BAN 502 Project Phase#2

Prafull Pandey

2023-06-25

# Continuation of code from Phase#1.

## Sourcing Libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.1.0 ──  
## ✔ broom 1.0.4 ✔ rsample 1.1.1  
## ✔ dials 1.2.0 ✔ tune 1.1.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.3  
## ✔ modeldata 1.1.0 ✔ workflowsets 1.0.1  
## ✔ parsnip 1.1.0 ✔ yardstick 1.2.0  
## ✔ recipes 1.0.6   
## ── 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()  
## • Search for functions across packages at https://www.tidymodels.org/find/

library(GGally)

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

library(gridExtra) #used for a little fancy arranging 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(glmnet)

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loaded glmnet 4.1-7

library(skimr)  
library(ggcorrplot) #create an alternative to ggcorr plots  
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) #for the dw test

## 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)

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

library(VIM)

## Loading required package: colorspace  
## Loading required package: grid  
## The legacy packages maptools, rgdal, and rgeos, underpinning this package  
## will retire shortly. Please refer to R-spatial evolution reports on  
## https://r-spatial.org/r/2023/05/15/evolution4.html for details.  
## This package is now running under evolution status 0   
## 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)

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

library(UpSetR)   
library(rpart)

##   
## Attaching package: 'rpart'  
##   
## The following object is masked from 'package:dials':  
##   
## prune

library(RColorBrewer)   
library(rattle)

## Loading required package: bitops  
##   
## Attaching package: 'bitops'  
##   
## The following object is masked from 'package:Matrix':  
##   
## %&%  
##   
## Rattle: A free graphical interface for data science with R.  
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.  
##   
## Attaching package: 'rattle'  
##   
## The following object is masked from 'package:VIM':  
##   
## wine

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'lattice'  
##   
## The following object is masked from 'package:UpSetR':  
##   
## histogram  
##   
##   
## Attaching package: 'caret'  
##   
## The following objects are masked from 'package:yardstick':  
##   
## precision, recall, sensitivity, specificity  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(ranger)

##   
## Attaching package: 'ranger'  
##   
## The following object is masked from 'package:rattle':  
##   
## importance

library(vip)

##   
## Attaching package: 'vip'  
##   
## The following object is masked from 'package:utils':  
##   
## vi

library(randomForest)

## randomForest 4.7-1.1  
## Type rfNews() to see new features/changes/bug fixes.  
##   
## Attaching package: 'randomForest'  
##   
## The following object is masked from 'package:ranger':  
##   
## importance  
##   
## The following object is masked from 'package:rattle':  
##   
## importance  
##   
## The following object is masked from 'package:gridExtra':  
##   
## combine  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine  
##   
## The following object is masked from 'package:ggplot2':  
##   
## margin

library(e1071)

##   
## Attaching package: 'e1071'  
##   
## The following object is masked from 'package:tune':  
##   
## tune  
##   
## The following object is masked from 'package:rsample':  
##   
## permutations  
##   
## The following object is masked from 'package:parsnip':  
##   
## tune

library(ROCR)

## Read-in data

ames=read\_csv("ames\_student-1.csv")

## Rows: 2053 Columns: 81  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (47): MS\_SubClass, MS\_Zoning, Street, Alley, Lot\_Shape, Land\_Contour, Ut...  
## dbl (34): Lot\_Frontage, Lot\_Area, Year\_Built, Year\_Remod\_Add, Mas\_Vnr\_Area, ...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Changed variables selection based on comments received from Dr. Hill during Phase#1 analysis.

## Data preparation

str(ames)

## 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(ames)

## 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(ames)

## 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(ames)

Data summary

|  |  |
| --- | --- |
| Name | ames |
| 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 | ▂▂▇▇▇ |

ames = ames %>% mutate\_if(is.character, as\_factor)   
  
  
  
ames\_selected = ames %>% dplyr::select("Overall\_Qual","Neighborhood","BsmtFin\_Type\_1","Heating\_QC","Garage\_Finish","Year\_Built","Year\_Remod\_Add","Gr\_Liv\_Area","Full\_Bath","TotRms\_AbvGrd","Fireplaces","Garage\_Cars","Garage\_Area","Above\_Median"  
)

## Split data

skim(ames\_selected)

Data summary

|  |  |
| --- | --- |
| Name | ames\_selected |
| Number of rows | 2053 |
| Number of columns | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 6 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Overall\_Qual | 0 | 1 | FALSE | 10 | Ave: 587, Abo: 518, Goo: 411, Ver: 237 |
| Neighborhood | 0 | 1 | FALSE | 28 | Nor: 327, Col: 183, Old: 181, Edw: 129 |
| BsmtFin\_Type\_1 | 0 | 1 | FALSE | 7 | Unf: 602, GLQ: 578, ALQ: 298, Rec: 216 |
| Heating\_QC | 0 | 1 | FALSE | 5 | Exc: 1040, Typ: 618, Goo: 333, Fai: 61 |
| Garage\_Finish | 0 | 1 | FALSE | 4 | Unf: 872, RFn: 563, Fin: 509, No\_: 109 |
| 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 | ▅▂▂▃▇ |
| Gr\_Liv\_Area | 0 | 1 | 1498.56 | 487.84 | 480 | 1137 | 1447 | 1737 | 5095 | ▇▇▁▁▁ |
| Full\_Bath | 0 | 1 | 1.56 | 0.55 | 0 | 1 | 2 | 2 | 4 | ▁▇▇▁▁ |
| TotRms\_AbvGrd | 0 | 1 | 6.44 | 1.54 | 3 | 5 | 6 | 7 | 15 | ▅▇▃▁▁ |
| 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 | ▅▇▂▁▁ |
| Garage\_Area | 0 | 1 | 471.96 | 213.43 | 0 | 320 | 478 | 576 | 1488 | ▃▇▂▁▁ |

set.seed(123)  
data\_split = initial\_split(ames\_selected, prop = 0.70, strata = Above\_Median)  
train = training(data\_split)  
test = testing(data\_split)

# Logistic Regression Model

log\_ames\_model =   
 logistic\_reg(mode = "classification") %>%   
 set\_engine("glm")   
  
log\_ames\_recipe = recipe(Above\_Median ~ ., train) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
logreg\_ames\_wf = workflow() %>%  
 add\_recipe(log\_ames\_recipe) %>%   
 add\_model(log\_ames\_model)  
  
log\_ames\_fit = fit(logreg\_ames\_wf, train)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(log\_ames\_fit$fit$fit$fit)

##   
## Call:  
## stats::glm(formula = ..y ~ ., family = stats::binomial, data = data)  
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) 5.239e+01 2.190e+01  
## Year\_Built -1.300e-02 9.109e-03  
## Year\_Remod\_Add -8.564e-03 8.049e-03  
## Gr\_Liv\_Area -2.853e-03 6.406e-04  
## Full\_Bath -1.052e+00 2.897e-01  
## TotRms\_AbvGrd -1.169e-01 1.550e-01  
## Fireplaces -1.323e+00 2.248e-01  
## Garage\_Cars -1.040e+00 4.477e-01  
## Garage\_Area -1.016e-04 1.519e-03  
## Overall\_Qual\_Average 5.157e-01 3.034e-01  
## Overall\_Qual\_Good -1.403e+00 3.897e-01  
## Overall\_Qual\_Very\_Good -5.785e+00 1.440e+00  
## Overall\_Qual\_Excellent 4.489e-01 1.230e+00  
## Overall\_Qual\_Below\_Average 2.360e+00 9.633e-01  
## Overall\_Qual\_Fair 1.567e+01 1.766e+03  
## Overall\_Qual\_Poor 1.161e+01 3.912e+03  
## Overall\_Qual\_Very\_Excellent -1.395e+01 2.075e+03  
## Overall\_Qual\_Very\_Poor 1.296e+01 7.039e+03  
## Neighborhood\_Gilbert -1.951e+00 7.845e-01  
## Neighborhood\_Stone\_Brook -1.422e+01 1.657e+03  
## Neighborhood\_Northwest\_Ames -1.331e+00 5.599e-01  
## Neighborhood\_Somerset -1.491e+00 7.278e-01  
## Neighborhood\_Briardale 1.523e+01 2.864e+03  
## Neighborhood\_Northpark\_Villa 2.055e+01 2.896e+03  
## Neighborhood\_Northridge\_Heights -9.422e-01 1.232e+00  
## Neighborhood\_Bloomington\_Heights -1.716e+01 2.650e+03  
## Neighborhood\_Northridge -1.325e+01 1.317e+03  
## Neighborhood\_Sawyer\_West -1.312e+00 6.960e-01  
## Neighborhood\_Sawyer 4.302e-03 5.601e-01  
## Neighborhood\_Greens -1.415e+01 4.100e+03  
## Neighborhood\_Old\_Town 1.028e+00 7.150e-01  
## Neighborhood\_Brookside -1.365e+00 7.207e-01  
## Neighborhood\_Iowa\_DOT\_and\_Rail\_Road -7.777e-01 1.049e+00  
## Neighborhood\_Clear\_Creek -3.301e+00 9.807e-01  
## Neighborhood\_South\_and\_West\_of\_Iowa\_State\_University -3.571e-01 8.565e-01  
## Neighborhood\_Edwards -7.187e-02 5.363e-01  
## Neighborhood\_College\_Creek -1.215e+00 6.095e-01  
## Neighborhood\_Crawford -2.540e+00 6.570e-01  
## Neighborhood\_Mitchell -1.419e+00 5.699e-01  
## Neighborhood\_Timberland -2.728e+00 1.126e+00  
## Neighborhood\_Meadow\_Village 1.390e+01 2.332e+03  
## Neighborhood\_Veenker -1.836e-01 1.353e+00  
## Neighborhood\_Blueste 1.857e+00 1.685e+00  
## Neighborhood\_Landmark 1.829e+01 1.075e+04  
## Neighborhood\_Green\_Hills -2.373e+01 1.075e+04  
## BsmtFin\_Type\_1\_Rec 1.903e-01 4.901e-01  
## BsmtFin\_Type\_1\_ALQ -9.898e-01 4.573e-01  
## BsmtFin\_Type\_1\_GLQ -4.251e-01 5.178e-01  
## BsmtFin\_Type\_1\_Unf 1.026e+00 4.778e-01  
## BsmtFin\_Type\_1\_LwQ 1.856e-01 6.129e-01  
## BsmtFin\_Type\_1\_No\_Basement 4.377e+00 1.686e+00  
## Heating\_QC\_Typical 6.169e-02 6.567e-01  
## Heating\_QC\_Excellent -4.574e-01 6.592e-01  
## Heating\_QC\_Good -9.045e-01 6.861e-01  
## Heating\_QC\_Poor NA NA  
## Garage\_Finish\_Unf 1.340e+00 3.847e-01  
## Garage\_Finish\_RFn -1.297e-02 3.880e-01  
## Garage\_Finish\_No\_Garage -6.801e-01 9.611e-01  
## z value Pr(>|z|)   
## (Intercept) 2.392 0.016760 \*   
## Year\_Built -1.427 0.153542   
## Year\_Remod\_Add -1.064 0.287323   
## Gr\_Liv\_Area -4.454 8.43e-06 \*\*\*  
## Full\_Bath -3.632 0.000281 \*\*\*  
## TotRms\_AbvGrd -0.754 0.450673   
## Fireplaces -5.887 3.94e-09 \*\*\*  
## Garage\_Cars -2.322 0.020245 \*   
## Garage\_Area -0.067 0.946641   
## Overall\_Qual\_Average 1.700 0.089215 .   
## Overall\_Qual\_Good -3.601 0.000317 \*\*\*  
## Overall\_Qual\_Very\_Good -4.017 5.89e-05 \*\*\*  
## Overall\_Qual\_Excellent 0.365 0.715119   
## Overall\_Qual\_Below\_Average 2.450 0.014279 \*   
## Overall\_Qual\_Fair 0.009 0.992918   
## Overall\_Qual\_Poor 0.003 0.997632   
## Overall\_Qual\_Very\_Excellent -0.007 0.994637   
## Overall\_Qual\_Very\_Poor 0.002 0.998530   
## Neighborhood\_Gilbert -2.487 0.012899 \*   
## Neighborhood\_Stone\_Brook -0.009 0.993153   
## Neighborhood\_Northwest\_Ames -2.376 0.017479 \*   
## Neighborhood\_Somerset -2.049 0.040482 \*   
## Neighborhood\_Briardale 0.005 0.995756   
## Neighborhood\_Northpark\_Villa 0.007 0.994339   
## Neighborhood\_Northridge\_Heights -0.765 0.444250   
## Neighborhood\_Bloomington\_Heights -0.006 0.994836   
## Neighborhood\_Northridge -0.010 0.991972   
## Neighborhood\_Sawyer\_West -1.884 0.059506 .   
## Neighborhood\_Sawyer 0.008 0.993872   
## Neighborhood\_Greens -0.003 0.997245   
## Neighborhood\_Old\_Town 1.437 0.150585   
## Neighborhood\_Brookside -1.894 0.058278 .   
## Neighborhood\_Iowa\_DOT\_and\_Rail\_Road -0.742 0.458331   
## Neighborhood\_Clear\_Creek -3.366 0.000762 \*\*\*  
## Neighborhood\_South\_and\_West\_of\_Iowa\_State\_University -0.417 0.676685   
## Neighborhood\_Edwards -0.134 0.893395   
## Neighborhood\_College\_Creek -1.993 0.046264 \*   
## Neighborhood\_Crawford -3.867 0.000110 \*\*\*  
## Neighborhood\_Mitchell -2.490 0.012786 \*   
## Neighborhood\_Timberland -2.422 0.015448 \*   
## Neighborhood\_Meadow\_Village 0.006 0.995244   
## Neighborhood\_Veenker -0.136 0.892022   
## Neighborhood\_Blueste 1.102 0.270433   
## Neighborhood\_Landmark 0.002 0.998643   
## Neighborhood\_Green\_Hills -0.002 0.998240   
## BsmtFin\_Type\_1\_Rec 0.388 0.697853   
## BsmtFin\_Type\_1\_ALQ -2.165 0.030420 \*   
## BsmtFin\_Type\_1\_GLQ -0.821 0.411632   
## BsmtFin\_Type\_1\_Unf 2.148 0.031695 \*   
## BsmtFin\_Type\_1\_LwQ 0.303 0.761950   
## BsmtFin\_Type\_1\_No\_Basement 2.596 0.009433 \*\*   
## Heating\_QC\_Typical 0.094 0.925147   
## Heating\_QC\_Excellent -0.694 0.487757   
## Heating\_QC\_Good -1.318 0.187434   
## Heating\_QC\_Poor NA NA   
## Garage\_Finish\_Unf 3.483 0.000495 \*\*\*  
## Garage\_Finish\_RFn -0.033 0.973327   
## Garage\_Finish\_No\_Garage -0.708 0.479174   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1991.74 on 1436 degrees of freedom  
## Residual deviance: 474.39 on 1380 degrees of freedom  
## AIC: 588.39  
##   
## Number of Fisher Scoring iterations: 18

## Logistic regression model train prediction and accuracy

predictions = predict(log\_ames\_fit, train, type="prob")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

head(predictions)

## # A tibble: 6 × 2  
## .pred\_Yes .pred\_No  
## <dbl> <dbl>  
## 1 3.76e- 1 0.624  
## 2 5.66e- 3 0.994  
## 3 6.27e-10 1.00   
## 4 4.12e-10 1.00   
## 5 4.19e- 2 0.958  
## 6 2.32e- 1 0.768

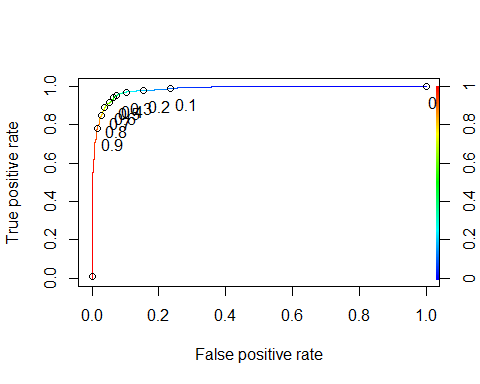
predictions = predict(log\_ames\_fit, train, type="prob")[1]

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(\*, "non-estim") has doubtful cases

head(predictions)

## # A tibble: 6 × 1  
## .pred\_Yes  
## <dbl>  
## 1 3.76e- 1  
## 2 5.66e- 3  
## 3 6.27e-10  
## 4 4.12e-10  
## 5 4.19e- 2  
## 6 2.32e- 1

ROCRpred = prediction(predictions, train$Above\_Median)   
  
  
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



as.numeric(performance(ROCRpred, "auc")@y.values)

## [1] 0.9831567

opt.cut = function(perf, pred){  
 cut.ind = mapply(FUN=function(x, y, p){  
 d = (x - 0)^2 + (y-1)^2  
 ind = which(d == min(d))  
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])  
 }, perf@x.values, perf@y.values, pred@cutoffs)  
}  
  
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.9479452  
## specificity 0.9349364  
## cutoff 0.4860099

## Check thresholds to evaluate accuracy for train

t1 = table(train$Above\_Median,predictions > 0.4860099)  
t1

##   
## FALSE TRUE  
## Yes 38 692  
## No 661 46

(t1[1,2]+t1[2,1])/nrow(train) # Calculate accuracy for train dataset

## [1] 0.9415449

## Logistic regression model test prodiction and accuracy

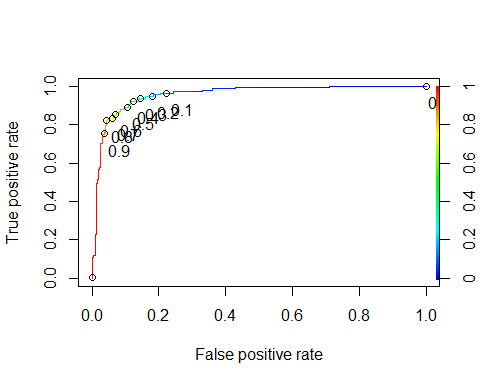
predictions = predict(log\_ames\_fit, test, type="prob")   
head(predictions)

## # A tibble: 6 × 2  
## .pred\_Yes .pred\_No  
## <dbl> <dbl>  
## 1 0.00101 9.99e- 1  
## 2 0.999 1.18e- 3  
## 3 1.00 2.05e-10  
## 4 1.00 7.27e- 9  
## 5 1.00 2.71e- 8  
## 6 0.988 1.24e- 2

predictions = predict(log\_ames\_fit, test, type="prob")[1]   
head(predictions)

## # A tibble: 6 × 1  
## .pred\_Yes  
## <dbl>  
## 1 0.00101  
## 2 0.999   
## 3 1.00   
## 4 1.00   
## 5 1.00   
## 6 0.988

ROCRpred = prediction(predictions, test$Above\_Median)   
  
  
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



as.numeric(performance(ROCRpred, "auc")@y.values)

## [1] 0.9570219

opt.cut = function(perf, pred){  
 cut.ind = mapply(FUN=function(x, y, p){  
 d = (x - 0)^2 + (y-1)^2  
 ind = which(d == min(d))  
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])  
 }, perf@x.values, perf@y.values, pred@cutoffs)  
}  
  
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.9201278  
## specificity 0.8811881  
## cutoff 0.4203653

## Check thresholds to evaluate accuracy for test

t1 = table(test$Above\_Median,predictions > 0.4203653)  
t1

##   
## FALSE TRUE  
## Yes 25 288  
## No 267 36

(t1[1,2]+t1[2,1])/nrow(test) # Accuracy for test dataset

## [1] 0.900974

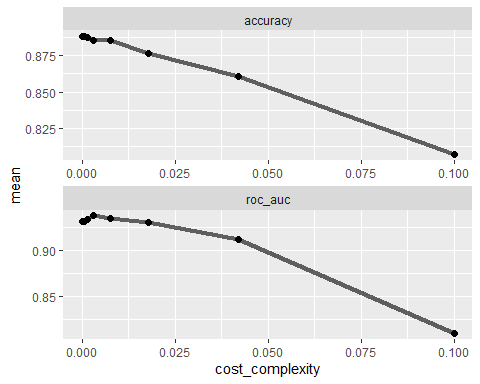
# Classification Tree Model

ames\_folds = vfold\_cv(train, v = 5)  
  
ames\_tree\_recipe = recipe(Above\_Median ~., train) %>%  
 step\_dummy(all\_nominal(),-all\_outcomes())  
  
ames\_tree\_model = decision\_tree(cost\_complexity = tune()) %>%   
 set\_engine("rpart", model = TRUE) %>%  
 set\_mode("classification")  
  
ames\_tree\_grid = grid\_regular(cost\_complexity(),  
 levels = 25)   
  
ames\_tree\_wflow =   
 workflow() %>%   
 add\_model(ames\_tree\_model) %>%   
 add\_recipe(ames\_tree\_recipe)  
  
ames\_tree\_res =   
 ames\_tree\_wflow %>%   
 tune\_grid(  
 resamples = ames\_folds,  
 grid = ames\_tree\_grid  
 )  
  
ames\_tree\_res

## # Tuning results  
## # 5-fold cross-validation   
## # A tibble: 5 × 4  
## splits id .metrics .notes   
## <list> <chr> <list> <list>   
## 1 <split [1149/288]> Fold1 <tibble [50 × 5]> <tibble [0 × 3]>  
## 2 <split [1149/288]> Fold2 <tibble [50 × 5]> <tibble [0 × 3]>  
## 3 <split [1150/287]> Fold3 <tibble [50 × 5]> <tibble [0 × 3]>  
## 4 <split [1150/287]> Fold4 <tibble [50 × 5]> <tibble [0 × 3]>  
## 5 <split [1150/287]> Fold5 <tibble [50 × 5]> <tibble [0 × 3]>

ames\_tree\_res %>%  
 collect\_metrics() %>%  
 ggplot(aes(cost\_complexity, mean)) +  
 geom\_line(size = 1.5, alpha = 0.6) +  
 geom\_point(size = 2) +  
 facet\_wrap(~ .metric, scales = "free", nrow = 2)

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



best\_tree = ames\_tree\_res %>%  
 select\_best("accuracy")  
  
best\_tree

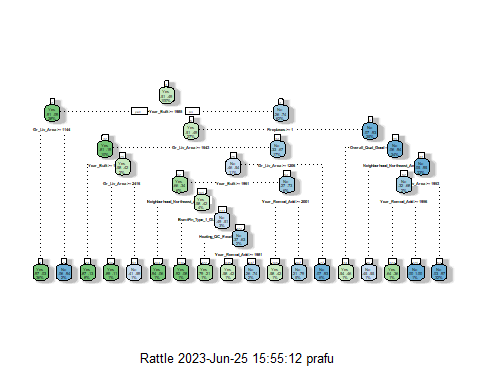
## # A tibble: 1 × 2  
## cost\_complexity .config   
## <dbl> <chr>   
## 1 0.0000000001 Preprocessor1\_Model01

final\_wf =   
 ames\_tree\_wflow %>%   
 finalize\_workflow(best\_tree)  
  
final\_fit = fit(final\_wf, train)  
  
tree = final\_fit %>%   
 pull\_workflow\_fit() %>%   
 pluck("fit")

## Warning: `pull\_workflow\_fit()` was deprecated in workflows 0.2.3.  
## ℹ Please use `extract\_fit\_parsnip()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

fancyRpartPlot(tree, tweak = 1.5)

## Warning: labs do not fit even at cex 0.15, there may be some overplotting



final\_fit$fit$fit$fit$cptable

## CP nsplit rel error xerror xstd  
## 1 0.6096181047 0 1.0000000 1.0396040 0.02680183  
## 2 0.0622347949 1 0.3903819 0.3917963 0.02115052  
## 3 0.0410183876 3 0.2659123 0.3111740 0.01930673  
## 4 0.0183875530 4 0.2248939 0.2432815 0.01740451  
## 5 0.0042432815 6 0.1881188 0.2121641 0.01639408  
## 6 0.0028288543 9 0.1753890 0.2135785 0.01644225  
## 7 0.0021216407 11 0.1697313 0.2277228 0.01691183  
## 8 0.0014144272 13 0.1654880 0.2333805 0.01709377  
## 9 0.0000000001 17 0.1598303 0.2503536 0.01762074

## Classification Tree Model for train

# Predicting training dataset  
  
treepred\_train = predict(final\_fit, train, type = "class")  
head(treepred\_train)

## # A tibble: 6 × 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 No   
## 3 No   
## 4 No   
## 5 No   
## 6 No

confusionMatrix(treepred\_train$.pred\_class,train$Above\_Median,positive="Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 674 57  
## No 56 650  
##   
## Accuracy : 0.9214   
## 95% CI : (0.9062, 0.9348)  
## No Information Rate : 0.508   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8427   
##   
## Mcnemar's Test P-Value : 1   
##   
## Sensitivity : 0.9233   
## Specificity : 0.9194   
## Pos Pred Value : 0.9220   
## Neg Pred Value : 0.9207   
## Prevalence : 0.5080   
## Detection Rate : 0.4690   
## Detection Prevalence : 0.5087   
## Balanced Accuracy : 0.9213   
##   
## 'Positive' Class : Yes   
##

## Classification Tree Model for test

#predicting testing dataset  
  
treepred\_test = predict(final\_fit, test, type = "class")  
head(treepred\_test)

## # A tibble: 6 × 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 Yes   
## 3 Yes   
## 4 Yes   
## 5 Yes   
## 6 Yes

confusionMatrix(treepred\_test$.pred\_class,test$Above\_Median,positive="Yes")

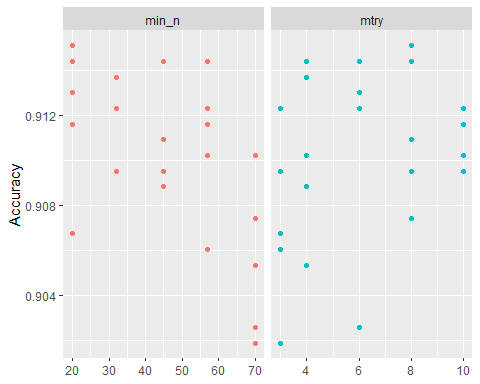
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 273 37  
## No 40 266  
##   
## Accuracy : 0.875   
## 95% CI : (0.8463, 0.9001)  
## No Information Rate : 0.5081   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.75   
##   
## Mcnemar's Test P-Value : 0.8197   
##   
## Sensitivity : 0.8722   
## Specificity : 0.8779   
## Pos Pred Value : 0.8806   
## Neg Pred Value : 0.8693   
## Prevalence : 0.5081   
## Detection Rate : 0.4432   
## Detection Prevalence : 0.5032   
## Balanced Accuracy : 0.8750   
##   
## 'Positive' Class : Yes   
##

# Random Forest Model

set.seed(123)  
rf\_folds = vfold\_cv(train, v = 5)  
  
ames\_rf\_recipe = recipe(Above\_Median ~., train) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
ames\_rf\_model = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
ames\_rf\_wflow =   
 workflow() %>%   
 add\_model(ames\_rf\_model) %>%   
 add\_recipe(ames\_rf\_recipe)  
  
  
rf\_grid = grid\_regular(  
 mtry(range = c(3, 10)),   
 min\_n(range = c(20, 70)),   
 levels = 5  
)  
  
ames\_rf\_res = tune\_grid(  
 ames\_rf\_wflow,  
 resamples = rf\_folds,  
 grid = rf\_grid   
)

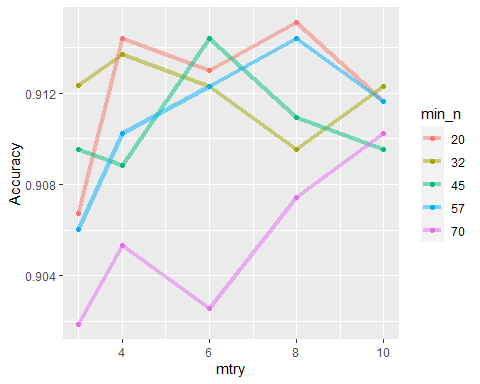
## Random Forest min\_n and mtry checks

ames\_rf\_res %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 dplyr::select(mean, min\_n, mtry) %>%  
 pivot\_longer(min\_n:mtry,  
 values\_to = "value",  
 names\_to = "parameter"  
 ) %>%  
 ggplot(aes(value, mean, color = parameter)) +  
 geom\_point(show.legend = FALSE) +  
 facet\_wrap(~parameter, scales = "free\_x") +  
 labs(x = NULL, y = "Accuracy")



## Random Forest Alternative view

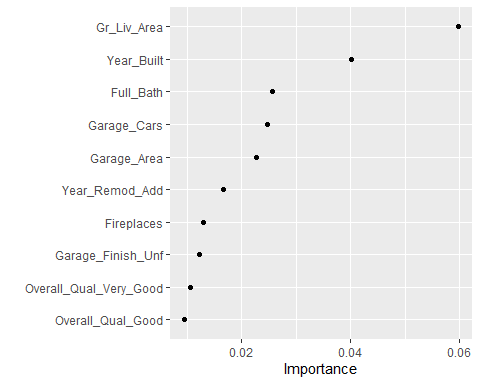
#Alternative view  
ames\_rf\_res %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 mutate(min\_n = factor(min\_n)) %>%  
 ggplot(aes(mtry, mean, color = min\_n)) +  
 geom\_line(alpha = 0.5, size = 1.5) +  
 geom\_point() +  
 labs(y = "Accuracy")



ames\_best\_rf = select\_best(ames\_rf\_res, "accuracy")  
  
ames\_final\_rf = finalize\_workflow(  
 ames\_rf\_wflow,  
 ames\_best\_rf  
)  
  
ames\_final\_rf

## ══ Workflow ════════════════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 1 Recipe Step  
##   
## • step\_dummy()  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## mtry = 8  
## trees = 100  
## min\_n = 20  
##   
## Engine-Specific Arguments:  
## importance = permutation  
##   
## Computational engine: ranger

ames\_final\_rf\_fit = fit(ames\_final\_rf, train)  
  
ames\_final\_rf\_fit %>% pull\_workflow\_fit() %>% vip(geom = "point")



## Random Forest train check

## predicting training dataset  
  
trainpredrf = predict(ames\_final\_rf\_fit, train)  
head(trainpredrf)

## # A tibble: 6 × 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 No   
## 3 No   
## 4 No   
## 5 No   
## 6 No

confusionMatrix(trainpredrf$.pred\_class, train$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 708 23  
## No 22 684  
##   
## Accuracy : 0.9687   
## 95% CI : (0.9583, 0.9771)  
## No Information Rate : 0.508   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9374   
##   
## Mcnemar's Test P-Value : 1   
##   
## Sensitivity : 0.9699   
## Specificity : 0.9675   
## Pos Pred Value : 0.9685   
## Neg Pred Value : 0.9688   
## Prevalence : 0.5080   
## Detection Rate : 0.4927   
## Detection Prevalence : 0.5087   
## Balanced Accuracy : 0.9687   
##   
## 'Positive' Class : Yes   
##

## Random Forest test check

## Predicting testing dataset  
  
testpredrf = predict(ames\_final\_rf\_fit, test)  
head(testpredrf)

## # A tibble: 6 × 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 Yes   
## 3 Yes   
## 4 Yes   
## 5 Yes   
## 6 Yes

confusionMatrix(testpredrf$.pred\_class, test$Above\_Median,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 271 20  
## No 42 283  
##   
## Accuracy : 0.8994   
## 95% CI : (0.8728, 0.922)  
## No Information Rate : 0.5081   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7989   
##   
## Mcnemar's Test P-Value : 0.007653   
##   
## Sensitivity : 0.8658   
## Specificity : 0.9340   
## Pos Pred Value : 0.9313   
## Neg Pred Value : 0.8708   
## Prevalence : 0.5081   
## Detection Rate : 0.4399   
## Detection Prevalence : 0.4724   
## Balanced Accuracy : 0.8999   
##   
## 'Positive' Class : Yes   
##