**כריית נתונים ב- R**

***מיני תרגיל מספר 4***

**שאלות:**

בתרגיל זה נתשמש במסד ToyotaCorolla.csv.

1. קרא את הקובץ לתוך מסד ב- R
2. חלק את המסד ל – training (60%) ו- validation (40%). השתמש ב- seed = 1
3. הורד ממסד ה- training רכבים שמשקלם גבוה מ- 1200, ובנה מודל רגרסיה על הנתונים המטויבים
4. בדוק את ביצועי המודל על ה- validation. האם ביצועי המודל השתפרו?
5. חשב את המחיר החזוי של רכב שמשקלו Weight = 1400

**פתרון:**

setwd("C:/Users/user/Desktop/R")

library(ggplot2)

library(forecast)

toyota.df <- read.csv("ToyotaCorolla.csv", stringsAsFactors = TRUE)

head(toyota.df)

## partition

set.seed(1)

train.index <- sample(1:dim(toyota.df)[1], dim(toyota.df)[1]\*0.6)

train.df <- toyota.df[train.index, ]

valid.df <- toyota.df[-train.index,]

## remove outliers

train.df <- train.df[train.df$Weight < 1200, ]

## simple regression

reg <- lm(Price ~ Weight, data = train.df)

summary(reg)

# plot regression

ggplot(train.df, aes(x = Weight, y = Price)) + geom\_point() +

geom\_smooth(method = lm)

## predict

pred.train <- predict(reg)

accuracy(pred.train, train.df$Price)

pred.valid <- predict(reg, newdata = valid.df)

accuracy(pred.valid, valid.df$Price)

## new car

new.car <- data.frame(Weight = 1400)

pred.new.car <- predict(reg, newdata = new.car)

pred.new.car