Rough

Sonaxy Mohanty

10/3/2022

Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
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(cowplot)
library(ggplot2)
library(GGally) ##ggcorr function
```

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

```
library(caTools) #hold-out validation
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(regclass)
```

```
## Loading required package: bestglm
## Loading required package: leaps
## Loading required package: VGAM
## Loading required package: stats4
## Loading required package: splines
## Loading required package: rpart
## Loading required package: 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:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
##
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
library(Metrics) #RMSE calculation
library(broom) #get the p-value of  the model
library(car) #ncvTest function
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:VGAM':
##
##     logit
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

General Data Prep

Read Data

```
# Convert all character data to factor
hd <- read.csv('housingData.csv', stringsAsFactors = TRUE) %>%

# creates new variables age, ageSinceRemodel, and ageofGarage and
  dplyr::mutate(age = YrSold - YearBuilt,
                ageSinceRemodel = YrSold - YearRemodAdd,
                ageofGarage = ifelse(is.na(GarageYrBlt), age, YrSold - GarageYrBlt)) %>%

# remove some columns used in the above calculations
  dplyr::select(!c(Id,YrSold ,
                   MoSold, YearBuilt, YearRemodAdd))


#str(hd)
```

Impute Missing Values with `PMM`

Make data set of `numeric` variables

```
hd.numericRaw <- hd %>%

  #selecting all the numeric data
  dplyr::select_if(is.numeric) %>%

  #converting the data frame to tibble
  as_tibble()
```

Make data set of `factor` variables

```
hd.factorRaw <- hd %>%

  #selecting all the numeric data
  dplyr::select_if(is.factor) %>%

  #converting the data frame to tibble
  as_tibble()
```

For each column with missing data, impute missing values with `PMM`

- Done with function `imputeWithPMM()` function
- Applies function via `dplyr` logic
- Note `seeImputation()` function to visualize the imputation from prior homework 4, not shown for simplicity in viewing

Create function to impute via `PMM`

```
imputeWithPMM <- function(colWithMissingData) {

  # Using the mice package
  #suppressMessages(library(mice))
  #?suppressMessages
  # Discover the missing rows
  isMissing <- is.na(colWithMissingData)

  # Create data frame to pass to PMM imputation function from mic package
  df <- data.frame(x       = rexp(length(colWithMissingData)), # meaningless x to help show variation
                   y       = colWithMissingData,
                   missing = isMissing)

  # imputation by PMM
  df[isMissing, "y"] <- mice.impute.pmm( df$y,
                                         !df$missing,
                                          df$x)

  return(df$y)
}
```

Apply PMM function to numeric data containing null values

```
# Data to store imputed values with PMM method
hd.Imputed <- hd

# Which columns has Na's?
colNamesWithNulls <- colnames(hd.numericRaw[ , colSums(is.na(hd.numericRaw)) != 0])
colNamesWithNulls
```

```
## [1] "LotFrontage" "MasVnrArea"  "GarageYrBlt"
```

```
numberOfColsWithNulls = length(colNamesWithNulls)


# For each of the numeric columns with null values
for (colWithNullsNum in 1:numberOfColsWithNulls) {

  # The name of the column with null values
  nameOfThisColumn <- colNamesWithNulls[colWithNullsNum]

  # Get the actual data of the column with nulls
  colWithNulls <- hd[, nameOfThisColumn]

  # Impute the missing values with PMM
  imputedValues <- imputeWithPMM(colWithNulls)

  # Now store the data in the original new frame
  hd.Imputed[, nameOfThisColumn] <- imputedValues

  # Save a visualization of the imputation
  pmmVisual <- seeImputation(data.frame(y = colWithNulls),
                             data.frame(y = imputedValues),
```

```
                              nameOfThisColumn )

  fileToSave = paste0('OutputPMM/Imputation_With_PMM_', nameOfThisColumn, '.pdf')
  print(paste0('For imputation results of ', nameOfThisColumn, ', see ', fileToSave))
  dir.create("OutputPMM/")
  ggsave(pmmVisual, filename = fileToSave,
         height = 11, width = 8.5)

  #hd.Imputed[!complete.cases(is.numeric(hd.Imputed)), ]

}
```

```
## [1] "For imputation results of LotFrontage, see OutputPMM/Imputation_With_PMM_LotFrontage.pdf"
```

```
## [1] "For imputation results of MasVnrArea, see OutputPMM/Imputation_With_PMM_MasVnrArea.pdf"
```

```
## [1] "For imputation results of GarageYrBlt, see OutputPMM/Imputation_With_PMM_GarageYrBlt.pdf"
```

Factor Level Collapse - Create `Other` Bin for Columns over 4 Unique Values

```
hd.Cleaned <- hd.Imputed # For final cleaned data

# Get list of factors and the number of unique values
factorCols <- as.data.frame(t(hd.factorRaw %>% summarise_all(n_distinct)))

# We are going to factor collapse factor columns with more than 4 columns
# So there will be 4 of the original, and 1 containing 'other'
# This is the threshold
factorThreshold = 4

# Get a list of the factors we are going to collapse
colsWithManyFactors <- rownames(factorCols %>% filter(V1 > factorThreshold))

# Show a summary of how many factors will be collapsed
numberOfColsWithManyFactors = length(colsWithManyFactors)
paste('Before cleaning, there are', numberOfColsWithManyFactors, 'factor columns with more than',
      factorThreshold, 'unique values')
```

```
## [1] "Before cleaning, there are 14 factor columns with more than 4 unique values"
```

```
# Collapse the affected factors in the original data (the one that already has imputation)

## for each factor column that we are about to collapse
for (collapsedColNum in 1:numberOfColsWithManyFactors) {

  # The name of the column with null values
  nameOfThisColumn <- colsWithManyFactors[collapsedColNum]

  # Get the actual data of the column with nulls
  colWithManyFactors <- hd[, nameOfThisColumn]

  # lumps all levels except for the n most frequent
  hd.Cleaned[, nameOfThisColumn] <- fct_lump_n(colWithManyFactors,
                                               n=factorThreshold)
}

# Check to see if the factor lumping worked
factorColsCleaned <- t(hd.Cleaned %>%
                       select_if(is.factor) %>%
                       summarise_all(n_distinct))
paste('After cleaning, there are', sum(factorColsCleaned > factorThreshold, na.rm = TRUE),
      "columns with more than", factorThreshold, "unique values (omitting NA's)")
```

```
## [1] "After cleaning, there are 14 columns with more than 4 unique values (omitting NA's)"
```

Remove Outliers from Numeric Data

- Since there are so many outliers, we are only going to remove some outliers
- If you count the number of outliers by column, the 75% of columns contain less than 50 outliers.
- However, some contain up to 200. Since remove ALL outliers would reduce the size of the data to less than 300 observations, we are removing up to 50 per column.

```
hd.CleanedNoOutliers <- hd.Cleaned

# Remove up to 75% of the outliers in the data set
# this is the 3rd quartile of number of outliers.
k_outliers = 50
numOutliers = c() # to store the number of outliers per column

theColNames <- colnames(hd.Cleaned)

for (colNum in 1:ncol(hd.Cleaned)) {

  theCol <- hd.Cleaned[, colNum]
  nrowBefore = length(theCol)
  colName <- theColNames[colNum]


  # Only consider numeric
  if (is.numeric(theCol)) {

    # Identify the outliers in the column
    # Source: https://www.geeksforgeeks.org/remove-outliers-from-data-set-in-r/
    columnOutliers <- boxplot.stats(hd.CleanedNoOutliers[, colNum])$out
    numOutliers <- c(numOutliers, length(columnOutliers))

    # Now remove k outliers from the column
    if (length(columnOutliers) < k_outliers) {

      hd.CleanedNoOutliers  <- hd.CleanedNoOutliers %>%

        # If this syntax looks weird, it is just referencing a column in the
        # data set using dplyr piping. See below for more info:
        # https://stackoverflow.com/questions/48062213/dplyr-using-column-names-as-function-arguments
        # https://stackoverflow.com/questions/72673381/column-names-as-variables-in-dplyr-select-v-filt
        filter( !( get({{colName}}) %in% columnOutliers ) )
    }
  }
}
paste0('Of the columns with outliers, removed up to 75th percentile of num. outliers.')
```

```
## [1] "Of the columns with outliers, removed up to 75th percentile of num. outliers."
```

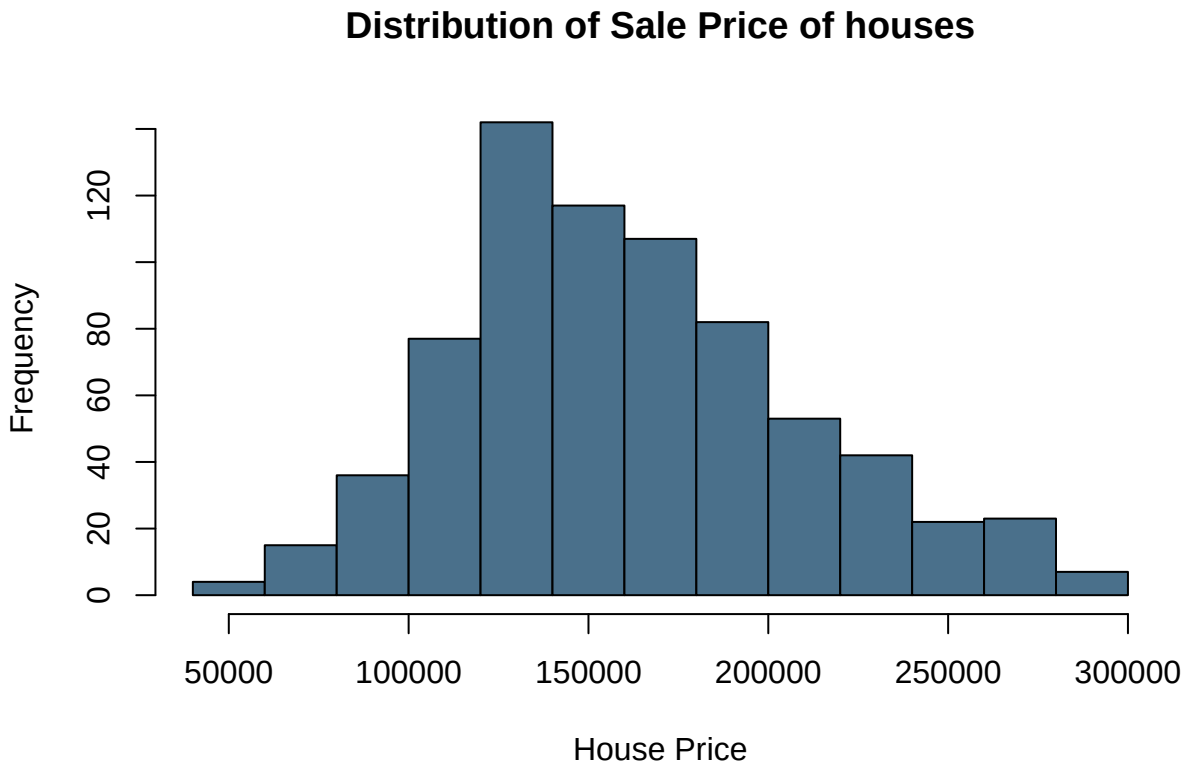
```
paste0('See that the 75th percentile of columns with outliers contain ',
       paste0(summary(numOutliers)[5]), ' outliers')
```

```
## [1] "See that the 75th percentile of columns with outliers contain 42 outliers"
```

Exploratory Data Analysis

Checking the distribution of Sale Price of houses

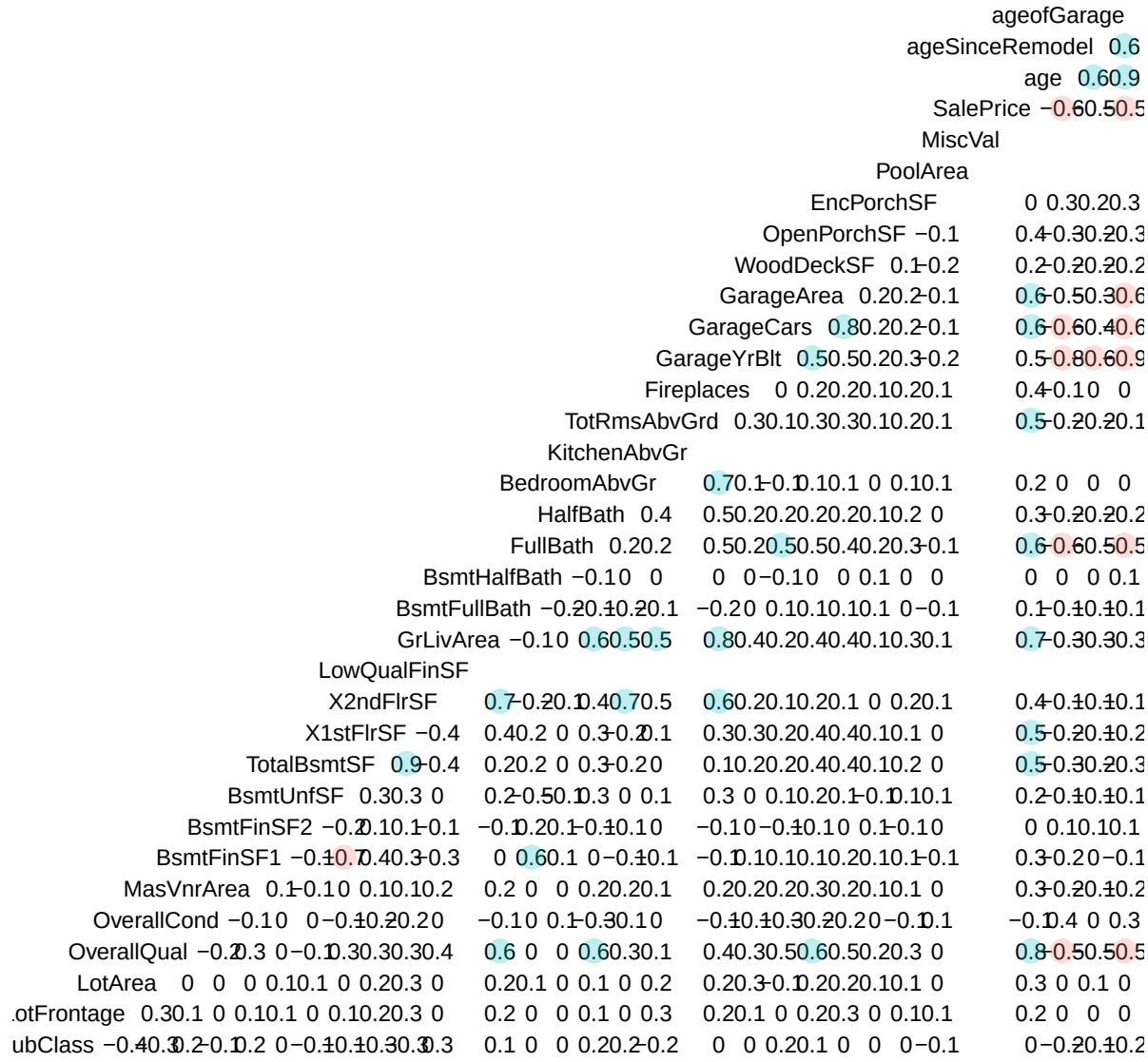
```
hist(hd.CleanedNoOutliers$SalePrice,
     col = 'skyblue4',
     main = 'Distribution of Sale Price of houses',
     xlab = 'House Price')
```



- After removing the desired outliers, we can see that the distribution of Sale Price looks like a normal distribution with few outliers on the right tail.

Correlation between features in the dataset

```
ggcorr(hd.CleanedNoOutliers, geom='blank', label=T, label_size=3, hjust=1,
       size=3, layout.exp=2) +
  geom_point(size = 4, aes(color = coefficient > 0, alpha = abs(coefficient) >= 0.5)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = F, alpha = F)
```



- We can see that `SalePrice` has strong correlations with `GarageArea`, `GarageCars`, `TotRmsAbvGrd`, `FullBath`, `GrLivArea`, `X1stFlrSF`, `TotalBsmtSF`, `OverallQual`.

1 (a) - OLS Model

i.

Hold-out validation set

- Since, we have deleted some of the outlier values during data pre-processing, using 10% of the data as test and remaining 90% as train

```
idx <- sample(nrow(hd.CleanedNoOutliers), nrow(hd.CleanedNoOutliers)*0.1)
test <- hd.CleanedNoOutliers[idx,]
train <- hd.CleanedNoOutliers[-idx,]
```

Fit the OLS Model

`Model 1`:
* Linear model containing:
- *Independent variables:* `GarageArea + GarageCars + TotRmsAbvGrd + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual` - *Predicted variable:* `SalePrice`

```
olsMdl1 <- lm(SalePrice ~ GarageArea + GarageCars + TotRmsAbvGrd
              + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual, data=train)
```

```
VIF(olsMdl1)
```

- **For Model 1**: Adjusted R-squared is `0.8153`, AIC is `16689.91` and BIC is `16735.88` and RMSE is `20995.73`.
- Still trying to improve the existing model.
- No multicolinearity detected.

`Model 2`:
* Linear model containing:
- *Independent variables:* `GarageArea * GarageCars * TotRmsAbvGrd * FullBath * GrLivArea * X1stFlrSF * TotalBsmtSF * OverallQual`
- *Predicted variable:* `SalePrice`

```
olsMdl2 <- lm(SalePrice ~ GarageArea * GarageCars * TotRmsAbvGrd
              * FullBath * GrLivArea * X1stFlrSF * TotalBsmtSF * OverallQual, data=train)
```

```
AIC(olsMdl2)
BIC(olsMdl2)

olsMdl2_RMSE <- rmse(actual=train$SalePrice, predicted=olsMdl2$fitted.values)
olsMdl2_RMSE
```

- **For Model 2**: Adjusted R-squared is `0.8475`, AIC is `16737.27`, BIC is `17914.13` and RMSE is `15502.76`.
- This model works better than the previous one.
- The next model created is based on `Principal Component Analysis`.
  - Uses `numeric` data for Principal Component Analysis
  - Then appends the `factor` data to the data *without* `NULL` *values*
  - Finally, uses `stepAIC()` to best model data

`Model 3`:

Get cleaned `numeric` and `factor data frames`

```r
# After cleaning, two data sets that contain..

## Numeric data ----------------------------------------------------
hd.numericClean <- train %>% select_if(is.numeric)

## Factors ---------------------------------------------------------
hd.factorClean  <- train %>% dplyr::select(where(is.factor))

# Removing any columns with NA
removeColsWithNA <- function(df) {
  return( df[ , colSums(is.na(df)) == 0] )
}
hd.factorClean <- removeColsWithNA(hd.factorClean)

paste('Num. factor cols. removed due to null values:',
      ncol(train %>% dplyr::select(where(is.factor)) ) - ncol(hd.factorClean) )
```

```
## [1] "Num. factor cols. removed due to null values: 12"
```

```r
paste(ncol(hd.factorClean), 'factor cols. remain')
```

```
## [1] "26 factor cols. remain"
```

Perform PCA

```r
# Principal component analysis on numeric data
#to remove zero variance columns from the dataset, using the apply expression,
#setting variance not equal to zero
pc.house <- prcomp(hd.numericClean[ , which(apply(hd.numericClean, 2, var) != 0)] %>%
                     dplyr::select(-SalePrice), # do not include response var
                   center = TRUE, # Mean centered
                   scale  = TRUE  # Z-Score standardized
                   )

# See first 10 cumulative proportions
pc.house.summary <- summary(pc.house)
pc.house.summary$importance[, 1:10]
```

```
##                             PC1      PC2      PC3      PC4     PC5      PC6
## Standard deviation     2.602965 1.879689 1.716776 1.410237 1.17504 1.105497
## Proportion of Variance 0.233640 0.121840 0.101630 0.068580 0.04761 0.042140
## Cumulative Proportion  0.233640 0.355470 0.457100 0.525680 0.57329 0.615430
##                             PC7      PC8      PC9      PC10
## Standard deviation     1.062365 1.044185 1.009602 0.9608487
## Proportion of Variance 0.038920 0.037600 0.035150 0.0318400
## Cumulative Proportion  0.654350 0.691950 0.727100 0.7589300
```

Now we choose number of PC's that explain 75% of the variation

- Note this threshold is just a judgement call. No significance behind 75%

```
cumPropThreshold = 0.75 # The threshold

numPCs <- sum(pc.house.summary$importance['Cumulative Proportion', ] < cumPropThreshold)
paste0('There are ', numPCs, ' principal components that explain up to ', cumPropThreshold*100,
       '% of the variation in the data')
```

```
## [1] "There are 9 principal components that explain up to 75% of the variation in the data"
```

```
chosenPCs <- as.data.frame(pc.house$x[, 1:numPCs])
```

Join on the factor data

```
df.ols <- cbind(SalePrice = hd.numericClean$SalePrice, chosenPCs, hd.factorClean)
```

Fit the Model

- Linear model containing:
  - Principal components explaining 75% of variation in `numeric` data
  - Non-null `factor` data
  - *Predicted variable:* `SalePrice`
- Then use `stepAIC()` to identify which variables are actually important for model

```
# Fit data using PC's, non-null factors
fit.ols <- lm(SalePrice ~ ., data = df.ols)

# Reduce to only important variables
olsMdl3 <- stepAIC(fit.ols, direction="both")
```

```
## Start:  AIC=12750.9
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LotShape + LandContour + LotConfig + LandSlope +
##     Neighborhood + Condition1 + BldgType + HouseStyle + RoofStyle +
##     Exterior1st + Exterior2nd + ExterQual + ExterCond + Foundation +
##     BsmtQual + BsmtCond + BsmtFinType1 + BsmtFinType2 + Heating +
##     HeatingQC + CentralAir + KitchenQual + Functional + PavedDrive +
##     SaleType
##
##                Df  Sum of Sq        RSS   AIC
## - HeatingQC     2 3.0940e+07 1.4477e+11 12747
## - BsmtCond      2 2.9968e+08 1.4504e+11 12748
## - LotShape      3 8.1724e+08 1.4556e+11 12749
## - CentralAir    1 2.9355e+07 1.4477e+11 12749
## - Foundation    3 9.3521e+08 1.4568e+11 12749
## - BsmtQual      2 5.2133e+08 1.4526e+11 12749
## - PC2           1 1.1841e+08 1.4486e+11 12749
## - SaleType      1 1.1878e+08 1.4486e+11 12749
## - PC9           1 1.5322e+08 1.4490e+11 12750
## - PC7           1 1.7135e+08 1.4491e+11 12750
## - HouseStyle    4 1.6110e+09 1.4635e+11 12750
```

```
## - MSZoning      3 1.2258e+09 1.4597e+11 12750
## - KitchenQual   2 7.9287e+08 1.4554e+11 12750
## - Neighborhood  4 1.7016e+09 1.4644e+11 12751
## <none>                       1.4474e+11 12751
## - LandContour   3 1.3375e+09 1.4608e+11 12751
## - Heating       1 5.7634e+08 1.4532e+11 12752
## - ExterCond     2 1.2396e+09 1.4598e+11 12752
## - LandSlope     2 1.4102e+09 1.4615e+11 12753
## - LotConfig     3 2.1711e+09 1.4691e+11 12755
## - BldgType      4 3.0982e+09 1.4784e+11 12757
## - PC5           1 1.7680e+09 1.4651e+11 12757
## - Exterior1st   4 3.1781e+09 1.4792e+11 12757
## - Exterior2nd   4 3.2051e+09 1.4795e+11 12757
## - PavedDrive    2 2.4506e+09 1.4719e+11 12758
## - BsmtFinType2  4 4.0044e+09 1.4875e+11 12761
## - RoofStyle     2 3.7558e+09 1.4850e+11 12764
## - BsmtFinType1  4 4.7198e+09 1.4946e+11 12764
## - Condition1    4 6.2136e+09 1.5096e+11 12770
## - ExterQual     2 5.7328e+09 1.5048e+11 12772
## - Functional    5 7.1611e+09 1.5190e+11 12772
## - PC6           1 6.6484e+09 1.5139e+11 12778
## - PC4           1 8.4733e+09 1.5322e+11 12786
## - PC8           1 1.5560e+10 1.6030e+11 12816
## - PC3           1 2.8295e+10 1.7304e+11 12866
## - PC1           1 8.3343e+10 2.2809e+11 13047
##
## Step:  AIC=12747.04
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LotShape + LandContour + LotConfig + LandSlope +
##     Neighborhood + Condition1 + BldgType + HouseStyle + RoofStyle +
##     Exterior1st + Exterior2nd + ExterQual + ExterCond + Foundation +
##     BsmtQual + BsmtCond + BsmtFinType1 + BsmtFinType2 + Heating +
##     CentralAir + KitchenQual + Functional + PavedDrive + SaleType
##
##                Df  Sum of Sq        RSS   AIC
## - BsmtCond      2 2.8536e+08 1.4506e+11 12744
## - LotShape      3 8.4268e+08 1.4562e+11 12745
## - CentralAir    1 2.8348e+07 1.4480e+11 12745
## - Foundation    3 9.2144e+08 1.4570e+11 12745
## - BsmtQual      2 5.1156e+08 1.4529e+11 12745
## - PC2           1 1.2056e+08 1.4489e+11 12746
## - SaleType      1 1.2756e+08 1.4490e+11 12746
## - PC9           1 1.5548e+08 1.4493e+11 12746
## - PC7           1 1.6588e+08 1.4494e+11 12746
## - HouseStyle    4 1.5998e+09 1.4637e+11 12746
## - MSZoning      3 1.2036e+09 1.4598e+11 12746
## - KitchenQual   2 7.6649e+08 1.4554e+11 12746
## - Neighborhood  4 1.6836e+09 1.4646e+11 12747
## - LandContour   3 1.3278e+09 1.4610e+11 12747
## <none>                       1.4477e+11 12747
## - Heating       1 5.9168e+08 1.4537e+11 12748
## - ExterCond     2 1.2619e+09 1.4604e+11 12749
## - LandSlope     2 1.4268e+09 1.4620e+11 12750
## - LotConfig     3 2.1493e+09 1.4692e+11 12751
```

```
## + HeatingQC     2 3.0940e+07 1.4474e+11 12751
## - BldgType      4 3.0688e+09 1.4784e+11 12753
## - PC5           1 1.7773e+09 1.4655e+11 12753
## - Exterior1st   4 3.1825e+09 1.4796e+11 12753
## - Exterior2nd   4 3.1944e+09 1.4797e+11 12753
## - PavedDrive    2 2.4531e+09 1.4723e+11 12754
## - BsmtFinType2  4 4.0091e+09 1.4878e+11 12757
## - RoofStyle     2 3.7395e+09 1.4851e+11 12760
## - BsmtFinType1  4 4.7345e+09 1.4951e+11 12760
## - Condition1    4 6.2698e+09 1.5104e+11 12767
## - ExterQual     2 5.7096e+09 1.5048e+11 12768
## - Functional    5 7.1882e+09 1.5196e+11 12769
## - PC6           1 6.6368e+09 1.5141e+11 12774
## - PC4           1 8.4841e+09 1.5326e+11 12782
## - PC8           1 1.5620e+10 1.6039e+11 12812
## - PC3           1 2.8438e+10 1.7321e+11 12862
## - PC1           1 8.4568e+10 2.2934e+11 13046
##
## Step:  AIC=12744.33
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LotShape + LandContour + LotConfig + LandSlope +
##     Neighborhood + Condition1 + BldgType + HouseStyle + RoofStyle +
##     Exterior1st + Exterior2nd + ExterQual + ExterCond + Foundation +
##     BsmtQual + BsmtFinType1 + BsmtFinType2 + Heating + CentralAir +
##     KitchenQual + Functional + PavedDrive + SaleType
##
##                Df  Sum of Sq        RSS   AIC
## - LotShape      3 8.0409e+08 1.4586e+11 12742
## - BsmtQual      2 4.6597e+08 1.4553e+11 12742
## - Foundation    3 9.4738e+08 1.4601e+11 12743
## - CentralAir    1 7.1215e+07 1.4513e+11 12743
## - SaleType      1 8.5830e+07 1.4515e+11 12743
## - PC2           1 1.0536e+08 1.4516e+11 12743
## - PC9           1 1.5937e+08 1.4522e+11 12743
## - PC7           1 1.6440e+08 1.4522e+11 12743
## - Neighborhood  4 1.5844e+09 1.4664e+11 12743
## - HouseStyle    4 1.6134e+09 1.4667e+11 12744
## - KitchenQual   2 7.7522e+08 1.4583e+11 12744
## - MSZoning      3 1.2468e+09 1.4631e+11 12744
## <none>                       1.4506e+11 12744
## - LandContour   3 1.4219e+09 1.4648e+11 12745
## - Heating       1 7.1319e+08 1.4577e+11 12746
## - LandSlope     2 1.4330e+09 1.4649e+11 12747
## - ExterCond     2 1.4625e+09 1.4652e+11 12747
## + BsmtCond      2 2.8536e+08 1.4477e+11 12747
## - LotConfig     3 2.1824e+09 1.4724e+11 12748
## + HeatingQC     2 1.6614e+07 1.4504e+11 12748
## - BldgType      4 2.9261e+09 1.4799e+11 12749
## - PC5           1 1.7552e+09 1.4681e+11 12750
## - Exterior2nd   4 3.3590e+09 1.4842e+11 12751
## - Exterior1st   4 3.4102e+09 1.4847e+11 12752
## - PavedDrive    2 2.5872e+09 1.4765e+11 12752
## - BsmtFinType2  4 4.0179e+09 1.4908e+11 12754
## - RoofStyle     2 3.7262e+09 1.4879e+11 12757
```

```
## - BsmtFinType1  4 4.8501e+09 1.4991e+11 12758
## - Condition1    4 6.4658e+09 1.5152e+11 12765
## - ExterQual     2 5.6328e+09 1.5069e+11 12765
## - Functional    5 7.1559e+09 1.5222e+11 12766
## - PC6           1 6.5636e+09 1.5162e+11 12771
## - PC4           1 8.6294e+09 1.5369e+11 12780
## - PC8           1 1.5910e+10 1.6097e+11 12810
## - PC3           1 2.8414e+10 1.7347e+11 12860
## - PC1           1 8.4532e+10 2.2959e+11 13043
##
## Step:  AIC=12741.95
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LandContour + LotConfig + LandSlope + Neighborhood +
##     Condition1 + BldgType + HouseStyle + RoofStyle + Exterior1st +
##     Exterior2nd + ExterQual + ExterCond + Foundation + BsmtQual +
##     BsmtFinType1 + BsmtFinType2 + Heating + CentralAir + KitchenQual +
##     Functional + PavedDrive + SaleType
##
##                Df  Sum of Sq        RSS   AIC
## - Foundation    3 9.3033e+08 1.4679e+11 12740
## - BsmtQual      2 5.1251e+08 1.4638e+11 12740
## - SaleType      1 6.9071e+07 1.4593e+11 12740
## - CentralAir    1 9.5164e+07 1.4596e+11 12740
## - PC2           1 1.0644e+08 1.4597e+11 12740
## - PC9           1 1.4569e+08 1.4601e+11 12741
## - PC7           1 1.5181e+08 1.4602e+11 12741
## - HouseStyle    4 1.5831e+09 1.4745e+11 12741
## - Neighborhood  4 1.7396e+09 1.4760e+11 12742
## - KitchenQual   2 8.5715e+08 1.4672e+11 12742
## <none>                       1.4586e+11 12742
## - MSZoning      3 1.3767e+09 1.4724e+11 12742
## - LandContour   3 1.4752e+09 1.4734e+11 12742
## - Heating       1 7.0042e+08 1.4656e+11 12743
## - ExterCond     2 1.3907e+09 1.4725e+11 12744
## + LotShape      3 8.0409e+08 1.4506e+11 12744
## - LandSlope     2 1.4763e+09 1.4734e+11 12744
## + BsmtCond      2 2.4677e+08 1.4562e+11 12745
## + HeatingQC     2 3.4673e+07 1.4583e+11 12746
## - BldgType      4 3.1523e+09 1.4902e+11 12748
## - Exterior1st   4 3.2858e+09 1.4915e+11 12748
## - Exterior2nd   4 3.3349e+09 1.4920e+11 12749
## - LotConfig     3 2.8931e+09 1.4876e+11 12749
## - PC5           1 1.9891e+09 1.4785e+11 12749
## - PavedDrive    2 2.6134e+09 1.4848e+11 12750
## - BsmtFinType2  4 4.3197e+09 1.5018e+11 12753
## - RoofStyle     2 3.5919e+09 1.4946e+11 12754
## - BsmtFinType1  4 4.7238e+09 1.5059e+11 12755
## - ExterQual     2 5.7557e+09 1.5162e+11 12763
## - Functional    5 7.1815e+09 1.5304e+11 12763
## - Condition1    4 7.1966e+09 1.5306e+11 12766
## - PC6           1 6.7568e+09 1.5262e+11 12770
## - PC4           1 8.9569e+09 1.5482e+11 12779
## - PC8           1 1.6216e+10 1.6208e+11 12809
## - PC3           1 2.8353e+10 1.7422e+11 12856
```

```
## - PC1           1 8.5555e+10 2.3142e+11 13042
##
## Step:  AIC=12740.12
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LandContour + LotConfig + LandSlope + Neighborhood +
##     Condition1 + BldgType + HouseStyle + RoofStyle + Exterior1st +
##     Exterior2nd + ExterQual + ExterCond + BsmtQual + BsmtFinType1 +
##     BsmtFinType2 + Heating + CentralAir + KitchenQual + Functional +
##     PavedDrive + SaleType
##
##                Df  Sum of Sq        RSS   AIC
## - SaleType      1 6.6940e+07 1.4686e+11 12738
## - CentralAir    1 1.0961e+08 1.4690e+11 12739
## - PC7           1 1.1671e+08 1.4691e+11 12739
## - BsmtQual      2 5.8142e+08 1.4738e+11 12739
## - PC9           1 1.3533e+08 1.4693e+11 12739
## - PC2           1 1.6960e+08 1.4696e+11 12739
## <none>                       1.4679e+11 12740
## - Neighborhood  4 1.8194e+09 1.4861e+11 12740
## - HouseStyle    4 1.8217e+09 1.4862e+11 12740
## - MSZoning      3 1.3787e+09 1.4817e+11 12740
## - KitchenQual   2 9.4182e+08 1.4774e+11 12740
## - LandContour   3 1.4989e+09 1.4829e+11 12741
## - Heating       1 6.4027e+08 1.4743e+11 12741
## + Foundation    3 9.3033e+08 1.4586e+11 12742
## - LandSlope     2 1.4242e+09 1.4822e+11 12742
## + LotShape      3 7.8705e+08 1.4601e+11 12743
## - ExterCond     2 1.5002e+09 1.4829e+11 12743
## + BsmtCond      2 2.7250e+08 1.4652e+11 12743
## + HeatingQC     2 1.2631e+07 1.4678e+11 12744
## - Exterior1st   4 3.2342e+09 1.5003e+11 12746
## - PC5           1 1.9267e+09 1.4872e+11 12747
## - LotConfig     3 2.8931e+09 1.4969e+11 12747
## - Exterior2nd   4 3.3937e+09 1.5019e+11 12747
## - BldgType      4 3.4503e+09 1.5024e+11 12747
## - PavedDrive    2 2.6210e+09 1.4941e+11 12748
## - BsmtFinType2  4 4.3767e+09 1.5117e+11 12751
## - RoofStyle     2 3.6364e+09 1.5043e+11 12752
## - BsmtFinType1  4 4.9217e+09 1.5172e+11 12754
## - ExterQual     2 5.7779e+09 1.5257e+11 12761
## - Functional    5 7.1878e+09 1.5398e+11 12761
## - Condition1    4 7.4407e+09 1.5423e+11 12764
## - PC6           1 6.5995e+09 1.5339e+11 12767
## - PC4           1 8.9759e+09 1.5577e+11 12777
## - PC8           1 1.6579e+10 1.6337e+11 12808
## - PC3           1 2.7982e+10 1.7478e+11 12852
## - PC1           1 9.1270e+10 2.3806e+11 13055
##
## Step:  AIC=12738.41
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LandContour + LotConfig + LandSlope + Neighborhood +
##     Condition1 + BldgType + HouseStyle + RoofStyle + Exterior1st +
##     Exterior2nd + ExterQual + ExterCond + BsmtQual + BsmtFinType1 +
##     BsmtFinType2 + Heating + CentralAir + KitchenQual + Functional +
```

```
##     PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - CentralAir    1 1.0308e+08 1.4696e+11 12737
## - PC7           1 1.0769e+08 1.4697e+11 12737
## - BsmtQual      2 5.8994e+08 1.4745e+11 12737
## - PC9           1 1.4511e+08 1.4701e+11 12737
## - PC2           1 1.5745e+08 1.4702e+11 12737
## <none>                       1.4686e+11 12738
## - HouseStyle    4 1.8297e+09 1.4869e+11 12738
## - KitchenQual   2 9.3353e+08 1.4779e+11 12739
## - Neighborhood  4 1.8419e+09 1.4870e+11 12739
## - MSZoning      3 1.3917e+09 1.4825e+11 12739
## - LandContour   3 1.4829e+09 1.4834e+11 12739
## - Heating       1 6.9380e+08 1.4755e+11 12740
## + SaleType      1 6.6940e+07 1.4679e+11 12740
## + Foundation    3 9.2820e+08 1.4593e+11 12740
## - LandSlope     2 1.4210e+09 1.4828e+11 12741
## + LotShape      3 7.7432e+08 1.4609e+11 12741
## - ExterCond     2 1.4734e+09 1.4833e+11 12741
## + BsmtCond      2 2.3503e+08 1.4663e+11 12741
## + HeatingQC     2 1.7683e+07 1.4684e+11 12742
## - Exterior1st   4 3.2287e+09 1.5009e+11 12745
## - LotConfig     3 2.8602e+09 1.4972e+11 12745
## - PC5           1 1.9530e+09 1.4881e+11 12745
## - BldgType      4 3.4135e+09 1.5027e+11 12746
## - Exterior2nd   4 3.4314e+09 1.5029e+11 12746
## - PavedDrive    2 2.6790e+09 1.4954e+11 12746
## - BsmtFinType2  4 4.4697e+09 1.5133e+11 12750
## - RoofStyle     2 3.6180e+09 1.5048e+11 12750
## - BsmtFinType1  4 4.9688e+09 1.5183e+11 12752
## - Functional    5 7.1772e+09 1.5404e+11 12760
## - ExterQual     2 5.7941e+09 1.5265e+11 12760
## - Condition1    4 7.4586e+09 1.5432e+11 12763
## - PC6           1 6.6577e+09 1.5352e+11 12766
## - PC4           1 9.1676e+09 1.5603e+11 12776
## - PC8           1 1.6513e+10 1.6337e+11 12806
## - PC3           1 2.7993e+10 1.7485e+11 12851
## - PC1           1 9.1206e+10 2.3807e+11 13053
##
## Step:  AIC=12736.87
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LandContour + LotConfig + LandSlope + Neighborhood +
##     Condition1 + BldgType + HouseStyle + RoofStyle + Exterior1st +
##     Exterior2nd + ExterQual + ExterCond + BsmtQual + BsmtFinType1 +
##     BsmtFinType2 + Heating + KitchenQual + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - BsmtQual      2 5.5782e+08 1.4752e+11 12735
## - PC7           1 1.1727e+08 1.4708e+11 12735
## - PC9           1 1.4688e+08 1.4711e+11 12736
## - PC2           1 1.6436e+08 1.4713e+11 12736
## - HouseStyle    4 1.7776e+09 1.4874e+11 12737
## - KitchenQual   2 8.9708e+08 1.4786e+11 12737
```

```
## <none>                       1.4696e+11 12737
## - Neighborhood  4 1.8444e+09 1.4881e+11 12737
## - MSZoning      3 1.4080e+09 1.4837e+11 12737
## - LandContour   3 1.4938e+09 1.4846e+11 12738
## - Heating       1 5.9260e+08 1.4756e+11 12738
## + CentralAir    1 1.0308e+08 1.4686e+11 12738
## + SaleType      1 6.0415e+07 1.4690e+11 12739
## + Foundation    3 9.4281e+08 1.4602e+11 12739
## - LandSlope     2 1.4216e+09 1.4839e+11 12739
## + LotShape      3 8.0209e+08 1.4616e+11 12739
## - ExterCond     2 1.4572e+09 1.4842e+11 12739
## + BsmtCond      2 2.8565e+08 1.4668e+11 12740
## + HeatingQC     2 1.5615e+07 1.4695e+11 12741
## - Exterior1st   4 3.2501e+09 1.5021e+11 12743
## - PC5           1 1.9071e+09 1.4887e+11 12743
## - LotConfig     3 2.8592e+09 1.4982e+11 12744
## - Exterior2nd   4 3.4873e+09 1.5045e+11 12744
## - BldgType      4 3.5697e+09 1.5053e+11 12745
## - PavedDrive    2 2.7492e+09 1.4971e+11 12745
## - RoofStyle     2 3.6339e+09 1.5060e+11 12749
## - BsmtFinType2  4 4.5926e+09 1.5156e+11 12749
## - BsmtFinType1  4 5.0114e+09 1.5198e+11 12751
## - ExterQual     2 5.7438e+09 1.5271e+11 12758
## - Functional    5 7.2029e+09 1.5417e+11 12758
## - Condition1    4 7.5798e+09 1.5454e+11 12762
## - PC6           1 6.8315e+09 1.5380e+11 12765
## - PC4           1 9.1872e+09 1.5615e+11 12775
## - PC8           1 1.6970e+10 1.6393e+11 12806
## - PC3           1 2.8282e+10 1.7525e+11 12850
## - PC1           1 9.2701e+10 2.3966e+11 13055
##
## Step:  AIC=12735.36
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
##     MSZoning + LandContour + LotConfig + LandSlope + Neighborhood +
##     Condition1 + BldgType + HouseStyle + RoofStyle + Exterior1st +
##     Exterior2nd + ExterQual + ExterCond + BsmtFinType1 + BsmtFinType2 +
##     Heating + KitchenQual + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - PC7           1 1.2031e+08 1.4764e+11 12734
## - PC9           1 1.5156e+08 1.4767e+11 12734
## - PC2           1 2.1532e+08 1.4774e+11 12734
## - Neighborhood  4 1.7889e+09 1.4931e+11 12735
## <none>                       1.4752e+11 12735
## - KitchenQual   2 9.5224e+08 1.4847e+11 12736
## - HouseStyle    4 1.9242e+09 1.4945e+11 12736
## - Heating       1 5.8858e+08 1.4811e+11 12736
## - LandContour   3 1.6182e+09 1.4914e+11 12736
## + BsmtQual      2 5.5782e+08 1.4696e+11 12737
## + Foundation    3 9.9601e+08 1.4653e+11 12737
## - MSZoning      3 1.7404e+09 1.4926e+11 12737
## + CentralAir    1 7.0958e+07 1.4745e+11 12737
## + SaleType      1 6.8559e+07 1.4745e+11 12737
## + LotShape      3 8.6394e+08 1.4666e+11 12738
```

```
## - LandSlope     2 1.4236e+09 1.4895e+11 12738
## - ExterCond     2 1.5191e+09 1.4904e+11 12738
## + BsmtCond      2 2.3361e+08 1.4729e+11 12738
## + HeatingQC     2 7.8537e+06 1.4751e+11 12739
## - LotConfig     3 2.8053e+09 1.5033e+11 12742
## - PC5           1 1.9725e+09 1.4949e+11 12742
## - Exterior1st   4 3.3548e+09 1.5088e+11 12742
## - Exterior2nd   4 3.3684e+09 1.5089e+11 12742
## - BldgType      4 3.4545e+09 1.5098e+11 12742
## - PavedDrive    2 3.0317e+09 1.5055e+11 12745
## - RoofStyle     2 3.4003e+09 1.5092e+11 12746
## - BsmtFinType2  4 5.1011e+09 1.5262e+11 12750
## - BsmtFinType1  4 5.4892e+09 1.5301e+11 12751
## - Functional    5 7.2592e+09 1.5478e+11 12757
## - ExterQual     2 5.8580e+09 1.5338e+11 12757
## - Condition1    4 7.7835e+09 1.5530e+11 12761
## - PC6           1 6.6952e+09 1.5422e+11 12762
## - PC4           1 9.4196e+09 1.5694e+11 12774
## - PC8           1 1.6646e+10 1.6417e+11 12803
## - PC3           1 2.8052e+10 1.7557e+11 12847
## - PC1           1 1.0723e+11 2.5476e+11 13091
##
## Step:  AIC=12733.89
## SalePrice ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC8 + PC9 + MSZoning +
##     LandContour + LotConfig + LandSlope + Neighborhood + Condition1 +
##     BldgType + HouseStyle + RoofStyle + Exterior1st + Exterior2nd +
##     ExterQual + ExterCond + BsmtFinType1 + BsmtFinType2 + Heating +
##     KitchenQual + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - PC2           1 2.1867e+08 1.4786e+11 12733
## - PC9           1 2.5869e+08 1.4790e+11 12733
## - Neighborhood  4 1.7805e+09 1.4942e+11 12734
## <none>                       1.4764e+11 12734
## - KitchenQual   2 9.6068e+08 1.4860e+11 12734
## - HouseStyle    4 1.8898e+09 1.4953e+11 12734
## - Heating       1 5.7344e+08 1.4822e+11 12734
## - LandContour   3 1.6018e+09 1.4924e+11 12735
## + PC7           1 1.2031e+08 1.4752e+11 12735
## + BsmtQual      2 5.6085e+08 1.4708e+11 12735
## - MSZoning      3 1.7289e+09 1.4937e+11 12736
## + CentralAir    1 7.8496e+07 1.4756e+11 12736
## + SaleType      1 5.9399e+07 1.4758e+11 12736
## + Foundation    3 9.5294e+08 1.4669e+11 12736
## + LotShape      3 8.6229e+08 1.4678e+11 12736
## - LandSlope     2 1.4460e+09 1.4909e+11 12736
## - ExterCond     2 1.5074e+09 1.4915e+11 12736
## + BsmtCond      2 2.3501e+08 1.4741e+11 12737
## + HeatingQC     2 6.9838e+06 1.4763e+11 12738
## - LotConfig     3 2.8211e+09 1.5046e+11 12740
## - PC5           1 1.9456e+09 1.4959e+11 12740
## - Exterior1st   4 3.3243e+09 1.5097e+11 12740
## - Exterior2nd   4 3.3705e+09 1.5101e+11 12741
## - BldgType      4 3.4928e+09 1.5113e+11 12741
```

```
## - PavedDrive    2 3.0340e+09 1.5068e+11 12743
## - RoofStyle     2 3.4834e+09 1.5113e+11 12745
## - BsmtFinType2  4 5.1266e+09 1.5277e+11 12748
## - BsmtFinType1  4 5.4746e+09 1.5312e+11 12750
## - ExterQual     2 5.9270e+09 1.5357e+11 12756
## - Functional    5 7.6567e+09 1.5530e+11 12757
## - Condition1    4 7.7412e+09 1.5538e+11 12759
## - PC6           1 7.3300e+09 1.5497e+11 12764
## - PC4           1 9.5271e+09 1.5717e+11 12773
## - PC8           1 1.6526e+10 1.6417e+11 12801
## - PC3           1 2.8375e+10 1.7602e+11 12847
## - PC1           1 1.0735e+11 2.5499e+11 13090
##
## Step:  AIC=12732.86
## SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + PC9 + MSZoning +
##     LandContour + LotConfig + LandSlope + Neighborhood + Condition1 +
##     BldgType + HouseStyle + RoofStyle + Exterior1st + Exterior2nd +
##     ExterQual + ExterCond + BsmtFinType1 + BsmtFinType2 + Heating +
##     KitchenQual + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - PC9           1 2.6651e+08 1.4813e+11 12732
## - HouseStyle    4 1.6910e+09 1.4955e+11 12732
## <none>                       1.4786e+11 12733
## - KitchenQual   2 9.1142e+08 1.4877e+11 12733
## - Neighborhood  4 1.8267e+09 1.4969e+11 12733
## - Heating       1 5.1018e+08 1.4837e+11 12733
## + PC2           1 2.1867e+08 1.4764e+11 12734
## + BsmtQual      2 6.1212e+08 1.4725e+11 12734
## - MSZoning      3 1.6621e+09 1.4952e+11 12734
## + Foundation    3 1.0263e+09 1.4683e+11 12734
## + PC7           1 1.2366e+08 1.4774e+11 12734
## - LandContour   3 1.7080e+09 1.4957e+11 12734
## + CentralAir    1 8.4551e+07 1.4778e+11 12734
## + SaleType      1 4.5728e+07 1.4781e+11 12735
## + LotShape      3 8.8062e+08 1.4698e+11 12735
## - LandSlope     2 1.4534e+09 1.4931e+11 12735
## + BsmtCond      2 2.2429e+08 1.4764e+11 12736
## - ExterCond     2 1.5935e+09 1.4945e+11 12736
## + HeatingQC     2 7.2833e+06 1.4785e+11 12737
## - Exterior2nd   4 3.3315e+09 1.5119e+11 12740
## - PC5           1 1.9559e+09 1.4982e+11 12740
## - BldgType      4 3.3599e+09 1.5122e+11 12740
## - Exterior1st   4 3.3946e+09 1.5125e+11 12740
## - LotConfig     3 2.9485e+09 1.5081e+11 12740
## - PavedDrive    2 3.0714e+09 1.5093e+11 12742
## - RoofStyle     2 3.5249e+09 1.5139e+11 12744
## - BsmtFinType2  4 5.2314e+09 1.5309e+11 12748
## - BsmtFinType1  4 6.2051e+09 1.5407e+11 12752
## - ExterQual     2 6.0347e+09 1.5390e+11 12755
## - Functional    5 7.5647e+09 1.5543e+11 12756
## - Condition1    4 7.6414e+09 1.5550e+11 12758
## - PC6           1 7.3261e+09 1.5519e+11 12762
## - PC4           1 1.1676e+10 1.5954e+11 12781
```

```
## - PC8           1 1.7013e+10 1.6487e+11 12802
## - PC3           1 2.9065e+10 1.7693e+11 12848
## - PC1           1 1.2822e+11 2.7608e+11 13140
##
## Step:  AIC=12732.04
## SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +
##     LotConfig + LandSlope + Neighborhood + Condition1 + BldgType +
##     HouseStyle + RoofStyle + Exterior1st + Exterior2nd + ExterQual +
##     ExterCond + BsmtFinType1 + BsmtFinType2 + Heating + KitchenQual +
##     Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - HouseStyle    4 1.6919e+09 1.4982e+11 12732
## - KitchenQual   2 8.5378e+08 1.4898e+11 12732
## - Neighborhood  4 1.7774e+09 1.4990e+11 12732
## <none>                       1.4813e+11 12732
## - Heating       1 5.2954e+08 1.4866e+11 12732
## + PC9           1 2.6651e+08 1.4786e+11 12733
## + PC7           1 2.3400e+08 1.4789e+11 12733
## + PC2           1 2.2648e+08 1.4790e+11 12733
## + BsmtQual      2 6.2272e+08 1.4750e+11 12733
## - MSZoning      3 1.7177e+09 1.4984e+11 12734
## + CentralAir    1 8.8664e+07 1.4804e+11 12734
## + Foundation    3 9.8652e+08 1.4714e+11 12734
## + SaleType      1 5.5678e+07 1.4807e+11 12734
## - LandContour   3 1.8331e+09 1.4996e+11 12734
## - LandSlope     2 1.3845e+09 1.4951e+11 12734
## + LotShape      3 8.6810e+08 1.4726e+11 12734
## - ExterCond     2 1.5614e+09 1.4969e+11 12735
## + BsmtCond      2 2.3177e+08 1.4790e+11 12735
## + HeatingQC     2 6.9975e+06 1.4812e+11 12736
## - PC5           1 1.9590e+09 1.5009e+11 12739
## - BldgType      4 3.3482e+09 1.5148e+11 12739
## - LotConfig     3 2.9271e+09 1.5105e+11 12739
## - Exterior2nd   4 3.4468e+09 1.5157e+11 12739
## - Exterior1st   4 3.4511e+09 1.5158e+11 12739
## - PavedDrive    2 3.0250e+09 1.5115e+11 12741
## - RoofStyle     2 3.7653e+09 1.5189e+11 12744
## - BsmtFinType2  4 4.9940e+09 1.5312e+11 12746
## - BsmtFinType1  4 5.9590e+09 1.5409e+11 12750
## - ExterQual     2 6.0632e+09 1.5419e+11 12754
## - Functional    5 7.5340e+09 1.5566e+11 12754
## - Condition1    4 7.5731e+09 1.5570e+11 12757
## - PC6           1 7.1967e+09 1.5532e+11 12761
## - PC4           1 1.1417e+10 1.5954e+11 12779
## - PC8           1 1.8009e+10 1.6614e+11 12805
## - PC3           1 2.8932e+10 1.7706e+11 12847
## - PC1           1 1.3135e+11 2.7947e+11 13146
##
## Step:  AIC=12731.48
## SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +
##     LotConfig + LandSlope + Neighborhood + Condition1 + BldgType +
##     RoofStyle + Exterior1st + Exterior2nd + ExterQual + ExterCond +
##     BsmtFinType1 + BsmtFinType2 + Heating + KitchenQual + Functional +
```

```
##     PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - Neighborhood  4 1.4853e+09 1.5130e+11 12730
## - Heating       1 3.2834e+08 1.5015e+11 12731
## - KitchenQual   2 8.8320e+08 1.5070e+11 12731
## <none>                       1.4982e+11 12732
## + HouseStyle    4 1.6919e+09 1.4813e+11 12732
## + PC9           1 2.6748e+08 1.4955e+11 12732
## + BsmtQual      2 6.9001e+08 1.4913e+11 12732
## + PC7           1 1.7767e+08 1.4964e+11 12733
## + Foundation    3 1.0764e+09 1.4874e+11 12733
## - LandContour   3 1.7297e+09 1.5155e+11 12733
## + SaleType      1 8.6685e+07 1.4973e+11 12733
## + CentralAir    1 2.8273e+07 1.4979e+11 12733
## + PC2           1 2.4124e+07 1.4979e+11 12733
## + LotShape      3 9.0109e+08 1.4892e+11 12734
## - MSZoning      3 2.0487e+09 1.5187e+11 12734
## + BsmtCond      2 2.3183e+08 1.4959e+11 12734
## - ExterCond     2 1.6654e+09 1.5148e+11 12735
## - LandSlope     2 1.7047e+09 1.5152e+11 12735
## + HeatingQC     2 1.4408e+07 1.4980e+11 12735
## - PC5           1 1.6607e+09 1.5148e+11 12737
## - BldgType      4 3.1018e+09 1.5292e+11 12737
## - Exterior1st   4 3.2139e+09 1.5303e+11 12737
## - Exterior2nd   4 3.2417e+09 1.5306e+11 12738
## - LotConfig     3 2.8908e+09 1.5271e+11 12738
## - PavedDrive    2 2.8725e+09 1.5269e+11 12740
## - RoofStyle     2 3.8078e+09 1.5363e+11 12744
## - BsmtFinType2  4 5.1050e+09 1.5492e+11 12745
## - BsmtFinType1  4 5.3511e+09 1.5517e+11 12746
## - Functional    5 7.7643e+09 1.5758e+11 12755
## - ExterQual     2 6.3573e+09 1.5618e+11 12755
## - Condition1    4 7.4383e+09 1.5726e+11 12755
## - PC6           1 6.8297e+09 1.5665e+11 12759
## - PC8           1 1.7813e+10 1.6763e+11 12803
## - PC4           1 1.8706e+10 1.6853e+11 12806
## - PC3           1 3.3044e+10 1.8286e+11 12860
## - PC1           1 1.8617e+11 3.3599e+11 13258
##
## Step:  AIC=12729.94
## SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +
##     LotConfig + LandSlope + Condition1 + BldgType + RoofStyle +
##     Exterior1st + Exterior2nd + ExterQual + ExterCond + BsmtFinType1 +
##     BsmtFinType2 + Heating + KitchenQual + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## - KitchenQual   2 7.9501e+08 1.5210e+11 12729
## - Heating       1 3.4284e+08 1.5165e+11 12729
## <none>                       1.5130e+11 12730
## + Foundation    3 1.1569e+09 1.5015e+11 12731
## + PC9           1 2.2495e+08 1.5108e+11 12731
## + BsmtQual      2 6.3712e+08 1.5067e+11 12731
## + PC7           1 1.5912e+08 1.5115e+11 12731
```

```
## + Neighborhood  4 1.4853e+09 1.4982e+11 12732
## + LotShape      3 1.0217e+09 1.5028e+11 12732
## + SaleType      1 9.7995e+07 1.5121e+11 12732
## + CentralAir    1 3.5652e+07 1.5127e+11 12732
## + HouseStyle    4 1.3998e+09 1.4990e+11 12732
## + PC2           1 8.2027e+06 1.5130e+11 12732
## - LandContour   3 1.9095e+09 1.5321e+11 12732
## + BsmtCond      2 1.6093e+08 1.5114e+11 12733
## - ExterCond     2 1.7475e+09 1.5305e+11 12734
## + HeatingQC     2 1.1104e+07 1.5129e+11 12734
## - LandSlope     2 1.8635e+09 1.5317e+11 12734
## - PC5           1 1.5954e+09 1.5290e+11 12735
## - Exterior1st   4 3.0647e+09 1.5437e+11 12735
## - Exterior2nd   4 3.2065e+09 1.5451e+11 12736
## - BldgType      4 3.2802e+09 1.5458e+11 12736
## - LotConfig     3 2.8201e+09 1.5412e+11 12736
## - PavedDrive    2 2.4929e+09 1.5380e+11 12737
## - MSZoning      3 3.2799e+09 1.5458e+11 12738
## - RoofStyle     2 3.8008e+09 1.5510e+11 12742
## - BsmtFinType2  4 5.7965e+09 1.5710e+11 12747
## - BsmtFinType1  4 6.1604e+09 1.5746e+11 12748
## - Functional    5 7.6802e+09 1.5898e+11 12752
## - ExterQual     2 6.3553e+09 1.5766e+11 12753
## - Condition1    4 7.8300e+09 1.5913e+11 12755
## - PC6           1 6.7836e+09 1.5809e+11 12757
## - PC8           1 1.7578e+10 1.6888e+11 12800
## - PC4           1 2.0278e+10 1.7158e+11 12810
## - PC3           1 3.4141e+10 1.8544e+11 12861
## - PC1           1 1.9653e+11 3.4783e+11 13273
##
## Step:  AIC=12729.37
## SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +
##     LotConfig + LandSlope + Condition1 + BldgType + RoofStyle +
##     Exterior1st + Exterior2nd + ExterQual + ExterCond + BsmtFinType1 +
##     BsmtFinType2 + Heating + Functional + PavedDrive
##
##                Df  Sum of Sq        RSS   AIC
## <none>                       1.5210e+11 12729
## - Heating       1 5.0649e+08 1.5261e+11 12730
## + Foundation    3 1.2666e+09 1.5083e+11 12730
## + KitchenQual   2 7.9501e+08 1.5130e+11 12730
## + BsmtQual      2 6.9205e+08 1.5141e+11 12730
## + PC9           1 1.6804e+08 1.5193e+11 12731
## + PC7           1 1.5456e+08 1.5194e+11 12731
## + LotShape      3 1.0553e+09 1.5104e+11 12731
## + SaleType      1 8.4215e+07 1.5201e+11 12731
## + HouseStyle    4 1.4350e+09 1.5066e+11 12731
## + CentralAir    1 1.5185e+07 1.5208e+11 12731
## + Neighborhood  4 1.3971e+09 1.5070e+11 12731
## + PC2           1 4.8868e+06 1.5209e+11 12731
## - LandContour   3 1.9602e+09 1.5406e+11 12732
## + BsmtCond      2 1.5479e+08 1.5194e+11 12733
## + HeatingQC     2 2.7924e+07 1.5207e+11 12733
## - ExterCond     2 1.8723e+09 1.5397e+11 12733
```

```
## - PC5           1 1.5393e+09 1.5364e+11 12734
## - LandSlope     2 2.0529e+09 1.5415e+11 12734
## - Exterior1st   4 3.1179e+09 1.5522e+11 12735
## - Exterior2nd   4 3.2217e+09 1.5532e+11 12735
## - LotConfig     3 2.7775e+09 1.5488e+11 12735
## - PavedDrive    2 2.3665e+09 1.5447e+11 12736
## - BldgType      4 3.4981e+09 1.5560e+11 12736
## - MSZoning      3 3.4659e+09 1.5557e+11 12738
## - RoofStyle     2 3.8020e+09 1.5590e+11 12742
## - BsmtFinType2  4 6.0404e+09 1.5814e+11 12747
## - BsmtFinType1  4 6.6452e+09 1.5874e+11 12749
## - Functional    5 7.9976e+09 1.6010e+11 12753
## - Condition1    4 7.6216e+09 1.5972e+11 12753
## - PC6           1 7.6695e+09 1.5977e+11 12760
## - ExterQual     2 9.4681e+09 1.6157e+11 12765
## - PC8           1 2.0023e+10 1.7212e+11 12808
## - PC4           1 2.0230e+10 1.7233e+11 12809
## - PC3           1 3.3942e+10 1.8604e+11 12859
## - PC1           1 2.0920e+11 3.6130e+11 13294
```

- Reporting `all the variables` of the best model (`Model 3`):

Coefficient estimates:

```
##                       Estimate Std. Error     t value      Pr(>|t|)
## (Intercept)       120507.6294 10893.4648  11.0623784  5.184157e-26
## PC1               -12448.6218   433.7043 -28.7030201 1.253520e-114
## PC3                 7354.2818   636.0972  11.5615687  4.770789e-28
## PC4                 5281.5707   591.7233   8.9257442  5.379382e-18
## PC5                -1792.5971   728.0682  -2.4621281  1.409204e-02
## PC6                 3811.7671   693.5751   5.4958248  5.759266e-08
## PC8                -6237.3050   702.3917  -8.8800948  7.736115e-18
## MSZoningRH        -14175.7537  9248.7404  -1.5327226  1.258722e-01
## MSZoningRL         -7645.3973  3777.7018  -2.0238224  4.343186e-02
## MSZoningRM        -13546.9726  4084.5403  -3.3166456  9.661844e-04
## LandContourHLS      9690.2907  5312.2317   1.8241469  6.862794e-02
## LandContourLow      6164.2200  6617.9009   0.9314464  3.519979e-01
## LandContourLvl     -2275.6269  3473.0937  -0.6552161  5.125802e-01
## LotConfigCulDSac    4512.9866  3016.5654   1.4960679  1.351626e-01
## LotConfigInside    -1556.3609  1847.6996  -0.8423236  3.999430e-01
## LotConfigother     -8274.3106  3551.2208  -2.3299904  2.013795e-02
## LandSlopeMod       10933.2789  4231.7977   2.5836015  1.001321e-02
## LandSlopeSev       -9320.6999 12589.1214  -0.7403773  4.593612e-01
## Condition1Feedr     6335.4157  4914.1123   1.2892289  1.978161e-01
## Condition1Norm     14140.4651  4034.7765   3.5046464  4.913560e-04
## Condition1RR        2363.3515  5409.5616   0.4368841  6.623528e-01
## Condition1Other     3346.7249  6818.4173   0.4908360  6.237222e-01
## BldgType2fmCon    -17038.1014  6683.5045  -2.5492766  1.104277e-02
## BldgTypeDuplex    -12992.6920  8245.8087  -1.5756723  1.156293e-01
## BldgTypeTwnhs     -10043.4035  4573.7231  -2.1958923  2.848231e-02
## BldgTypeTwnhsE       -893.7624  3421.1918  -0.2612430  7.939949e-01
## RoofStyleHip         2548.2228  1888.0357   1.3496687  1.776321e-01
## RoofStyleother      18721.0445  5059.1037   3.7004666  2.350300e-04
```

```
## Exterior1stMetalSd   5254.7299  6730.5080   0.7807330  4.352680e-01
## Exterior1stVinylSd   3886.2301  7251.2510   0.5359393  5.921995e-01
## Exterior1stWd Sdng  -5598.9758  5209.1334  -1.0748382  2.828799e-01
## Exterior1stOther     6975.9086  3762.5854   1.8540200  6.422766e-02
## Exterior2ndMetalSd   1474.6586  6792.8068   0.2170912  8.282112e-01
## Exterior2ndVinylSd   3705.9481  7393.7259   0.5012288  6.163943e-01
## Exterior2ndWd Sdng  10672.8218  5289.1943   2.0178540  4.405183e-02
## Exterior2ndOther    -3779.3208  3663.2825  -1.0316761  3.026403e-01
## ExterQualAvg       -12040.9747  2112.4979  -5.6998754  1.882766e-08
## ExterQualBelowAvg    3556.0867  8588.7046   0.4140423  6.789914e-01
## ExterCondAvg         5813.5916  2238.5871   2.5969915  9.635409e-03
## ExterCondBelowAvg   11221.5520  6909.4794   1.6240807  1.048847e-01
## BsmtFinType1BLQ     -1871.5993  2392.2906  -0.7823462  4.343203e-01
## BsmtFinType1GLQ      7015.9586  2213.9672   3.1689532  1.607868e-03
## BsmtFinType1Unf       437.5524  2476.5468   0.1766784  8.598207e-01
## BsmtFinType1Other   -5489.8470  2397.4143  -2.2899033  2.237372e-02
## BsmtFinType2LwQ      7224.3407  5287.5857   1.3662834  1.723627e-01
## BsmtFinType2Rec       613.0138  5327.4991   0.1150660  9.084314e-01
## BsmtFinType2Unf     10707.9131  4198.7999   2.5502318  1.101289e-02
## BsmtFinType2Other   25801.7579  6541.8509   3.9441067  8.957527e-05
## Heatingother         7376.1340  5222.6743   1.4123289  1.583724e-01
## FunctionalMaj2      -4952.4903 14013.5718  -0.3534067  7.239079e-01
## FunctionalMin1      14724.9821  8370.3739   1.7591785  7.905769e-02
## FunctionalMin2       6730.3446  8231.2658   0.8176561  4.138787e-01
## FunctionalMod       34746.2363 21885.1281   1.5876643  1.128900e-01
## FunctionalTyp       23602.7752  6641.6433   3.5537553  4.096898e-04
## PavedDriveP         -7245.0061  5314.6643  -1.3632105  1.733284e-01
## PavedDriveY          5599.5048  3351.2048   1.6708930  9.526515e-02
```

p-values:

```
##         value
## 4.379136e-259
```

Adjusted R-squared:

```
## [1] 0.8877827
```

AIC:

```
## [1] 14590.18
```

BIC:

```
## [1] 14845.81
```

VIF:

```
VIF(olsMdl3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## PC1             3.282478  1        1.811761
## PC3             3.071509  1        1.752572
## PC4             1.793490  1        1.339212
## PC5             1.885068  1        1.372978
## PC6             1.514187  1        1.230523
## PC8             1.385452  1        1.177052
## MSZoning        2.382067  3        1.155648
## LandContour     2.671815  3        1.177970
## LotConfig       1.476468  3        1.067097
## LandSlope       3.307846  2        1.348610
## Condition1      1.586586  4        1.059395
## BldgType        4.659022  4        1.212096
## RoofStyle       1.504389  2        1.107491
## Exterior1st  4778.808279  4        2.883467
## Exterior2nd  4839.900411  4        2.888049
## ExterQual       3.238195  2        1.341454
## ExterCond       1.661982  2        1.135420
## BsmtFinType1    4.286295  4        1.199528
## BsmtFinType2    2.209995  4        1.104203
## Heating         1.265426  1        1.124912
## Functional      3.553579  5        1.135185
## PavedDrive      1.572996  2        1.119907
```

RMSE:

```
## [1] 15238.52
```

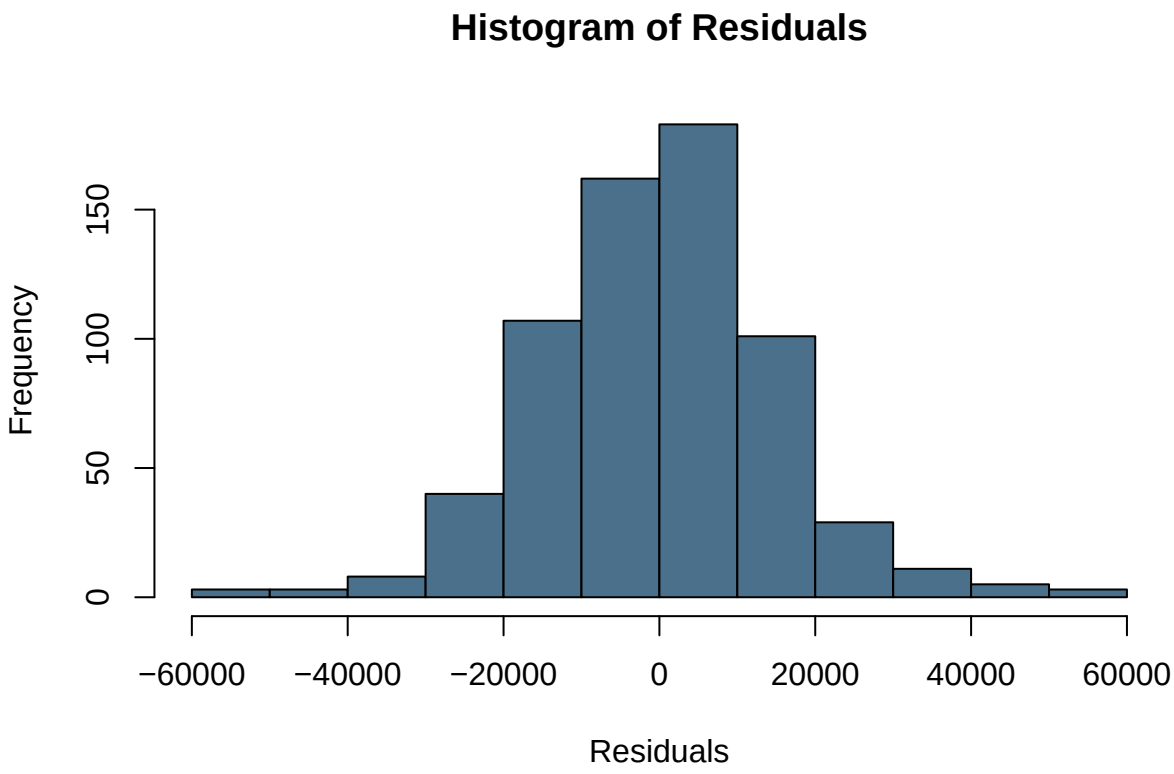
- So, we can say that using PCA followed by stepAIC the OLS regression model is better as compared to the other OLS models built.
- There is also no multicolinearity found in the model as the VIF values are less than 10.

ii. Complete analysis of the residuals

A linear regression model is considered fit if the below assumptions are met:

- **Residuals should follow normal distribution**
- **There should be no heteroscedasticity**
- **There should be no multicollinearity**

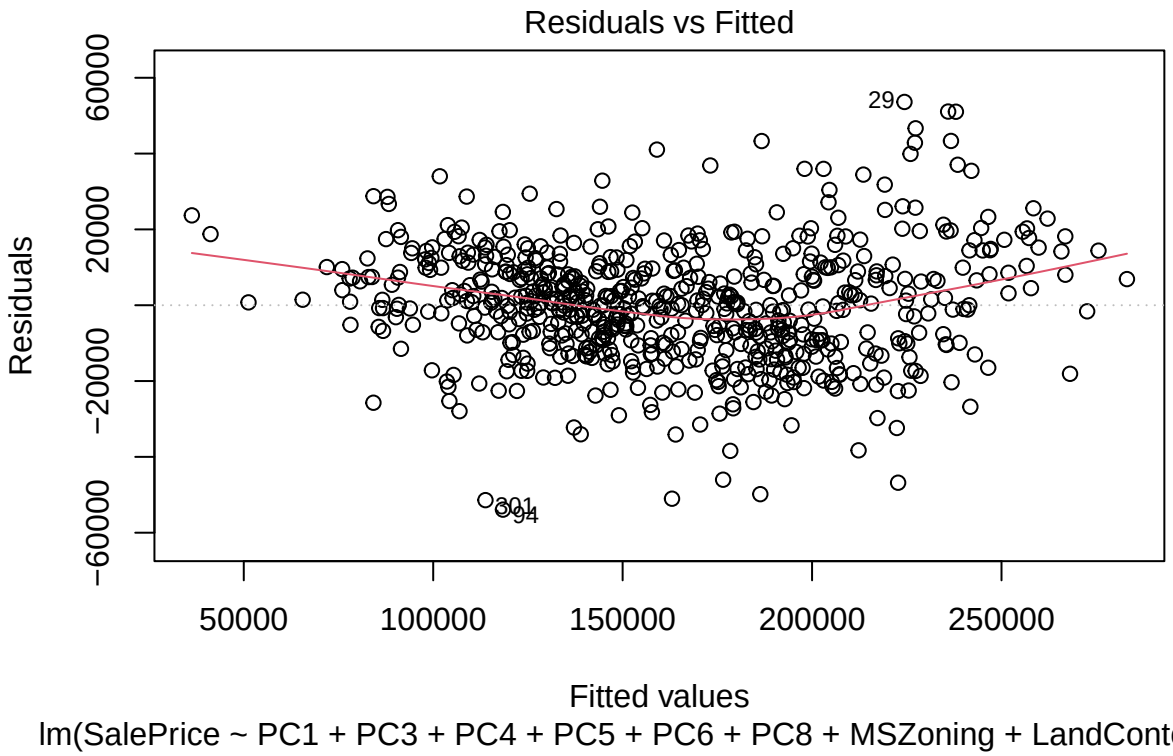
```
hist(olsMdl3$residuals,
     col = 'skyblue4',
     main = 'Histogram of Residuals',
     xlab = 'Residuals')
```

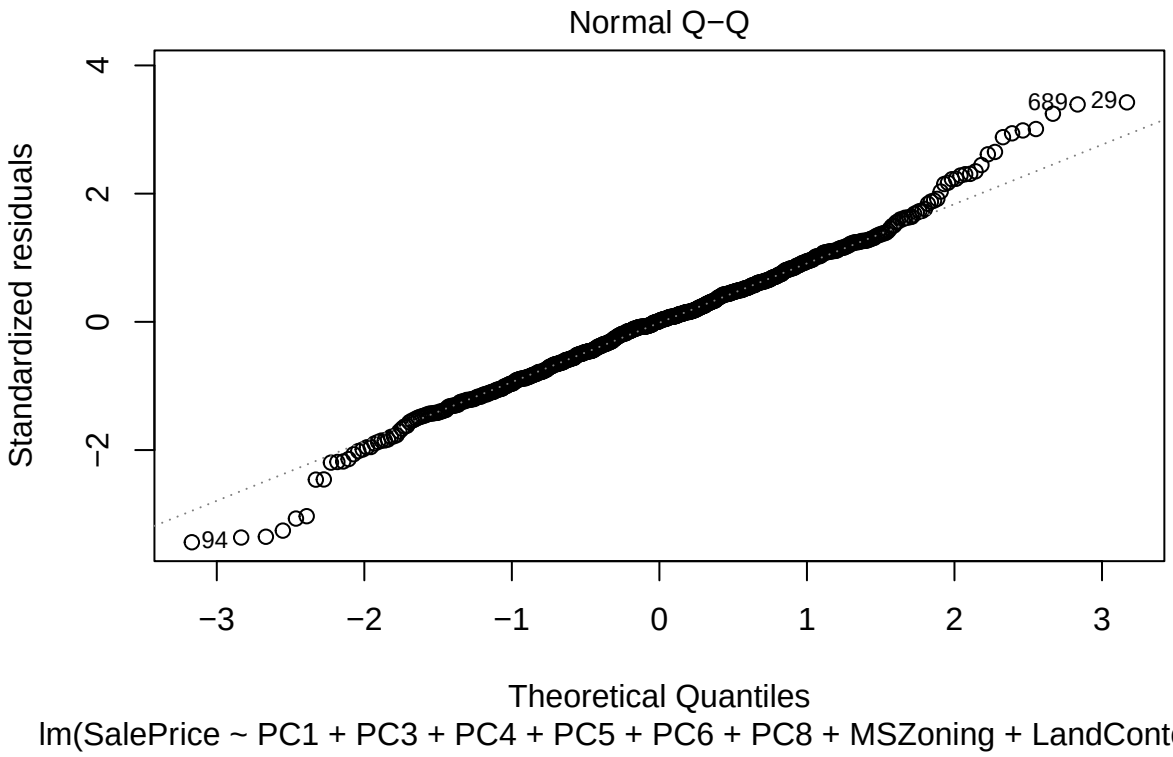


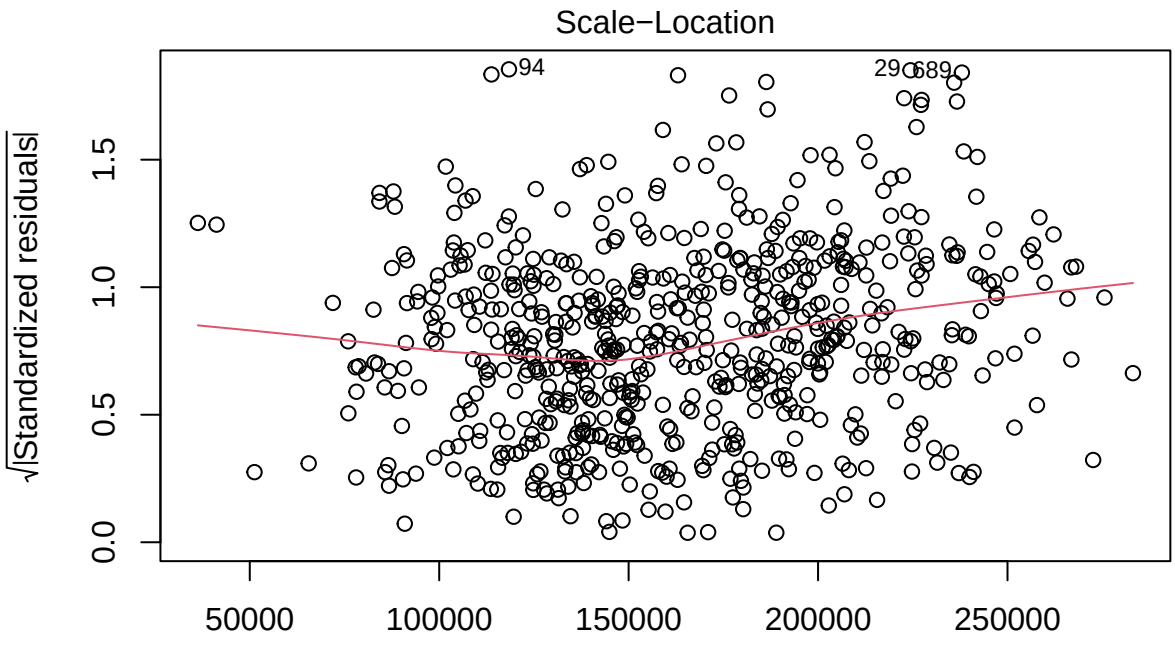
We can see that the residuals are `normally distributed`.

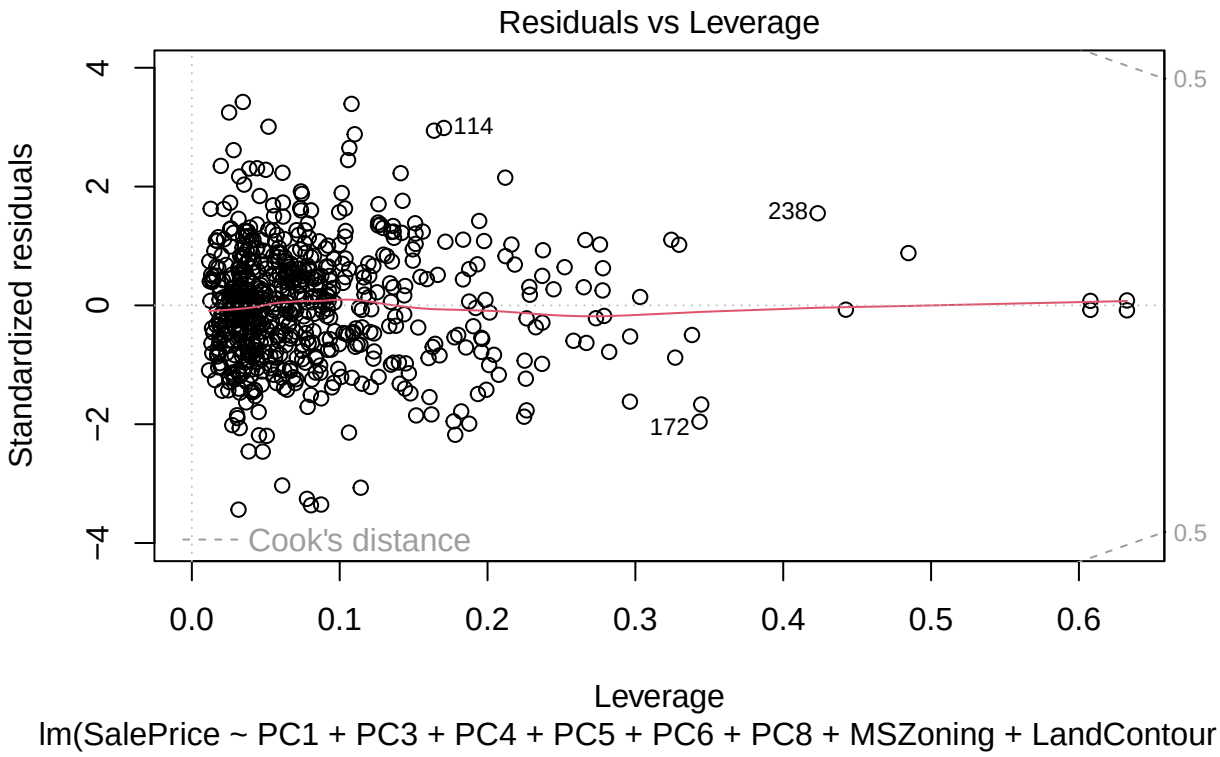
```
plot(olsMdl3)
```

```
## Warning: not plotting observations with leverage one:
##   333
```


Residuals vs Fitted
Residuals
60000
20000
−20000
−60000
29
301
94
50000
100000
150000
200000
250000
Fitted values
lm(SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +

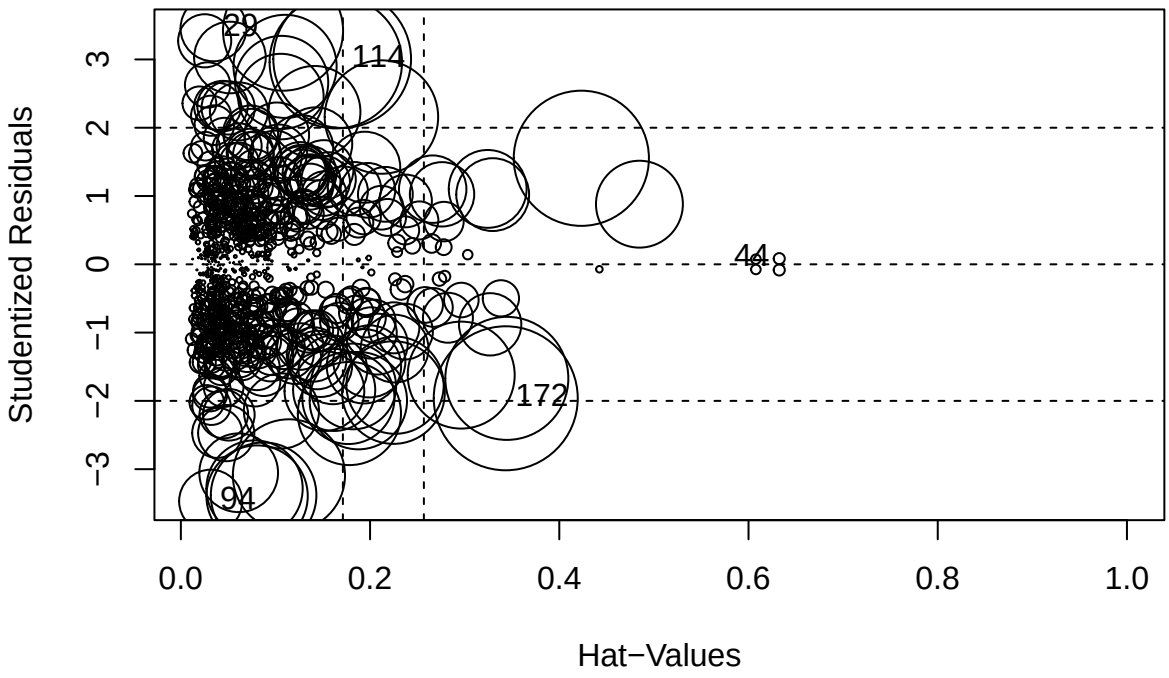

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +
689
29
94
4
2
0
−2
−3
−2
−1
0
1
2
3


Scale−Location
94
29
689
√|Standardized residuals|
0.0
0.5
1.0
1.5
50000
100000
150000
200000
250000
Fitted values
lm(SalePrice ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + MSZoning + LandContour +



- From the *Residuals vs Fitted* plot, we can see there are points above and below the 0 line.
- There is also a pattern seen like a `slight curvature pattern` which indicates that there maybe a systematic lack of fit.
- From the *Normal Q-Q* plot, we can see that most of the points are `very close to the dotted line`, indicating that the residuals follow a normal distribution, except some points which might be outliers which maybe affecting the regression line fit of data.
- Here the *Scale-Location* plot suggests that the red line is roughly horizontal across the plot and the spread of magnitude looks unequal, at some fitted values there are more residuals as compared to other like the ones in between 100000 and 150000, indicating some heteroskedasticity.
- From the *Residuals vs Leverage* plot, we can see that there are no influential points in our regression model. We need to check `influencePlot` to see if we are missing any leverage.

```
influencePlot(olsMdl3)
```




```
##         StudRes        Hat        CookD
## 29   3.45536218 0.03448629 0.0074787191
## 44   0.08314887 0.63248464 0.0002128236
## 94  -3.46964200 0.03156057 0.0068789688
## 114  3.00607347 0.17050932 0.0327310394
## 172 -1.96186593 0.34340931 0.0357772780
## 372         NaN 1.00000000          NaN
```

- We can now see some high influential points for the fitted values - 741, 684, 712.

```
#ncv Test
ncvTest(olsMdl3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 27.33894, Df = 1, p = 1.7074e-07
```

Since `p-value` is less than significance level (α) of `0.05`, that means we `reject the null hypothesis` of constant error variance which indicates heteroscedasticity.

```
VIF(olsMdl3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## PC1             3.282478  1        1.811761
```

```
## PC3             3.071509  1        1.752572
## PC4             1.793490  1        1.339212
## PC5             1.885068  1        1.372978
## PC6             1.514187  1        1.230523
## PC8             1.385452  1        1.177052
## MSZoning        2.382067  3        1.155648
## LandContour     2.671815  3        1.177970
## LotConfig       1.476468  3        1.067097
## LandSlope       3.307846  2        1.348610
## Condition1      1.586586  4        1.059395
## BldgType        4.659022  4        1.212096
## RoofStyle       1.504389  2        1.107491
## Exterior1st  4778.808279  4        2.883467
## Exterior2nd  4839.900411  4        2.888049
## ExterQual       3.238195  2        1.341454
## ExterCond       1.661982  2        1.135420
## BsmtFinType1    4.286295  4        1.199528
## BsmtFinType2    2.209995  4        1.104203
## Heating         1.265426  1        1.124912
## Functional      3.553579  5        1.135185
## PavedDrive      1.572996  2        1.119907
```

Generally, VIF values which are greater than 5 or 7 are the cause of multicollinearity which we do not see in our model.

Improving the current model:
* To improve our model, we need to remove some influential observations from our model and then fit the regression model to the data.
* We can re-build the model with new predictors.
* We can also perform variable transformation such as Box-Cox or use better evolved models like SVR, PCR etc., and see how it works.

References

1. https://rpubs.com/staneaurelius/house_price_prediction