מטלת 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.

<u>מטרות המשימה הראשונה:</u>

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

קובץ נתונים מתאים

אתם רשאים לבחור כל אחד משלושת הקבצים הבאים (ללא קבלת אישור מסגל הקורס):

- **נתונים על סטודנטים והצלחה אקדמית**
https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

- **נתוני ביקוש מלונות** –
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

- **הסקר החברתי האירופאי** –
https://ess.sikt.no/en/datafile/bcc624a3-edbb-4df3-ab18-9a96702c92ae/67?tab=0

תוכלו לבחור גם בקבצי נתונים נוספים ובלבד שיעמדו בדרישות להלן ויהיו נגישים לסגל הקורס, בכפוף לאישור. **לא יאושרו קבצי נתונים שניתנו בשנים קודמות לשימוש בפרויקט**.

שימו לב שלא ניתן להחליף את קובץ הנתונים עליו תעבדו לאחר מטלת הפרויקט הראשונה, בחרו בתבונה.

<u>הקובץ הסופי בו אתם משתמשים חייב לענות על הדרישות הבאות:</u>

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

<u>שיקולים מנחים בעריכת הנתונים:</u>

1. אנו ממליצים, לצורך הנוחות החישובית, לא לעבוד עם קבצי נתונים הגדולים מסדר גודל של 10,000 דגימות. במקרה שמספר הדגימות גדול בהרבה, ניתן לסנן חלק מהדגימות על

ידי בחירה באקראי או על ידי סינון של תת אוכלוסייה מתוך האוכלוסייה הנדגמת (למשל, שימוש בנתונים לגבי עיר ספציפית או חודש ספציפי). אם בחרתם בתת אוכלוסייה מסוימת יש לציין זאת במפורש.

2. בחרו משתנים כך שהמאגר יהיה עשיר מספיק כדי לשאול עליו שאלות מעניינות (ראו סעיף ניסוח שאלות מחקר) אך מצומצם מספיק כדי שלא יכביד עליכם במהלך ביצוע הפרויקט – לרוב כעשרה משתנים יספיקו.
3. ניתן לבצע מניפולציות על המשתנים (למשל, אפשר להחליף את משתנה הגיל במשתנה קטגוריאלי עם שני ערכים "ילד/ה" / "מבוגר/ת").

<u>הגשת חלק זה תכלול את החלקים הבאים:</u>

- פסקה שתכיל תיאור קצר של קובץ הנתונים.
- קישור לקובץ. במידה והשתמשתם בחלק מהנתונים או ביצעתם טרנספורמציות לחלק מהמשתנים, צרפו קטעי קוד שמבצעים זאת.
- עבור קובץ הנתונים המתקבל לאחר הטרנספורמציות שביצעתם, צרפו רשימת עמודות בקובץ, סוג המשתנים בעמודה, תיאור קצר של משמעות העמודה, מספר הרשומות הכולל בקובץ.

שימו לב שקובץ הנתונים המתקבל לאחר הטרנספורמציות שבצעתם ישמש אותנו לכל בדיקה עתידית של מטלות הפרויקט.

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 **<u>לפחות</u>** 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 שבחרתם מתאימים לשאלות מסוג זה.

פורמט הגשה

כל 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.

סך הכל יש להגיש שני קבצים. אין להגיש קובץ zip.