094202 - מבוא לניתוח נתונים בפייתון

אביב תשפ"ג - תרגיל בית 2

הנחיות

- הגשת תרגיל הבית תיעשה עד לתאריך 8.6.2023 בשעה 23:55.
  - שימו לב: תיבת ההגשה במודל תיסגר 48 שעות לאחר מועד זה, זאת על מנת להימנע מהורדת נקודות על איחורים (כמפורט בסילבוס), עליכם להגיש לפי המועד המצוין כאן.
- הגשת התרגיל היא בזוגות בלבד (פרט למקרים חריגים באישור מתרגל אחראי).
- ההגשה תכלול (לפחות) שני קבצים (לא קובץ **ZIP** יחיד) - מחברת ג'ופיטר עבור שבה יהיה הקוד והתשובות לתרגיל, וקובץ PDF המכיל את תוכן מחברת ה - jupyter (עם הפלטים).
  - ניתן לייצא מחברת jupyter ל - PDF ע"י הדפסת המחברת (ctrl+P או דרך האפשרויות בדפדפן) ובחירה באפשרות "save as pdf".
  - ניתן למזג את קובץ הPDF המכיל את ייצוא החלק הראשון עם קובץ ה - PDF המכיל את פתרון החלק השני באמצעות כלי merge PDF המוצעים בחינם באינטרנט, כמו למשל: https://tools.pdf24.org/en/merge-pdf
  - כמפורט בסילבוס, על סטודנטים המשתמשים בכלי בינה מלאכותית גנרטיביים להגיש בנוסף קובץ docx המפרט את השימוש שנעשה. אנא פנו לסילבוס להנחיות פרטניות בנושא זה.
- כל תשובה חייבת להיות מגובה בפלט קוד, אלא אם כן נאמר אחרת. כל תשובה בחלק הויזואליזציה חייבת להיות מלווה בויזואליזציה. בחלק ב' חובה לצרף ויזואליזציה בסעיפים הדורשים זאת.
- על שמות הקבצים המוגשים להיות בפורמט הבא: 'ID1_ID2_HW1.pdf', 'ID1_ID2_HW1.ipynb' כאשר ID1 ו – ID2 הם מספרי תעודות הזהות של המגישים/ות. אין צורך להגיש את קובץ הנתונים.
- הסילבוס כולל נהלים מפורטים הנוגעים להכנה והגשה של תרגילי הבית. חובה לעמוד בנהלים אלו.
- חריגה מהנחיות התרגיל ו/או איחור בהגשה יגררו הורדת ניקוד, בהתאם למפורט בסילבוס הקורס.

חלק א' - עיבוד נתונים

תיאור קובץ נתונים

קובץ הנתונים הוא תיעוד של תאונות אוויריות שנלקחו ממאגר ה-Aviation Safety Network, פירוש העמודות הוא כדלהלן:

שם עמודה	תיאור
date	התאריך בו התרחשה התאונה
type	סוג המטוס
operator	החברה/צבא המפעילה את המטוס
fatalities	מספר ההרוגים בתאונה
country	המדינה בה התרחשה התאונה
cat	הקטגוריה של התאונה לפי NAS: <u>Letter:</u> A = Accident I = Incident H = Hijacking C = Criminal occurrence (sabotage, shoot down) O= other occurrence (ground fire, sabotage) U= type of occurrence unknown <u>Number:</u> 1 = hull-loss 2 = repairable damage
year	השנה בה התרחשה התאונה

שאלות - כללי

עבור כל אחת מן השאלות יש לכתוב תשובה ברורה בטקסט בתא markdown, התשובה חייבת להיות מבוססת על פלט קוד שיופיע בתא קוד.

1. כמה רשומות יש בקובץ הנתונים?
2. כמה ערכים חסרים יש בכל עמודה?
3. עבור אילו סוגי מטוסים היו לפחות 500 תאונות מתועדות?
4. מהו אחוז התאונות המתועדות שהתרחשו בארה"ב?
5. איזה קטגוריית תאונה היא הקטלנית ביותר, בממוצע?

שאלות - ויזואליזציה

בחלק זה עליכם ללוות כל תשובה בויזואליזציה (מצג גרפי). על הויזואליזציות להיות ברורות ולכלול כותרות ושמות מתאימים לכל הצירים עם גודל גופן קריא. אין צורך להשתמש בשיטות של בחינת השערות שלמדנו בהרצאות האחרונות בשאלות אלה.

6. א. הראו את ההתפלגות של מספר ההרוגים בתאונות לפי מדינות (עבור 5 המדינות בהן התרחש מספר התאונות הגדול ביותר ועבור המטוסים שעבורם קיימים לפחות 500 תאונות מתועדות - מסעיף 3). השתמשו בשיטה המסכמת את ההתפלגות ומאפשרת השוואה בין ההתפלגויות השונות (רמז: בתרגולים ראינו שיטה שמציגה חמישה ערכים מסכמים על ההתפלגות (מינימום, רבעון ראשון, חציון, רבעון שלישי ומקסימום).
   ב. לפי הגרף שנוצר, באיזו מדינה חציון מספר ההרוגים הוא הגבוה ביותר?
7. האם יש קורלציה בין מספר התאונות בכל שנה ולשנה עצמה? בסעיף זה אין צורך להראות חישוב של קו רגרסיה, ויזואליזציה מספיקה.
8. מהם שמות עשרת המטוסים "המסוכנים" ביותר (מטוסים שעבורם התרחש מספר התאונות הגדול ביותר)? הציגו גרף של מספר התאונות של כל אחד.
9. מי הם שלושת המפעילים "המסוכנים" ביותר (מפעילים שבצי שלהם יש לפחות מטוס מסוכן אחד מעשרת המטוסים המסוכנים ביותר שביצע תאונה ומספר התאונות שהתבצעו בהם הוא הגדול ביותר)? הציגו את מספר התאונות שהתרחשו עבור כל מפעיל.
10. א. האם מספר התאונות של שלושת המטוסים המסוכנים ביותר פוחת במשך השנים? ספקו שני הסברים אפשריים לתשובה שמצאתם.
    ב. נסו להסביר את התצפיות החריגות (רמז: התשובה לא נמצאת בקובץ הנתונים, תצטרכו לבצע מחקר קצר ולחשוב על ההסבר)

חלק ב' - בדיקת השערות וסימולציה

בתרגיל זה נרצה לייצר היסטוגרמה של ערכי p-value ממבחני בדיקת השערות מסוימים כמפורט מטה. לשם כך, עליכם לבצע את השלבים הבאים:

1. גובה אדם בוגר באוכלוסייה מסוימת מתפלג נורמלית עם ממוצע (למעשה, תוחלת) 175 ס"מ וסטיית תקן של 6 ס"מ. השתמשו בפקודה np.random.normal כדי להגריל מדגם בגודל $n = 40$ של גבהים מאוכלוסייה זו. הציגו בגרף את ההתפלגות של הנתונים.
   a. האם זהו גרף של ההתפלגות האמיתית (probability distribution) או ההתפלגות האמפירית (empirical distribution)?
   b. מהו ממוצע הגבהים במדגם שקיבלתם?
   c. מהי צורת ההתפלגות שקיבלתם? בפרט, מהם ה-skewness (התפלגות סימטרית, מוטה לימין או לשמאל) וה-modality (התפלגות עם שיא יחיד, שני שיאים, שיאים רבים, אחידה)? האם ואיך תשובתכם צפויה להשתנות אם הייתם לוקחים מדגם בגודל $n = 1000$?
2. נניח כי אינכם יודעים מה מקור המדגם שהגרלתם בסעיף 1 ואתם מעוניינים לבדוק את ההשערה שממוצע הגבהים באוכלוסייה ממנה הוא נלקח שווה ל-175 ס"מ או שהוא נמוך מ-175.
   a. נסחו את השערת האפס וההשערה האלטרנטיבית
   b. בדקו את ההשערה באמצעות סימולציות. הניחו שסטיית התקן באוכלוסייה ידועה ועומדת על 6 ס"מ. השתמשו ב-2000 סימולציות ($num\_repetitions = 2000$)
   c. מהו ערך ה-p-value שקיבלתם?
   d. מהי מסקנתכם אם רמת המובהקות הנדרשת היא 0.1? ואם היא 0.01?
3. ממשו את הפונקציה $get\_p\_value\_heights(sample\_heights, n, mean\_0)$ אשר מקבלת מדגם של גבהים sample_heights באורך n ומחזירה p-value של המבחן הבודק את ההשערה שממוצע הגבהים באוכלוסייה שווה ל-mean_0 או קטן ממנו, תחת ההנחה שסטיית התקן באוכלוסייה עומדת על 6 ס"מ.
4. כעת, באמצעות הפונקציה מסעיף 3, נרצה לחזור 5000 פעמים על התהליך מסעיפים 1 ו-2. בכל איטרציה של התהליך נגריל מדגם חדש בגודל $n = 40$ של גבהים מהאוכלוסייה ואז נמצא p-value של מבחן הבודק באמצעות המדגם (החדש) את ההשערה שממוצע הגבהים באוכלוסייה ממנה הוא נלקח שווה ל-175 (שוב, תחת ההנחה שסטיית התקן היא 6).
   a. בלי לכתוב קוד, דונו במספר משפטים: איך לדעתכם תראה היסטוגרמה של 5000 ערכי p-value כאלה? מדוע אתם חושבים כך? (כל עוד התשובה מנומקת היטב, אין תשובה שגויה לשאלה זו. דונו לפני הרצת הסימולציה בפועל)
   b. ממשו את התהליך המתואר והציגו את ההיסטוגרמה של ערכי p-value שהתקבלו. (שימו לב: זמן הרצת הקוד עלול להיות ארוך, כדאי לנסות ראשית להריץ עם מספר קטן יותר של איטרציות כדי לראות שהקוד עובד).
   c. מהי (פחות או יותר) צורת ההתפלגות של ערכי p-value שנוצרו? האם היא תואמת לציפיותיכם?

d. מבין ערכי p-value שקיבלתם, מהו אחוז הערכים הקטנים מ-0.05? הסבירו בקצרה מדוע זה הגיוני תוך התייחסות להגדרה של p-value.

5. מה לדעתכם יקרה להיסטוגרמה אם גודל כל אחד ממדגמי הגבהים יגדל מ-$n = 40$ ל-$n = 200$? הסבירו במספר משפטים ולאחר מכן בדקו באמצעות שינוי הקוד. (שימו לב: זמן הרצת הקוד יהיה ארוך עוד יותר)