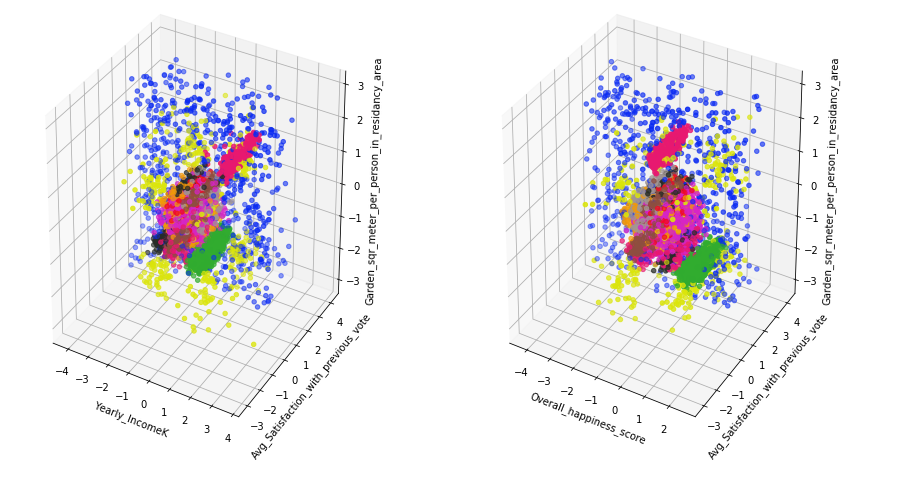
# Exercise 4

# Stable coalition

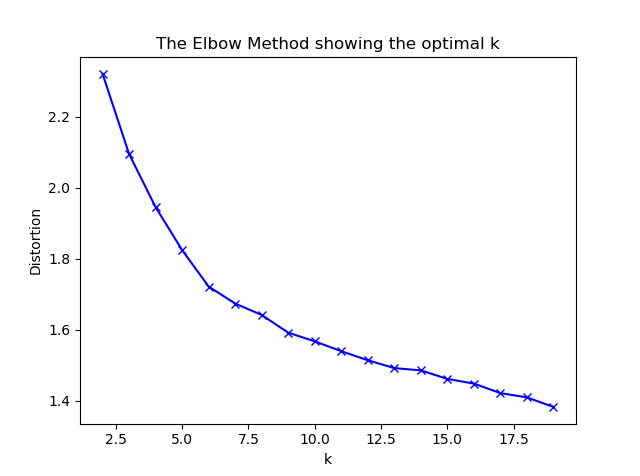
First, we try to find clustering in the data. By identifying clusters which have a "good" mix of two or more parties, we might be able to identify parties which are 'close together' in their features.

First and foremost, we took another look at the clustering we did for the previous assignment, which gave us a good "feel" of what clustering to expect, at least for the big parties:

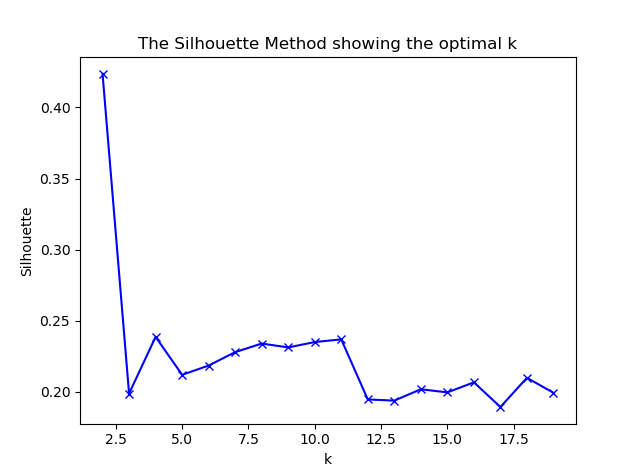


After that, we decided to train k-mean. For that, we want to figure out how many clusters we should pick, as it's not guaranteed that each party is clustered around one point.

For that we use both the elbow method:



And the Silhouette method:



From this we conclude we should probably use 8-11 clusters.

After training the k-means model with 11 clusters, we want to see the spread of the parties across the clusters:

## Example output of count of parties in each cluster:

'0': {'Reds': 320, 'Yellows': 36, 'Greys': 322, 'Oranges': 307},

'1': {'Pinks': 8, 'Yellows': 39, 'Purples': 13, 'Greens': 983},

'2': {'Whites': 78, 'Browns': 665, 'Pinks': 241, 'Yellows': 13, 'Purples': 68},

'3': {'Blues': 116, 'Yellows': 9},

'4': {'Pinks': 394, 'Yellows': 18},

'5': {'Blues': 126, 'Yellows': 1},

'6': {'Whites': 109, 'Browns': 434, 'Pinks': 2, 'Purples': 185},

'7': {'Blues': 111, 'Yellows': 2},

'8': {'Whites': 13, 'Yellows': 25, 'Oranges': 1, 'Browns': 4, 'Pinks': 13, 'Purples': 967}

'9': {'Blues': 70, 'Yellows': 104},

'10': {'Blues': 122, 'Yellows': 81},

## Example output of spread of parties across clusters:

'Reds': {'0': 320},

'Greens': {'1': 983},

'Whites': {'8': 13, '2': 78, '6': 109},

'Yellows': {'10': 81, '1': 39, '0': 36, '3': 9, '2': 13, '5': 1, '4': 18, '7': 2, '9': 104, '8': 25},

'Greys': {'0': 322},

'Oranges': {'0': 307, '8': 1},

'Browns': {'8': 4, '2': 665, '6': 434},

'Blues': {'9': 70, '10': 122, '3': 116, '5': 126, '7': 111},

'Pinks': {'1': 8, '8': 13, '2': 241, '4': 394, '6': 2},

'Purples': {'1': 13, '8': 967, '2': 68, '6': 185}}

We also calculated the distances between the clusters' centers, to figure out which clusters were close to each other and which were far (by using the median distance as a yardstick).

## Example output clusters distances:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster distances** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **0** | Close | Close | Close | Far | Close | Far | Close | Far | Close | Close | Close |
| **1** | Close | Close | Close | Far | Far | Far | Close | Far | Close | Close | Far |
| **2** | Close | Close | Close | Far | Far | Far | Close | Far | Close | Close | Far |
| **3** | Far | Far | Far | Close | Far | Far | Far | Far | Far | Far | Far |
| **4** | Close | Far | Far | Far | Close | Far | Close | Far | Close | Far | Far |
| **5** | Far | Far | Far | Far | Far | Close | Far | Far | Far | Close | Far |
| **6** | Close | Close | Close | Far | Close | Far | Close | Far | Close | Far | Close |
| **7** | Far | Far | Far | Far | Far | Far | Far | Close | Far | Far | Far |
| **8** | Close | Close | Close | Far | Close | Far | Close | Far | Close | Far | Close |
| **9** | Close | Close | Close | Far | Far | Close | Far | Far | Far | Close | Close |
| **10** | Close | Far | Far | Far | Far | Far | Close | Far | Close | Close | Close |

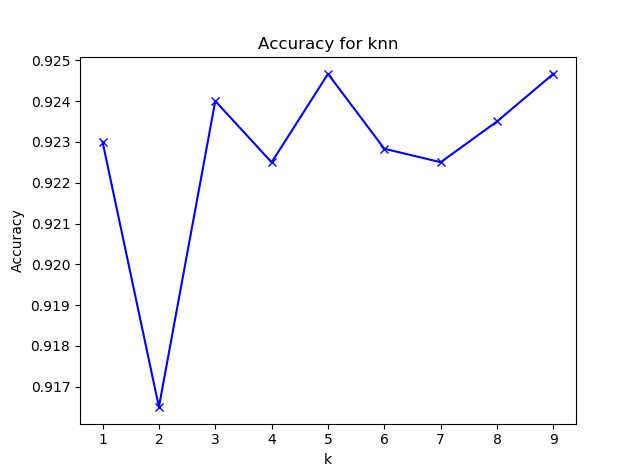
We ran this process several times to see the results.

### From this we conclude the following:

1. Yellows and Blues are pretty spread out, but tend to cluster together.
2. Reds, greys and oranges are very similar, almost always end up in one cluster, and are not too far away from the other two main groups.
3. Whites, browns, pinks and purples are pretty close together.
4. Greens seems to be in their own cluster usually and close to browns.

Next, we try the k-nearest-neighbors clustering algorithm.

First, we run it with in classification mode, with 5-fold cross validation to figure out what would be a good K to set:



We decide to select K = 5, to get good accuracy on one side, but ability to "spread out" the neighbors for each point.

Then, we count for each point how many neighbors it has from each party, and sum over all the parties:

## Example output for KNN:

'Reds': {'Reds': 1146, 'Greys': 64, 'Oranges': 70}

'Greens': {'Greens': 3932}

'Whites': {'Whites': 516, 'Browns': 219, 'Pinks': 40, 'Purples': 25}

'Yellows': {'Blues': 243, 'Yellows': 1069}

'Greys': {'Reds': 69, 'Greys': 1145, 'Oranges': 74}

'Oranges': {'Reds': 145, 'Greys': 113, 'Oranges': 974}

'Browns': {'Whites': 111, 'Browns': 4182, 'Pinks': 98, 'Purples': 21}

'Blues': {'Blues': 1847, 'Yellows': 333}

'Pinks': {'Whites': 50, 'Browns': 345, 'Pinks': 2190, 'Purples': 47}

'Purples': {'Whites': 29, 'Browns': 29, 'Pinks': 64, 'Greens': 13, 'Purples': 4797}

We see that we get similar results to the first method, which we use to strengthen our conclusions.

Finally, we trained a Gaussian Naive Bayes model, and from it we spread out the feature distribution for each party (See below).

From that we concluded which features are similar between the parties are which are dissimilar, which can support our selection of coalition.

## Coalition:

We suggest a narrow coalition of browns-purples-pinks-whites which should be higher than 50% of the votes, and contains 4 parties which are pretty close together.

This, in our opinion, would be stronger than the alternatives:

Greens are a large party, and pretty distinct and even their closest "ally" Browns seems to not be that close to them.

Red-Gray-Orange could almost be seen as one party.

Blue and Yellows are medium sized parties, but tend to be spread out in their locations, so it's hard for the parties to be close to them in view. Any coalition containing them would need much more support from bigger, "tighter" parties.

Brown and Purple are large and close together and so can provide a solid core for a coalition, and they have enough of close "allies", clustered with them or very close, to get above 51%.

# Major features

To identify the major features for each party, we trained a Naive Bayes model, and for each feature, checked the mean and the variance for each party.

We tried to understand which features have distinct means for each party, preferably with low variance.

## Sample output with important features marked:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Blues** | **Browns** | **Greens** | **Grays** | **Oranges** | **Pinks** | **Purples** | **Reds** | **Whites** | **Yellows** |
| **Kneset\_members** | 0.738 | -0.9 | 0.689 | -0.85 | -0.769 | -0.7 | 0.759 | -0.51 | -0.498 | 0.718 |
| **Yearly\_IncomeK** | -0.91 | -0.293 | 0.648 | 0.11 | -0.382 | 0.99 | -0.17 | -0.25 | -0.545 | 0.121 |
| **happiness\_score** | -1.1 | -0.564 | 1.362 | -0.68 | -0.362 | 0.02 | 0.114 | -0.01 | 0.08 | 0.277 |
| **previous\_vote** | 2.28 | -0.245 | -0.238 | -0.22 | -0.216 | -0.28 | -0.243 | -0.26 | -0.228 | 0.042 |
| **only\_large\_party** | 1 | -1 | -1 | 0 | 0 | -1 | -1 | 0 | -1 | 1 |
| **Garden\_sq** | -0.08 | -0.112 | -0.864 | 0.33 | -0.005 | 1.06 | 0.062 | 0.366 | 0.205 | -0.001 |
| **education\_rank** | 1.883 | -0.344 | 0.476 | -0.73 | -0.29 | 0.33 | -0.666 | -0.89 | -0.453 | 0.728 |
| **Issue\_0** | 0.198 | 0.198 | 0.173 | 0 | 0 | 0.24 | 0.201 | 0 | 0.245 | 0.22 |
| **Issue\_1** | 0.2 | 0.2 | 0.214 | 0 | 0 | 0.18 | 0.221 | 0 | 0.215 | 0.223 |
| **Issue\_2** | 0.213 | 0.217 | 0.189 | 0 | 0 | 0.21 | 0.194 | 0 | 0.205 | 0.201 |
| **Issue\_3** | 0 | 0 | 0 | 0.47 | 0 | 0 | 0 | 0.469 | 0 | 0 |
| **Issue\_4** | 0.161 | 0.186 | 0.209 | 0 | 0 | 0.17 | 0.196 | 0 | 0.155 | 0.162 |
| **Issue\_5** | 0 | 0 | 0 | 0.53 | 0.555 | 0 | 0 | 0 | 0 | 0 |
| **Issue\_6** | 0 | 0 | 0 | 0 | 0.445 | 0 | 0 | 0.531 | 0 | 0 |
| **Issue\_7** | 0.228 | 0.199 | 0.216 | 0 | 0 | 0.2 | 0.188 | 0 | 0.18 | 0.195 |

We can observe the following:

1. Number of valued Knesset members is a good divider into two groups, and only vote to a large party is a good divider into 3 groups. They are probably important to all the parties.
2. High yearly income is distinctively a Pink feature. At first it seems like low income was a Blue feature, but its variance turned out to be too high.
3. Will only vote for a large party and most important issue separated Grays, Oranges and Reds from the rest. Inside this sub group, Issue 3 separates the Oranges, Issue 5 the reds, and Yearly income the Grays.
4. High satisfaction with previous vote is pretty distinctively a Blue feature, which is partly shared with Yellows.
5. Greens are distinctively happy, blues are distinctively unhappy.
6. Greens have low garden size, pinks have high.

# Feature manipulation

We used two types of inference for this part, one of which we've done in the extra assignment for the previous part.

One option which we used in the extra assignment, is training a decision tree to identify the most significant features, as it tends to split on the most decisive features.

(The original winner we predicted was purples).

The features we identified were:

After increasing Overall\_happiness\_score by 1.75 (All values are normalized), the predicted winner is Greens.

After decreasing Overall\_happiness\_score by 0.25, the predicted winner is Browns.

After changing Will\_vote\_only\_large\_party\_int to 1, the predicted winner is Yellows.

After changing Will\_vote\_only\_large\_party\_int to 0, the predicted winner is Oranges.

After increasing Garden\_sqr\_meter\_per\_person\_in\_residancy\_area by 2, the predicted winner is Whites.

After decreasing Number\_of\_valued\_Kneset\_members by 0.25, the predicted winner is Browns.

The second option is to consult the table above to see which features are most likely to affect one large party to "move near" another one.

From looking at it, we can see several options:

1. Move the purples towards the greens: Increase happiness and decrease garden size.

After increasing Overall\_happiness\_score by 1 and decreasing Garden\_sqr\_meter\_per\_person\_in\_residancy\_area by 1, the predicted winner is Greens.

1. Move the purples towards the browns: Decrease Knesset members and increase education.

After decreasing Number\_of\_valued\_Kneset\_members by 0.15 and increasing Weighted\_education\_rank by 1, the predicted winner is Pinks.

# Strengthening the coalition

The coalition we selected is rather similar in its most distinctive features. The main differentiator is the Purples high number of valued Knesset members, where the rest of the coalition has it pretty low. We could decrease its values to move the purples closer to the rest of the coalition.

Another possibility is to move the red-gray-orange cluster closer to the pinks, by "resetting" the most important issue which will move their cluster closer to the coalition cluster.