Modeling housing prices

Chase Mathis

Import Data/libraries

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.8
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1

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

```
library(corrr)
library(tidymodels)
```

```
Registered S3 method overwritten by 'tune':
  method                   from
  required_pkgs.model_spec parsnip

-- Attaching packages -------------------------------------- tidymodels 0.1.4 --

v broom        0.7.11     v rsample      0.1.1
v dials        0.1.0      v tune         0.1.6
v infer        1.0.0      v workflows    0.2.4
v modeldata    0.1.1      v workflowsets 0.1.0
v parsnip      0.1.7      v yardstick    0.0.9
v recipes      0.2.0
```

```
-- Conflicts ----------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```r
library(knitr)
ggplot2::theme_set(ggplot2::theme_minimal(base_size = 16))
```

```r
test_house <- read_csv("data/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.
```

```r
train_house <- read_csv("data/train.csv")
```

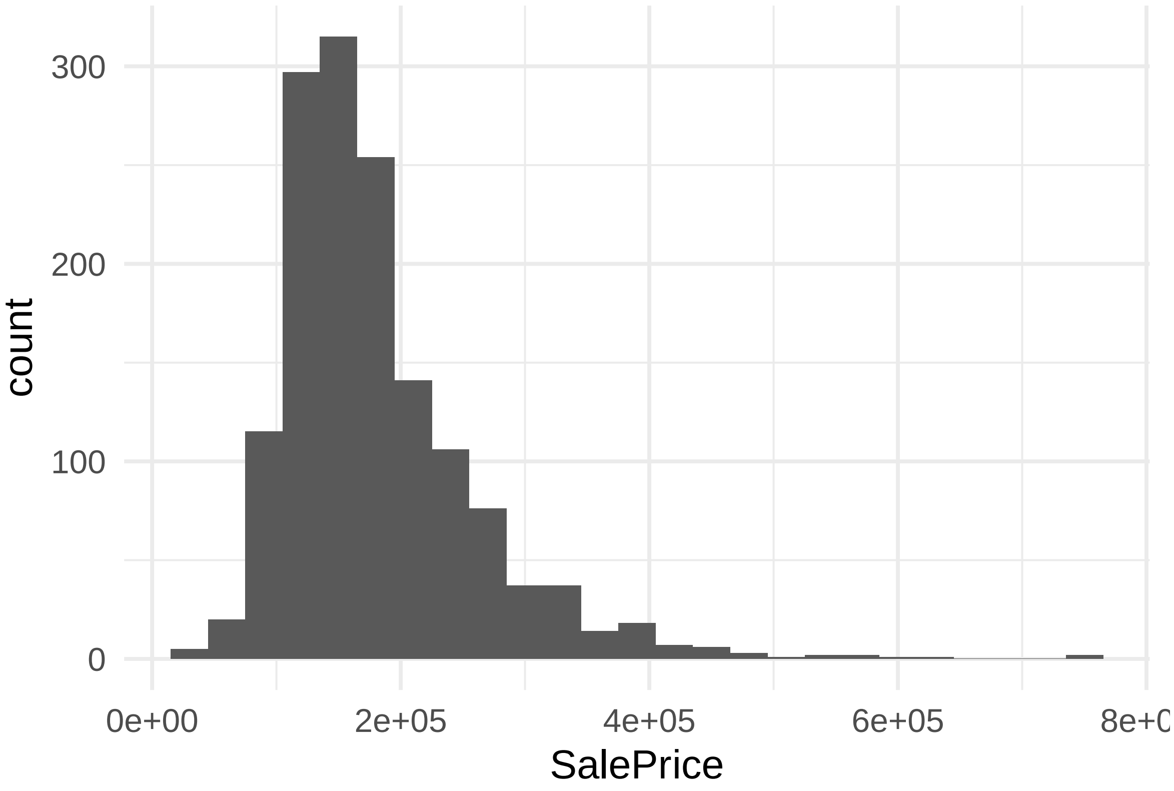
```
Rows: 1460 Columns: 81
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...

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.
```

Goal

Create a model predicting sales price.

EDA

```
train_house %>%
  ggplot(aes(x = SalePrice)) +
  geom_histogram(bins = 25)
```



```
summary(train_house$SalePrice) %>%
  tidy() %>%
  kable()
```

```
Warning: `tidy.summaryDefault()` is deprecated. Please use `skimr::skim()`
instead.
```

minimum	q1	median	mean	q3	maximum
34900	129975	163000	180921.2	214000	755000

Selecting variables

Looking at the columns and how many values are n/a, I am going to take out the top 5 in this table for these predictors have so many n/a values.

One character has only one unique factor, which causes an error in the regression. Also n/a values should be ommited.

```r
lst <- train_house %>%
  select(where(is.character)) %>%
  sapply(unique)


train_house <- train_house %>%
  select(!Utilities)

train_house %>%
  correlate() %>%
  select(term,SalePrice) %>%
  arrange(desc(SalePrice)) %>%
  top_n(5) %>%
  select(term)
```

```
Non-numeric variables removed from input: `MSZoning`, `Street`, `Alley`, `LotShape`, `LandCon
Correlation computed with
* Method: 'pearson'
* Missing treated using: 'pairwise.complete.obs'
Selecting by SalePrice
```

```
# A tibble: 5 x 1
  term
  <chr>
1 OverallQual
2 GrLivArea
3 GarageCars
4 GarageArea
5 TotalBsmtSF
```

Modeling

```
house_spec <- linear_reg() %>%
  set_engine("lm")
```

Model 1

```
house_rec_1 <- recipe(SalePrice ~ OverallQual + GrLivArea + GarageCars + GarageArea + TotalB
  step_corr(removals = TRUE) %>% # removes variables with correlation above 0.9
step_center(all_numeric_predictors()) %>% # mean center
step_dummy(all_nominal_predictors()) %>% # dummy coding
step_zv(all_predictors())# remove zero variance variables
```

```
house_wflow1 <- workflow() %>%
  add_model(house_spec) %>%
  add_recipe(house_rec_1)
```

```
house_fit_1 <- house_wflow1 %>%
  fit(train_house)

house_fit_1%>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	180921.196	1017.941	177.733	0.000
OverallQual	23635.007	1072.532	22.037	0.000
GrLivArea	45.346	2.489	18.218	0.000
GarageCars	14544.315	3022.681	4.812	0.000
GarageArea	17.133	10.468	1.637	0.102
TotalBsmtSF	31.501	2.904	10.848	0.000

Model 2

Only first 3 variables

```r
house_rec_2 <- recipe(SalePrice ~ OverallQual + GrLivArea + GarageCars, data = train_house) %
  step_corr(removals = TRUE) %>% # removes variables with correlation above 0.9
step_center(all_numeric_predictors()) %>% # mean center
step_dummy(all_nominal_predictors()) %>% # dummy coding
step_zv(all_predictors())# remove zero variance variables
```

```r
house_wflow2 <- workflow() %>%
  add_model(house_spec) %>%
  add_recipe(house_rec_2)
```

```r
house_fit_2 <- house_wflow2 %>%
  fit(train_house)

house_fit_2%>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	180921.196	1063.129	170.178	0
OverallQual	27104.826	1072.182	25.280	0
GrLivArea	50.674	2.552	19.859	0
GarageCars	21298.960	1807.065	11.786	0

Model 3... Interaction Terms?

```r
house_rec_3 <- recipe(SalePrice ~ OverallQual + GrLivArea + GarageCars + Id, data = train_hou
  update_role(Id, new_role = "id variable") %>%
  step_interact(terms = ~ OverallQual:GrLivArea +OverallQual:GarageCars + GrLivArea:GarageCa
  step_corr(removals = TRUE) %>% # removes variables with correlation above 0.9
step_center(all_numeric_predictors()) %>% # mean center
step_dummy(all_nominal_predictors()) %>% # dummy coding
step_zv(all_predictors())# remove zero variance variables

house_rec_3
```

```
Recipe

Inputs:
        role #variables
 id variable          1
     outcome          1
   predictor          3

Operations:

Interactions with OverallQual:GrLivArea + OverallQual:GarageCars + G...
Correlation filter on <none>
Centering for all_numeric_predictors()
Dummy variables from all_nominal_predictors()
Zero variance filter on all_predictors()
```

```r
house_wflow3 <- workflow() %>%
  add_model(house_spec) %>%
  add_recipe(house_rec_3)
```

```r
house_fit_3 <- house_wflow3 %>%
  fit(train_house)

house_fit_3%>%
  tidy() %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	180921.196	956.912	189.068	0.000
OverallQual	-1124.017	2715.953	-0.414	0.679
GrLivArea	11.712	8.331	1.406	0.160
GarageCars	-57113.838	6022.491	-9.483	0.000
OverallQual_x_GrLivArea	2.856	1.313	2.175	0.030
OverallQual_x_GarageCars	11799.301	1050.270	11.235	0.000
GrLivArea_x_GarageCars	8.280	3.068	2.699	0.007

Model Evaluation

```r
glance(house_fit_1) %>%
  select(r.squared,adj.r.squared,AIC)
```

```
# A tibble: 1 x 3
  r.squared adj.r.squared    AIC
      <dbl>         <dbl>  <dbl>
1     0.761         0.760 35012.
```

```r
glance(house_fit_2) %>%
  select(r.squared,adj.r.squared,AIC)
```

```
# A tibble: 1 x 3
  r.squared adj.r.squared    AIC
      <dbl>         <dbl>  <dbl>
1     0.739         0.739 35137.
```

```r
glance(house_fit_3) %>%
  select(r.squared,adj.r.squared,AIC)
```

```
# A tibble: 1 x 3
  r.squared adj.r.squared    AIC
      <dbl>         <dbl>  <dbl>
1     0.789         0.788 34832.
```

Model 3 has the highest adj.r.squared of 0.788 and the lowest AIC. We will choose this model.

```r
house_aug <- augment(house_fit_3, new_data = test_house) %>%
  select(Id, .pred) %>%
  rename(SalePrice = .pred) %>%
  mutate(SalePrice = if_else(is.na(SalePrice), mean(SalePrice, na.rm = TRUE), SalePrice))

write_csv(house_aug, "houseprices_files/submit.csv")
```

Issue with this row as Garage Cars have an n/a value. As we have already modeled our data using the test data, we will return the na value with the mean sale price.