

## כריית נתונים ב-R // תרגיל בית מספר 2

בתרגיל זה נשתמש בקובץ נתונים BostonHousing.csv המכיל מידע על שכונות בבוסטון, ובכללו חציון מחיר בתים בשכונה, רמת פשיעה בשכונה, מצב סוציו-אקונומי, ועוד.

הקובץ נלקח מתוך המאמר:

Harrison and Rubinfeld, "Hedonic prices and the demand for clean air", *Journal of Environmental Economics & Management*, vol. 5, pages 81-102, 1978.

להלן תיאור משתני המסד:

| שם משתנה | תיאור  |
|----------|--|
| CRIM     | אחוז פשיעה   |
| ZN       | אחוז שטחים פרטיים הגדולים מ 25,000 ft <sup>2</sup>           |
| INDUS    | אחוז שטח שאינו מסחרי   |
| CHAS     | האם ממוקם ליד נהר צ'רלס (1 אם כן)                            |
| NOX      | ריכוז תחמוצת החנקן   |
| RM       | מספר חדרים ממוצע לבית  |
| AGE      | אחוז דירות שנבנו לפני 1940                                   |
| DIS      | מרחק משוקלל ממרכז בוסטון                                     |
| RAD      | גישה לכבישים היקפיים   |
| TAX      | ערך ארנונה   |
| LSTAT    | אחוז אוכלסיה במצב סוציו אקונומי נמוך                         |
| MEDV     | חציון ערך בתים באלפי דולרים                                  |
| CAT.MEDV | האם MEDV גדול מחציון הבתים הכללי<br>שערכו \$30,000 (1 אם כן) |

עבור כל אחד מהקשרים הבאים, בנה את הגרף המתאים, והסבר את אופי הקשר בין המשתנים. יש לעבוד עם גרפים מספריית ggplot2. וודא כי הגרף מסוגן היטב: שמות צירים, כותרות, log-scale במידת הצורך (2.5 נקודות לסעיף, תורד נקודה אחת על שימוש בגרפים שאינם מספריית ggplot2):

1. הקשר בין MEDV ל-LSTAT
2. הקשר בין CHAS לממוצע MEDV
3. התפלגות CRIM לפי MEDV.CAT
4. גרף נוסף לבחירתך

## אופן הגשה:

- ✓ הגשה דרך אתר למידה
- ✓ יש להגיש קובץ R מתועד, וקובץ word עם הגרפים והסבר קצר על כל גרף.