Quiz 4

Chaeeun Shin

03/05/2024

```
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## Warning: package 'stringr' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)
library(bestglm)
library(MASS)
```

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

#Part 1

```
data1 <- read.csv("Ames_Housing_Data.csv")
data1$CentralAir <- ifelse(data1$CentralAir=='Y',1,0)
str(data1)
```

```
## 'data.frame':    1460 obs. of  20 variables:
##  $ Id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ LotArea     : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ OverallQual : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond : int  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 ...
##  $ CentralAir  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ X1stFlrSF   : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF   : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ GrLivArea   : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ 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 ...
##  $ TotRmsAbvGrd: int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Fireplaces  : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ GarageCars  : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea  : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ YrSold      : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SalePrice   : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

#1-1. Splitting training and testing data

```
set.seed(123)
training_samples <- data1$SalePrice %>%
  createDataPartition(p=.75,list=FALSE)
train_data <- data1[training_samples,]
test_data <- data1[-training_samples,]
nrow(train_data)
```

```
## [1] 1097
```

```
nrow(test_data)
```

```
## [1] 363
```

#1-2. Stepwise variable selection

```
full <- lm(SalePrice~., data=train_data)
null <- lm(SalePrice~1, data=train_data)
n = nrow(train_data)
bestmodel1 <- stats::step(null,scope=list(lower=null,upper=full),direction='both',k=log(n))
```

```
## Start:  AIC=24805.6
## SalePrice ~ 1
##
##                Df  Sum of Sq        RSS   AIC
## + OverallQual   1 4.5678e+12 2.6399e+12 23711
## + GrLivArea     1 3.9340e+12 3.2737e+12 23947
## + GarageCars    1 2.9822e+12 4.2255e+12 24227
## + GarageArea    1 2.8966e+12 4.3111e+12 24249
## + X1stFlrSF     1 2.8508e+12 4.3569e+12 24260
## + FullBath      1 2.3153e+12 4.8924e+12 24388
## + TotRmsAbvGrd  1 2.2046e+12 5.0031e+12 24412
## + YearBuilt     1 1.9200e+12 5.2877e+12 24473
## + YearRemodAdd  1 1.8773e+12 5.3304e+12 24482
```

```
## + Fireplaces    1 1.5633e+12 5.6444e+12 24544
## + X2ndFlrSF     1 7.3966e+11 6.4680e+12 24694
## + LotArea       1 5.9110e+11 6.6166e+12 24719
## + HalfBath      1 5.4828e+11 6.6594e+12 24726
## + CentralAir    1 4.3689e+11 6.7708e+12 24744
## + BedroomAbvGr  1 2.5739e+11 6.9503e+12 24773
## + KitchenAbvGr  1 1.3210e+11 7.0756e+12 24792
## <none>                       7.2077e+12 24806
## + OverallCond   1 2.7572e+10 7.1801e+12 24808
## + YrSold        1 1.4182e+10 7.1935e+12 24810
## + Id            1 1.0171e+10 7.1975e+12 24811
##
## Step:  AIC=23710.75
## SalePrice ~ OverallQual
##
##                Df  Sum of Sq        RSS   AIC
## + GrLivArea     1 7.7987e+11 1.8600e+12 23334
## + X1stFlrSF     1 6.1545e+11 2.0244e+12 23426
## + TotRmsAbvGrd  1 3.8077e+11 2.2591e+12 23547
## + GarageArea    1 3.6672e+11 2.2731e+12 23554
## + LotArea       1 3.0237e+11 2.3375e+12 23584
## + GarageCars    1 2.9868e+11 2.3412e+12 23586
## + Fireplaces    1 1.9849e+11 2.4414e+12 23632
## + FullBath      1 1.5636e+11 2.4835e+12 23651
## + YearRemodAdd  1 5.9554e+10 2.5803e+12 23693
## + X2ndFlrSF     1 5.6070e+10 2.5838e+12 23694
## + YearBuilt     1 5.1999e+10 2.5879e+12 23696
## + BedroomAbvGr  1 5.0790e+10 2.5891e+12 23696
## + HalfBath      1 2.8967e+10 2.6109e+12 23706
## <none>                       2.6399e+12 23711
## + CentralAir    1 4.9566e+09 2.6349e+12 23716
## + YrSold        1 2.7451e+09 2.6371e+12 23717
## + Id            1 2.0545e+09 2.6378e+12 23717
## + KitchenAbvGr  1 3.0071e+08 2.6396e+12 23718
## + OverallCond   1 1.6954e+08 2.6397e+12 23718
## - OverallQual   1 4.5678e+12 7.2077e+12 24806
##
## Step:  AIC=23333.63
## SalePrice ~ OverallQual + GrLivArea
##
##                Df  Sum of Sq        RSS   AIC
## + X1stFlrSF     1 2.5018e+11 1.6098e+12 23182
## + X2ndFlrSF     1 2.2963e+11 1.6304e+12 23196
## + GarageArea    1 2.0246e+11 1.6575e+12 23214
## + YearBuilt     1 1.8228e+11 1.6777e+12 23228
## + GarageCars    1 1.4881e+11 1.7112e+12 23249
## + BedroomAbvGr  1 1.3008e+11 1.7299e+12 23261
## + LotArea       1 1.1039e+11 1.7496e+12 23274
## + YearRemodAdd  1 8.9774e+10 1.7702e+12 23286
## + KitchenAbvGr  1 3.9493e+10 1.8205e+12 23317
## + Fireplaces    1 3.3799e+10 1.8262e+12 23320
## + CentralAir    1 2.5828e+10 1.8342e+12 23325
## + TotRmsAbvGrd  1 1.8578e+10 1.8414e+12 23330
## + HalfBath      1 1.8117e+10 1.8419e+12 23330
```

```
## <none>                       1.8600e+12 23334
## + Id            1 1.6321e+09 1.8584e+12 23340
## + YrSold        1 7.7611e+08 1.8592e+12 23340
## + FullBath      1 1.5251e+08 1.8598e+12 23340
## + OverallCond   1 1.0533e+08 1.8599e+12 23341
## - GrLivArea     1 7.7987e+11 2.6399e+12 23711
## - OverallQual   1 1.4137e+12 3.2737e+12 23947
##
## Step:  AIC=23182.16
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF
##
##                Df  Sum of Sq        RSS   AIC
## + YearBuilt     1 1.4114e+11 1.4687e+12 23088
## + GarageArea    1 1.1360e+11 1.4962e+12 23109
## + YearRemodAdd  1 8.9261e+10 1.5205e+12 23127
## + GarageCars    1 8.7310e+10 1.5225e+12 23128
## + BedroomAbvGr  1 8.4463e+10 1.5253e+12 23130
## + KitchenAbvGr  1 6.8749e+10 1.5411e+12 23141
## + LotArea       1 5.7740e+10 1.5521e+12 23149
## + CentralAir    1 1.6956e+10 1.5928e+12 23178
## + HalfBath      1 1.6947e+10 1.5929e+12 23178
## + TotRmsAbvGrd  1 1.4013e+10 1.5958e+12 23180
## + X2ndFlrSF     1 1.0678e+10 1.5991e+12 23182
## + Fireplaces    1 1.0413e+10 1.5994e+12 23182
## <none>                       1.6098e+12 23182
## + OverallCond   1 5.1535e+09 1.6046e+12 23186
## + Id            1 1.8068e+09 1.6080e+12 23188
## + YrSold        1 1.1149e+09 1.6087e+12 23188
## + FullBath      1 2.1842e+08 1.6096e+12 23189
## - X1stFlrSF     1 2.5018e+11 1.8600e+12 23334
## - GrLivArea     1 4.1460e+11 2.0244e+12 23426
## - OverallQual   1 1.1075e+12 2.7173e+12 23750
##
## Step:  AIC=23088.49
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt
##
##                Df  Sum of Sq        RSS   AIC
## + BedroomAbvGr  1 8.1383e+10 1.3873e+12 23033
## + KitchenAbvGr  1 6.5353e+10 1.4033e+12 23046
## + LotArea       1 6.1704e+10 1.4070e+12 23048
## + OverallCond   1 5.7453e+10 1.4112e+12 23052
## + GarageArea    1 5.6093e+10 1.4126e+12 23053
## + GarageCars    1 3.0749e+10 1.4379e+12 23072
## + FullBath      1 2.8789e+10 1.4399e+12 23074
## + YearRemodAdd  1 2.5236e+10 1.4434e+12 23076
## + Fireplaces    1 1.7214e+10 1.4514e+12 23083
## + TotRmsAbvGrd  1 1.0446e+10 1.4582e+12 23088
## <none>                       1.4687e+12 23088
## + X2ndFlrSF     1 1.5606e+09 1.4671e+12 23094
## + Id            1 1.2813e+09 1.4674e+12 23094
## + HalfBath      1 7.6763e+08 1.4679e+12 23095
## + CentralAir    1 6.5649e+08 1.4680e+12 23095
## + YrSold        1 5.9223e+08 1.4681e+12 23095
## - YearBuilt     1 1.4114e+11 1.6098e+12 23182
```

```
## - X1stFlrSF     1 2.0904e+11 1.6777e+12 23228
## - OverallQual   1 4.5892e+11 1.9276e+12 23380
## - GrLivArea     1 5.1073e+11 1.9794e+12 23409
##
## Step:  AIC=23032.96
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr
##
##                Df  Sum of Sq        RSS   AIC
## + LotArea       1 6.2522e+10 1.3248e+12 22989
## + OverallCond   1 6.1741e+10 1.3255e+12 22990
## + KitchenAbvGr  1 5.6535e+10 1.3307e+12 22994
## + GarageArea    1 4.5453e+10 1.3418e+12 23003
## + GarageCars    1 2.3081e+10 1.3642e+12 23022
## + FullBath      1 1.6326e+10 1.3709e+12 23027
## + YearRemodAdd  1 1.5700e+10 1.3716e+12 23028
## + Fireplaces    1 1.0635e+10 1.3766e+12 23032
## <none>                       1.3873e+12 23033
## + X2ndFlrSF     1 3.0503e+09 1.3842e+12 23038
## + CentralAir    1 1.7023e+09 1.3856e+12 23039
## + TotRmsAbvGrd  1 1.0514e+09 1.3862e+12 23039
## + YrSold        1 7.3566e+08 1.3865e+12 23039
## + Id            1 5.6152e+08 1.3867e+12 23040
## + HalfBath      1 4.2095e+07 1.3872e+12 23040
## - BedroomAbvGr  1 8.1383e+10 1.4687e+12 23088
## - YearBuilt     1 1.3806e+11 1.5253e+12 23130
## - X1stFlrSF     1 1.6923e+11 1.5565e+12 23152
## - OverallQual   1 3.6408e+11 1.7514e+12 23282
## - GrLivArea     1 5.5910e+11 1.9464e+12 23397
##
## Step:  AIC=22989.37
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr + LotArea
##
##                Df  Sum of Sq        RSS   AIC
## + OverallCond   1 5.4224e+10 1.2705e+12 22950
## + KitchenAbvGr  1 4.6223e+10 1.2785e+12 22957
## + GarageArea    1 3.7101e+10 1.2877e+12 22965
## + GarageCars    1 1.7448e+10 1.3073e+12 22982
## + YearRemodAdd  1 1.5016e+10 1.3097e+12 22984
## + FullBath      1 1.4826e+10 1.3099e+12 22984
## <none>                       1.3248e+12 22989
## + Fireplaces    1 4.0513e+09 1.3207e+12 22993
## + X2ndFlrSF     1 2.2243e+09 1.3225e+12 22994
## + TotRmsAbvGrd  1 2.1955e+09 1.3226e+12 22994
## + CentralAir    1 6.6297e+08 1.3241e+12 22996
## + YrSold        1 3.9053e+08 1.3244e+12 22996
## + Id            1 2.0118e+08 1.3246e+12 22996
## + HalfBath      1 6.5241e+07 1.3247e+12 22996
## - LotArea       1 6.2522e+10 1.3873e+12 23033
## - BedroomAbvGr  1 8.2201e+10 1.4070e+12 23048
## - X1stFlrSF     1 1.2633e+11 1.4511e+12 23082
## - YearBuilt     1 1.4200e+11 1.4667e+12 23094
## - OverallQual   1 3.8668e+11 1.7114e+12 23263
```

```
## - GrLivArea     1 4.9887e+11 1.8236e+12 23333
##
## Step:  AIC=22950.52
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr + LotArea + OverallCond
##
##                Df  Sum of Sq        RSS   AIC
## + KitchenAbvGr  1 3.5189e+10 1.2353e+12 22927
## + GarageArea    1 3.5147e+10 1.2354e+12 22927
## + GarageCars    1 1.7736e+10 1.2528e+12 22942
## + FullBath      1 1.3148e+10 1.2574e+12 22946
## <none>                       1.2705e+12 22950
## + Fireplaces    1 3.6154e+09 1.2669e+12 22954
## + TotRmsAbvGrd  1 3.2341e+09 1.2673e+12 22955
## + CentralAir    1 2.2357e+09 1.2683e+12 22956
## + X2ndFlrSF     1 1.4131e+09 1.2691e+12 22956
## + YearRemodAdd  1 1.3546e+09 1.2692e+12 22956
## + YrSold        1 1.0520e+09 1.2695e+12 22957
## + Id            1 3.3072e+08 1.2702e+12 22957
## + HalfBath      1 5.2399e+06 1.2705e+12 22958
## - OverallCond   1 5.4224e+10 1.3248e+12 22989
## - LotArea       1 5.5005e+10 1.3255e+12 22990
## - BedroomAbvGr  1 8.6190e+10 1.3567e+12 23016
## - X1stFlrSF     1 1.4095e+11 1.4115e+12 23059
## - YearBuilt     1 1.9195e+11 1.4625e+12 23098
## - OverallQual   1 3.0762e+11 1.5781e+12 23181
## - GrLivArea     1 5.2521e+11 1.7957e+12 23323
##
## Step:  AIC=22926.71
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr + LotArea + OverallCond + KitchenAbvGr
##
##                Df  Sum of Sq        RSS   AIC
## + GarageArea    1 3.7118e+10 1.1982e+12 22900
## + GarageCars    1 2.1811e+10 1.2135e+12 22914
## + TotRmsAbvGrd  1 1.2850e+10 1.2225e+12 22922
## <none>                       1.2353e+12 22927
## + CentralAir    1 6.2453e+09 1.2291e+12 22928
## + FullBath      1 5.7701e+09 1.2296e+12 22929
## + X2ndFlrSF     1 2.6668e+09 1.2327e+12 22931
## + Fireplaces    1 1.2374e+09 1.2341e+12 22933
## + YearRemodAdd  1 1.1883e+09 1.2342e+12 22933
## + YrSold        1 6.9978e+08 1.2346e+12 22933
## + Id            1 4.5523e+08 1.2349e+12 22933
## + HalfBath      1 1.3264e+08 1.2352e+12 22934
## - KitchenAbvGr  1 3.5189e+10 1.2705e+12 22950
## - OverallCond   1 4.3189e+10 1.2785e+12 22957
## - LotArea       1 4.7178e+10 1.2825e+12 22961
## - BedroomAbvGr  1 7.8317e+10 1.3137e+12 22987
## - X1stFlrSF     1 1.5813e+11 1.3935e+12 23052
## - YearBuilt     1 1.8035e+11 1.4157e+12 23069
## - OverallQual   1 2.5296e+11 1.4883e+12 23124
## - GrLivArea     1 5.4864e+11 1.7840e+12 23323
##
```

```
## Step:  AIC=22900.25
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr + LotArea + OverallCond + KitchenAbvGr + GarageArea
##
##                 Df  Sum of Sq        RSS   AIC
## + TotRmsAbvGrd   1 1.0361e+10 1.1879e+12 22898
## <none>                        1.1982e+12 22900
## + CentralAir     1 7.3618e+09 1.1909e+12 22900
## + FullBath       1 5.8701e+09 1.1924e+12 22902
## + X2ndFlrSF      1 2.9525e+09 1.1953e+12 22904
## + Fireplaces     1 1.8239e+09 1.1964e+12 22906
## + YearRemodAdd   1 1.4747e+09 1.1967e+12 22906
## + Id             1 8.1008e+08 1.1974e+12 22906
## + YrSold         1 5.2751e+08 1.1977e+12 22907
## + GarageCars     1 8.2145e+07 1.1981e+12 22907
## + HalfBath       1 2.8317e+07 1.1982e+12 22907
## - GarageArea     1 3.7118e+10 1.2353e+12 22927
## - KitchenAbvGr   1 3.7159e+10 1.2354e+12 22927
## - LotArea        1 3.9966e+10 1.2382e+12 22929
## - OverallCond    1 4.1145e+10 1.2394e+12 22930
## - BedroomAbvGr   1 6.8461e+10 1.2667e+12 22954
## - YearBuilt      1 1.2425e+11 1.3225e+12 23002
## - X1stFlrSF      1 1.2515e+11 1.3234e+12 23002
## - OverallQual    1 2.1877e+11 1.4170e+12 23077
## - GrLivArea      1 4.8757e+11 1.6858e+12 23268
##
## Step:  AIC=22897.72
## SalePrice ~ OverallQual + GrLivArea + X1stFlrSF + YearBuilt +
##     BedroomAbvGr + LotArea + OverallCond + KitchenAbvGr + GarageArea +
##     TotRmsAbvGrd
##
##                 Df  Sum of Sq        RSS   AIC
## <none>                        1.1879e+12 22898
## + CentralAir     1 7.3575e+09 1.1805e+12 22898
## + FullBath       1 5.5856e+09 1.1823e+12 22900
## - TotRmsAbvGrd   1 1.0361e+10 1.1982e+12 22900
## + X2ndFlrSF      1 3.3058e+09 1.1846e+12 22902
## + Fireplaces     1 1.8307e+09 1.1860e+12 22903
## + YearRemodAdd   1 1.0114e+09 1.1868e+12 22904
## + Id             1 9.4975e+08 1.1869e+12 22904
## + YrSold         1 5.3972e+08 1.1873e+12 22904
## + GarageCars     1 1.7513e+08 1.1877e+12 22905
## + HalfBath       1 5.1538e+07 1.1878e+12 22905
## - GarageArea     1 3.4628e+10 1.2225e+12 22922
## - LotArea        1 4.1413e+10 1.2293e+12 22928
## - OverallCond    1 4.1555e+10 1.2294e+12 22928
## - KitchenAbvGr   1 4.5535e+10 1.2334e+12 22932
## - BedroomAbvGr   1 7.7845e+10 1.2657e+12 22960
## - X1stFlrSF      1 1.2544e+11 1.3133e+12 23001
## - YearBuilt      1 1.2784e+11 1.3157e+12 23003
## - OverallQual    1 2.1117e+11 1.3990e+12 23070
## - GrLivArea      1 2.4514e+11 1.4330e+12 23096
```

```
bestmodel1$call
```

```
## lm(formula = SalePrice ~ OverallQual + GrLivArea + X1stFlrSF +
##     YearBuilt + BedroomAbvGr + LotArea + OverallCond + KitchenAbvGr +
##     GarageArea + TotRmsAbvGrd, data = train_data)
```

```
bestmodel1$coefficients
```

```
##   (Intercept)   OverallQual     GrLivArea     X1stFlrSF     YearBuilt
## -1.069990e+06  1.664525e+04  6.456689e+01  3.694842e+01  5.073670e+02
##  BedroomAbvGr       LotArea   OverallCond  KitchenAbvGr    GarageArea
## -1.464607e+04  7.430582e-01  6.103115e+03 -3.214772e+04  3.523431e+01
##  TotRmsAbvGrd
##  4.054139e+03
```

#1-2. Prediction: bestmodel1

```
pred1 <- predict(bestmodel1,newdata=test_data)
rmse1 <- sqrt(mean((pred1-test_data$SalePrice)^2))
r_squared1 <- cor(test_data$SalePrice,pred1)^2
print(paste("RMSE: ",rmse1))
```

```
## [1] "RMSE:  48820.768871431"
```

```
print(paste("R^2: ",r_squared1))
```

```
## [1] "R^2:  0.659587940625723"
```

#1-3. Best subset selection

```
bestmodel2 <- regsubsets(SalePrice~., data=train_data,nvmax=ncol(train_data)-1,method="exhaustive")
bestmodel2_summary <- summary(bestmodel2)
bestmodel2_size <- which.min(bestmodel2_summary$rss)
coef(bestmodel2,id=bestmodel2_size)
```

```
##   (Intercept)            Id       LotArea   OverallQual   OverallCond
## -2.723513e+05 -2.319630e+00  7.055228e-01  1.647482e+04  6.329801e+03
##     YearBuilt  YearRemodAdd    CentralAir     X1stFlrSF     X2ndFlrSF
##  5.848652e+02  1.102337e+02 -1.422130e+04  7.598137e+01  4.277682e+01
##     GrLivArea      FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr
##  2.741528e+01 -8.982671e+03 -3.842912e+03 -1.325852e+04 -3.084836e+04
##  TotRmsAbvGrd    Fireplaces    GarageCars    GarageArea        YrSold
##  4.004474e+03  3.203474e+03 -1.270691e+03  4.001495e+01 -5.724910e+02
```

```
bestmodel2_formula <- as.formula(paste("SalePrice~",paste(names(coef(bestmodel2,id=bestmodel2_size))[-1
print(bestmodel2_formula)
```

```
## SalePrice ~ Id + LotArea + OverallQual + OverallCond + YearBuilt +
##     YearRemodAdd + CentralAir + X1stFlrSF + X2ndFlrSF + GrLivArea +
##     FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd +
##     Fireplaces + GarageCars + GarageArea + YrSold
```

#1-3. Prediction: bestmodel2

```
fitted_model <- lm(bestmodel2_formula, data=train_data)
pred2 <- predict(fitted_model, test_data)
rmse2 <- sqrt(mean((pred2-test_data$SalePrice)^2))
r_squared <- cor(test_data$SalePrice,pred2)^2
print(paste("RMSE: ",rmse2))
```

```
## [1] "RMSE:  49437.9773926597"
```

```
print(paste("R^2: ",r_squared))
```

```
## [1] "R^2:  0.654938262613752"
```

#1-4. Compare BIC : Smaller BIC is better

```
(bic1 <- BIC(bestmodel1))
```

```
## [1] 26017.87
```

```
(bic2 <- BIC(fitted_model))
```

```
## [1] 26055.94
```

#Part 2. #2-1.

```
data2 <- read.csv("Titanic2.csv")
data2 <- data2[,-c(1,4,9,11)]
data2 <- data2[!is.na(data2$Age),]
cat("There are",nrow(data2),"passengers left.")
```

```
## There are 714 passengers left.
```

#2-2.

```
str(data2)
```

```
## 'data.frame':    714 obs. of  8 variables:
##  $ Survived: int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 1 3 3 2 3 ...
##  $ Sex     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
##  $ SibSp   : int  1 1 0 1 0 0 3 0 1 1 ...
##  $ Parch   : int  0 0 0 0 0 0 1 2 0 1 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: chr  "S" "C" "S" "S" ...
```

```
data2$Survived <- as.factor(data2$Survived)
data2$Pclass <- as.factor(data2$Pclass)
```

#2-3.

```
set.seed(123)
training_samples <- data2$Survived %>%
  createDataPartition(p=0.8,list=FALSE)
train_data <- data2[training_samples,]
test_data <- data2[-training_samples,]
nrow(train_data)
```

```
## [1] 572
```

```
nrow(test_data)
```

```
## [1] 142
```

#2-4.

```
log_model <- glm(Survived~., data=train_data, family=binomial)
summary(log_model)
```

```
##
```

```
## Call:
## glm(formula = Survived ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5369  -0.6665  -0.4107   0.6445   2.4178
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.128749   0.604277   6.833 8.34e-12 ***
## Pclass2     -1.012074   0.370945  -2.728 0.006365 **
## Pclass3     -2.076496   0.384989  -5.394 6.90e-08 ***
## Sexmale     -2.556944   0.242699 -10.535  < 2e-16 ***
## Age         -0.031944   0.009200  -3.472 0.000516 ***
## SibSp       -0.350734   0.144212  -2.432 0.015012 *
## Parch       -0.115459   0.139457  -0.828 0.407717
## Fare         0.002723   0.003202   0.851 0.395005
## EmbarkedQ   -1.080480   0.641079  -1.685 0.091910 .
## EmbarkedS   -0.718802   0.320819  -2.241 0.025057 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 772.45  on 571  degrees of freedom
## Residual deviance: 518.19  on 562  degrees of freedom
## AIC: 538.19
##
## Number of Fisher Scoring iterations: 5
```

#2-5

```
probabilities <- log_model %>%
  predict(test_data, type="response")
log_pred <- as.factor(ifelse(probabilities>0.5,1,0))
confusionMatrix(log_pred,test_data$Survived,positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 71 15
##          1 13 43
##
##                Accuracy : 0.8028
##                  95% CI : (0.7278, 0.8648)
##     No Information Rate : 0.5915
##     P-Value [Acc > NIR] : 6.975e-08
##
##                   Kappa : 0.5898
##
##  Mcnemar's Test P-Value : 0.8501
##
##             Sensitivity : 0.7414
##             Specificity : 0.8452
```

```
##          Pos Pred Value : 0.7679
##          Neg Pred Value : 0.8256
##              Prevalence : 0.4085
##          Detection Rate : 0.3028
##    Detection Prevalence : 0.3944
##       Balanced Accuracy : 0.7933
##
##        'Positive' Class : 1
##
```

#2-6

```
new <- test_data[1,]
new$Pclass <- '3'
new$Sex <- 'male'
new$Age <- 23
new$SibSp <- 0
new$Fare <- 8.25
new$Embarked <- 'Q'
p <- log_model %>%
  predict(new, type='response')
ifelse(p>0.5,'Survived','Not survived')
```

```
##              1
## "Not survived"
```