -9346.0                  -11719.8
##               lot_frontage                   lot_area
##                     3328.0                     5015.9
##                  longitude                   latitude
##                     -923.3                     1071.9
##                   misc_val                  year_sold
##                      541.4                     -832.0
```

And can measure our prediction as well by trying the model on our test data, finding the same RMSE as we

found using the first method.

```
predict_lm_caret = predict(fit_lm_caret, newdata = testing_matrix_all)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
rmse_lm_caret = RMSE(predict_lm_caret, test_data$sale_price)

rmse_lm_caret
```

```
## [1] 21149.18
```

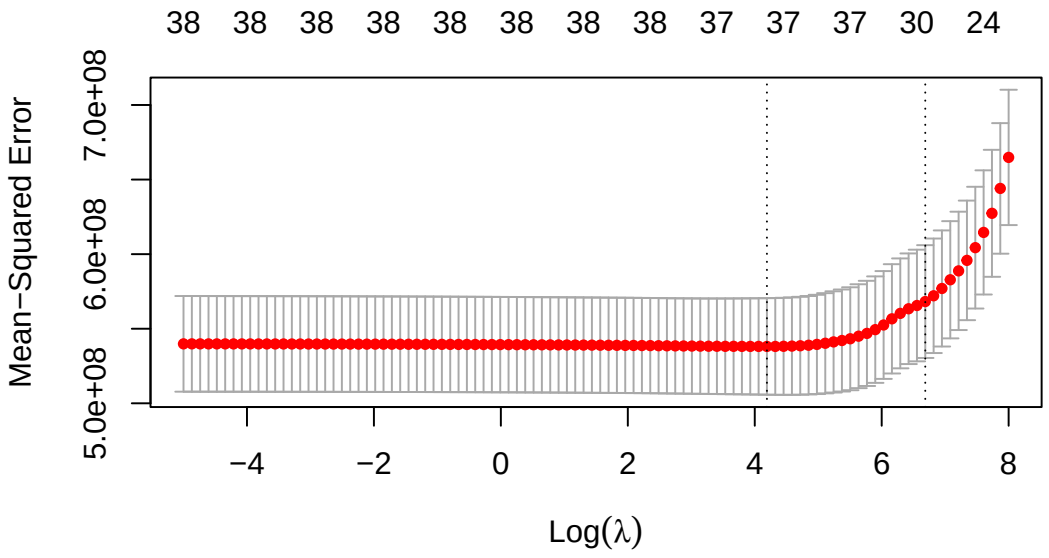
There are a number of potential disadvantages with the linear model. For instance, TO DO

Lasso model

The Lasso model utilizes an L1 penalty that, at times, forces some of the coefficient estimates to zero with sufficiently large lambda. This provides a useful variable selection process that may improve interpretability of the our model.

Using `glmnet`

```
cv_lasso_glmnet = cv.glmnet(x = training_predictors_matrix,
                      y = training_outcomes_vector,
                      standardize = TRUE,
                      alpha = 1,
                      lambda = exp(seq(8, -5, length = 100)))

plot(cv_lasso_glmnet)
```



```
# Use 1se rule
cv_lasso_glmnet$lambda.1se
```

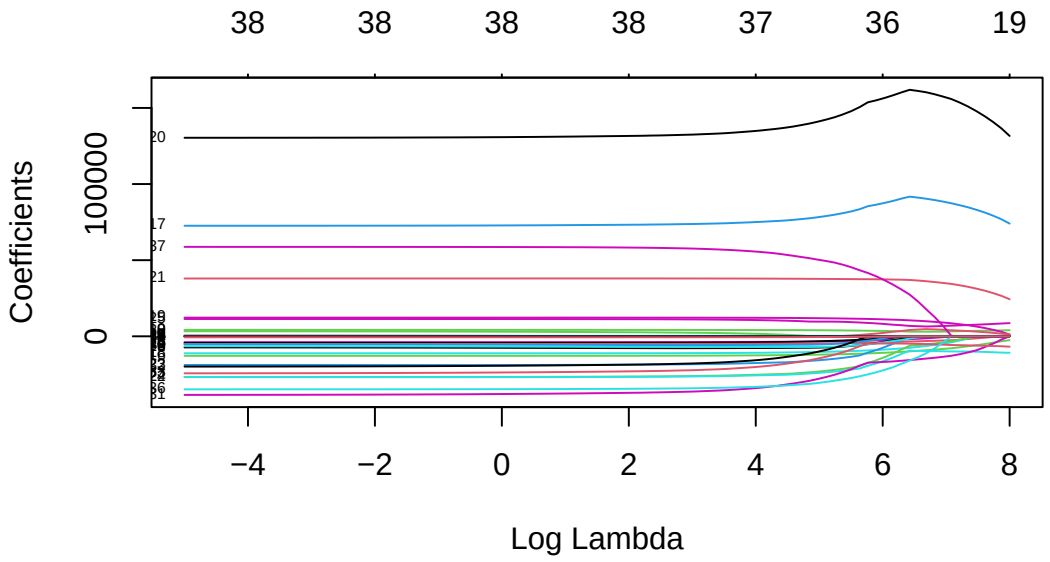
```
## [1] 801.8076
```

When the 1SE rule is applied, our lambda is 801.8. Our optimal coefficients are as follows:

```
# Obtain coefficients
lasso_glmnet_coefs = coef(cv_lasso_glmnet, s = "lambda.1se")

# Determine non-zero predictors (included in model)
num_pred = length(which(lasso_glmnet_coefs != 0))
```

Ultimately, we include 31 predictors in our model using this lasso method with 1SE rule for our lambda tuning parameter.

```
# Note: does this use lambda.1se?
# Note: this isn't standardized either
plot(cv_lasso_glmnet$glmnet.fit, "lambda", label = TRUE)
```



Finally, we perform prediction on our test data as follows, and obtain our RMSE.

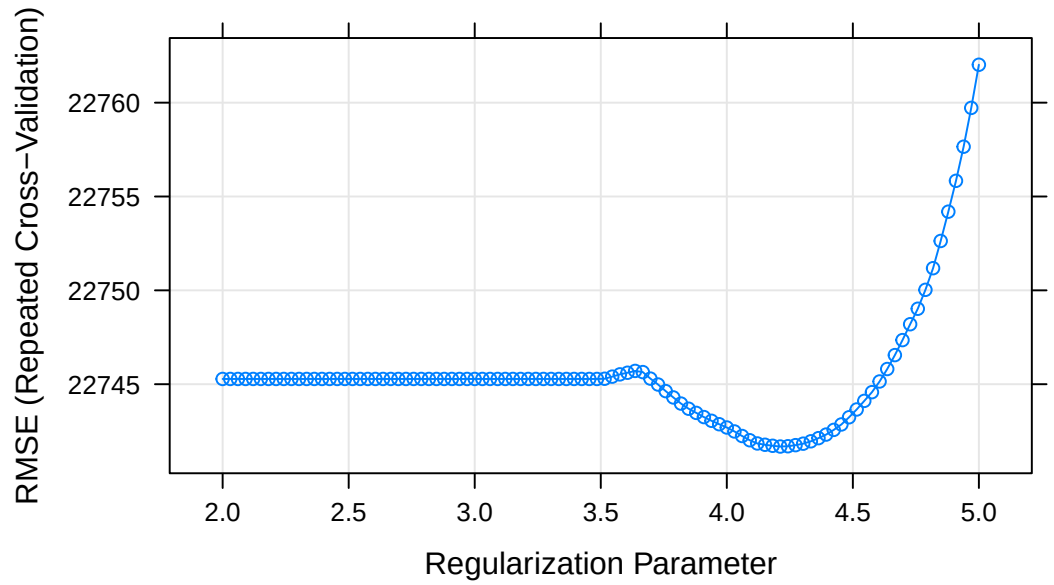
```
# Make predictions using glmnet object
lasso_predict = predict(cv_lasso_glmnet, newx = testing_predictors_matrix, s = "lambda.1se", type = "res

rmse_lasso_glmnet = RMSE(lasso_predict, test_data$sale_price)

rmse_lasso_glmnet
```

```
## [1] 20509.32
```

Using `caret`

Rather than using `glmnet`, we can also use the `caret` package.

```
set.seed(2132)

# Fit model on training data
lasso_caret_fit = train(x = training_predictors_matrix,
                        y = training_outcomes_vector,
                    method = "glmnet",
                    tuneGrid = expand.grid(alpha = 1,
                                           lambda = exp(seq(5, 2, length = 100))),
                    preProcess = c("center", "scale"),
                    trControl = control_cv
)

# Plot RMSE against tuning parameter
plot(lasso_caret_fit, xTrans = log)
```




```
# Note: how can we use lambda1se instead of just lambda?
# Optimal lambda
lasso_caret_fit$bestTune$lambda
```

```
## [1] 67.49957
```

Our optimal lambda is 67.499569.

We can also obtain our coefficients in the optimal model, and then make predictions using our test data to determine model performance (RMSE).

```r
# Note: how can we use lambda1se instead of just lambda?
# Obtain coefficients for final model
coef(lasso_caret_fit$finalModel, lasso_caret_fit$bestTune$lambda)
```

```
## 40 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                 177568.5021
## gr_liv_area                  31689.2792
## first_flr_sf                   295.7739
## second_flr_sf                    .
## total_bsmt_sf                14660.0703
## low_qual_fin_sf              -1813.4996
## wood_deck_sf                  1555.4681
## open_porch_sf                  978.2544
## bsmt_unf_sf                  -8668.3873
## mas_vnr_area                  1833.3855
## garage_cars                   2948.7469
## garage_area                   1648.6164
## year_built                    9492.0974
## tot_rms_abv_grd              -5533.4731
## full_bath                    -2066.0419
## overall_qualAverage          -2211.7939
## overall_qualBelow_Average    -3222.7537
## overall_qualExcellent        12720.8755
## overall_qualFair             -1317.5762
## overall_qualGood              4934.0200
## overall_qualVery_Excellent   12852.9289
## overall_qualVery_Good        11577.5069
## kitchen_qualFair             -3174.2015
## kitchen_qualGood             -8362.5423
## kitchen_qualTypical         -12587.0003
## fireplaces                    6840.4647
## fireplace_quFair             -1274.3705
## fireplace_quGood                 .
## fireplace_quNo_Fireplace      699.9602
## fireplace_quPoor             -736.1779
## fireplace_quTypical          -2875.4747
## exter_qualFair               -3372.5767
## exter_qualGood               -7017.9100
## exter_qualTypical            -9337.3081
## lot_frontage                  3235.5374
## lot_area                      5016.4029
## longitude                     -871.9000
## latitude                      1002.8658
## misc_val                       487.0711
## year_sold                     -718.8373
```

```r
set.seed(2132)

# Note: how can we use lambda1se instead of just lambda?
# Make predictions on test data set
lasso_pred_glmnet = predict(lasso_caret_fit, newdata = testing_predictors_matrix)
```

```
# Find RMSE
rmse_lasso_caret = RMSE(lasso_pred_glmnet, test_data$sale_price)
```

Elastic net

Elastic net is a more recent method that tends to be more effective when dealing with groups of highly correlated predictors. It includes two types of penalty, drawing on both the lasso and ridge methodologies, permitting us to create an optimal model using two tuning parameters. Here, we show its implementation using the `caret` package.

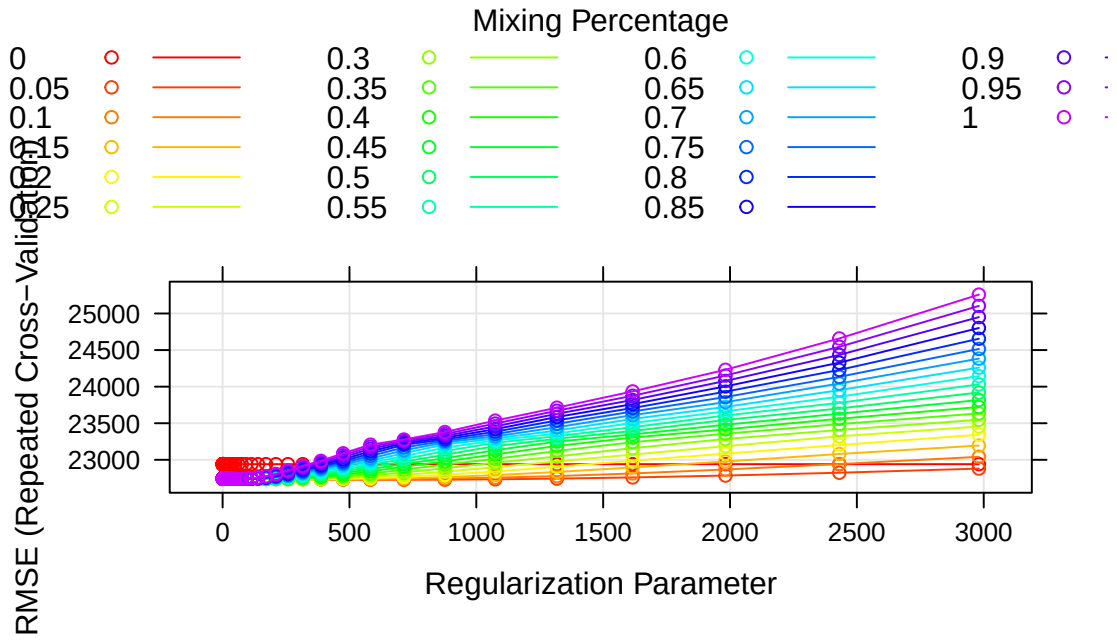
As always, we first train our model on standardized predictors:

```
set.seed(2132)

# Train model
enet_fit = train(x = training_predictors_matrix, y = training_outcomes_vector,
                 method = "glmnet",
                 tuneGrid = expand.grid(alpha = seq(0, 1, length = 21),
                                        lambda = exp(seq(-2, 8, length = 50))),
                 preProcess = c("center", "scale"),
                 trControl = control_cv
)

# Optimal tuning parameters
enet_fit$bestTune
```

```
##    alpha   lambda
## 93  0.05 714.3897
```

Using elastic net, we find that our optimal alpha is 0.05 and our optimal lambda is . This means that our ideal elastic net model is much closer to a ridge model than a lasso model. We can visualize as follows

```
# Rainbow plot settings
myCol = rainbow(25)
myPar = list(superpose.symbol = list(col = myCol),
             superpose.line = list(col = myCol))

# Plot RMSE against lambda, stratified by alpha
plot(enet_fit, par.settings = myPar)
```



And here are the coefficients for our optimal elastic net model:

```
# Coefficients for ideal elastic net model
coef(enet_fit$finalModel, enet_fit$bestTune$lambda)
```

```
## 40 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                 177568.50208
## gr_liv_area                  18703.61694
## first_flr_sf                  9788.71612
## second_flr_sf                10673.15283
## total_bsmt_sf                14436.78633
## low_qual_fin_sf               -706.40908
## wood_deck_sf                  1659.29223
## open_porch_sf                 1080.93987
## bsmt_unf_sf                  -8583.61167
## mas_vnr_area                  1998.43446
## garage_cars                   2905.14307
## garage_area                   1836.66264
## year_built                    9330.81279
## tot_rms_abv_grd              -5154.57597
## full_bath                    -1931.92056
## overall_qualAverage          -2341.95830
## overall_qualBelow_Average    -3283.47198
## overall_qualExcellent        12834.29027
## overall_qualFair             -1413.81617
## overall_qualGood              4846.46258
## overall_qualVery_Excellent   12982.97382
## overall_qualVery_Good        11471.34561
## kitchen_qualFair             -2964.06039
## kitchen_qualGood             -7613.96252
## kitchen_qualTypical         -11795.33734
```

```
## fireplaces                   6960.34493
## fireplace_quFair            -1315.80600
## fireplace_quGood              50.24277
## fireplace_quNo_Fireplace     778.59071
## fireplace_quPoor            -764.52935
## fireplace_quTypical         -2869.61043
## exter_qualFair              -3266.29824
## exter_qualGood              -6451.80364
## exter_qualTypical           -8862.54628
## lot_frontage                 3239.32678
## lot_area                     5002.48221
## longitude                    -935.91698
## latitude                     1050.65080
## misc_val                      505.72550
## year_sold                    -723.58661
```

Finally, we test our model on the test data set and determine our RMSE accuracy.

```
set.seed(2132)

# Elastic net predictions
enet_pred = predict(enet_fit, newdata = testing_predictors_matrix)

# Elastic net RMSE
rmse_elastic_net = RMSE(enet_pred, test_data$sale_price)

rmse_elastic_net
```

```
## [1] 20908.4
```

Partial least squares model

TO DO

Model comparison

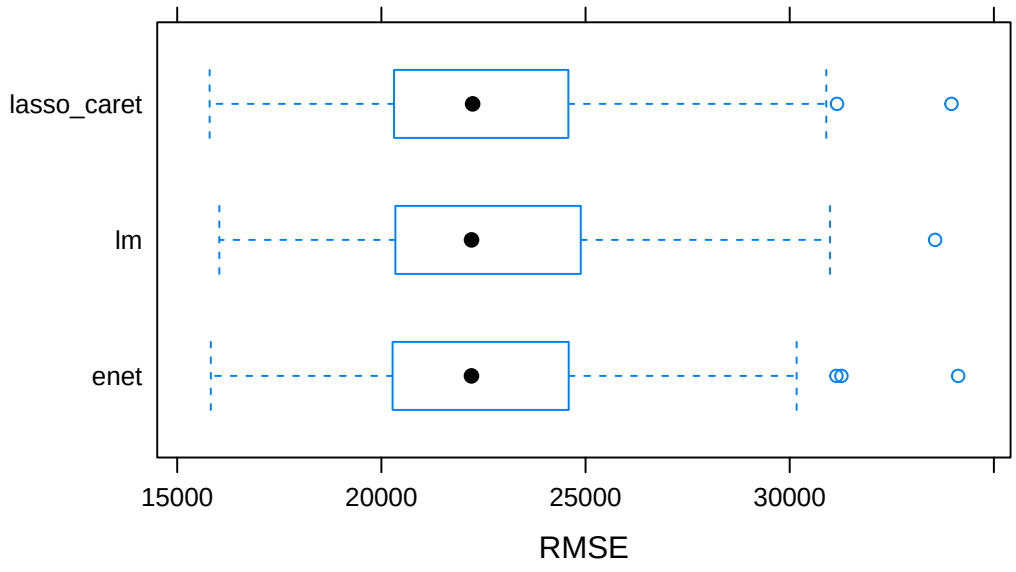
Finally, we use a resampling method to check how our final models perform on RMSE.

```
set.seed(2132)

# CHECK: Comparing different models based only on training error?
# TODO: includes only caret model for lasso, but we probably want to use glmnet for lambda1se...how do
# Resampling testing
resamp = resamples(list(lm = fit_lm_caret, lasso_caret = lasso_caret_fit, enet = enet_fit))

# Summary statistics on resampling
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
```

```
##
## Models: lm, lasso_caret, enet
## Number of resamples: 100
##
## MAE
##                   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm           12602.86 15462.72 16563.67 16697.00 17811.04 22257.38    0
## lasso_caret 12528.70 15396.12 16530.17 16636.65 17825.36 22275.44    0
## enet         12565.81 15359.97 16490.82 16597.01 17753.53 22294.78    0
##
## RMSE
##                   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm           16031.12 20344.04 22207.97 22777.76 24777.60 33561.15    0
## lasso_caret 15793.85 20310.87 22234.54 22741.68 24467.85 33960.61    0
## enet         15823.23 20279.61 22206.06 22727.14 24488.44 34125.78    0
##
## Rsquared
##                    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm           0.8322394 0.8924542 0.9061546 0.9050941 0.9222373 0.9532063    0
## lasso_caret 0.8326279 0.8925997 0.9059775 0.9054194 0.9219974 0.9543052    0
## enet         0.8310350 0.8928202 0.9058368 0.9055924 0.9220324 0.9545525    0
```

```
# Box plots of resampling for RMSE
bwplot(resamp, metric = "RMSE")
```



Based on this, we would likely use the elastic net model to predict our response because it minimizes the mean RMSE over resamples.