

Machine Learning Project

Michal Heydel

29 November 2017

```
setwd("C:/Users/Michal/Documents/01- Master Degree/GitHub/ST443-Project-group9/Housing price data")
getwd()
```

```
## [1] "C:/Users/Michal/Documents/01- Master Degree/GitHub/ST443-Project-group9/Housing price data"
```

```
train = read.csv("train.csv", row.names = "Id", stringsAsFactors=FALSE)
testing_kaggle = read.csv("test.csv", row.names = "Id", stringsAsFactors=FALSE)
```

```
#combining train and test data for quicker data prep
```

```
testing_kaggle$SalePrice <- NA
```

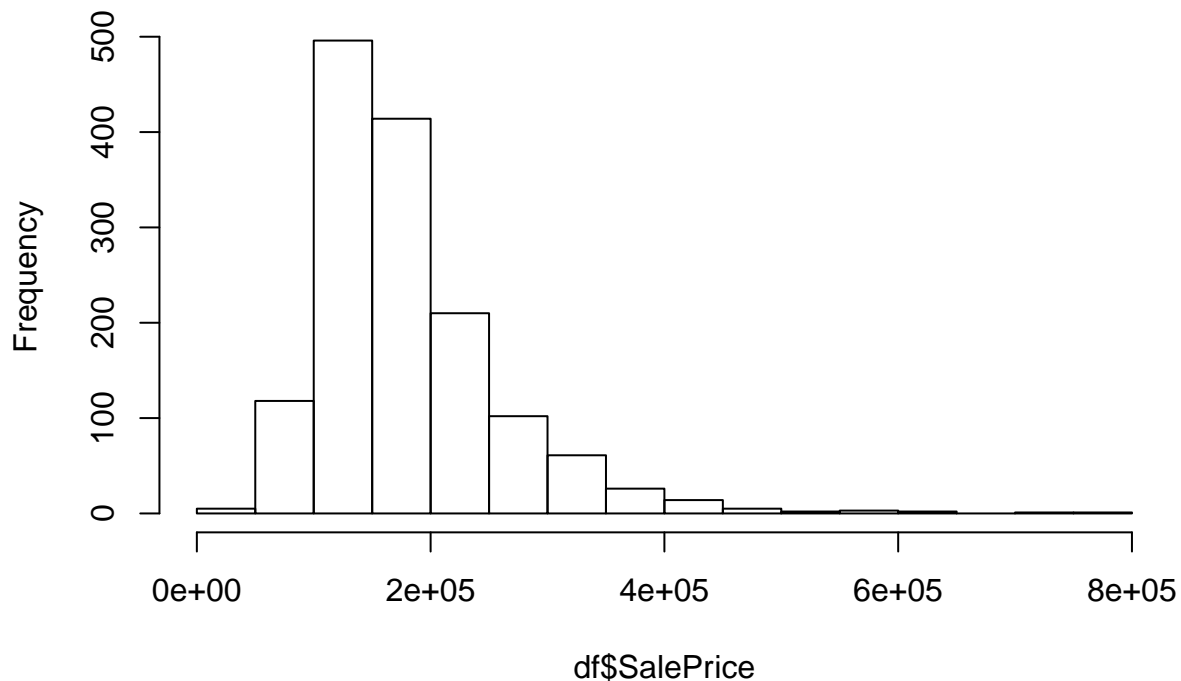
```
train$isTrain <- 1
```

```
testing_kaggle$isTrain <- 0
```

```
df <- rbind(train,testing_kaggle)
```

```
hist(df$SalePrice)
```

Histogram of df\$SalePrice



```
colSums(sapply(df, is.na))
```

```
##    MSSubClass    MSZoning  LotFrontage    LotArea    Street
##         0         4         486         0         0
##    Alley    LotShape  LandContour  Utilities    LotConfig
```

```
##      2721      0      0      2      0
##      LandSlope Neighborhood Condition1 Condition2 BldgType
##      0      0      0      0      0
##      HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
##      0      0      0      0      0
##      RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
##      0      0      1      1      24
##      MasVnrArea ExterQual ExterCond Foundation BsmtQual
##      23      0      0      0      81
##      BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
##      82      82      79      1      80
##      BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC
##      1      1      1      0      0
##      CentralAir Electrical X1stFlrSF X2ndFlrSF LowQualFinSF
##      0      1      0      0      0
##      GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
##      0      2      2      0      0
##      BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
##      0      0      1      0      2
##      Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish
##      0      1420      157      159      159
##      GarageCars GarageArea GarageQual GarageCond PavedDrive
##      1      1      159      159      0
##      WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch
##      0      0      0      0      0
##      PoolArea PoolQC Fence MiscFeature MiscVal
##      0      2909      2348      2814      0
##      MoSold YrSold SaleType SaleCondition SalePrice
##      0      0      1      0      1459
##      isTrain
##      0
```

```
for(i in colnames(df[,sapply(df, is.character)])){
  df[,i][which(is.na(df[,i]))] <- "None"
}
```

```
colSums(sapply(df, is.na))
```

```
##      MSSubClass MSZoning LotFrontage LotArea Street
##      0      0      486      0      0
##      Alley LotShape LandContour Utilities LotConfig
##      0      0      0      0      0
##      LandSlope Neighborhood Condition1 Condition2 BldgType
##      0      0      0      0      0
##      HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
##      0      0      0      0      0
##      RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
##      0      0      0      0      0
##      MasVnrArea ExterQual ExterCond Foundation BsmtQual
##      23      0      0      0      0
##      BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
##      0      0      0      1      0
##      BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC
##      1      1      1      0      0
```

```
##      CentralAir      Electrical      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##           0           0           0           0           0
##      GrLivArea      BsmtFullBath      BsmtHalfBath      FullBath      HalfBath
##           0           2           2           0           0
##      BedroomAbvGr      KitchenAbvGr      KitchenQual      TotRmsAbvGrd      Functional
##           0           0           0           0           0
##      Fireplaces      FireplaceQu      GarageType      GarageYrBlt      GarageFinish
##           0           0           0           159           0
##      GarageCars      GarageArea      GarageQual      GarageCond      PavedDrive
##           1           1           0           0           0
##      WoodDeckSF      OpenPorchSF      EnclosedPorch      X3SsnPorch      ScreenPorch
##           0           0           0           0           0
##      PoolArea      PoolQC      Fence      MiscFeature      MiscVal
##           0           0           0           0           0
##      MoSold      YrSold      SaleType      SaleCondition      SalePrice
##           0           0           0           0           1459
##      isTrain
##           0
```

```
df$LotFrontage[which(is.na(df$LotFrontage))] <- mean(df$LotFrontage,na.rm = T)
```

```
df$MasVnrArea[which(is.na(df$MasVnrArea))] <- mean(df$LotFrontage,na.rm = T)
```

```
x = c("BsmtFinSF1","BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF", "BsmtFullBath", "BsmtHalfBath", "GarageYrBlt")
```

```
for(i in x){
```

```
  df[,i][which(is.na(df[,i]))] <- 0
}
```

```
colSums(sapply(df, is.na))
```

```
##      MSSubClass      MSZoning      LotFrontage      LotArea      Street
##           0           0           0           0           0
##      Alley      LotShape      LandContour      Utilities      LotConfig
##           0           0           0           0           0
##      LandSlope      Neighborhood      Condition1      Condition2      BldgType
##           0           0           0           0           0
##      HouseStyle      OverallQual      OverallCond      YearBuilt      YearRemodAdd
##           0           0           0           0           0
##      RoofStyle      RoofMatl      Exterior1st      Exterior2nd      MasVnrType
##           0           0           0           0           0
##      MasVnrArea      ExterQual      ExterCond      Foundation      BsmtQual
##           0           0           0           0           0
##      BsmtCond      BsmtExposure      BsmtFinType1      BsmtFinSF1      BsmtFinType2
##           0           0           0           0           0
##      BsmtFinSF2      BsmtUnfSF      TotalBsmtSF      Heating      HeatingQC
##           0           0           0           0           0
##      CentralAir      Electrical      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##           0           0           0           0           0
##      GrLivArea      BsmtFullBath      BsmtHalfBath      FullBath      HalfBath
##           0           0           0           0           0
##      BedroomAbvGr      KitchenAbvGr      KitchenQual      TotRmsAbvGrd      Functional
##           0           0           0           0           0
##      Fireplaces      FireplaceQu      GarageType      GarageYrBlt      GarageFinish
```

```
##           0           0           0           0           0
##   GarageCars   GarageArea   GarageQual   GarageCond   PavedDrive
##           0           0           0           0           0
##   WoodDeckSF   OpenPorchSF   EnclosedPorch   X3SsnPorch   ScreenPorch
##           0           0           0           0           0
##   PoolArea     PoolQC       Fence     MiscFeature     MiscVal
##           0           0           0           0           0
##   MoSold       YrSold       SaleType   SaleCondition     SalePrice
##           0           0           0           0           1459
##   isTrain
##           0
```

```
for(i in colnames(df[,sapply(df, is.character)])){
  df[,i] <- as.factor(df[,i])
}
```

```
# These are also categorical Variables
df$MSSubClass <- as.factor(df$MSSubClass)
df$OverallCond <- as.factor(df$OverallCond)
df$OverallQual <- as.factor(df$OverallQual)
```

```
str(df)
```

```
## 'data.frame':   2919 obs. of  81 variables:
## $ MSSubClass    : Factor w/ 16 levels "20","30","40",...: 6 1 6 7 6 5 1 6 5 16 ...
## $ MSZoning      : Factor w/ 6 levels "C (all)","FV",...: 5 5 5 5 5 5 5 5 6 5 ...
## $ LotFrontage   : num  65 80 68 60 84 ...
## $ LotArea       : int   8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
## $ Street        : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 ...
## $ Alley         : Factor w/ 3 levels "Grvl","None",...: 2 2 2 2 2 2 2 2 2 ...
## $ LotShape      : Factor w/ 4 levels "IR1","IR2","IR3",...: 4 4 1 1 1 1 1 4 1 4 4 ...
## $ LandContour   : Factor w/ 4 levels "Bnk","HLS","Low",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Utilities     : Factor w/ 3 levels "AllPub","None",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ LotConfig     : Factor w/ 5 levels "Corner","CulDSac",...: 5 3 5 1 3 5 5 1 5 1 ...
## $ LandSlope     : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
## $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",...: 6 25 6 7 14 12 21 17 18 4 ...
## $ Condition1    : Factor w/ 9 levels "Artery","Feedr",...: 3 2 3 3 3 3 3 5 1 ...
## $ Condition2    : Factor w/ 8 levels "Artery","Feedr",...: 3 3 3 3 3 3 3 3 1 ...
## $ BldgType      : Factor w/ 5 levels "1Fam","2fmCon",...: 1 1 1 1 1 1 1 1 1 2 ...
## $ HouseStyle    : Factor w/ 8 levels "1.5Fin","1.5Unf",...: 6 3 6 6 6 1 3 6 1 2 ...
## $ OverallQual   : Factor w/ 10 levels "1","2","3","4",...: 7 6 7 7 8 5 8 7 7 5 ...
## $ OverallCond   : Factor w/ 9 levels "1","2","3","4",...: 5 8 5 5 5 5 5 6 5 6 ...
## $ YearBuilt     : int   2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
## $ YearRemodAdd  : int   2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
## $ RoofStyle     : Factor w/ 6 levels "Flat","Gable",...: 2 2 2 2 2 2 2 2 2 ...
## $ RoofMatl      : Factor w/ 8 levels "ClyTile","CompShg",...: 2 2 2 2 2 2 2 2 2 ...
## $ Exterior1st   : Factor w/ 16 levels "AsbShng","AsphShn",...: 14 9 14 15 14 14 14 7 4 9 ...
## $ Exterior2nd   : Factor w/ 17 levels "AsbShng","AsphShn",...: 15 9 15 17 15 15 15 7 17 9 ...
## $ MasVnrType    : Factor w/ 4 levels "BrkCmn","BrkFace",...: 2 3 2 3 2 3 4 4 3 3 ...
## $ MasVnrArea    : num   196 0 162 0 350 0 186 240 0 0 ...
## $ ExterQual     : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 4 3 4 3 4 3 4 4 4 ...
## $ ExterCond     : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 5 5 ...
## $ Foundation    : Factor w/ 6 levels "BrkTil","CBlock",...: 3 2 3 1 3 6 3 2 1 1 ...
## $ BsmtQual      : Factor w/ 5 levels "Ex","Fa","Gd",...: 3 3 3 5 3 3 1 3 5 5 ...
## $ BsmtCond      : Factor w/ 5 levels "Fa","Gd","None",...: 5 5 5 2 5 5 5 5 5 5 ...
```

```

## $ BsmtExposure : Factor w/ 5 levels "Av","Gd","Mn",...: 4 2 3 4 1 4 1 3 4 4 ...
## $ BsmtFinType1 : Factor w/ 7 levels "ALQ","BLQ","GLQ",...: 3 1 3 1 3 3 3 1 7 3 ...
## $ BsmtFinSF1 : num 706 978 486 216 655 ...
## $ BsmtFinType2 : Factor w/ 7 levels "ALQ","BLQ","GLQ",...: 7 7 7 7 7 7 7 2 7 7 ...
## $ BsmtFinSF2 : num 0 0 0 0 0 0 0 32 0 0 ...
## $ BsmtUnfSF : num 150 284 434 540 490 64 317 216 952 140 ...
## $ TotalBsmtSF : num 856 1262 920 756 1145 ...
## $ Heating : Factor w/ 6 levels "Floor","GasA",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ HeatingQC : Factor w/ 5 levels "Ex","Fa","Gd",...: 1 1 1 3 1 1 1 1 3 1 ...
## $ CentralAir : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
## $ Electrical : Factor w/ 6 levels "FuseA","FuseF",...: 6 6 6 6 6 6 6 6 6 2 6 ...
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
## $ BsmtFullBath : num 1 0 1 1 1 1 1 1 0 1 ...
## $ BsmtHalfBath : num 0 1 0 0 0 0 0 0 0 0 ...
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...
## $ KitchenQual : Factor w/ 5 levels "Ex","Fa","Gd",...: 3 5 3 3 3 5 3 5 5 5 ...
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...
## $ Functional : Factor w/ 8 levels "Maj1","Maj2",...: 8 8 8 8 8 8 8 8 3 8 ...
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...
## $ FireplaceQu : Factor w/ 6 levels "Ex","Fa","Gd",...: 4 6 6 3 6 4 3 6 6 6 ...
## $ GarageType : Factor w/ 7 levels "2Types","Attchd",...: 2 2 2 6 2 2 2 2 6 2 ...
## $ GarageYrBlt : num 2003 1976 2001 1998 2000 ...
## $ GarageFinish : Factor w/ 4 levels "Fin","None","RFn",...: 3 3 3 4 3 4 3 3 4 3 ...
## $ GarageCars : num 2 2 2 3 3 2 2 2 2 1 ...
## $ GarageArea : num 548 460 608 642 836 480 636 484 468 205 ...
## $ GarageQual : Factor w/ 6 levels "Ex","Fa","Gd",...: 6 6 6 6 6 6 6 6 6 2 3 ...
## $ GarageCond : Factor w/ 6 levels "Ex","Fa","Gd",...: 6 6 6 6 6 6 6 6 6 6 ...
## $ PavedDrive : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...
## $ EnclosedPorch : int 0 0 0 272 0 0 0 228 205 0 ...
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...
## $ PoolQC : Factor w/ 4 levels "Ex","Fa","Gd",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Fence : Factor w/ 5 levels "GdPrv","GdWo",...: 5 5 5 5 5 3 5 5 5 5 ...
## $ MiscFeature : Factor w/ 5 levels "Gar2","None",...: 2 2 2 2 2 4 2 4 2 2 ...
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...
## $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...
## $ YrSold : int 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
## $ SaleType : Factor w/ 10 levels "COD","Con","ConLD",...: 10 10 10 10 10 10 10 10 10 10 ...
## $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",...: 5 5 5 1 5 5 5 5 1 5 ...
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
## $ isTrain : num 1 1 1 1 1 1 1 1 1 1 ...

```