```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

Q1. Please use the random seed 123 to divide the data into 75% training and 25% testing.

```r
set.seed(123)

ames_data <- read.csv("Ames_Housing_Data.csv")


if (any(is.na(ames_data))) {
  ames_data <- na.omit(ames_data)  # Remove rows with missing values
}

ames_data$CentralAir <- as.factor(ames_data$CentralAir)

train_index <- createDataPartition(ames_data$SalePrice, p = 0.75, list = FALSE)
train_data <- ames_data[train_index, ]
test_data <- ames_data[-train_index, ]
```

Q2. Please find the best model using the stepwise variable selection method (based on the BIC criterion) using the training data.Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R.

```r
# 2. Stepwise variable selection using BIC criterion
# Fit the full model
full_model <- lm(SalePrice ~ ., data = train_data)
# Perform stepwise selection based on BIC
stepwise_model <- stepAIC(full_model, direction = "both", trace = FALSE, k = log(nrow(train_data)))

# (a) Display the coefficients of the fitted model
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenAbvGr +
##     TotRmsAbvGrd + GarageArea, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124802  -17854   -1674   15227  245317
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.017e+06  9.199e+04 -11.060  < 2e-16 ***
## LotArea       7.340e-01  1.207e-01   6.079 1.68e-09 ***
## OverallQual   1.667e+04  1.196e+03  13.943  < 2e-16 ***
## OverallCond   5.943e+03  9.889e+02   6.010 2.53e-09 ***
## YearBuilt     4.813e+02  4.667e+01  10.313  < 2e-16 ***
## X1stFlrSF     1.013e+02  4.872e+00  20.802  < 2e-16 ***
## X2ndFlrSF     6.478e+01  4.309e+00  15.033  < 2e-16 ***
## BedroomAbvGr -1.488e+04  1.737e+03  -8.569  < 2e-16 ***
## KitchenAbvGr -3.343e+04  4.972e+03  -6.723 2.88e-11 ***
## TotRmsAbvGrd  4.361e+03  1.302e+03   3.349  0.00084 ***
## GarageArea    3.560e+01  6.255e+00   5.691 1.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33050 on 1086 degrees of freedom
## Multiple R-squared:  0.8354, Adjusted R-squared:  0.8339
## F-statistic: 551.3 on 10 and 1086 DF,  p-value: < 2.2e-16
```

```
# (b) Make predictions on the testing data and report RMSE and R^2
stepwise_predictions <- predict(stepwise_model, newdata = test_data)
stepwise_rmse <- sqrt(mean((test_data$SalePrice - stepwise_predictions)^2))
stepwise_r2 <- cor(test_data$SalePrice, stepwise_predictions)^2

cat("Stepwise Model RMSE:", stepwise_rmse, "\n")
```

```
## Stepwise Model RMSE: 48727.68
```

```
cat("Stepwise Model R^2:", stepwise_r2, "\n")
```

```
## Stepwise Model R^2: 0.6601152
```

Q3. Please find the best model using the best subset variable selection method (based on the SSE criterion) using the training data. Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R2

```
subset_model <- regsubsets(SalePrice ~ ., data = train_data, nvmax = 20)

best_subset_index <- which.min(summary(subset_model)$bic)
best_subset_vars <- names(coef(subset_model, id = best_subset_index))[-1]  # Exclude intercept

best_subset_formula <- as.formula(paste("SalePrice ~", paste(best_subset_vars, collapse = " + ")))
best_subset_model <- lm(best_subset_formula, data = train_data)

cat("\nBest Subset Model Coefficients:\n")
```

```
##
## Best Subset Model Coefficients:
```

```
print(coef(best_subset_model))
```

```
##   (Intercept)       LotArea   OverallQual   OverallCond     YearBuilt
## -1.017464e+06  7.339645e-01  1.667461e+04  5.943478e+03  4.812917e+02
##     X1stFlrSF     X2ndFlrSF  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##  1.013470e+02  6.477953e+01 -1.488137e+04 -3.342536e+04  4.361293e+03
##    GarageArea
##  3.559750e+01
```

```
best_subset_predictions <- predict(best_subset_model, newdata = test_data)

best_subset_rmse <- sqrt(mean((test_data$SalePrice - best_subset_predictions)^2))
best_subset_r_squared <- cor(test_data$SalePrice, best_subset_predictions)^2

cat("\nBest Subset Model Performance:\n")
```

```
##
## Best Subset Model Performance:
```

```
cat("RMSE:", best_subset_rmse, "\n")
```

```
## RMSE: 48727.68
```

```
cat("R^2:", best_subset_r_squared, "\n")
```

```
## R^2: 0.6601152
```

Q4. Which model selection method among the 2 we have used above is the best? (a)Please compare the BIC of these models using the training data, as well as display these two models so we can see the parameter estimators and model goodness of fit measures. (b) Furthermore, please compare the RMSE and R2 of these models using the test data. (c) Please discuss any modifications you can do to further improve your model(s).

```
step_bic <- BIC(stepwise_model)
best_subset_bic <- BIC(best_subset_model)

cat("\nModel Comparison (BIC):\n")
```

```
##
## Model Comparison (BIC):
```

```
cat("Stepwise Model BIC:", step_bic, "\n")
```

```
## Stepwise Model BIC: 26016.31
```

```
cat("Best Subset Model BIC:", best_subset_bic, "\n")
```

```
## Best Subset Model BIC: 26016.31
```

```
cat("\nModel Comparison (Test Data):\n")
```

```
##
## Model Comparison (Test Data):
```

```
cat("Stepwise Model RMSE:", stepwise_rmse, "\n")
```

```
## Stepwise Model RMSE: 48727.68
```

```
cat("Best Subset Model RMSE:", best_subset_rmse, "\n")
```

```
## Best Subset Model RMSE: 48727.68
cat("Stepwise Model R^2:", stepwise_r2, "\n")

## Stepwise Model R^2: 0.6601152
cat("Best Subset Model R^2:", best_subset_r_squared, "\n")

## Best Subset Model R^2: 0.6601152
cat("\nPotential Improvements:\n")

##
## Potential Improvements:
cat("1. Incorporate interaction or polynomial terms to model non-linear relationships.\n")

## 1. Incorporate interaction or polynomial terms to model non-linear relationships.
cat("2. Apply regularization methods such as Ridge or Lasso regression to address multicollinearity and

## 2. Apply regularization methods such as Ridge or Lasso regression to address multicollinearity and p
cat("3. Perform feature engineering, including log transformations, to handle skewed predictors.\n")

## 3. Perform feature engineering, including log transformations, to handle skewed predictors.
cat("4. Explore advanced machine learning models like Random Forest or Gradient Boosting for enhanced p

## 4. Explore advanced machine learning models like Random Forest or Gradient Boosting for enhanced pre
```