

מעבדה לסטטיסטיקה 52568 - 2019-20, מטלה 2. להגשה והצגה ב-10.11

תיאור המשימה:

במשימה זו נציג את הקשרים בין הצבעה למפלגות שונות בבחירות לכנסת ה-22 (ספטמבר 2019) וכן את השפעת אחוז החסימה על ההצבעה למפלגות השונות

- א. כתבו פונקציית scatter-plot של שתי מפלגות כפי שהודגם בכיתה ונמצא בקובץ לדוגמא במודל, אבל כך שהשטח של כל עיגול בפלוט יהיה פרופורציוני למספר הקולות הכללי ביישוב. לצורך כך השתמשו בפרמטר s בפונקציית scatter-plot של matplotlib. כיילו את השטח כך שהישוב עם מספר הקולות המקסימלי יקבל ערך 200.
ב. בחרו 3 זוגות של מפלגות וציירו scatter-plot בעזרת הפונקציה שכתבתם עבור כל זוג. תארו במילים מה רואים בגרף והאם ומה זה מעיד על הקשר בין המפלגות.
- א. חשבו עבור כל זוג מפלגות מבין 10 המפלגות הגדולות ביותר את מקדם המתאם של **pearson** בין שכיחות המצביעים שלהם על פני כל הישובים כפי שהודגם בכיתה. הציגו heatmap המתאר את מטריצת מקדמי המתאם בגודל 10×10 כאשר המפלגות ממויינות על פי גדלן. הוסיפו colormap לתמונה. תארו מה קיבלתם.
ב. הציגו שוב את heatmap אבל הפעם סדרו את המפלגות על פי נטייתן הפוליטית, מימין לשמאל (גם בשורות וגם בעמודות, באותו הסדר). אם אינכם בטוחים בסדר בין 2 מפלגות הפעילו את שיקול דעתכם והסבירו. תארו את התמונה החדשה שקיבלתם. איזה סידור אינפורמטיבי יותר?
ג. חזרו על סעיף ב' אבל הפעם חשבו והציגו את מקדם מתאם הדרגות של **spearman**. זהו מקדם מתאם א-פרמטרי שלא מניח הנחות התפלגותיות על הנתונים. תוכלו לקרוא עליו כאן: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
תוכלו להשתמש בפונקציות `spearmanr`, `pearsonr` מתוך המודול `scipy.stats` (יש להתקין את המודול `scipy` אם הוא לא קיים אצלכם).
3. בסעיף זה ננסה לבדוק ולתקן את השפעת הבדלי אחוזי ההצבעה על שכיחות ההצבעה למפלגות.
א. תקנו עבור כל ישוב את מספר הקולות לכל מפלגה בהנחה שבכל ישוב אחוז ההצבעה היה זהה עבור בוחרי כל המפלגות (אבל כמובן שונה בין ישוב ליישוב), כפי שנלמד בכיתה.
חשבו עבור כל מפלגה את שכיחות הקולות הארצית שהיתה מתקבלת עבור המקרה בו היו 100 אחוזי הצבעה ועל פי תיקון זה. הציגו ב-bar-plot זה לצד זה בצבעים שונים שכיחויות אלו מול שכיחות הקולות הארצית עבור כל מפלגה בפועל.
ב. חזרו על סעיף א', אבל הפעם השתמשו בקובץ ההצבעה על פי **קלפיות** (ולא על פי ערים) שהועלה למודל לצורך התיקון בשם **votes per ballot - 2019b elections**. ניתן לקרוא קובץ זה באמצעות פקודת `read_csv` בדומה לקובץ על פי ערים אך שימו לב לפורמט הקובץ שיש בו שדות שונים מעט מאשר הקובץ על פי ערים - עליכם לנפות את העמודות הלא רלוונטיות.

הערות:

- כפי שתואר היום בכיתה, ועדת הבחירות עדכנה את התוצאות מאז שקובץ ההצבעה לפי ערים הורד מאתרה. הורידו את הקובץ המעודכן **votes per city- 2019b elections** מהמודל - מעכשיו כל ניתוחי הנתונים ישתמשו בו.
- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו' כדי להדגיש נקודות חשובות.
- שימו לב שישנו בקובץ "ישוב" בשם "מעטפות חיצוניות" המתאר את המעטפות הכפולות ואינו מתאר ישוב מסוים. יש להסיר אותו כאשר מציגים נתונים על ישובים.

- מכיוון שקשה לזכור את אותיות ההצבעה לכל מפלגה ולשייכן למפלגה, הציגו את שמות המפלגה הבאים בכל פעם שתדרשו לכך במקום באותיות: (ניתן להמיר באמצעות בניית dictionary בפייטון)

פתק	פה	מחל	ודעם	שס	ל	ג	טב	אמת	מרץ	כף
מפלגה	כחול לבן	הליכוד	הרשימה המשותפת	שס	ישראל ביתנו	יהדות התורה	ימינה	העבודה גשר	המחנה הדמוקרטי	עוצמה יהודית

- מותר לכם להיות יצירתיים; נסו לחשוב על שיטות אחרות לתיקון אחוזי ההצבעה.