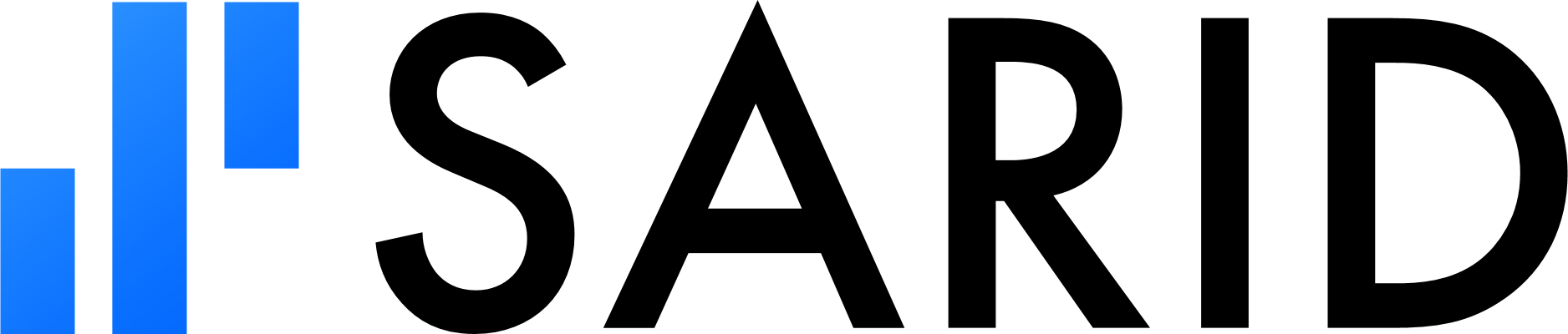
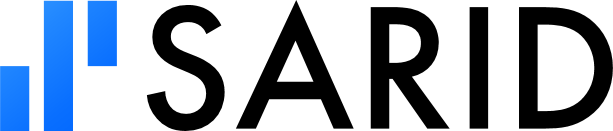


המדריך העברי למשתמש ב-R

עדי שריד

בהוצאת מכון שריד – שרותי מחקר והדרכה





© 2025 כל הזכויות שמורות למכון שריד

ספר זה זמין בגרסה אינטרנטית תחת רישיון Creative Commons BY-NC-ND 4.0 בכתובת: <https://www.hebrewr.co.il>

אתם חופשיים לשתף את הספר בכל צורה שתרצו תוך ציטוט המקור, הוספת הקישור למקור (שנמצא לעיל), באופן שלא ישתמע שהכותב תומך או אחראי לשימוש שאתם מבצעים בספר.

לא תוכלו להשתמש בחומרים לצרכים מסחריים.

אם אתם משנים את החומרים באופן כלשהו, אסור לכם להפיץ את הגרסה ששונתה.

התנאים המלאים בכתובת: <https://creativecommons.org/licenses/by-nc-nd/4.0>

בהקדשה,

לאשתי, אנה, שתמיד תומכת ומעודדת

לילדי, איתן, נעמי, ואביב, תמשיכו להיות סקרנים

להורי, שהיוו ומהווים מקור השראה

תוכן עניינים

הקדמה

R שפה נפלאה, היא התחילה כשפה סטטיסטית, וגם כיום זה השימוש המרכזי שלה. בפועל אפשר לעשות בה כמעט הכל. החל מניתוח נתונים ועד בניית אפליקציות ואוטומציות מורכבות. אצלנו בעסק (מכון שריד) אנחנו משתמשים ב-R לניתוח נתונים, הצגת תוצאות עם דשבורדים בזמן אמת, הפקת דוחות Power Point בלחיצת כפתור, אוטומציות של תהליכים שונים בעסק, ועוד הרבה - היא הפכה לכלי עבודה שימושי והחליפה הרבה מאוד מהדברים שהיינו עושים בעבר עם כלים אחרים.

ללומדים את השפה יש מקורות רבים ומגוונים - ספרים, סרטוני YouTube, בלוגים, ועוד. העניין הוא שרוב המקורות שישנם כתובים באנגלית. זה מה שהביא אותי לכתוב את הספר הזה, ולעזור לקהל הקוראים שיותר נוח לו ללמוד R בעברית.

התוכן שבספר שלפניכם מבוסס על מספר קורסים שלימדתי בשנים האחרונות, כולל הכשרות מקצועיות בחברות הייטק, וקורסים שלימדתי באוניברסיטת תל-אביב.

התמונה שעל כריכת הספר היא תמונה של דוכיפת, הציפור הלאומית של ישראל. התמונה הופקה על ידי ChatGPT (Dalle3). הרעיון לשימוש בציפור על כריכת הספר הוא מחווה לספר R for Data Science (R4DS) שהיווה השראה לכתיבת ספר זה, וגם עליו יש תמונה של ציפור (תוכי ניו-זילנדי).

# למי מיועד הספר?

כל אחד יכול ללמוד לעבוד עם R. אם אתם עובדים היום עם נתונים ומשתמשים ב-Excel או בכל כלי אחר, ומעוניינים לעבור לכלי חזק משמעותית שיגדיל את הגמישות, סל היכולות שלכם, ומהירות העבודה שלכם.

אם אתם סטודנטים שלומדים מקצוע כמותי, כגון מבוא לסטטיסטיקה או קורס מתקדם בסטטיסטיקה.

אם אתם עוברים מכלי סטטיסטי אחר (כגון SAS, SPSS, או אחרים) ומעוניינים למנף את הכלים והיכולות שיש ל-R להציע.

הספר מתאים גם לכאלו שיודעים R ברמה בינונית ורוצים לחזק את הבסיס, או כאלו שרוצים להשתפשף בידיעות שלהם על החבילות שנלמדות בספר (כמו tidyverse).

הספר אינו נועד להחליף ספרי תיאוריה סטטיסטית, הוא אינו מכסה תיאוריה סטטיסטית לעומק, אלא בעיקר מדגים כיצד ניתן לממש כלים סטטיסטיים שונים באמצעות R.

# למה R ולא Python?

כמו שמרמזת הכותרת, יש היום שתי שפות שמובילות בתחום ה-Data Science, הן R ו-Python. בנקודה הזו אין שחור ולבן, טוב יותר או פחות. אני עובד בשתי השפות, ולי באופן אישי יש העדפה חזקה לשימוש ב-R, פשוט כי היא מרגישה לי יותר טבעית, זה כנראה נגזרת של העובדה שזה מה שלמדתי בהתחלה, וזה מה שהתמקצעתי בו. בפועל בשתי השפות יש כלים רבים ושימושיים, ובכל הקשור במשימות של ניתוח נתונים, שתי השפות נותנות מענה טוב מאוד.

מבין שתיהן Python יותר נפוצה משום שהיא נפוצה כשפת תכנות (באופן כללי), בעוד ש-R מוכוונת יותר לתחום הסטטיסטיקה. בפועל, שתיהן עושות את העבודה, וכדאי להתאים את השפה לצורך. לדוגמה R הרבה פעמים נוחה יותר לביצוע ניתוח נתונים, אבל ל-Python עשוי להיות יתרון במעבר ל-Production (הטמעה במערכות).

הדיון של בחירה בשפה מסוימת צריך להיגזר מהמשימה הנדרשת, ההקשר העסקי, ועוד הרבה שיקולים אחרים. ככלל למי שמשתלב בתוך ארגון כדאי לדעת שיש ב-R חבילות שמאפשרות להפעיל גם קוד Python מתוך R (חבילת reticulate), שתאפשר להשתלב עם R גם בתוך ארגון שעובד מסורתית עם Python.

# מה בספר?

בפרק 1 "מבוא" נראה מבוא קצר לתהליך ניתוח הנתונים ונלמד מהם הכלים הנדרשים על מנת להתקין את R. לאחר מכן נצלול, בפרק 2 "תחביר בסיסי" לבסיס של R כשפת תכנות. הפרק מתייחס לחלקים ב-R שמגיעים בגרסה ה”ערומה ביותר” מה שמכונה גם Base R.

בפרק 3 "הכנת נתונים" נלמד את הבסיס של חבילות tidyverse ואיך הן מרחיבות את Base-R והופכות את הקוד שלנו ליותר קריא, ואת הכתיבה ליותר נוחה. אנחנו נתמקד בעיקר בשלבי הכנת הנתונים לקראת ניתוח, ואז בפרק 4 "ויז’ואליזציות" נלמד על ויז’ואליזציות באמצעות חבילת ggplot2. בפרק 5 "מבנה וחיבור נתונים" נטפל במניפולציות קצת יותר מתקדמות כמו חיבור טבלאות ושינוי מבנה טבלאות.

לעולם הסטטיסטי ניכנס בפרק 6 עם מבחני השערה, ובפרק 7 נדון ברגרסיה לינארית.

לאחר מכן נצלול קצת יותר לעומק עם מודלים מתקדמים מעולם ה-Data Science תוך שימוש בחבילת tidymodels בפרק 8 "מודלים מונחים".

בפרק 9 נחזור מהעולם הסטטיסטי לתכנות, ונתעמק קצת יותר בלולאות מתוכמות. בפרק 10 נציג כיצד ניתן להפעיל את פיית’ון מתוך R, ובפרק 11 נדגים את השימוש בחבילות של מודלי שפה (הפעלת מודלי LLM כגון ChatGPT) מתוך R.

בפרק 12 "עברית ב-R" ניתן טיפים לכל מי שמשתמש ב-R וצריך לשלב את העבודה שלו עם עברית.

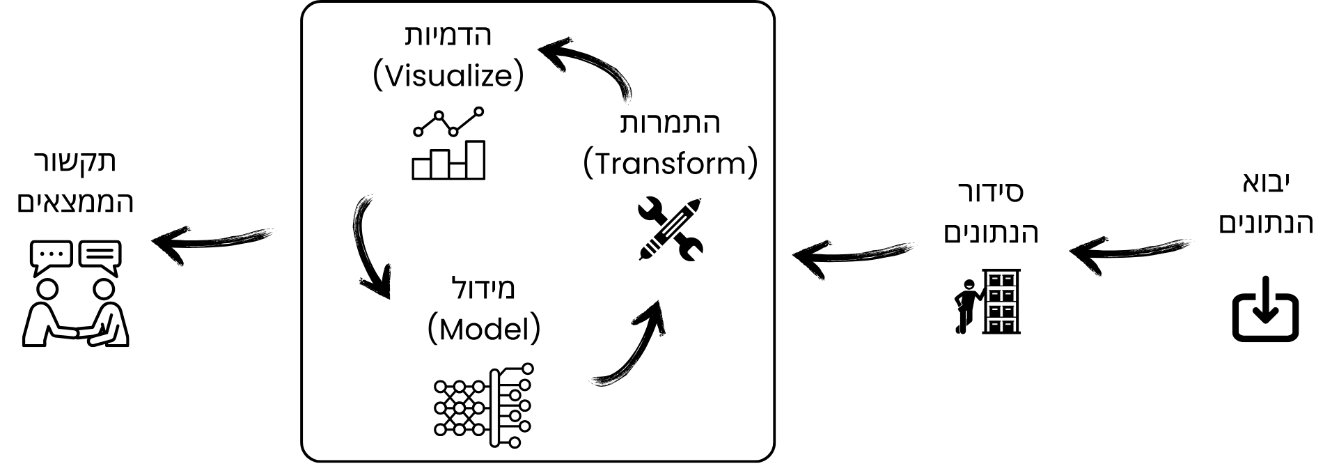
ספר זה זמין באופן חופשי באינטרנט, ללא עלות, תחת רישיון Creative Commons BY-NC-ND 4.0, מה שאומר שאפשר לשתף את הקישור לספר או לשתף ציטוטים מהספר תוך ציון המקור עם קישור למקור. לא ניתן להשתמש בתוכן למטרות מסחריות.

# מבוא

## תהליך ניתוח נתונים

בספר הזה אני מאמץ את נקודת ההשקפה של (Wickham, et al., 2023)

שמציגים את אבני הבניין של עבודת מדען הנתונים כפי שמומחש באיור הבא:



איור 1: תהליך העבודה של מדען הנתונים

כל תהליך ניתוח נתונים מורכב מיבוא סט הנתונים, סידור שלהם בצורה שנוח לעבוד איתה, לאחר מכן מחקר על הנתונים, שהוא תהליך איטרטיבי (חוזר על עצמו) שמורכב מהתמרות שונות על הנתונים (טרנספורמציות), המחשות בתרשימים (ויז’ואליזציות), ומידול של הנתונים. לבסוף, התהליך עובר למסמך, מצגת או אמצעי אחר, ומוצג למקבלי החלטות.

בספר זה נדון בשלבים של יבוא הנתונים, סידור, וכלל שלבי המחקר (טרנספורמציות, ויז’ואליזציות, ומידול). אך ראשית, לפני שנתחיל בצעדים הראשונים (למידת בסיס השפה), נראה כיצד להתקין את R.

## איך מתקינים את R?

התקנת התוכנה פשוטה מאוד. R היא שפה בקוד פתוח (Open Source), כלומר, כל אחד יכול להתקין אותה ללא עלות. כמו כן, ישנה סביבת פיתוח, גם היא חינמית ובקוד פתוח, הנקראת RStudio IDE.

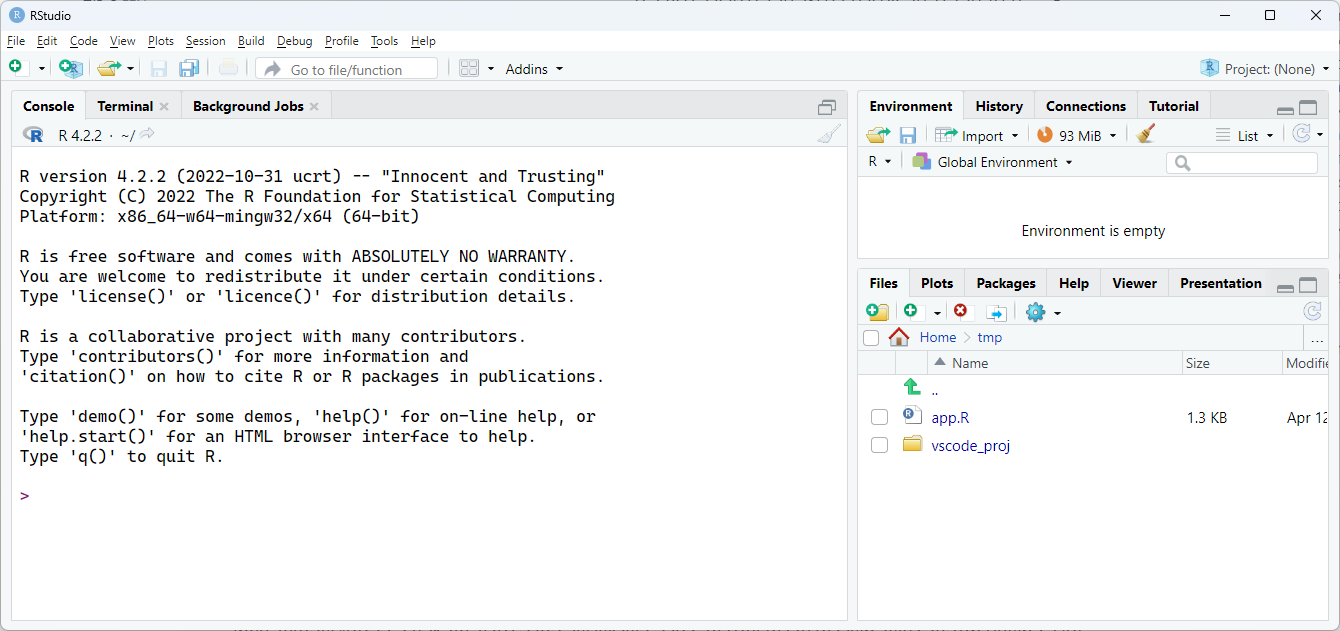
על מנת להתקין את R עליכם להיכנס לכתובת <https://cran.r-project.org> ולבחור בגרסה המתאימה למערכת ההפעלה שלכם. לאחר מכן, כדי להתקין את RStudio היכנסו לכתובת <https://posit.co/download/rstudio-desktop> הגרסה החינמית של RStudio נקראת RStudio Desktop והיא זמינה לכל צורך (כולל לשימוש במסגרת מסחרית).

הסיבה שאנחנו מתקינים גם את R וגם את RStudio היא שבהתקנת R בלבד, מקבלים ממשק משתמש בסיסי ביותר, בעוד שלאחר התקנת RStudio ניתן יהיה לעבוד עם ממשק משתמש נוח שיש לו הרבה אפשרויות שיקלו עליכם את הלמידה ואת העבודה השוטפת ב-R.

נציין שבעת כתיבת שורות אלו, יש גם סביבת פיתוח הנקראת Positron, גם היא בפיתוח של חברת Posit (אותו בית יוצר של RStudio), אך עדיין בגרסת בטא (לא שוחררה בגרסה סופית). פוזיטרון נועדה לתת מענה משופר של סביבת העבודה, ויש לה הרבה יתרונות, חידושים, ותוספות ביחס ל-RStudio. אם אתם רוצים, תוכלו להוריד ולנסות את גרסת הבטא שלה מהכתובת: <https://positron.posit.co>.

## איך משתמשים ב-RStudio?

לאחר ההתקנה, כשתפעילו את RStudio, תראו חלון המפוצל לשלושה (או אולי ארבעה) חלקים.



איור 2: חלון RStudio מיד לאחר ההתקנה.

החלק הגדול בצד שמאל הוא ה-Console, הוא סביבת ה-R המריצה את הקוד שלכם. אתם יכולים לנסות לרשום בו קוד ולראות מה קורה, לדוגמה תנסו לרשום 2+2 וללחוץ Enter או לרשום pi וללחוץ Enter.

2+2

[1] 4

pi

[1] 3.141593

בצד ימין למעלה ישנה חלונית המחולקת לטאבים. בין הטאבים שמופיעים בה תוכלו לראות את:

Environment - משתנים או פונקציות שנטענו לסביבה

History - היסטורית הפקודות שהרצתם

Connections - חיבורים למקורות נתונים

Tutorial - הצגה של מדריכים שונים שמלמדים איך להשתמש בחבילות של R

בצד ימין למטה תופיעה חלונית נוספת המחולקת גם היא לטאבים:

Files - תוכן התיקייה הנוכחית

Plots - תרשימים שנכין יופיעו בטאב זה

Packages - חבילות שזמינות לנו או שנטענו לסביבה (יסומנו ב-V)

Help - חלונית עזרה על פקודות

Viewer - צפיה בתכנים שונים שנכין (בעיקר תכנים אינטראקטיביים כמו טבלאות html תרשימים אינטראקטיביים, ואפליקציות).

בתור משתמשים מתחילים נשתמש בחלק קטן מאלו, אך ככל שיעבור זמן תראו שאתם משתמשים ביותר ויותר אפשרויות, וגם אפשרויות נוספות שכרגע אינן מופיעות לכם.

עוד כמה פינות חשובות שנמצאות ב-RStudio:

אם תלחצו על Help->Cheat sheets תוכלו לראות רשימה של “שליפים”. דפים שימושיים שמרכזים פקודות שונות של R וחבילות של R. לדוגמה, אחד מהשליפים ששם נקרא RStudio IDE Cheat Sheet, והוא מכיל הרבה הסברים נוספים על הממשק משתמש של RStudio.

בתפריט File -> New File ישנה אפשרות לפתוח קובץ חדש. מומלץ שתנסו לפתוח קובץ חדש מסוג R Script, שישמש אותנו לפרק הבא.

שימו לב: לאחר שפתחתם Script חדש, וכתבתם בו קוד, על מנת להריץ את הקוד צריך לסמן אותו וללחוץ Ctrl+Enter. זה יגרום ל-R להריץ את כל הקוד המסומן.

כעת אנחנו מוכנים לפרק הבא, שבו נלמד את בסיס השפה, מה שמכונה גם Base R.