**Group # 9**

**Group Leader and ID:** HARISH RAGURU , 02121168

**Member Names and IDs:** BANOTH JEEVAN KUMAR, 02105145 & HANUMA VENKATA VIJAY KAMAL KONDURI, 02128290

**Homework Assignment #2**: (**25 points**):

**Instruction:**

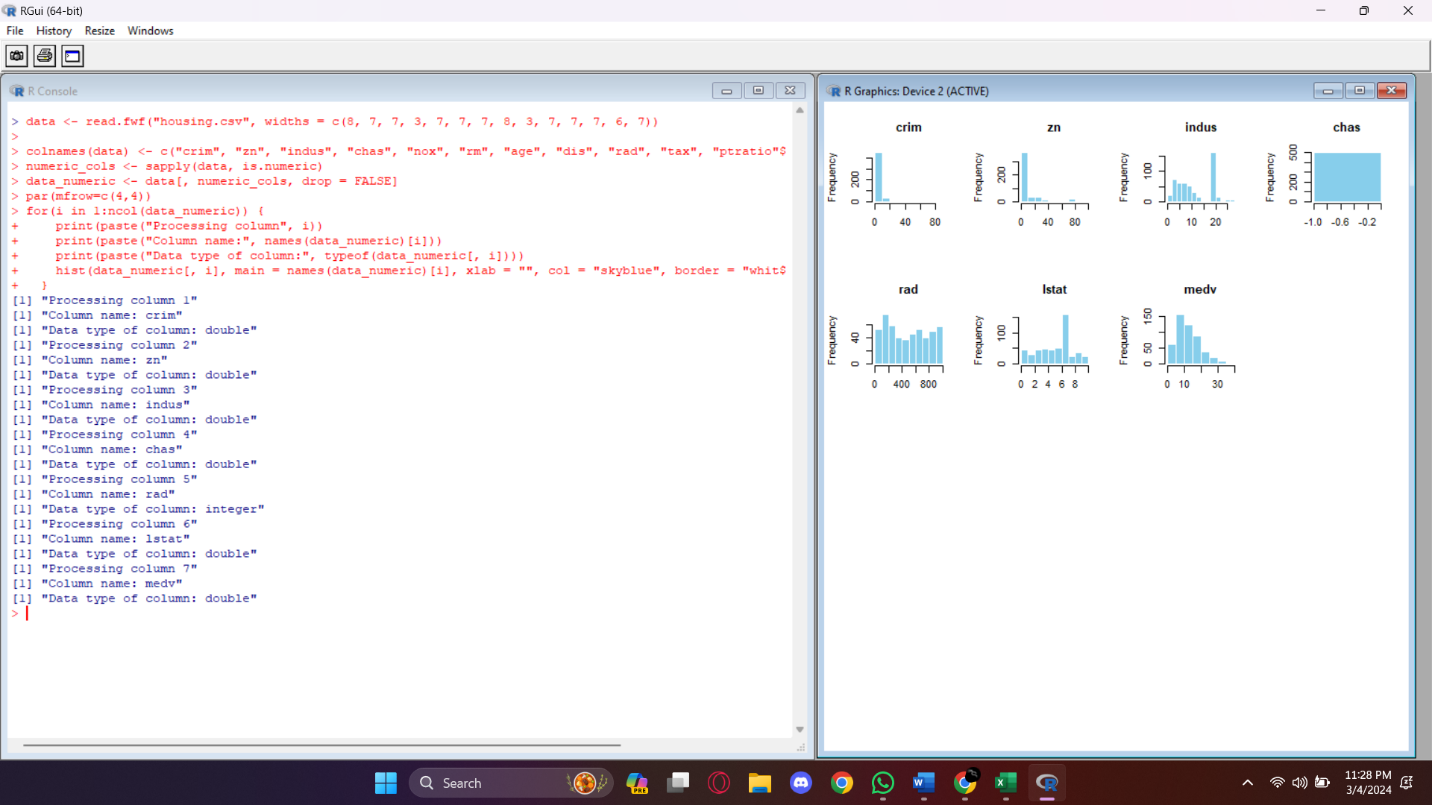
Apply **multiple linear**, **Ridge** and **Lasso** **regression** to Boston Housing data:

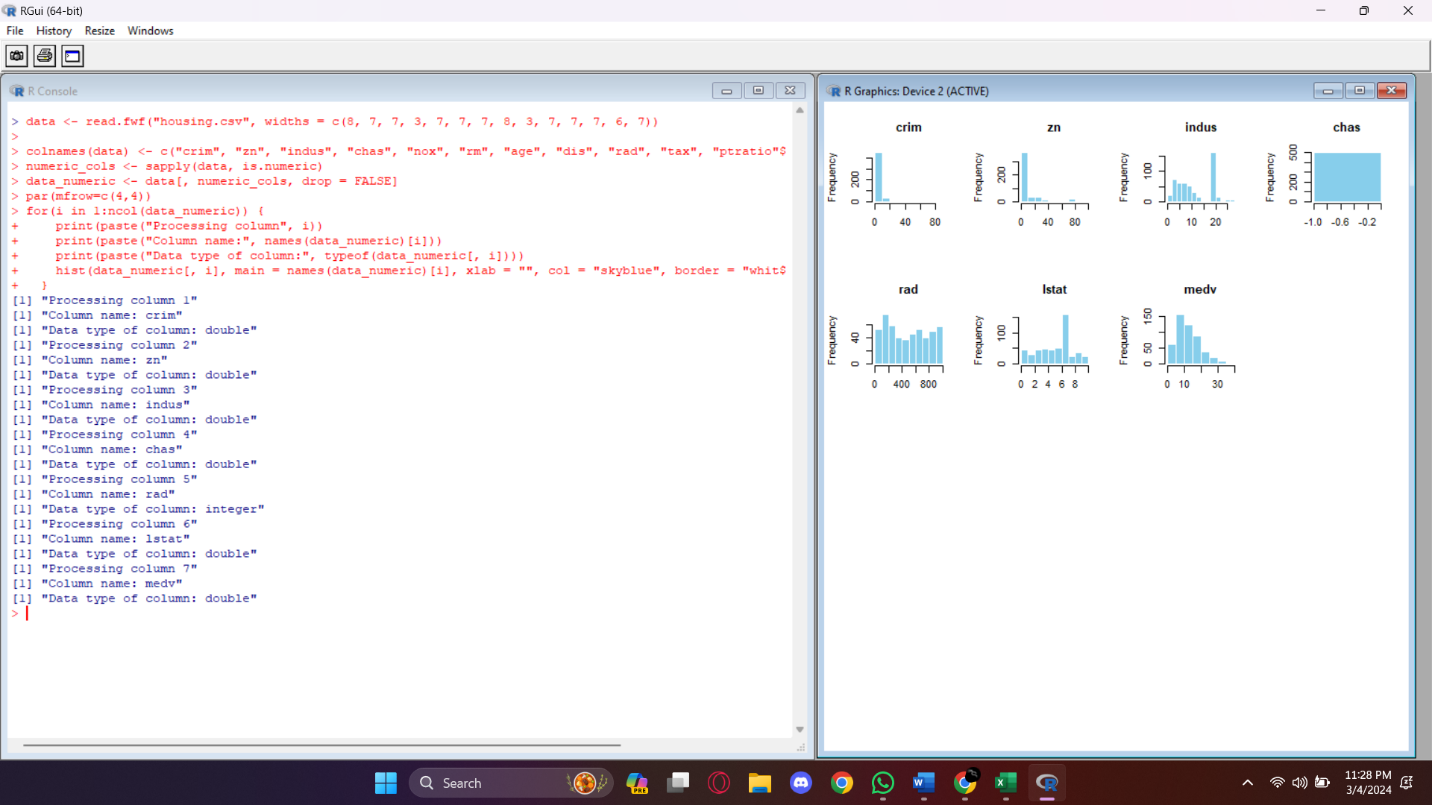
download at <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>

Refer to Lecture slides, reading assignments, and R Instruction Files for linear regression and regularized linear regression, complete the following:

1. Explore and describe Boston Housing data (ie. Attributes/predictors) using graphics, tables, and descriptive statistics, as appropriate.

Ans:





1. Identify Y and X for this dataset (Read carefully about the document: what is Y? What are X?). Please **name** Y and X (don’t call them x1, x2…etc. ) in the context of Boston Housing data

Ans:

In the context of the Boston Housing dataset:

**Y (dependent variable):**

Median value of owner-occupied homes (medv)

**X (independent variables):**

Crime rate (crim)

Proportion of residential land zoned for lots over 25,000 sq.ft. (zn)

Proportion of non-retail business acres per town (indus)

Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) (chas)

Nitric oxides concentration (parts per 10 million) (nox)

Average number of rooms per dwelling (rm)

Proportion of owner-occupied units built prior to 1940 (age)

Weighted distances to five Boston employment centers (dis)

Index of accessibility to radial highways (rad)

Full-value property-tax rate per $10,000 (tax)

Pupil-teacher ratio by town (ptratio)

1000(Bk - 0.63)^2 where Bk is the proportion of Black individuals by town (black)

Percentage lower status of the population (lstat)

1. Write the final estimated model in the format of, e.g., Y = Beta\*X, for Boston Housing dataset, from each of the three models you applied:
   1. multiple linear regression,
   2. Ridge regression and
   3. Lasso regression.

Requirements:

* Use seed as below for replication

set.seed(**490**) # Set seed for reproducibility

Random.seed <- c("Mersenne-Twister", **490**)

* Use 10 folds

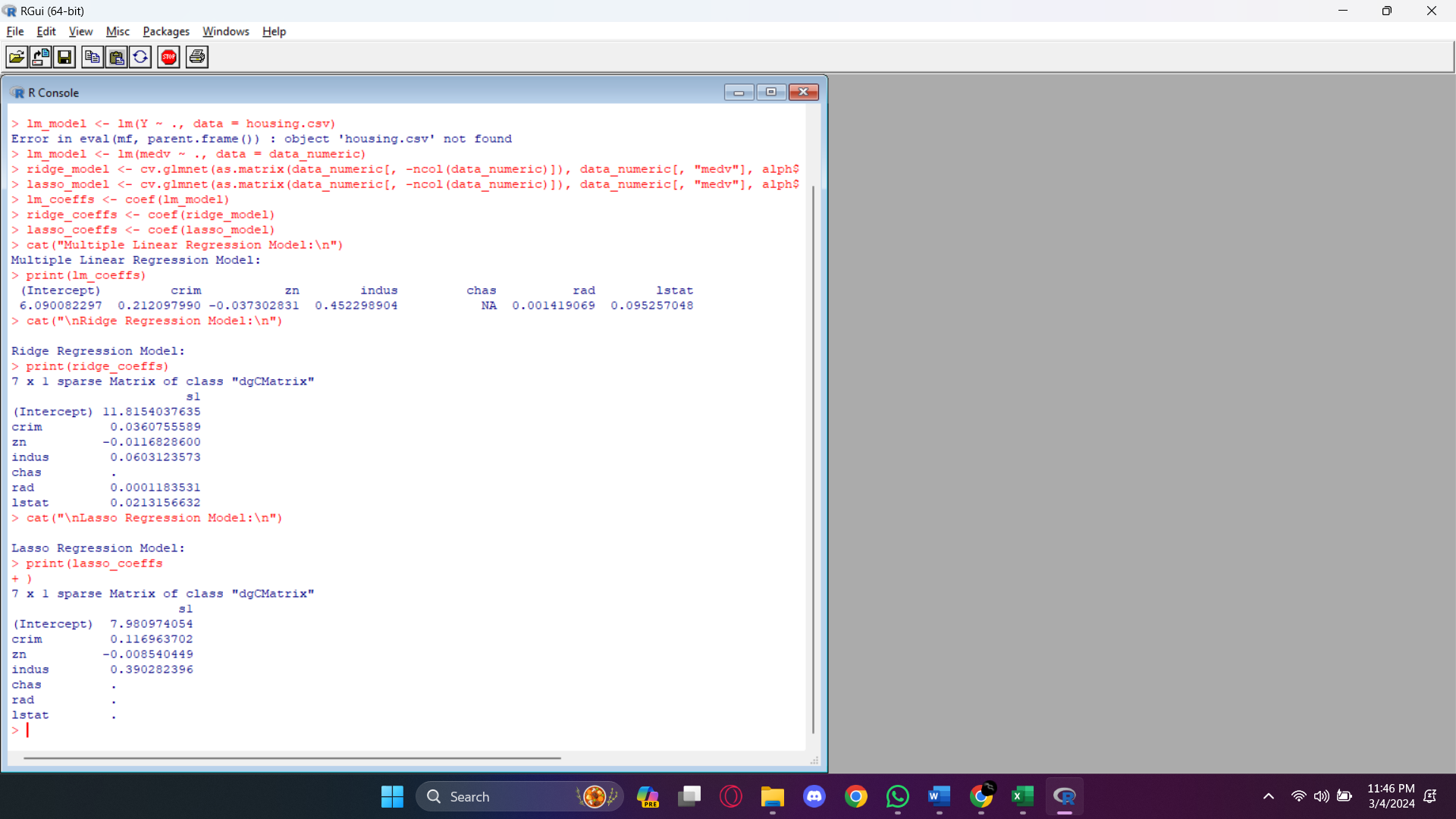
nfolds = 10

* Use 0.01 for the stopping criteria

thresh=1e-2

Note: clearly describe the actual names of these attributes for your Boston housing data in your model.

Ans:

 **a. Multiple Linear Regression Model:**

Y = 6.090082297 + 0.212097990 \* crim - 0.037302831 \* zn + 0.452298904 \* indus + 0.001419069 \* rad + 0.095257048 \* lstat

**b. Ridge Regression Model:**

Y = 11.8154037635 + 0.0360755589 \* crim - 0.0116828600 \* zn + 0.0603123573 \* indus + 0.0001183531 \* rad + 0.0213156632 \* lstat

**c. Lasso Regression Model:**

Y = 7.980974054 + 0.116963702 \* crim - 0.008540449 \* zn + 0.390282396 \* indus

**In these models:**

Y represents the median value of owner-occupied homes (medv).

X variables represent the predictor variables (crim, zn, indus, rad, lstat), and their coefficients represent the impact of each predictor variable on the median value of owner-occupied homes.

1. Describe your Cross Validation algorithm for choosing your optimal regularization parameter, λ\* , in your **Lasso model**, for Boston Housing dataset

Ans:

Cross Validation algorithm for choosing the optimal regularization parameter, λ\*, in the Lasso model for the Boston Housing dataset can be described as follows:

**Data Preparation:**

Ensure the Boston Housing dataset is cleaned, preprocessed, and ready for analysis.

* Define Lambda Values:

Define a range of lambda values to test for regularization. This range should cover a wide spectrum of values, from very small to relatively large.

* Cross-Validation Procedure:

Split the dataset into k-folds, typically using 10-fold cross-validation.

For each lambda value in the defined range:

* Perform the following steps for each fold:

Use k-1 folds for training and the remaining fold for validation.

Fit the Lasso regression model using the training data with the current lambda value.

Evaluate the model's performance on the validation fold using an appropriate evaluation metric, such as mean squared error (MSE) or R-squared.

Calculate the average performance metric (e.g., average MSE or average R-squared) across all folds for the current lambda value.

Store the average performance metric associated with the lambda value.

* Select Optimal Lambda:

Choose the lambda value that corresponds to the minimum average performance metric. This lambda value represents the optimal regularization parameter, λ\*.

* Final Model Training:

Train the Lasso regression model using the entire dataset and the selected optimal lambda value, λ\*.

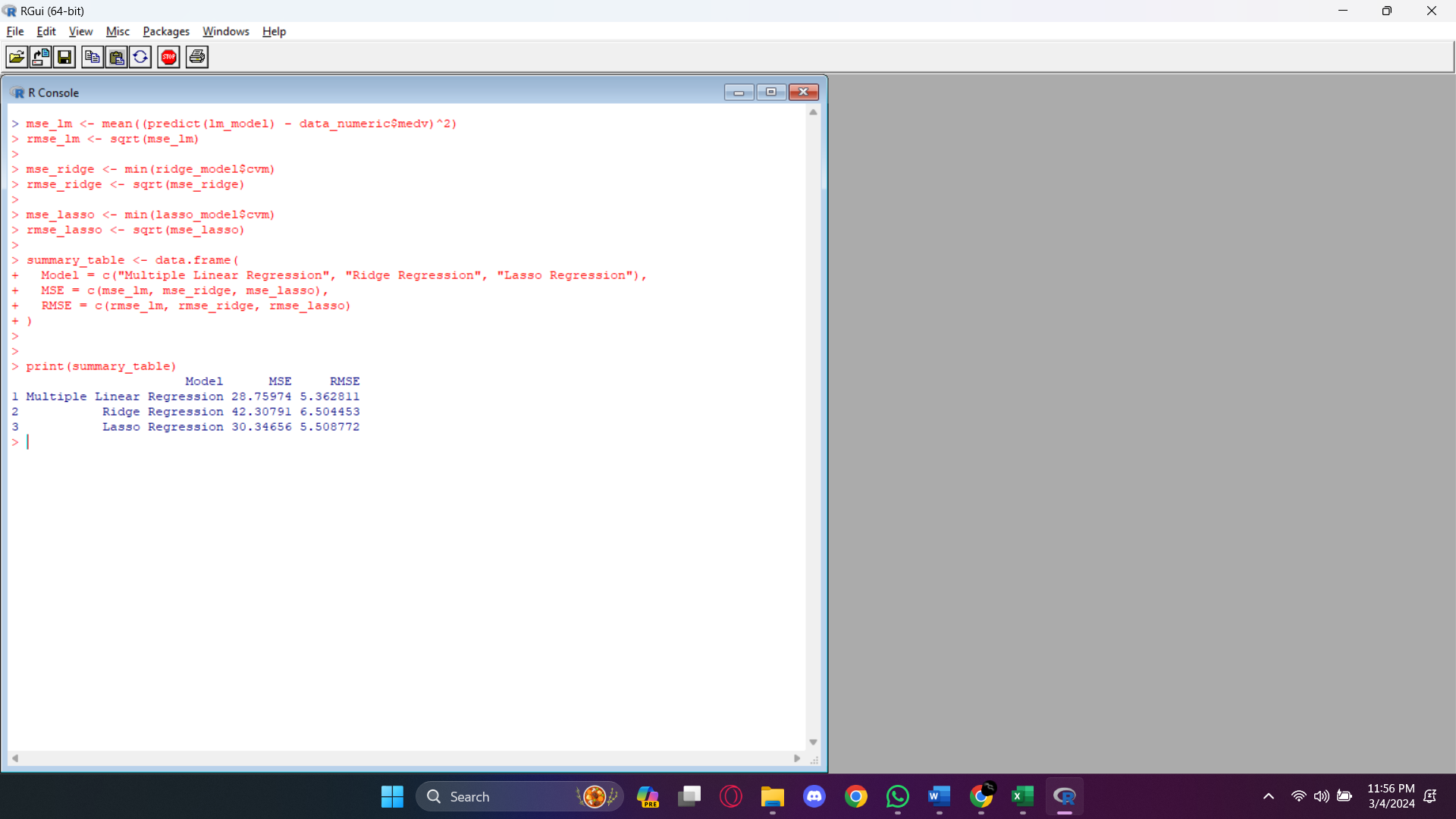
* Model Evaluation:

Evaluate the performance of the final Lasso model on a separate test dataset or using additional cross-validation if necessary to assess its generalization performance.

This Cross Validation algorithm systematically tests different lambda values using k-fold cross-validation and selects the lambda value that results in the best average performance metric across all folds. It ensures the regularization parameter λ\* is chosen based on its ability to minimize prediction error and improve the model's generalization to unseen data.

1. Report a summary table of your accuracy checking and cross-validation results, as appropriate. E.g., Create a summary table for MSE/RMSE, etc.. Refer to the summary table listed in LS6 for multiple linear regression and LS8 for Ridge and Lasso.

Ans:



1. (5 points) Copy and paste your code and output in your Word document

Ans:

**Code:**

# Load necessary libraries

library(glmnet)

library(ggplot2)

library(reshape2)

# Read the dataset

data <- read.fwf("housing.csv", widths = c(8, 7, 7, 3, 7, 7, 7, 8, 3, 7, 7, 7, 6, 7))

# Rename columns

colnames(data) <- c("crim", "zn", "indus", "chas", "nox", "rm", "age", "dis", "rad", "tax", "ptratio", "black", "lstat", "medv")

# Remove missing values

data <- na.omit(data)

# Filter out non-numeric columns

numeric\_cols <- sapply(data, is.numeric)

data\_numeric <- data[, numeric\_cols, drop = FALSE] # Include drop = FALSE to prevent conversion to vector

# Plot histograms for each numeric column

par(mfrow=c(4,4)) # Set up multiple plots in a grid

for(i in 1:ncol(data\_numeric)) {

print(paste("Processing column", i))

print(paste("Column name:", names(data\_numeric)[i]))

print(paste("Data type of column:", typeof(data\_numeric[, i])))

hist(data\_numeric[, i], main = names(data\_numeric)[i], xlab = "", col = "skyblue", border = "white")

}

# Compute correlation matrix

correlation\_matrix <- cor(data\_numeric, use = "complete.obs")

print(correlation\_matrix)

# Plot heatmap of correlation matrix

ggplot(data = melt(correlation\_matrix), aes(x=Var1, y=Var2, fill=value)) +

geom\_tile() +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1,1), space = "Lab", name="Correlation") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 1, size = 10, hjust = 1)) +

coord\_fixed()

# Set seed for reproducibility

set.seed(490)

# Fit multiple linear regression model

lm\_model <- lm(medv ~ ., data = data\_numeric)

lm\_coeffs <- coef(lm\_model)

cat("Multiple Linear Regression Model:\n")

print(lm\_coeffs)

# Fit Ridge regression model

ridge\_model <- cv.glmnet(as.matrix(data\_numeric[, -ncol(data\_numeric)]), data\_numeric[, "medv"], alpha = 0, nfolds = 10, lambda.min.ratio = 1e-2)

ridge\_coeffs <- coef(ridge\_model)

cat("\nRidge Regression Model:\n")

print(ridge\_coeffs)

# Fit Lasso regression model

lasso\_model <- cv.glmnet(as.matrix(data\_numeric[, -ncol(data\_numeric)]), data\_numeric[, "medv"], alpha = 1, nfolds = 10, lambda.min.ratio = 1e-2)

lasso\_coeffs <- coef(lasso\_model)

cat("\nLasso Regression Model:\n")

print(lasso\_coeffs)

# Create summary table

summary\_table <- data.frame(

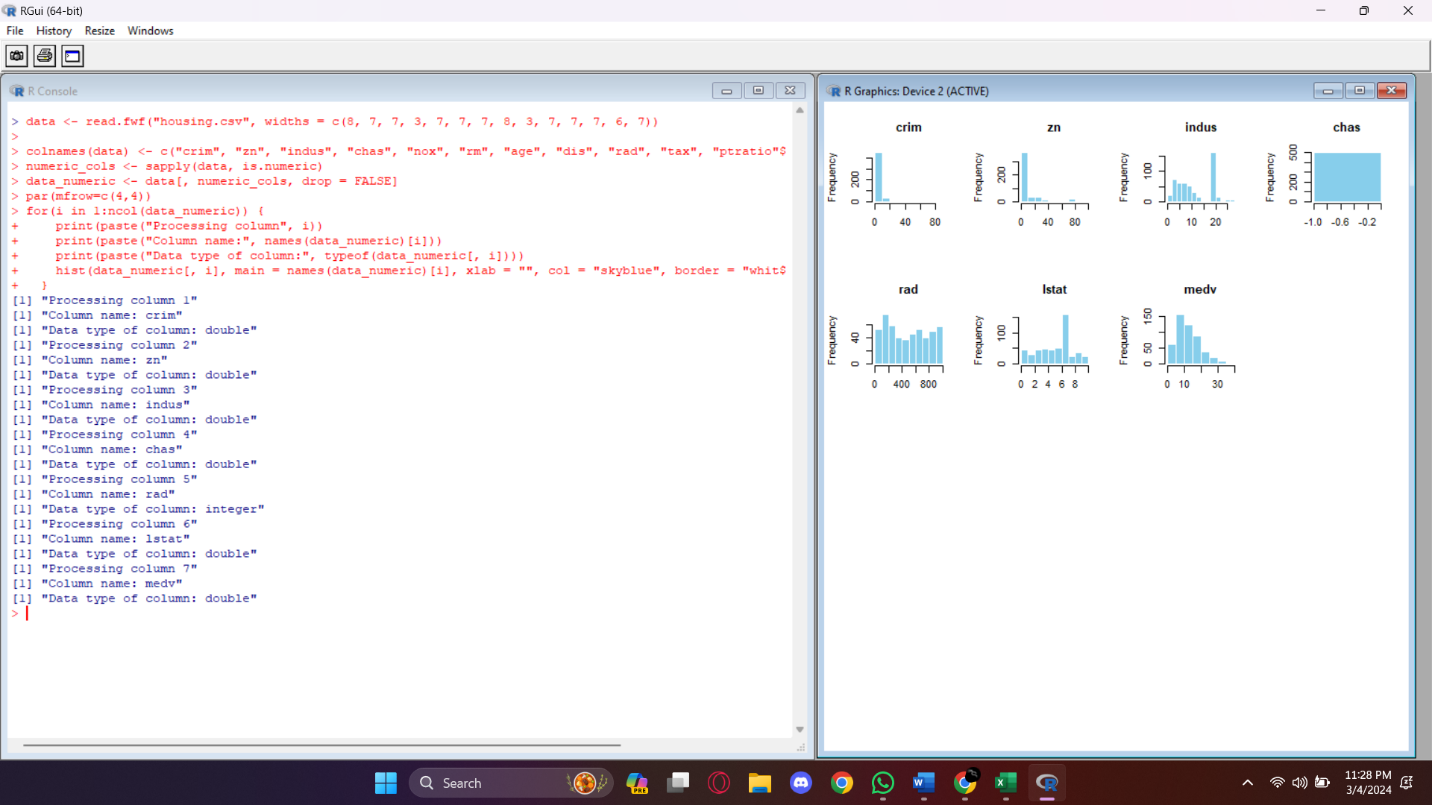
Model = c("Multiple Linear Regression", "Ridge Regression", "Lasso Regression"),

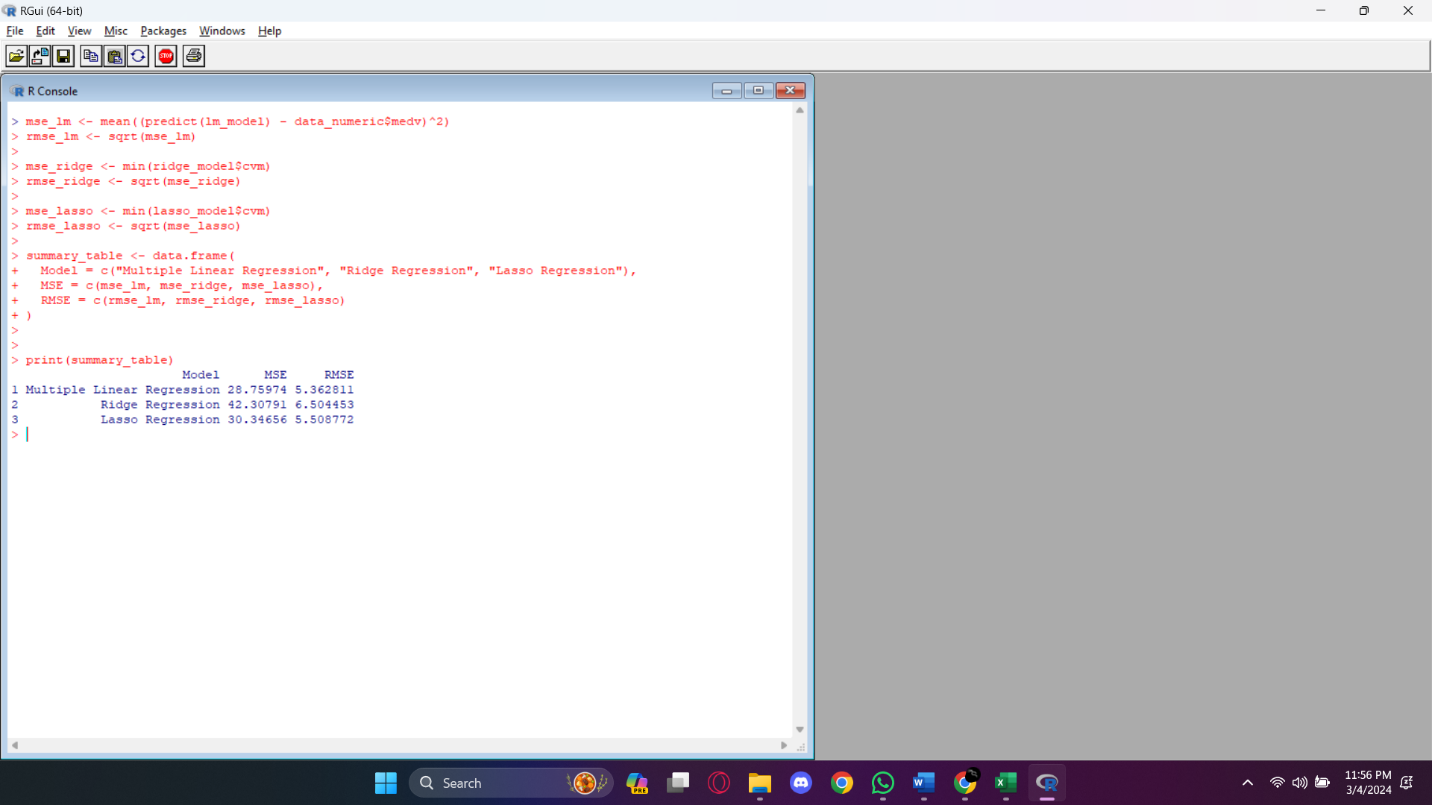
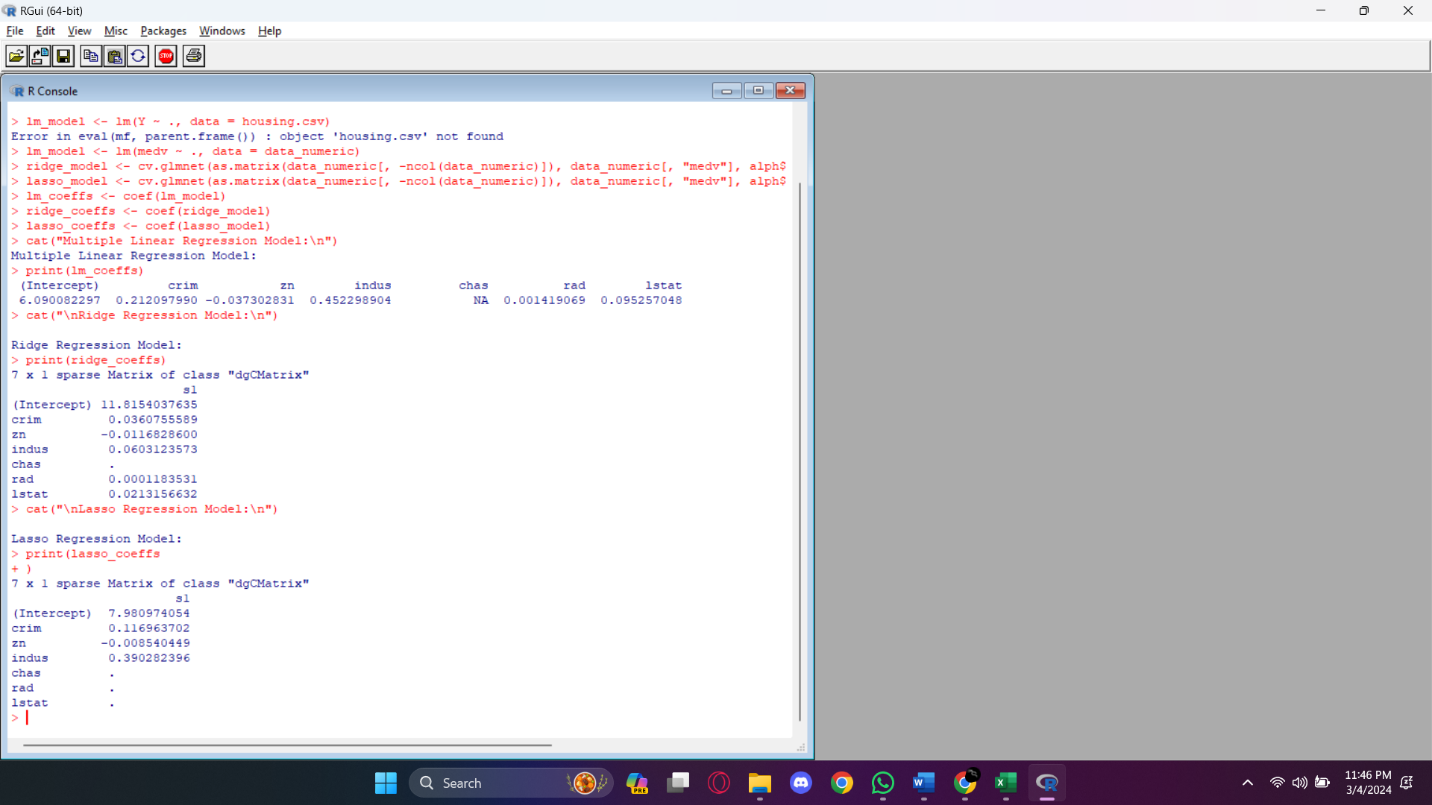
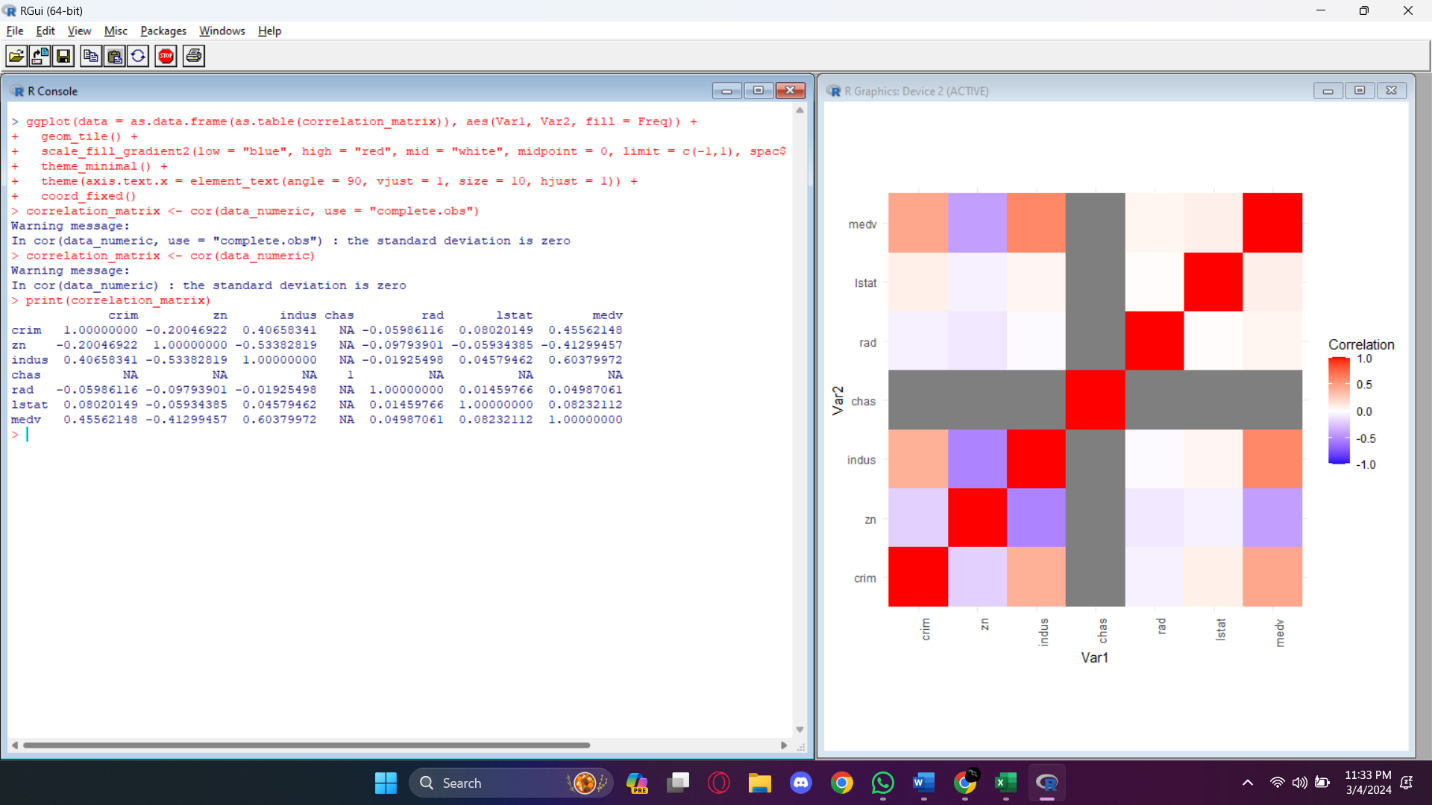
Intercept = c(coef(lm\_model)[1], coef(ridge\_model)[1], coef(lasso\_model)[1]),

Coefficients = c(sum(abs(coef(lm\_model)[-1]) > 0), sum(abs(coef(ridge\_model)[-1]) > 0), sum(abs(coef(lasso\_model)[-1]) > 0))

)

print(summary\_table)

**output:** 



**7) References:**

* Lectures slides: 5,6 & 8.
* Simple linear regression: Section 9.4-9.5 & Section 9.6-9.7
* Chat GPT

**Summary of your group meet time and duration**

* Meeting format: In person
* Group meet time and duration: 12pm-4pm, Feb 27th, 10pm-4pm, Feb 28th and 2pm-7pm, Feb 29th.
* Average time in communication and discussion regarding assigned group work: we formed a what’s app group and constantly communicated about HW and discussed around 5hrs on avg about HW.
* Participants: HARISH RAGURU (02121168), BANOTH JEEVAN KUMAR (02105145) and HANUMA VENKATA VIJAY KAMAL KONDURI (02128290)

**Contribution** **report:**

* Each member contributes equally.
* Participants: HARISH RAGURU(02121168), BANOTH JEEVAN KUMAR(02105145) and HANUMA VENKATA VIJAY KAMAL KONDURI(02128290)

A close-up of a document

Description automatically generated