

## תוכנות מתקדם - מטלה 4

נושא: עיבוד נתונים

הנחיות:

1. נא לכתוב קוד מסודר, תוך שימוש בשמות משתנים בעלי משמעות והערות היכן שנדרש.
2. כל שאלה צריכה להיות מוגשת בנפרד
3. בתחילת כל מענה, נא להוסיף כהערה את שמות המגישים והאם התיעצרתם נעזרתם בסטודנטים נוספים.

הגשה:

1. יש להגיש את העבודות בזוגות (עפ"י הקבוצות שנרשמהם).
2. יש לשתף מחברת בודדת (Jupyter notebook) עם המטלה.
3. שם הקובץ המוגש צריך להיות: `student1_id_student2_id.ipynb`

**בצלחה!**

## הקדמה

במקרה זה, המשימות ניתנות למימוש במספר דרכים שונות. בסעיפים בהם ניתן, יש להימנע משימוש באיטרציות (לולאות) ובמקום יש להשתמש בפונקציונות של `Pandas` / `Numpy`.  
אוסף הנתונים המצורף, "StudentsPerformance.csv", הוא קובץ הנתונים למקרה זה, אשר מטרתו היא עיבוד הנתונים לҚරאת ניתוח והסקת מסקנות. מצורפות 5 שורות הראשונות בטבלה:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group C	bachelor's degree	standard	completed	100	100	100
1	female	group A	associate's degree	free/reduced	none	45	65	58
2	male	group D	high school	free/reduced	completed	87	81	87
3	female	group B	bachelor's degree	free/reduced	none	80	87	85
4	male	group D	some high school	standard	none	81	82	70

gender: female, male

ethnicity: group A, B, C, D, E

education: some college, associate's degree, high school, some high school, bachelor's degree, master's degree

lunch: standard, free/reduced

pre-course: none, completed

math, reading, writing: grades 0 - 100

## חלק א' (50 נקודות) - טיפול בערכים חסרים ומשתנים קטגוריאליים

1. הדפס תחקור ראשוני עבור הנתונים: גודל הנתונים, שמות העמודות ומדדים סטטיסטיים לעמודות נומריות (מצורף: describe).
  2. הציג דיאגרמת bar המציגת כמות הנתונים אשר קיימים עבור כל קבוצה ב-ethnicity.
  3. מהו אחוז הסטודנטים אשר לומדים בתיכון ביחס לכל הנתונים?
  4. במידה וקיימים ערכים חסרים, הדפס את שמות העמודות בהן חסרים ערכים ואת כמות הערכים החסרים. אם לא, הסביר כיצד הבחנתת כי אין ערכים חסרים נתונים.
  5. המר את עמודת gender לקידוד one-hot-vector. כלומר, לאחר פעולה זו, יתווסף שני עמודות חדשות נתונים: (gender\_male, gender\_female) המכילים ערכים בינהירים בלבד.(0/1)
  6. בהתבסס על התשובה של סעיף 5 בלבד, כמה גברים וכמה נשים קיימים נתונים?
  7. השתמש בפקודות Pandas, ושנה את שם העמודה "race/ethnicity" ל - "ethnicity".
  8. כתוב את הפונקציה get\_last\_char אשר מקבלת כקלט מחרוזת ומחזיר אותה את התו האחרון בחרוזת.
  9. הפעיל את הפונקציה הנ"ל על עמודות ethnicity (שים לו: לא ליצור עמודה חדשה, אלא **לדריש** את הקיימת).
  10. עבור עמודת test preparation course, המר את הערך "completed" לערך 1, ואת השאר ל-0. (כלומר, לאחר סעיף זה, עמודת test preparation course היא עמודה בינהירית).
  11. השתמש בקידוד one-hot-vector encoding והמר את עמודות "lunch" בהתאם.
  12. עבור עמודת education, המר את ערכיה כך ש:
    - a. אם הערך הוא some college, bachelor's degree, master's degree - המר את ערך העמודה להיות "higher education".
    - b. אם הערך הוא associate's degree - המר את ערך העמודה להיות "degree".
    - c. אחרת, המר את ערך העמודה להיות "high school".
  13. המר את עמודת "ethnicity" לקידוד מספרי **לא** שימוש Pandas.
- הՃרכה:** הגדר מילון אשר מיפה את קבוצות-holiday למספרים רצופים (למשל, 1=A, 2=B...) והפעיל את מיפוי הערכים על העמודה הרלוונטית.

## חלק ב' (20 נקודות) - נרמול נתונים + שאלות

1. עברו עמודות math, reading, writing - המר אותם עפ"י הכלל הבא:  
נסמן ב-z את ממוצע העמודה וב-s את סטיית התקן של העמודה. הערך החדש בכל אחת מהעמודות הוא הערך המקורי פחות z, לפחות ל-s. במילים אחרות, הפחיתו מכל ערך את ממוצע העמודה ולאחר מכן יש לחלק בסטיית התקן.
2. עברו כלל קבוצות ה-"ethnicity" הקיימות, מי היא הקבוצה אשר קיימים בה היא הרבה סטודנטים אשר השתלימו קורס\מכינה לפני הלימודים?
3. הצג בדיאגרמה (לבחירתך) את היחס בין ממוצע ציוני הסטודנטים במתמטיקה עפ"י כל אחת מקבוצות ההצטיינות. כלומר, חשב לכל קבוצה את ממוצע הציוניים במתמטיקה והציג תוצאה זו בצורה גרפית.
4. צור DataFrame חדש המכיל את העמודות המנווארות של ציוני הסטודנטים במתמטיקה, קרייה וכתיבה (מסעיף 1). הפעיל את הפקודה הבאה:

```
new_df.hist()
```

מה הוא הפלט? הסבירו במשפט אחד.

5. הציגו שתי דיאגרמות Pie המציגות את התפלגות הציוניים במתמטיקה, קרייה וכתיבה עבור נשים \ גברים. במילים אחרות, גרפ' Pie ראשון עברו התפלגות הנשים והשני עברו התפלגות הגברים.

## חלק ג' (30 נקודות)

חלק זה במלה אינו קשור לחלקים הקודמים.

- כתבו את הפונקציה `distance`, המתקבלת שתי רשימות (זהות באורך) ומחזירה את המרחק בין רשימות אלו.

марחק בין רשימות הוא כמו מרחק בין נקודות. למשל, אם איברי הרשימה הראשונה מסומנים ב-X ואיברי הרשימה השנייה מסומנים ב-Y, אז המרחק הוא:

$$d(lst1, lst2) = \sqrt{(x_0 - y_0)^2 + (x_1 - y_1)^2 + \dots + (x_{n-1} - y_{n-1})^2}$$

- צור DataFrame בגודל 100 עם שתי עמודות: הראשונה `id` (מספר אקראי שלם כלשהו ללא כפליות), והשנייה עמודת `item`, אשר מכילה את רנדומלית כלשי מהקבוצה: A,B,C,D,E.
- הדפס את `value_counts` עבור עמודת `item` ובדוק כי אכן מופיעים רק האותיות הנ"ל.
- המר את עמודת `item` ל-one-hot-vector encoding.
- כעת, אוסף הנתונים אמר לhiveות (בערך) כך:

<code>id</code>	<code>item_A</code>	<code>item_B</code>	<code>item_C</code>	<code>item_D</code>	<code>item_E</code>
91	0	1	0	0	0
90	1	0	0	0	0
33	0	0	0	1	0

- עבור כל שורה, הוסף עמודה חדשה המכילה את מרחק השורה מהרשימה [0,1,0,0,0].  
למשל, עבור השורה השנייה, נבדק את המרחק בין [0,1,0,0,0] לבין [1,0,0,0,0].
- הפעל `groupby` על העמודה החדשה, ובדוק מי הם השורות אשר מרחקם הוא בדיק 0. מה ניתן להגיד על שורות אלה?
- במידה ומרחק השורה הוא לא אפס, מה הוא המרחק המקסימלי אשר מתאפשר מסעיף 4? הסבר במשפט אחד מדווג.