Final Project Methods

```
data_1990 <- read.csv("C:/Users/wildb/Documents/stonks/573_project/filtered_data.csv")
```

```
#install.packages("tidyr")
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
#install.packages("dplyr")
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## 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
```

```
#install.packages("zoo")
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
#install.packages("lubridate")
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
##
## Attaching package: 'lubridate'
```

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

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v readr   2.1.5
## v ggplot2 3.4.4     v stringr 1.5.1
## v purrr   1.0.2     v tibble  3.2.1
```

```
## -- 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
```

```
#install.packages("googledrive")
library(googledrive)
```

```
## Warning: package 'googledrive' was built under R version 4.3.3
```

```
#install.packages("data")
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.3.3
```

```
##
## Attaching package: 'data.table'
##
## The following object is masked from 'package:purrr':
##
##     transpose
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:zoo':
##
##     yearmon, yearqtr
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```r
#install.packages("fst")
library(fst)
```

```
## Warning: package 'fst' was built under R version 4.3.3
```

```r
#install.packages("getPass")
library(getPass)
```

```
## Warning: package 'getPass' was built under R version 4.3.3
```

```r
#install.packages("RPostgres")
library(RPostgres)
```

```
## Warning: package 'RPostgres' was built under R version 4.3.3
```

```r
#data_1990 <- read.csv("C:/Users/wildb/Documents/stonks/573_project/filtered_data.csv")

library(tidyr)
library(dplyr)

remove_sparse_columns <- function(data, threshold = 0.15) {
  # Calculate the number of rows
  num_rows <- nrow(data)

  # Calculate the maximum number of NA allowed per column
  max_na <- num_rows * threshold

  # Identify columns with NA counts greater than the threshold
  sparse_cols <- colSums(is.na(data)) > max_na

  # Remove sparse columns
  data[, !sparse_cols]
}

# Apply the function
data_1990 <- remove_sparse_columns(data_1990)


#Now fill down and up for empty values
data_1990 <- data_1990 %>%
  dplyr::group_by(permno) %>%
  fill(everything(), .direction = "downup") %>%
  dplyr::ungroup()


#Now we have to make all NA an average of the values in the column
#Since there's a lot of variation in the data between what's available for
#Different securities
#we do this then run the same code essentially but without the grouping
#to prioritize within permno average
data_1990 <- data_1990 %>%
```

```
  group_by(permno) %>%
  mutate(across(-c(date, yyyymm), ~if_else(is.na(.), mean(., na.rm = TRUE), .))) %>%
  ungroup()

#Now we do it generally
data_1990 <- data_1990 %>%
  mutate(across(-c(date, yyyymm), ~if_else(is.na(.), mean(., na.rm = TRUE), .)))



#Cols we keep
keep_cols <- !(names(data_1990) %in% c("date", "permno", "yyyymm"))

# Apply the filtering operation, excluding the specified columns
data_1990 <- data_1990[!is.infinite(rowSums(data_1990[, keep_cols, drop = FALSE])), ]


#Hopefully this should do nothing, because at this point we don't want any dropped columns
#Everything should have NO NA's in it!
colnames(data_1990)
```

```
##  [1] "permno"          "yyyymm"          "BidAskSpread"    "Coskewness"
##  [5] "DolVol"          "High52"          "Idiovol3F"       "IdioVolAHT"
##  [9] "MaxRet"          "PriceDelayRsq"   "PriceDelaySlope" "RealizedVol"
## [13] "ReturnSkew"      "ReturnSkew3F"    "VolMkt"          "betaVIX"
## [17] "zerotrade"       "zerotradeAlt1"   "STreversal"      "Price"
## [21] "adj_price"       "price_prc"       "Size"            "date"
```

```
data_1990 <- data_1990 %>% select_if(~ !any(is.na(.)))
colnames(data_1990)
```

```
##  [1] "permno"          "yyyymm"          "BidAskSpread"    "Coskewness"
##  [5] "DolVol"          "High52"          "Idiovol3F"       "IdioVolAHT"
##  [9] "MaxRet"          "PriceDelayRsq"   "PriceDelaySlope" "RealizedVol"
## [13] "ReturnSkew"      "ReturnSkew3F"    "VolMkt"          "betaVIX"
## [17] "zerotrade"       "zerotradeAlt1"   "STreversal"      "Price"
## [21] "adj_price"       "price_prc"       "Size"            "date"
```

```
library(dplyr)
library(lubridate)


avg_price_1990 <- data_1990 %>%
  filter(substr(yyyymm, 1, 4) == "1990") %>%
  group_by(permno) %>%
  summarise(avg_price_1990 = mean(adj_price, na.rm = TRUE))

# Join the average price data with the original data
data_with_avg_price <- left_join(data_1990, avg_price_1990, by = "permno")


avg_price_2000 <- data_1990 %>%
```

```
  filter(substr(yyyymm, 1, 4) == "2000") %>%
  group_by(permno) %>%
  summarise(avg_price_2000 = mean(adj_price, na.rm = TRUE))

# Join the average price data with the original data
data_with_avg_price <- left_join(data_with_avg_price, avg_price_2000, by = "permno")

avg_price_2010 <- data_1990 %>%
  filter(substr(yyyymm, 1, 4) == "2010") %>%
  group_by(permno) %>%
  summarise(avg_price_2010 = mean(adj_price, na.rm = TRUE))

# Join the average price data with the original data
data_with_avg_price <- left_join(data_with_avg_price, avg_price_2010, by = "permno")

#Get a year identifier, we'll use it with the percentage calculations
data_with_avg_price <- data_with_avg_price %>%
  mutate(YEAR = substr(yyyymm, 1, 4))


# Percentage change in price from 1990 to 2000
data_with_avg_price <- data_with_avg_price %>%
  group_by(permno, YEAR) %>%
  mutate(percentage_change_1990_2000 = ifelse(is.numeric(avg_price_2000) & is.numeric(avg_price_1990),
                                    ((avg_price_2000 - avg_price_1990) / avg_price_1990) * 100,
                                    NA))


# Percentage change in price from 2000 to 2010
data_with_avg_price <- data_with_avg_price %>%
  group_by(permno, YEAR) %>%
  mutate(percentage_change_2000_2010 = ifelse(is.numeric(avg_price_2000) & is.numeric(avg_price_2010),
                                              ((avg_price_2010 - avg_price_2000) / avg_price_2000) * 10
                                              NA))

# Percentage change in price from 1990 to 2010
data_with_avg_price <- data_with_avg_price %>%
  group_by(permno, YEAR) %>%
  mutate(percentage_change_1990_2010 = ifelse(is.numeric(avg_price_1990) & is.numeric(avg_price_2010),
                                              ((avg_price_2010 - avg_price_1990) / avg_price_1990) * 10
                                              NA))



#View(data_with_avg_price)

write.csv(data_with_avg_price, file = "cleaned_avg_perc_change.csv", row.names = FALSE)
#View(data_with_avg_price)
```

```
#GRABS SUBSETS

#Get a subset where 1990 to 2000 change is NOT empty
data_1990_2000 <- data_with_avg_price[!is.na(data_with_avg_price$percentage_change_1990_2000), ]
```

```r
#Need to drop rows that are not the year 1990, since we map 1990 to 2000, not both years to results of
data_1990_2000 <- data_1990_2000[data_1990_2000$YEAR == 1990, ]
#Need to drop rows that are not the year 1990, since we map 1990 to 2000, not both years to results of
data_1990_2000 <- data_1990_2000[, !(names(data_1990_2000) %in% c("avg_price_2010", "YEAR", "percentage_


#Get a subset where 2000 to 2010 change is NOT empty
data_2000_2010 <-data_with_avg_price[!is.na(data_with_avg_price$percentage_change_2000_2010), ]

data_2000_2010 <- data_2000_2010[data_2000_2010$YEAR == 2000, ]

data_2000_2010 <- data_2000_2010[, !(names(data_2000_2010) %in% c("avg_price_2010", "YEAR", "percentage_

#Get a subset where 1990 to 2010 change is NOT empty
data_1990_2010 <-data_with_avg_price[!is.na(data_with_avg_price$percentage_change_1990_2010), ]
data_1990_2010 <- data_1990_2010[data_1990_2010$YEAR == 1990, ]
data_1990_2010 <- data_1990_2010[, !(names(data_1990_2010) %in% c("avg_price_2010", "YEAR", "percentage_

#View(data_1990_2000)
```

```r
set.seed(1)
#install.packages("e1071")
library(e1071) # for sum()
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```r
#install.packages("caret")
library(caret) # for train/test split
```

```
## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
trainIndex <- createDataPartition(data_1990_2000$percentage_change_1990_2000, p = 0.15, list = FALSE)
trainData <- data_1990_2000[trainIndex, ]
testData <- data_1990_2000[-trainIndex, ]

# Create feature matrix and target vector for training data
x_train <- trainData[, -which(names(trainData) == "percentage_change_1990_2000")]
y_train <- trainData$percentage_change_1990_2000
# Create feature matrix and target vector for test data
x_test <- testData[, -which(names(testData) == "percentage_change_1990_2000")]
y_test <- testData$percentage_change_1990_2000
# Train the SVM model
```

```r
svm_model <- svm(x_train, y_train)
# Predict on the test data
predictions <- predict(svm_model, x_test)
# Calculate accuracy
accuracy <- sum(predictions == y_test) / length(y_test)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0
```

```r
mse <- mean((predictions - y_test)^2)
cat("MSE:", mse, "\n")
```

```
## MSE: 1990543
```

```r
set.seed(1)
#install.packages("e1071")
library(e1071) # for svm()
#install.packages("caret")
library(caret) # for train/test split
trainIndex <- createDataPartition(data_2000_2010$percentage_change_2000_2010, p = 0.15, list = FALSE)
trainData <- data_2000_2010[trainIndex, ]
testData <- data_2000_2010[-trainIndex, ]

# Create feature matrix and target vector for training data
x_train <- trainData[, -which(names(trainData) == "percentage_change_2000_2010")]
y_train <- trainData$percentage_change_2000_2010
# Create feature matrix and target vector for test data
x_test <- testData[, -which(names(testData) == "percentage_change_2000_2010")]
y_test <- testData$percentage_change_2000_2010
# Train the SVM model
svm_model <- svm(x_train, y_train)
# Predict on the test data
predictions <- predict(svm_model, x_test)
# Calculate accuracy
accuracy <- sum(predictions == y_test) / length(y_test)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0
```

```r
mse <- mean((predictions - y_test)^2)
cat("MSE:", mse, "\n")
```

```
## MSE: 107747.9
```

```r
set.seed(1)
#install.packages("e1071")
library(e1071) # for svm()
#install.packages("caret")
library(caret) # for train/test split
trainIndex <- createDataPartition(data_1990_2010$percentage_change_1990_2010, p = 0.15, list = FALSE)
```

```r
trainData <- data_1990_2010[trainIndex, ]
testData <- data_1990_2010[-trainIndex, ]

# Create feature matrix and target vector for training data
x_train <- trainData[, -which(names(trainData) == "percentage_change_1990_2010")]
y_train <- trainData$percentage_change_1990_2010
# Create feature matrix and target vector for test data
x_test <- testData[, -which(names(testData) == "percentage_change_1990_2010")]
y_test <- testData$percentage_change_1990_2010
# Train the SVM model
svm_model <- svm(x_train, y_train)
# Predict on the test data
predictions <- predict(svm_model, x_test)
# Calculate accuracy
accuracy <- sum(predictions == y_test) / length(y_test)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0
```

```r
mse <- mean((predictions - y_test)^2)
cat("MSE:", mse, "\n")
```

```
## MSE: 2672302
```

```r
library(gbm)
```

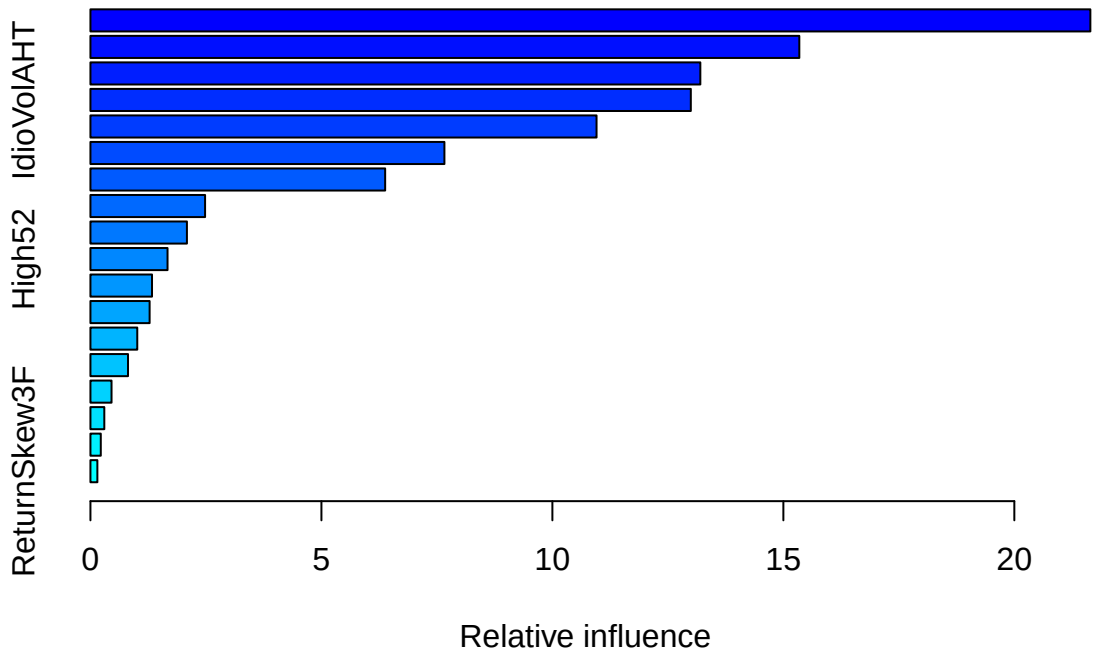
```
## Warning: package 'gbm' was built under R version 4.3.3
```

```
## Loaded gbm 2.1.9
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.co
```

```r
set.seed(1)
# Assuming all other columns are features and 'percentage_change_1990_2000' is the target
features <- setdiff(names(data_1990_2000), "percentage_change_1990_2000")
X <- data_1990_2000[, features]
y <- data_1990_2000$percentage_change_1990_2000
set.seed(1)  # for reproducibility
# Summary of the model
gbm_model <- gbm(percentage_change_1990_2000 ~ ., data=data_1990_2000, distribution="gaussian", n.trees=

# Summary of the model
summary(gbm_model)
```




```
##                              var    rel.inf
## PriceDelayRsq     PriceDelayRsq 21.6457190
## VolMkt                   VolMkt 15.3455076
## PriceDelaySlope PriceDelaySlope 13.2026247
## IdioVolAHT           IdioVolAHT 12.9965042
## zerotrade             zerotrade 10.9559952
## Coskewness           Coskewness  7.6627909
## Size                       Size  6.3801970
## BidAskSpread       BidAskSpread  2.4814031
## DolVol                   DolVol  2.0884856
## High52                   High52  1.6691052
## zerotradeAlt1     zerotradeAlt1  1.3326957
## RealizedVol         RealizedVol  1.2808605
## STreversal           STreversal  1.0139303
## betaVIX                 betaVIX  0.8134390
## MaxRet                   MaxRet  0.4564824
## IdioVol3F             IdioVol3F  0.3000919
## ReturnSkew           ReturnSkew  0.2242554
## ReturnSkew3F       ReturnSkew3F  0.1499121
```

```r
#data_matrix <- xgboost::xgb.DMatrix(data =
predictions <- predict(gbm_model, data_1990_2000, n.trees = 500)
mse <- mean((predictions - data_1990_2000$percentage_change_1990_2000)^2)
rmse <- sqrt(mse)

print(paste("MSE:", mse))
```

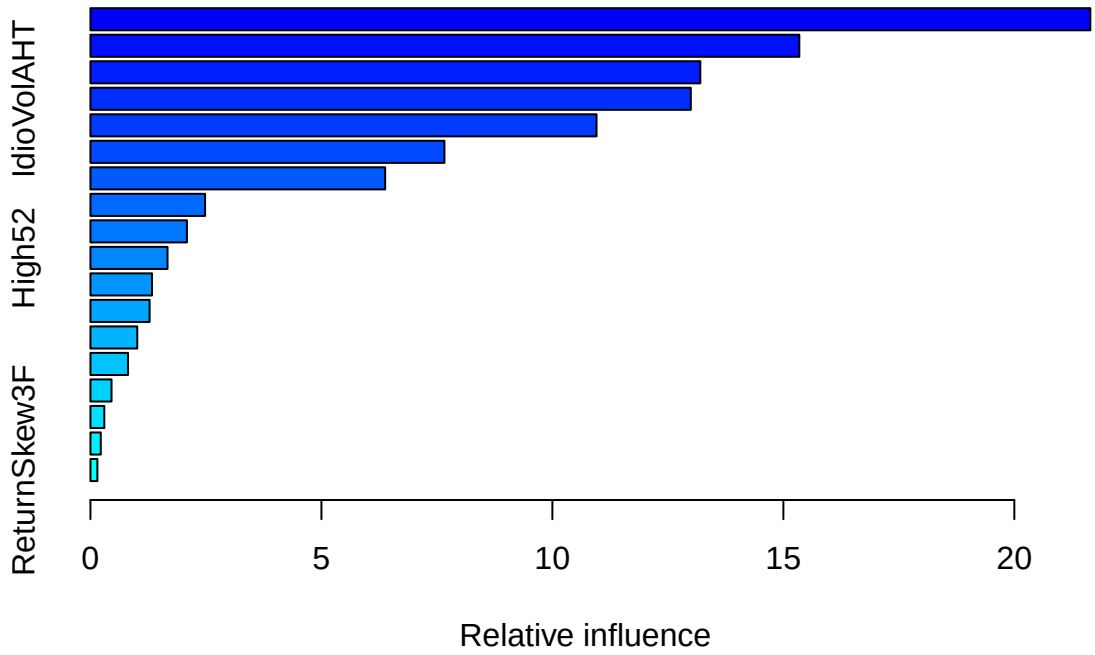
```
## [1] "MSE: 1731643.16561029"
```

```
print(paste("RMSE:", rmse))
```

```
## [1] "RMSE: 1315.91913338559"
```

```
library(gbm)
set.seed(1)
# Assuming all other columns are features and 'percentage_change_1990_2000' is the target
features <- setdiff(names(data_2000_2010), "percentage_change_1990_2000")
X <- data_2000_2010[, features]
y <- data_2000_2010$percentage_change_2000_2010
set.seed(1)  # for reproducibility
# Summary of the model


# Summary of the model
summary(gbm_model)
```




```
##                             var    rel.inf
## PriceDelayRsq     PriceDelayRsq 21.6457190
## VolMkt                   VolMkt 15.3455076
```

```
## PriceDelaySlope PriceDelaySlope 13.2026247
## IdioVolAHT           IdioVolAHT 12.9965042
## zerotrade             zerotrade 10.9559952
## Coskewness           Coskewness  7.6627909
## Size                       Size  6.3801970
## BidAskSpread       BidAskSpread  2.4814031
## DolVol                   DolVol  2.0884856
## High52                   High52  1.6691052
## zerotradeAlt1     zerotradeAlt1  1.3326957
## RealizedVol         RealizedVol  1.2808605
## STreversal           STreversal  1.0139303
## betaVIX                 betaVIX  0.8134390
## MaxRet                   MaxRet  0.4564824
## IdioVol3F             IdioVol3F  0.3000919
## ReturnSkew           ReturnSkew  0.2242554
## ReturnSkew3F       ReturnSkew3F  0.1499121
```

```
#data_matrix <- xgboost::xgb.DMatrix(data =
predictions <- predict(gbm_model, data_2000_2010, n.trees = 500)
mse <- mean((predictions - data_2000_2010$percentage_change_2000_2010)^2)
rmse <- sqrt(mse)

print(paste("MSE:", mse))
```

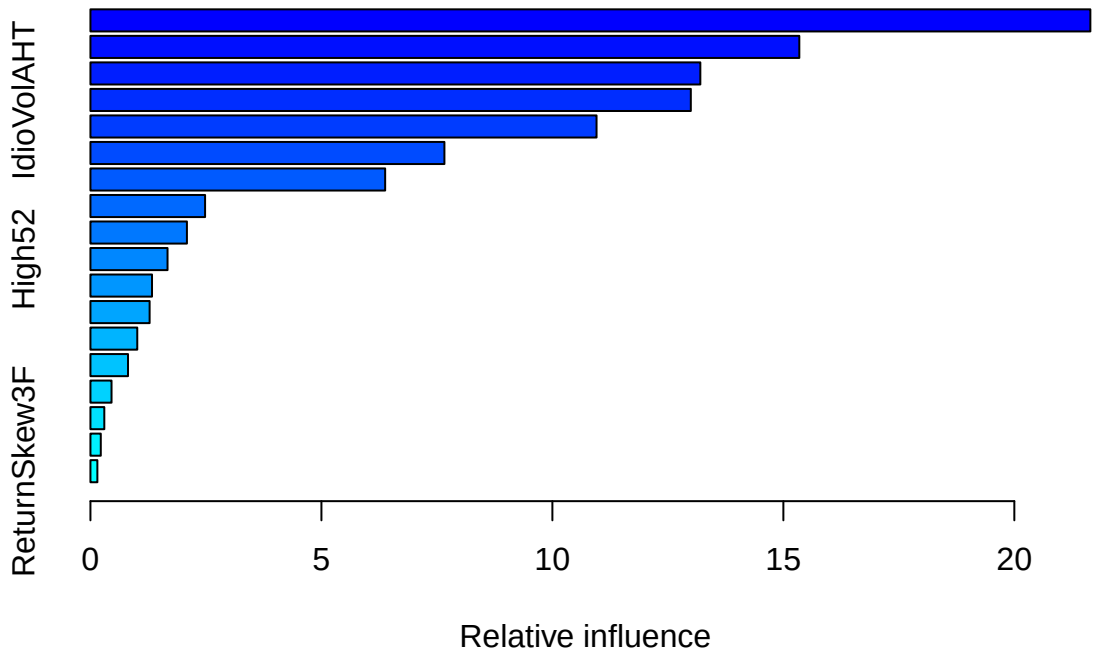
```
## [1] "MSE: 1886207.54633089"
```

```
print(paste("RMSE:", rmse))
```

```
## [1] "RMSE: 1373.39271380436"
```

```
library(gbm)
set.seed(1)
# Assuming all other columns are features and 'percentage_change_1990_2010' is the target
features <- setdiff(names(data_1990_2010), "percentage_change_1990_2010")
X <- data_1990_2010[, features]
y <- data_1990_2010$percentage_change_1990_2010
set.seed(1)  # for reproducibility
# Summary of the model


# Summary of the model
summary(gbm_model)
```




```
##                              var    rel.inf
## PriceDelayRsq     PriceDelayRsq 21.6457190
## VolMkt                   VolMkt 15.3455076
## PriceDelaySlope PriceDelaySlope 13.2026247
## IdioVolAHT           IdioVolAHT 12.9965042
## zerotrade             zerotrade 10.9559952
## Coskewness           Coskewness  7.6627909
## Size                       Size  6.3801970
## BidAskSpread       BidAskSpread  2.4814031
## DolVol                   DolVol  2.0884856
## High52                   High52  1.6691052
## zerotradeAlt1     zerotradeAlt1  1.3326957
## RealizedVol         RealizedVol  1.2808605
## STreversal           STreversal  1.0139303
## betaVIX                 betaVIX  0.8134390
## MaxRet                   MaxRet  0.4564824
## IdioVol3F             IdioVol3F  0.3000919
## ReturnSkew           ReturnSkew  0.2242554
## ReturnSkew3F       ReturnSkew3F  0.1499121
```

```
#data_matrix <- xgboost::xgb.DMatrix(data =
predictions <- predict(gbm_model, data_1990_2010, n.trees = 500)
mse <- mean((predictions - data_1990_2000$percentage_change_1990_2010)^2)
```

```
## Warning: Unknown or uninitialised column: 'percentage_change_1990_2010'.
```

```r
rmse <- sqrt(mse)

print(paste("MSE:", mse))
```

```
## [1] "MSE: NaN"
```

```r
print(paste("RMSE:", rmse))
```

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

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```r
# Perform LASSO regression for 1990 to 2000
x_1990_2000 <- as.matrix(data_1990_2000[, !(names(data_1990_2000) %in% "percentage_change_1990_2000")])
y_1990_2000 <- data_1990_2000$percentage_change_1990_2000
lasso_model_1990_2000 <- glmnet(x_1990_2000, y_1990_2000, alpha = 1)
cv_lasso_1990_2000 <- cv.glmnet(x_1990_2000, y_1990_2000, alpha = 1)
best_lambda_1990_2000 <- cv_lasso_1990_2000$lambda.min
lasso_best_1990_2000 <- glmnet(x_1990_2000, y_1990_2000, alpha = 1, lambda = best_lambda_1990_2000)
coef(lasso_best_1990_2000)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)       605.2666121
## BidAskSpread    -1082.8914479
## Coskewness        -42.9157085
## DolVol            -41.6741024
## High52            102.7868381
## IdioVol3F        6669.1135174
## IdioVolAHT      -7028.1816770
## MaxRet            497.1579269
## PriceDelayRsq    -144.4645495
## PriceDelaySlope    -0.0252276
## RealizedVol     -7181.8612533
## ReturnSkew         -3.6576683
## ReturnSkew3F       -7.8601328
## VolMkt            234.1691018
```

```
## betaVIX           756.1550329
## zerotrade          -6.1600417
## zerotradeAlt1       8.9622153
## STreversal         -2.4492021
## Size               40.2521986
```

```
# Perform LASSO regression for 2000 to 2010
x_2000_2010 <- as.matrix(data_2000_2010[, !(names(data_2000_2010) %in% "percentage_change_2000_2010")])
y_2000_2010 <- data_2000_2010$percentage_change_2000_2010
lasso_model_2000_2010 <- glmnet(x_2000_2010, y_2000_2010, alpha = 1)
cv_lasso_2000_2010 <- cv.glmnet(x_2000_2010, y_2000_2010, alpha = 1)
best_lambda_2000_2010 <- cv_lasso_2000_2010$lambda.min
lasso_best_2000_2010 <- glmnet(x_2000_2010, y_2000_2010, alpha = 1, lambda = best_lambda_2000_2010)
coef(lasso_best_2000_2010)
```

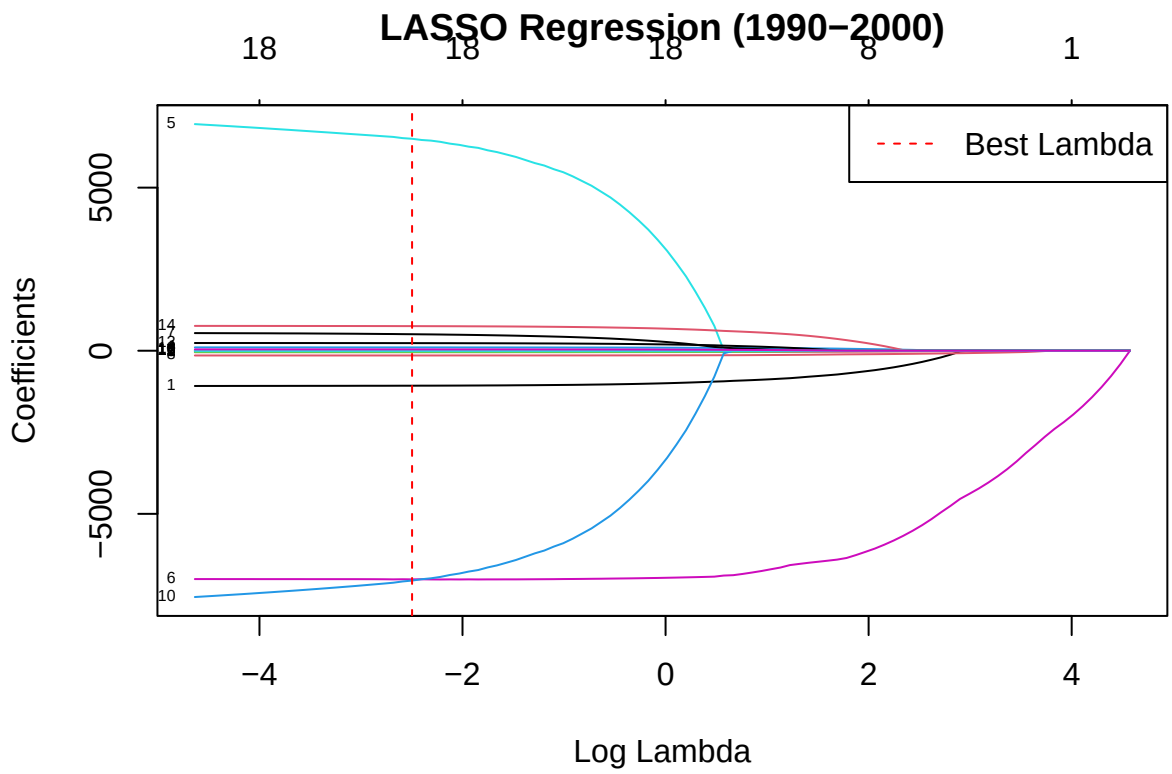
```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)      2.884664e+02
## BidAskSpread     6.268994e+02
## Coskewness       3.198936e+01
## DolVol           1.573116e+01
## High52          -4.204663e+01
## IdioVol3F       -2.445757e+03
## IdioVolAHT       4.030545e+02
## MaxRet           2.615575e+01
## PriceDelayRsq   -4.548751e+01
## PriceDelaySlope -1.675984e-02
## RealizedVol      1.983656e+03
## ReturnSkew       9.114841e-01
## ReturnSkew3F     1.321840e+00
## VolMkt           3.709491e+01
## betaVIX         -3.181206e+02
## zerotrade       -1.993662e-01
## zerotradeAlt1   -3.213020e+00
## STreversal      -3.494203e-01
## Size             7.726990e+00
```

```
# Perform LASSO regression for 1990 to 2010
x_1990_2010 <- as.matrix(data_1990_2010[, !(names(data_1990_2010) %in% "percentage_change_1990_2010")])
y_1990_2010 <- data_1990_2010$percentage_change_1990_2010
lasso_model_1990_2010 <- glmnet(x_1990_2010, y_1990_2010, alpha = 1)
cv_lasso_1990_2010 <- cv.glmnet(x_1990_2010, y_1990_2010, alpha = 1)
best_lambda_1990_2010 <- cv_lasso_1990_2010$lambda.min
lasso_best_1990_2010 <- glmnet(x_1990_2010, y_1990_2010, alpha = 1, lambda = best_lambda_1990_2010)
coef(lasso_best_1990_2010)
```

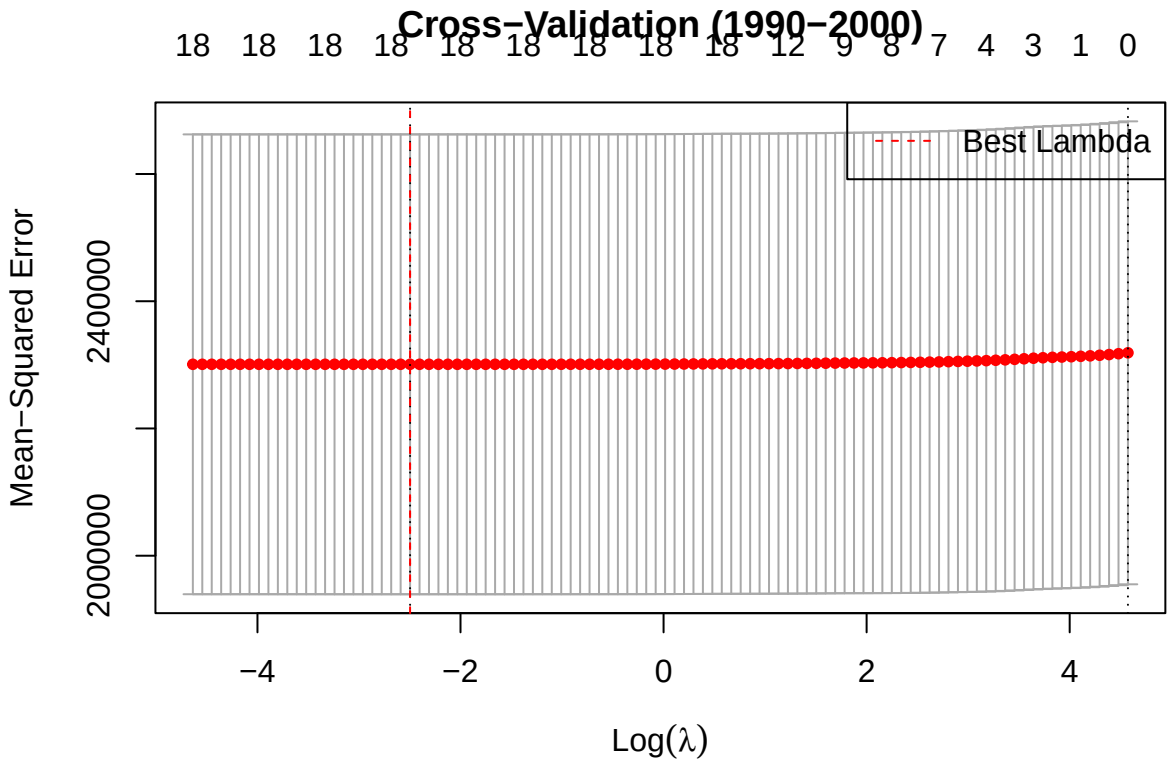
```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)      1.146960e+03
## BidAskSpread     2.542403e+02
## Coskewness       9.299306e+01
## DolVol          -5.620261e+01
## High52           1.685362e+02
```

```
## IdioVol3F         6.146116e+03
## IdioVolAHT       -9.109189e+03
## MaxRet            1.251621e+03
## PriceDelayRsq    -9.255003e+01
## PriceDelaySlope -5.909943e-02
## RealizedVol      -1.006188e+04
## ReturnSkew       -8.694692e+00
## ReturnSkew3F     -2.825483e+01
## VolMkt            4.695024e+02
## betaVIX           2.612484e+03
## zerotrade        -1.164587e+01
## zerotradeAlt1     1.203325e+01
## STreversal       -3.227663e+00
## Size              8.907226e+01
```

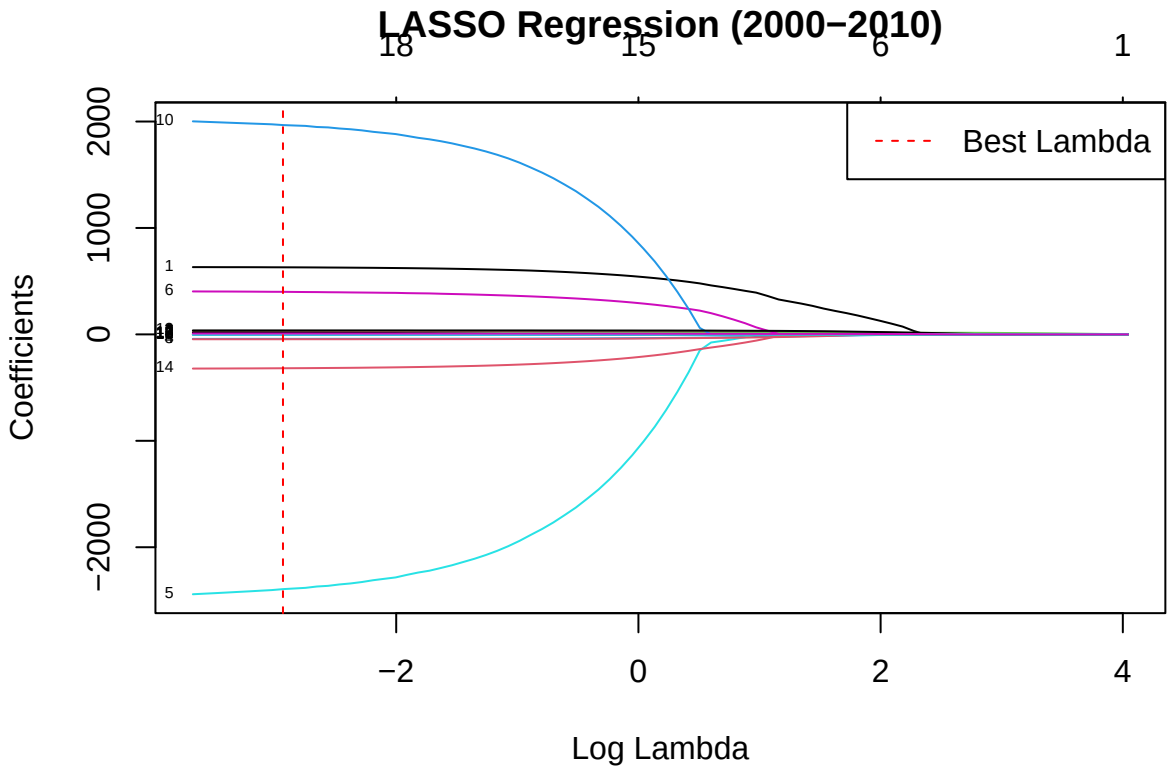
```r
# Plot the LASSO regression results for 1990 to 2000
plot(lasso_model_1990_2000, xvar = "lambda", label = TRUE, main = "LASSO Regression (1990-2000)")
abline(v = log(best_lambda_1990_2000), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```



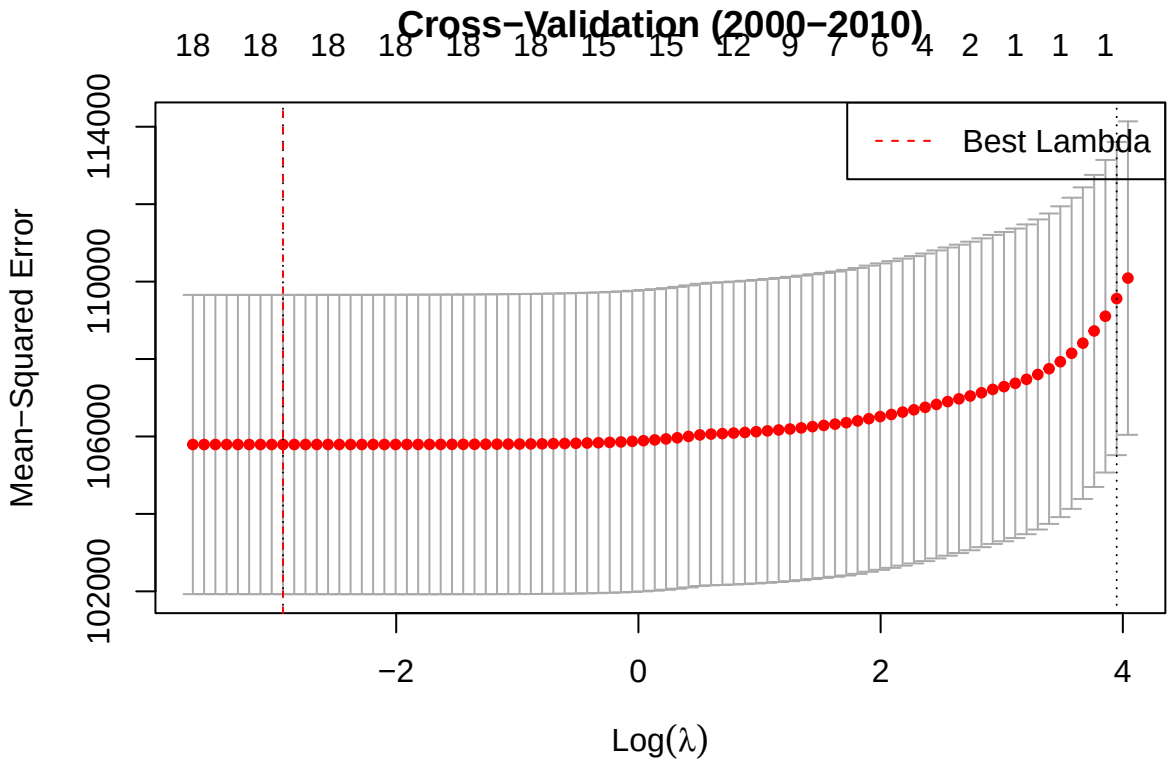

```r
# Plot the cross-validation results for 1990 to 2000
plot(cv_lasso_1990_2000, main = "Cross-Validation (1990-2000)")
abline(v = log(best_lambda_1990_2000), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```



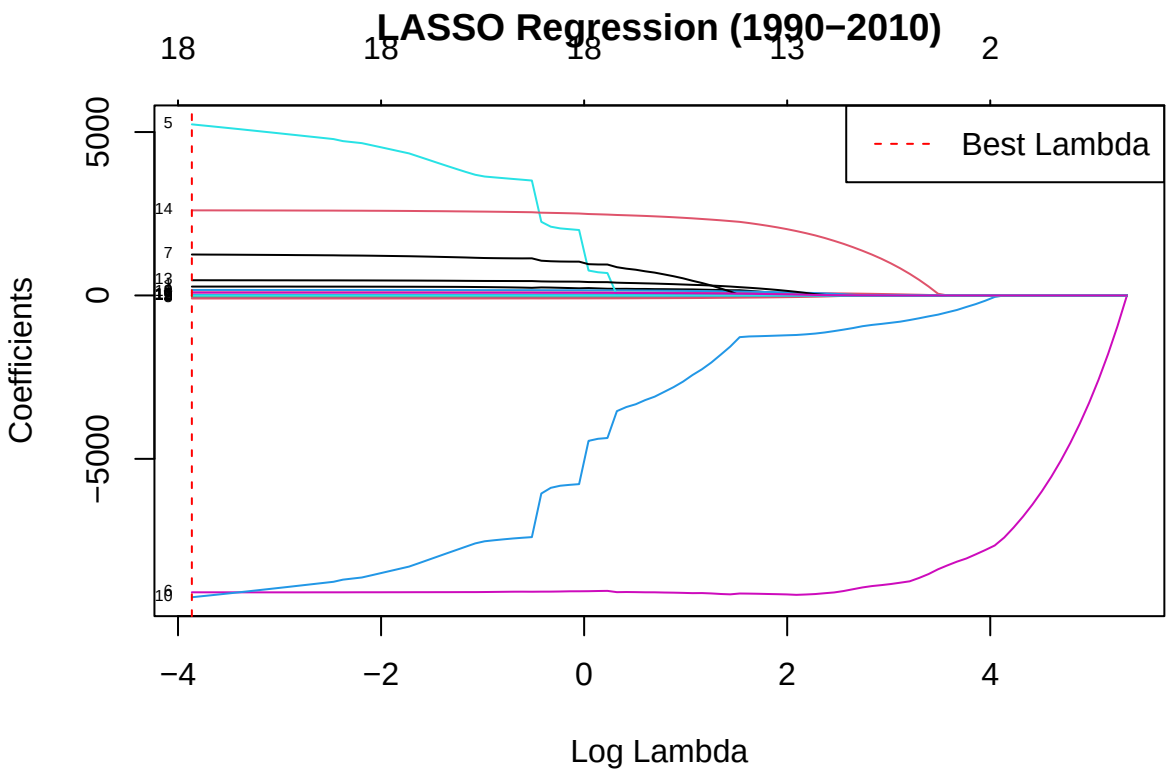

```
# Plot the LASSO regression results for 2000 to 2010
plot(lasso_model_2000_2010, xvar = "lambda", label = TRUE, main = "LASSO Regression (2000-2010)")
abline(v = log(best_lambda_2000_2010), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```



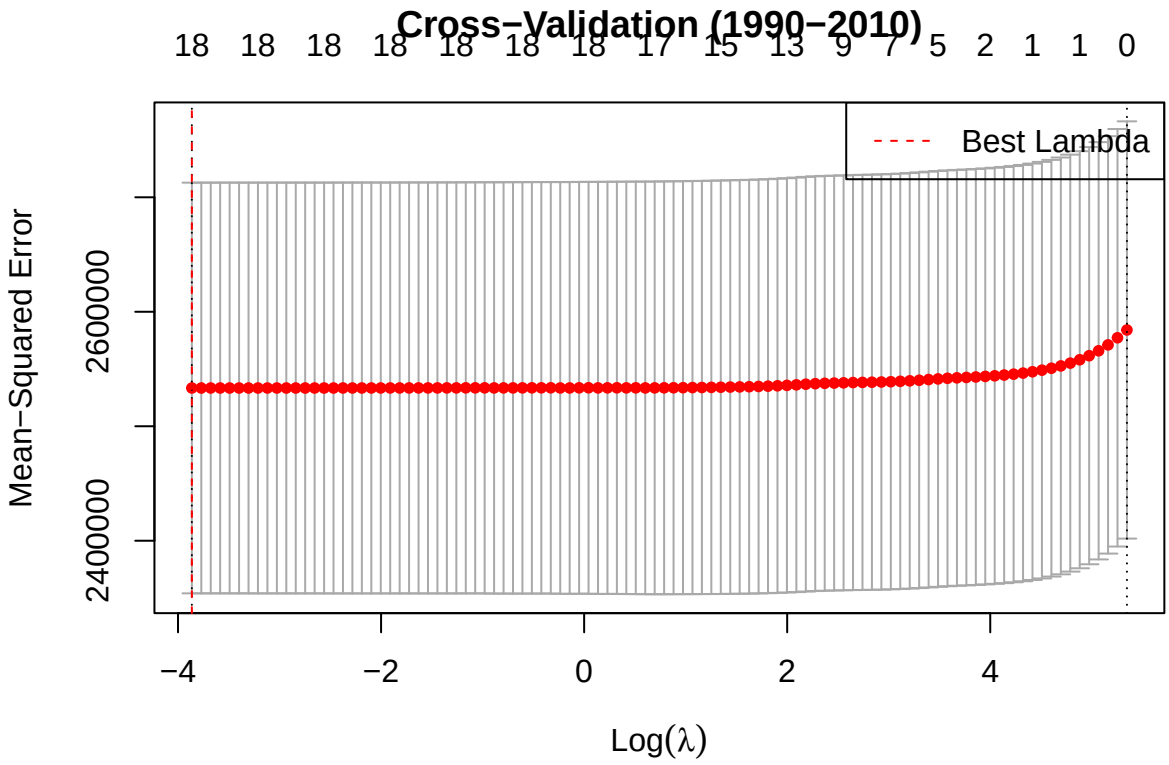

```
# Plot the cross-validation results for 2000 to 2010
plot(cv_lasso_2000_2010, main = "Cross-Validation (2000-2010)")
abline(v = log(best_lambda_2000_2010), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```




```r
# Plot the LASSO regression results for 1990 to 2010
plot(lasso_model_1990_2010, xvar = "lambda", label = TRUE, main = "LASSO Regression (1990-2010)")
abline(v = log(best_lambda_1990_2010), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```




```r
# Plot the cross-validation results for 1990 to 2010
plot(cv_lasso_1990_2010, main = "Cross-Validation (1990-2010)")
abline(v = log(best_lambda_1990_2010), col = "red", lty = 2)
legend("topright", legend = "Best Lambda", col = "red", lty = 2)
```




```r
x_1990_2000 <- data_1990_2000[, !(names(data_1990_2000) %in% "percentage_change_1990_2000")]
y_1990_2000 <- data_1990_2000$percentage_change_1990_2000
train_control_1990_2000 <- trainControl(method = "cv", number = 45)
knn_model_1990_2000 <- train(x_1990_2000, y_1990_2000, method = "knn", trControl = train_control_1990_2
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
print(knn_model_1990_2000)
```

```
## k-Nearest Neighbors
##
## 41069 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
## Summary of sample sizes: 40157, 40156, 40157, 40157, 40156, 40156, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared    MAE
##   5  1418.014  0.08510161  438.6380
##   7  1391.823  0.09313409  438.1590
##   9  1391.551  0.08304157  437.4274
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
# Perform KNN regression for 2000 to 2010
x_2000_2010 <- data_2000_2010[, !(names(data_2000_2010) %in% "percentage_change_2000_2010")]
y_2000_2010 <- data_2000_2010$percentage_change_2000_2010
train_control_2000_2010 <- trainControl(method = "cv", number = 45)
knn_model_2000_2010 <- train(x_2000_2010, y_2000_2010, method = "knn", trControl = train_control_2000_2
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
print(knn_model_2000_2010)
```

```
## k-Nearest Neighbors
##
## 46920 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
## Summary of sample sizes: 45876, 45877, 45877, 45877, 45878, 45876, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  315.9186  0.1285519  153.8316
##   7  309.6148  0.1395802  152.3364
##   9  306.8171  0.1440635  151.3620
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
# Perform KNN regression for 1990 to 2010
x_1990_2010 <- data_1990_2010[, !(names(data_1990_2010) %in% "percentage_change_1990_2010")]
y_1990_2010 <- data_1990_2010$percentage_change_1990_2010
train_control_1990_2010 <- trainControl(method = "cv", number = 45)
knn_model_1990_2010 <- train(x_1990_2010, y_1990_2010, method = "knn", trControl = train_control_1990_2
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
print(knn_model_1990_2010)
```

```
## k-Nearest Neighbors
##
## 20093 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
## Summary of sample sizes: 19647, 19646, 19646, 19647, 19646, 19648, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared    MAE
##   5  1551.151  0.10208789  679.6633
##   7  1532.338  0.09938889  681.1169
##   9  1528.827  0.09390334  682.7665
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
# Updated: Split each dataset into training and testing sets

library(caret)

# Perform KNN regression for 1990 to 2000
set.seed(1)
train_index_1990_2000 <- createDataPartition(data_1990_2000$percentage_change_1990_2000, p = 0.9, list =
train_1990_2000 <- data_1990_2000[train_index_1990_2000, ]
test_1990_2000 <- data_1990_2000[-train_index_1990_2000, ]
x_train_1990_2000 <- train_1990_2000[, !(names(train_1990_2000) %in% "percentage_change_1990_2000")]
y_train_1990_2000 <- train_1990_2000$percentage_change_1990_2000
train_control_1990_2000 <- trainControl(method = "cv", number = 45)
knn_model_1990_2000 <- train(x_train_1990_2000, y_train_1990_2000, method = "knn", trControl = train_co
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```r
print(knn_model_1990_2000)
```

```
## k-Nearest Neighbors
##
## 36964 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
## Summary of sample sizes: 36143, 36144, 36141, 36141, 36140, 36144, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared    MAE
##   5  1381.467  0.09619420  438.3583
##   7  1372.905  0.08164952  438.6767
##   9  1357.025  0.07819236  436.2905
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```r
x_test_1990_2000 <- test_1990_2000[, !(names(test_1990_2000) %in% "percentage_change_1990_2000")]
predictions_knn_1990_2000 <- predict(knn_model_1990_2000, newdata = x_test_1990_2000)
residuals_knn_1990_2000 <- test_1990_2000$percentage_change_1990_2000 - predictions_knn_1990_2000
mse_knn_1990_2000 <- mean(residuals_knn_1990_2000^2)
rmse_knn_1990_2000 <- sqrt(mse_knn_1990_2000)
cat("Mean Squared Error (KNN 1990-2000):", mse_knn_1990_2000, "\n")
```

```
## Mean Squared Error (KNN 1990-2000): 2904546
```

```r
cat("Root Mean Squared Error (KNN 1990-2000):", rmse_knn_1990_2000, "\n")
```

```
## Root Mean Squared Error (KNN 1990-2000): 1704.273
```

```r
# Perform KNN regression for 2000 to 2010
set.seed(1)
train_index_2000_2010 <- createDataPartition(data_2000_2010$percentage_change_2000_2010, p = 0.9, list =
train_2000_2010 <- data_2000_2010[train_index_2000_2010, ]
test_2000_2010 <- data_2000_2010[-train_index_2000_2010, ]
x_train_2000_2010 <- train_2000_2010[, !(names(train_2000_2010) %in% "percentage_change_2000_2010")]
y_train_2000_2010 <- train_2000_2010$percentage_change_2000_2010
train_control_2000_2010 <- trainControl(method = "cv", number = 45)
knn_model_2000_2010 <- train(x_train_2000_2010, y_train_2000_2010, method = "knn", trControl = train_co
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```r
print(knn_model_2000_2010)
```

```
## k-Nearest Neighbors
##
## 42230 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
```

```
## Summary of sample sizes: 41292, 41291, 41292, 41292, 41291, 41291, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  310.7995  0.1323178  153.3062
##   7  305.1150  0.1417268  151.5951
##   9  302.5902  0.1460361  150.9013
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
x_test_2000_2010 <- test_2000_2010[, !(names(test_2000_2010) %in% "percentage_change_2000_2010")]
predictions_knn_2000_2010 <- predict(knn_model_2000_2010, newdata = x_test_2000_2010)
residuals_knn_2000_2010 <- test_2000_2010$percentage_change_2000_2010 - predictions_knn_2000_2010
mse_knn_2000_2010 <- mean(residuals_knn_2000_2010^2)
rmse_knn_2000_2010 <- sqrt(mse_knn_2000_2010)
cat("Mean Squared Error (KNN 2000-2010):", mse_knn_2000_2010, "\n")
```

```
## Mean Squared Error (KNN 2000-2010): 124949.7
```

```
cat("Root Mean Squared Error (KNN 2000-2010):", rmse_knn_2000_2010, "\n")
```

```
## Root Mean Squared Error (KNN 2000-2010): 353.4822
```

```
# Perform KNN regression for 1990 to 2010
set.seed(1)
train_index_1990_2010 <- createDataPartition(data_1990_2010$percentage_change_1990_2010, p = 0.9, list =
train_1990_2010 <- data_1990_2010[train_index_1990_2010, ]
test_1990_2010 <- data_1990_2010[-train_index_1990_2010, ]
x_train_1990_2010 <- train_1990_2010[, !(names(train_1990_2010) %in% "percentage_change_1990_2010")]
y_train_1990_2010 <- train_1990_2010$percentage_change_1990_2010
train_control_1990_2010 <- trainControl(method = "cv", number = 45)
knn_model_1990_2010 <- train(x_train_1990_2010, y_train_1990_2010, method = "knn", trControl = train_co
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
print(knn_model_1990_2010)
```

```
## k-Nearest Neighbors
##
## 18085 samples
##    18 predictor
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (45 fold)
## Summary of sample sizes: 17684, 17683, 17683, 17684, 17683, 17683, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared    MAE
##   5  1568.579  0.08967684  691.7214
##   7  1547.413  0.08359648  689.6979
##   9  1542.891  0.07914185  691.4089
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
x_test_1990_2010 <- test_1990_2010[, !(names(test_1990_2010) %in% "percentage_change_1990_2010")]
predictions_knn_1990_2010 <- predict(knn_model_1990_2010, newdata = x_test_1990_2010)
residuals_knn_1990_2010 <- test_1990_2010$percentage_change_1990_2010 - predictions_knn_1990_2010
mse_knn_1990_2010 <- mean(residuals_knn_1990_2010^2)
rmse_knn_1990_2010 <- sqrt(mse_knn_1990_2010)
cat("Mean Squared Error (KNN 1990-2010):", mse_knn_1990_2010, "\n")
```

```
## Mean Squared Error (KNN 1990-2010): 2214319
```

```
cat("Root Mean Squared Error (KNN 1990-2010):", rmse_knn_1990_2010, "\n")
```

```
## Root Mean Squared Error (KNN 1990-2010): 1488.059
```

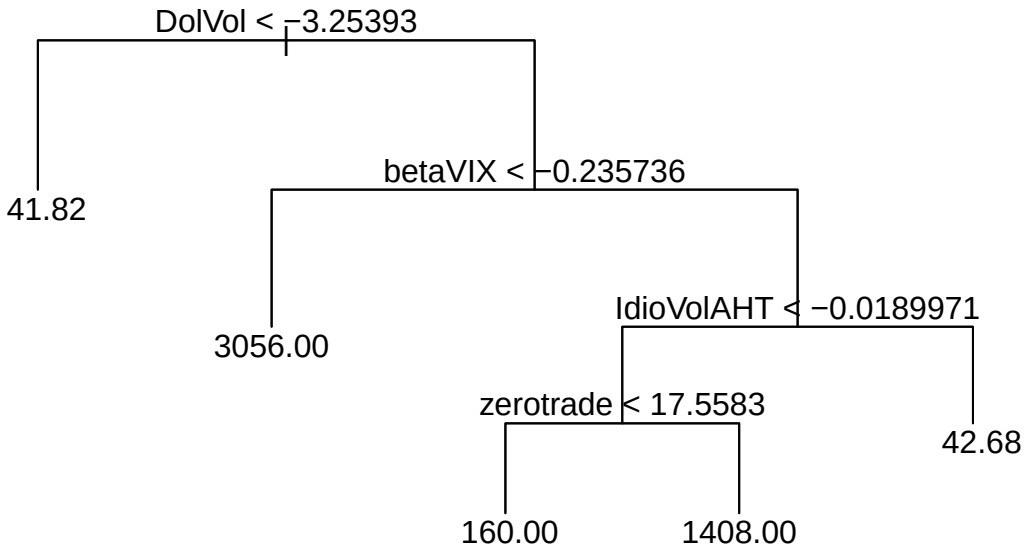
```
set.seed(1)
train1990_2000 = sample(1:nrow(data_1990_2000), 0.9*nrow(data_1990_2000)) #We must first make a trainin
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.3.3
```

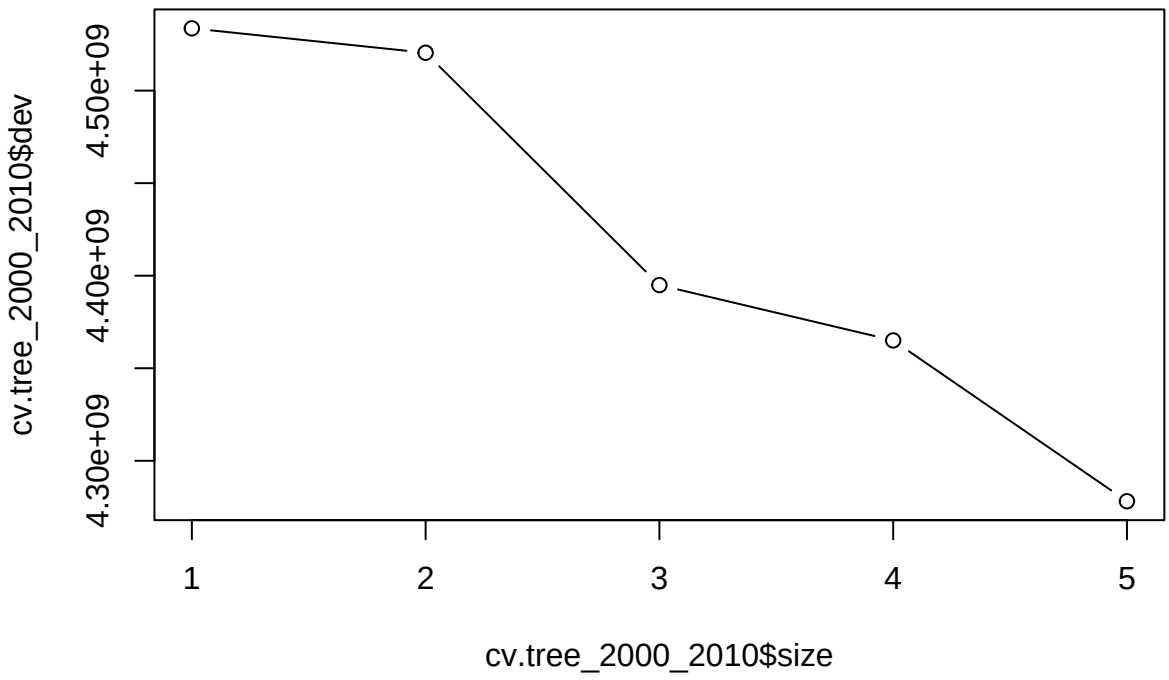
```
tree.data_1990_2000 <- tree(percentage_change_1990_2000 ~ ., data_1990_2000, subset= train1990_2000) #T
#plot(tree.data_1990_2000)
#text(tree.data_1990_2000)
summary(tree.data_1990_2000)
```

```
##
## Regression tree:
## tree(formula = percentage_change_1990_2000 ~ ., data = data_1990_2000,
##     subset = train1990_2000)
## Variables actually used in tree construction:
## character(0)
## Number of terminal nodes:  1
## Residual mean deviance:  2360000 = 8.721e+10 / 36960
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  -433.80  -348.00  -272.60     0.00   -75.69 59170.00
```

```
train2000_2010 = sample(1:nrow(data_2000_2010), 0.9*nrow(data_2000_2010))
tree.data_2000_2010 <- tree(percentage_change_2000_2010 ~ ., data_2000_2010, subset= train2000_2010)
plot(tree.data_2000_2010)
text(tree.data_2000_2010)
```



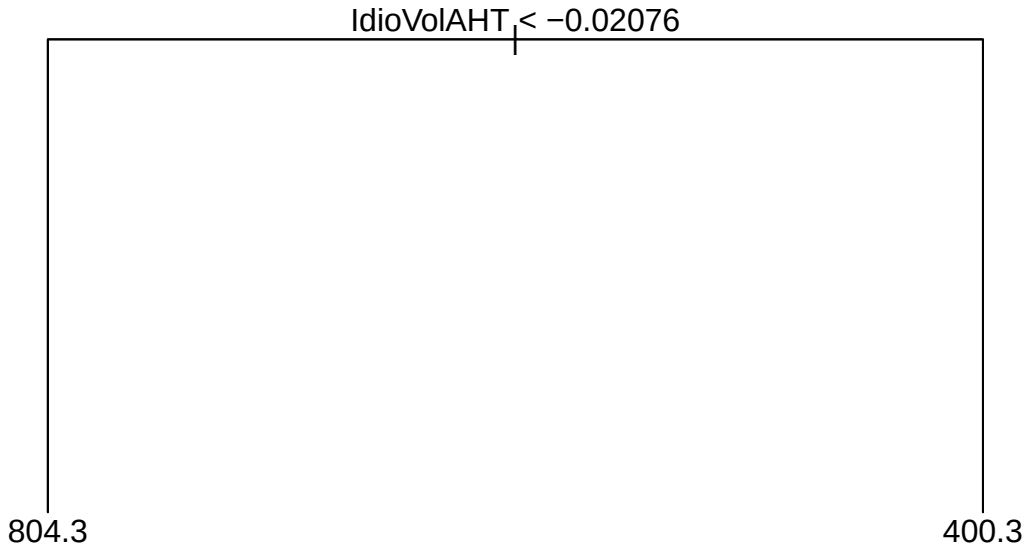

```
cv.tree_2000_2010 <- cv.tree(tree.data_2000_2010)
plot(cv.tree_2000_2010$size, cv.tree_2000_2010$dev, type = "b")
```



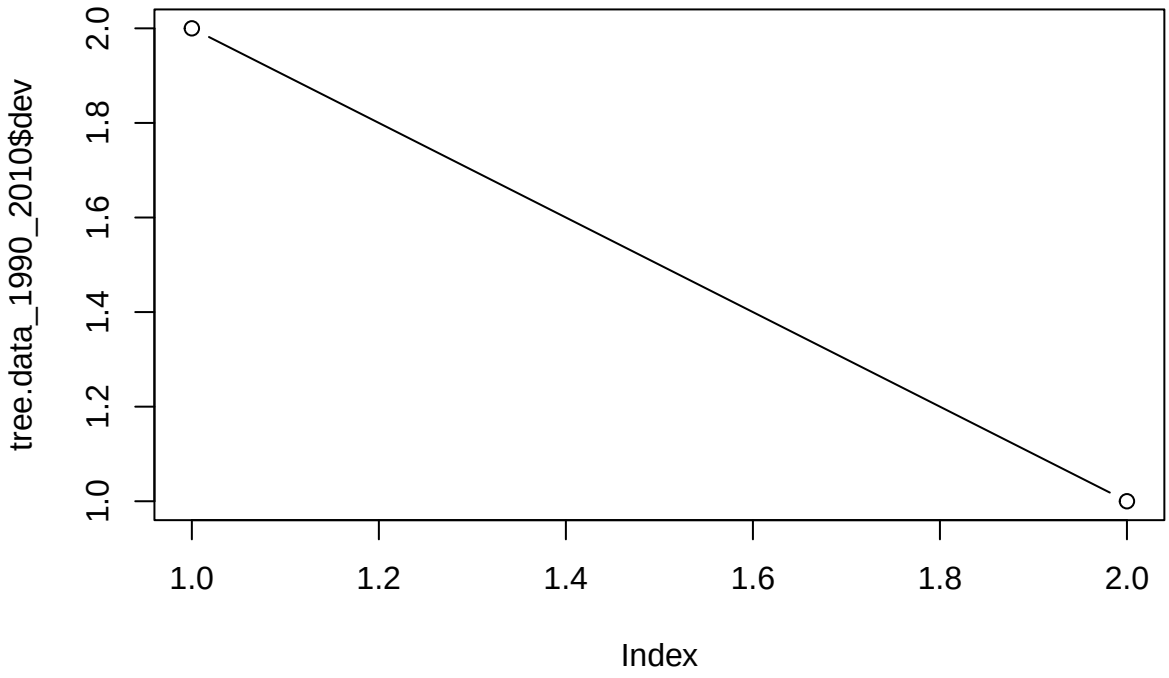

```
summary(tree.data_2000_2010)
```

```
##
## Regression tree:
## tree(formula = percentage_change_2000_2010 ~ ., data = data_2000_2010,
##     subset = train2000_2010)
## Variables actually used in tree construction:
## [1] "DolVol"     "betaVIX"    "IdioVolAHT" "zerotrade"
## Number of terminal nodes:  5
## Residual mean deviance:  100400 = 4.239e+09 / 42220
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1480.00  -134.10   -64.19     0.00    33.80  7676.00
```

```
train1990_2010 = sample(1:nrow(data_1990_2010), 0.9*nrow(data_1990_2010))
tree.data_1990_2010 <- tree(percentage_change_1990_2010 ~ ., data_1990_2010, subset= train1990_2010)
plot(tree.data_1990_2010)
text(tree.data_1990_2010)
```




```
cv.tree_1990_2010 <- cv.tree(tree.data_1990_2010)
plot(cv.tree_1990_2010$size, tree.data_1990_2010$dev, type = "b")
```




```
summary(tree.data_1990_2010)
```

```
##
## Regression tree:
## tree(formula = percentage_change_1990_2010 ~ ., data = data_1990_2010,
##     subset = train1990_2010)
## Variables actually used in tree construction:
## [1] "IdioVolAHT"
## Number of terminal nodes:  2
## Residual mean deviance:  2499000 = 4.519e+10 / 18080
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  -904.20  -632.00  -343.20     0.00    14.11 26380.00
```

```
#This cell and the following 2 cells include the code for random forests.
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

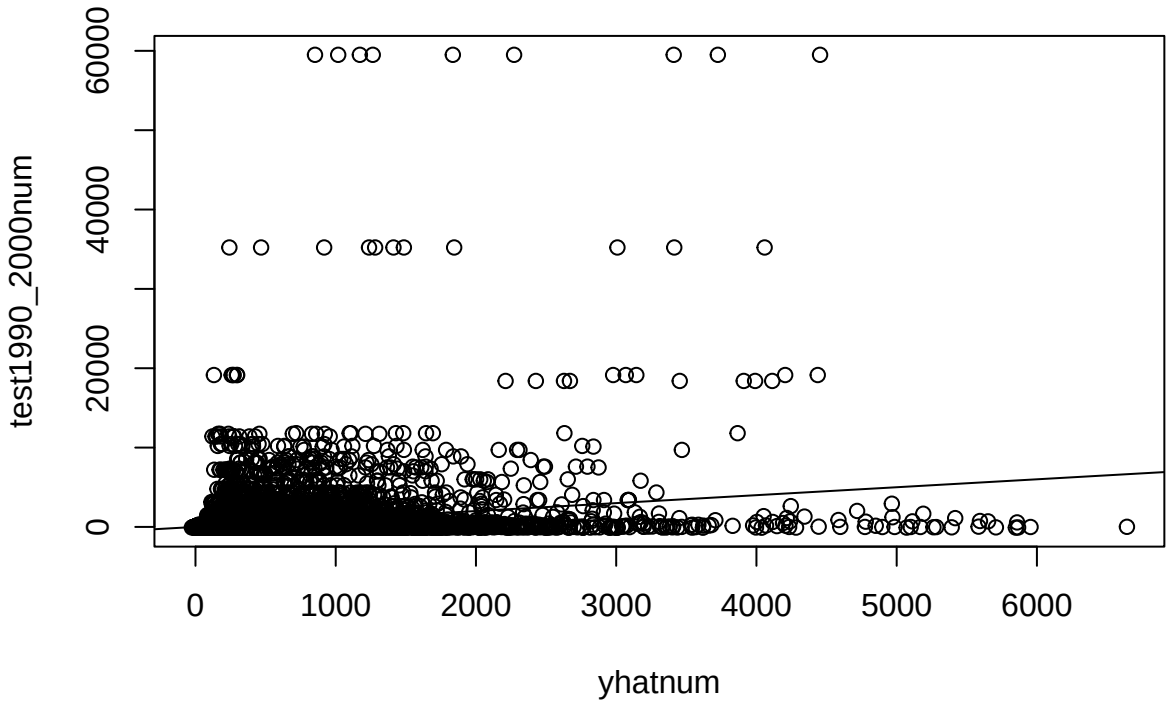
```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin

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

```r
#Random Forests 1990-2000
library(caret) # for train/test split
trainIndex <- createDataPartition(data_1990_2000$percentage_change_1990_2000, p = 0.15, list = FALSE)
trainData <- data_1990_2000[trainIndex, ]
testData <- data_1990_2000[-trainIndex, ]
set.seed(1)
bag.1990_2000 <- randomForest(percentage_change_1990_2000 ~ ., data = data_1990_2000, subset = trainInd
test1990_2000 <- data_1990_2000[-trainIndex, "percentage_change_1990_2000"]
yhat.bag <- predict(bag.1990_2000, newdata = data_1990_2000[-trainIndex, ])

test1990_2000matrix_data <- as.matrix(test1990_2000)
yhat.data.frame_data <- as.data.frame(yhat.bag)
yhat_data <- lapply(yhat.data.frame_data, as.numeric)
test1990_2000num <- as.numeric(test1990_2000matrix_data)
yhatnum <- as.numeric(unlist(yhat_data))
plot(yhatnum,test1990_2000num)
abline(0, 1)
```



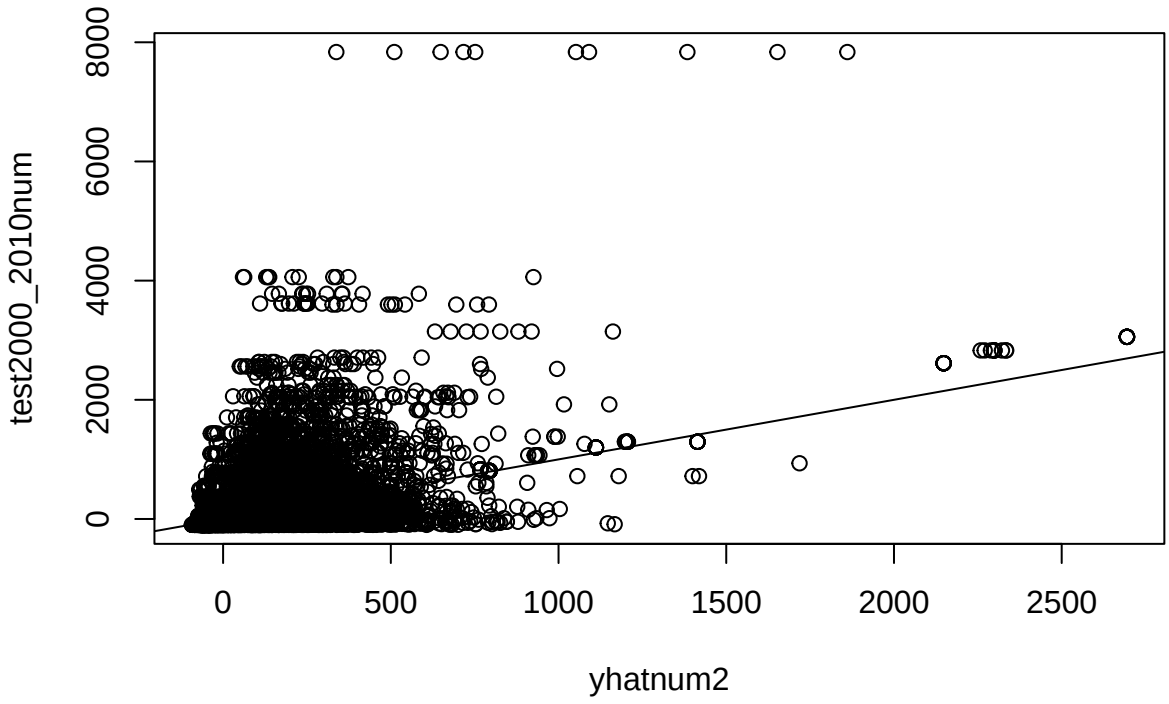
```r
print(mean((yhatnum - test1990_2000num)^2))
```

```
## [1] 2113724
```

```r
#Random Forest 2000-2010
trainIndex2 <- createDataPartition(data_2000_2010$percentage_change_2000_2010, p = 0.15, list = FALSE)
trainData2 <- data_2000_2010[trainIndex2, ]
testData2 <- data_2000_2010[-trainIndex2, ]

bag.2000_2010 <- randomForest(percentage_change_2000_2010 ~ ., data = data_2000_2010, subset = trainInd
test2000_2010 <- data_2000_2010[-trainIndex2, "percentage_change_2000_2010"]
yhat.bag2 <- predict(bag.2000_2010, newdata = data_2000_2010[-trainIndex2, ])

test2000_2010matrix_data <- as.matrix(test2000_2010)
yhat.data.frame_data2 <- as.data.frame(yhat.bag2)
yhat_data2 <- lapply(yhat.data.frame_data2, as.numeric)
test2000_2010num <- as.numeric(test2000_2010matrix_data)
yhatnum2 <- as.numeric(unlist(yhat_data2))
plot(yhatnum2,test2000_2010num)
abline(0, 1)
```



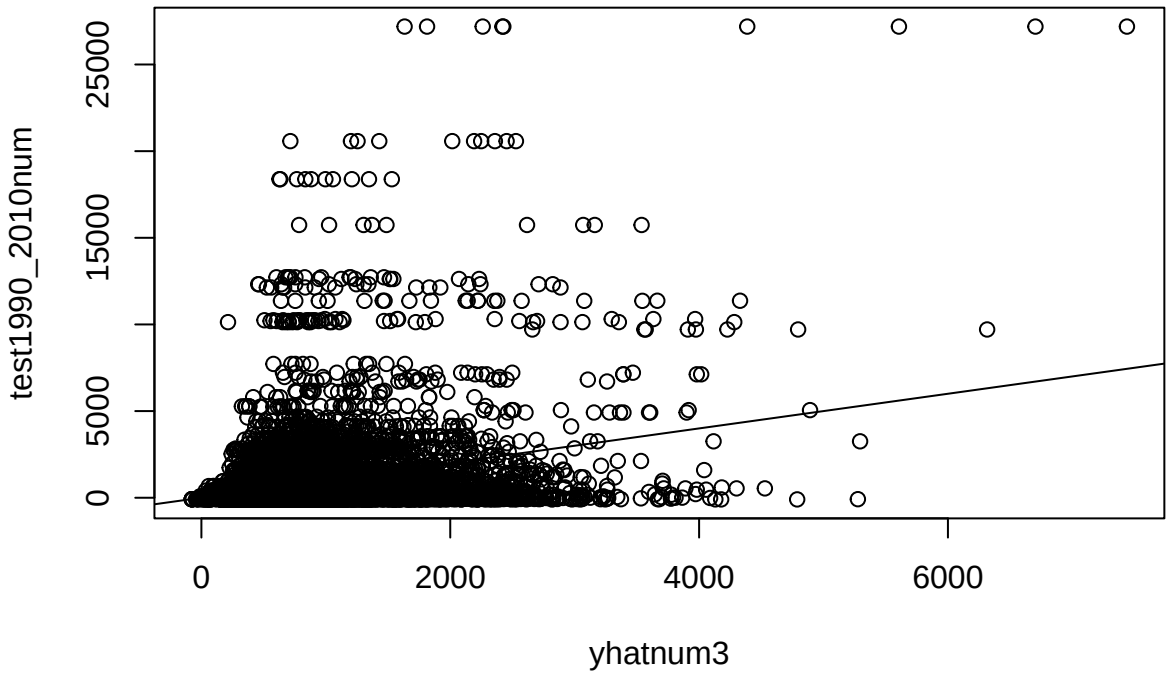

```r
print(mean((yhatnum2 - test2000_2010num)^2))
```

```
## [1] 89519.8
```

```r
#Random Forest 1990-2010
trainIndex3 <- createDataPartition(data_1990_2010$percentage_change_1990_2010, p = 0.15, list = FALSE)
trainData3 <- data_1990_2010[trainIndex3, ]
testData3 <- data_1990_2010[-trainIndex3, ]

bag.1990_2010 <- randomForest(percentage_change_1990_2010 ~ ., data = data_1990_2010, subset = trainInd
test1990_2010 <- data_1990_2010[-trainIndex3, "percentage_change_1990_2010"]
yhat.bag3 <- predict(bag.1990_2010, newdata = data_1990_2010[-trainIndex3, ])

test1990_2010matrix_data <- as.matrix(test1990_2010)
yhat.data.frame_data3 <- as.data.frame(yhat.bag3)
yhat_data3 <- lapply(yhat.data.frame_data3, as.numeric)
test1990_2010num <- as.numeric(test1990_2010matrix_data)
yhatnum3 <- as.numeric(unlist(yhat_data3))
plot(yhatnum3,test1990_2010num)
abline(0, 1)
```




```r
print(mean((yhatnum3 - test1990_2010num)^2))
```

```
## [1] 2293268
```