משימת פרויקט 1

מטרת הפרויקט בקורס היא תרגול מעשי של הכלים והשיטות הנלמדים בשיעורי הקורס. במהלכו תנתחו קובץ נתונים בשיטות שונות בהתאם לנושאים הנלמדים לאורך הסמסטר.

מטרות המשימה הראשונה:

1. בחירת קובץ נתונים מתאים להנחיות בהמשך
2. ניתוח תיאורי ראשוני של הנתונים
3. ניסוח שאלות מחקר רלוונטיות

קובץ נתונים מתאים

אתם רשאים לבחור כל אחד מ-6 הקבצים הבאים (ללא קבלת אישור מסגל הקורס):

- https://archive.ics.uci.edu/ml/datasets/wine+quality
- https://www.kaggle.com/wenruliu/adult-income-dataset?select=adult.csv
- https://www.kaggle.com/henriqueyamahata/bank-marketing
- https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset
- https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease-prediction
- https://www.kaggle.com/rio2016/olympic-games

תוכלו לבחור גם בקבצי נתונים נוספים ובלבד שיעמדו בדרישות להלן ויהיו נגישים לסגל הקורס, בכפוף לאישור. **שימו לב שלא ניתן להחליף את קובץ הנתונים עליו תעבדו לאחר מטלת הפרויקט הראשונה, בחרו בתבונה.**

במקרה שהקובץ גדול, ניתן להשתמש רק בחלק שלו (למשל, רק בנתונים לגבי עיר ספציפית או חודש ספציפי). ניתן לבצע מניפולציות על המשתנים (למשל, אפשר להחליף את משתנה הגיל במשתנה קטגוריאלי עם שני ערכים "ילד/ה" / "מבוגר/ת").

הקובץ הסופי בו אתם משתמשים חייב לענות על הדרישות הבאות:

- מכיל לפחות 2 משתנים נומריים (רציפים, גם ערכים בין 1 ל100 לצורך העניין יכולים להיחשב כרציפים).
- מכיל לפחות 2 משתנים בינאריים.
- יש בו לפחות 4000 רשומות.
- המשמעות של כל המשתנים ברורה לכם.

הגשת חלק זה תכלול את החלקים הבאים:

- פסקה שתכיל תיאור קצר של קובץ הנתונים.
- קישור לקובץ. במידה והשתמשתם בחלק מהנתונים או ביצעתם טרנספורמציות לחלק מהמשתנים, צרפו קטעי קוד שמבצעים זאת.

- רשימת העמודות בקובץ, סוג המשתנים בעמודה, תיאור קצר של משמעות העמודה, מספר הרשומות הכולל בקובץ.

ניתוח תיאורי ראשוני של הנתונים

לכל משתנה נומרי הציגו סיכום ערכים סטטיסטיים משמעותיים והציגו את התפלגות הערכים בצורה גרפית (היסטוגרמה, או boxplot). לכל משתנה קטגוריאלי, תארו בעזרת ייצוג גרפי הולם את התפלגות הערכים בקטגוריות. בדקו האם יש נתונים חסרים או חריגים ודווחו על כך.

חשבו על דרכים חכמות ויצירתיות להציג את הנתונים כדי שהנמען של הויזואלציות יפיק את מיטב ההבנה תוך שימוש יעיל בגרפים.

ניסוח שאלות מחקר

נסחו **לפחות** שלוש שאלות מחקר.

א. נסחו שאלת רגרסיה שבה יש משתנה מסביר רציף ומשתנה מוסבר רציף (למשל האם עליה של משתנה X גורמת לירידה במשתנה Y).

ב. נסחו שאלת רגרסיה שבה יש משתנה מסביר רציף ומשתנה מוסבר בינארי (למשל האם עליה של משתנה X גורמת לירידה בהסתברות שהמשתנה Y שווה לאחד).

ג. נסחו שאלת מבחן – האם הערך של משתנה רציף X שונה בין קטגוריות שונות של משתנה בינארי Y.

במהלך הפרויקט ניתן להחליף את שאלות המחקר, אבל השאלות מוודאות שהנתונים מתאים לשאלות מסוג זה.

פורמט הגשה

כלל מטלות הפרויקט יוגשו בקובץ ipynb. (קובץ Jupyter notebook).
ת"ז המגישים יופיעו בראש הקובץ וכן שם הקובץ יהיה בפורמט **ProjectEx1_ID1_ID2.ipynb**.

בנוסף, יש להגיש קובץ PDF/HTML שמכיל הרצה של המחברת שלכם, בפורמט השם הנ"ל, בתוספת הסיומת הרלוונטית.