R for Data Science Project

Adding Packages - tidyverse

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Background

As a statistical consultant working for a real estate investment firm, your task is to develop a model to predict the selling price of a given home in Ames, Iowa. Your employer hopes to use this information to help assess whether the asking price of a house is higher or lower than the true value of the house. If the home is undervalued, it may be a good investment for the firm.

Training Data and relevant packages

In order to better assess the quality of the model you will produce, the data have been randomly divided into three separate pieces: a training data set, a testing data set, and a validation data set. For now we will load the training data set, the others will be loaded and used later.

```
load("ames_train.Rdata")
```

```
ames_train
```

```
## # A tibble: 1,000 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int>       <int> <fct>            <int>    <int> <fct>  <fct>
##  1  9.09e8   856 126000          30 RL                  NA     7890 Pave   <NA>
##  2  9.05e8  1049 139500         120 RL                  42     4235 Pave   <NA>
##  3  9.11e8  1001 124900          30 C (all)             60     6060 Pave   <NA>
##  4  5.35e8  1039 114000          70 RL                  80     8146 Pave   <NA>
##  5  5.34e8  1665 227000          60 RL                  70     8400 Pave   <NA>
##  6  9.08e8  1922 198500          85 RL                  64     7301 Pave   <NA>
##  7  9.02e8   936  93000          20 RM                  60     6000 Pave   Pave
##  8  5.28e8  1246 187687          20 RL                  53     3710 Pave   <NA>
##  9  9.23e8   889 137500          20 RL                  74    12395 Pave   <NA>
## 10  9.08e8  1072 140000         180 RM                  35     3675 Pave   <NA>
## # ... with 990 more rows, and 72 more variables: Lot.Shape <fct>,
```

```
## #   Land.Contour <fct>, Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>,
## #   Neighborhood <fct>, Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>,
## #   House.Style <fct>, Overall.Qual <int>, Overall.Cond <int>,
## #   Year.Built <int>, Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

```
?load
```

Use the code block below to load any necessary packages

```
library(statsr)
library(dplyr)
library(BAS)
library(tidyverse)
library(MASS)
```

Look at data

```
ames_train
```

```
## # A tibble: 1,000 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int>       <int> <fct>            <int>    <int> <fct>  <fct>
##  1  9.09e8   856 126000          30 RL                  NA     7890 Pave   <NA>
##  2  9.05e8  1049 139500         120 RL                  42     4235 Pave   <NA>
##  3  9.11e8  1001 124900          30 C (all)             60     6060 Pave   <NA>
##  4  5.35e8  1039 114000          70 RL                  80     8146 Pave   <NA>
##  5  5.34e8  1665 227000          60 RL                  70     8400 Pave   <NA>
##  6  9.08e8  1922 198500          85 RL                  64     7301 Pave   <NA>
##  7  9.02e8   936  93000          20 RM                  60     6000 Pave   Pave
##  8  5.28e8  1246 187687          20 RL                  53     3710 Pave   <NA>
##  9  9.23e8   889 137500          20 RL                  74    12395 Pave   <NA>
## 10  9.08e8  1072 140000         180 RM                  35     3675 Pave   <NA>
## # ... with 990 more rows, and 72 more variables: Lot.Shape <fct>,
## #   Land.Contour <fct>, Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>,
## #   Neighborhood <fct>, Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>,
## #   House.Style <fct>, Overall.Qual <int>, Overall.Cond <int>,
## #   Year.Built <int>, Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
```

```
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

Look at all the columns of the dataset

```
names(ames_train)
```

```
##  [1] "PID"             "area"            "price"           "MS.SubClass"
##  [5] "MS.Zoning"       "Lot.Frontage"    "Lot.Area"        "Street"
##  [9] "Alley"           "Lot.Shape"       "Land.Contour"    "Utilities"
## [13] "Lot.Config"      "Land.Slope"      "Neighborhood"    "Condition.1"
## [17] "Condition.2"     "Bldg.Type"       "House.Style"     "Overall.Qual"
## [21] "Overall.Cond"    "Year.Built"      "Year.Remod.Add"  "Roof.Style"
## [25] "Roof.Matl"       "Exterior.1st"    "Exterior.2nd"    "Mas.Vnr.Type"
## [29] "Mas.Vnr.Area"    "Exter.Qual"      "Exter.Cond"      "Foundation"
## [33] "Bsmt.Qual"       "Bsmt.Cond"       "Bsmt.Exposure"   "BsmtFin.Type.1"
## [37] "BsmtFin.SF.1"    "BsmtFin.Type.2"  "BsmtFin.SF.2"    "Bsmt.Unf.SF"
## [41] "Total.Bsmt.SF"   "Heating"         "Heating.QC"      "Central.Air"
## [45] "Electrical"      "X1st.Flr.SF"     "X2nd.Flr.SF"     "Low.Qual.Fin.SF"
## [49] "Bsmt.Full.Bath"  "Bsmt.Half.Bath"  "Full.Bath"       "Half.Bath"
## [53] "Bedroom.AbvGr"   "Kitchen.AbvGr"   "Kitchen.Qual"    "TotRms.AbvGrd"
## [57] "Functional"      "Fireplaces"      "Fireplace.Qu"    "Garage.Type"
## [61] "Garage.Yr.Blt"   "Garage.Finish"   "Garage.Cars"     "Garage.Area"
## [65] "Garage.Qual"     "Garage.Cond"     "Paved.Drive"     "Wood.Deck.SF"
## [69] "Open.Porch.SF"   "Enclosed.Porch"  "X3Ssn.Porch"     "Screen.Porch"
## [73] "Pool.Area"       "Pool.QC"         "Fence"           "Misc.Feature"
## [77] "Misc.Val"        "Mo.Sold"         "Yr.Sold"         "Sale.Type"
## [81] "Sale.Condition"
```

Find all the variables that has a word "price" in it

```
dplyr::select(ames_train, contains("price"))
```

```
## # A tibble: 1,000 x 1
##     price
##     <int>
##  1 126000
##  2 139500
##  3 124900
##  4 114000
##  5 227000
##  6 198500
##  7  93000
```

```
##  8 187687
##  9 137500
## 10 140000
## # ... with 990 more rows
```

```
head(ames_train, n = 6)
```

```
## # A tibble: 6 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int>       <int> <fct>            <int>    <int> <fct>  <fct>
## 1   9.09e8   856 126000          30 RL                  NA     7890 Pave   <NA>
## 2   9.05e8  1049 139500         120 RL                  42     4235 Pave   <NA>
## 3   9.11e8  1001 124900          30 C (all)             60     6060 Pave   <NA>
## 4   5.35e8  1039 114000          70 RL                  80     8146 Pave   <NA>
## 5   5.34e8  1665 227000          60 RL                  70     8400 Pave   <NA>
## 6   9.08e8  1922 198500          85 RL                  64     7301 Pave   <NA>
## # ... with 72 more variables: Lot.Shape <fct>, Land.Contour <fct>,
## #   Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>, Neighborhood <fct>,
## #   Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>, House.Style <fct>,
## #   Overall.Qual <int>, Overall.Cond <int>, Year.Built <int>,
## #   Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

```
tail(ames_train, n = 5)
```

```
## # A tibble: 5 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int>       <int> <fct>            <int>    <int> <fct>  <fct>
## 1   9.07e8   848 145000         120 RM                  NA     4426 Pave   <NA>
## 2   5.28e8  1576 197000          60 FV                  65     8125 Pave   <NA>
## 3   5.34e8  1728  84900          90 RL                  98    13260 Pave   <NA>
## 4   9.05e8  1352 158000          60 RL                  80     9364 Pave   <NA>
## 5   9.14e8   912 156000          85 RL                  NA     7540 Pave   <NA>
## # ... with 72 more variables: Lot.Shape <fct>, Land.Contour <fct>,
## #   Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>, Neighborhood <fct>,
## #   Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>, House.Style <fct>,
## #   Overall.Qual <int>, Overall.Cond <int>, Year.Built <int>,
## #   Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
```

```
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

```
nrow(ames_train)
```

```
## [1] 1000
```

```
ncol(ames_train)
```

```
## [1] 81
```

Part 1 - Exploratory Data Analysis (EDA)

When you first get your data, it's very tempting to immediately begin fitting models and assessing how they perform. However, before you begin modeling, it's absolutely essential to explore the structure of the data and the relationships between the variables in the data set.

Do a detailed EDA of the ames_train data set, to learn about the structure of the data and the relationships between the variables in the data set (refer to Introduction to Probability and Data, Week 2, for a reminder about EDA if needed). Your EDA should involve creating and reviewing many plots/graphs and considering the patterns and relationships you see.

After you have explored completely, submit the three graphs/plots that you found most informative during your EDA process, and briefly explain what you learned from each (why you found each informative).

```
summary(ames_train)
```

```
##       PID                 area          price          MS.SubClass
##  Min.   :5.263e+08   Min.   : 334   Min.   : 12789   Min.   : 20.00
##  1st Qu.:5.285e+08   1st Qu.:1092   1st Qu.:129762   1st Qu.: 20.00
##  Median :5.354e+08   Median :1411   Median :159467   Median : 50.00
##  Mean   :7.059e+08   Mean   :1477   Mean   :181190   Mean   : 57.15
##  3rd Qu.:9.071e+08   3rd Qu.:1743   3rd Qu.:213000   3rd Qu.: 70.00
##  Max.   :1.007e+09   Max.   :4676   Max.   :615000   Max.   :190.00
##
##    MS.Zoning    Lot.Frontage       Lot.Area       Street     Alley
##  A (agr):  0   Min.   : 21.00   Min.   :  1470   Grvl:  3   Grvl: 33
##  C (all):  9   1st Qu.: 57.00   1st Qu.:  7314   Pave:997   Pave: 34
##  FV     : 56   Median : 69.00   Median :  9317              NA's:933
##  I (all):  1   Mean   : 69.21   Mean   : 10352
##  RH     :  7   3rd Qu.: 80.00   3rd Qu.: 11650
```

```
##  RL     :772   Max.   :313.00   Max.   :215245
##  RM     :155   NA's   :167
##  Lot.Shape Land.Contour  Utilities      Lot.Config  Land.Slope  Neighborhood
##  IR1:338   Bnk: 33      AllPub:1000   Corner :173   Gtl:962    NAmes  :155
##  IR2: 30   HLS: 38      NoSeWa:   0   CulDSac: 76   Mod: 33    CollgCr: 85
##  IR3:  3   Low: 20      NoSewr:   0   FR2    : 36   Sev:  5    Somerst: 74
##  Reg:629   Lvl:909                    FR3    :  5              OldTown: 71
##                                       Inside :710              Sawyer : 61
##                                                                Edwards: 60
##                                                                (Other):494
##   Condition.1   Condition.2   Bldg.Type    House.Style   Overall.Qual
##  Norm   :875   Norm   :988   1Fam  :823   1Story :521   Min.   : 1.000
##  Feedr  : 53   Feedr  :  6   2fmCon: 20   2Story :286   1st Qu.: 5.000
##  Artery : 23   Artery :  2   Duplex: 35   1.5Fin : 98   Median : 6.000
##  RRAn   : 14   PosN   :  2   Twnhs : 38   SLvl   : 41   Mean   : 6.095
##  PosN   : 11   PosA   :  1   TwnhsE: 84   SFoyer : 36   3rd Qu.: 7.000
##  RRAe   : 11   RRNn   :  1                2.5Unf : 10   Max.   :10.000
##  (Other): 13   (Other):  0                (Other):  8
##   Overall.Cond     Year.Built   Year.Remod.Add   Roof.Style    Roof.Matl
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  9   CompShg:984
##  1st Qu.:5.000   1st Qu.:1955   1st Qu.:1966   Gable  :775   Tar&Grv: 11
##  Median :5.000   Median :1975   Median :1992   Gambrel:  8   WdShake:  2
##  Mean   :5.559   Mean   :1972   Mean   :1984   Hip    :204   WdShngl:  2
##  3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   Mansard:  4   Metal  :  1
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :  0   ClyTile:  0
##                                                             (Other):  0
##   Exterior.1st  Exterior.2nd  Mas.Vnr.Type  Mas.Vnr.Area    Exter.Qual
##  VinylSd:349   VinylSd:345          :  7   Min.   :   0.0   Ex: 39
##  HdBoard:164   HdBoard:150   BrkCmn :  8   1st Qu.:   0.0   Fa: 11
##  MetalSd:147   MetalSd:148   BrkFace:317   Median :   0.0   Gd:337
##  Wd Sdng:138   Wd Sdng:130   CBlock :  0   Mean   : 104.1   TA:613
##  Plywood: 74   Plywood: 96   None   :593   3rd Qu.: 160.0
##  CemntBd: 40   CmentBd: 40   Stone  : 75   Max.   :1290.0
##  (Other): 88   (Other): 91                 NA's   :7
##  Exter.Cond  Foundation  Bsmt.Qual  Bsmt.Cond  Bsmt.Exposure BsmtFin.Type.1
##  Ex:  4     BrkTil:102       :  1       :  1       :  2      GLQ    :294
##  Fa: 19     CBlock:430   Ex  : 87   Ex  :  2   Av  :157      Unf    :279
##  Gd:116     PConc :453   Fa  : 28   Fa  : 23   Gd  : 98      ALQ    :163
##  Po:  0     Slab  : 12   Gd  :424   Gd  : 44   Mn  : 87      Rec    :107
##  TA:861     Stone :  3   Po  :  1   Po  :  1   No  :635      BLQ    : 87
##             Wood  :  0   TA  :438   TA  :908   NA's: 21      (Other): 49
##                          NA's: 21   NA's: 21                 NA's   : 21
##   BsmtFin.SF.1    BsmtFin.Type.2  BsmtFin.SF.2      Bsmt.Unf.SF
##  Min.   :   0.0   Unf    :863    Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   0.0   LwQ    : 31    1st Qu.:   0.00   1st Qu.: 223.5
##  Median : 400.0   Rec    : 29    Median :   0.00   Median : 461.0
##  Mean   : 464.1   BLQ    : 24    Mean   :  48.07   Mean   : 547.0
##  3rd Qu.: 773.0   ALQ    : 20    3rd Qu.:   0.00   3rd Qu.: 783.0
##  Max.   :2260.0   (Other): 12    Max.   :1526.00   Max.   :2336.0
##  NA's   :1        NA's   : 21    NA's   :1         NA's   :1
##  Total.Bsmt.SF     Heating    Heating.QC Central.Air Electrical
##  Min.   :   0.0   Floor:  0   Ex:516     N: 55            :  0
##  1st Qu.: 797.5   GasA :988   Fa: 22     Y:945       FuseA: 54
##  Median : 998.0   GasW :  8   Gd:157                 FuseF: 12
```

```
##  Mean   :1059.2   Grav :  2   Po:  1                 FuseP:  2
##  3rd Qu.:1301.0   OthW :  1   TA:304                 Mix  :  0
##  Max.   :3138.0   Wall :  1                          SBrkr:932
##  NA's   :1
##   X1st.Flr.SF      X2nd.Flr.SF     Low.Qual.Fin.SF   Bsmt.Full.Bath
##  Min.   : 334.0   Min.   :   0.0   Min.   :   0.00   Min.   :0.0000
##  1st Qu.: 876.2   1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.:0.0000
##  Median :1080.5   Median :   0.0   Median :   0.00   Median :0.0000
##  Mean   :1157.1   Mean   : 315.2   Mean   :   4.32   Mean   :0.4474
##  3rd Qu.:1376.2   3rd Qu.: 688.2   3rd Qu.:   0.00   3rd Qu.:1.0000
##  Max.   :3138.0   Max.   :1836.0   Max.   :1064.00   Max.   :3.0000
##                                                      NA's   :1
##  Bsmt.Half.Bath      Full.Bath       Half.Bath     Bedroom.AbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.000   Median :3.000
##  Mean   :0.06106   Mean   :1.541   Mean   :0.378   Mean   :2.806
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :4.000   Max.   :2.000   Max.   :6.000
##  NA's   :1
##  Kitchen.AbvGr   Kitchen.Qual TotRms.AbvGrd     Functional    Fireplaces
##  Min.   :0.000   Ex: 67       Min.   : 2.00   Typ    :935   Min.   :0.000
##  1st Qu.:1.000   Fa: 20       1st Qu.: 5.00   Min2   : 24   1st Qu.:0.000
##  Median :1.000   Gd:403       Median : 6.00   Min1   : 18   Median :1.000
##  Mean   :1.039   Po:  1       Mean   : 6.34   Mod    : 16   Mean   :0.597
##  3rd Qu.:1.000   TA:509       3rd Qu.: 7.00   Maj1   :  4   3rd Qu.:1.000
##  Max.   :2.000                Max.   :13.00   Maj2   :  2   Max.   :4.000
##                                               (Other):  1
##  Fireplace.Qu  Garage.Type  Garage.Yr.Blt  Garage.Finish  Garage.Cars
##  Ex  : 16     2Types : 10   Min.   :1900       :  2      Min.   :0.000
##  Fa  : 24     Attchd :610   1st Qu.:1961   Fin :247      1st Qu.:1.000
##  Gd  :232     Basment: 11   Median :1979   RFn :278      Median :2.000
##  Po  : 18     BuiltIn: 56   Mean   :1978   Unf :427      Mean   :1.767
##  TA  :219     CarPort:  1   3rd Qu.:2002   NA's: 46      3rd Qu.:2.000
##  NA's:491     Detchd :266   Max.   :2010                 Max.   :5.000
##               NA's   : 46   NA's   :48                   NA's   :1
##   Garage.Area     Garage.Qual Garage.Cond Paved.Drive  Wood.Deck.SF
##  Min.   :   0.0       :  1        :  1    N: 67       Min.   :  0.00
##  1st Qu.: 312.0   Ex  :  1    Ex  :  1    P: 29       1st Qu.:  0.00
##  Median : 480.0   Fa  : 37    Fa  : 21    Y:904       Median :  0.00
##  Mean   : 475.4   Gd  :  7    Gd  :  6                Mean   : 93.84
##  3rd Qu.: 576.0   Po  :  3    Po  :  6                3rd Qu.:168.00
##  Max.   :1390.0   TA  :904    TA  :918                Max.   :857.00
##  NA's   :1        NA's: 47    NA's: 47
##  Open.Porch.SF    Enclosed.Porch    X3Ssn.Porch       Screen.Porch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
##  Median : 28.00   Median :  0.00   Median :  0.000   Median :  0.00
##  Mean   : 48.93   Mean   : 23.48   Mean   :  3.118   Mean   : 14.77
##  3rd Qu.: 74.00   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.00
##  Max.   :742.00   Max.   :432.00   Max.   :508.000   Max.   :440.00
##
##    Pool.Area       Pool.QC      Fence     Misc.Feature    Misc.Val
##  Min.   :  0.000   Ex  :  1   GdPrv: 43   Elev:  0     Min.   :    0.00
```

```
##  1st Qu.:  0.000   Fa  :  1   GdWo : 37   Gar2:  2     1st Qu.:    0.00
##  Median :  0.000   Gd  :  1   MnPrv:120   Othr:  1     Median :    0.00
##  Mean   :  1.463   TA  :  0   MnWw :  2   Shed: 25     Mean   :   45.81
##  3rd Qu.:  0.000   NA's:997   NA's :798   TenC:  1     3rd Qu.:    0.00
##  Max.   :800.000                          NA's:971     Max.   :15500.00
##
##     Mo.Sold          Yr.Sold       Sale.Type   Sale.Condition
##  Min.   : 1.000   Min.   :2006   WD     :863   Abnorml: 61
##  1st Qu.: 4.000   1st Qu.:2007   New    : 79   AdjLand:  2
##  Median : 6.000   Median :2008   COD    : 27   Alloca :  4
##  Mean   : 6.243   Mean   :2008   ConLD  :  7   Family : 17
##  3rd Qu.: 8.000   3rd Qu.:2009   ConLw  :  6   Normal :834
##  Max.   :12.000   Max.   :2010   Con    :  5   Partial: 82
##                                  (Other): 13
```

```
ames_train %>% dplyr::select(Pool.QC, Fence, Misc.Feature) %>% filter(!is.na(Misc.Feature))
```

```
## # A tibble: 29 x 3
##    Pool.QC Fence Misc.Feature
##    <fct>   <fct> <fct>
##  1 <NA>    <NA>  Shed
##  2 <NA>    <NA>  Shed
##  3 <NA>    MnPrv Othr
##  4 <NA>    <NA>  Shed
##  5 <NA>    <NA>  Gar2
##  6 <NA>    MnPrv Shed
##  7 <NA>    <NA>  Shed
##  8 <NA>    MnPrv Shed
##  9 <NA>    MnPrv Gar2
## 10 <NA>    MnPrv Shed
## # ... with 19 more rows
```

Let's first clean the data.

The categorical variables which are encoded as type int have to be converted to factors first.

- MS.SubClass
- Overall.Qual
- Overall.Cond

```
str(ames_train$MS.SubClass)
```

```
##  int [1:1000] 30 120 30 70 60 85 20 20 20 180 ...
```

```
str(ames_train$Overall.Cond)
```

```
##  int [1:1000] 6 5 9 8 6 5 4 5 6 5 ...
```

```
str(ames_train$Overall.Qual)
```

```
##  int [1:1000] 6 5 5 4 8 7 4 7 5 6 ...
```

```
count(ames_train, Overall.Qual)
```

```
## # A tibble: 10 x 2
##    Overall.Qual     n
##  *        <int> <int>
##  1            1     1
##  2            2     8
```

```
##  3            3     9
##  4            4    68
##  5            5   305
##  6            6   238
##  7            7   200
##  8            8   122
##  9            9    40
## 10           10     9
```

Convert the above three variables to factors:

```
(ames_train <- ames_train %>% mutate(MS.SubClass = as.factor(MS.SubClass), Overall.Qual = as.factor(Ove
```

```
## # A tibble: 1,000 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int> <fct>       <fct>            <int>    <int> <fct>  <fct>
##  1  9.09e8   856 126000 30          RL                  NA     7890 Pave   <NA>
##  2  9.05e8  1049 139500 120         RL                  42     4235 Pave   <NA>
##  3  9.11e8  1001 124900 30          C (all)             60     6060 Pave   <NA>
##  4  5.35e8  1039 114000 70          RL                  80     8146 Pave   <NA>
##  5  5.34e8  1665 227000 60          RL                  70     8400 Pave   <NA>
##  6  9.08e8  1922 198500 85          RL                  64     7301 Pave   <NA>
##  7  9.02e8   936  93000 20          RM                  60     6000 Pave   Pave
##  8  5.28e8  1246 187687 20          RL                  53     3710 Pave   <NA>
##  9  9.23e8   889 137500 20          RL                  74    12395 Pave   <NA>
## 10  9.08e8  1072 140000 180         RM                  35     3675 Pave   <NA>
## # ... with 990 more rows, and 72 more variables: Lot.Shape <fct>,
## #   Land.Contour <fct>, Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>,
## #   Neighborhood <fct>, Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>,
## #   House.Style <fct>, Overall.Qual <fct>, Overall.Cond <fct>,
## #   Year.Built <int>, Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

```
str(ames_train$MS.SubClass)
```

```
##  Factor w/ 15 levels "20","30","40",..: 2 12 2 7 6 10 1 1 1 14 ...
```

```
count(ames_train, MS.SubClass)
```

```
## # A tibble: 15 x 2
##    MS.SubClass     n
```

```
##  * <fct>       <int>
##  1 20            379
##  2 30             49
##  3 40              1
##  4 45              7
##  5 50             93
##  6 60            195
##  7 70             34
##  8 75              6
##  9 80             39
## 10 85             21
## 11 90             35
## 12 120            69
## 13 160            46
## 14 180             7
## 15 190            19
```

Transformation of NA's to a new category will avoid bias in the data and the modelling by removing data from the dataset.

Lot.Frontage variable is a continuous variable which has 167 NA's (missing data). Hence, we shall not transform Lot.Frontage variable.

But other variables such as , Alley, Bsmt.Qual, Bsmt.Cond, Bsmt.Exposure, BsmtFin.Type.1, BsmtFin.Type.2, Fireplace.Qu, Garage.Type, Garage.Finish,Garage.Qual,Garage.Cond,Pool.QC, Fence, Misc.Feature are categorical variables which has NA's that should be converted to a new category.

```
ames_train %>% count(Alley)
```

```
## # A tibble: 3 x 2
##   Alley     n
## * <fct> <int>
## 1 Grvl     33
## 2 Pave     34
## 3 <NA>    933
```

```
str(ames_train$Alley)
```

```
##  Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA 2 NA NA NA ...
```

```
ames_train <- ames_train %>% mutate(
  Alley = if_else(is.na(Alley), 'No Alley', as.character(Alley)),
  Bsmt.Qual = if_else(is.na(Bsmt.Qual), 'No Basement', as.character(Bsmt.Qual)),
  Bsmt.Cond = if_else(is.na(Bsmt.Cond), 'No Basement', as.character(Bsmt.Cond)),
  Bsmt.Exposure = if_else(is.na(Bsmt.Exposure), 'No Basement', as.character(Bsmt.Cond)),
  BsmtFin.Type.1 = if_else(is.na(BsmtFin.Type.1), 'No Basement', as.character(BsmtFin.Type.1)),
  BsmtFin.Type.2 = if_else(is.na(BsmtFin.Type.2), 'No Basement', as.character(BsmtFin.Type.2)),
  Fireplace.Qu = if_else(is.na(Fireplace.Qu), 'No Fireplace', as.character(Fireplace.Qu)),
  Garage.Type = if_else(is.na(Garage.Type), 'No Garage', as.character(Garage.Type)),
  Garage.Finish = if_else(is.na(Garage.Finish), 'No Garage', as.character(Garage.Finish)),
  Garage.Qual = if_else(is.na(Garage.Qual), 'No Garage', as.character(Garage.Qual)),
  Garage.Cond = if_else(is.na(Garage.Cond), 'No Garage', as.character(Garage.Cond)),
  Pool.QC = if_else(is.na(Pool.QC), 'No Pool', as.character(Pool.QC)),
  Fence = if_else(is.na(Fence), 'No Fence', as.character(Fence)),
  Misc.Feature = if_else(is.na(Misc.Feature), 'No MiscFeature', as.character(Misc.Feature))

)
```

```
count(ames_train,Alley)
```

```
## # A tibble: 3 x 2
##   Alley        n
## * <chr>    <int>
## 1 Grvl        33
## 2 No Alley   933
## 3 Pave        34
```

```
ames_train%>%count(Sale.Condition
                   )
```

```
## # A tibble: 6 x 2
##   Sale.Condition     n
## * <fct>          <int>
## 1 Abnorml           61
## 2 AdjLand            2
## 3 Alloca             4
## 4 Family            17
## 5 Normal           834
## 6 Partial           82
```

```
(ames_train <- ames_train %>% dplyr::filter(Sale.Condition=="Normal"))
```

```
## # A tibble: 834 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int> <fct>       <fct>            <int>    <int> <fct>  <chr>
##  1  9.09e8   856 126000 30          RL                  NA     7890 Pave   No A~
##  2  9.05e8  1049 139500 120         RL                  42     4235 Pave   No A~
##  3  9.11e8  1001 124900 30          C (all)             60     6060 Pave   No A~
##  4  5.35e8  1039 114000 70          RL                  80     8146 Pave   No A~
##  5  5.34e8  1665 227000 60          RL                  70     8400 Pave   No A~
##  6  9.08e8  1922 198500 85          RL                  64     7301 Pave   No A~
##  7  9.02e8   936  93000 20          RM                  60     6000 Pave   Pave
##  8  9.23e8   889 137500 20          RL                  74    12395 Pave   No A~
##  9  9.08e8  1072 140000 180         RM                  35     3675 Pave   No A~
## 10  5.28e8  1342 219500 120         RL                  48     6240 Pave   No A~
## # ... with 824 more rows, and 72 more variables: Lot.Shape <fct>,
## #   Land.Contour <fct>, Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>,
## #   Neighborhood <fct>, Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>,
## #   House.Style <fct>, Overall.Qual <fct>, Overall.Cond <fct>,
## #   Year.Built <int>, Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <chr>, Bsmt.Cond <chr>, Bsmt.Exposure <chr>,
## #   BsmtFin.Type.1 <chr>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <chr>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <chr>, Garage.Type <chr>, Garage.Yr.Blt <int>,
## #   Garage.Finish <chr>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <chr>, Garage.Cond <chr>, Paved.Drive <fct>,
```

```
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <chr>,
## #   Fence <chr>, Misc.Feature <chr>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

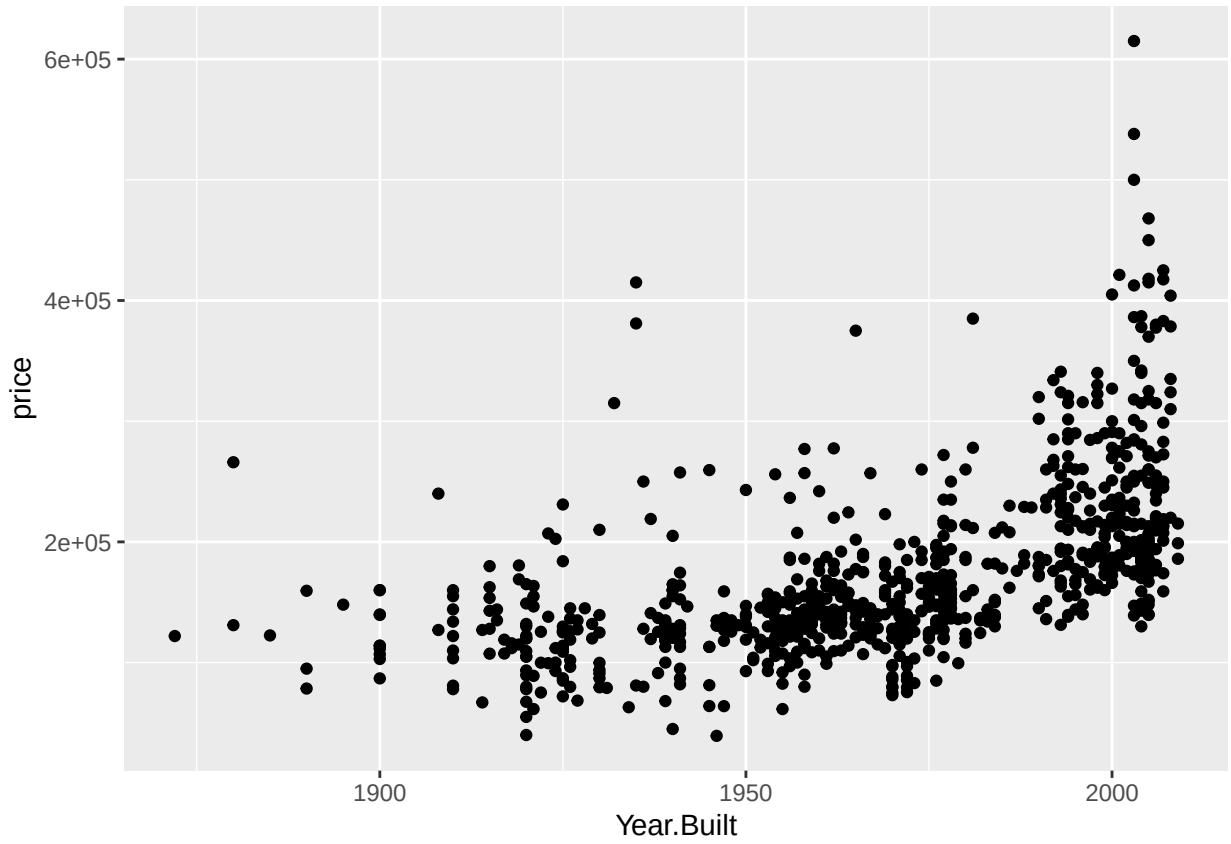
```
ames_train
```

```
## # A tibble: 834 x 81
##        PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##      <int> <int>  <int> <fct>       <fct>            <int>    <int> <fct>  <chr>
##  1  9.09e8   856 126000 30          RL                  NA     7890 Pave   No A~
##  2  9.05e8  1049 139500 120         RL                  42     4235 Pave   No A~
##  3  9.11e8  1001 124900 30          C (all)             60     6060 Pave   No A~
##  4  5.35e8  1039 114000 70          RL                  80     8146 Pave   No A~
##  5  5.34e8  1665 227000 60          RL                  70     8400 Pave   No A~
##  6  9.08e8  1922 198500 85          RL                  64     7301 Pave   No A~
##  7  9.02e8   936  93000 20          RM                  60     6000 Pave   Pave
##  8  9.23e8   889 137500 20          RL                  74    12395 Pave   No A~
##  9  9.08e8  1072 140000 180         RM                  35     3675 Pave   No A~
## 10  5.28e8  1342 219500 120         RL                  48     6240 Pave   No A~
## # ... with 824 more rows, and 72 more variables: Lot.Shape <fct>,
## #   Land.Contour <fct>, Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>,
## #   Neighborhood <fct>, Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>,
## #   House.Style <fct>, Overall.Qual <fct>, Overall.Cond <fct>,
## #   Year.Built <int>, Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <chr>, Bsmt.Cond <chr>, Bsmt.Exposure <chr>,
## #   BsmtFin.Type.1 <chr>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <chr>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <chr>, Garage.Type <chr>, Garage.Yr.Blt <int>,
## #   Garage.Finish <chr>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <chr>, Garage.Cond <chr>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <chr>,
## #   Fence <chr>, Misc.Feature <chr>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

```
nrow(ames_train)
```

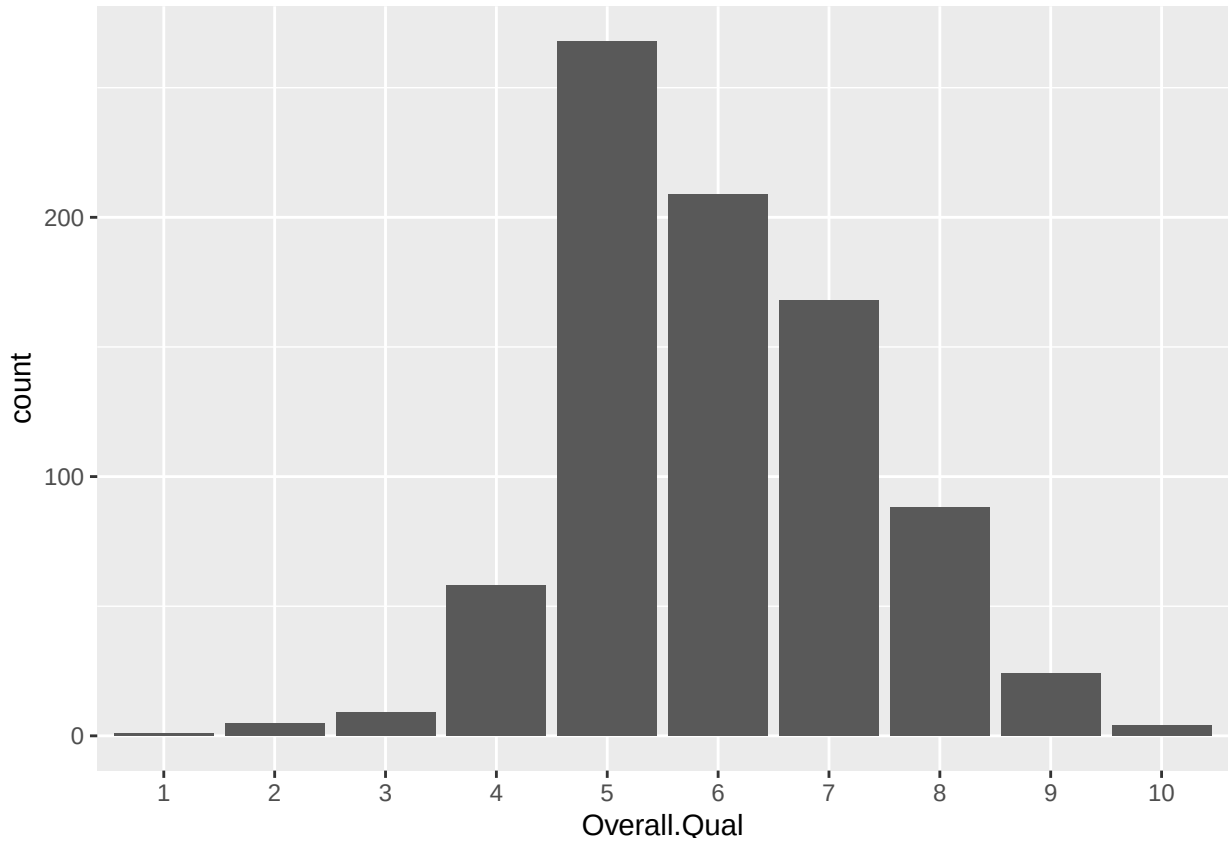
```
## [1] 834
```

```
ames_train %>% ggplot()+geom_point(aes(x=Year.Built,y=price))
```




```
ames_train %>% ggplot()+geom_bar(aes(x=Overall.Qual,stat=price))
```

```
## Warning: Ignoring unknown aesthetics: stat
```




```
(model <- lm(price~Year.Built+Overall.Qual+area+Bedroom.AbvGr+Lot.Area+Overall.Cond,data = ames_train))
```

```
##
## Call:
## lm(formula = price ~ Year.Built + Overall.Qual + area + Bedroom.AbvGr +
##     Lot.Area + Overall.Cond, data = ames_train)
##
## Coefficients:
##    (Intercept)      Year.Built   Overall.Qual2   Overall.Qual3   Overall.Qual4
##     -1.375e+06       7.010e+02      -7.595e+03      -3.078e+02      -6.729e+03
##  Overall.Qual5   Overall.Qual6   Overall.Qual7   Overall.Qual8   Overall.Qual9
##      2.943e+03       5.647e+03       1.855e+04       5.842e+04       1.400e+05
## Overall.Qual10            area   Bedroom.AbvGr        Lot.Area   Overall.Cond2
##      1.931e+05       7.783e+01      -1.011e+04       1.016e+00       4.183e+04
##  Overall.Cond3   Overall.Cond4   Overall.Cond5   Overall.Cond6   Overall.Cond7
##      2.966e+04       3.847e+04       5.323e+04       5.937e+04       6.976e+04
##  Overall.Cond8   Overall.Cond9
##      7.713e+04       7.933e+04
```

```
summary(model)
```

```
##
## Call:
## lm(formula = price ~ Year.Built + Overall.Qual + area + Bedroom.AbvGr +
##     Lot.Area + Overall.Cond, data = ames_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -105562  -15266     210   12705  144943
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.375e+06  8.730e+04 -15.756  < 2e-16 ***
## Year.Built      7.010e+02  4.221e+01  16.607  < 2e-16 ***
## Overall.Qual2  -7.595e+03  2.932e+04  -0.259 0.795692
## Overall.Qual3  -3.078e+02  2.715e+04  -0.011 0.990957
## Overall.Qual4  -6.729e+03  2.626e+04  -0.256 0.797794
## Overall.Qual5   2.943e+03  2.616e+04   0.112 0.910464
## Overall.Qual6   5.647e+03  2.627e+04   0.215 0.829844
## Overall.Qual7   1.855e+04  2.638e+04   0.703 0.482108
## Overall.Qual8   5.842e+04  2.653e+04   2.202 0.027923 *
## Overall.Qual9   1.400e+05  2.696e+04   5.190 2.65e-07 ***
## Overall.Qual10  1.931e+05  2.932e+04   6.587 8.05e-11 ***
## area            7.783e+01  2.971e+00  26.199  < 2e-16 ***
## Bedroom.AbvGr  -1.011e+04  1.458e+03  -6.934 8.35e-12 ***
## Lot.Area        1.016e+00  8.638e-02  11.761  < 2e-16 ***
## Overall.Cond2   4.183e+04  3.122e+04   1.340 0.180623
## Overall.Cond3   2.966e+04  2.056e+04   1.443 0.149471
## Overall.Cond4   3.847e+04  1.925e+04   1.998 0.046028 *
## Overall.Cond5   5.323e+04  1.914e+04   2.781 0.005546 **
## Overall.Cond6   5.937e+04  1.913e+04   3.104 0.001974 **
## Overall.Cond7   6.976e+04  1.912e+04   3.648 0.000281 ***
## Overall.Cond8   7.713e+04  1.934e+04   3.987 7.28e-05 ***
## Overall.Cond9   7.933e+04  2.056e+04   3.858 0.000123 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24790 on 812 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8824
## F-statistic: 298.5 on 21 and 812 DF,  p-value: < 2.2e-16
```

```
stepAIC(model,k=2,direction = "backward",trace = FALSE)
```

```
##
## Call:
## lm(formula = price ~ Year.Built + Overall.Qual + area + Bedroom.AbvGr +
##     Lot.Area + Overall.Cond, data = ames_train)
##
## Coefficients:
##    (Intercept)      Year.Built   Overall.Qual2   Overall.Qual3   Overall.Qual4
##     -1.375e+06       7.010e+02      -7.595e+03      -3.078e+02      -6.729e+03
##  Overall.Qual5   Overall.Qual6   Overall.Qual7   Overall.Qual8   Overall.Qual9
##      2.943e+03       5.647e+03       1.855e+04       5.842e+04       1.400e+05
## Overall.Qual10            area   Bedroom.AbvGr        Lot.Area   Overall.Cond2
##      1.931e+05       7.783e+01      -1.011e+04       1.016e+00       4.183e+04
##  Overall.Cond3   Overall.Cond4   Overall.Cond5   Overall.Cond6   Overall.Cond7
##      2.966e+04       3.847e+04       5.323e+04       5.937e+04       6.976e+04
##  Overall.Cond8   Overall.Cond9
##      7.713e+04       7.933e+04
```

```
sqrt(mean(model$residuals^2))
```

```
## [1] 24457.73
```

```
ames_train %>% dplyr::select(Year.Built,Overall.Qual , area , Bedroom.AbvGr ,
    Lot.Area ,Overall.Cond,price,)
```

```
## # A tibble: 834 x 7
##    Year.Built Overall.Qual  area Bedroom.AbvGr Lot.Area Overall.Cond  price
##         <int> <fct>        <int>         <int>    <int> <fct>         <int>
##  1       1939 6              856             2     7890 6            126000
##  2       1984 5             1049             2     4235 5            139500
##  3       1930 5             1001             2     6060 9            124900
##  4       1900 4             1039             2     8146 8            114000
##  5       2001 8             1665             3     8400 6            227000
##  6       2003 7             1922             4     7301 5            198500
##  7       1953 4              936             2     6000 4             93000
##  8       1984 5              889             3    12395 6            137500
##  9       2005 6             1072             2     3675 5            140000
## 10       2006 8             1342             2     6240 5            219500
## # ... with 824 more rows
```

```
(df <- tibble(Year.Built=2010,Overall.Qual=as.character(5),area=900,Bedroom.AbvGr=4,Lot.Area=7654,Overal
```

```
## # A tibble: 1 x 6
##   Year.Built Overall.Qual  area Bedroom.AbvGr Lot.Area Overall.Cond
##        <dbl> <chr>        <dbl>         <dbl>    <dbl> <chr>
## 1       2010 5              900             4     7654 5
```

```
predict(model,df)
```

```
##        1
## 127090.1
```

```
103203.8 -126000
```

```
## [1] -22796.2
```

```
model$residuals[1]
```

```
##        1
## 22796.21
```