**עבודה מסכמת ב-R**

מבוא לסטטיסטיקה, 2023-2024

**1. הנחיות כלליות**

יש להגיש את העבודה בזוגות עד לתאריך **29.04.2024 ב-23:59**. ההגשה תתבצע לתיבת הגשה ייעודית במודל, אליה יש להגיש שלושה קבצים:

* **קובץ קוד** (R Script).
* **קובץ Word**, שיכלול הסברים מפורטים לגבי סעיפי העבודה, יחד עם תוצאות וגרפים מה-R.
* **קובץ נתונים**, שבו נעשה שימוש לצורך העבודה.

שמות כל הקבצים יהיו מספרי תעודות הזהות של המגישים/ות (למשל: 123456789\_987654321). **רק נציג/ה אחד/ת מתבקש/ת להגיש את הקבצים לתיבת ההגשה במודל**.

שימו לב, מטרת העבודה היא לבדוק כיצד אתם/ן מתמודדים/ות עם למידה עצמאית של R, בהתבסס על השיעורים והתרגולים. לפיכך, צוות הקורס לא יענה על שאלות הקשורות בעבודה. בכל מקרה של התלבטות או ספק, החליטו עצמאית כיצד לפעול ונמקו זאת תחת הסעיף המתאים בקובץ ה-Word.

**1.1. מניעת הונאה אקדמית בהגשת עבודות**

במטרה למנוע העתקה, מרמה, ו/או מצב שבו רק אחד/ת מהמגישים/ות ביצע/ה את העבודה, אנו נדגום באופן אקראי סטודנטים/ות שיידרשו להגן על העבודה (להגן = שיחה עם מתרגל/ת, שבה יישאלו שאלות על העבודה). מטרת שיחות אלה היא לוודא שאתם/ן כתבתם/ן את העבודה עצמאית.

אם יתעורר חשד לאחד מהמצבים שתואר לעיל, הדבר ייחשב לעבירת משמעת מצד הזוג, גם אם רק אחד/ת מהם/ן ביצע/ה את העבירה בפועל.

**2. מקורות למציאת נתונים**

עליכם/ן לחפש קובץ נתונים שמעניין אתכם/ן וכולל מעל ל-5 משתנים, מתוכם לפחות שניים קטגוריאליים (בסולם שמי או סדר). כמו כן, רצוי שהקובץ יכיל מספר רב של תצפיות. ניתן לחפש קובץ נתונים באופן עצמאי או במקורות המצורפים להלן:

* קבצי נתונים הנמצאים בחבילות שונות ב-R – [לחצו כאן](https://vincentarelbundock.github.io/Rdatasets/datasets.html).
* [Kaggle](https://www.kaggle.com/datasets).
* [Google Dataset Search](https://datasetsearch.research.google.com/).

שימו לב – עליכם/ן לצרף את הקובץ לעבודה, לכתוב את מקורו ולצרף במסגרת מסמך ה-Word תיאור קצר של המשתנים בקובץ.

**על כל זוג לעבוד על קובץ נתונים אחר** – לשם כך, עליכם/ן לבדוק שהקובץ שברצונכם/ן לעבוד עליו לא נתפס ע"י זוג אחר ב[**טבלה זו**](https://docs.google.com/spreadsheets/d/1UXGgTYW1ZAlQ64YeV7HRQGkHEdxUnqWBtO7nIKD-IOw/edit?usp=sharing). אם הקובץ פנוי, יש למלא את [**הטופס המקוון**](https://forms.gle/H5iHHST2w3jzexMH9)על מנת לשריין את קובץ הנתונים.

**3. אופן כתיבת הקוד**

הקפידו על כתיבת קוד מסודרת – כותרת לכל סעיף, בשילוב הערות, שמבהירות לכם/ן ולנו את החשיבה שהובילה לכתיבת הקוד (להוספת הערות – השתמשו ב-**#**). בנוסף, יש לוודא שהקוד רץ במלואו ללא שגיאות.

אם תוך כדי העבודה תצוץ הודעת שגיאה שאינכם/ן מצליחים/ות לפתור, חפשו אותה ב-Google. רוב הסיכויים שתמצאו את הפתרון שם.

**4. סעיפי העבודה**

**א.** תארו את קובץ הנתונים עליו תעבדו באופן הבא: מה הנתונים מייצגים ואיך נאספו (ניסוי, סקר, איסוף ממאגרי מידע, מקורות אחרים...), כמה שורות (תצפיות) וכמה עמודות (משתנים) יש בקובץ, **במילים ספורות**: עבור כל משתנה, מה סוגו (שמי/קטגוריאלי, מספרי/רציף) ומה הוא מייצג (מגדר, גיל, זמן תגובה וכדומה).

**ב.** בחרו שני משתנים רציפים מקובץ הנתונים וחשבו את המתאם ביניהם. כעת צרו וקטור חדש המכיל 14000 ערכים חסרים. צרו לולאה שתרוץ 14000 פעמים ובכל פעם תעשה את הפעולות הבאות:

* דגימה מקרית בלי החזרה של 28% מהשורות בטבלה (שימו לב שייתכן ותצטרכו לעגל את המספר, בחרו אם למעלה או למטה והשתמשו בפונקציה המתאימה כדי לעשות זאת).
* יצירת טבלה חדשה המכילה רק את השורות שמספרן נדגם.
* חישוב המתאם בין אותם משתנים בטבלה החדשה.
* שמירת מקדם המתאם שחושב במקום המתאים (המקום ה-i) בוקטור הריק שיצרתם קודם.

חשבו את הממוצע וסטיית התקן של הוקטור המכיל את מקדמי המתאם, והציגו את הנתונים בווקטור בתרשים.

**ג.** בחרו אחד מהמשתנים הרציפים בסעיף ב' וצרו על בסיס הערכים שלו משתנה קטגוריאלי חדש בטבלה שמקבל את הערכים: High אם הערך במשתנה הרציף גדול או שווה מהממוצע, ו-Low אם הוא קטן מהממוצע. השתמשו בחבילה ggplot2 על מנת להציג את שני המשתנים הרציפים והמשתנה הקטגוריאלי באותו הגרף (רמז: צבעים!).

**ד.** בהמשך לסעיף ג', חשבו את אחוז השגיאות מהשכיח במשתנה הקטגוריאלי החדש.

**ה.** בהמשך לסעיפים ב' ו-ג', צרו משתנה קטגוריאלי נוסף על בסיס המשתנה הרציף השני מסעיף א' באותו אופן. צרו עבורם טבלת שכיחויות יחסיות המכילה את השכיחות היחסית של החיתוכים בין הרמות השונות של שני המשתנים. האם יש תלות בין המשתנים? השתמשו בטבלת השכיחויות היחסיות כדי לבדוק תלות, והסבירו איך התוצאה שקיבלתם/ן קשורה לזו של סעיף א'.

**ו.** השתמשו בחבילה dplyr מתוך tidyverse על מנת להחליף את כל הערכים הקיצוניים במשתנה רציף כלשהו (הגדירו בעצמכם מהו "קיצוני" וכתבו בהערה באמצעות הסימן #) בערך חסר – NA. צרו טבלה (מסגרת נתונים) חדשה המכילה את כל העמודות מהטבלה המקורית, ללא השורות בהן יש כעת ערך חסר (השתמשו בפונקציה filter). כיצד השתנתה השונות של המשתנה ומדוע?

**ז.** בחרו משתנה רציף מתוך הקובץ וצרו על בסיסו שני משתנים חדשים:

(א) טרנספורמציה ליניארית כלשהי של המשתנה המקורי.

(ב) אותה הטרנספורמציה ליניארית, ובנוסף הוסיפו לכל אחד מהערכים מספר שנדגם מהתפלגות נורמלית עם תוחלת 0 וסטיית תקן השווה לסטיית התקן של המשתנה הרציף המקורי.

מה המתאם בין המשתנה המקורי למשתנה א'? מה המתאם בין המשתנה המקורי למשתנה ב'? מה לדעתכם הסיבה להבדלים?

**ח.** צרו וקטור המכיל 14,000 ערכים אקראיים מהתפלגות מעריכית (ככה: 'rexp(14000)', לא למדתן אותה אך אין בכך צורך כדי לפתור את התרגיל) וקראו לו Y. הציגו את ההתפלגות בצורה גרפית – כיצד ההתפלגות נראית? כעת צרו וקטור שנקרא means באורך 14,000 המכיל בינתיים רק ערכים חסרים (NA). כעת כתבו לולאה שתרוץ 14,000 פעמים, בכל פעם הלולאה תדגום מקרית עם החזרה 5 ערכים מתוך הוקטור Y, ותשמור את הממוצע שלהם במקום ה-i בוקטור means.  
הציגו את הוקטור means בצורה גרפית. איך התפלגות הממוצעים נראית?  
כעת חזרו על התהליך, רק שבמקום לדגום מקרית כל פעם 5 ערכים, הלולאה תדגום מקרית 80 ערכים.  
כיצד נראית התפלגות הממוצעים כעת? כתבו במילים שלכם את ההבדל שאתם רואים.

**ט.** חזרו על התהליך בסעיף הקודם עם 5 דגימות אקראיות וצרו וקטור ריק חדש באורך 14,000. כעת כתבו לולאה שעבור כל ערך בוקטור means תכניס לוקטור החדש את המילה "bigger" אם הערך גדול או שווה מהממוצע, ואת המילה "smaller" אם הוא קטן מהממוצע. אתם אמורים לקבל וקטור באורך 14,000 המכיל את המילים "bigger" ו-"smaller". איזה ערך שכיח יותר בוקטור?  
כעת חזרו על התהליך עם גודל מדגם של 80. מה החלוקה בין "bigger" ו-"smaller" עכשיו? הסבירו כיצד זה מתיישב עם התפלגויות הממוצעים שהצגתן בסעיף ז'.

**י.** חשבו על שני ניתוחים סטטיסטיים מעניינים שאתם יכולים לבצע על קובץ הנתונים שבחרתם. בצעו אותם והציגו אותם גרפית. לדוגמה: הבדל בין ממוצעים של משתנה אחד כתלות בערכים של משתנה קטגוריאלי אחר, הבדלים בין מקדמי מתאם וכדומה. היו יצירתיים (!), והסבירו את הניתוח, התוצאות שלו והמסקנות שלכם ממנו.

**יא.** עשו שימוש בשתי פונקציות **חדשות** שלא נלמדו בקורס על מנת לנתח את הנתונים בצורה מסוימת, להציג אותם בצורה גרפית, או עבור פתרון בעיה סטטיסטית אחרת. הסבירו מה הפונקציה עושה וכיצד משתמשים בה.

ב ה צ ל ח ה !