094202 - מבוא לניתוח נתונים בפייתון

תרגיל בית 3

הנחיות

- הגשת תרגיל הבית תיעשה עד לתאריך 29/06/2023 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_HW3.pdf', 'ID1_ID2_HW3.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'

תיאור קובץ הנתונים

קובץ הנתונים מכיל מדגם של משכורות של עובדים במקצועות הקשורים בעיבוד נתונים (לצורך קיצור נקרא להם מעתה עובדי "מקצועות הנתונים") שנאסף בתחילת שנת 2023 בארצות הברית.

feature	**Description**
experience_level	The experience level in the job during the year. (SE : Senior, EN : Entry level, EX : Executive level, MI : Mid/Intermediate level)
employment_type	The type of employment for the role. (CT: Contract worker, FL: Freelancer, FT: Full Time, PT: Part Time)
job_title	The role worked in during the year.
Salary_in_usd	The salary in USD ($)
remote_ratio	The overall amount of work done remotely. (Fully Remote: 100% of work is remote, Hybrid: less than 100% remote work)
company_size	The median number of people that worked for the company during the year

שאלות

*הערה - כאשר נבקש להסביר את קטע הקוד שכתבתם נצפה לתיעוד הקוד במידה מספקת.

פונקציות: הסבירו מי הם המשתנים ומה סוגם, מה הפונקציה עושה ומהי התוצאה המתקבלת (משמעותה וסוג המשתנה).

מקטעי קוד/תאים - הסבר מילולי קצר (משפט/שני משפטים).

1. סקר שנערך בחודש פברואר 2023 קבע כי אחוז האמריקאים שעובדים מהבית באופן מוחלט הוא 35%:
https://www.pewresearch.org/short-reads/2023/03/30/about-a-third-of-us-workers-who-can-work-from-home-do-so-all-the-time/
נניח שנתון זה נמדד גם בקרב עובדי "מקצועות הנתונים" בארה"ב ונמצא דומה. האם קובץ הנתונים מייצג היטב את האוכלוסייה של עובדי "מקצועות הנתונים" בארה"ב או שהוא כולל תת ייצוג לעובדים שעובדים מהבית ב-100% מהזמן? בדקו באמצעות המשתנה *remote_ratio* (רמת מובהקות נדרשת: 0.05)
   א. ציינו באופן ברור את השערת האפס וההשערה האלטרנטיבית

ב. מהו סטטיסטי המבחן

ג. כתבו קוד לבחינת ההשערה באמצעות סימולציות. הסבירו את הקוד שכתבתם.

ד. מהי המסקנה שלכם? הציגו תוצאה מספרית וכן גרף המדגים את תוצאת המבחן

2. קיימת טענה בתעשייה לפיה Data Scientists מקבלים בממוצע משכורת שווה למשכורת שמקבלים בממוצע Data Engineers. בדקו את הטענה בהסתמך על הדאטה הנתון, בהנחה שהוא מייצג את אוכלוסיית עובדי "מקצועות הנתונים" בארה"ב.

א. ציינו באופן ברור את השערת האפס וההשערה האלטרנטיבית.

ב. מהו סטטיסטי המבחן?

ג. כתבו קוד לבחינת ההשערה באמצעות רווח סמך (עם לפחות 5000 רפליקציות). הסבירו את הקוד שכתבתם.

ד. מצאו רווחי סמך לערכי הפרמטר שחיפשתם ברמות ביטחון של 0.99 ו-0.95.

ה. מהי המסקנה שלכם? הציגו תוצאות מספרית וכן גרפים המדגימים את תוצאת המבחן לשתי רמות הביטחון המבוקשות.

3. חשבו רווח סמך של 95% עבור ממוצע המשכורות של עובדים בחברות בינוניות (M). הסבירו את הקוד שכתבתם וציינו באופן מפורש את רווח הסמך (ערך עליון ותחתון).

4. חברת ניתוח נתונים כלכליים עולמית רוצה לדעת מה המשכורת החציונית שמשולמת לעובדים בתחום הדאטה. מנתח נתונים חדש ומפוזר בחברה מחק בטעות את קובץ הנתונים. כדי למצוא את הערך החציוני, הוא החליט לערוך מדגם ולחשב את החציון באמצעות בוטסטראפ. הוא מצא קובץ עם נתונים על עובדים בתחום הדאטה באחד המיילים שקיבל, מבלי שידע שהאנשים המופיעים בקובץ הם רק אנשים שאינם עובדים מהבית לחלוטין (hybrid). העובד בחר באקראי 150 אנשים מתוך קובץ העובדים ושאל אותם מהי המשכורת הנוכחית שלהם. מכיוון שהבטיח פרס כספי למשיבים, כל האנשים שקיבלו את המייל ענו על השאלה. הניחו שהנתונים עליהם דיווחו המשיבים הם נתוני אמת התואמים את קובץ הנתונים המקורי.

א. כדי לבדוק מה הסיכוי שהעובד הצליח ליצור מתוך המדגם (של העובדים ההיברידיים) רווח סמך של 95% המכיל את החציון האמיתי של משכורות כל העובדים, הריצו 100 סימולציות הדוגמות כל פעם 150 אנשים מתוך אוכלוסיית העובדים ההיברידיים ומחשבות על כל אחד מהמדגמים רווחי סמך של 95% עבור הערך החציוני של המשכורת שלהם. בכמה מתוך 100 רווחי הסמך שקיבלתם מוכל החציון האמיתי של עובדים בתחום הדאטה כפי שניתן לחשבו מקובץ הנתונים המקורי המלא?

ב. מנתח הנתונים השתמש במדגם שלא נלקח באקראי מהאוכלוסייה אותה הוא בוחן (אלא מאוכלוסייה אחרת). האם, למרות זאת, מתקיימת בפועל ההבטחה התיאורטית של שיטת הבוטסטראפ ליצירת רווח סמך? מדוע ההבטחה מתקיימת או לא מתקיימת? הסבירו את תשובתכם. בסעיף זה אין צורך בכתיבת קוד.

ג. איך היו משתנות תשובותיכם לסעיפים א' ו-ב' אם חברת ניתוח הנתונים הייתה מעוניינת לדעת מהי המשכורת הרבעונית של עובדים בתחום (כלומר במקום לחשב חציון, מנתח הנתונים היה מחשב את הרבעון הראשון של המשכורות). הסבירו.

רמז: ציירו את התפלגויות המשכורות או סיכום שלהן בנפרד עבור שתי קבוצות העובדים.

חלק ב'

1. מדוע בעייתי להשתמש בבוטסטראפ כדי לחשב רווח סמך בהתבסס על מדגם מאוד קטן? הסבירו ב-2-4 משפטים.
2. חוקרים בארה"ב בחנו את ההשפעה של גזע על ההחלטות של שופטים לדון למוות אדם שהורשע ברצח. הנתונים שלהם כללו 702 הרשעות ברצח במדינת לואיזיאנה. נמצא כי, בסך הכל, 16.4% מהמורשעים הלבנים נידונו למוות, לעומת 15.2% מהמורשעים השחורים. עם זאת, כאשר הם כללו באנליזה גם את גזע הקורבן, הם מצאו כי מורשעים שחורים נידונים למוות באחוזים גבוהים יותר מאשר מורשעים לבנים כאשר הקורבן לבן וגם כאשר הקורבן שחור. ראו פירוט בטבלה:

		נידון למוות?		
גזע הקורבן	**גזע הנאשם המורשע**	**כן**	**לא**	**אחוז נידונים למוות**
לבן	לבן	42	176	19.2
	שחור	29	84	25.7
שחור	לבן	1	43	2.3
	שחור	38	289	11.6
סך הכל	לבן	43	219	16.4
	שחור	67	373	15.2

א. איזו תופעה (סטטיסטית) יכולה להסביר את הפער בתוצאות בין המצב בו מתייחסים למצב בו לא מתייחסים לגזע הקורבן? הסבירו ב-2-3 משפטים.

ב. בהתבסס על הטבלה:

1. איזו נטייה של השופטים יכולה לעזור להסביר את הפער בתוצאות? רמז: באיזה מצב יותר סביר שיינתן גזר דין מוות?
2. איזו נטייה של המורשעים יכולה לעזור להסביר את הפער בתוצאות? רמז: מה הקשר בין גזע הקורבן לבין גזע הנאשם?