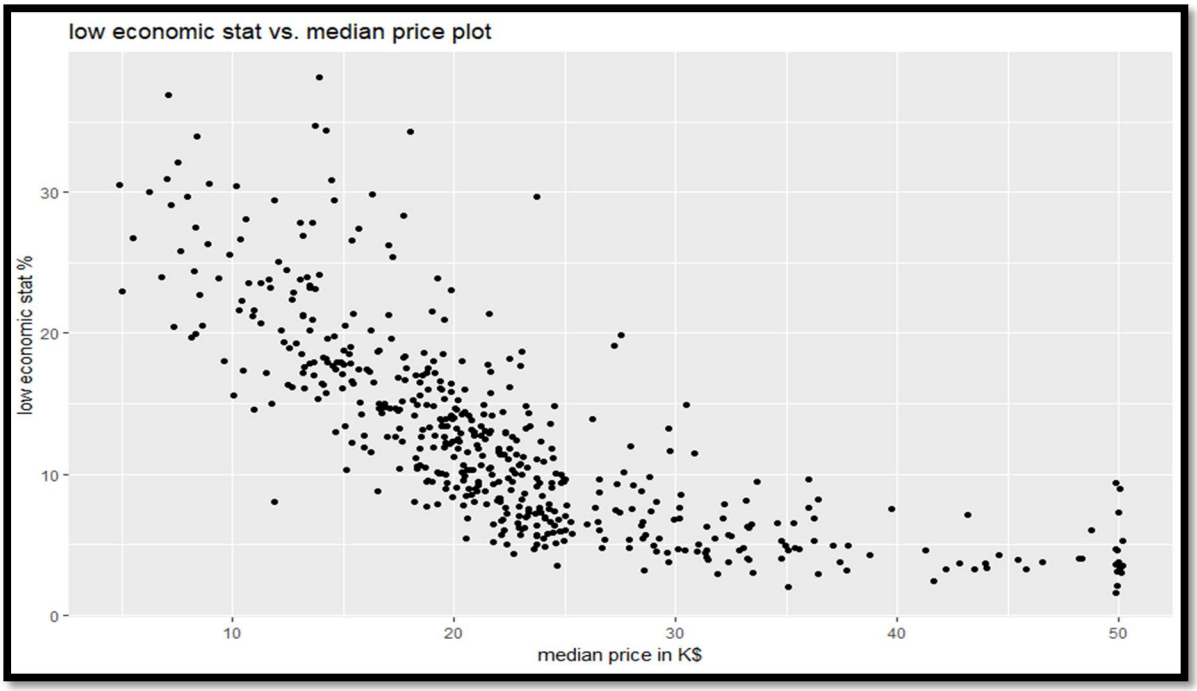
יישומי R בכריית נתונים, תרגיל 2 – קובץ פלט

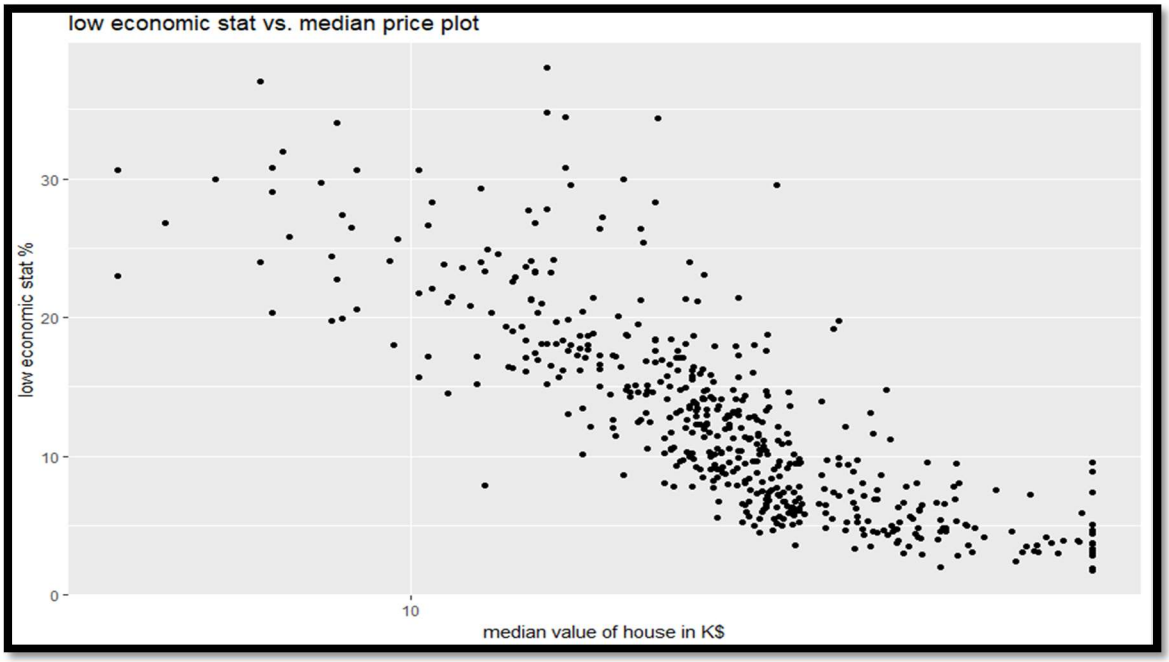
1. הקשר בין MEDV ל- LSTAT

אנחנו רואים שככל שאחוז האוכלוסייה במצב סוציואקונומי יורד, כך החציון של ערך הבית עולה, הנ"ל נכון עבור ערכים נמוכים של החציון ומתייצב לפלטו באזור ה-25 אלף דולרים.

ללא הפעלת לוג על ציר X

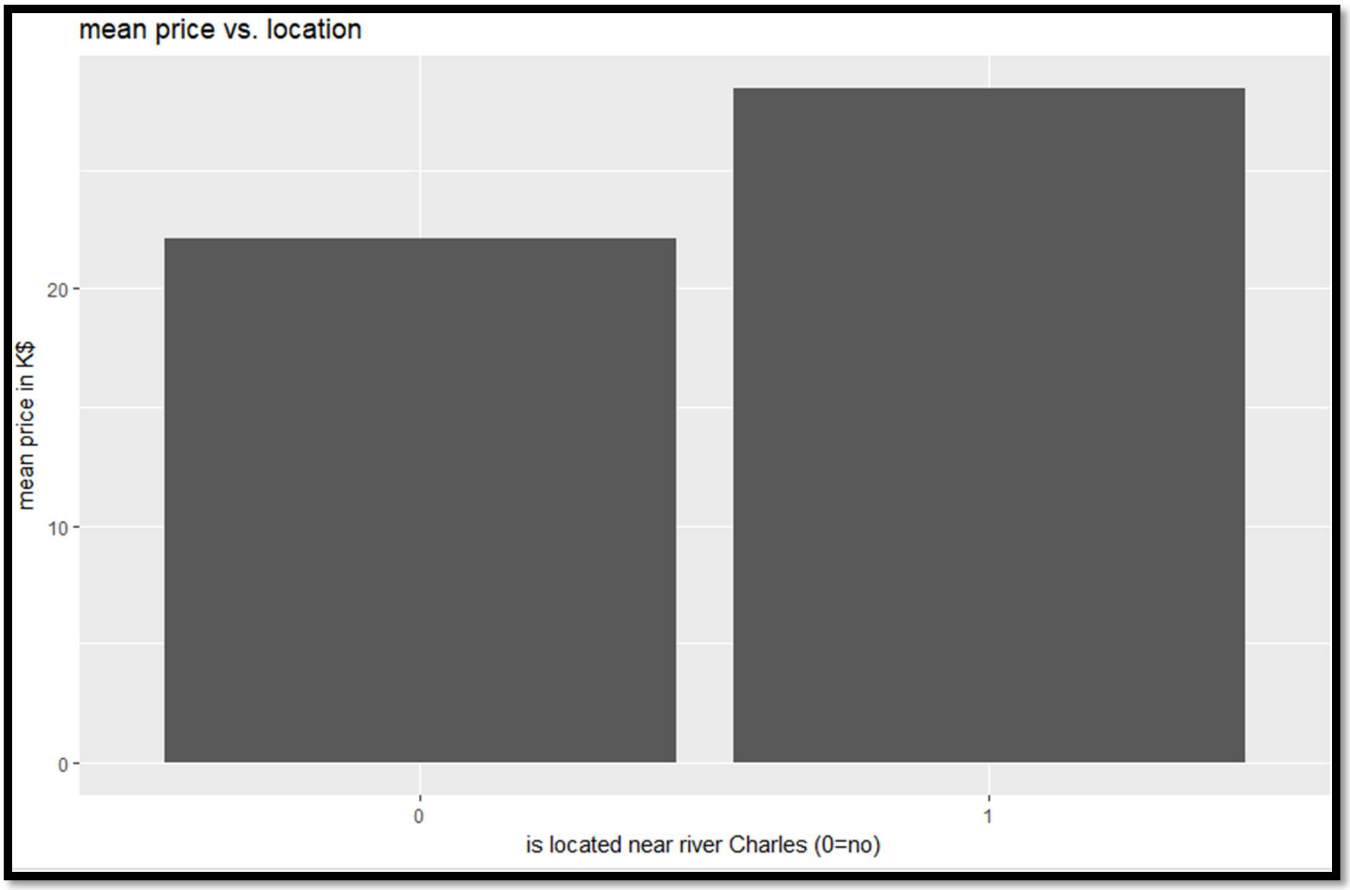


עם הפעלת לוג על ציר ה-X



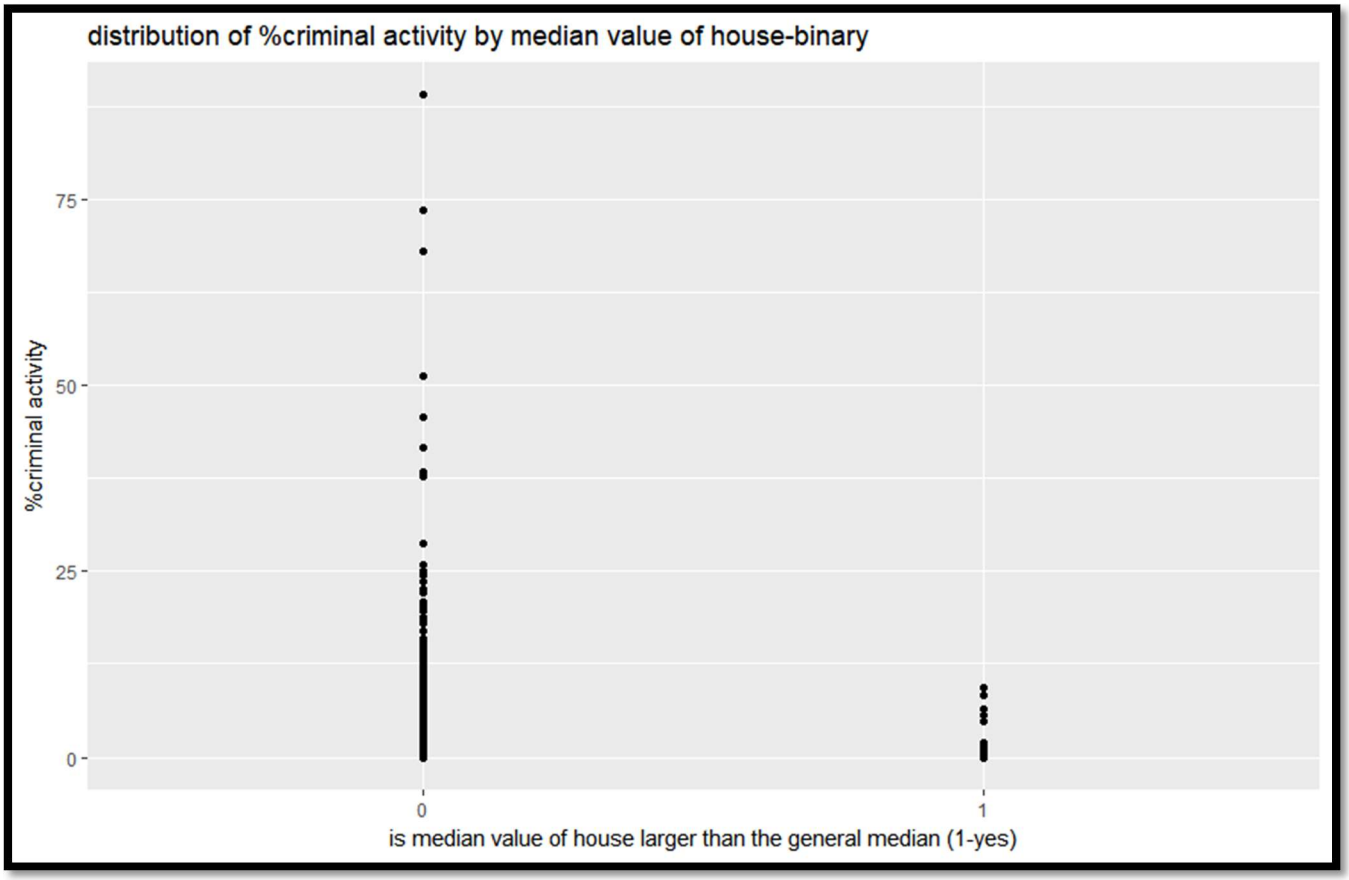
2. הקשר בין CHAS לממוצע MED

אנו רואים שממוצע ערך הבית גבוה באזורים הממוקמים ליד נהר צ'רלס. האם יש לכך משמעות סטטיסטית? יש להריץ מבחן t להשוואה בין 2 מדגמים ב"ת תלויים



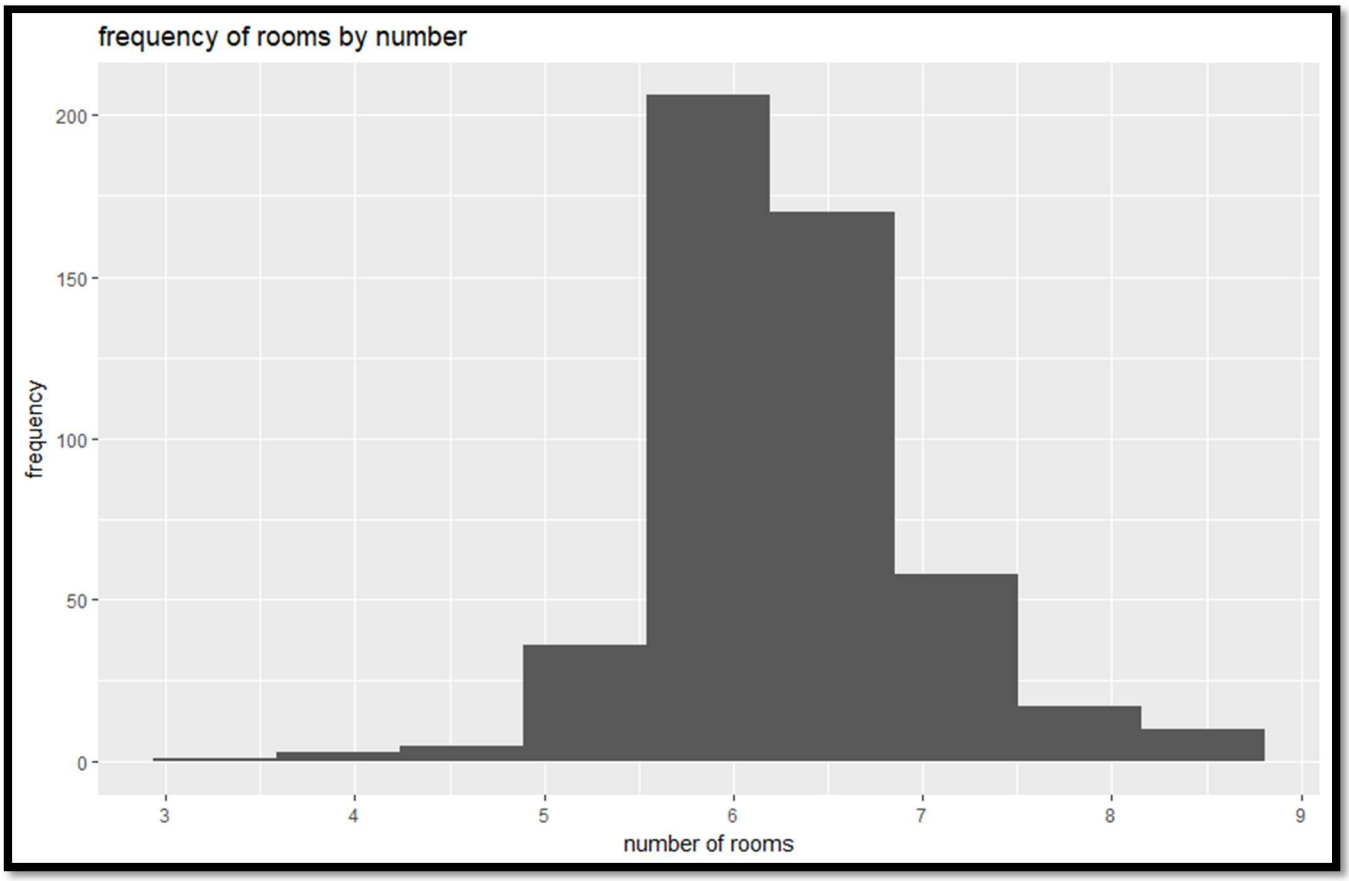
3. התפלגות CRIM לפי MEDV.CAT

אנו רואים כי במקרה בו חציון ערך הבית נמוך מחציון הבתים הכללי, הפשיעה מתפלגת על פני טווח אחוזים גבוה הרבה יותר מאשר במקרה בו חציון הבית גבוה מהחציון הללי.



4. גרף נוסף לבחירתך-היסטוגרמה של מספר חדרים בבית

אנו רואים ששכיחות החדרים נעה סביב הערך 6.



מאחר ובהגשה ניתן לצרף רק קובץ 1 למודל- מצ"ב קובץ R

```
library(ggplot2)

 setwd("C:/Users/dom/Desktop/r files")

 Boston.df<-read.csv("BostonHousing.csv")

View(Boston.df)

p<- ggplot(Boston.df, aes(x=MEDV, y =LSTAT))

p + geom_point()+ylab("low economic stat %")+xlab("median value of house in K$")+ scale_x_log10() +ggtitle("low economic stat vs. median price plot")

meanMEDV <- aggregate(MEDV ~ CHAS,
                          data = Boston.df,
                          FUN = "mean")
View(meanMEDV)

ggplot(meanMEDV, aes(x = as.factor(CHAS), y = MEDV)) +
  geom_bar(stat="identity") +
  ylab("mean price in K$")+xlab("is located near river Charles (0=no)")+
  ggtitle("mean price vs. location")

ggplot(Boston.df, aes(x = as.factor(CAT..MEDV), y = CRIM )) + geom_point()+
  xlab("is median value of house larger than the general median (1-yes)")+
  ylab("%criminal activity")+
  ggtitle("distribution of %criminal activity by median value of house-binary")

ggplot(Boston.df, aes(x=RM))+geom_histogram(bins=9)+
  xlab("number of rooms")+
  ylab("frequency")+
ggtitle("frequency of rooms by number")
```