In-Class 8

```r
# load packages
library(tidyverse)
```

Load data

The data we will be working with contains information about the housing market in Ames, TX.

We need to load the following RData file, which contains a training and testing dataset.

```r
# read in training and testing data
load('regression-data.RData')

# preview training data
glimpse(data_train)
```

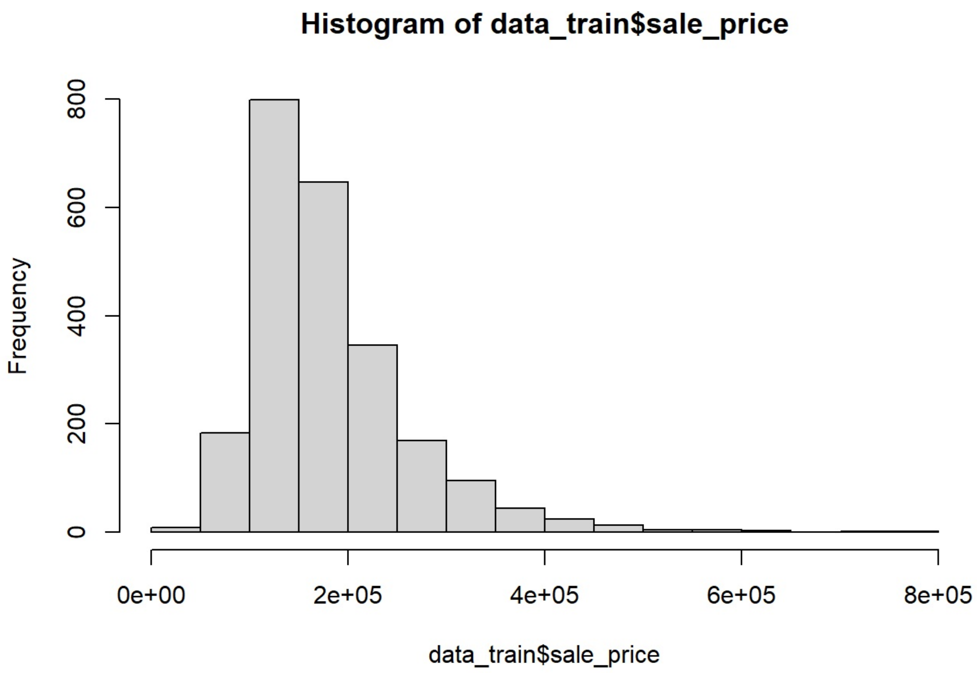
```
Rows: 2,344
Columns: 61
$ lot_frontage    <dbl> 46, 60, 0, 78, 74, 43, 0, 0, 80, 80, 100, 60, 21, 75, …
$ lot_area        <int> 20544, 7200, 9555, 15600, 11988, 3182, 10464, 4426, 92…
$ year_built      <int> 1986, 1949, 1979, 1949, 1934, 2005, 1980, 2004, 1965, …
$ year_remod_add  <int> 1991, 1950, 1979, 2005, 1995, 2006, 1980, 2004, 1965, …
$ mas_vnr_area    <dbl> 123, 0, 0, 0, 0, 16, 130, 169, 0, 252, 0, 0, 0, 0, 0, …
$ bsmt_fin_sf_1   <dbl> 7, 5, 5, 2, 4, 3, 3, 3, 6, 1, 3, 7, 3, 3, 1, 7, 7, 7, …
$ bsmt_unf_sf     <dbl> 791, 0, 0, 248, 389, 1357, 138, 186, 244, 467, 172, 85…
$ total_bsmt_sf   <dbl> 791, 0, 0, 1067, 715, 1373, 988, 848, 1136, 1165, 924,…
$ first_flr_sf    <int> 1236, 1040, 1100, 986, 849, 1555, 1102, 848, 1136, 116…
$ second_flr_sf   <int> 857, 0, 1133, 537, 811, 0, 0, 0, 0, 896, 0, 0, 546, 14…
$ gr_liv_area     <int> 2093, 1040, 2233, 1523, 1660, 1555, 1102, 848, 1136, 2…
$ bsmt_full_bath  <dbl> 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, …
$ full_bath       <int> 2, 2, 2, 2, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 2, 2, 2, …
$ half_bath       <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, …
$ bedroom_abv_gr  <int> 3, 2, 5, 3, 3, 2, 2, 1, 3, 4, 2, 2, 3, 4, 4, 4, 3, 3, …
$ kitchen_abv_gr  <int> 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ tot_rms_abv_grd <int> 7, 6, 11, 7, 6, 7, 5, 3, 5, 8, 6, 5, 5, 11, 8, 8, 10, …
$ fireplaces      <int> 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, …
$ garage_cars     <dbl> 2, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, 0, 2, 2, 2, 3, 3, …
$ garage_area     <dbl> 542, 420, 579, 295, 240, 430, 582, 420, 384, 498, 528,…
$ wood_deck_sf    <int> 364, 0, 0, 0, 0, 143, 140, 160, 426, 0, 0, 0, 200, 208…
$ open_porch_sf   <int> 63, 0, 0, 0, 0, 20, 22, 0, 0, 77, 36, 0, 26, 364, 207,…
$ enclosed_porch  <int> 0, 0, 0, 81, 0, 0, 0, 0, 0, 0, 0, 116, 0, 0, 0, 0, 0, …
$ screen_porch    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 196, 0, 0, 0, 0, 224, 0, 0,…
$ sale_price      <int> 215000, 90000, 141000, 158000, 188700, 192500, 169000,…
$ longitude       <dbl> -93.63915, -93.60890, -93.67433, -93.64076, -93.64141,…
$ latitude        <dbl> 42.05602, 42.03584, 42.01917, 42.01494, 42.01844, 42.0…
$ ms_zoning       <fct> Residential_Low_Density, Residential_Low_Density, Resi…
$ street          <fct> Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, …
$ alley           <fct> No_Alley_Access, No_Alley_Access, No_Alley_Access, No_…
$ lot_shape       <fct> Slightly_Irregular, Regular, Slightly_Irregular, Regul…
$ land_contour    <fct> Lvl, Lvl, Lvl, Bnk, HLS, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl,…
$ lot_config      <fct> CulDSac, Inside, CulDSac, Inside, Inside, Inside, FR3,…
$ land_slope      <fct> Gtl, Gtl, Gtl, Gtl, Mod, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl,…
$ condition_1     <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, …
$ condition_2     <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, …
$ bldg_type       <fct> OneFam, Duplex, Duplex, OneFam, OneFam, TwnhsE, OneFam…
$ house_style     <fct> Two_Story, One_Story, Two_Story, One_and_Half_Fin, Two…
$ overall_cond    <fct> Above_Average, Average, Above_Average, Good, Good, Ave…
$ roof_style      <fct> Gable, Gable, Gable, Gable, Hip, Gable, Gable, Gable, …
$ roof_matl       <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg, …
$ mas_vnr_type    <fct> BrkFace, None, None, None, None, BrkFace, BrkFace, Brk…
$ exter_cond      <fct> Good, Typical, Typical, Typical, Typical, Typical, Typ…
$ foundation      <fct> CBlock, Slab, Slab, BrkTil, CBlock, PConc, CBlock, PCo…
$ bsmt_cond       <fct> Typical, No_Basement, No_Basement, Typical, Typical, T…
$ bsmt_exposure   <fct> No, No_Basement, No_Basement, No, No, Av, Av, Av, No, …
$ bsmt_fin_type_1 <fct> Unf, No_Basement, No_Basement, BLQ, LwQ, GLQ, GLQ, GLQ…
$ bsmt_fin_type_2 <fct> Unf, No_Basement, No_Basement, Rec, Unf, Unf, Unf, Unf…
$ heating         <fct> GasA, Wall, GasA, GasW, GasA, GasA, GasA, GasA, GasA, …
$ heating_qc      <fct> Good, Fair, Typical, Fair, Fair, Excellent, Typical, E…
$ central_air     <fct> Y, N, Y, N, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, …
$ electrical      <fct> SBrkr, FuseF, SBrkr, SBrkr, FuseA, SBrkr, SBrkr, SBrkr…
$ functional      <fct> Typ, Typ, Typ, Maj2, Typ, Typ, Typ, Typ, Typ, Typ, Typ…
```

```
$ garage_type     <fct> Attchd, Detchd, Attchd, Attchd, Detchd, Attchd, Attchd…
$ garage_finish   <fct> Fin, Unf, Fin, Unf, Unf, Fin, RFn, RFn, RFn, RFn, Unf,…
$ garage_cond     <fct> Typical, Typical, Good, Typical, Typical, Typical, Typ…
$ paved_drive     <fct> Paved, Paved, Paved, Paved, Paved, Paved, Paved, Paved…
$ pool_qc         <fct> No_Pool, No_Pool, No_Pool, No_Pool, No_Pool, No_Pool, …
$ fence           <fct> Minimum_Privacy, No_Fence, No_Fence, No_Fence, No_Fenc…
$ misc_feature    <fct> None, None, None, None, None, None, None, None, None, …
$ sale_condition  <fct> Normal, Normal, Normal, Normal, Normal, Normal, Normal…
```

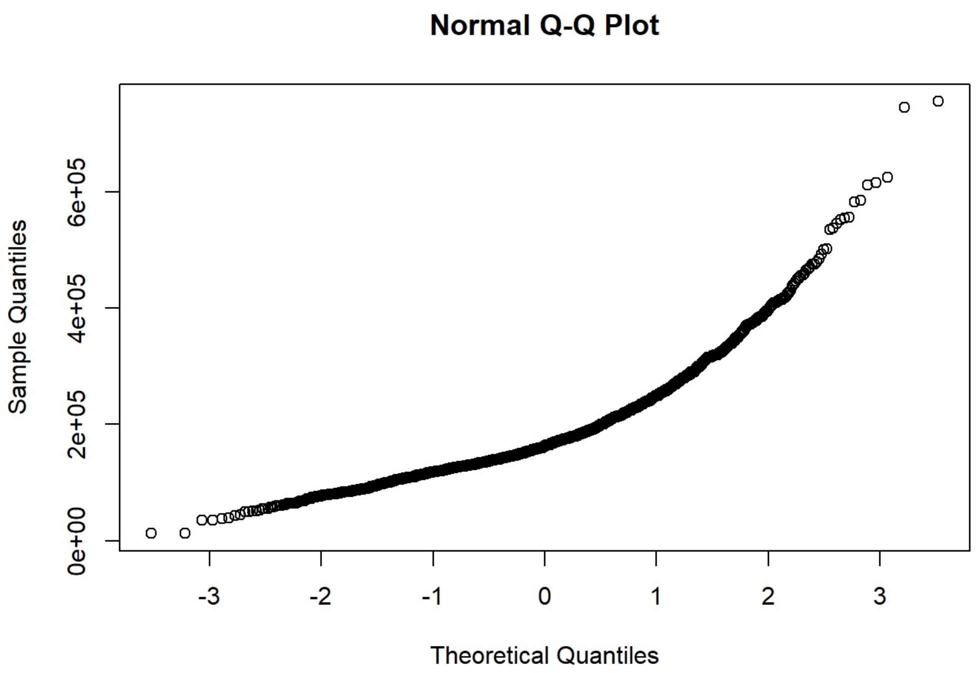
Intial EDA

The model we want to create is: `sale_price ~ < set of Xs >`. Let's first inspect the response variable.

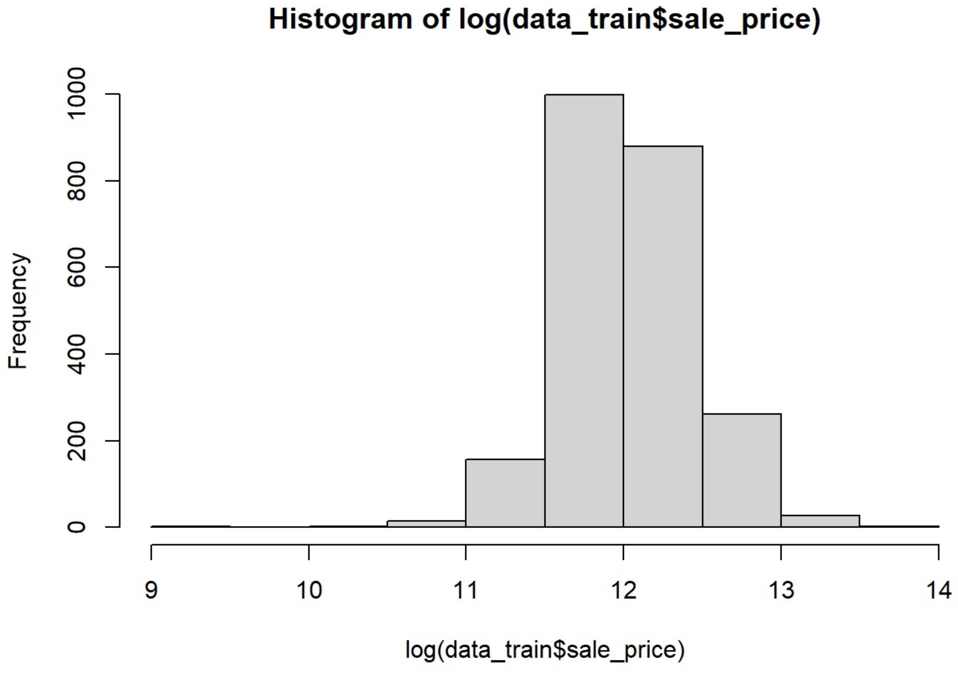
```
# plot response variable
# -> seems skewed right and not normal
hist(data_train$sale_price)
```



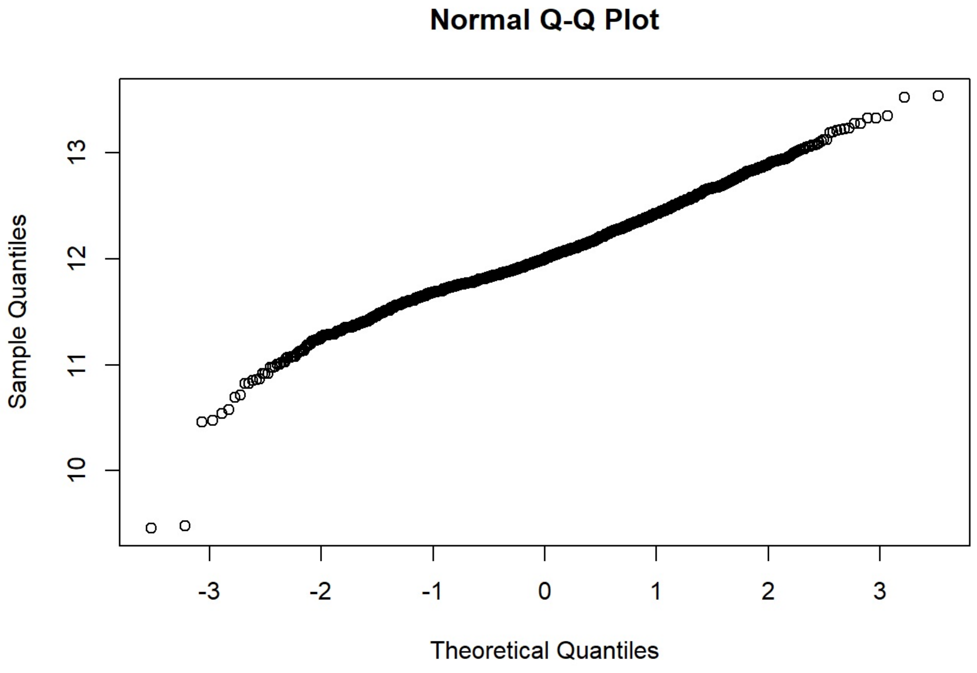

```
qqnorm(data_train$sale_price)
```



```
# -> try log transformation and recheck
hist(log(data_train$sale_price))
```




```
qqnorm(log(data_train$sale_price))
```



The distribution of `sale_price` appears to be much more normal after transforming. So let's use `log(sale_price) ~ < set of Xs >` as our model.

This is technically a regression with a transformed response, so it is a generalized linear model (GLM). Once we make the transformation, everything works as usual, only the units and interpretations change.

Variable selection

Now try to find some \(X\) variables to use in a model.

```
summary(data_train)
```

```
  lot_frontage       lot_area        year_built    year_remod_add
 Min.   :  0.00   Min.   :  1476   Min.   :1872   Min.   :1950
 1st Qu.: 43.00   1st Qu.:  7420   1st Qu.:1954   1st Qu.:1966
```

```
Median : 63.00   Median :  9431   Median :1974   Median :1993
Mean   : 57.79   Mean   : 10229   Mean   :1972   Mean   :1985
3rd Qu.: 78.00   3rd Qu.: 11590   3rd Qu.:2001   3rd Qu.:2004
Max.   :313.00   Max.   :215245   Max.   :2010   Max.   :2010

 mas_vnr_area    bsmt_fin_sf_1    bsmt_unf_sf     total_bsmt_sf
Min.   :   0.0   Min.   :0.000   Min.   :   0.0   Min.   :   0
1st Qu.:   0.0   1st Qu.:3.000   1st Qu.: 216.0   1st Qu.: 796
Median :   0.0   Median :3.000   Median : 455.5   Median : 992
Mean   : 100.8   Mean   :4.124   Mean   : 551.3   Mean   :1055
3rd Qu.: 163.2   3rd Qu.:7.000   3rd Qu.: 785.0   3rd Qu.:1302
Max.   :1600.0   Max.   :7.000   Max.   :2336.0   Max.   :6110

 first_flr_sf  second_flr_sf     gr_liv_area   bsmt_full_bath    full_bath
Min.   : 334   Min.   :   0.0   Min.   : 334   Min.   :0.000   Min.   :0.000
1st Qu.: 876   1st Qu.:   0.0   1st Qu.:1126   1st Qu.:0.000   1st Qu.:1.000
Median :1090   Median :   0.0   Median :1448   Median :0.000   Median :2.000
Mean   :1162   Mean   : 337.3   Mean   :1504   Mean   :0.439   Mean   :1.575
3rd Qu.:1390   3rd Qu.: 713.2   3rd Qu.:1749   3rd Qu.:1.000   3rd Qu.:2.000
Max.   :5095   Max.   :2065.0   Max.   :5642   Max.   :3.000   Max.   :4.000

  half_bath      bedroom_abv_gr  kitchen_abv_gr  tot_rms_abv_grd
Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   : 2.000
1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 5.000
Median :0.0000   Median :3.000   Median :1.000   Median : 6.000
Mean   :0.3891   Mean   :2.844   Mean   :1.045   Mean   : 6.433
3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.: 7.000
Max.   :2.0000   Max.   :6.000   Max.   :3.000   Max.   :15.000

  fireplaces      garage_cars     garage_area      wood_deck_sf
Min.   :0.0000   Min.   :0.000   Min.   :   0.0   Min.   :   0.00
1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 323.8   1st Qu.:   0.00
Median :1.0000   Median :2.000   Median : 480.0   Median :   0.00
Mean   :0.6139   Mean   :1.775   Mean   : 475.0   Mean   :  97.34
3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.: 172.00
Max.   :4.0000   Max.   :4.000   Max.   :1488.0   Max.   :1424.00

open_porch_sf    enclosed_porch     screen_porch      sale_price
Min.   :  0.00   Min.   :   0.00   Min.   :  0.00   Min.   : 12789
1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:130000
Median : 28.00   Median :   0.00   Median :  0.00   Median :162500
Mean   : 47.29   Mean   :  22.84   Mean   : 16.25   Mean   :182270
3rd Qu.: 70.00   3rd Qu.:   0.00   3rd Qu.:  0.00   3rd Qu.:215000
Max.   :742.00   Max.   :1012.00   Max.   :576.00   Max.   :755000

  longitude         latitude                            ms_zoning
Min.   :-93.69   Min.   :41.99   Floating_Village_Residential: 122
1st Qu.:-93.66   1st Qu.:42.02   Residential_High_Density    :  23
Median :-93.64   Median :42.03   Residential_Low_Density     :1826
Mean   :-93.64   Mean   :42.03   Residential_Medium_Density  : 351
3rd Qu.:-93.62   3rd Qu.:42.05   A_agr                       :   2
Max.   :-93.58   Max.   :42.06   C_all                       :  19
                                 I_all                       :   1
 street                 alley                       lot_shape    land_contour
Grvl:  10   Gravel         :  87   Regular             :1467   Bnk:  99
Pave:2334   No_Alley_Access:2189   Slightly_Irregular  : 804   HLS:  97
            Paved          :  68   Moderately_Irregular:  60   Low:  51
                                   Irregular           :  13   Lvl:2097



  lot_config   land_slope  condition_1    condition_2      bldg_type
Corner : 403   Gtl:2226   Norm   :2035   Norm   :2320   OneFam  :1936
CulDSac: 141   Mod: 103   Feedr  : 121   Feedr  :  10   TwoFmCon:  47
FR2    :  73   Sev:  15   Artery :  71   Artery :   4   Duplex  :  88
FR3    :  11              RRAn   :  41   PosN   :   4   Twnhs   :  85
Inside :1716              PosN   :  29   PosA   :   2   TwnhsE  : 188
                          RRAe   :  20   RRNn   :   2
                          (Other):  27   (Other):   2
          house_style          overall_cond    roof_style     roof_matl
One_Story       :1193   Average      :1332   Flat   :  16   CompShg:2311
Two_Story       : 705   Above_Average: 434   Gable  :1855   Tar&Grv:  18
One_and_Half_Fin: 239   Good         : 306   Gambrel:  19   WdShake:   6
SLvl            : 102   Very_Good    : 110   Hip    : 440   WdShngl:   6
SFoyer          :  66   Below_Average:  82   Mansard:   9   ClyTile:   1
```

```
 Two_and_Half_Unf:  18   Fair         :  37   Shed   :   5   Membran:   1  
 (Other)         :  21   (Other)      :  43                  (Other):   1  
  mas_vnr_type      exter_cond    foundation          bsmt_cond   
 BrkCmn :  17   Excellent:   9   BrkTil: 244   Excellent  :   2  
 BrkFace: 707   Fair     :  49   CBlock: 982   Fair       :  76  
 CBlock :   1   Good     : 238   PConc :1068   Good       :  94  
 None   :1416   Poor     :   3   Slab  :  36   No_Basement:  63  
 Stone  : 203   Typical  :2045   Stone :  10   Poor       :   2  
                                 Wood  :   4   Typical    :2107  
                                                                 
     bsmt_exposure     bsmt_fin_type_1    bsmt_fin_type_2  heating    
 Av         : 338   ALQ        :349    ALQ        :  48   Floor:   1  
 Gd         : 239   BLQ        :215    BLQ        :  56   GasA :2308  
 Mn         : 191   GLQ        :706    GLQ        :  31   GasW :  19  
 No         :1510   LwQ        :132    LwQ        :  75   Grav :   9  
 No_Basement:  66   No_Basement: 63    No_Basement:  64   OthW :   2  
                    Rec        :222    Rec        :  85   Wall :   5  
                    Unf        :657    Unf        :1985                
     heating_qc   central_air   electrical     functional  
 Excellent:1223   N: 147      FuseA  : 143   Typ    :2188  
 Fair     :  78   Y:2197      FuseF  :  34   Min2   :  52  
 Good     : 371               FuseP  :   8   Min1   :  51  
 Poor     :   3               Mix    :   0   Mod    :  29  
 Typical  : 669               SBrkr  :2158   Maj1   :  16  
                              Unknown:   1   Maj2   :   6  
                                             (Other):   2  
             garage_type     garage_finish    garage_cond  
 Attchd            :1398   Fin       :584   Excellent:   3  
 Basment           :  28   No_Garage:123   Fair     :  59  
 BuiltIn           : 148   RFn       :671   Good     :  10  
 CarPort           :  14   Unf       :966   No_Garage: 123  
 Detchd            : 614                    Poor     :   9  
 More_Than_Two_Types:  20                   Typical  :2140  
 No_Garage         : 122                                    
          paved_drive        pool_qc                   fence      misc_feature
 Dirt_Gravel     : 160   Excellent:   3   Good_Privacy     : 101   Elev:   1   
 Partial_Pavement:  45   Fair     :   2   Good_Wood        :  80   Gar2:   3   
 Paved           :2139   Good     :   4   Minimum_Privacy  : 248   None:2259   
                         No_Pool  :2332   Minimum_Wood_Wire:  10   Othr:   3   
                         Typical  :   3   No_Fence         :1905   Shed:  77   
                                                                   TenC:   1   
                                                                               
 sale_condition
 Abnorml: 149  
 AdjLand:   7  
 Alloca :  21  
 Family :  33  
 Normal :1930  
 Partial: 204  
```

```r
mod_data_train = lm(log(sale_price) ~ foundation+sale_condition, data=data_train)
summary(mod_data_train)
```

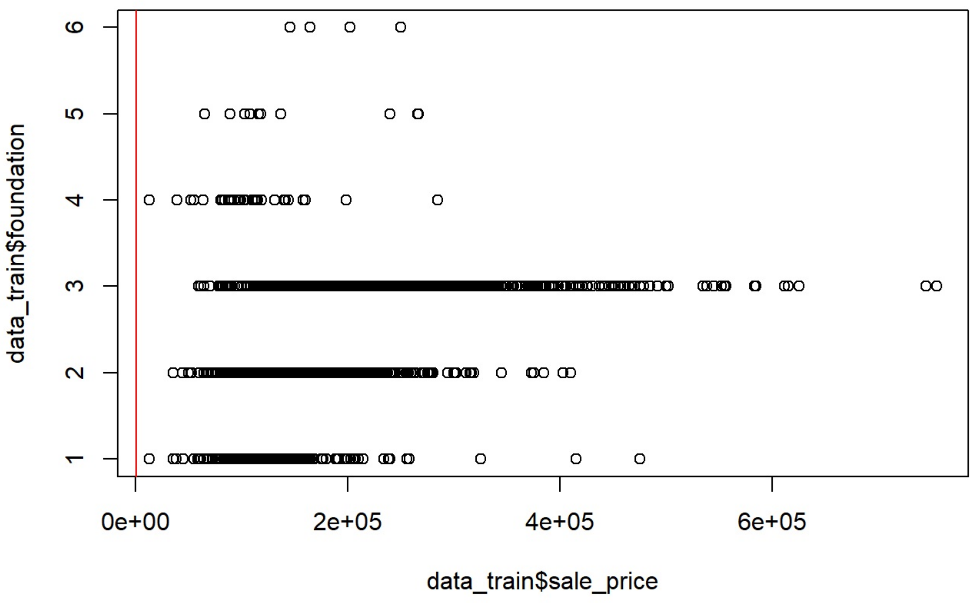
```

Call:
lm(formula = log(sale_price) ~ foundation + sale_condition, data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.07275 -0.17803 -0.01073  0.18459  1.46126 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           11.52909    0.03285 350.957  < 2e-16 ***
foundationCBlock       0.17966    0.02337   7.687 2.20e-14 ***
foundationPConc        0.53079    0.02363  22.458  < 2e-16 ***
foundationSlab        -0.19778    0.05880  -3.363 0.000782 ***
foundationStone        0.12406    0.10538   1.177 0.239233    
foundationWood         0.42690    0.16465   2.593 0.009580 ** 
sale_conditionAdjLand -0.19456    0.12633  -1.540 0.123680    
sale_conditionAlloca   0.22281    0.07677   2.902 0.003739 ** 
sale_conditionFamily   0.05781    0.06290   0.919 0.358183    
sale_conditionNormal   0.17920    0.02788   6.427 1.57e-10 ***
```

```
sale_conditionPartial  0.40264    0.03665  10.987  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3266 on 2333 degrees of freedom
Multiple R-squared:  0.3612,	Adjusted R-squared:  0.3585 
F-statistic: 131.9 on 10 and 2333 DF,  p-value: < 2.2e-16
```

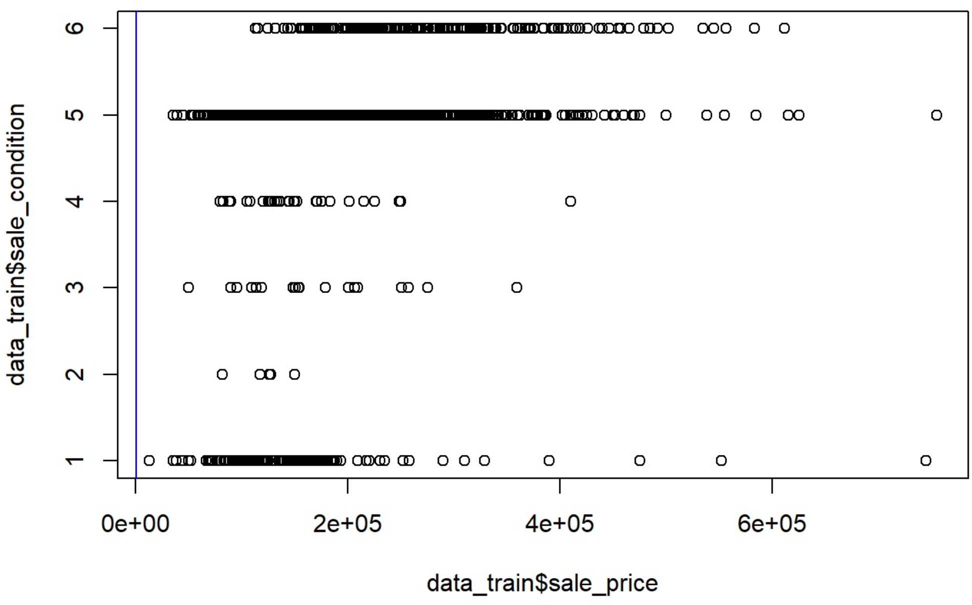
```r
mod_data_train1 = lm(log(sale_price) ~ condition_1+condition_2+functional, data=data_train)
summary(mod_data_train1)
```

```
Call:
lm(formula = log(sale_price) ~ condition_1 + condition_2 + functional, 
    data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.60537 -0.24868 -0.01815  0.23305  1.47277 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      11.362018   0.222698  51.020  < 2e-16 ***
condition_1Feedr  0.109348   0.059431   1.840  0.06591 .  
condition_1Norm   0.317814   0.047977   6.624 4.32e-11 ***
condition_1PosA   0.532824   0.116012   4.593 4.61e-06 ***
condition_1PosN   0.493073   0.091462   5.391 7.71e-08 ***
condition_1RRAe   0.080438   0.099635   0.807  0.41956    
condition_1RRAn   0.378638   0.079548   4.760 2.06e-06 ***
condition_1RRNe   0.207146   0.201393   1.029  0.30379    
condition_1RRNn   0.450728   0.140453   3.209  0.00135 ** 
condition_2Feedr -0.068045   0.236776  -0.287  0.77385    
condition_2Norm   0.185620   0.200316   0.927  0.35421    
condition_2PosA   1.035131   0.345007   3.000  0.00273 ** 
condition_2PosN   0.639372   0.290835   2.198  0.02802 *  
condition_2RRAe   0.487157   0.440494   1.106  0.26887    
condition_2RRAn   0.159421   0.440494   0.362  0.71745    
condition_2RRNn  -0.232293   0.342611  -0.678  0.49783    
functionalMaj2   -0.453092   0.187505  -2.416  0.01575 *  
functionalMin1    0.029872   0.112282   0.266  0.79023    
functionalMin2   -0.002344   0.112133  -0.021  0.98332    
functionalMod     0.016439   0.122117   0.135  0.89293    
functionalSal    -1.556472   0.294425  -5.286 1.36e-07 ***
functionalTyp     0.196256   0.098330   1.996  0.04606 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3915 on 2322 degrees of freedom
Multiple R-squared:  0.08617,	Adjusted R-squared:  0.07791 
F-statistic: 10.43 on 21 and 2322 DF,  p-value: < 2.2e-16
```

```r
plot(x=data_train$sale_price, y=data_train$foundation)
abline(mod_data_train, col="red")
```




```r
plot(x=data_train$sale_price, y=data_train$sale_condition)
abline(mod_data_train, col="blue")
```




```r
coef(mod_data_train)
```

```
         (Intercept)      foundationCBlock       foundationPConc
         11.52908998            0.17965980            0.53078682
      foundationSlab       foundationStone        foundationWood
         -0.19777956            0.12405636            0.42690341
sale_conditionAdjLand  sale_conditionAlloca  sale_conditionFamily
         -0.19456005            0.22281232            0.05781038
 sale_conditionNormal sale_conditionPartial
          0.17920468            0.40264278
```

```r
coef(mod_data_train1)
```

```
     (Intercept) condition_1Feedr  condition_1Norm  condition_1PosA
     11.36201795       0.10934840       0.31781403       0.53282389
 condition_1PosN  condition_1RRAe  condition_1RRAn  condition_1RRNe
```

```
      0.49307337       0.08043815       0.37863812       0.20714648
 condition_1RRNn condition_2Feedr  condition_2Norm  condition_2PosA
      0.45072798      -0.06804493       0.18561953       1.03513105
 condition_2PosN  condition_2RRAe  condition_2RRAn  condition_2RRNn
      0.63937237       0.48715736       0.15942055      -0.23229297
  functionalMaj2   functionalMin1   functionalMin2    functionalMod
     -0.45309152       0.02987241      -0.00234413       0.01643905
   functionalSal    functionalTyp
     -1.55647159       0.19625564
```

Prediction

Once you have some candidate models, see which one is the best by calculating the \(RMSE\).

```
area <- lm(log(sale_price)~foundation, data=data_train)
log_area <- lm(log(sale_price)~sale_condition, data=data_train)

yardstick::rmse_vec(truth=log(data_test$sale_price), estimate= predict(log_area, newdata=data_test))
```

```
[1] 0.3692361
```

```
yardstick::rmse_vec(truth=log(data_test$sale_price), estimate= predict(area, newdata=data_test))
```

```
[1] 0.3374639
```