MBissonnette Phase 1 Project

me

2023-02-22

Needed libraries

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──  
## ✔ broom 1.0.2 ✔ rsample 1.1.1  
## ✔ dials 1.1.0 ✔ tune 1.0.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.2  
## ✔ modeldata 1.1.0 ✔ workflowsets 1.0.0  
## ✔ parsnip 1.0.3 ✔ yardstick 1.1.0  
## ✔ recipes 1.0.4   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(dplyr)  
library(GGally) #ggcorr and ggpairs

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggcorrplot) #correlation plot alternative  
library(glmnet) #for Lasso, ridge, and elastic net models

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loaded glmnet 4.1-6

library(gridExtra) #create grids of plots

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

library(car) #for the VIF function

## Loading required package: carData  
##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode  
##   
## The following object is masked from 'package:purrr':  
##   
## some

library(MASS) #access to forward and backward selection algorithms

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(leaps) #best subset selection  
library(lmtest)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(splines) #for nonlinear fitting  
library(mice) #package for imputation

##   
## Attaching package: 'mice'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM) #visualizing missingness

## Loading required package: colorspace  
## Loading required package: grid  
## VIM is ready to use.  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues  
##   
## Attaching package: 'VIM'  
##   
## The following object is masked from 'package:recipes':  
##   
## prepare  
##   
## The following object is masked from 'package:datasets':  
##   
## sleep

library(naniar) #visualizing missingness  
library(skimr) #alternative way to view dataset summaries

##   
## Attaching package: 'skimr'  
##   
## The following object is masked from 'package:naniar':  
##   
## n\_complete

library(UpSetR) #visualizing missingness

Read-in the dataset

amesstudent = read\_csv("ames\_student-1.csv",show\_col\_types = FALSE)

Structure and summary.

str(amesstudent)

## spc\_tbl\_ [2,053 × 81] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ MS\_SubClass : chr [1:2053] "One\_Story\_1946\_and\_Newer\_All\_Styles" "One\_Story\_1946\_and\_Newer\_All\_Styles" "One\_Story\_1946\_and\_Newer\_All\_Styles" "One\_Story\_1946\_and\_Newer\_All\_Styles" ...  
## $ MS\_Zoning : chr [1:2053] "Residential\_Low\_Density" "Residential\_High\_Density" "Residential\_Low\_Density" "Residential\_Low\_Density" ...  
## $ Lot\_Frontage : num [1:2053] 141 80 81 93 74 78 43 39 0 85 ...  
## $ Lot\_Area : num [1:2053] 31770 11622 14267 11160 13830 ...  
## $ Street : chr [1:2053] "Pave" "Pave" "Pave" "Pave" ...  
## $ Alley : chr [1:2053] "No\_Alley\_Access" "No\_Alley\_Access" "No\_Alley\_Access" "No\_Alley\_Access" ...  
## $ Lot\_Shape : chr [1:2053] "Slightly\_Irregular" "Regular" "Slightly\_Irregular" "Regular" ...  
## $ Land\_Contour : chr [1:2053] "Lvl" "Lvl" "Lvl" "Lvl" ...  
## $ Utilities : chr [1:2053] "AllPub" "AllPub" "AllPub" "AllPub" ...  
## $ Lot\_Config : chr [1:2053] "Corner" "Inside" "Corner" "Corner" ...  
## $ Land\_Slope : chr [1:2053] "Gtl" "Gtl" "Gtl" "Gtl" ...  
## $ Neighborhood : chr [1:2053] "North\_Ames" "North\_Ames" "North\_Ames" "North\_Ames" ...  
## $ Condition\_1 : chr [1:2053] "Norm" "Feedr" "Norm" "Norm" ...  
## $ Condition\_2 : chr [1:2053] "Norm" "Norm" "Norm" "Norm" ...  
## $ Bldg\_Type : chr [1:2053] "OneFam" "OneFam" "OneFam" "OneFam" ...  
## $ House\_Style : chr [1:2053] "One\_Story" "One\_Story" "One\_Story" "One\_Story" ...  
## $ Overall\_Qual : chr [1:2053] "Above\_Average" "Average" "Above\_Average" "Good" ...  
## $ Overall\_Cond : chr [1:2053] "Average" "Above\_Average" "Above\_Average" "Average" ...  
## $ Year\_Built : num [1:2053] 1960 1961 1958 1968 1997 ...  
## $ Year\_Remod\_Add : num [1:2053] 1960 1961 1958 1968 1998 ...  
## $ Roof\_Style : chr [1:2053] "Hip" "Gable" "Hip" "Hip" ...  
## $ Roof\_Matl : chr [1:2053] "CompShg" "CompShg" "CompShg" "CompShg" ...  
## $ Exterior\_1st : chr [1:2053] "BrkFace" "VinylSd" "Wd Sdng" "BrkFace" ...  
## $ Exterior\_2nd : chr [1:2053] "Plywood" "VinylSd" "Wd Sdng" "BrkFace" ...  
## $ Mas\_Vnr\_Type : chr [1:2053] "Stone" "None" "BrkFace" "None" ...  
## $ Mas\_Vnr\_Area : num [1:2053] 112 0 108 0 0 20 0 0 0 0 ...  
## $ Exter\_Qual : chr [1:2053] "Typical" "Typical" "Typical" "Good" ...  
## $ Exter\_Cond : chr [1:2053] "Typical" "Typical" "Typical" "Typical" ...  
## $ Foundation : chr [1:2053] "CBlock" "CBlock" "CBlock" "CBlock" ...  
## $ Bsmt\_Qual : chr [1:2053] "Typical" "Typical" "Typical" "Typical" ...  
## $ Bsmt\_Cond : chr [1:2053] "Good" "Typical" "Typical" "Typical" ...  
## $ Bsmt\_Exposure : chr [1:2053] "Gd" "No" "No" "No" ...  
## $ BsmtFin\_Type\_1 : chr [1:2053] "BLQ" "Rec" "ALQ" "ALQ" ...  
## $ BsmtFin\_SF\_1 : num [1:2053] 2 6 1 1 3 3 1 3 1 3 ...  
## $ BsmtFin\_Type\_2 : chr [1:2053] "Unf" "LwQ" "Unf" "Unf" ...  
## $ BsmtFin\_SF\_2 : num [1:2053] 0 144 0 0 0 0 0 0 0 0 ...  
## $ Bsmt\_Unf\_SF : num [1:2053] 441 270 406 1045 137 ...  
## $ Total\_Bsmt\_SF : num [1:2053] 1080 882 1329 2110 928 ...  
## $ Heating : chr [1:2053] "GasA" "GasA" "GasA" "GasA" ...  
## $ Heating\_QC : chr [1:2053] "Fair" "Typical" "Typical" "Excellent" ...  
## $ Central\_Air : chr [1:2053] "Y" "Y" "Y" "Y" ...  
## $ Electrical : chr [1:2053] "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...  
## $ First\_Flr\_SF : num [1:2053] 1656 896 1329 2110 928 ...  
## $ Second\_Flr\_SF : num [1:2053] 0 0 0 0 701 678 0 0 0 0 ...  
## $ Low\_Qual\_Fin\_SF : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Gr\_Liv\_Area : num [1:2053] 1656 896 1329 2110 1629 ...  
## $ Bsmt\_Full\_Bath : num [1:2053] 1 0 0 1 0 0 0 1 1 1 ...  
## $ Bsmt\_Half\_Bath : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Full\_Bath : num [1:2053] 1 1 1 2 2 2 2 2 2 1 ...  
## $ Half\_Bath : num [1:2053] 0 0 1 1 1 1 0 0 0 1 ...  
## $ Bedroom\_AbvGr : num [1:2053] 3 2 3 3 3 3 2 2 3 2 ...  
## $ Kitchen\_AbvGr : num [1:2053] 1 1 1 1 1 1 1 1 1 1 ...  
## $ Kitchen\_Qual : chr [1:2053] "Typical" "Typical" "Good" "Excellent" ...  
## $ TotRms\_AbvGrd : num [1:2053] 7 5 6 8 6 7 5 5 6 5 ...  
## $ Functional : chr [1:2053] "Typ" "Typ" "Typ" "Typ" ...  
## $ Fireplaces : num [1:2053] 2 0 0 2 1 1 0 1 0 1 ...  
## $ Fireplace\_Qu : chr [1:2053] "Good" "No\_Fireplace" "No\_Fireplace" "Typical" ...  
## $ Garage\_Type : chr [1:2053] "Attchd" "Attchd" "Attchd" "Attchd" ...  
## $ Garage\_Finish : chr [1:2053] "Fin" "Unf" "Unf" "Fin" ...  
## $ Garage\_Cars : num [1:2053] 2 1 1 2 2 2 2 2 2 2 ...  
## $ Garage\_Area : num [1:2053] 528 730 312 522 482 470 506 608 420 506 ...  
## $ Garage\_Qual : chr [1:2053] "Typical" "Typical" "Typical" "Typical" ...  
## $ Garage\_Cond : chr [1:2053] "Typical" "Typical" "Typical" "Typical" ...  
## $ Paved\_Drive : chr [1:2053] "Partial\_Pavement" "Paved" "Paved" "Paved" ...  
## $ Wood\_Deck\_SF : num [1:2053] 210 140 393 0 212 360 0 237 483 192 ...  
## $ Open\_Porch\_SF : num [1:2053] 62 0 36 0 34 36 82 152 21 0 ...  
## $ Enclosed\_Porch : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Three\_season\_porch: num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Screen\_Porch : num [1:2053] 0 120 0 0 0 0 144 0 0 0 ...  
## $ Pool\_Area : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Pool\_QC : chr [1:2053] "No\_Pool" "No\_Pool" "No\_Pool" "No\_Pool" ...  
## $ Fence : chr [1:2053] "No\_Fence" "Minimum\_Privacy" "No\_Fence" "No\_Fence" ...  
## $ Misc\_Feature : chr [1:2053] "None" "None" "Gar2" "None" ...  
## $ Misc\_Val : num [1:2053] 0 0 12500 0 0 0 0 0 500 0 ...  
## $ Mo\_Sold : num [1:2053] 5 6 6 4 3 6 1 3 3 2 ...  
## $ Year\_Sold : num [1:2053] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...  
## $ Sale\_Type : chr [1:2053] "WD" "WD" "WD" "WD" ...  
## $ Sale\_Condition : chr [1:2053] "Normal" "Normal" "Normal" "Normal" ...  
## $ Longitude : num [1:2053] -93.6 -93.6 -93.6 -93.6 -93.6 ...  
## $ Latitude : num [1:2053] 42.1 42.1 42.1 42.1 42.1 ...  
## $ Above\_Median : chr [1:2053] "Yes" "No" "Yes" "Yes" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. MS\_SubClass = col\_character(),  
## .. MS\_Zoning = col\_character(),  
## .. Lot\_Frontage = col\_double(),  
## .. Lot\_Area = col\_double(),  
## .. Street = col\_character(),  
## .. Alley = col\_character(),  
## .. Lot\_Shape = col\_character(),  
## .. Land\_Contour = col\_character(),  
## .. Utilities = col\_character(),  
## .. Lot\_Config = col\_character(),  
## .. Land\_Slope = col\_character(),  
## .. Neighborhood = col\_character(),  
## .. Condition\_1 = col\_character(),  
## .. Condition\_2 = col\_character(),  
## .. Bldg\_Type = col\_character(),  
## .. House\_Style = col\_character(),  
## .. Overall\_Qual = col\_character(),  
## .. Overall\_Cond = col\_character(),  
## .. Year\_Built = col\_double(),  
## .. Year\_Remod\_Add = col\_double(),  
## .. Roof\_Style = col\_character(),  
## .. Roof\_Matl = col\_character(),  
## .. Exterior\_1st = col\_character(),  
## .. Exterior\_2nd = col\_character(),  
## .. Mas\_Vnr\_Type = col\_character(),  
## .. Mas\_Vnr\_Area = col\_double(),  
## .. Exter\_Qual = col\_character(),  
## .. Exter\_Cond = col\_character(),  
## .. Foundation = col\_character(),  
## .. Bsmt\_Qual = col\_character(),  
## .. Bsmt\_Cond = col\_character(),  
## .. Bsmt\_Exposure = col\_character(),  
## .. BsmtFin\_Type\_1 = col\_character(),  
## .. BsmtFin\_SF\_1 = col\_double(),  
## .. BsmtFin\_Type\_2 = col\_character(),  
## .. BsmtFin\_SF\_2 = col\_double(),  
## .. Bsmt\_Unf\_SF = col\_double(),  
## .. Total\_Bsmt\_SF = col\_double(),  
## .. Heating = col\_character(),  
## .. Heating\_QC = col\_character(),  
## .. Central\_Air = col\_character(),  
## .. Electrical = col\_character(),  
## .. First\_Flr\_SF = col\_double(),  
## .. Second\_Flr\_SF = col\_double(),  
## .. Low\_Qual\_Fin\_SF = col\_double(),  
## .. Gr\_Liv\_Area = col\_double(),  
## .. Bsmt\_Full\_Bath = col\_double(),  
## .. Bsmt\_Half\_Bath = col\_double(),  
## .. Full\_Bath = col\_double(),  
## .. Half\_Bath = col\_double(),  
## .. Bedroom\_AbvGr = col\_double(),  
## .. Kitchen\_AbvGr = col\_double(),  
## .. Kitchen\_Qual = col\_character(),  
## .. TotRms\_AbvGrd = col\_double(),  
## .. Functional = col\_character(),  
## .. Fireplaces = col\_double(),  
## .. Fireplace\_Qu = col\_character(),  
## .. Garage\_Type = col\_character(),  
## .. Garage\_Finish = col\_character(),  
## .. Garage\_Cars = col\_double(),  
## .. Garage\_Area = col\_double(),  
## .. Garage\_Qual = col\_character(),  
## .. Garage\_Cond = col\_character(),  
## .. Paved\_Drive = col\_character(),  
## .. Wood\_Deck\_SF = col\_double(),  
## .. Open\_Porch\_SF = col\_double(),  
## .. Enclosed\_Porch = col\_double(),  
## .. Three\_season\_porch = col\_double(),  
## .. Screen\_Porch = col\_double(),  
## .. Pool\_Area = col\_double(),  
## .. Pool\_QC = col\_character(),  
## .. Fence = col\_character(),  
## .. Misc\_Feature = col\_character(),  
## .. Misc\_Val = col\_double(),  
## .. Mo\_Sold = col\_double(),  
## .. Year\_Sold = col\_double(),  
## .. Sale\_Type = col\_character(),  
## .. Sale\_Condition = col\_character(),  
## .. Longitude = col\_double(),  
## .. Latitude = col\_double(),  
## .. Above\_Median = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(amesstudent)

## MS\_SubClass MS\_Zoning Lot\_Frontage Lot\_Area   
## Length:2053 Length:2053 Min. : 0.00 Min. : 1300   
## Class :character Class :character 1st Qu.: 43.00 1st Qu.: 7500   
## Mode :character Mode :character Median : 62.00 Median : 9548   
## Mean : 57.38 Mean : 10258   
## 3rd Qu.: 78.00 3rd Qu.: 11600   
## Max. :313.00 Max. :215245   
## Street Alley Lot\_Shape Land\_Contour   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Utilities Lot\_Config Land\_Slope Neighborhood   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Condition\_1 Condition\_2 Bldg\_Type House\_Style   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Overall\_Qual Overall\_Cond Year\_Built Year\_Remod\_Add  
## Length:2053 Length:2053 Min. :1875 Min. :1950   
## Class :character Class :character 1st Qu.:1953 1st Qu.:1965   
## Mode :character Mode :character Median :1972 Median :1993   
## Mean :1971 Mean :1984   
## 3rd Qu.:2000 3rd Qu.:2004   
## Max. :2010 Max. :2010   
## Roof\_Style Roof\_Matl Exterior\_1st Exterior\_2nd   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Mas\_Vnr\_Type Mas\_Vnr\_Area Exter\_Qual Exter\_Cond   
## Length:2053 Min. : 0.0 Length:2053 Length:2053   
## Class :character 1st Qu.: 0.0 Class :character Class :character   
## Mode :character Median : 0.0 Mode :character Mode :character   
## Mean : 103.8   
## 3rd Qu.: 164.0   
## Max. :1600.0   
## Foundation Bsmt\_Qual Bsmt\_Cond Bsmt\_Exposure   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## BsmtFin\_Type\_1 BsmtFin\_SF\_1 BsmtFin\_Type\_2 BsmtFin\_SF\_2   
## Length:2053 Min. :1.00 Length:2053 Min. : 0.00   
## Class :character 1st Qu.:3.00 Class :character 1st Qu.: 0.00   
## Mode :character Median :3.00 Mode :character Median : 0.00   
## Mean :4.21 Mean : 52.57   
## 3rd Qu.:7.00 3rd Qu.: 0.00   
## Max. :7.00 Max. :1526.00   
## Bsmt\_Unf\_SF Total\_Bsmt\_SF Heating Heating\_QC   
## Min. : 0.0 Min. : 0 Length:2053 Length:2053   
## 1st Qu.: 226.0 1st Qu.: 793 Class :character Class :character   
## Median : 460.0 Median : 988 Mode :character Mode :character   
## Mean : 561.2 Mean :1055   
## 3rd Qu.: 801.0 3rd Qu.:1304   
## Max. :2336.0 Max. :5095   
## Central\_Air Electrical First\_Flr\_SF Second\_Flr\_SF   
## Length:2053 Length:2053 Min. : 432 Min. : 0.0   
## Class :character Class :character 1st Qu.: 882 1st Qu.: 0.0   
## Mode :character Mode :character Median :1088 Median : 0.0   
## Mean :1168 Mean : 326.1   
## 3rd Qu.:1402 3rd Qu.: 701.0   
## Max. :5095 Max. :1862.0   
## Low\_Qual\_Fin\_SF Gr\_Liv\_Area Bsmt\_Full\_Bath Bsmt\_Half\_Bath   
## Min. : 0.000 Min. : 480 Min. :0.0000 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:1137 1st Qu.:0.0000 1st Qu.:0.00000   
## Median : 0.000 Median :1447 Median :0.0000 Median :0.00000   
## Mean : 4.973 Mean :1499 Mean :0.4301 Mean :0.05796   
## 3rd Qu.: 0.000 3rd Qu.:1737 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1064.000 Max. :5095 Max. :3.0000 Max. :2.00000   
## Full\_Bath Half\_Bath Bedroom\_AbvGr Kitchen\_AbvGr   
## Min. :0.000 Min. :0.0000 Min. :0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000 1st Qu.:1.000   
## Median :2.000 Median :0.0000 Median :3.000 Median :1.000   
## Mean :1.564 Mean :0.3751 Mean :2.855 Mean :1.047   
## 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :4.000 Max. :2.0000 Max. :6.000 Max. :3.000   
## Kitchen\_Qual TotRms\_AbvGrd Functional Fireplaces   
## Length:2053 Min. : 3.000 Length:2053 Min. :0.000   
## Class :character 1st Qu.: 5.000 Class :character 1st Qu.:0.000   
## Mode :character Median : 6.000 Mode :character Median :1.000   
## Mean : 6.442 Mean :0.603   
## 3rd Qu.: 7.000 3rd Qu.:1.000   
## Max. :15.000 Max. :4.000   
## Fireplace\_Qu Garage\_Type Garage\_Finish Garage\_Cars   
## Length:2053 Length:2053 Length:2053 Min. :0.000   
## Class :character Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Mode :character Median :2.000   
## Mean :1.774   
## 3rd Qu.:2.000   
## Max. :5.000   
## Garage\_Area Garage\_Qual Garage\_Cond Paved\_Drive   
## Min. : 0 Length:2053 Length:2053 Length:2053   
## 1st Qu.: 320 Class :character Class :character Class :character   
## Median : 478 Mode :character Mode :character Mode :character   
## Mean : 472   
## 3rd Qu.: 576   
## Max. :1488   
## Wood\_Deck\_SF Open\_Porch\_SF Enclosed\_Porch Three\_season\_porch  
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.00 Median : 27.00 Median : 0.00 Median : 0.000   
## Mean : 93.52 Mean : 48.17 Mean : 23.02 Mean : 2.799   
## 3rd Qu.: 168.00 3rd Qu.: 72.00 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :1424.00 Max. :742.00 Max. :584.00 Max. :407.000   
## Screen\_Porch Pool\_Area Pool\_QC Fence   
## Min. : 0.00 Min. : 0.000 Length:2053 Length:2053   
## 1st Qu.: 0.00 1st Qu.: 0.000 Class :character Class :character   
## Median : 0.00 Median : 0.000 Mode :character Mode :character   
## Mean : 16.68 Mean : 1.339   
## 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :576.00 Max. :800.000   
## Misc\_Feature Misc\_Val Mo\_Sold Year\_Sold   
## Length:2053 Min. : 0.00 Min. : 1.000 Min. :2006   
## Class :character 1st Qu.: 0.00 1st Qu.: 4.000 1st Qu.:2007   
## Mode :character Median : 0.00 Median : 6.000 Median :2008   
## Mean : 60.12 Mean : 6.189 Mean :2008   
## 3rd Qu.: 0.00 3rd Qu.: 8.000 3rd Qu.:2009   
## Max. :17000.00 Max. :12.000 Max. :2010   
## Sale\_Type Sale\_Condition Longitude Latitude   
## Length:2053 Length:2053 Min. :-93.69 Min. :41.99   
## Class :character Class :character 1st Qu.:-93.66 1st Qu.:42.02   
## Mode :character Mode :character Median :-93.64 Median :42.03   
## Mean :-93.64 Mean :42.03   
## 3rd Qu.:-93.62 3rd Qu.:42.05   
## Max. :-93.58 Max. :42.06   
## Above\_Median   
## Length:2053   
## Class :character   
## Mode :character   
##   
##   
##

glimpse(amesstudent)

## Rows: 2,053  
## Columns: 81  
## $ MS\_SubClass <chr> "One\_Story\_1946\_and\_Newer\_All\_Styles", "One\_Story\_1…  
## $ MS\_Zoning <chr> "Residential\_Low\_Density", "Residential\_High\_Densit…  
## $ Lot\_Frontage <dbl> 141, 80, 81, 93, 74, 78, 43, 39, 0, 85, 0, 47, 152,…  
## $ Lot\_Area <dbl> 31770, 11622, 14267, 11160, 13830, 9978, 5005, 5389…  
## $ Street <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pa…  
## $ Alley <chr> "No\_Alley\_Access", "No\_Alley\_Access", "No\_Alley\_Acc…  
## $ Lot\_Shape <chr> "Slightly\_Irregular", "Regular", "Slightly\_Irregula…  
## $ Land\_Contour <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "HLS", "L…  
## $ Utilities <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "…  
## $ Lot\_Config <chr> "Corner", "Inside", "Corner", "Corner", "Inside", "…  
## $ Land\_Slope <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "G…  
## $ Neighborhood <chr> "North\_Ames", "North\_Ames", "North\_Ames", "North\_Am…  
## $ Condition\_1 <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "N…  
## $ Condition\_2 <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "No…  
## $ Bldg\_Type <chr> "OneFam", "OneFam", "OneFam", "OneFam", "OneFam", "…  
## $ House\_Style <chr> "One\_Story", "One\_Story", "One\_Story", "One\_Story",…  
## $ Overall\_Qual <chr> "Above\_Average", "Average", "Above\_Average", "Good"…  
## $ Overall\_Cond <chr> "Average", "Above\_Average", "Above\_Average", "Avera…  
## $ Year\_Built <dbl> 1960, 1961, 1958, 1968, 1997, 1998, 1992, 1995, 199…  
## $ Year\_Remod\_Add <dbl> 1960, 1961, 1958, 1968, 1998, 1998, 1992, 1996, 200…  
## $ Roof\_Style <chr> "Hip", "Gable", "Hip", "Hip", "Gable", "Gable", "Ga…  
## $ Roof\_Matl <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompSh…  
## $ Exterior\_1st <chr> "BrkFace", "VinylSd", "Wd Sdng", "BrkFace", "VinylS…  
## $ Exterior\_2nd <chr> "Plywood", "VinylSd", "Wd Sdng", "BrkFace", "VinylS…  
## $ Mas\_Vnr\_Type <chr> "Stone", "None", "BrkFace", "None", "None", "BrkFac…  
## $ Mas\_Vnr\_Area <dbl> 112, 0, 108, 0, 0, 20, 0, 0, 0, 0, 0, 603, 0, 350, …  
## $ Exter\_Qual <chr> "Typical", "Typical", "Typical", "Good", "Typical",…  
## $ Exter\_Cond <chr> "Typical", "Typical", "Typical", "Typical", "Typica…  
## $ Foundation <chr> "CBlock", "CBlock", "CBlock", "CBlock", "PConc", "P…  
## $ Bsmt\_Qual <chr> "Typical", "Typical", "Typical", "Typical", "Good",…  
## $ Bsmt\_Cond <chr> "Good", "Typical", "Typical", "Typical", "Typical",…  
## $ Bsmt\_Exposure <chr> "Gd", "No", "No", "No", "No", "No", "No", "No", "No…  
## $ BsmtFin\_Type\_1 <chr> "BLQ", "Rec", "ALQ", "ALQ", "GLQ", "GLQ", "ALQ", "G…  
## $ BsmtFin\_SF\_1 <dbl> 2, 6, 1, 1, 3, 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, …  
## $ BsmtFin\_Type\_2 <chr> "Unf", "LwQ", "Unf", "Unf", "Unf", "Unf", "Unf", "U…  
## $ BsmtFin\_SF\_2 <dbl> 0, 144, 0, 0, 0, 0, 0, 0, 0, 0, 1120, 0, 0, 0, 0, 0…  
## $ Bsmt\_Unf\_SF <dbl> 441, 270, 406, 1045, 137, 324, 1017, 415, 233, 663,…  
## $ Total\_Bsmt\_SF <dbl> 1080, 882, 1329, 2110, 928, 926, 1280, 1595, 1168, …  
## $ Heating <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", "Ga…  
## $ Heating\_QC <chr> "Fair", "Typical", "Typical", "Excellent", "Good", …  
## $ Central\_Air <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "…  
## $ Electrical <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr…  
## $ First\_Flr\_SF <dbl> 1656, 896, 1329, 2110, 928, 926, 1280, 1616, 1187, …  
## $ Second\_Flr\_SF <dbl> 0, 0, 0, 0, 701, 678, 0, 0, 0, 0, 0, 1589, 672, 0, …  
## $ Low\_Qual\_Fin\_SF <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ Gr\_Liv\_Area <dbl> 1656, 896, 1329, 2110, 1629, 1604, 1280, 1616, 1187…  
## $ Bsmt\_Full\_Bath <dbl> 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, …  
## $ Bsmt\_Half\_Bath <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ Full\_Bath <dbl> 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1, 2, 2, 1, …  
## $ Half\_Bath <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, …  
## $ Bedroom\_AbvGr <dbl> 3, 2, 3, 3, 3, 3, 2, 2, 3, 2, 1, 4, 4, 1, 3, 3, 2, …  
## $ Kitchen\_AbvGr <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ Kitchen\_Qual <chr> "Typical", "Typical", "Good", "Excellent", "Typical…  
## $ TotRms\_AbvGrd <dbl> 7, 5, 6, 8, 6, 7, 5, 5, 6, 5, 4, 12, 8, 8, 7, 7, 5,…  
## $ Functional <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "T…  
## $ Fireplaces <dbl> 2, 0, 0, 2, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, …  
## $ Fireplace\_Qu <chr> "Good", "No\_Fireplace", "No\_Fireplace", "Typical", …  
## $ Garage\_Type <chr> "Attchd", "Attchd", "Attchd", "Attchd", "Attchd", "…  
## $ Garage\_Finish <chr> "Fin", "Unf", "Unf", "Fin", "Fin", "Fin", "RFn", "R…  
## $ Garage\_Cars <dbl> 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, …  
## $ Garage\_Area <dbl> 528, 730, 312, 522, 482, 470, 506, 608, 420, 506, 5…  
## $ Garage\_Qual <chr> "Typical", "Typical", "Typical", "Typical", "Typica…  
## $ Garage\_Cond <chr> "Typical", "Typical", "Typical", "Typical", "Typica…  
## $ Paved\_Drive <chr> "Partial\_Pavement", "Paved", "Paved", "Paved", "Pav…  
## $ Wood\_Deck\_SF <dbl> 210, 140, 393, 0, 212, 360, 0, 237, 483, 192, 0, 50…  
## $ Open\_Porch\_SF <dbl> 62, 0, 36, 0, 34, 36, 82, 152, 21, 0, 54, 36, 12, 0…  
## $ Enclosed\_Porch <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ Three\_season\_porch <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ Screen\_Porch <dbl> 0, 120, 0, 0, 0, 0, 144, 0, 0, 0, 140, 210, 0, 0, 0…  
## $ Pool\_Area <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ Pool\_QC <chr> "No\_Pool", "No\_Pool", "No\_Pool", "No\_Pool", "No\_Poo…  
## $ Fence <chr> "No\_Fence", "Minimum\_Privacy", "No\_Fence", "No\_Fenc…  
## $ Misc\_Feature <chr> "None", "None", "Gar2", "None", "None", "None", "No…  
## $ Misc\_Val <dbl> 0, 0, 12500, 0, 0, 0, 0, 0, 500, 0, 0, 0, 0, 0, 0, …  
## $ Mo\_Sold <dbl> 5, 6, 6, 4, 3, 6, 1, 3, 3, 2, 6, 6, 6, 6, 1, 1, 3, …  
## $ Year\_Sold <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 201…  
## $ Sale\_Type <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD…  
## $ Sale\_Condition <chr> "Normal", "Normal", "Normal", "Normal", "Normal", "…  
## $ Longitude <dbl> -93.61975, -93.61976, -93.61939, -93.61732, -93.638…  
## $ Latitude <dbl> 42.05403, 42.05301, 42.05266, 42.05125, 42.06090, 4…  
## $ Above\_Median <chr> "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Ye…

skim(amesstudent)

Data summary

|  |  |
| --- | --- |
| Name | amesstudent |
| Number of rows | 2053 |
| Number of columns | 81 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 47 |
| numeric | 34 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MS\_SubClass | 0 | 1 | 11 | 41 | 0 | 16 | 0 |
| MS\_Zoning | 0 | 1 | 5 | 28 | 0 | 7 | 0 |
| Street | 0 | 1 | 4 | 4 | 0 | 2 | 0 |
| Alley | 0 | 1 | 5 | 15 | 0 | 3 | 0 |
| Lot\_Shape | 0 | 1 | 7 | 20 | 0 | 4 | 0 |
| Land\_Contour | 0 | 1 | 3 | 3 | 0 | 4 | 0 |
| Utilities | 0 | 1 | 6 | 6 | 0 | 2 | 0 |
| Lot\_Config | 0 | 1 | 3 | 7 | 0 | 5 | 0 |
| Land\_Slope | 0 | 1 | 3 | 3 | 0 | 3 | 0 |
| Neighborhood | 0 | 1 | 6 | 39 | 0 | 28 | 0 |
| Condition\_1 | 0 | 1 | 4 | 6 | 0 | 9 | 0 |
| Condition\_2 | 0 | 1 | 4 | 6 | 0 | 8 | 0 |
| Bldg\_Type | 0 | 1 | 5 | 8 | 0 | 5 | 0 |
| House\_Style | 0 | 1 | 4 | 16 | 0 | 8 | 0 |
| Overall\_Qual | 0 | 1 | 4 | 14 | 0 | 10 | 0 |
| Overall\_Cond | 0 | 1 | 4 | 13 | 0 | 9 | 0 |
| Roof\_Style | 0 | 1 | 3 | 7 | 0 | 6 | 0 |
| Roof\_Matl | 0 | 1 | 4 | 7 | 0 | 6 | 0 |
| Exterior\_1st | 0 | 1 | 5 | 7 | 0 | 16 | 0 |
| Exterior\_2nd | 0 | 1 | 5 | 7 | 0 | 17 | 0 |
| Mas\_Vnr\_Type | 0 | 1 | 4 | 7 | 0 | 5 | 0 |
| Exter\_Qual | 0 | 1 | 4 | 9 | 0 | 4 | 0 |
| Exter\_Cond | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| Foundation | 0 | 1 | 4 | 6 | 0 | 6 | 0 |
| Bsmt\_Qual | 0 | 1 | 4 | 11 | 0 | 6 | 0 |
| Bsmt\_Cond | 0 | 1 | 4 | 11 | 0 | 6 | 0 |
| Bsmt\_Exposure | 0 | 1 | 2 | 11 | 0 | 5 | 0 |
| BsmtFin\_Type\_1 | 0 | 1 | 3 | 11 | 0 | 7 | 0 |
| BsmtFin\_Type\_2 | 0 | 1 | 3 | 11 | 0 | 7 | 0 |
| Heating | 0 | 1 | 4 | 5 | 0 | 6 | 0 |
| Heating\_QC | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| Central\_Air | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| Electrical | 0 | 1 | 5 | 7 | 0 | 5 | 0 |
| Kitchen\_Qual | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| Functional | 0 | 1 | 3 | 4 | 0 | 8 | 0 |
| Fireplace\_Qu | 0 | 1 | 4 | 12 | 0 | 6 | 0 |
| Garage\_Type | 0 | 1 | 6 | 19 | 0 | 7 | 0 |
| Garage\_Finish | 0 | 1 | 3 | 9 | 0 | 4 | 0 |
| Garage\_Qual | 0 | 1 | 4 | 9 | 0 | 6 | 0 |
| Garage\_Cond | 0 | 1 | 4 | 9 | 0 | 6 | 0 |
| Paved\_Drive | 0 | 1 | 5 | 16 | 0 | 3 | 0 |
| Pool\_QC | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| Fence | 0 | 1 | 8 | 17 | 0 | 5 | 0 |
| Misc\_Feature | 0 | 1 | 4 | 4 | 0 | 5 | 0 |
| Sale\_Type | 0 | 1 | 2 | 5 | 0 | 10 | 0 |
| Sale\_Condition | 0 | 1 | 6 | 7 | 0 | 6 | 0 |
| Above\_Median | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lot\_Frontage | 0 | 1 | 57.38 | 33.20 | 0.00 | 43.00 | 62.00 | 78.00 | 313.00 | ▇▇▁▁▁ |
| Lot\_Area | 0 | 1 | 10258.40 | 8427.38 | 1300.00 | 7500.00 | 9548.00 | 11600.00 | 215245.00 | ▇▁▁▁▁ |
| Year\_Built | 0 | 1 | 1970.64 | 30.40 | 1875.00 | 1953.00 | 1972.00 | 2000.00 | 2010.00 | ▁▂▃▆▇ |
| Year\_Remod\_Add | 0 | 1 | 1984.08 | 20.96 | 1950.00 | 1965.00 | 1993.00 | 2004.00 | 2010.00 | ▅▂▂▃▇ |
| Mas\_Vnr\_Area | 0 | 1 | 103.75 | 183.59 | 0.00 | 0.00 | 0.00 | 164.00 | 1600.00 | ▇▁▁▁▁ |
| BsmtFin\_SF\_1 | 0 | 1 | 4.21 | 2.24 | 1.00 | 3.00 | 3.00 | 7.00 | 7.00 | ▅▆▁▁▇ |
| BsmtFin\_SF\_2 | 0 | 1 | 52.57 | 175.99 | 0.00 | 0.00 | 0.00 | 0.00 | 1526.00 | ▇▁▁▁▁ |
| Bsmt\_Unf\_SF | 0 | 1 | 561.19 | 441.72 | 0.00 | 226.00 | 460.00 | 801.00 | 2336.00 | ▇▅▂▁▁ |
| Total\_Bsmt\_SF | 0 | 1 | 1054.57 | 435.33 | 0.00 | 793.00 | 988.00 | 1304.00 | 5095.00 | ▇▇▁▁▁ |
| First\_Flr\_SF | 0 | 1 | 1167.52 | 391.79 | 432.00 | 882.00 | 1088.00 | 1402.00 | 5095.00 | ▇▃▁▁▁ |
| Second\_Flr\_SF | 0 | 1 | 326.07 | 422.44 | 0.00 | 0.00 | 0.00 | 701.00 | 1862.00 | ▇▂▂▁▁ |
| Low\_Qual\_Fin\_SF | 0 | 1 | 4.97 | 49.09 | 0.00 | 0.00 | 0.00 | 0.00 | 1064.00 | ▇▁▁▁▁ |
| Gr\_Liv\_Area | 0 | 1 | 1498.56 | 487.84 | 480.00 | 1137.00 | 1447.00 | 1737.00 | 5095.00 | ▇▇▁▁▁ |
| Bsmt\_Full\_Bath | 0 | 1 | 0.43 | 0.53 | 0.00 | 0.00 | 0.00 | 1.00 | 3.00 | ▇▆▁▁▁ |
| Bsmt\_Half\_Bath | 0 | 1 | 0.06 | 0.24 | 0.00 | 0.00 | 0.00 | 0.00 | 2.00 | ▇▁▁▁▁ |
| Full\_Bath | 0 | 1 | 1.56 | 0.55 | 0.00 | 1.00 | 2.00 | 2.00 | 4.00 | ▁▇▇▁▁ |
| Half\_Bath | 0 | 1 | 0.38 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 2.00 | ▇▁▅▁▁ |
| Bedroom\_AbvGr | 0 | 1 | 2.86 | 0.82 | 0.00 | 2.00 | 3.00 | 3.00 | 6.00 | ▁▃▇▂▁ |
| Kitchen\_AbvGr | 0 | 1 | 1.05 | 0.22 | 1.00 | 1.00 | 1.00 | 1.00 | 3.00 | ▇▁▁▁▁ |
| TotRms\_AbvGrd | 0 | 1 | 6.44 | 1.54 | 3.00 | 5.00 | 6.00 | 7.00 | 15.00 | ▅▇▃▁▁ |
| Fireplaces | 0 | 1 | 0.60 | 0.65 | 0.00 | 0.00 | 1.00 | 1.00 | 4.00 | ▇▇▁▁▁ |
| Garage\_Cars | 0 | 1 | 1.77 | 0.76 | 0.00 | 1.00 | 2.00 | 2.00 | 5.00 | ▅▇▂▁▁ |
| Garage\_Area | 0 | 1 | 471.96 | 213.43 | 0.00 | 320.00 | 478.00 | 576.00 | 1488.00 | ▃▇▂▁▁ |
| Wood\_Deck\_SF | 0 | 1 | 93.52 | 127.71 | 0.00 | 0.00 | 0.00 | 168.00 | 1424.00 | ▇▁▁▁▁ |
| Open\_Porch\_SF | 0 | 1 | 48.17 | 69.51 | 0.00 | 0.00 | 27.00 | 72.00 | 742.00 | ▇▁▁▁▁ |
| Enclosed\_Porch | 0 | 1 | 23.02 | 60.59 | 0.00 | 0.00 | 0.00 | 0.00 | 584.00 | ▇▁▁▁▁ |
| Three\_season\_porch | 0 | 1 | 2.80 | 25.65 | 0.00 | 0.00 | 0.00 | 0.00 | 407.00 | ▇▁▁▁▁ |
| Screen\_Porch | 0 | 1 | 16.68 | 57.94 | 0.00 | 0.00 | 0.00 | 0.00 | 576.00 | ▇▁▁▁▁ |
| Pool\_Area | 0 | 1 | 1.34 | 27.74 | 0.00 | 0.00 | 0.00 | 0.00 | 800.00 | ▇▁▁▁▁ |
| Misc\_Val | 0 | 1 | 60.12 | 662.76 | 0.00 | 0.00 | 0.00 | 0.00 | 17000.00 | ▇▁▁▁▁ |
| Mo\_Sold | 0 | 1 | 6.19 | 2.70 | 1.00 | 4.00 | 6.00 | 8.00 | 12.00 | ▅▆▇▃▃ |
| Year\_Sold | 0 | 1 | 2007.75 | 1.30 | 2006.00 | 2007.00 | 2008.00 | 2009.00 | 2010.00 | ▇▇▇▇▃ |
| 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 | ▂▂▇▇▇ |

Select some important variables.

amesstudent2 = amesstudent %>% dplyr::select("Neighborhood", "Bldg\_Type", "House\_Style", "Overall\_Qual", "Year\_Built", "Year\_Remod\_Add", "Total\_Bsmt\_SF", "Full\_Bath", "Bedroom\_AbvGr", "Fireplaces", "Garage\_Type", "Garage\_Cars", "Above\_Median")

Structure and summary.

str(amesstudent2)

## tibble [2,053 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Neighborhood : chr [1:2053] "North\_Ames" "North\_Ames" "North\_Ames" "North\_Ames" ...  
## $ Bldg\_Type : chr [1:2053] "OneFam" "OneFam" "OneFam" "OneFam" ...  
## $ House\_Style : chr [1:2053] "One\_Story" "One\_Story" "One\_Story" "One\_Story" ...  
## $ Overall\_Qual : chr [1:2053] "Above\_Average" "Average" "Above\_Average" "Good" ...  
## $ Year\_Built : num [1:2053] 1960 1961 1958 1968 1997 ...  
## $ Year\_Remod\_Add: num [1:2053] 1960 1961 1958 1968 1998 ...  
## $ Total\_Bsmt\_SF : num [1:2053] 1080 882 1329 2110 928 ...  
## $ Full\_Bath : num [1:2053] 1 1 1 2 2 2 2 2 2 1 ...  
## $ Bedroom\_AbvGr : num [1:2053] 3 2 3 3 3 3 2 2 3 2 ...  
## $ Fireplaces : num [1:2053] 2 0 0 2 1 1 0 1 0 1 ...  
## $ Garage\_Type : chr [1:2053] "Attchd" "Attchd" "Attchd" "Attchd" ...  
## $ Garage\_Cars : num [1:2053] 2 1 1 2 2 2 2 2 2 2 ...  
## $ Above\_Median : chr [1:2053] "Yes" "No" "Yes" "Yes" ...

summary(amesstudent2)

## Neighborhood Bldg\_Type House\_Style Overall\_Qual   
## Length:2053 Length:2053 Length:2053 Length:2053   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Year\_Built Year\_Remod\_Add Total\_Bsmt\_SF Full\_Bath Bedroom\_AbvGr   
## Min. :1875 Min. :1950 Min. : 0 Min. :0.000 Min. :0.000   
## 1st Qu.:1953 1st Qu.:1965 1st Qu.: 793 1st Qu.:1.000 1st Qu.:2.000   
## Median :1972 Median :1993 Median : 988 Median :2.000 Median :3.000   
## Mean :1971 Mean :1984 Mean :1055 Mean :1.564 Mean :2.855   
## 3rd Qu.:2000 3rd Qu.:2004 3rd Qu.:1304 3rd Qu.:2.000 3rd Qu.:3.000   
## Max. :2010 Max. :2010 Max. :5095 Max. :4.000 Max. :6.000   
## Fireplaces Garage\_Type Garage\_Cars Above\_Median   
## Min. :0.000 Length:2053 Min. :0.000 Length:2053   
## 1st Qu.:0.000 Class :character 1st Qu.:1.000 Class :character   
## Median :1.000 Mode :character Median :2.000 Mode :character   
## Mean :0.603 Mean :1.774   
## 3rd Qu.:1.000 3rd Qu.:2.000   
## Max. :4.000 Max. :5.000

glimpse(amesstudent2)

## Rows: 2,053  
## Columns: 13  
## $ Neighborhood <chr> "North\_Ames", "North\_Ames", "North\_Ames", "North\_Ames",…  
## $ Bldg\_Type <chr> "OneFam", "OneFam", "OneFam", "OneFam", "OneFam", "OneF…  
## $ House\_Style <chr> "One\_Story", "One\_Story", "One\_Story", "One\_Story", "Tw…  
## $ Overall\_Qual <chr> "Above\_Average", "Average", "Above\_Average", "Good", "A…  
## $ Year\_Built <dbl> 1960, 1961, 1958, 1968, 1997, 1998, 1992, 1995, 1992, 1…  
## $ Year\_Remod\_Add <dbl> 1960, 1961, 1958, 1968, 1998, 1998, 1992, 1996, 2007, 1…  
## $ Total\_Bsmt\_SF <dbl> 1080, 882, 1329, 2110, 928, 926, 1280, 1595, 1168, 1300…  
## $ Full\_Bath <dbl> 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1, 2, 2, 1, 1, 1…  
## $ Bedroom\_AbvGr <dbl> 3, 2, 3, 3, 3, 3, 2, 2, 3, 2, 1, 4, 4, 1, 3, 3, 2, 3, 3…  
## $ Fireplaces <dbl> 2, 0, 0, 2, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1…  
## $ Garage\_Type <chr> "Attchd", "Attchd", "Attchd", "Attchd", "Attchd", "Attc…  
## $ Garage\_Cars <dbl> 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 1…  
## $ Above\_Median <chr> "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", …

skim(amesstudent2)

Data summary

|  |  |
| --- | --- |
| Name | amesstudent2 |
| Number of rows | 2053 |
| Number of columns | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 6 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Neighborhood | 0 | 1 | 6 | 39 | 0 | 28 | 0 |
| Bldg\_Type | 0 | 1 | 5 | 8 | 0 | 5 | 0 |
| House\_Style | 0 | 1 | 4 | 16 | 0 | 8 | 0 |
| Overall\_Qual | 0 | 1 | 4 | 14 | 0 | 10 | 0 |
| Garage\_Type | 0 | 1 | 6 | 19 | 0 | 7 | 0 |
| Above\_Median | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year\_Built | 0 | 1 | 1970.64 | 30.40 | 1875 | 1953 | 1972 | 2000 | 2010 | ▁▂▃▆▇ |
| Year\_Remod\_Add | 0 | 1 | 1984.08 | 20.96 | 1950 | 1965 | 1993 | 2004 | 2010 | ▅▂▂▃▇ |
| Total\_Bsmt\_SF | 0 | 1 | 1054.57 | 435.33 | 0 | 793 | 988 | 1304 | 5095 | ▇▇▁▁▁ |
| Full\_Bath | 0 | 1 | 1.56 | 0.55 | 0 | 1 | 2 | 2 | 4 | ▁▇▇▁▁ |
| Bedroom\_AbvGr | 0 | 1 | 2.86 | 0.82 | 0 | 2 | 3 | 3 | 6 | ▁▃▇▂▁ |
| Fireplaces | 0 | 1 | 0.60 | 0.65 | 0 | 0 | 1 | 1 | 4 | ▇▇▁▁▁ |
| Garage\_Cars | 0 | 1 | 1.77 | 0.76 | 0 | 1 | 2 | 2 | 5 | ▅▇▂▁▁ |

amesstudent2 = amesstudent2 %>% mutate(across(where(is.character),as\_factor))

Structure and summary.

str(amesstudent2)

## tibble [2,053 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Neighborhood : Factor w/ 28 levels "North\_Ames","Gilbert",..: 1 1 1 1 2 2 3 3 2 2 ...  
## $ Bldg\_Type : Factor w/ 5 levels "OneFam","TwnhsE",..: 1 1 1 1 1 1 2 2 1 1 ...  
## $ House\_Style : Factor w/ 8 levels "One\_Story","Two\_Story",..: 1 1 1 1 2 2 1 1 1 1 ...  
## $ Overall\_Qual : Factor w/ 10 levels "Above\_Average",..: 1 2 1 3 2 1 4 4 1 3 ...  
## $ Year\_Built : num [1:2053] 1960 1961 1958 1968 1997 ...  
## $ Year\_Remod\_Add: num [1:2053] 1960 1961 1958 1968 1998 ...  
## $ Total\_Bsmt\_SF : num [1:2053] 1080 882 1329 2110 928 ...  
## $ Full\_Bath : num [1:2053] 1 1 1 2 2 2 2 2 2 1 ...  
## $ Bedroom\_AbvGr : num [1:2053] 3 2 3 3 3 3 2 2 3 2 ...  
## $ Fireplaces : num [1:2053] 2 0 0 2 1 1 0 1 0 1 ...  
## $ Garage\_Type : Factor w/ 7 levels "Attchd","BuiltIn",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Garage\_Cars : num [1:2053] 2 1 1 2 2 2 2 2 2 2 ...  
## $ Above\_Median : Factor w/ 2 levels "Yes","No": 1 2 1 1 1 1 1 1 1 1 ...

summary(amesstudent2)

## Neighborhood Bldg\_Type House\_Style   
## North\_Ames : 327 OneFam :1706 One\_Story :1052   
## College\_Creek: 183 TwnhsE : 157 Two\_Story : 590   
## Old\_Town : 181 Twnhs : 67 One\_and\_Half\_Fin: 225   
## Edwards : 129 Duplex : 76 SLvl : 90   
## Somerset : 119 TwoFmCon: 47 SFoyer : 56   
## Gilbert : 109 Two\_and\_Half\_Unf: 19   
## (Other) :1005 (Other) : 21   
## Overall\_Qual Year\_Built Year\_Remod\_Add Total\_Bsmt\_SF   
## Average :587 Min. :1875 Min. :1950 Min. : 0   
## Above\_Average:518 1st Qu.:1953 1st Qu.:1965 1st Qu.: 793   
## Good :411 Median :1972 Median :1993 Median : 988   
## Very\_Good :237 Mean :1971 Mean :1984 Mean :1055   
## Below\_Average:169 3rd Qu.:2000 3rd Qu.:2004 3rd Qu.:1304   
## Excellent : 70 Max. :2010 Max. :2010 Max. :5095   
## (Other) : 61   
## Full\_Bath Bedroom\_AbvGr Fireplaces Garage\_Type   
## Min. :0.000 Min. :0.000 Min. :0.000 Attchd :1204   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:0.000 BuiltIn : 127   
## Median :2.000 Median :3.000 Median :1.000 Basment : 29   
## Mean :1.564 Mean :2.855 Mean :0.603 Detchd : 549   
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:1.000 No\_Garage : 108   
## Max. :4.000 Max. :6.000 Max. :4.000 CarPort : 15   
## More\_Than\_Two\_Types: 21   
## Garage\_Cars Above\_Median  
## Min. :0.000 Yes:1043   
## 1st Qu.:1.000 No :1010   
## Median :2.000   
## Mean :1.774   
## 3rd Qu.:2.000   
## Max. :5.000   
##

glimpse(amesstudent2)

## Rows: 2,053  
## Columns: 13  
## $ Neighborhood <fct> North\_Ames, North\_Ames, North\_Ames, North\_Ames, Gilbert…  
## $ Bldg\_Type <fct> OneFam, OneFam, OneFam, OneFam, OneFam, OneFam, TwnhsE,…  
## $ House\_Style <fct> One\_Story, One\_Story, One\_Story, One\_Story, Two\_Story, …  
## $ Overall\_Qual <fct> Above\_Average, Average, Above\_Average, Good, Average, A…  
## $ Year\_Built <dbl> 1960, 1961, 1958, 1968, 1997, 1998, 1992, 1995, 1992, 1…  
## $ Year\_Remod\_Add <dbl> 1960, 1961, 1958, 1968, 1998, 1998, 1992, 1996, 2007, 1…  
## $ Total\_Bsmt\_SF <dbl> 1080, 882, 1329, 2110, 928, 926, 1280, 1595, 1168, 1300…  
## $ Full\_Bath <dbl> 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1, 2, 2, 1, 1, 1…  
## $ Bedroom\_AbvGr <dbl> 3, 2, 3, 3, 3, 3, 2, 2, 3, 2, 1, 4, 4, 1, 3, 3, 2, 3, 3…  
## $ Fireplaces <dbl> 2, 0, 0, 2, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1…  
## $ Garage\_Type <fct> Attchd, Attchd, Attchd, Attchd, Attchd, Attchd, Attchd,…  
## $ Garage\_Cars <dbl> 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 1…  
## $ Above\_Median <fct> Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Y…

skim(amesstudent2)

Data summary

|  |  |
| --- | --- |
| Name | amesstudent2 |
| Number of rows | 2053 |
| Number of columns | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 6 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

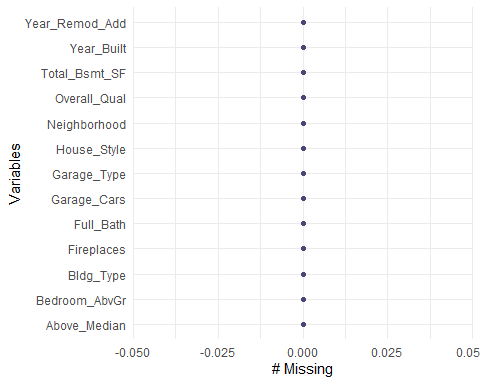
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Neighborhood | 0 | 1 | FALSE | 28 | Nor: 327, Col: 183, Old: 181, Edw: 129 |
| Bldg\_Type | 0 | 1 | FALSE | 5 | One: 1706, Twn: 157, Dup: 76, Twn: 67 |
| House\_Style | 0 | 1 | FALSE | 8 | One: 1052, Two: 590, One: 225, SLv: 90 |
| Overall\_Qual | 0 | 1 | FALSE | 10 | Ave: 587, Abo: 518, Goo: 411, Ver: 237 |
| Garage\_Type | 0 | 1 | FALSE | 7 | Att: 1204, Det: 549, Bui: 127, No\_: 108 |
| Above\_Median | 0 | 1 | FALSE | 2 | Yes: 1043, No: 1010 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year\_Built | 0 | 1 | 1970.64 | 30.40 | 1875 | 1953 | 1972 | 2000 | 2010 | ▁▂▃▆▇ |
| Year\_Remod\_Add | 0 | 1 | 1984.08 | 20.96 | 1950 | 1965 | 1993 | 2004 | 2010 | ▅▂▂▃▇ |
| Total\_Bsmt\_SF | 0 | 1 | 1054.57 | 435.33 | 0 | 793 | 988 | 1304 | 5095 | ▇▇▁▁▁ |
| Full\_Bath | 0 | 1 | 1.56 | 0.55 | 0 | 1 | 2 | 2 | 4 | ▁▇▇▁▁ |
| Bedroom\_AbvGr | 0 | 1 | 2.86 | 0.82 | 0 | 2 | 3 | 3 | 6 | ▁▃▇▂▁ |
| Fireplaces | 0 | 1 | 0.60 | 0.65 | 0 | 0 | 1 | 1 | 4 | ▇▇▁▁▁ |
| Garage\_Cars | 0 | 1 | 1.77 | 0.76 | 0 | 1 | 2 | 2 | 5 | ▅▇▂▁▁ |

Simple view of missingess. As you can see, no missing data.

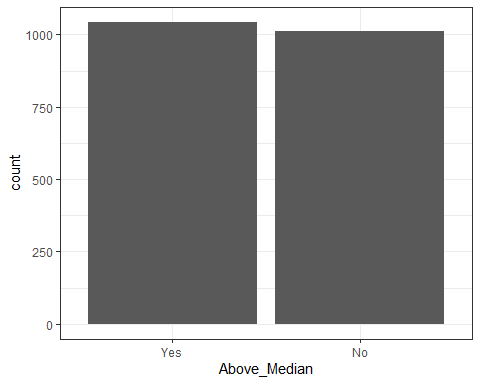
gg\_miss\_var(amesstudent2)



### Data Exploration

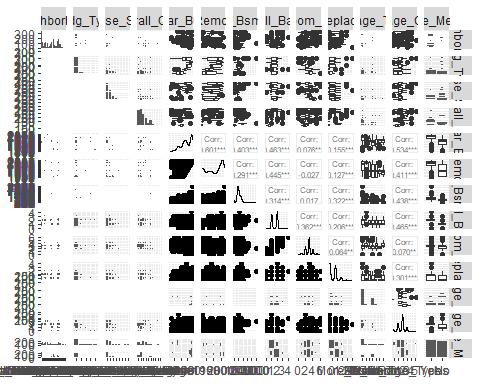
Begin exploring the data by looking at a plot of our response variable only (choose a histogram for a single quantitative variable)

ggplot(amesstudent2, aes(x=Above\_Median)) + geom\_bar() + theme\_bw()

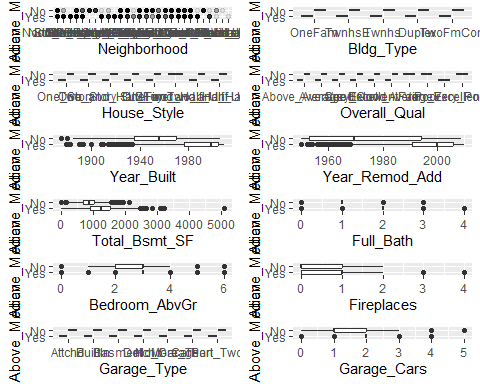
 Next we look at correlation. This is a logical step since almost all of our variables are quantitative.Year\_Built has highest correlation

Our Y (response) variable in this dataset is “Above\_Median”. Let’s look at ggpairs plot for visualization and correlation. There is a strong correlation between Year\_Remod\_Add and Above\_Median

ggpairs(amesstudent2,cardinality\_threshold = 28,upper = list(continuous=wrap("cor", size=2.0)),lower = list(combo="count"))

 Alternatively: Year\_Built, Year\_Remod\_Add, and Total\_Bmst\_SF all appear to impact Above\_Median.

p1 = ggplot(amesstudent2, aes(x=Neighborhood,y=Above\_Median)) + geom\_point(alpha = 0.1) #changing alpha is helpful when many points may overlap  
p2 = ggplot(amesstudent2, aes(x=Bldg\_Type,y=Above\_Median)) + geom\_boxplot()  
p3 = ggplot(amesstudent2, aes(x=House\_Style,y=Above\_Median)) + geom\_boxplot()  
p4 = ggplot(amesstudent2, aes(x=Overall\_Qual,y=Above\_Median)) + geom\_boxplot()  
p5 = ggplot(amesstudent2, aes(x=Year\_Built,y=Above\_Median)) + geom\_boxplot()  
p6 = ggplot(amesstudent2, aes(x=Year\_Remod\_Add,y=Above\_Median)) + geom\_boxplot()  
p7 = ggplot(amesstudent2, aes(x=Total\_Bsmt\_SF,y=Above\_Median)) + geom\_boxplot()  
p8 = ggplot(amesstudent2, aes(x=Full\_Bath,y=Above\_Median)) + geom\_boxplot()  
p9 = ggplot(amesstudent2, aes(x=Bedroom\_AbvGr,y=Above\_Median)) + geom\_boxplot()  
p10 = ggplot(amesstudent2, aes(x=Fireplaces,y=Above\_Median)) + geom\_boxplot()  
p11 = ggplot(amesstudent2, aes(x=Garage\_Type,y=Above\_Median)) + geom\_boxplot()  
p12 = ggplot(amesstudent2, aes(x=Garage\_Cars,y=Above\_Median)) + geom\_boxplot()  
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,ncol=2)

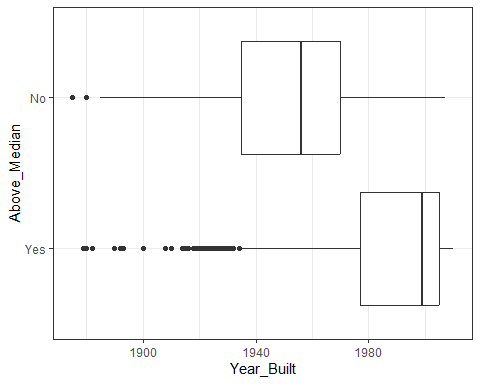
 ##Create a table to look at correlation between Neighborhood and Above\_Median. Can see there is a correlation between value of home and neighborhood.

amesstudent3<- table(amesstudent2$Neighborhood, amesstudent2$Above\_Median)  
 amesstudent3

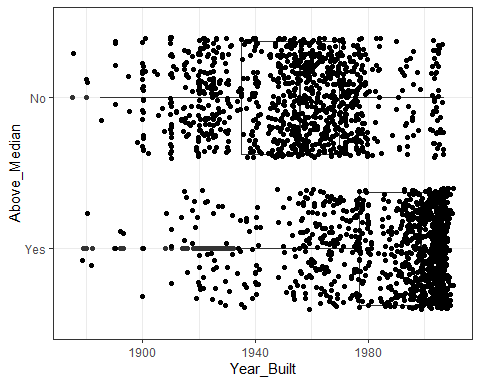
##   
## Yes No  
## North\_Ames 72 255  
## Gilbert 106 3  
## Stone\_Brook 39 0  
## Northwest\_Ames 76 19  
## Somerset 110 9  
## Briardale 0 22  
## Northpark\_Villa 0 15  
## Northridge\_Heights 104 1  
## Bloomington\_Heights 19 2  
## Northridge 50 0  
## Sawyer\_West 57 25  
## Sawyer 17 92  
## Greens 6 1  
## Old\_Town 18 163  
## Brookside 10 64  
## Iowa\_DOT\_and\_Rail\_Road 4 53  
## Clear\_Creek 24 7  
## South\_and\_West\_of\_Iowa\_State\_University 9 26  
## Edwards 20 109  
## College\_Creek 134 49  
## Crawford 56 21  
## Mitchell 36 43  
## Timberland 59 1  
## Meadow\_Village 0 24  
## Veenker 15 1  
## Blueste 1 4  
## Landmark 0 1  
## Green\_Hills 1 0

Closer look at “Year\_Built” and “Above\_Median”. This shows that the newer the home, the above\_median is higher.

ggplot(amesstudent2, aes(x=Year\_Built,y=Above\_Median)) + geom\_boxplot() + theme\_bw()

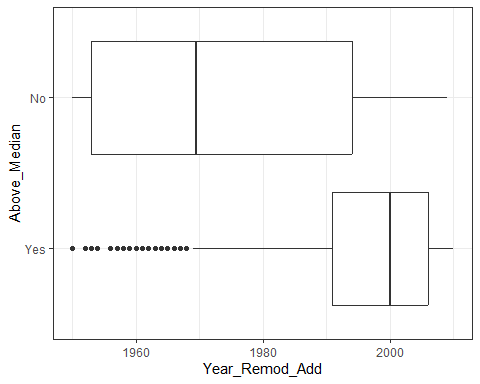


ggplot(amesstudent2, aes(x=Year\_Built,y=Above\_Median)) + geom\_boxplot() +   
 geom\_jitter() + theme\_bw()

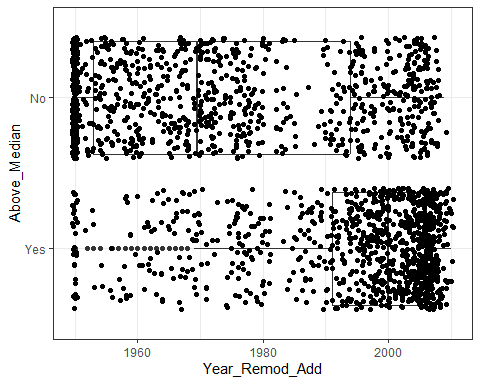


Closer look at “Year\_Remod\_Add” and “Above\_Median”. This shows that the newer the remodel, the above\_median is higher.

ggplot(amesstudent2, aes(x=Year\_Remod\_Add,y=Above\_Median)) + geom\_boxplot() + theme\_bw()

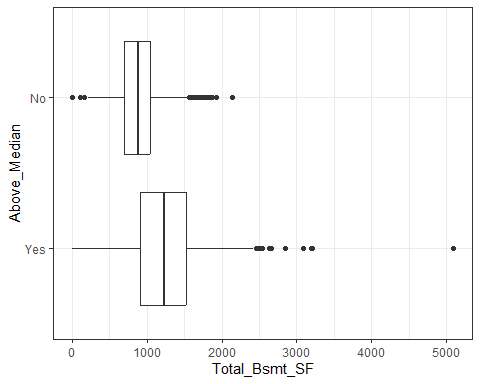


ggplot(amesstudent2, aes(x=Year\_Remod\_Add,y=Above\_Median)) + geom\_boxplot() +   
 geom\_jitter() + theme\_bw()

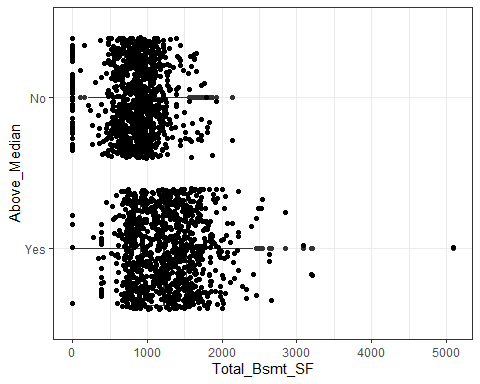


Closer look at “Total\_Bsmt\_SF” and “Above\_Median”. This shows that the more basement square footage, the above\_median is higher.

ggplot(amesstudent2, aes(x=Total\_Bsmt\_SF,y=Above\_Median)) + geom\_boxplot() + theme\_bw()



ggplot(amesstudent2, aes(x=Total\_Bsmt\_SF,y=Above\_Median)) + geom\_boxplot() +   
 geom\_jitter() + theme\_bw()



##As seen above, Above\_Median is highly correlated to Neighborhood, Year\_Built, Year\_Remod\_Add, and Total\_Bsmt\_SF.