מעבדה לסטטיסטיקה 52568 - 2019-20 ,מטלה 1. להגשה והצגה ב-3.11

במטלה זו נקרא את קובץ ההצבעה בבחירות לכנסת ה-22 (ספטמבר 2019) ונציג את דפוסי ההצבעה בישובים

הדרכה: את כל חלקי המשימה ניתן לבצע על ידי שינויים קלים של קובץ הפייתון election_data_analysis.py שהודגם בכיתה. תחילה, הורידו את קובץ זה, וכן את קובץ הנתונים 'Votes per city - 2019b elections' מהמודל ושימו אותם בתיקיית עבודה לבחירתכם. עליכם לשנות את ה-DATA_PATH המופיע בקובץ הפייתון לתיקיה בה נמצאים הנתונים. כמו כן, יש להתקין את המודולים numpy, pandas, os, matplotlib כדי לעבוד עם הקובץ. תוכלו לקרוא את הקובץ לתוך data-frame באמצעות פקודת read_csv כפי שהראינו בכיתה ומתואר בקובץ הפייתון.

1. מצאו את העמודות המציינות את מספר הקולות הפסולים ואת מספר כלל המצביעים בכל ישוב וציירו היסטוגרמה של **אחוז** הקולות הפסולים בכל ישוב. השתמשו ב-100 תאים שווי מרחק בין 0 לבין אחוז הקולות הפסולים המקסימלי.
בנוסף, מצאו את הישוב עם אחוז הקולות הפסולים המקסימלי וציינו את שמו ואת האחוז.
2. כתבו פונקציה המציגה עבור שני ישובים bar-plot של ההצבעה ל-9 המפלגות שעברו את אחוז החסימה (מעל 3.25% הצבעה כלל ארצית) שניהם אחד ליד השני בצבעים שונים כפי שהודגם בכיתה. בחרו 3 זוגות של ערים גדולות (מעל 50000 מצביעים) והציגו את ההצבעה ב-3 זוגות הערים בעזרת הפונקציה שכתבתם. תארו את ההבדלים והדמיון בין הערים בהצבעה למפלגות השונות.
3. עבור שתי התפלגויות מולטינומיות עם n ערכים עם ההסתברויות:
$$\mathbf{p}=[p_1,\ldots,p_n],\ \mathbf{q}=[q_1,\ldots,q_n]$$
נגדיר את המרחק הריבועי להיות:
$$d(\mathbf{p},\mathbf{q})=(p_1-q_1)^2+\ldots+(p_n-q_n)^2$$
(זוהי דוגמא לפונקציית מרחק בין התפלגויות).
א. מצאו את הישוב המייצג ביותר את אחוז ההצבעה בכלל ישראל - כלומר הישוב שהתפלגות ההצבעה בו (על פני **כל** המפלגות) היא **הקרובה** ביותר להתפלגות ההצבעה הארצית עבור מדד זה (ממזערת את המרחק הריבועי מהתפלגות ההצבעה הארצית) והציגו bar-plot של הישוב ביחד עם התפלגות ההצבעה הארצית. הסבירו את תשובתכם.
ב. חזרו על א' עבור הישוב שהתפלגות ההצבעה בו היא **הרחוקה** ביותר מהתפלגות ההצבעה הארצית.

הערות:

- כדי לשמור גרף לקובץ השתמשו בפקודה savefig במודול matplotlib. יש דוגמא לשימוש כזה בקובץ הפייתון המצורף.
- החישוב שהוצג בכיתה ומופיע בקובץ הפייתון עבור מספר המצביעים הכללי אינו נכון, מכיוון שהוא סיכם את מספר המצביעים רק עבור המפלגות הגדולות ביותר, והתעלם ממפלגות קטנות ומקולות פסולים. השתמשו במספר המצביעים הכללי הנתון בקובץ בכל ישוב במקום זאת.
- חשבו על עיצוב הגרפים. תנו כותרת לצירים וכותרת כללית לגרף. תנו כותרות לעמודות במידת הצורך. הוסיפו legend המתאר כל קו במידת הצורך. שימו לב לגבולות הצירים. השתמשו בצבעים וכו׳ כדי להדגיש הבדלים.
- מותר להיות יצירתיים; נסו לחפש ישובים הדומים זה לזה בהצבעה על פני כל המפלגות, לאפיין ישובים עם אחוז קולות פסולים גבוה וכו'