Math 218: Project Proposal

Team Haydoja

Payoja Adhikari and Hayden Hunskor

11/13/2022

```
library(tidyverse)
library(dplyr)
library(tidyr)
```

Data description

We picked the Ames Housing dataset compiled by Dean De Cock, which is publicly available on Kaggle. The dataset includes extensive information on the sale of residential properties in Ames, Iowa from 2006 to 2010. It contains 1460 observations with 79 variables on various house specifications. The variables are a mix of continuous, categorical and discrete types. Here is a link to the dataset itself which includes a description of all the data fields:

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Define variables you want to explore

Rather than list out all 79 variables from our dataset, we selected a few that we found particularly interesting and listed them below. That said, we are interested in starting out with models that include all available predictors and applying various shrinkage techniques.

Categorical Variable:

- Street: Type of road access to property (Gravel/Paved)
- Utilities: Type of Utilities available
- Neighborhood: Physical locations within Ames city limits
- Heating: Type of heating
- CentralAir: Central air conditioning (Yes/No)
- PavedDrive: Paved driveway
- GarageFinish: Interior finish of the garage
- SaleType: Type of sale
- SaleCondition: Condition of sale

Continuous Variable:

- SalePrice: Price of house
- PoolArea: Pool area in square feet
- Lot Area: Lot size in square feet
- TotalBsmtSF: Total square feet of basement area
- Garage Area :Size of garage in square feet
- GrLivArea: Above ground living area square feet

Discrete Variable

- FullBath: Number of full bathrooms above ground
- HalfBath: Number of half bathrooms above ground
- BedroomAbvGr: Number of bedrooms above ground
- KitchensAbvGr: Number of kitchens above ground
- FirePlaces: Number of fireplaces
- YrBuilt: Original construction date
- YrSold: Year house was sold

EDA

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   Id = col_double(),
##   MSSubClass = col_double(),
##   LotFrontage = col_double(),
##   LotArea = col_double(),
##   OverallQual = col_double(),
##   OverallCond = col_double(),
##   YearBuilt = col_double(),
##   YearRemodAdd = col_double(),
##   MasVnrArea = col_double(),
##   BsmtFinSF1 = col_double(),
##   BsmtFinSF2 = col_double(),
##   BsmtUnfSF = col_double(),
##   TotalBsmtSF = col_double(),
##   `1stFlrSF` = col_double(),
##   `2ndFlrSF` = col_double(),
##   LowQualFinSF = col_double(),
##   GrLivArea = col_double(),
```

```
##   BsmtFullBath = col_double(),
##   BsmtHalfBath = col_double(),
##   FullBath = col_double()
##   # ... with 18 more columns
## )
## i Use `spec()` for the full column specifications.
```

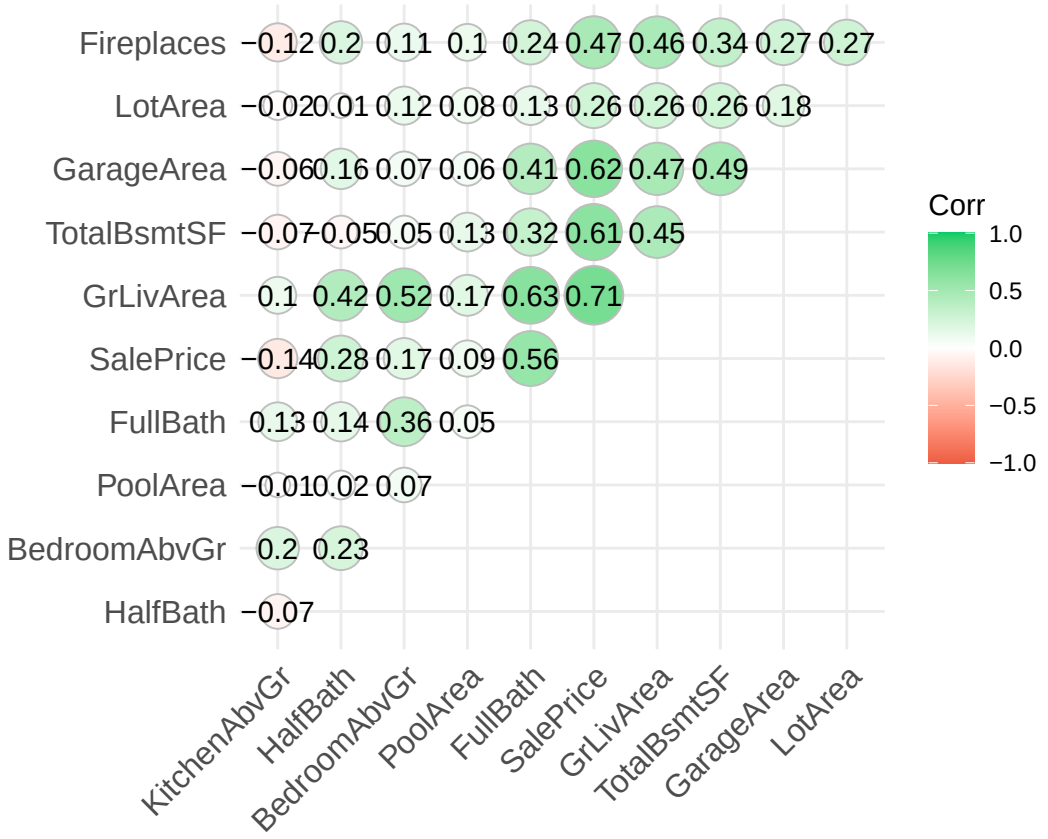
```
#find variables that are numeric
numeric_var<-names(train)[which(sapply(train, is.numeric))]
#find variables that are categorical
categorical_var<- names(train)[-which(sapply(train, is.numeric))]
#numeric variable that we want to work with
var<- numeric_var[-c(1:3,5:6,8,10:12,16,18:19,27,35:36)]
#summary table for numeric variables. The table includes some variable not mentioned earlier
train %>%
  select(var)%>%
  summarise_all(list(mean = mean, std = sd), na.rm=TRUE)%>%
  pivot_longer(cols = everything(),
               names_to = c('col', '.value'),
               names_sep = '_')%>%
print(n=34)
```

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

```
## # A tibble: 23 x 3
##    col                mean       std
##    <chr>             <dbl>     <dbl>
##  1 LotArea        10517.     9981.
##  2 YearBuilt       1971.       30.2
##  3 MasVnrArea       104.      181.
##  4 TotalBsmtSF     1057.      439.
##  5 1stFlrSF        1163.      387.
##  6 2ndFlrSF         347.      437.
##  7 GrLivArea       1515.      525.
##  8 FullBath           1.57      0.551
##  9 HalfBath           0.383     0.503
## 10 BedroomAbvGr       2.87      0.816
## 11 KitchenAbvGr       1.05      0.220
## 12 TotRmsAbvGrd       6.52      1.63
## 13 Fireplaces         0.613     0.645
## 14 GarageYrBlt     1979.       24.7
## 15 GarageArea       473.      214.
## 16 WoodDeckSF        94.2     125.
## 17 OpenPorchSF       46.7      66.3
## 18 EnclosedPorch     22.0      61.1
## 19 3SsnPorch          3.41     29.3
## 20 ScreenPorch       15.1      55.8
## 21 PoolArea           2.76     40.2
## 22 YrSold          2008.        1.33
## 23 SalePrice     180921.    79443.
```

Correlation between housing prices and some of the discrete and categorical data:

```
library("ggcorrplot")
correlation<-train%>%
  select(SalePrice, PoolArea,LotArea, TotalBsmtSF, GarageArea,GrLivArea, FullBath,HalfBath,BedroomAbvGr
corr_mtx <- cor(correlation, use = "pairwise.complete.obs")
ggcorrplot(corr_mtx, # input the correlation matrix
           hc.order = TRUE,
           type = "upper",
           lab = TRUE,
           method = "circle",
           colors = c("tomato2", "white", "springgreen3"))
```



As seen from above, housing prices seem to be the most correlated (positively) with above ground living area (in sq. ft), total basement area(in sq. ft) and garage area (in sq. ft). The price is also moderately correlated with number of full bathrooms.

Potential Research Questions

How do housing prices differ in Ames, Iowa based on the different housing features?

Which property data fields are the most accurate predictors of a house's sale price?