מעבדה בסטטיסטיקה 52568 - 2020-21 ,מטלה 6. להגשה עד ה-6.12.

תיאור המשימה: המעבדה עוסקת בהשוואה בין 2 מערכות בחירות ומציאת קלפיות חשודות.

שאלות:

1. השוו את תוצאות הבחירות השניות ב-2019 (ספטמבר) לתוצאות בחירות 2020 (מרץ) באופן הבא:

א. אחדו את קבצי **הישובים** של 2 מערכות הבחירות כאשר לכל ישוב יהיו שני וקטורי תצפיות שונים (אחד עבור כל מערכת בחירות). אפשר להשתמש בפקודות append או concatenate של pandas. כדי להתמודד עם השינויים בשמות והרכבי המפלגות בין שתי מערכות הבחירות נשתמש בנתונים רק עבור 9 המפלגות הגדולות בבחירות 2020 (המפלגות שעברו את אחוז החסימה ומפלגת עוצמה יהודית). מפו את הקולות מ-10 מפלגות הגדולות בבחירות ספטמבר 2019 ל-9 מפלגות הגדולות, ע"י איחוד קולות המפלגות המחנה הדמוקרטי + העבודה-גשר מספטמבר 2019 כקולות העבודה-גשר-מרץ במרץ 2020. לאחר מכן נרמלו את הוקטור של כל ישוב כך שיכיל את שכיחויות ההצבעה למפלגות מתוך כלל הקולות שקיבלו 9 המפלגות הגדולות בישוב.

ב. הציגו תמונה של הישובים (לאחר איחוד שתי מערכות הבחירות) במישור הנפרש ע"י שני המרכיבים הראשיים המובילים (PCA), כששטחי עיגולי הישובים פרופורציונים לגודלם. לכל ישוב יהיו שני עגולים (אחד עבור כל מערכת בחירות) - צבעו את העיגולים לפי מערכת הבחירות. תנו משמעות לצירים המייצגים את שני ה-PCs הראשונים.

ג. עבור כל אחד מהישובים עם מעל 10,000 בוחרים בספטמבר 2019, הוסיפו לתמונה חץ המתחיל בעיגול של הישוב בבחירות ספטמבר 2019 ומסתיים בעיגול הישוב בבחירות 2020. תוכלו להשתמש בפקודה plt.arrow מתוך matplotlib. תארו את המתקבל - אילו מגמות מראים החצים?

ד. מצאו את שלושת הישובים עבורם השינוי בין שתי מערכות בחירות היה מקסימלי - כאשר השינוי מוגדר להיות המרחק הריבועי בין ייצוגי הישוב במישור הנפרש ע"י שני המרכיבים הראשיים המובילים ב-2 מערכות הבחירות. הוסיפו חצים ל-3 ישובים אלו כמו בסעיף ג' אך בצבע אחר וציינו את שמות הישובים ליד החצים. בנוסף, עבור כל אחד מ-3 הישובים האלו ציירו גרף ובו bar-plot כפול המתאר את שכיחויות ההצבעה ל-10 המפלגות בישוב ב-2 מערכות הבחירות (לכל מפלגה 2 עמודות בצבעים שונים). תארו את השינויים העיקריים בהצבעה בישובים אלו.

2. א. התאימו את קבצי **הקלפיות** בין מערכת הבחירות של מרץ 2020 למערכת הבחירות של ספטמבר 2019 כך שתקבלו תוצאות עבור קלפיות משותפות (שם ישוב ומספר קלפי זהה). ניתן להשמש בפונקציה adapt_df וכן ב-join של pandas.

שימו לב: בשאלה זו יש בחירה: ענו על 2 סעיפים מבין סעיפים ב'-ה', ובכן בהתאמה על סעיף ו' עבורם:

ב. מצאו את 10 הקלפיות עבורן סכום המרחקים הריבועיים בין שכיחויות ההצבעה ל-9 המפלגות הגדולות בבחירות מועד ב למועד ג היה מקסימלי

ג. מצאו את 10 הקלפיות בהן אחוז ההצבעה הכללי (ממוצע על פני 2 מערכות הבחירות) היה מקסימלי.

ד. מצאו את 10 הקלפיות בהן חל השינוי הרב ביותר באחוז ההצבעה הכללי בין 2 מערכות הבחירות.

ה. מצאו את 10 הקלפיות בהן חל השינוי הרב ביותר בשכיחות ההצבעה לחמש מפלגות הימין ('מחל', 'טב', 'שס', 'ג', 'כף/נץ') בין 2 מערכות הבחירות.

ו. עבור כל אחד מהסעיפים אותו עשיתם מבין ב'-ה', ציירו גרף bars כפול של שכיחויות ההצבעה ל-9 המפלגות הגדולות בשתי מערכות הבחירות (כל מערכת בחירות בצבע אחר) עבור 10 הקלפיות שמצאתם תוך שימוש ב-subplot , עם 5X2 גרפים. ציינו בהם את שמות הישובים והקלפיות, וכן מספר הקולות הכשרים ומספר בעלי זכות הבחירה בכל אחת מ-2 מערכות הבחירות. חפשו הסברים לשינויים עבור 10 הקלפיות. האם סביר שהשינויים משקפים מגמות אמיתיות בין 2 מערכות הבחירות? אם כן, מהן? האם יש אינדיקציה לטעויות או אי סדרים ברישום הקולות? אם כן, באיזה ממערכות הבחירות?

[בונוס] האם יש לכם עוד דרכים למצוא מפלגות חשודות? מותר להיות יצירתיים

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.