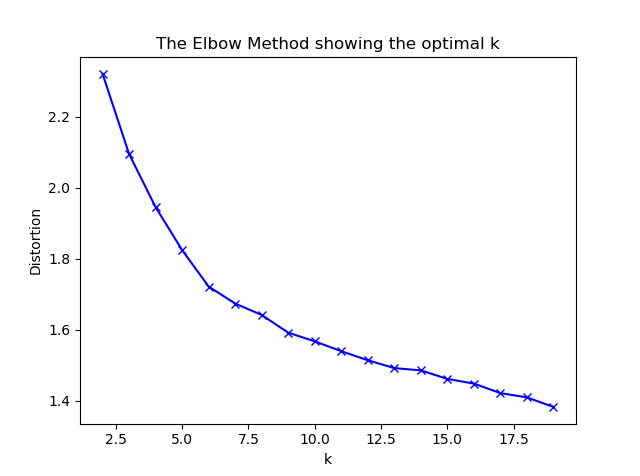
# Exercise 4

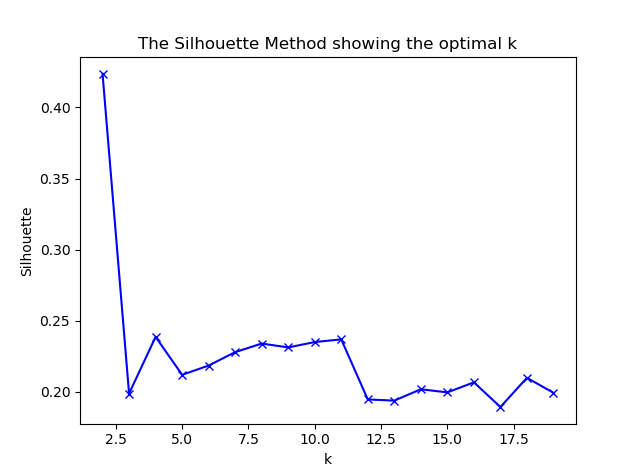
First, we try to find clustering in the data. By identifying clusters which have a "good" mix of two or more parties, we might be able to identify parties which are 'close together' in their features.

We first 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:

# Count of parties in each cluster:

'0': {'Whites': 109, 'Browns': 434, 'Pinks': 1, 'Purples': 184}

'1': {'Pinks': 11, 'Blues': 1, 'Yellows': 17, 'Purples': 14, 'Greens': 473}

'2': {'Whites': 78, 'Browns': 665, 'Pinks': 240, 'Yellows': 16, 'Purples': 65}

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

'4': {'Blues': 155, 'Yellows': 33}

'5': {'Blues': 4, 'Yellows': 52, 'Purples': 5, 'Greens': 510}

'6': {'Blues': 130, 'Yellows': 3}

'7': {'Whites': 13, 'Yellows': 26, 'Oranges': 1, 'Browns': 4, 'Pinks': 12, 'Purples': 965}

'8': {'Blues': 133, 'Yellows': 83}

'9': {'Reds': 320, 'Yellows': 39, 'Greys': 322, 'Oranges': 307}

'10': {'Blues': 122, 'Yellows': 41}

# Spread of parties across clusters:

'Reds': {'9': 320}

'Greens': {'1': 473, '5': 510}

'Whites': {'0': 109, '2': 78, '7': 13}

'Yellows': {'10': 41, '1': 17, '3': 18, '2': 16, '5': 52, '4': 33, '7': 26, '6': 3, '9': 39, '8': 83}

'Greys': {'9': 322}

'Oranges': {'9': 307, '7': 1}

'Browns': {'0': 434, '2': 665, '7': 4}

'Blues': {'10': 122, '1': 1, '5': 4, '4': 155, '6': 130, '8': 133}

'Pinks': {'1': 11, '0': 1, '3': 394, '2': 240, '7': 12}

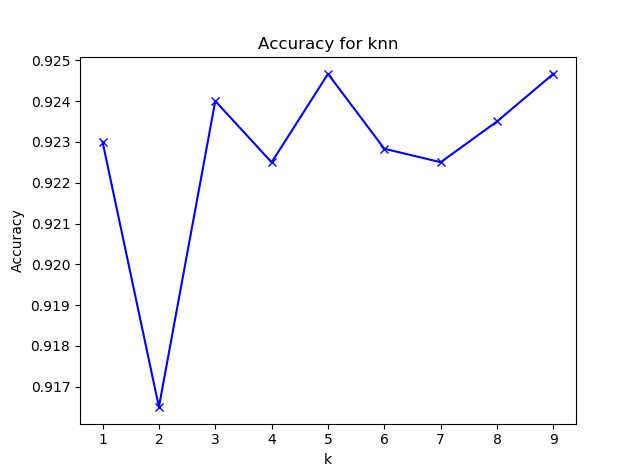
'Purples': {'1': 14, '0': 184, '2': 65, '5': 5, '7': 965}

From this we conclude the following:

1. Yellows and blues tend to cluster together.
2. Reds, greys and oranges are very much alike, and pretty close to yellow.
3. Whites, browns, pinks and purples are pretty close together.
4. Greens seems to be pretty distinct, and closest to yellows.

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:

'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.

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.