

מטלת פרויקט 1

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

מטרות המשימה הראשונה:

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

קובץ נתונים מתאים

אתם רשאים לבחור אחד משני הקבצים הבאים (לא קבלת אישור מסלול הקורס):

- **נתונים על תאונות בארה"ב**
 - <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>
 - בשימוש נתונים אלו מומלץ ליצור משתנה בינהרי מוסבר בדוגמה "האם חומרת התאונה גדולה מ[ערך לבחירתכם]".
- **2021 - Behavioral Risk Factor Surveillance System (BRFSS)**
 - ה-BRFSS הוא סקר שנתי רחב היקף של ה-Centers for Disease Control and Prevention (CDC) בארצות הברית. הסקר אוסף נתונים על הרגלי בריאות, מחלות ברוניות וטיפול מונע, בהתקפס על דיווח עצמי של מאות אלפי מוגרים מכל רחבי ארצות הברית.
 - הנתונים מאפשרים בחינה של הקשרים בין מצב סוציאו-דמוגרפי, הרגלי חיים ומדדים בריאותיים.
 - הנתונים המלאים זמינים לציבור באתר ה-CDC:
https://www.cdc.gov/brfss/annual_data/annual_2021.html
 - בנוסף, הכלל תיאורים של כל המשתנים וקוד הערכים שלהם, במסמך-*Codebook* המלא, אשר מופיע בקישור הבא:
https://www.cdc.gov/brfss/annual_data/2021/pdf/codebook21-2-v2-508.pdf
 - מאגר נתונים מצומצם: כדי להקל על שילוף הנתונים, מצורף למטרה קוד `python` אשר מוריד את הנתונים המקוריים ישירות מתוך ה-CDC ומסנן לתת-מודגם עם כ-20 משתנים אשר בחרנו עבורכם. תוכלו להעתיק את הקוד וישירות ל-`Colab` / `jupyter notebook` ולהשתמש בו. הקוד וטבלה מסכמת של הנתונים שנבחרו מצורפים בקבצים נפרדים למטרת הפרויקט.
 - אם אתם מעוניינים לבצע את הפרויקט ב-R, כתבו מייל למתרגלת האחראית כדי לקבל קוד מקביל.
 - אתם יכולים להיעזר ב-*codebook* המלא של הנתונים ולהוסיף משתנים נוספים או להחליף חלק מהמשתנים שנבחרו. לשם כך, עדכנו את רשימת המשתנים `cols` בקובץ הקוד המצורף בהתאם לבחירתכם. אין צורך לקבל בקבלת אישור מראש עבור שינוי זה.

בקשה לאישור קובץ נתונים חלופי

1. יש לשלוח את הבקשה במיל למתרגלת האחראית ביצירוף קישור לקובץ הנתונים. הקובץ חייב לעמוד בכל הדרישות המפורטות במסמך זה.
2. לבקשתה יש לצרף שתי שאלות מחקר על הנתונים – אחת לרוגסיה לינארית וחת משתנים מסוימים לפחות, כך שלפחות אחד מהמשתנים המסבירים יהיה רציף ולפחות אחד יהיה בדיד (קטגוריאלי או ביןארי). השאלות חיברות להיוות בעלות היון מחקרי בסיסי; לא יתקבלו שאלות טריויאליות או לא סבירות (למשל "בצד גיל משפייע על מין").
3. **לא יאשרו קבצי נתונים שכבר נעשו בהם שימוש בפרויקטם בקורס בסמסטרים קודמות.**

שימוש לב שלא ניתן להחליפּ את קובץ הנתונים עליו תעבור לאחר מטלת הפרויקט הראשונה, בחרו בתבונה.

הקובץ הסופי בו אתם משתמשים חייב לענות על הדרישות הבאות:

1. מכל לפחות 2 משתנים נומריים (רציפים, גם ערכים בין 1 ל-100 לצורך העניין יכולים להיחשב רציפים).
2. מכל לפחות 2 משתנים ביןaries.
3. יש בו לפחות 4000 רשומות.
4. המשמעות של כל המשתנים ברורה לכם.

שיקולים מנהיים בעריכת הנתונים:

1. אם ממליצים, לצורך הנוחות החישובית, לא לعباد עם קבצי נתונים הגודלים 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 המדגמת (למשל שימוש נתונים לגבי עיר ספציפית או חדש ספציפי). אם בחרתם בתת אוכלוסייה מסוימת יש לציין זאת במפורש.
2. בחרו משתנים כך שהמארג יהיה עשיר מספיק כדי לשאול עליו שאלות מעניינות (ראו סעיף ניסוח שאלות מחקר) אך מצומצם מספיק כדי שלא יכיד עליהם במהלך ביצוע הפרויקט – לרוב בעשרה משתנים יספיקו.
3. ניתן לבצע מניפולציות על הנתונים (למשל, אפשר להחליף את משתנה הגיל במשתנה קטגוריאלי עם שני ערכים "ילד/ה" / "מבוגר/ת").

הגשת חלק זה תכלול את החלקים הבאים:

- פסקה שתכילה תיאור קצר של קובץ הנתונים.
- קישור לקובץ. במידה והשתמשם בחלק מהנתונים או ביצעתם טרנספורמציות לחלק מהמשתנים, צרפו קטעי קוד שמבצעים זאת.
- עברו קובץ הנתונים המתkeletal לאחר הטרנספורמציות שביצעתם, צרפו רשימה מעודדת בקובץ, סוג הנתונים בעמודה, תיאור קצר של משמעות העמודה, מספר הרשומות הכלל בקובץ.

שימוש לב שקובץ הנתונים המתkeletal לאחר הטרנספורמציות שבצעתם ישמש אותנו לכל בדיקה עתידית של מטלות הפרויקט.

ניתוח תיאורי ורפואי של הנתונים

כל משתנה ממרי הציגו סיבום ערכיים סטטיסטיים משמעותיים והציגו את התפלגות הערכים בΖ' (היסטוגרמה או boxplot). כל משתנה קטגוריאלי, תארו בעזרת ייצוג גרפי הולם את התפלגות הערכים בקטגוריות. בדקו האם יש נתונים חסרים או חריגים ודוחו על כך.

חשבו על דרכי חבמות ויצירתיות להציג את הנתונים כדי שהנמען של היזואלציות יפיק את מיטב ההבנה תוך שימושיעיל בגרפים.

ניסוח שאלות מחקר

- נסחו פחות שלוש שאלות מחקר.
- נסחו שאלת גרסיה שבה יש משתנה מסביר רציף ומשתנה מוסבר רציף (למשל האם עליה של משתנה X גורמת לירידה במשתנה Y).
 - נסחו שאלת גרסיה שבה יש משתנה מסביר רציף ומשתנה מוסבר בינארי (למשל האם עליה של משתנה X גורמת לירידה בהסתברות שהמשתנה Y שווה לאחד).
 - נסחו שאלת מבחן – האם הערך של משתנה רציף X שונה בין קטגוריות שונות של משתנה בינארי Y.

במהלך הפרויקט ניתן להחליף את שאלות הממחקר, אבל השאלות מודאות שהנתונים שבחרתם מתאימים לשאלות מסווג זה.

פורמט הגשה

כל מטלות הפרויקט יוגשו בקובץ **ipynb**. (קובץ Jupyter notebook). (קובץ notebook).

ת"ז המציגים יופיעו בראש הקובץ וכן שם הקובץ יהיה בפורמט **ProjectEx1_ID1_ID2.ipynb**.

スク הצל יש להגיש שני קבצים. אין להגיש קובץ דוח.