094202 - מבוא לניתוח נתונים בפייתון

אביב תשפ"ג - תרגיל בית 1

הנחיות

- הגשת תרגיל הבית תיעשה עד לתאריך 18.5.23 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 החלק הראשון של התרגיל, וקובץ PDF המכיל את תוכן מחברת ה - jupyter (עם הפלטים) ואת הפתרון עבור החלק השני של התרגיל.
  - 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.
- בחלק הראשון של התרגיל, כל תשובה חייבת להיות מגובה בפלט קוד, אלא אם כן נאמר אחרת. בחלק השני של התרגיל אין דרישה לכתיבת קוד.
- על שמות הקבצים המוגשים להיות בפורמט הבא: 'ID1_ID2_HW1.ipynb', 'ID1_ID2_HW1.pdf' 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׳ - חקירת מערך נתונים

בחלק זה תתבקשו לנתח מערך נתונים המציג נתונים על 6 ליגות מובילות בכדורגל. הורידו את קובץ מערך הנתונים HW1_data.csv מאתר הקורס במודל.

ענו על כל אחת מהשאלות הבאות. עבור כל אחת מן השאלות יש לכתוב תשובה ברורה בטקסט בתא markdown. התשובה חייבת להיות מבוססת על פלט קוד שיופיע בתא קוד. ניתן לכתוב את הטקסט בתא ה - markdown בעברית או באנגלית. אם אתם כותבים בעברית, אנא ודאו שהטקסט לא מתערבב (רדו שורה אם אתם עוברים בין מילים באנגלית לעברית).

להלן פירוט של השדות במערך המידע:

פירוט	שדה
שם הליגה	league
שנת התיעוד	year
מיקום הקבוצה בליגה (ראשון, שני וכו')	position
שם הקבוצה	team
כמות המשחקים ששיחקה הקבוצה בליגה	matches
כמות ניצחונות הקבוצה בליגה	wins
כמות התיקו של הקבוצה בליגה	draws
כמות ההפסדים של הקבוצה בליגה	loses
כמות השערים שכבשה הקבוצה בליגה	scored
כמות השערים שספגה הקבוצה בליגה	conceded
כמות הנקודות שקיבלה הקבוצה	pts

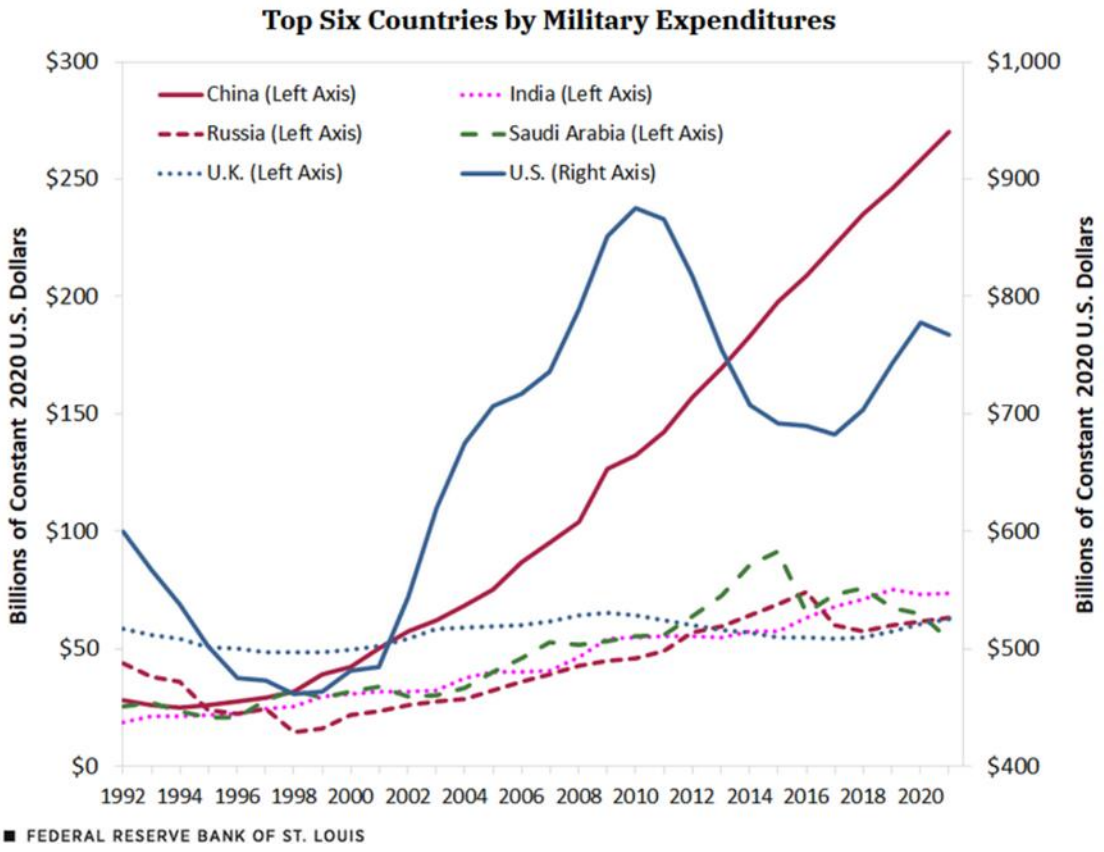
1. כמה רשומות קיימות בתוך מערך הנתונים?
2. אלו מן השדות במערך המידע הם מספריים (numerical)? אלו הם קטגוריאלים (categorical)?
3. כמה ערכים ייחודיים (unique) קיימים בכל אחד מהשדות הקטגוריאלים במערך הנתונים?

4. באילו מהשדות של מערך הנתונים ישנם ערכים חסרים (ערכי null), וכמה ערכים חסרים יש בשדות אלו?
5. כמה קבוצות שונות שיחקו בכל אחת מהליגות במהלך השנים שהנתונים מתייחסים אליהם? איזו ליגה כללה את המספר הקטן ביותר של קבוצות שונות?
6. הציגו את ממוצע כמות השערים שכבשה כל קבוצה במשחק עבור כל שנה. הציגו את הקורלציה בין ממוצע כמות השערים שכבשה כל קבוצה לבין מיקום הקבוצה בליגה. הסבירו את משמעות התוצאה שקיבלתם.
7. הציגו, לכל ליגה ולכל שנה בנפרד, את חציון מספר השערים שכבשה קבוצה באותה ליגה ושנה. בהתבסס על הפלט שהצגתם/ן בלבד (כלומר, אין צורך לכתוב קוד נוסף), באיזו ליגה היה הפער הגדול ביותר בין חציון מספר השערים הגבוה ביותר מבין השנים לחציון מספר השערים הנמוך ביותר מבין השנים? בליגה זו, באיזו שנה היה החציון הנמוך ביותר ובאיזו שנה החציון הגבוה ביותר?
8. מספר הנקודות שצוברת קבוצה מחושב כך: ניצחון במשחק שווה 3 נקודות, תיקו שווה נקודה אחת, והפסד שווה 0 נקודות. העמודה pts אמורה לתאר חישוב זה. חשבו בעצמכם את מספר הנקודות שקיבלה כל קבוצה על סמך נתוני המשחקים והשוו את מספר הנקודות שחישבתם לאלו הנתונים בעמודה pts. האם נראה שיש טעויות? אם כן, באיזו ליגה, באיזו שנה ועבור איזו קבוצה? האם לדעתכם מקור הטעות בעמודה pts או באחת או יותר מהעמודות באמצעותן חישבתם את מספר הנקודות שצברה הקבוצה?
9. נגדיר קבוצה "נכשלת" כקבוצה שספגה בשנה מסוימת יותר שערים מכמות השערים שהבקיעה. שימו לב כי קבוצה עבורה הנתון של מספר השערים שנספגו חסר (null) איננה קבוצה נכשלת. מהי הליגה בה היו הכי הרבה קבוצות נכשלות (במצטבר, לאורך כל השנים)?

חלק ב׳ - הצגת מידע

שאלה 1

בחנו את הגרף הבא וענו על השאלות שאחריו



A. מה תוכלו להסיק מן הגרף המוצג לגבי ההוצאה הצבאית של מדינות שונות בעולם לאורך השנים 1992 - 2020?

B. הצביעו על בעיה בגרף שעלולה לגרום לצופה בו להסיק מסקנות שגויות לגבי ההוצאה הצבאית של מדינות בעולם לאורך השנים 1992 - 2020.

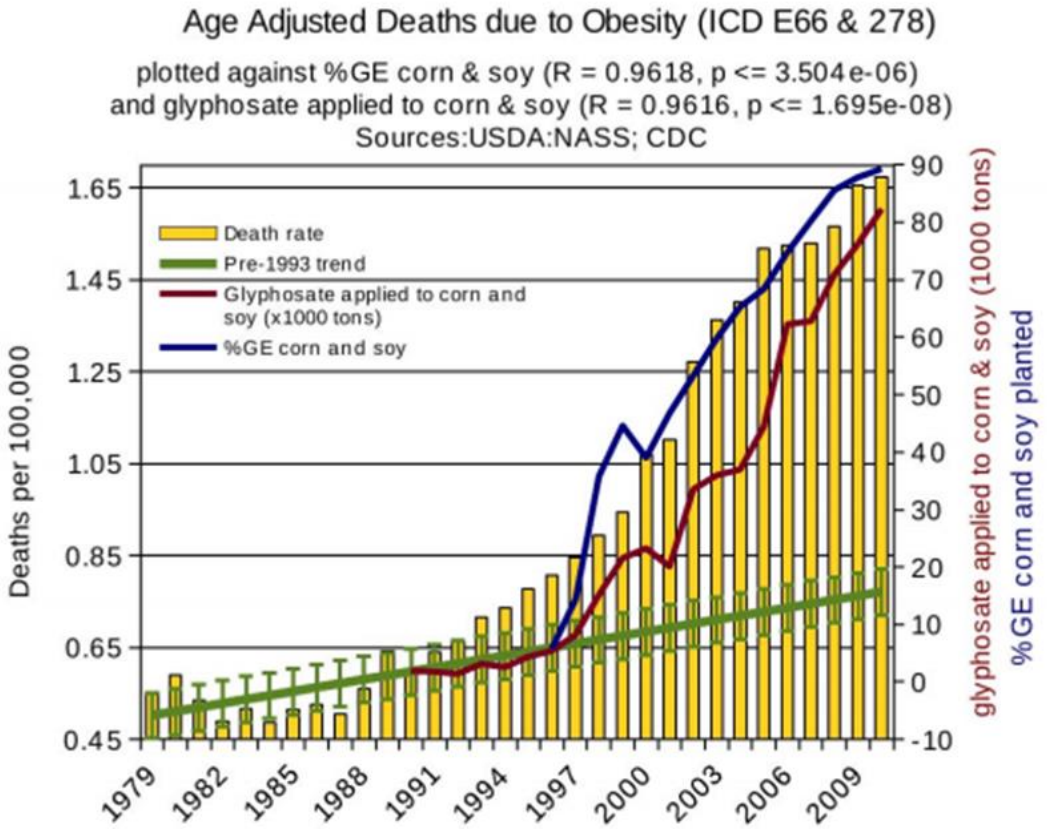
C. הציעו ויזואוליזציה חלופית שתציג את המידע שאמור להיות מועבר לצופה ואת הנתונים המוצגים בצורה אפקטיבית. אין צורך בכתיבת קוד, אך הוסיפו סקיצה פשוטה שתבהיר כיצד הויזואליזציה החלופית שהצעתם תיראה.

שאלה 2

בחנו את הויזאוליזציה הבאה וענו על השאלות שאחריה.

שימו לב: Glyphosate הוא סוג של קוטל עשבים הנפוץ בגידולים חקלאיים רבים;

GE = Genetically engineered = מהונדס גנטית; obesity = עודף משקל.



A. מה ניתן להסיק לגבי הקשר שבין מוות מעודף משקל, השימוש ב-Glyphosate בגידולי תירס וסויה, ואחוז גידולי תירס וסויה המהונדסים גנטית?

B. בעיצוב הגרף, יוצריו קיבלו מספר החלטות לא סטנדרטיות על מנת להעביר מסר מסוים. מהו לדעתכם המסר שיוצרי הגרף ביקשו לייצר באמצעותו? הסבירו. האם אתם מסכימים עם המסר?

C. מצאו בויזואליזציה לפחות 2 בעיות שגורמות לה להיות פחות אפקטיבית. הסבירו מהי הבעיה ומדוע זו נחשבת בעיה.

שאלה 3

בחנו את הויזואליזציה הבאה וענו על השאלות שאחריה. **שימו לב:** CAGR = קצב צמיחה שנתי ממוצע

A. מהו המידע שאמור להיות מועבר לצופה באמצעות הויזואליזציה הזו?

B. מצאו בויזואליזציה לפחות 3 בעיות שגורמות לה להיות פחות אפקטיבית. הסבירו מהי הבעיה ומדוע זו בעיה.

C. הציעו ויזואוליזציה חלופית שתציג את המידע שאמור להיות מועבר לצופה ואת הנתונים המוצגים בצורה אפקטיבית. אין צורך בכתיבת קוד, אך הוסיפו סקיצה פשוטה שתבהיר כיצד הויזואליזציה החלופית שהצעתם תיראה.

שאלה 4

מהי הבעיה בכותרת הבאה (NC = North Carolina):

Average NC teacher pay is nearly $58,000, state says. But educators argue many earn less.