תכנות מתקדם - מטלה 3

נושא: הספריות Numpy, Pandas, Matplotlib

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 התייעצתם נעזרתם בסטודנטים נוספים.

הגשה:

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

בהצלחה!

(18 נק׳) - חלק א'

צרו DataFrame המכיל 3 עמודות: **x, y, z**. הערך של המשתנה x הוא מערך רנדומלי עם 10,000 איברים. יש להשתמש בפקודה np.random.normal(mean, std, size) עם ממוצע של 0 וסטיית תקן של 2. בנוסף,

$$y = x^3 - 2x^2$$

$$z = \sqrt{e^y} \cdot sin(y)$$

עבור כל אחת מהשאלות הבאות, יש להדפיס תשובה מלאה. למשל, עבור שאלה 1, נדפיס "The maximum value of y is …"

ענה על השאלות הבאות:

1. מה הוא הערך המקסימלי של y?
2. מהו הערך המינימלי של z?
3. הדפס את הערכים עבורם מתקיים $y = z$
4. הדפס את 5 השורות הראשונות של עמודות x ו-z.
5. צור את העמודה $3y - z^3$ וחשב את ערכה (יש להוסיף עמודה זו ל-DataFrame הקיים).
6. הצג גרף scatter עבור עמודות x,y.
7. הצג גרף scatter עבור עמודות x,z.
8. הוסף לגרפים ב6,7: כותרת, שמות לצירים, צבעים, ועוד.
9. השתמש בפקודה df.sample לבחירת סט ערכים אקראיים מהנתונים (**<u>טיפ</u>**: העזר בקובץ במודל "פקודות נפוצות"). בחר, בצורה אקראית, 35% מהנתונים והצג גרף של עמודת $3y - z^3$.

(60 נק׳) - חלק ב' - ניתוח והבנת נתונים

לחלק זה, מצורף קובץ **flights.csv** המתאר אוסף טיסות. הקובץ מכיל את הנתונים הבאים:

	Carrier	Day	DepTime	Dest	Origin	Weather	Delayed
0	OH	WED	1455	JFK	BWI	0	0
1	DH	WED	1640	JFK	DCA	0	0
2	DH	WED	1245	LGA	IAD	0	0
3	DH	WED	1709	LGA	IAD	0	0
4	DH	WED	1035	LGA	IAD	0	0

Carrier - למי שייכת הטיסה? מיוצג באמצעות ראשי תיבות של המחוז בארצות הברית

Day - יום בשבוע

DepTime - שעת המראה

Dest - יעד

Origin - מקור

Weather - 0 (Sunny) or 1 (Clouds) or 2 (Rainy)

Delayed - 0 if the flight departed on time, otherwise 1.

בנוסף, קיימת עמודה נוספת, flightId אשר מייצגת את המספר המזהה של הטיסה.

(8 נק׳) - א. תחקור ראשוני

1. קראו את הקובץ הנתון לאובייקט DataFrame.
2. הדפיסו את מספר השורות והעמודות בנתונים.
3. הדפיסו את שמות העמודות ובדקו עבור כל עמודה האם קיימים בה ערכים שהם ריקים (NaN).
4. הדפיסו את היום בשבוע אשר התרחשו בו הכי הרבה טיסות.

(40 נק׳) - ב. שאילתות

עבור כל אחת מהשאילתות, יש לענות תשובה מלאה הכוללת את תשובתכם.

למשל, עבור חיפוש מחוז בארה"ב:

“The state in the USA that has the most number of flights is…”

ענו על כל אחת מהשאילתות הבאות:

1. מהו המחוז בארה"ב אשר מנהל הכי הרבה טיסות?
2. כמה טיסות התבצעו **<u>אל</u>** ניו-יורק (EWR)?
3. כמה טיסות המריאו מDCA ונחתו ב-JFK?
4. מהו אחוז הטיסות המאחרות מתוך סך הטיסות?
5. מהו אחוז הטיסות המאחרות מתוך סך הטיסות, אשר טסו תחת מזג אוויר גשום?
6. לכמה יעדים שונים טסה חברת התעופה של DH?
7. כמה טיסות המריאו בין השעות 10:00 70 ל-15:00?
8. כמה טיסות המריאו בין 14:00 ל-17:00 בימי רביעי?
9. כמה טיסות המריאו **<u>אל</u>** JFK בין 13:00 ל-17:00 בימי רביעי?
10. מה היא שעת ההמראה המאוחרת ביותר שקיימת בנתונים?

(12 נק׳) - ג. ויזואליזציה

בחרו 3 שאילתות (מתוך ה10 הקיימות) והציגו את תוצאתם בצורה ויזואלית. יש לכלול כותרות, צבעים, תוויות לצירים וכו'. הוסיפו משפט תיאור **אחד** לכל גרף.

(22 נק׳) - חלק ג'

כידוע, ערך בינארי הוא ערך המקבל 0/1, True/False. השתמשו באוסף הנתונים flights מחלק ב', ובצעו את הסעיפים הבאים:

1. עבור עמודת Deyaled, הציגו כמה פעמים הופיע 0 וכמה פעמים הופיע 1.
2. עבור אותה עמודה, הדפיסו את **סכום** העמודה.
3. מה ניתן להגיד על משתנים בינארים? כיצד הסכום קשור למספר הפעמים אשר כל איבר הופיע? ענו במשפט אחד.
4. הדפיסו את הממוצע של העמודה. בהתחשב על תוצאתכם לשאילתה 4 ("מהו אחוז הטיסות המאחרות מתוך סך הטיסות?"), מה ניתן להגיד על הממוצע של משתנים בינארים? הסבר במשפט אחד.
5. ללא קשר לסעיפים הקודמים, צרו מערך של 25,000 מספרים אקראיים בטווח בין 0 ל1. עבור כל ערך במערך אשר קטן ממש מ0.5, החליפו אותו ב0, אחרת - החליפו אותו ב1. חשבו את ממוצע המערך שקיבלתם וכתבו **במשפט אחד** מה ניתן להסיק מכך.
6. מיינו את המערך מסעיף 5 בסדר יורד, כלומר: קודם כל האחדות ולאחר מכן האפסים.