House Prices Kaggle

Load all the packages needed

library(tidymodels)

## Registered S3 method overwritten by 'tune':  
## method from   
## required\_pkgs.model\_spec parsnip

## -- Attaching packages -------------------------------------- tidymodels 0.1.3 --

## v broom 0.7.9 v recipes 0.1.16  
## v dials 0.0.10 v rsample 0.1.0   
## v dplyr 1.0.7 v tibble 3.1.4   
## v ggplot2 3.3.5 v tidyr 1.1.3   
## v infer 1.0.0 v tune 0.1.6   
## v modeldata 0.1.1 v workflows 0.2.3   
## v parsnip 0.1.7 v workflowsets 0.1.0   
## v purrr 0.3.4 v yardstick 0.0.8

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x purrr::discard() masks scales::discard()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x recipes::step() masks stats::step()  
## \* Use tidymodels\_prefer() to resolve common conflicts.

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v readr 2.0.1 v forcats 0.5.1  
## v stringr 1.4.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x readr::col\_factor() masks scales::col\_factor()  
## x purrr::discard() masks scales::discard()  
## x dplyr::filter() masks stats::filter()  
## x stringr::fixed() masks recipes::fixed()  
## x dplyr::lag() masks stats::lag()  
## x readr::spec() masks yardstick::spec()

library(skimr)  
library(parsnip)  
library(ranger)  
library(yardstick)  
library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-2

library(earth)

## Loading required package: Formula

## Loading required package: plotmo

## Loading required package: plotrix

##   
## Attaching package: 'plotrix'

## The following object is masked from 'package:scales':  
##   
## rescale

## Loading required package: TeachingDemos

Load the data sets

setwd("~/R projects/House Prices")  
train <- read\_csv("train.csv")

## Rows: 1460 Columns: 81

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...  
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...

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

test <- read\_csv("test.csv")

## Rows: 1459 Columns: 80

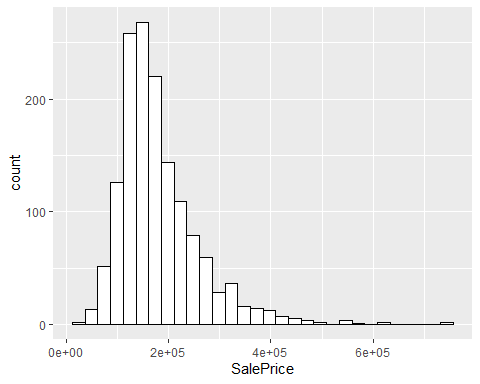
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...  
## dbl (37): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...

##   
## 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.

EDA Lets look at how the SalePrices are distributed

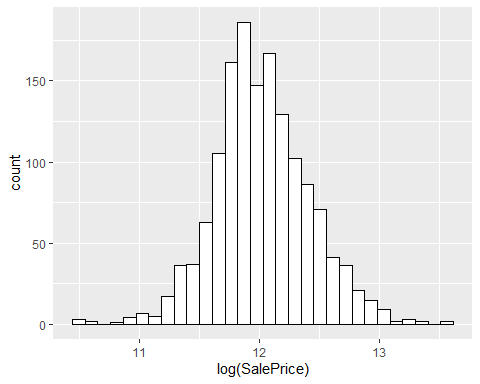
ggplot(train,  
 aes(x = SalePrice)) +  
 geom\_histogram(fill = "white", color = "black")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

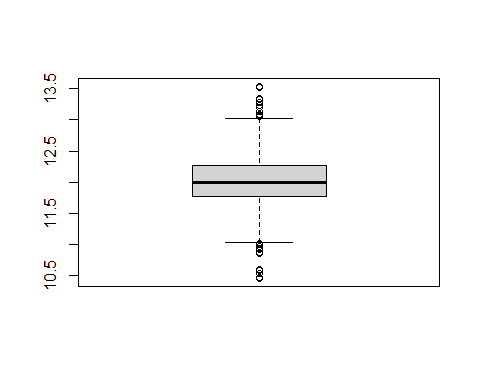
 I don’t like the shape of the distribution of SalePrice so lets try making it look more symmetric with a log transformation.

ggplot(train, aes(x = log(SalePrice))) +  
 geom\_histogram(fill = "white", color = "black")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 This is much better. We could try more kinds of transformations like inverse, power or BoxCox but I think this looks good enough. However, lets do the transformation and then remove observations that have a boxplot’s definition of outlier for SalePrice.

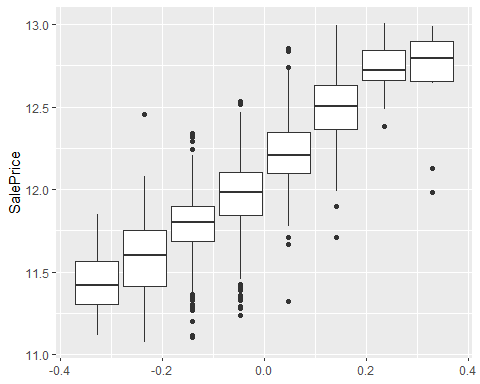
train$SalePrice <- log(train$SalePrice)  
sale\_upper <- boxplot(train$SalePrice)$stats[5]  
sale\_lower <- boxplot(train$SalePrice)$stats[1]



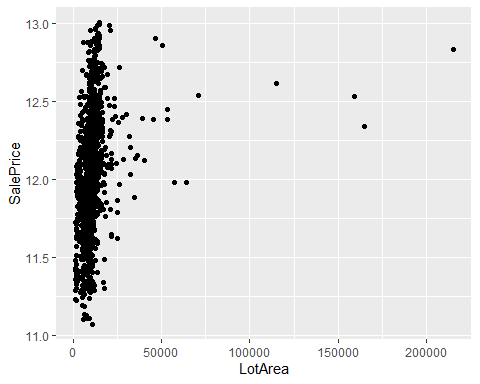
train <- train %>%  
 filter(SalePrice < sale\_upper, SalePrice > sale\_lower)

Lets check some other features and their relationship to SalePrice

ggplot(train,  
 aes(y = SalePrice,  
 group = OverallQual)) +  
 geom\_boxplot()



ggplot(train,  
 aes(y = SalePrice,  
 x = LotArea)) +  
 geom\_point()

 Data Cleaning

Now it is time to clean our datasets. For this I combine the train and test dataset. I am going to remove columns which have more that 25% missing values. Also I remove Street and Utilties because they have a very small variance.

test$SalePrice <- 0  
full <- rbind(test, train)  
skim(full)

Data summary

|  |  |
| --- | --- |
| Name | full |
| Number of rows | 2889 |
| Number of columns | 81 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 43 |
| numeric | 38 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| MSZoning | 4 | 1.00 | 2 | 7 | 0 | 5 | 0 |
| Street | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |
| Alley | 2694 | 0.07 | 4 | 4 | 0 | 2 | 0 |
| LotShape | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| LandContour | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| Utilities | 2 | 1.00 | 6 | 6 | 0 | 2 | 0 |
| LotConfig | 0 | 1.00 | 3 | 7 | 0 | 5 | 0 |
| LandSlope | 0 | 1.00 | 3 | 3 | 0 | 3 | 0 |
| Neighborhood | 0 | 1.00 | 5 | 7 | 0 | 25 | 0 |
| Condition1 | 0 | 1.00 | 4 | 6 | 0 | 9 | 0 |
| Condition2 | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| BldgType | 0 | 1.00 | 4 | 6 | 0 | 5 | 0 |
| HouseStyle | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| RoofStyle | 0 | 1.00 | 3 | 7 | 0 | 6 | 0 |
| RoofMatl | 0 | 1.00 | 4 | 7 | 0 | 8 | 0 |
| Exterior1st | 1 | 1.00 | 5 | 7 | 0 | 15 | 0 |
| Exterior2nd | 1 | 1.00 | 5 | 7 | 0 | 16 | 0 |
| MasVnrType | 23 | 0.99 | 4 | 7 | 0 | 4 | 0 |
| ExterQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| ExterCond | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| Foundation | 0 | 1.00 | 4 | 6 | 0 | 6 | 0 |
| BsmtQual | 79 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtCond | 80 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtExposure | 80 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtFinType1 | 77 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| BsmtFinType2 | 78 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| Heating | 0 | 1.00 | 4 | 5 | 0 | 6 | 0 |
| HeatingQC | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| CentralAir | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| Electrical | 1 | 1.00 | 3 | 5 | 0 | 5 | 0 |
| KitchenQual | 1 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| Functional | 2 | 1.00 | 3 | 4 | 0 | 7 | 0 |
| FireplaceQu | 1406 | 0.51 | 2 | 2 | 0 | 5 | 0 |
| GarageType | 148 | 0.95 | 6 | 7 | 0 | 6 | 0 |
| GarageFinish | 150 | 0.95 | 3 | 3 | 0 | 3 | 0 |
| GarageQual | 150 | 0.95 | 2 | 2 | 0 | 5 | 0 |
| GarageCond | 150 | 0.95 | 2 | 2 | 0 | 5 | 0 |
| PavedDrive | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| PoolQC | 2880 | 0.00 | 2 | 2 | 0 | 3 | 0 |
| Fence | 2325 | 0.20 | 4 | 5 | 0 | 4 | 0 |
| MiscFeature | 2786 | 0.04 | 4 | 4 | 0 | 4 | 0 |
| SaleType | 1 | 1.00 | 2 | 5 | 0 | 9 | 0 |
| SaleCondition | 0 | 1.00 | 6 | 7 | 0 | 6 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Id | 0 | 1.00 | 1467.00 | 843.50 | 1 | 737.00 | 1475.0 | 2197.00 | 2919.00 | ▇▇▇▇▇ |
| MSSubClass | 0 | 1.00 | 57.28 | 42.58 | 20 | 20.00 | 50.0 | 70.00 | 190.00 | ▇▅▂▁▁ |
| LotFrontage | 484 | 0.83 | 69.19 | 23.21 | 21 | 59.00 | 68.0 | 80.00 | 313.00 | ▇▃▁▁▁ |
| LotArea | 0 | 1.00 | 10138.54 | 7856.52 | 1300 | 7476.00 | 9452.0 | 11520.00 | 215245.00 | ▇▁▁▁▁ |
| OverallQual | 0 | 1.00 | 6.09 | 1.38 | 1 | 5.00 | 6.0 | 7.00 | 10.00 | ▁▁▇▅▁ |
| OverallCond | 0 | 1.00 | 5.57 | 1.11 | 1 | 5.00 | 5.0 | 6.00 | 9.00 | ▁▁▇▅▁ |
| YearBuilt | 0 | 1.00 | 1971.40 | 30.18 | 1872 | 1954.00 | 1973.0 | 2001.00 | 2010.00 | ▁▂▃▆▇ |
| YearRemodAdd | 0 | 1.00 | 1984.33 | 20.82 | 1950 | 1965.00 | 1993.0 | 2004.00 | 2010.00 | ▅▂▂▃▇ |
| MasVnrArea | 22 | 0.99 | 100.63 | 175.11 | 0 | 0.00 | 0.0 | 164.00 | 1600.00 | ▇▁▁▁▁ |
| BsmtFinSF1 | 1 | 1.00 | 439.64 | 451.19 | 0 | 0.00 | 370.0 | 732.25 | 5644.00 | ▇▁▁▁▁ |
| BsmtFinSF2 | 1 | 1.00 | 49.80 | 169.70 | 0 | 0.00 | 0.0 | 0.00 | 1526.00 | ▇▁▁▁▁ |
| BsmtUnfSF | 1 | 1.00 | 560.02 | 438.48 | 0 | 219.75 | 467.0 | 806.00 | 2336.00 | ▇▅▂▁▁ |
| TotalBsmtSF | 1 | 1.00 | 1049.46 | 433.68 | 0 | 793.00 | 989.5 | 1299.25 | 6110.00 | ▇▃▁▁▁ |
| 1stFlrSF | 0 | 1.00 | 1157.57 | 386.76 | 407 | 879.00 | 1082.0 | 1383.00 | 5095.00 | ▇▃▁▁▁ |
| 2ndFlrSF | 0 | 1.00 | 335.03 | 425.12 | 0 | 0.00 | 0.0 | 704.00 | 1862.00 | ▇▂▂▁▁ |
| LowQualFinSF | 0 | 1.00 | 4.55 | 45.42 | 0 | 0.00 | 0.0 | 0.00 | 1064.00 | ▇▁▁▁▁ |
| GrLivArea | 0 | 1.00 | 1497.15 | 492.11 | 407 | 1128.00 | 1444.0 | 1740.00 | 5642.00 | ▇▇▁▁▁ |
| BsmtFullBath | 2 | 1.00 | 0.43 | 0.53 | 0 | 0.00 | 0.0 | 1.00 | 3.00 | ▇▆▁▁▁ |
| BsmtHalfBath | 2 | 1.00 | 0.06 | 0.25 | 0 | 0.00 | 0.0 | 0.00 | 2.00 | ▇▁▁▁▁ |
| FullBath | 0 | 1.00 | 1.57 | 0.55 | 0 | 1.00 | 2.0 | 2.00 | 4.00 | ▁▇▇▁▁ |
| HalfBath | 0 | 1.00 | 0.38 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 2.00 | ▇▁▅▁▁ |
| BedroomAbvGr | 0 | 1.00 | 2.86 | 0.82 | 0 | 2.00 | 3.0 | 3.00 | 8.00 | ▁▇▂▁▁ |
| KitchenAbvGr | 0 | 1.00 | 1.04 | 0.21 | 0 | 1.00 | 1.0 | 1.00 | 3.00 | ▁▇▁▁▁ |
| TotRmsAbvGrd | 0 | 1.00 | 6.44 | 1.54 | 3 | 5.00 | 6.0 | 7.00 | 15.00 | ▅▇▃▁▁ |
| Fireplaces | 0 | 1.00 | 0.60 | 0.64 | 0 | 0.00 | 1.0 | 1.00 | 4.00 | ▇▇▁▁▁ |
| GarageYrBlt | 150 | 0.95 | 1978.07 | 25.54 | 1895 | 1960.00 | 1979.0 | 2002.00 | 2207.00 | ▂▇▁▁▁ |
| GarageCars | 1 | 1.00 | 1.77 | 0.75 | 0 | 1.00 | 2.0 | 2.00 | 5.00 | ▅▇▂▁▁ |
| GarageArea | 1 | 1.00 | 472.92 | 212.98 | 0 | 322.75 | 480.0 | 576.00 | 1488.00 | ▃▇▃▁▁ |
| WoodDeckSF | 0 | 1.00 | 93.73 | 126.27 | 0 | 0.00 | 0.0 | 168.00 | 1424.00 | ▇▁▁▁▁ |
| OpenPorchSF | 0 | 1.00 | 47.19 | 66.85 | 0 | 0.00 | 26.0 | 70.00 | 742.00 | ▇▁▁▁▁ |
| EnclosedPorch | 0 | 1.00 | 23.09 | 64.31 | 0 | 0.00 | 0.0 | 0.00 | 1012.00 | ▇▁▁▁▁ |
| 3SsnPorch | 0 | 1.00 | 2.63 | 25.32 | 0 | 0.00 | 0.0 | 0.00 | 508.00 | ▇▁▁▁▁ |
| ScreenPorch | 0 | 1.00 | 15.89 | 55.69 | 0 | 0.00 | 0.0 | 0.00 | 576.00 | ▇▁▁▁▁ |
| PoolArea | 0 | 1.00 | 2.08 | 34.34 | 0 | 0.00 | 0.0 | 0.00 | 800.00 | ▇▁▁▁▁ |
| MiscVal | 0 | 1.00 | 50.12 | 566.70 | 0 | 0.00 | 0.0 | 0.00 | 17000.00 | ▇▁▁▁▁ |
| MoSold | 0 | 1.00 | 6.22 | 2.71 | 1 | 4.00 | 6.0 | 8.00 | 12.00 | ▅▆▇▃▃ |
| YrSold | 0 | 1.00 | 2007.79 | 1.31 | 2006 | 2007.00 | 2008.0 | 2009.00 | 2010.00 | ▇▇▇▇▃ |
| SalePrice | 0 | 1.00 | 5.95 | 6.02 | 0 | 0.00 | 0.0 | 12.00 | 13.01 | ▇▁▁▁▇ |

remove\_cols <- colnames(full)[colSums(is.na(full)) > (0.25 \* nrow(full))]  
full <- full %>%  
 select(!remove\_cols)

## Note: Using an external vector in selections is ambiguous.  
## i Use `all\_of(remove\_cols)` instead of `remove\_cols` to silence this message.  
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.  
## This message is displayed once per session.

full <- full %>%  
 select(!c(Street, Utilities))  
  
train <- full %>%  
 filter(SalePrice != 0)  
test <- full %>%  
 filter(SalePrice == 0)

Now we split the training data

set.seed(135)  
data\_split <- initial\_split(train, strata = "SalePrice", prop = 0.80)  
  
house\_test <- testing(data\_split)  
house\_train <- training(data\_split)

Here I am doing all the preprocessing.

house\_rec <- recipe(SalePrice ~., data = house\_train) %>%  
 step\_impute\_mode(all\_nominal\_predictors()) %>%  
 step\_impute\_mean(all\_numeric\_predictors()) %>%  
 update\_role(Id, new\_role = "ID") %>%  
 step\_dummy(all\_nominal\_predictors()) %>%  
 step\_impute\_median(all\_predictors()) %>%  
 step\_BoxCox(all\_numeric\_predictors()) %>%  
 step\_nzv(all\_predictors()) %>%  
 step\_normalize(all\_numeric\_predictors())

Modeling

For some of the models I will be tuning the hyperparameters and will be doing so using a 5-fold crossvalidation.

set.seed(123)  
folds <- vfold\_cv(house\_train, v = 5)

Random Forest Model - Model

rf\_mod <-  
 rand\_forest(trees = 1000, mtry = tune(), min\_n = tune()) %>%  
 set\_mode("regression") %>%  
 set\_engine("ranger")

Random Forest - Workflow

rf\_wf <- workflow() %>%  
 add\_recipe(house\_rec) %>%  
 add\_model(rf\_mod)

Random Forest - Grid for tuning

rf\_grid <- grid\_regular(  
 mtry(range = c(10, 30)),  
 min\_n(range = c(2, 8)),  
 levels = 5  
)

Random Forest - Tune and update the parameters

set.seed(345)  
tune\_res <- tune\_grid(  
 rf\_wf,  
 resamples = folds,  
 grid = rf\_grid  
)

## ! Fold1: preprocessor 1/1, model 1/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 2/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 3/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 4/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 5/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 6/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 7/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 8/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 9/25 (predictions): There are new levels in a fa...

## ! Fold1: preprocessor 1/1, model 10/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 11/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 12/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 13/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 14/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 15/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 16/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 17/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 18/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 19/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 20/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 21/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 22/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 23/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 24/25 (predictions): There are new levels in a f...

## ! Fold1: preprocessor 1/1, model 25/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 1/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 2/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 3/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 4/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 5/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 6/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 7/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 8/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 9/25 (predictions): There are new levels in a fa...

## ! Fold2: preprocessor 1/1, model 10/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 11/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 12/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 13/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 14/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 15/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 16/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 17/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 18/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 19/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 20/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 21/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 22/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 23/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 24/25 (predictions): There are new levels in a f...

## ! Fold2: preprocessor 1/1, model 25/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 1/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 2/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 3/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 4/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 5/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 6/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 7/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 8/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 9/25 (predictions): There are new levels in a fa...

## ! Fold3: preprocessor 1/1, model 10/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 11/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 12/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 13/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 14/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 15/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 16/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 17/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 18/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 19/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 20/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 21/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 22/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 23/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 24/25 (predictions): There are new levels in a f...

## ! Fold3: preprocessor 1/1, model 25/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 1/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 2/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 3/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 4/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 5/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 6/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 7/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 8/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 9/25 (predictions): There are new levels in a fa...

## ! Fold4: preprocessor 1/1, model 10/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 11/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 12/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 13/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 14/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 15/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 16/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 17/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 18/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 19/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 20/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 21/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 22/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 23/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 24/25 (predictions): There are new levels in a f...

## ! Fold4: preprocessor 1/1, model 25/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 1/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 2/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 3/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 4/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 5/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 6/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 7/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 8/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 9/25 (predictions): There are new levels in a fa...

## ! Fold5: preprocessor 1/1, model 10/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 11/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 12/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 13/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 14/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 15/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 16/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 17/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 18/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 19/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 20/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 21/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 22/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 23/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 24/25 (predictions): There are new levels in a f...

## ! Fold5: preprocessor 1/1, model 25/25 (predictions): There are new levels in a f...

best\_rmse <- select\_best(tune\_res, "rmse")  
final\_rf <- finalize\_model(  
 rf\_mod,  
 best\_rmse  
)  
rf\_wf <- rf\_wf %>%  
 update\_model(final\_rf)

Random Forest - Fit the model

rf\_fit <- fit(rf\_wf, data = house\_train)

Random Forest - Predict and find the RMSE

rf\_pred <- rf\_fit %>%  
 predict(new\_data = house\_test)

## Warning: There are new levels in a factor: Artery

## Warning: There are new levels in a factor: Other

## Warning: There are new levels in a factor: Floor

rf\_pred <- bind\_cols(rf\_pred, house\_test %>% select(SalePrice))  
rf\_pred

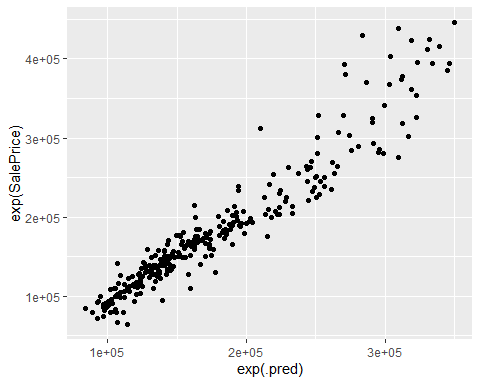
## # A tibble: 288 x 2  
## .pred SalePrice  
## <dbl> <dbl>  
## 1 12.5 12.6  
## 2 11.7 11.7  
## 3 11.8 11.8  
## 4 11.9 11.9  
## 5 11.8 11.8  
## 6 12.2 12.1  
## 7 12.1 12.0  
## 8 11.9 11.9  
## 9 11.7 11.6  
## 10 12.6 13.0  
## # ... with 278 more rows

rmse(rf\_pred, truth = exp(SalePrice), estimate = exp(.pred))

## # A tibble: 1 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 29029.

Random Forest - Plot the predictions against the actual SalePrice and see if there any postProcesses that can be done.

ggplot(data = rf\_pred, aes(x = exp(.pred), y = exp(SalePrice))) +  
 geom\_point()

 Does not look like any postProcessing is needed.

LASSO Model Same Process as with Random Forest

lasso\_model <- linear\_reg(penalty = tune(), mixture = tune()) %>%  
 set\_engine("glmnet")  
   
  
lasso\_wf <- workflow() %>%  
 add\_recipe(house\_rec) %>%  
 add\_model(lasso\_model)  
  
lasso\_grid <- grid\_regular(  
 penalty(), # The tune package has default values for penalty() and mixture() so no need to give them any  
 mixture(),  
 levels = 5  
)  
set.seed(345)  
tune\_res\_las <- tune\_grid(  
 lasso\_wf,  
 resamples = folds,  
 grid = lasso\_grid  
)

## ! Fold1: preprocessor 1/1, model 1/5 (predictions): There are new levels in a fac...

## ! Fold1: preprocessor 1/1, model 2/5 (predictions): There are new levels in a fac...

## ! Fold1: preprocessor 1/1, model 3/5 (predictions): There are new levels in a fac...

## ! Fold1: preprocessor 1/1, model 4/5 (predictions): There are new levels in a fac...

## ! Fold1: preprocessor 1/1, model 5/5 (predictions): There are new levels in a fac...

## ! Fold1: internal: A correlation computation is required, but `estimate` is const...

## ! Fold2: preprocessor 1/1, model 1/5 (predictions): There are new levels in a fac...

## ! Fold2: preprocessor 1/1, model 2/5 (predictions): There are new levels in a fac...

## ! Fold2: preprocessor 1/1, model 3/5 (predictions): There are new levels in a fac...

## ! Fold2: preprocessor 1/1, model 4/5 (predictions): There are new levels in a fac...

## ! Fold2: preprocessor 1/1, model 5/5 (predictions): There are new levels in a fac...

## ! Fold2: internal: A correlation computation is required, but `estimate` is const...

## ! Fold3: preprocessor 1/1, model 1/5 (predictions): There are new levels in a fac...

## ! Fold3: preprocessor 1/1, model 2/5 (predictions): There are new levels in a fac...

## ! Fold3: preprocessor 1/1, model 3/5 (predictions): There are new levels in a fac...

## ! Fold3: preprocessor 1/1, model 4/5 (predictions): There are new levels in a fac...

## ! Fold3: preprocessor 1/1, model 5/5 (predictions): There are new levels in a fac...

## ! Fold3: internal: A correlation computation is required, but `estimate` is const...

## ! Fold4: preprocessor 1/1, model 1/5 (predictions): There are new levels in a fac...

## ! Fold4: preprocessor 1/1, model 2/5 (predictions): There are new levels in a fac...

## ! Fold4: preprocessor 1/1, model 3/5 (predictions): There are new levels in a fac...

## ! Fold4: preprocessor 1/1, model 4/5 (predictions): There are new levels in a fac...

## ! Fold4: preprocessor 1/1, model 5/5 (predictions): There are new levels in a fac...

## ! Fold4: internal: A correlation computation is required, but `estimate` is const...

## ! Fold5: preprocessor 1/1, model 1/5 (predictions): There are new levels in a fac...

## ! Fold5: preprocessor 1/1, model 2/5 (predictions): There are new levels in a fac...

## ! Fold5: preprocessor 1/1, model 3/5 (predictions): There are new levels in a fac...

## ! Fold5: preprocessor 1/1, model 4/5 (predictions): There are new levels in a fac...

## ! Fold5: preprocessor 1/1, model 5/5 (predictions): There are new levels in a fac...

## ! Fold5: internal: A correlation computation is required, but `estimate` is const...

best\_rmse\_las <- select\_best(tune\_res\_las, "rmse")  
final\_las <- finalize\_model(  
 lasso\_model,  
 best\_rmse\_las  
)  
lasso\_wf <- lasso\_wf %>%  
 update\_model(final\_las)

LASSO - Fit the model

lasso\_fit <- fit(lasso\_wf, data = house\_train)

LASSO - Predict and evaluate using RMSE

lasso\_pred <- lasso\_fit %>%  
 predict(new\_data = house\_test)

## Warning: There are new levels in a factor: Artery

## Warning: There are new levels in a factor: Other

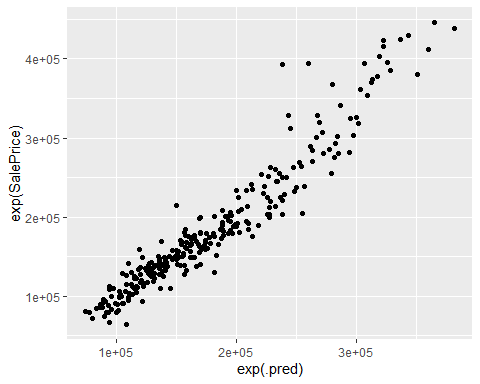
## Warning: There are new levels in a factor: Floor

lasso\_pred <- bind\_cols(lasso\_pred, house\_test %>% select(SalePrice))  
rmse(lasso\_pred, truth = exp(SalePrice), estimate = exp(.pred))

## # A tibble: 1 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 27997.

LASSO - Check for any potential postProcessing

ggplot(data = lasso\_pred, aes(x = exp(.pred), y = exp(SalePrice))) +  
 geom\_point()

 Looks good, no postProcessing required

MARS Model For the MARS model I am not going to use parameter tuning

mars\_model <- mars(mode = "regression") %>%  
 set\_engine("earth")  
  
mars\_wf <- workflow() %>%  
 add\_recipe(house\_rec) %>%  
 add\_model(mars\_model)  
  
mars\_fit <- fit(mars\_wf, data = house\_train)

MARS - Prediction and evaluate using RMSE

mars\_pred <- mars\_fit %>%  
 predict(new\_data = house\_test)

## Warning: There are new levels in a factor: Artery

## Warning: There are new levels in a factor: Other

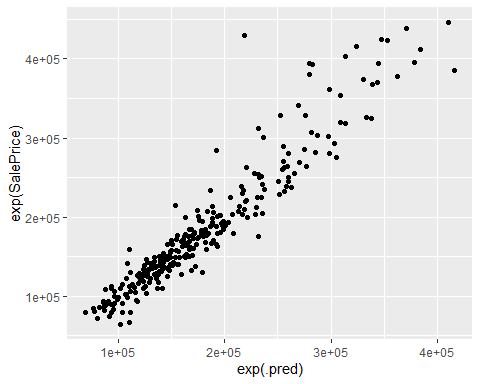
## Warning: There are new levels in a factor: Floor

mars\_pred <- bind\_cols(mars\_pred, house\_test %>% select(SalePrice))  
rmse(mars\_pred, truth = exp(SalePrice), estimate = exp(.pred))

## # A tibble: 1 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 rmse standard 28164.

MARS - Check for any postProcessing

ggplot(data = mars\_pred, aes(x = exp(.pred), y = exp(SalePrice))) +  
 geom\_point()

 Looks good, no postProcessing required.

Submission From the RMSE scores it looks like the LASSO model worked best so that is what I am going to use for the final prediction.

lasso\_final\_fit <- fit(lasso\_wf, data = train)  
lasso\_final\_pred <- predict(lasso\_final\_fit, new\_data = test)  
lasso\_final\_pred <- bind\_cols(test %>% select(Id), exp(lasso\_final\_pred))  
names(lasso\_final\_pred)[2] <- "SalePrice"  
  
write\_csv(lasso\_final\_pred, "tidymodels\_pred.csv")

This gave me a 0.13401 score on Kaggle.