למידה חישובית – תרגיל 4

מיכל אברמוב, 301834297
בוריס בורשבסקי, 311898746

מה מצורף

1. מחברת המפרטת בשלבים את מה שעשינו (clustering.ipynb)
2. קובץ clustering.py המכיל את כל הקוד של התרגיל (בעצם המחברת בצורה ניתנת להרצה מהקונסול)
3. קובץ clustering.html שמאפשר צפיה נוחה יותר במחברת
4. קבצי הdata.
5. כל הקבצים מהתרגילים הקודמים

התרגיל

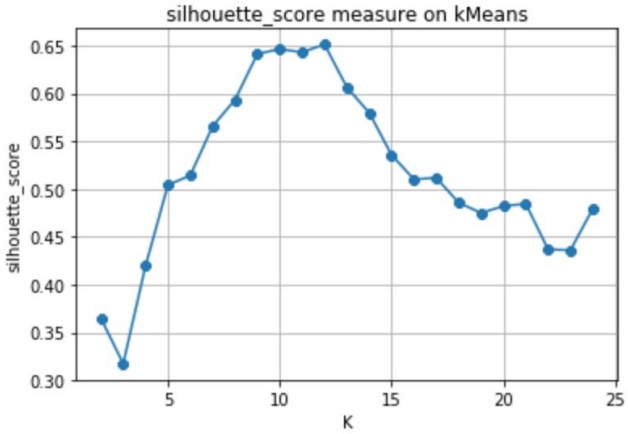
התרגיל מחולק 2 שלבים עיקריים:

1. מציאת clusters במידע ע״מ להרכיב קואליציה
    a. בחירת k אופטימלי
    b. ניתוח תוצאות קלאסטרים
    c. חיבור בין קלאסטרים
2. פיצ׳רים עיקריים
    a. הצגת פיצ׳רים עיקריים בכל מפלגה ושינוי התוצאות
    b. הצגת פיצ׳רים עיקריים לקואליציה ושינוי התוצאות

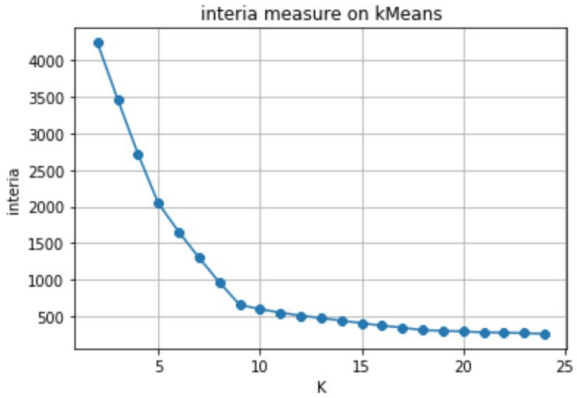
מציאת cluster-ים

K- Means

בשלב הראשון נסתכל על אלגוריתם k-means, ומבחר את הk האופטימלי מבחינותינו:
על מנת למצוא את הk האופטימאלי אימננו את k means על הדאטה שלנו עם k = 2 - 25, וניסינו לראות מה הk האופטימלי לפי מספר מדדים:
את הבדיקה עדינו בעזרת k-fold כדי למנוע overfittings
הסתכלנו שלפי מדד **silhouette** כדאי לנו לבחור k בין 9 ל 12



ראינו לפי מדד ה **inertia** שמעל k = 9 המרחקים ממרכזי הקלאסטרים יורדים הרבה פחות עבור כל קלאסטר נוסף כל שזה מצביע שזה מדד טוב.



הסתכלנו גם על מדדים אחרים ולבסוף הלכנו על **k = 10**

init	time	inertia	homo	compl	v-meas	ARI	AMI	calinski	silhouette
k-means k=2	0.04s	4247	0.273	1.000	0.429	0.193	0.273	2081.288	0.359
k-means k=3	0.07s	3468	0.273	0.593	0.374	0.173	0.273	1853.051	0.322
k-means k=4	0.06s	2710	0.274	0.432	0.335	0.151	0.273	2061.600	0.403
k-means k=5	0.07s	2045	0.274	0.353	0.308	0.127	0.273	2468.488	0.481
k-means k=6	0.07s	1656	0.274	0.312	0.292	0.111	0.272	2680.566	0.521
k-means k=7	0.07s	1300	0.302	0.318	0.310	0.116	0.300	3080.732	0.563
k-means k=8	0.10s	964	0.355	0.354	0.354	0.136	0.352	3816.999	0.603
k-means k=9	0.08s	655	0.375	0.359	0.367	0.133	0.357	5213.223	0.643
k-means k=10	0.10s	593	0.375	0.354	0.364	0.128	0.352	5177.127	0.638
k-means k=11	0.12s	548	0.375	0.351	0.362	0.125	0.348	5088.091	0.645
k-means k=12	0.11s	505	0.375	0.348	0.361	0.122	0.345	5051.127	0.655
k-means k=13	0.15s	470	0.413	0.365	0.388	0.135	0.362	5003.500	0.614
k-means k=14	0.14s	437	0.453	0.383	0.416	0.149	0.380	5005.534	0.575
k-means k=15	0.17s	402	0.490	0.396	0.438	0.165	0.393	5082.629	0.538
k-means k=16	0.17s	368	0.530	0.412	0.464	0.181	0.409	5201.828	0.513
k-means k=17	0.19s	341	0.531	0.410	0.462	0.182	0.406	5296.809	0.522
k-means k=18	0.23s	307	0.571	0.424	0.487	0.200	0.421	5565.701	0.471
k-means k=19	0.22s	296	0.574	0.418	0.483	0.197	0.414	5462.105	0.466
k-means k=20	0.25s	286	0.578	0.414	0.482	0.194	0.410	5365.094	0.473
k-means k=21	0.27s	279	0.610	0.434	0.508	0.207	0.431	5234.159	0.467
k-means k=22	0.30s	271	0.618	0.429	0.506	0.205	0.425	5133.281	0.450
k-means k=23	0.30s	268	0.636	0.440	0.520	0.211	0.436	4956.739	0.452
k-means k=24	0.31s	255	0.644	0.432	0.517	0.209	0.428	4984.663	0.465
k-means k=25	0.33s	249	0.656	0.439	0.526	0.215	0.435	4891.708	0.444

ניתוח התוצאות

בשלב זה הרצנו k means עם k = 10 על הדאטה שלנו ותחלנו לבדוק את התוצאות
קודם כל נסתכל על החלוקה של הlabels בין הקלאסטרים:

```
Group: 0, Distribution: ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Group: 1, Distribution: ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Group: 2, Distribution: ['Greys', 'Oranges']
Group: 3, Distribution: ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Group: 4, Distribution: ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Group: 5, Distribution: ['Oranges', 'Reds']
Group: 6, Distribution: ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Group: 7, Distribution: ['Blues', 'Yellows']
Group: 8, Distribution: ['Greys', 'Reds']
Group: 9, Distribution: ['Blues', 'Yellows']
```

אפשר לנחש לפי החלוקה הזו שיש לנו קבוצות של מפלגות שהמצביעים שלהם די דומים ולכן הולכים איתם לאותו קלאסטר, בגלל ש kmeans יוצר קלאסטרים על בסיס מרחק אוקלידי, יכול להיות מצב שפיצ׳רים דיסקרטיים משפעים על החלוקה הזו וכדאי לנרמל אותם, נבדוק את זה בהמשך.

כמו אפשר לראות ש:

- חומים, ורודים, סגולים, ירוקים ולבנים נוטים להיות ביחד
- כחולים וצהובים נוטים להיות ביחד
- אצל הכתומים, האדומים ואפורים ישנו ערבוב של זוגות ביניהם, יכול להיות שאפשר לחבר אותם יחד. נבדוק בהמשך

עכשיו נסתכל על החלוקה הזו מהצד השני:

```
Group: Reds, Distribution: [5, 8]
Group: Greens, Distribution: [0, 1, 3, 4, 6]
Group: Whites, Distribution: [0, 1, 3, 4, 6]
Group: Yellows, Distribution: [7, 9]
Group: Greys, Distribution: [2, 8]
Group: Oranges, Distribution: [2, 5]
Group: Browns, Distribution: [0, 1, 3, 4, 6]
Group: Blues, Distribution: [7, 9]
Group: Pinks, Distribution: [0, 1, 3, 4, 6]
Group: Purples, Distribution: [0, 1, 3, 4, 6]
```

פה אנו רואים את אותם מסקנות.
כעת נסתכל על ההתפלגות הפנימית לראות שאין לנו מקרים בהם מצביע אחד במקרה הלך לקלאסטר אחר ושיבש לנו את הנתונים.

```
_____________________dist per cluster_____________________
__________________________________________________________
0 Counter({'Purples': 231, 'Browns': 220, 'Greens': 199, 'Pinks': 97, 'Whites': 34})
1 Counter({'Purples': 253, 'Browns': 220, 'Greens': 210, 'Pinks': 90, 'Whites': 40})
2 Counter({'Greys': 177, 'Oranges': 176})
3 Counter({'Purples': 240, 'Browns': 223, 'Greens': 191, 'Pinks': 102, 'Whites': 36})
4 Counter({'Purples': 242, 'Browns': 212, 'Greens': 167, 'Pinks': 119, 'Whites': 44})
5 Counter({'Reds': 161, 'Oranges': 142})
6 Counter({'Purples': 231, 'Browns': 206, 'Greens': 185, 'Pinks': 94, 'Whites': 35})
7 Counter({'Yellows': 190, 'Blues': 18})
8 Counter({'Greys': 156, 'Reds': 155})
9 Counter({'Yellows': 58, 'Blues': 10})

_____________________ dist per party _____________________
__________________________________________________________
Reds Counter({5: 161, 8: 155})
Greens Counter({1: 210, 0: 199, 3: 191, 6: 185, 4: 167})
Whites Counter({4: 44, 1: 40, 3: 36, 6: 35, 0: 34})
Yellows Counter({7: 190, 9: 58})
Greys Counter({2: 177, 8: 156})
Oranges Counter({2: 176, 5: 142})
Browns Counter({3: 223, 0: 220, 1: 220, 4: 212, 6: 206})
Blues Counter({7: 18, 9: 10})
Pinks Counter({4: 119, 3: 102, 0: 97, 6: 94, 1: 90})
Purples Counter({1: 253, 4: 242, 3: 240, 0: 231, 6: 231})
```

נראה שמרבית המצביעים מתפזרים בצורה די אחידה בין הקלאסטרים בהם יש מצביעים באותו סוג, פרט לצהובים אבל הם מפלגה קטנה והם לא יפריעו לנו.

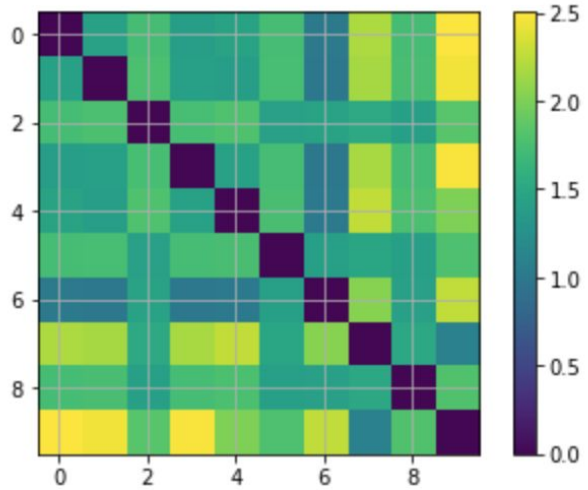
חיבור בין קלאסטרים

בצורה הנוכחית לא ניתן עדיין להרכיב קואליציה:
מה שננסה לעשות זה למצוא קלאסטרים דומים ונחבר ביניהם, הפרמטרים שלנו לחיבור בין קלאסטרים הם:

1. מרחק אוקלידי בין מרכזי הקלאסטרים
2. התפלגות מצביעים דומה בין המפלגות באותו קלאסטר

ההנחה היא שאם המרחק האוקלידי הוא קרוב, אזי המצבעים בין שני הקלאסטרים יחסית דומים וזה יאפשר לנו ליצור קואליציה יציבה:

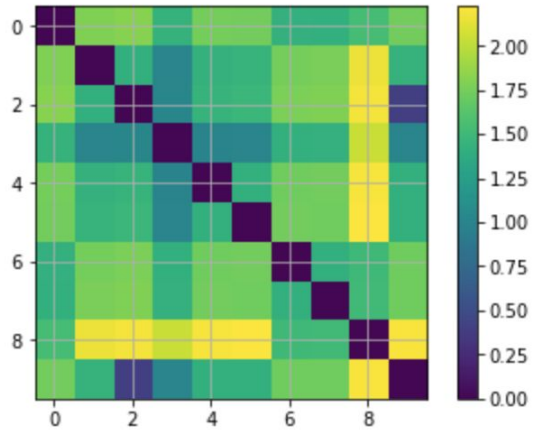
קודם כל מדפיס את המרחקים בין המרכזים:



או במספרים:

```
[[ 0.     1.44   1.738  1.417  1.458  1.737  1.018  2.192  1.736  2.502]]
[[ 1.44   0.     1.774  1.425  1.417  1.744  1.003  2.168  1.756  2.462]]
[[ 1.738  1.774  0.     1.746  1.792  1.427  1.447  1.533  1.418  1.836]]
[[ 1.417  1.425  1.746  0.     1.438  1.734  1.005  2.174  1.738  2.485]]
[[ 1.458  1.417  1.792  1.438  0.     1.754  1.014  2.274  1.77   2.006]]
[[ 1.737  1.744  1.427  1.734  1.754  0.     1.421  1.493  1.418  1.783]]
[[ 1.018  1.003  1.447  1.005  1.014  1.421  0.     2.046  1.431  2.26 ]]
[[ 2.192  2.168  1.533  2.174  2.274  1.493  2.046  0.     1.509  1.1  ]]
[[ 1.736  1.756  1.418  1.738  1.77   1.418  1.431  1.509  0.     1.805]]
[[ 2.502  2.462  1.836  2.485  2.006  1.783  2.26   1.1    1.805  0.   ]]
```

בשלב זה הסתכלנו גם על gmm כמודל לקלאסטרים:




```
[[ 0.     1.792  1.818  1.447  1.746  1.738  1.427  1.418  1.541  1.739]
 [ 1.792  0.     1.419  1.014  1.438  1.458  1.754  1.77   2.16   1.447]
 [ 1.818  1.419  0.     1.03   1.455  1.479  1.772  1.792  2.194  0.403]
 [ 1.447  1.014  1.03   0.     1.005  1.018  1.421  1.431  2.046  1.013]
 [ 1.746  1.438  1.455  1.005  0.     1.417  1.734  1.738  2.204  1.417]
 [ 1.738  1.458  1.479  1.018  1.417  0.     1.737  1.736  2.223  1.418]
 [ 1.427  1.754  1.772  1.421  1.734  1.737  0.     1.418  1.496  1.733]
 [ 1.418  1.77   1.792  1.431  1.738  1.736  1.418  0.     1.515  1.733]
 [ 1.541  2.16   2.194  2.046  2.204  2.223  1.496  1.515  0.     2.214]
 [ 1.739  1.447  0.403  1.013  1.417  1.418  1.733  1.733  2.214  0.   ]]
Mean distance: 1.418360
```

הסתכלנו על המרחקים בין מרכזי הקלאסטרים בgmm וראינו שממוצע המרחקים שם הוא 1.41, שזה פחות מ kmean, זה אומר שהקלאסטרים פחות מופרדים אחד מהשני והמשכנו עם k means.

עכשיו ננסה למצוא זוגות שיחסית קרובים ביניהם וננסה לחבר ביניהם:
בחרנו בחציון במרחקים כמדד שממנו נתחיל והדפסנו זוגות של קלאסטרים קרובים:

```
1 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
6 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Clusters: 1 --> 6, distance: 1.003
_____________________________________________________
3 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
6 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Clusters: 3 --> 6, distance: 1.005
_____________________________________________________
4 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
6 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Clusters: 4 --> 6, distance: 1.014
_____________________________________________________
0 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
6 ['Browns', 'Greens', 'Pinks', 'Purples', 'Whites']
Clusters: 0 --> 6, distance: 1.018
_____________________________________________________
7 ['Blues', 'Yellows']
9 ['Blues', 'Yellows']
Clusters: 7 --> 9, distance: 1.100
```

אפשר לראות לפי המדד הזה שלקלאסטרים 1,6,3,4,0 הם מאוד קרובים והמצביעים שלהם מצביעים לאותה קבוצה של מפלגות מכאן ניתן לחבר ביניהם.
ניתן לחבר בין עוד קבוצות אך אין טעם לפרט פה (אפשר לראות במחברת)

בעקבות החיבור הזה ננסה להציע את הקואליציה שכוללת את:

- **Browns**
- **Greens**
- **Pinks**
- **Purples**
- **Whites**

נבדוק שאכן מדובר במעל 50%

```
In Coualtion the are 3921 votes which are 75.93% percent
```

אכן קיבלנו קואליציה גדולה יחסית.

כעט נוודא את מידע מול הטסט:
לאחר חיבור הקלאסטרים הגענו למסקנה שהקלאסטרים הבאים יהיה בקואליציה

```
coalition_clusters = [0,1,3,4,6]
```

כעת נריץ predict על הטסט ונראה האם גם שם נקבל קואליציה kmeans, גם פה ראינו שהקואליציה נשמרת

```
Training...
Pridicting...
predicted winner is party ## Purples ##
In Coualtion the are 1315 votes which are 77.08% percent
Clustering - In Coualtion the are 1315 votes which are 77.08% percent
```

Random forest

KMeans

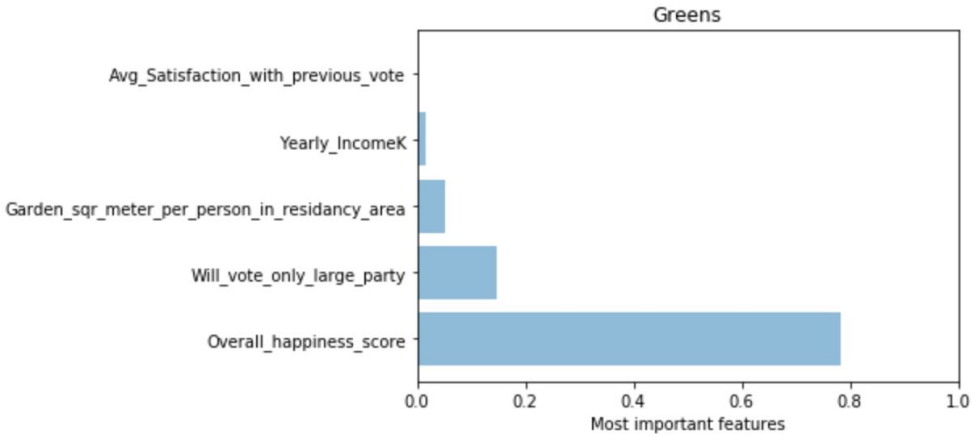
מסקנות

- הקואליציה שלנו מורכב מחומים, ירוקים, ורודים סגולים ולבנים.
- הכחולים והצהובים מאוד דומים אחד לשני ולכן אפשר לחבר אותם יחד.
- אדומים אפורים וכתומים גם מאוד דומים ביניהם אך שונים מהשאר.

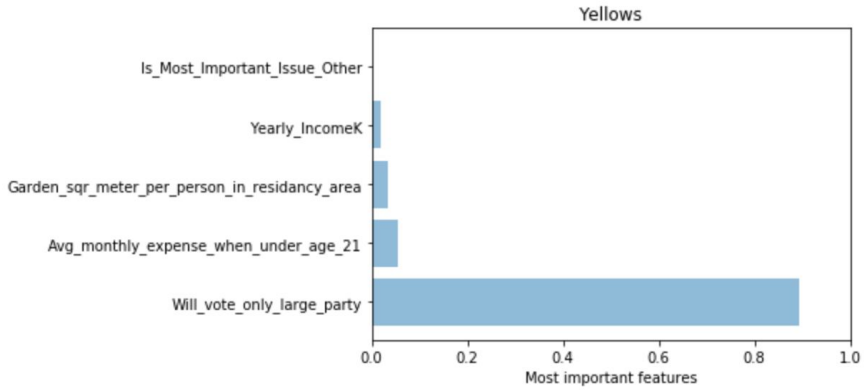
פיצ׳רים משמעותיים

ע״מ למצוא פיצ׳רים משמעותיים לכל מפלגה השתמשנו ב **feature_importances_** של RandomForestClassifier כאשר אימנו אותו לכל מפלגה בנפרד כך שעבור כל מפלגה יצרנו עמודה בוליאנית של האם בחרו בה או לא, והיא שימשה כ label שלנו.

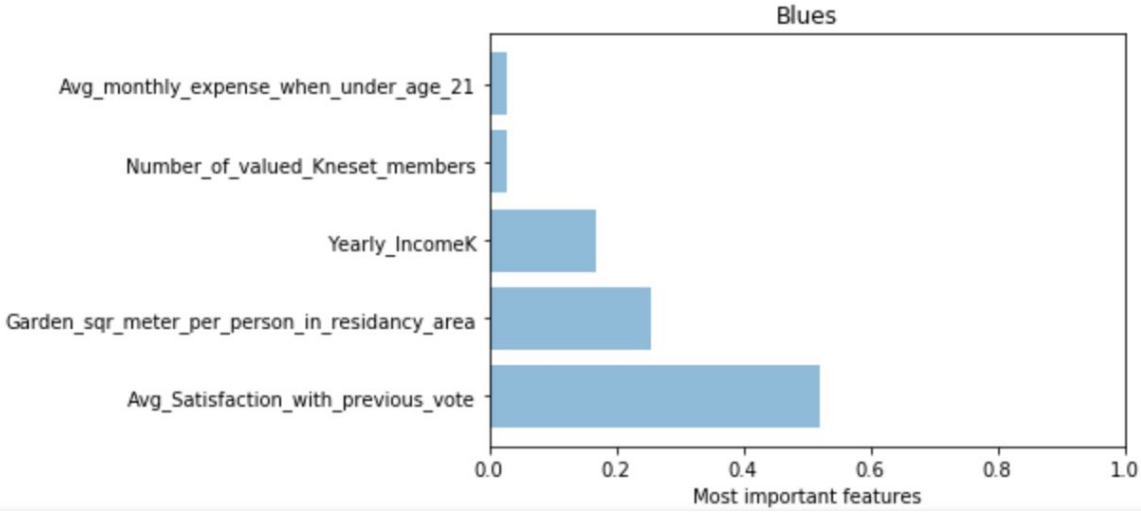
ניתן לראות שקיימות מפלגות שקיים פיצ׳ר שמשפיע עליהם כמעט בבלעדיות:
למשל הירוקים שמושפעים כמעט רק מhapiness



והצהובים שמושפעים כמעט רק מ will vote only large party



כמו כן ישנם מפלגות שמושפעות משני פיצרים משמעותיים כמו הכחולים:



ויש מפלגות שמושפעות מכמה פיצ׳רים במידה שווה.
אפשר לראות עבור כלמפלגה את הפיצ׳רים המשפיעים עליה ביותר במחברת

השפעה על פיצר ע״מ לקבל מנצח אחר

כדי לקבל מנצח אחר ניסינו להסתכל על הפיצ׳רים המשמעותיים לכל מפלגה ולהשפיע עליהם ע״מ שהמפלגה תזכה, הסתכלנו על ההתפלגות שלהם באותם מפלגות ושינינו בהתאם:

```
Original winner is party ## Purples ##

Garden_sqr_meter_per_person_in_residancy_area = 0.33
winner is party ## Greens ##

Will_vote_only_large_party = 0.9
winner is party ## Yellows ##

Number_of_valued_Kneset_members = 0.236734
winner is party ## Browns ##

Garden_sqr_meter_per_person_in_residancy_area /= 0.236734
Number_of_valued_Kneset_members -= 0.23
winner is party ## Pinks ##
```

השפעה על פיצ׳ר ע״מ לשנות את הקואליציה

ע״מ להשפיע על הקואליציה ראינו ש will_vote_only_large_party הוא מפריד ממש ביניהם.
המשמעות היא, שאם נגדיל את הרצון של הבוחרים להצביע למפלגות גדולות בלבד, יעברו קולות מהקואליציה ואילו אם נקטין אותו הקואליציה תגדל.