Data mining and ML Project

Tamir Ovadia

השאלה העיקרית הנחקרת בפרויקט- האם היה שינוי בדפוסי ההצבעה בין בחירות 24 לבחירות 25?

בפרויקט בוצע שימוש ב-4 מערכי נתונים- נתוני בחירות ברמת יישוב וברמת קלפי, לכל אחת ממערכות הבחירות

תיאור מקדים של הנתונים -EDA

	24 לפי יישוב	**25 לפי יישוב**	**24 לפי קלפי**	**25 לפי קלפי**
גודל מערך הנתונים	1215 שורות, 46 עמודות	1216 שורות, 47 עמודות	12926 שורות, 50 עמודות	12545 שורות, 51 עמודות
תיאור כל שורה	כל שורה מייצגת את התוצאות הכוללות עבור יישוב מסוים בבחירות לכנסת ה-24.	כל שורה מייצגת את התוצאות הכוללות עבור יישוב מסוים בבחירות לכנסת ה-25.	כל שורה מייצגת את התוצאות עבור קלפי בודדת בבחירות לכנסת ה-24.	כל שורה מייצגת את התוצאות עבור קלפי בודדת בבחירות לכנסת ה-25.
תיאור כל משתנה	• סמל ועדה: מזהה ייחודי של הוועדה האחראית על ספירת הקולות ביישוב. • שם יישוב: שם היישוב שבו ממוקמת הקלפי. • סמל יישוב: מזהה ייחודי של היישוב במאגרי המידע. • בז"ב: מספר בעלי זכות הבחירה הרשומים להצביע ביישוב. • מצביעים: מספר האנשים שהצביעו בפועל ביישוב. • פסולים: מספר הקולות שהוכרזו כפסולים ביישוב. • כשרים: מספר הקולות שהוכרזו ככשרים ביישוב, כלומר אלה שנחשבו לחישוב הקולות עבור המפלגות. • שאר העמודות: מתארות את אותיות המפלגות		• סמל ועדה: מזהה ייחודי של הוועדה האחראית על ספירת הקולות ביישוב בו ממוקם הקלפי. • ברזל: מזהה ייחודי לכל קלפי ברחבי הארץ • שם יישוב: שם היישוב שבו ממוקמת הקלפי. • סמל יישוב: מזהה ייחודי של היישוב במאגרי המידע. • קלפי: מספר מזהה של הקלפי בתוך היישוב. • ריכוז: מספר המייצג אזור גאוגרפי או לוגיסטי בתוך היישוב שבו ממוקמות קבוצות של קלפיות. • שופט: אינדיקציה אם יש נוכחות של שופט בקלפי (0 = אין שופט, 1 = יש שופט). • בז"ב: מספר בעלי זכות הבחירה הרשומים להצביע בקלפי. • מצביעים: מספר האנשים שהצביעו בפועל בקלפי.	

		• פסולים: מספר הקולות שהוכרזו כפסולים בקלפי זו. • כשרים: מספר הקולות שהוכרזו ככשרים בקלפי, כלומר אלה שנחשבו לחישוב הקולות עבור המפלגות. • שאר העמודות: מתארות את אותיות המפלגות

- הערה: בתוצאות לפי יישוב- יש שורה של "מעטפות חיצוניות", שורה זו מייצגת את התוצאות של הקולות שהתקבלו באמצעות מעטפות כפולות. מדובר בקולות שנאספו מחוץ למקום ההצבעה הרשמי של המצביעים, כגון:
חיילים :המצביעים בבסיסי צה"ל.
בתי חולים :חולים ואנשי צוות שהצביעו במקום אשפוזם.
אסירים :קולות שנאספו בבתי כלא.
נציגויות ישראל בחו"ל :אזרחים ישראלים שהצביעו בנציגויות דיפלומטיות.
אנשים עם מוגבלויות :אם הוקצו עבורם קלפיות מיוחדות.

- גם בתוצאות לפי קלפיות יש שורות שהן מעטפות חיצוניות, אך פה מדובר במספר שורות ולא בשורה אחת, כי זה מחולק לפי קלפיות. יש מספר קלפיות שהוקצו למעטפות חיצוניות ולכן יש כמה שורות.

<u>נתוני בחירות 25 ברמת קלפי</u>

- גודל: 11,707 שורות, 13 עמודות.
- סוג המשתנים:

העבודה	int64
הבית היהודי	int64
יהדות התורה	int64
בלד	int64
חדש תעל	int64
הציונות הדתית	int64
המחנה הממלכתי	int64
ישראל ביתנו	int64
הליכוד	int64
מרצ	int64
הרשימה הערבית המאוחדת	int64
יש עתיד	int64
ש"ס	int64

- אין ערכים חסרים.

סטטיסטיקה תיאורית:

Pandas.describe – מציג סטטיסטיקה תיאורית.

נדגים על שתי מפלגות:

דוגמה 1: ש"ס -

Count (מספר הקלפיות): יש 11,707 קלפיות, כלומר אין קלפי חסרה- לש"ס יש ערך מוזן בכל שורה (בכל קלפי).

mean (ממוצע קולות): ממוצע הקולות לש"ס בקלפי הוא 30.89 קולות.

std (סטיית תקן): סטיית התקן היא 44.32, כלומר יש פיזור רחב יחסית לממוצע. יש קלפיות עם הרבה מאוד קולות, ויש קלפיות עם מעט מאוד קולות (ביחס לממוצע).

min (ערך מינימלי): יש קלפיות שבהן ש"ס לא קיבלה אף קול (0 קולות).

25% (רבעון ראשון): ב-25% מהקלפיות, ש"ס קיבלה 3 קולות או פחות.

50% (חציון): ב-50% מהקלפיות (החציון), ש"ס קיבלה 15 קולות או פחות.

75% (רבעון שלישי): ב-25% העליונים של הקלפיות, ש"ס קיבלה 39 קולות או יותר.

מהאחוזונים ניתן להבין כי ההתפלגות על ההצבעות לש"ס בקלפיות היא התפלגות עם זנב ימני (החציון נמוך מאוד מהממוצע). משמע- ברוב הקלפיות בארץ ש"ס מקבלת מספר קטן של קולות. אך יש מספר קטן יחסית של קלפיות בהן לש"ס יש תמיכה גדולה מאוד. רוב ההצבעות מרוכזות במספר קטן של קלפיות.

max (ערך מקסימלי): הקלפי עם התמיכה הגבוהה ביותר נתנה לש"ס 417 קולות.

דוגמה 2: יש עתיד -

Count (מספר הקלפיות): יש 11,707 קלפיות, כלומר אין קלפי חסרה- ל"יש עתיד" יש ערך מוזן בכל שורה (בכל קלפי).

mean (ממוצע קולות): ממוצע הקולות ליש עתיד בקלפי הוא 65.75 קולות.

std (סטיית תקן): סטיית התקן היא 61.14, כלומר יש פיזור רחב יחסית לממוצע. יש קלפיות עם מעט קולות, ויש כאלה עם הרבה קולות (ביחס לממוצע).

min (ערך מינימלי): יש קלפיות שבהן יש עתיד לא קיבלה אף קול (0 קולות).

25% (רבעון ראשון): ב-25% מהקלפיות, יש עתיד קיבלה 12 קולות או פחות.

50% (חציון): ב-50% מהקלפיות, קיבלה 49 קולות או פחות.

75% (רבעון שלישי): ב-25% העליונים של הקלפיות, קיבלה 108 קולות או יותר.

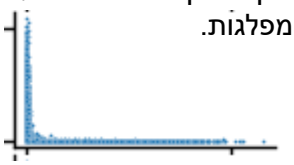
גם כאן ההתפלגות על ההצבעות ל"יש עתיד" היא עם זנב ימני- החציון קטן מהממוצע (אותו הסבר כמו לגבי ש"ס).

max (ערך מקסימלי): הקלפי עם התמיכה הגבוהה ביותר נתנה ליש עתיד 313 קולות.

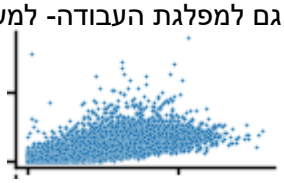
Seaborn.pairplot – הנקודות בגרף מייצגות את פיזור ההצבעות בקלפיות, בהתייחס לשתי מפלגות שונות (הצירים) בכל גרף.

*ההנחה היא שבכל קלפי יש דפוס שמייצג את אוכלוסיית המצביעים באזור שבו הקלפי ממוקמת. דפוס זה עשוי לשקף מאפיינים דמוגרפיים, חברתיים וגיאוגרפיים של האזור. לדוגמה, בקלפיות בירושלים, שבהן מתגוררת אוכלוסייה דתית וחרדית רבה, ניתן לצפות לתמיכה גבוהה במפלגות כמו יהדות התורה וש"ס. לעומת זאת, בקלפיות בתל אביב, המתאפיינות באוכלוסייה חילונית או מרכז-שמאל, ניתן לצפות לתמיכה משמעותית במפלגות כמו יש עתיד או העבודה. מכאן נובע השוני או הקורלציה בין הצבעות למפלגות שונות בקלפיות שונות.

דוגמה 1: בגרף המציג את הפיזור בהצבעות בקלפיות למפלגת יהדות התורה ולמפלגת בלד, ניתן לראות כי הפיזור מתפרס ממש על גבי שני הצירים. המשמעות היא שבקלפיות שבהן יש הצבעות גבוהות לבלד, יש מעט מאוד הצבעות ליהדות התורה, ולהפך- בקלפיות בהן יש הצבעות גבוהות ליהדות התורה, יש מספר נמוך מאוד של הצבעות לבלד. ניתן להבין שהקשר הזה הגיוני, מכיוון שאופיין הפוליטי של שתי המפלגות הוא שונה באופן קיצוני. בקלפיות שיש הצבעות גבוהות ליהדות התורה, ניתן להסיק שהן ממוקמות בערים/ שכונות של חרדים. כמובן שניתן להבין כי חרדים כמעט ואינם מצביעים למפלגות הערביות והדבר תומך בגרף הפיזור בין שתי המפלגות.



דוגמה 2: בגרף המציג את הפיזור בהצבעות בקלפיות למפלגת העבודה ולמפלגת יש עתיד, ניתן לראות כי יש קורולציה מסויימת בין ההצבעות בקלפיות למפלגות אלו. ניתן לראות שבקלפיות בהן יש הצבעות גבוהות למפלגת יש עתיד, יש גם הצבעות גבוהות יחסית למפלגת העבודה, ובקלפיות שיש הצבעות נמוכות ליש עתיד, יש גם הצבעות נמוכות למפלגת העבודה. ניתן לראות כי ישנה מגמת עלייה מסויימת מבחינת הפיזור. דבר זה עולה בקנה אחד עם אופיין הפוליטי של המפלגות- המפלגות הללו די דומות ברעיונות ובדברים שהם מייצגים, ולכן ניתן להבין שבקלפיות בהן יש מצביעים רבים ליש עתיד, יהיו גם מצביעים רבים למפלגת העבודה- לדוגמא בקלפיות בעיר תל אביב כנראה. לעומת זאת, בקלפיות שאין הרבה מצביעים ליש עתיד, אין הרבה מצביעים גם למפלגת העבודה- למשל בקלפיות בערים ערביות כנראה, או אפילו קלפיות בערים "ימניות".



matplotlib.pyplot.matshow(pandas.corr) - מטריצת הקורלציות מציגה את הקשרים (קורלציות) בין הקולות שניתנו לכל זוג מפלגות. הערכים נעים בין:

1 (אדום כהה): קורלציה חיובית חזקה – כאשר התמיכה במפלגה אחת עולה, גם השנייה עולה.

1 - (כחול כהה): קורלציה שלילית חזקה – כאשר התמיכה במפלגה אחת עולה, השנייה יורדת.

<u>הערה:</u> אצלנו בגרף הערך הכי נמוך שהוצג בגרף הצבעים הינו 0.4- כיוון שלא הייתה קורלציה נמוכה מזאת.

דוגמה 1: קורלציה חיובית – העבודה ויש עתיד:

התא בין "העבודה" ל"יש עתיד" מראה צבע אדום כהה יחסית עם ערך קרוב ל-1.

המשמעות: יש מתאם חזק מאוד בין הקולות שקיבלו שתי המפלגות. בקלפיות שבהן יש תמיכה גבוהה ב"יש עתיד", יש גם תמיכה גבוהה ב"עבודה". זה הגיוני, שכן שתי המפלגות מייצגות את אותו קהל יעד, ומאפיינות את אותן דעות בגדול.

דוגמה 2: קורלציה שלילית – בל"ד והליכוד:

התא בין "בל"ד" ל"הליכוד" מראה צבע כחול כהה עם ערך קרוב ל- 0.4-.

המשמעות: יש מתאם שלילי בין הקולות של שתי המפלגות. בקלפיות שבהן יש תמיכה גבוהה ב"בל"ד" (מפלגה ערבית), התמיכה ב"הליכוד" (מפלגה ימנית) נמוכה מאוד, ולהפך. זה מובן, כי שתי המפלגות פונות לקהלים שונים מאוד – אוכלוסייה ערבית מול אוכלוסיה ישראלית ימנית.

<u>נתוני בחירות 24 ברמת קלפי:</u>

- גודל: 12,127 שורות, 13 עמודות.
- סוג המשתנים:

העבודה	int64
הבית היהודי	int64
יהדות התורה	int64
הרשימה המשותפת	int64
הציונות הדתית	int64
המחנה הממלכתי	int64
ישראל ביתנו	int64
הליכוד	int64
מרצ	int64
הרשימה הערבית המאוחדת	int64
יש עתיד	int64
ש"ס	int64
תקווה חדשה	int64

- אין ערכים חסרים.

<u>סטטיסטיקה תיאורית</u>

Pandas.describe – מציג סטטיסטיקה תיאורית.

נדגים על שתי מפלגות:

דוגמה 1- ש"ס

Count (מספר הקלפיות): יש 12,127 קלפיות, כלומר אין קלפי חסרה- לש"ס יש ערך מוזן בכל שורה (בכל קלפי).

mean (ממוצע קולות): ממוצע הקולות לש"ס בקלפי הוא 24.08 קולות.

std (סטיית תקן): סטיית התקן היא 35.6, כלומר יש פיזור רחב יחסית לממוצע. יש קלפיות ספציפיות עם הרבה מאוד קולות, ויש קלפיות עם מעט מאוד קולות (ביחס לממוצע).

min (ערך מינימלי): יש קלפיות שבהן ש"ס לא קיבלה אף קול (0 קולות).

25% (רבעון ראשון): ב-25% מהקלפיות, ש"ס קיבלה 2 קולות או פחות.

50% (חציון): ב-50% מהקלפיות (החציון), ש"ס קיבלה 12 קולות או פחות.

75% (רבעון שלישי): ב-25% העליונים של הקלפיות, ש"ס קיבלה 30 קולות או יותר.

מהאחוזונים ניתן להבין כי ההתפלגות על ההצבעות לש"ס בקלפיות היא התפלגות עם זנב ימני (החציון נמוך מאוד מהממוצע). משמע- ברוב הקלפיות בארץ ש"ס מקבלת מספר קטן של קולות. אך יש מספר קטן יחסית של קלפיות בהן לש"ס יש תמיכה גדולה מאוד. רוב ההצבעות מרוכזות במספר קטן של קלפיות.

max (ערך מקסימלי): הקלפי עם התמיכה הגבוהה ביותר נתנה לש"ס 381 קולות.

דוגמה 2 – "יש עתיד"

Count (מספר הקלפיות): יש 12,127 קלפיות, כלומר אין קלפי חסרה- ל"יש עתיד" יש ערך מוזן בכל שורה (בכל קלפי).

mean (ממוצע קולות): ממוצע הקולות ליש עתיד בקלפי הוא 45.85 קולות.

std (סטיית תקן): סטיית התקן היא 42.21, כלומר יש פיזור רחב יחסית לממוצע. יש קלפיות עם מעט קולות, ויש כאלה עם הרבה קולות (ביחס לממוצע).

min (ערך מינימלי): יש קלפיות שבהן יש עתיד לא קיבלה אף קול (0 קולות).

25% (רבעון ראשון): ב-25% מהקלפיות, יש עתיד קיבלה 9 קולות או פחות.

50% (חציון): ב-50% מהקלפיות, קיבלה 35 קולות או פחות.

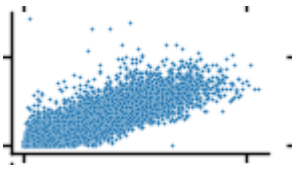
75% (רבעון שלישי): ב-25% העליונים של הקלפיות, קיבלה 75 קולות או יותר.

גם כאן ההתפלגות על ההצבעות ל"יש עתיד" היא עם זנב ימני- החציון קטן מהממוצע (אותו הסבר כמו לגבי ש"ס).

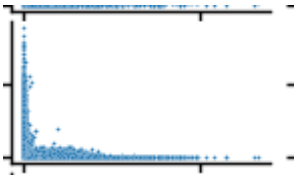
max (ערך מקסימלי): הקלפי עם התמיכה הגבוהה ביותר נתנה ליש עתיד 210 קולות.

Seaborn.pairplot – הנקודות בגרף מייצגות את פיזור ההצבעות בקלפיות, בהתייחס לשתי מפלגות שונות (הצירים) בכל גרף.

דוגמה 1: בגרף המציג את הפיזור בהצבעות בקלפיות למפלגת "יש עתיד" ולמפלגת "המחנה הממלכתי", ניתן לראות כי יש קורלציה גבוהה בין ההצבעות בקלפיות למפלגות אלו. בקלפיות שיש הרבה מצביעים ל"יש עתיד", יש גם הרבה מצביעים ל"מחנה הממלכתי", ולהפך. הדבר עולה בקנה אחד עם אופיין של המפלגות, שכן הוא מאוד דומה והם מייצגים את אותה אוכלוסייה, ועל כן ניתן לראות מגמת עליה- ככל שמספר ההצבעות ליש עתיד גדול, כך מספר ההצבעות למחנה הממלכתי גדול.



דוגמה 2: בגרף המציג את הפיזור בהצבעות בקלפיות למפלגת "הבית היהודי" ולמפלגת "הרשימה הערבית המאוחדת", ניתן לראות כי הפיזור מתפרס ממש על גבי שני הצירים. המשמעות היא שבקלפיות שבהן יש הצבעות גבוהות לבית היהודי, יש מעט מאוד הצבעות לרשימה הערבית המאוחדת, ולהפך. ניתן להבין שהקשר הזה הגיוני, מכיוון שאופיין הפוליטי של שתי המפלגות הוא שונה באופן קיצוני. בקלפיות שיש הצבעות גבוהות לבית היהודי, ניתן להסיק שהן ממוקמות בערים/ שכונות של דתיים לאומים/ ימנים. כמובן שניתן להבין כי אוכלוסייה זו כמעט ואינם מצביעים למפלגות הערביות והדבר תומך בגרף הפיזור בין שתי המפלגות.



- matplotlib.pyplot.matshow(pandas.corr)

דוגמה 1: קורלציה חיובית – המחנה הממלכתי ויש עתיד:

התא בין "המחנה הממלכתי" ל"יש עתיד" מראה צבע אדום כהה יחסית עם ערך קרוב ל-1.

המשמעות: יש מתאם חזק מאוד בין הקולות שקיבלו שתי המפלגות. בקלפיות שבהן יש תמיכה גבוהה ב"יש עתיד", יש גם תמיכה גבוהה ב"מחנה הממלכתי". זה הגיוני, שכן שתי המפלגות מייצגות את אותו קהל יעד, ומאפיינות את אותן דעות בגדול.

דוגמה 2: קורלציה שלילית – הרשימה המשותפת והליכוד:

התא בין "הרשימה המשותפת" ל"הליכוד" מראה צבע כחול כהה עם ערך קרוב ל- 0.4-.

המשמעות: יש מתאם שלילי בין הקולות של שתי המפלגות. בקלפיות שבהן יש תמיכה גבוהה ב"רשימה המשותפת" (מפלגה ערבית), התמיכה ב"הליכוד" (מפלגה ימנית) נמוכה מאוד, ולהפך. זה מובן, כי שתי המפלגות פונות לקהלים שונים מאוד – אוכלוסייה ערבית מול אוכלוסיה ישראלית ימנית.

נקודות שוני בין שני מסדי הנתונים:

- בבחירות לכנסת ה-25 היו 11,707 קלפיות, לעומת 12,127 קלפיות בבחירות לכנסת ה-24, מה שמשקף צמצום של כ-3.5% במספר הקלפיות בין מערכות הבחירות. ייתכן שהשינוי נובע מהתייעלות לוגיסטית, שינויי אוכלוסייה, או שינוי במדיניות הקמת הקלפיות.
- ממוצע הקולות לחלק מהמפלגות השתנה באופן ניכר בין מערכות הבחירות. לדוגמה, ממוצע הקולות ליש עתיד בבחירות לכנסת ה-25 נמוך משמעותית לעומת הבחירות לכנסת ה-24, מה שמעיד על ירידה בתמיכה הציבורית במפלגה.
- בבחירות לכנסת ה-25 נוספו מפלגות חדשות שלא היו בכנסת ה-24, בעוד שמפלגות אחרות התפרקו או נעלמו מהמפה הפוליטית. לדוגמה, 'תקווה חדשה' שהתמודדה בבחירות לכנסת ה-24 התמזגה עם 'המחנה הממלכתי', והרשימה הערבית המשותפת התפרקה לשתי מפלגות: בל"ד וחדש-תע"ל.

נקודות דמיון בין שני מסדי הנתונים:

- בשני מערכי הנתונים, מספר המפלגות שקיבלו יותר מ-1% מהקולות נשאר זהה – 13 מפלגות. נתון זה משקף יציבות בכמות המפלגות הגדולות בשתי מערכות הבחירות.
- בשני המערכים ניתן לזהות קורלציות חיוביות חזקות בין מפלגות שמייצגות אותו קהל יעד (למשל, ש"ס ויהדות התורה) וקורלציות שליליות בין מפלגות שפונות לקהלים שונים מאוד (למשל, בל"ד/ הרשימה המשותפת והליכוד).
- בשני מערכי הנתונים, התפלגות הקולות כמעט לכל המפלגות היא עם זנב ימני, כאשר רוב הקלפיות נותנות מעט קולות למפלגה מסוימת, אך יש קלפיות בודדות עם תמיכה גבוהה. רוב הקולות לכל מפלגה מתקבלים מקצת קלפיות.

ארגון הנתונים

לצורך זיהוי דמיון בין מפלגות לפי דפוסי ההצבעה, עדיף להפוך את המפלגות לשורות. בצורה זו, ה-PCA יתמקד במפלגות ובקשרים ביניהן, מה שמספק מענה ישיר לשאלה.

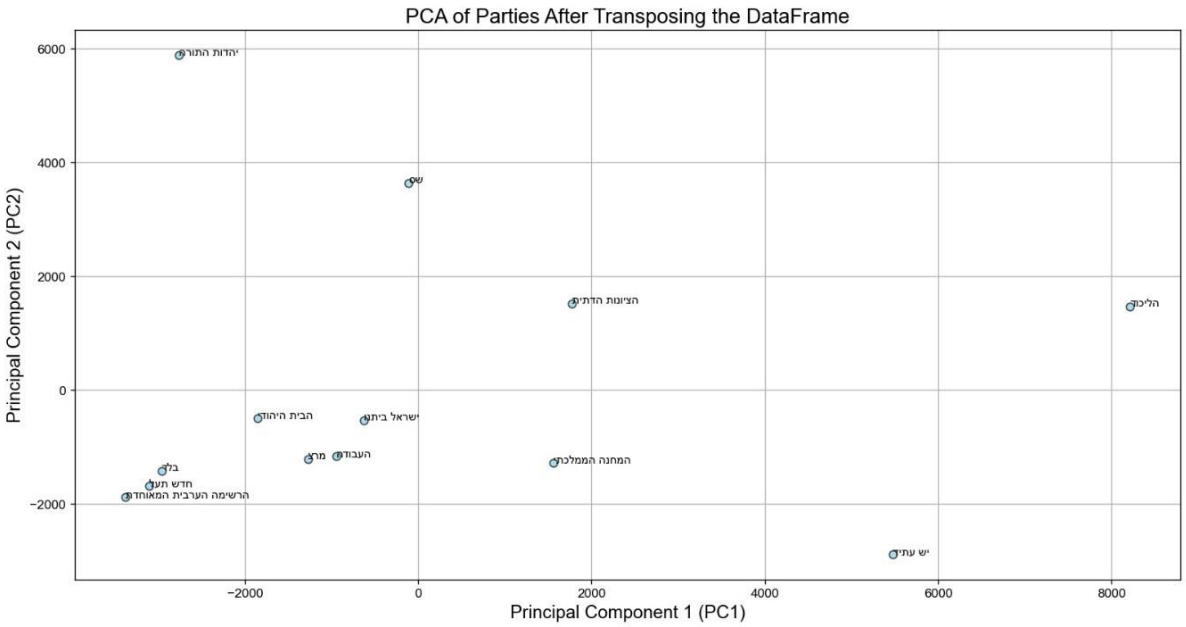
במידה ונשאיר את המבנה המקורי של מערך הנתונים כל קלפי היה הופך להיות נקודה בגרף והגרף היה מראה את הדמיון בין הקלפיות. כלומר, קלפיות קרובות בגרף – דפוסי הצבעה דומים (לדוגמה שכונות חרדיות יופיעו קרובות) וקלפיות רחוקות – דפוסי הצבעה שונים (לדוגמה קלפי ביישוב יהודי לעומת יישוב ערבי). במצב זה לא נזהה דמיון בין המפלגות כפי שנתון בשאלה.

כאשר נהפוך את המפלגות לשורות, כל מפלגה הופכת להיות נקודה בגרף. במצב זה הגרף מראה דמיון בין המפלגות. מפלגות קרובות בגרף – מפלגות עם דפוסי הצבעה דומים (מפלגות בעלות אופי פוליטי דומה) ומפלגות רחוקות – דפוסי הצבעה שונים (אופי פוליטי שונה).

כאשר המפלגות הן שורות, הצירים יתארו את הגורמים המרכזיים שמסבירים את ההבדלים בין המפלגות (למשל, אידיאולוגיה, גאוגרפיה, או מגזרים).

ניתוח דפוסי הצבעה עבור בחירות 25 – לפי מסד נתונים ברמת הקלפי

בוצע PCA- הורדת מימד ל-2 מימדים, והתוצאות הוצגו בגרף:



פרשנות לצירים לאחר PCA

Principal Component 1 (PC1) - הציר האופקי:

ציר זה מתאר את הפיזור של המפלגות שעברו את אחוז החסימה ברחבי הארץ.

מפלגות בצד הימני - לדוגמה, הליכוד: מייצגות מפלגות גדולות ובעלות בסיס תמיכה רחב יותר, כנראה עם פיזור גאוגרפי רחב. מפלגות בצד השמאלי - לדוגמה, חדש-תע"ל ובל"ד: מייצגות מפלגות קטנות יותר שממוקדות בקהלים מסוימים מאוד, כמו מגזרים ייחודיים.

Principal Component 2 (PC2) - הציר האנכי:

ציר זה מייצג את ההבדל בין מפלגות על בסיס מאפיינים אידיאולוגיים או דמוגרפיים, במיוחד בהקשר של דתיות מול חילוניות.

מפלגות למעלה - לדוגמה, ש"ס ויהדות התורה: מייצגות קהלים חרדיים או דתיים.

מפלגות למטה - לדוגמה, יש עתיד: מייצגות קהלים חילוניים יותר.

כעת, נבחן שתי גישות. נרמול הנתונים בצורת min- max ותקנון Z. נבצע PCA לאחר יישום כל גישה ונחקור את התוצאה המתקבלת.

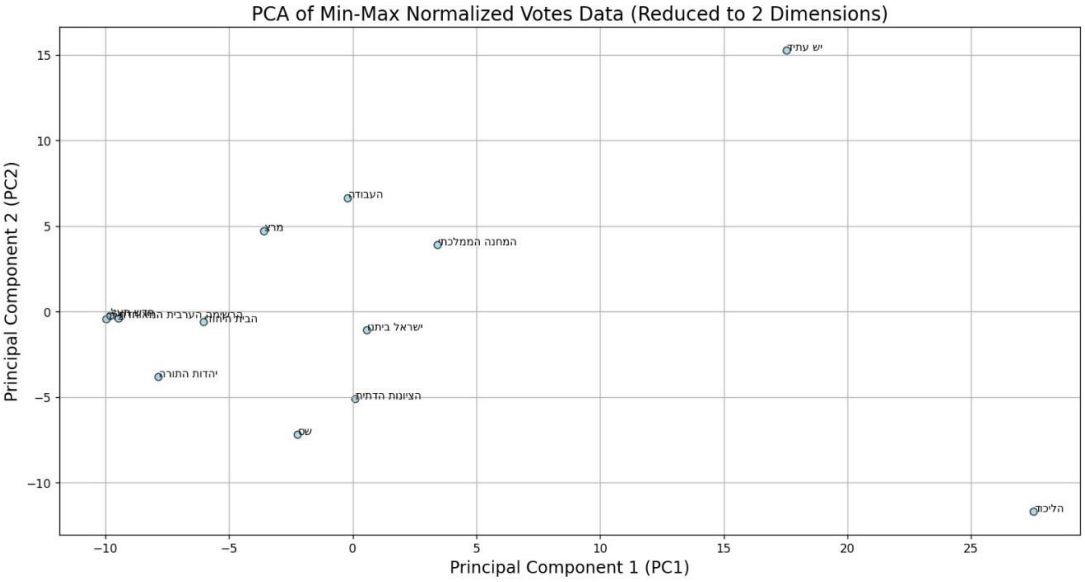
<u>ביצוע נרמול min-max לנתונים</u>

המשמעות של נרמול במערך הנתונים ממיר את הערכים של כל משתנה לטווח שבין 0 ל-1. הערך המינימלי בכל משתנה הופך ל-0, והערך המקסימלי הופך ל-1, וכל הערכים האחרים מחושבים באופן יחסי בין המינימום למקסימום. במערך הנתונים הנוכחי, לפני הנרמול הערכים המקוריים משתנים בקנה מידה שונה. במצב זה, התוצאה תהיה שטווחי הערכים שונים מאוד ונוצר קושי להשוות בין המשתנים (המפלגות).

המשמעות של הנרמול במערך הנתונים הנוכחי מביאה לכך שהערכים של כל מפלגה נעים בין 0-1, כך שהערכים מתארים יחסיות: ערך 0 – הקלפי שבו המפלגה קיבלה את מספר הקולות המינימלי. ערך 1 – הקלפי שבו המפלגה קיבלה את מספר הקולות המקסימלי. ערכים בין 0-1 – המיקום היחסי של הקלפי עבור אותה מפלגה.

הנרמול מאפשר להשוות את דפוסי ההצבעה לכל מפלגה בקלפיות השונות, בלי תלות בגודל המפלגה. בנוסף, PCA רגיש לערכים גדולים מאוד ולכן הנרמול עוזר למנוע מצב שבו משתנה אחד משפיע באופן מוגזם על התוצאה.

<u>ביצוע PCA מחדש ל-2 מימדים לאחר נרמול הנתונים (נרמול min- max)</u>



<u>פרשנות לצירים לאחר הצגת הגרף על הנתונים המנורמלים (min-max) ולאחר PCA</u>

הפרשנות לצירים מהסעיף הקודם אכן תקפה חלקית.

Principal Component 1 (PC1) - הציר האופקי:

ציר זה מייצג את ההבדל המרכזי בין המפלגות על בסיס גודל הקהל ויחסיות התמיכה.

מפלגות בצד הימני - לדוגמה, הליכוד: מייצגות מפלגות גדולות ובעלות בסיס תמיכה רחב יותר, כנראה עם פיזור גאוגרפי רחב. מפלגות בצד השמאלי - לדוגמה, חדש-תע"ל ובל"ד: מייצגות מפלגות קטנות יותר שממוקדות בקהלים מסוימים מאוד, כמו מגזרים ייחודיים.

Principal Component 2 (PC2) - הציר האנכי:

ציר זה מייצג את ההבדל בין מפלגות על בסיס מאפיינים אידיאולוגיים או דמוגרפיים, במיוחד בהקשר של דתיות מול חילוניות. ההבדל מהסעיף הקודם הוא שהכיוון של הציר PC2 השתנה מכיוון שהוא שרירותי ב-PCA, כתוצאה משינוי במבנה הנתונים ונרמול. המשמעות נשארת זהה – רק הכיוון השתנה.

מפלגות למטה - לדוגמה, ש"ס ויהדות התורה: מייצגות קהלים חרדיים או דתיים.
מפלגות למעלה - לדוגמה, יש עתיד: מייצגות קהלים חילוניים יותר.

<u>ביצוע תקנון Z לנתונים</u>

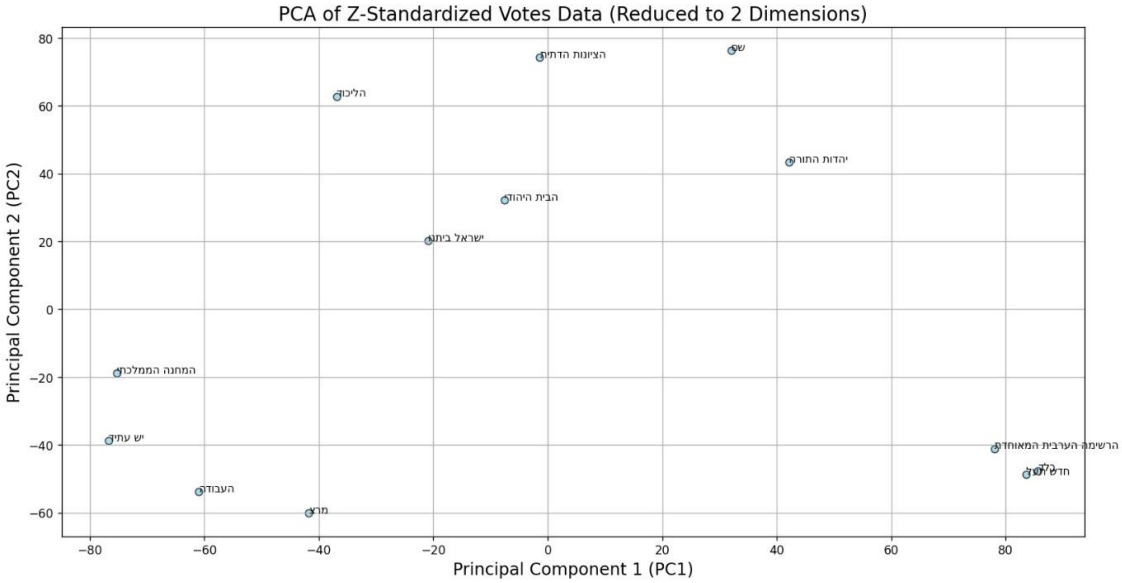
אחרי התקנון, ממוצע כל משתנה (עמודה) הופך ל-0. המשמעות: הערכים במשתנה מתואמים סביב הממוצע, כך שהנתונים אינם תלויים בערך המוחלט של הממוצע המקורי. אחרי התקנון, סטיית התקן של כל משתנה (עמודה) היא 1. המשמעות: הנתונים מנורמלים כך שאין השפעה של טווח הערכים המקורי (לדוגמה, אם מפלגה מסוימת קיבלה אלפי קולות ומפלגה אחרת קיבלה מאות).

לפני התקנון: מפלגות כמו "הליכוד" או "ש"ס" עשויים להיות בעלי טווחים שונים מאוד של קולות, דבר שיכול לגרום לכך שה-PCA יושפע בעיקר מהמשתנים עם הטווח הגדול ביותר. אחרי התקנון: כל המפלגות מיוצגות באופן יחסי עם ממוצע של 0 וסטיית תקן של 1: מפלגה גדולה כמו הליכוד ומפלגה

קטנה יותר כמו מרצ משפיעות באותה מידה על ה-PCA. הדפוסים היחסיים בין קלפיות שונים הופכים לחשובים יותר מהערכים האבסולוטיים.

הנרמול מדגיש את היחסיות בין מפלגות בתוך כל קלפי: מפלגות שמקבלות תמיכה יחסית דומה בקלפי מסוימת יהיו קרובות אחת לשנייה. מפלגות עם דפוסי הצבעה שונים יהיו רחוקות יותר זו מזו.

<u>ביצוע PCA מחדש ל-2 מימדים לאחר תקנון הנתונים (תקנון Z)</u>



<u>פרשנות לצירים לאחר הצגת הגרף על הנתונים המתוקננים (Z) ולאחר PCA</u>

לאחר תקנון Z, הנתונים עוברים טרנספורמציה כך שלכל משתנה (עמודה) יש ממוצע של 0 וסטיית תקן של 1. המשמעות היא שהערכים בכל משתנה מבוטאים ביחידות של סטיות תקן מהממוצע, מה שמדגיש את החריגות בדפוסי ההצבעה.
הפרשנות לצירים מהסעיף הקודם אכן תקפה חלקית.

Principal Component 1 (PC1) - הציר האופקי:

ציר זה מייצג את ההבדל בליברליות/שמרנות בין המפלגות.

מפלגות בצד הימני - לדוגמה, חד"ש תע"ל: מייצגות מפלגות שמרניות ולא ליברליות. מפלגות בצד השמאלי - לדוגמה, המחנה הממלכתי ויש עתיד: מייצגות מפלגות ליברליות.

Principal Component 2 (PC2) - הציר האנכי:

ציר זה מייצג את ההבדל בין מפלגות על בסיס מאפיינים אידיאולוגיים, במיוחד בהקשר של דעות ימין ושמאל.

מפלגות למעלה - לדוגמה, הליכוד וש"ס: מייצגות קהלים ימניים.

מפלגות למטה - לדוגמה, מרצ והעבודה: מייצגות קהלים שמאלניים יותר.

בחירת שיטה עדיפה

בחירה מועדפת - נרמול Z:

משמעות המשתנה בקלט: נרמול Z מבטל הטיות הנובעות מגודל נתונים אבסולוטי, מדגיש את דפוסי החריגה מהממוצע ומציג את היחסים בין מפלגות בצורה ברורה.

כמות השונות המוסברת: השונות מתפלגת בין הצירים בצורה יותר פרופורציונלית, מה שמאפשר לזהות דמיון בין מפלגות על בסיס דפוסי ההצבעה בצורה מדויקת.

פרשנות המשתנה (ציר): הצירים מבטאים קשרים יחסיים ודפוסי שונות בין המפלגות על בסיס מגמות כלליות, כמו חלוקה לפי מגזרים או אזורים.

נרמול Z עדיף כיוון שהוא מאפשר להסביר דמיון בין מפלגות בצורה רחבה ומדויקת יותר, תוך ביטול השפעת סדרי גודל מוחלטים.

בשיטה הראשונה ללא נרמול הערכים האבסולוטיים של מספרי הקולות משפיעים ישירו על ניתוח הPCA, כלומר, מפלגות גדולות כמו "הליכוד" או מקבלות דגש רב יותר בהשוואה למפלגות קטנות יותר כמו "מרצ". התוצאה מושפעת באופן משמעותי מהטווחים האבסולוטיים של הנתונים.

נרמול Min-Max: מתאים לערכים בטווח 0-1, אך מאבד מידע על שונות משמעותית בין קלפיות ומושפע באופן ישיר מערכים קיצוניים. לעומת זאת נרמול Z מדגיש חריגות יחסית ומפחית את השפעת הערכים הקיצוניים. במערכת הבחירות הישראלית קיימים פערי גודל בין המפלגות כאשר קיימות מפלגות גדולות ומפלגות קטנות. נרמול Z במצב זה עדיף כיוון שהוא מדגיש דפוסים ייחודים וחריגות בעוד ש Min-Max מאזן את הדפוסים באופן שעלול לטשטש מידע.

בנוסף, בנרמול Min-Max ערכים קיצוניים קובעים את גבולות הטווח, מה שמוביל לעיוות ההשוואה. נרמול Z מפחית את השפעתם דרך השימוש בסטיית תקן, מה שמאפשר זיהוי מדויק יותר של שונות.

לסיכום, נרמול Z עדיף מהסיבות הבאות:

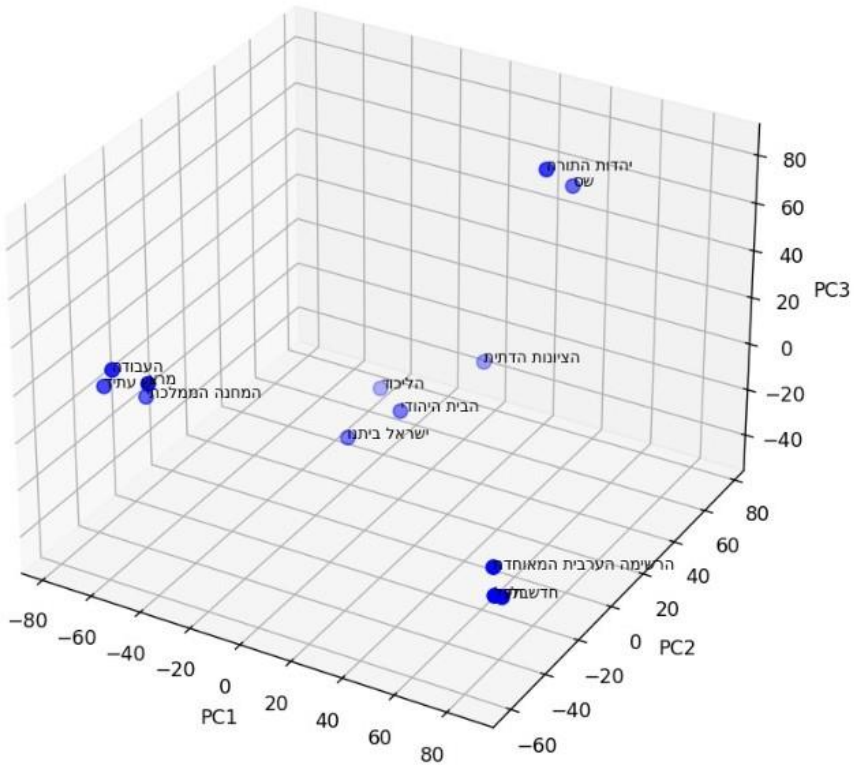
מדגיש את שונות החריגות בדפוסי ההצבעה בין קלפיות.

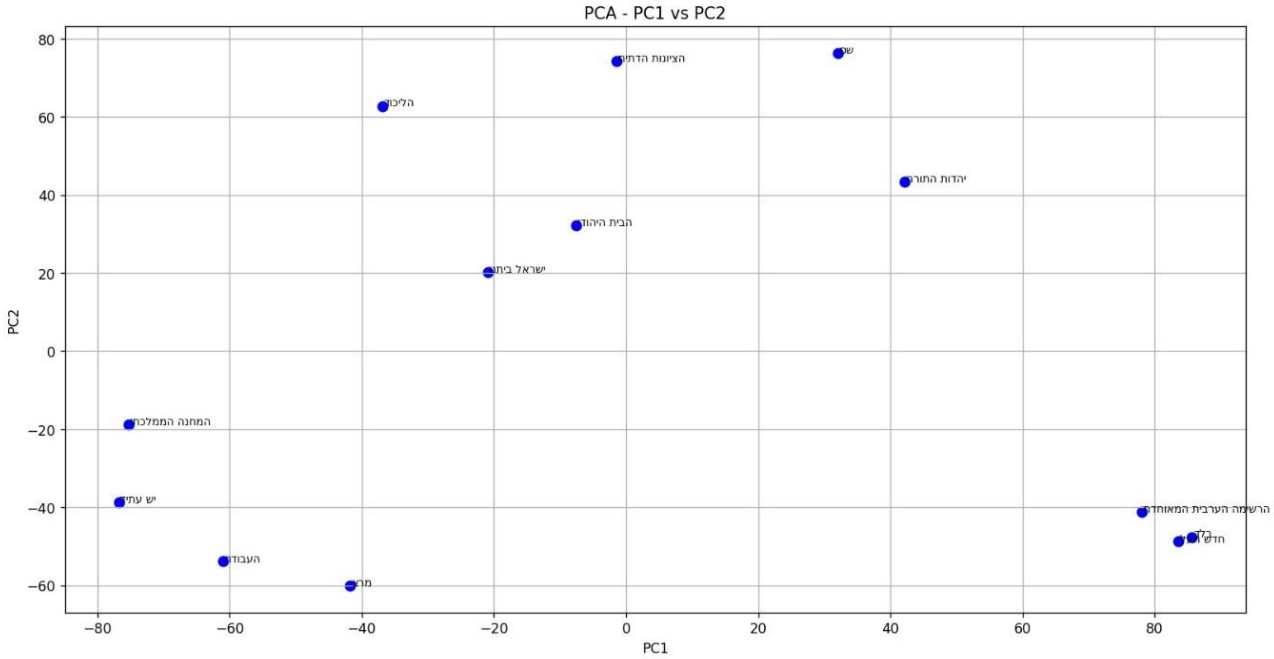
מייצג טוב יותר קשרים מהותיים ולא רק יחסים מוחלטים.

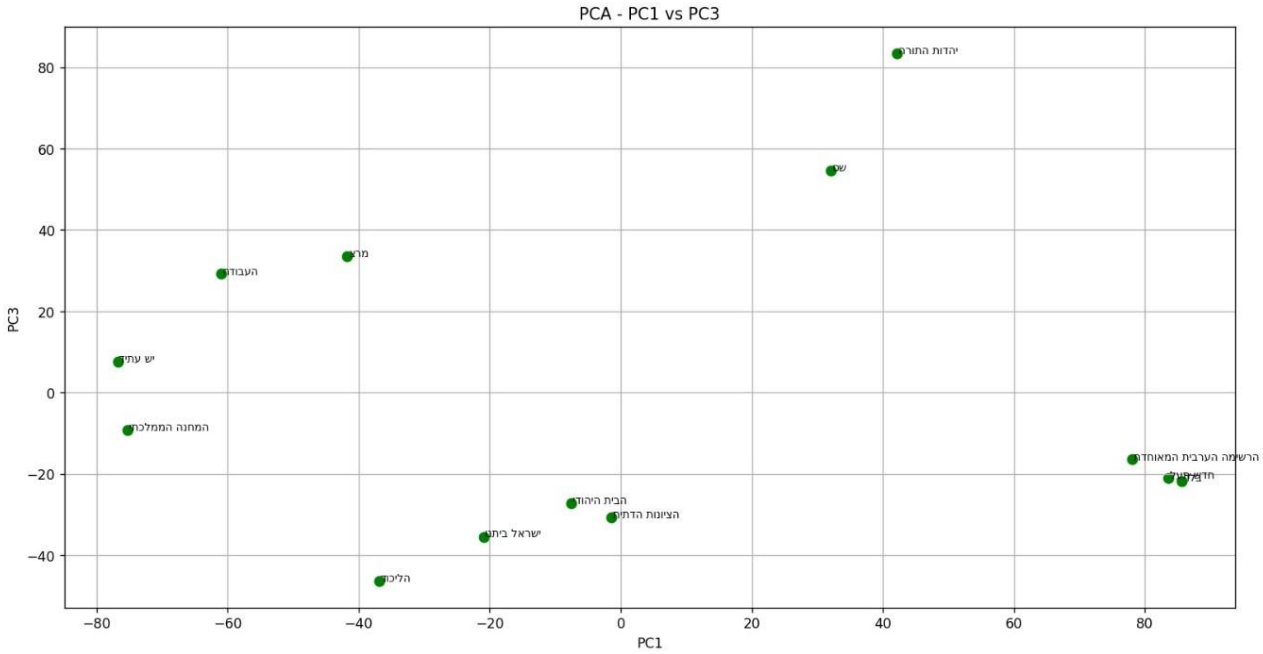
מתמודד טוב יותר עם ערכים קיצוניים, מה שעוזר לזהות דפוסים משמעותיים במערכת נתונים מורכבת כמו הצבעות בקלפיות.

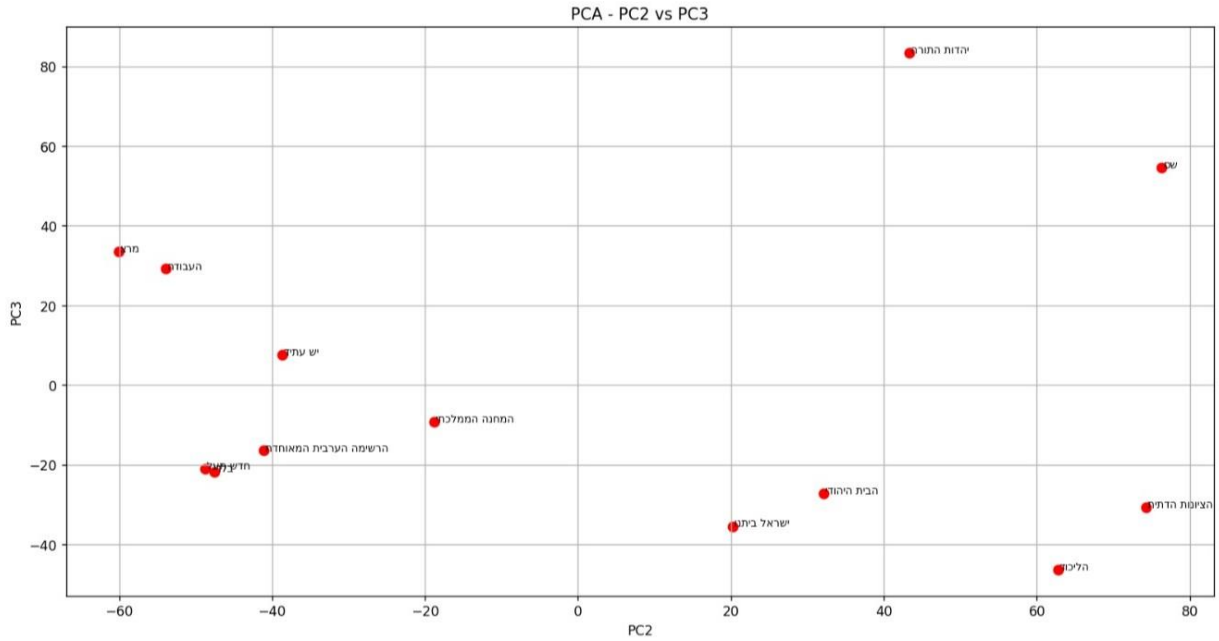
לאחר בחירת תקנון Z, וביצוע PCA ל- 3 מימדים – נתוני כנסת 25 ברמת קלפי

PCA - 3D Representation









<u>פרשנות לצירים לאחר הצגת הגרפים – תקנון Z עם PCA - 3</u>

פרשנות לצירים:
Principal Component 1 (PC1)- ציר זה מייצג את ההבדל בליברליות/שמרנות בין המפלגות.

Principal Component 2 (PC2)- ציר זה מייצג את ההבדל בין מפלגות על בסיס מאפיינים אידיאולוגיים, במיוחד בהקשר של דעות ימין ושמאל.

Principal Component 3 (PC3)- לדעתנו ציר זה מייצג אחוזי הצבעה בקלפיות ספציפיות. לדוגמא- ביישובים חרדים, אחוז ההצבעה למפלגות החרדיות הוא גבוה מאוד- כמעט כולם מצביעים, ולכן יהדות התורה וש"ס לדוגמא ממוקמות גבוה בציר זה. הליכוד לדוגמא, היא מפלגה שמצביעים אליה

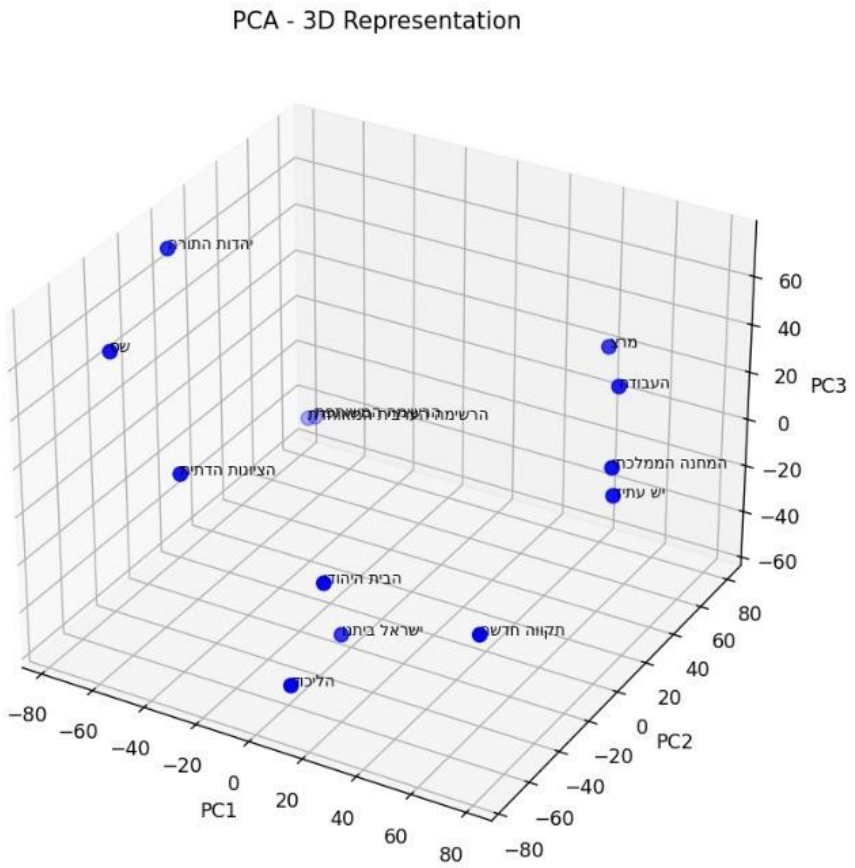
ממגוון רחב של קלפיות ברחבי הארץ, ואין קלפיות ספציפיות שבה היא מקבל אחוז משמעותי מהקולות. הפיזור של הקולות שלה רחב יותר.

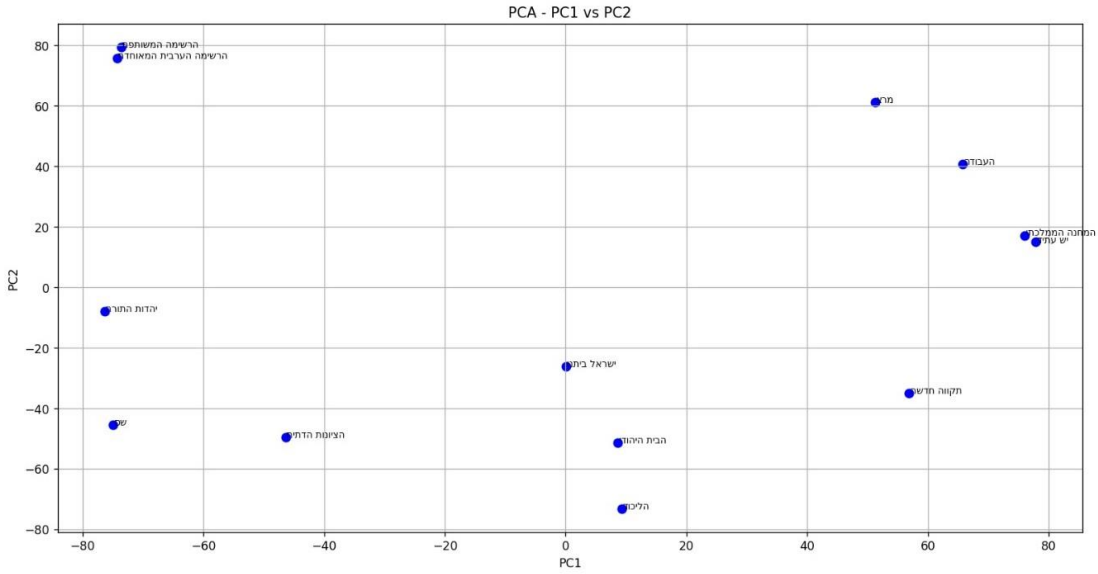
<u>זיהוי מפלגות עם דפוסי הצבעה דומים</u>

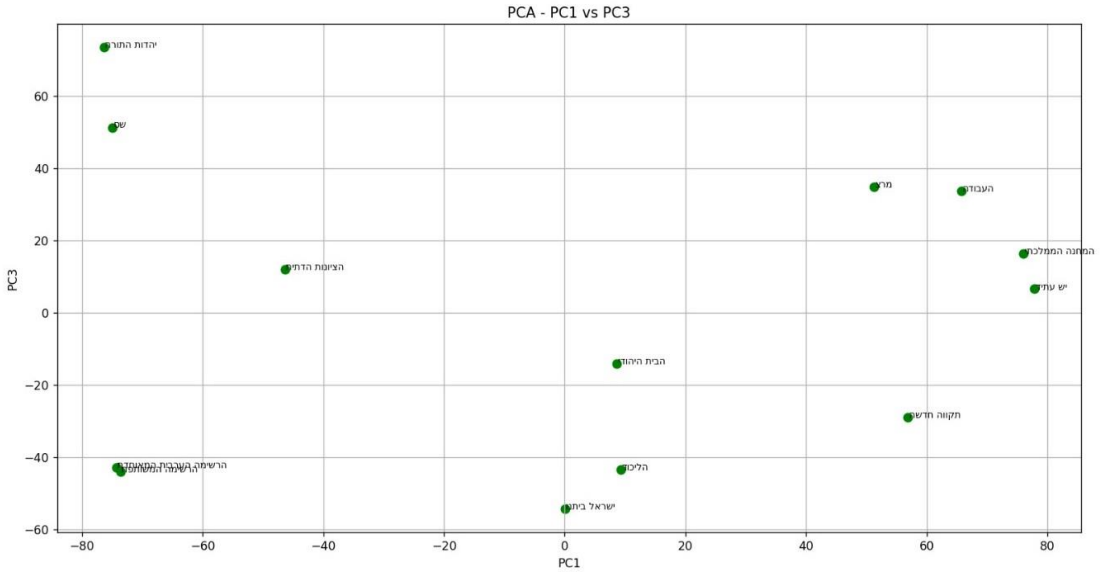
לפי התוצאות ניתן להבחין בכמה קבוצות של מפלגות. הקבוצות: מפלגות חילוניות, דתיות לאומיות, חרדיות וערביות.

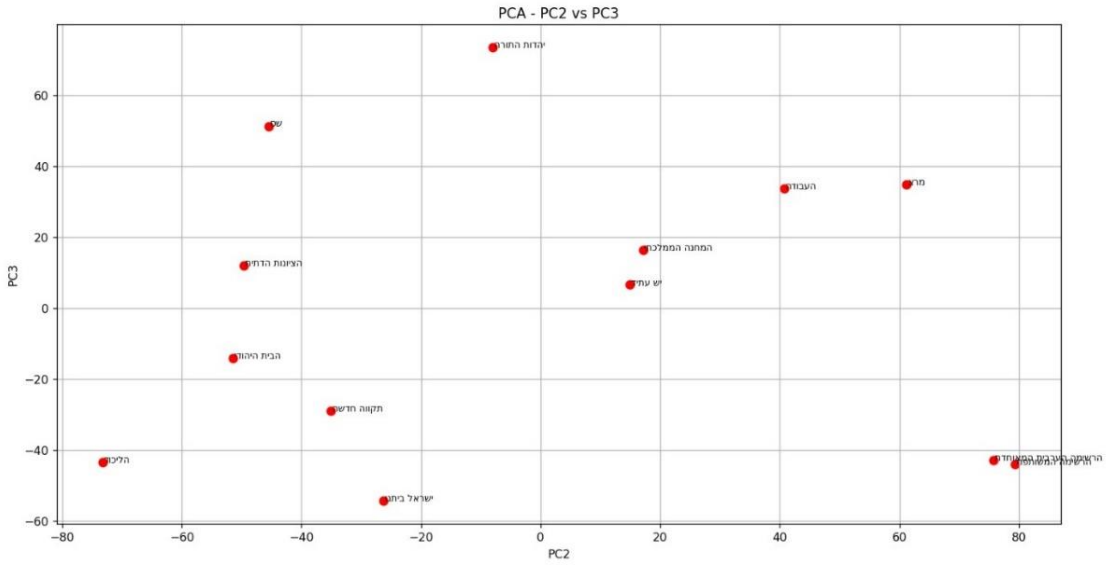
ניתן לראות בתרשים זיהוי לקבוצות הנ"ל אשר המפלגות מכל קבוצה מקובצות באותו האזור בגרף. לדוגמה: המפלגות הערביות (חד"ש תע"ל, הרשימה הערבית המאוחדת) מקבוצות יחדיו, וכך גם החרדיות (יהדות התורה וש"ס).

<u>לאחר בחירת תקנון Z, וביצוע PCA ל- 3 מימדים – נתוני כנסת 24 ברמת קלפי</u>




PCA - PC1 vs PC2
PC1
PC2
הרשימה המשותפת
הרשימה הערבית המאוחדת
מרצ
העבודה
המחנה הממלכתי
יש עתיד
יהדות התורה
ישראל ביתנו
תקווה חדשה
שס
הציונות הדתית
הבית היהודי
הליכוד


PCA - PC1 vs PC3
PC1
PC3
יהדות התורה
שס
מרצ
העבודה
המחנה הממלכתי
הציונות הדתית
יש עתיד
הבית היהודי
תקווה חדשה
הרשימה הערבית המאוחדת
הרשימה המשותפת
הליכוד
ישראל ביתנו


PCA - PC2 vs PC3
PC2
PC3
יהדות התורה
שס
העבודה
מרצ
המחנה הממלכתי
הציונות הדתית
יש עתיד
הבית היהודי
תקווה חדשה
הליכוד
הרשימה הערבית המאוחדת
הרשימה המשותפת
ישראל ביתנו

פרשנות לצירים לאחר הצגת הגרפים – תקנון Z עם PCA - 3

פרשנות לצירים:
(PC1) Principal Component 1- ציר זה מייצג את ההבדל בליברליות/שמרנות בין המפלגות.

(PC2) Principal Component 2- ציר זה מייצג את ההבדל בין מפלגות על בסיס מאפיינים אידיאולוגיים, במיוחד בהקשר של דעות ימין ושמאל.

(PC3) Principal Component 3- לדעתנו ציר זה מייצג אחוזי הצבעה בקלפיות ספציפיות. לדוגמא- ביישובים חרדים, אחוז ההצבעה למפלגות החרדיות הוא גבוה מאוד- כמעט כולם מצביעים, ולכן יהדות התורה וש"ס לדוגמא ממוקמות גבוה בציר זה. הליכוד לדוגמא, היא מפלגה שמצביעים אליה ממגוון רחב של קלפיות ברחבי הארץ, ואין קלפיות ספציפיות שבה היא מקבל אחוז משמעותי מהקולות. הפיזור של הקולות שלה רחב יותר.

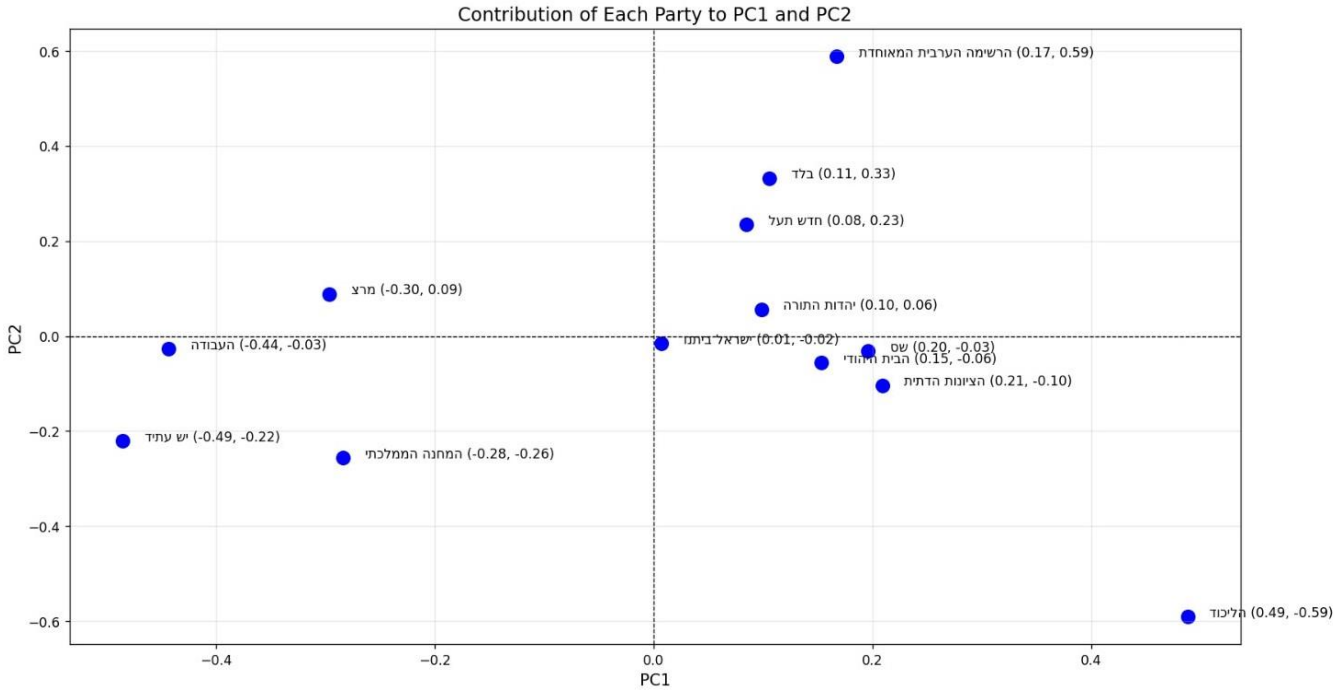
השוואה בין שתי מערכות הבחירות

ניתן לראות כי התוצאות דומות, מכיוון שמדובר כמעט באותן מפלגות, והן בעלות אותה אידיאולוגיה בשתי מערכות הבחירות, לכן אין שינוי משמעותי בנתונים ובפרשנות הצירים.

מבחינת הנתונים- כל המפלגות נמצאות בשתי מערכות הבחירות (למעט תקווה חדשה והמפלגות הערביות שהתפצלו), ולכן אין שינוי בפלט ובפרשנות שלנו.

כעת, נשתמש בשני מסדי הנתונים של מערכות הבחירות ברמת היישוב. ביצענו מניפולציה על הנתונים (ניתן לראות בקוד) כדי להתאים את המפלגות בשתי מערכות הבחירות.

ביצוע תקנון Z לנתונים מערכת בחירות 24 ברמת היישוב, ו- PCA ל-2 מימדים



<u>פרשנות לצירים</u>

הגרף מציג את התרומה של כל מפלגה לשני הצירים הראשיים (PC1 ו-PC2) שהתקבלו מניתוח PCA.

PC1 - ציר ה-X:

הליכוד תורמת חיובית משמעותית לציר זה (0.49), מה שעשוי להצביע על כך שהיא מפלגה בעלת מאפיינים ייחודיים מבחינת דפוסי ההצבעה, מייצגת את הימין מבחינה פוליטית.

יש עתיד, מרצ והעבודה תורמות שלילית לציר זה (-0.49 ו- -0.44 בהתאמה), והן מפלגות שמאל מרכז, הנמצאות בניגוד לליכוד.

לכן לדעתנו PC1 עשוי לייצג חלוקה בין ימין לשמאל מבחינה פוליטית.

PC2 - ציר ה-Y:

הרשימה הערבית המאוחדת תורמת חיובית משמעותית לציר זה (0.59), מה שעשוי להעיד על מאפיינים ייחודיים של מצביעיה.

שס ויהדות התורה תורמות חיובית מתונה, מה שיכול להעיד על דפוסי הצבעה דומים בקרב מצביעים דתיים או חרדיים.

הליכוד תורמת שלילית לציר זה (-0.59), בעוד מפלגות כמו יש עתיד והמחנה הממלכתי קרובות גם הן לתרומה שלילית.

מסקנה: PC2 עשוי לייצג חלוקה בין מפלגות המייצגות מגזרים שונים, למשל בין מפלגות המייצגות מצביעים דתיים/חרדים או ערבים לעומת מפלגות המייצגות מצביעים חילוניים או כלליים יותר (לא אוכלוסיה ספציפית).

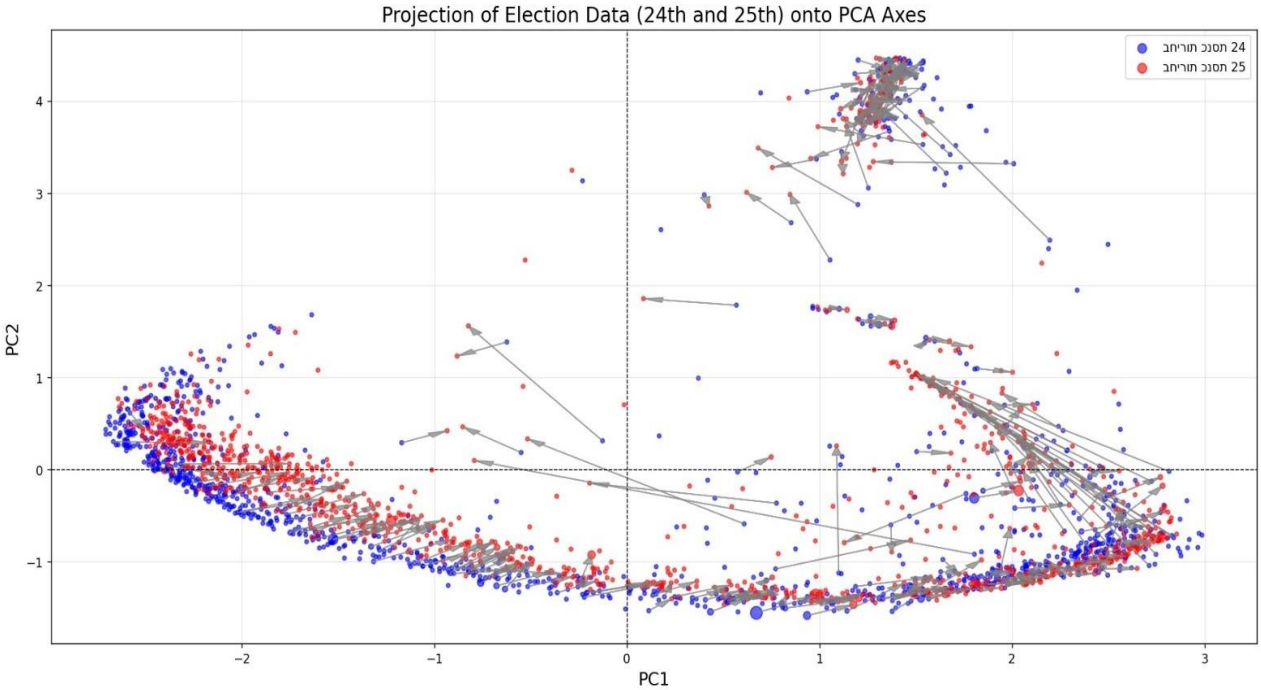
כעת, נטיל את התוצאות המתוקננות של שתי מערכות הבחירות, על מערכת הצירים החדשה שהתקבלה.

הסבר: במידה ונעשה מערכת צירים נפרדת לכל מערך נתונים נקבל "מתכון" שונה לכל ציר בכל PCA שנעשה. כלומר, לכל מערך נתונים תהיה מערכת צירים שונה ולא ניתן יהיה להשוות ישירות בין נתוני הבחירות.

במערכת צירים משותפת ניתן לאחד בין נתוני הבחירות כך שמערכת הצירים מתחשבת גם בנתוני הכנסת ה-24 וגם בנתוני הכנסת ה-25. כל מערך נתונים מיוצג באותה מערכת צירים.

המטרה כאן היא לבדוק האם יש שינוי בדפוסי ההצבעה או דמיון בין היישובים בבחירות 25 ביחס לבחירות 24, ולכן השיטה המתאימה היא לעשות PCA לפי נתוני 24, ולהציב את הנתונים של שתי הטבלאות במערכת הצירים הזו.

הצגת התוצאות



פרשנות לצירים וסיכום

כפי שציינו בסעיפים הקודמים: PC1 עשוי לייצג חלוקה בין ימין לשמאל מבחינה פוליטית ו - PC2 עשוי לייצג חלוקה בין מפלגות המייצגות מגזרים שונים, למשל בין מפלגות המייצגות מצביעים דתיים/חרדים או ערבים לעומת מפלגות המייצגות מצביעים חילוניים או כלליים.

מהגרף ניתן לראות כי דפוסי ההצבעה ב-2 מערכות הבחירות היו **יחסית דומים**.

ניתן לראות שפיזור הנקודות על ציר PC1 נשאר דומה בשתי מערכות הבחירות, מה שמראה על כך שמספר הישובים שהצביעו למפלגות ימין/שמאל יחסית נשמר. בנוסף, ניתן לראות שאין הרבה חצים ארוכים לאורך ציר זה, מה שמעיד על כך שישובים שהצביעו למפלגות ימין, המשיכו להצביע למפלגות ימין, וישובים שהצביעו למפלגות שמאל המשיכו להצביע למפלגות שמאל. מסקנה- קיימת יציבות בחלוקה האידאולוגית של המצביעים.
בציר PC2 גם פיזור היישובים נראה די דומה בשתי מערכות הבחירות, ועל כן ניתן לראות שישובים שהצביעו למפלגות מגזריות (מפלגות המייצגות מגזר ספציפי) במערכת בחירות 24, המשיכו להצביע למפלגות מגזריות במערכת בחירות 25. ניתן לראות שמספר החצים הארוכים לאורך ציר זה הוא קטן, ולכן נראה שלא היה שינוי מהותי בדפוסי ההצבעה למפלגות המגזריות.