Project work 2: Report

Insights from data

Viivi Väisänen

1. Data preprocessing

- Removing summary variables from the vaalikone_questions
- Exploring missing values: the precentage of NA-values accross all columns was about 20%. Alot of columns that where all awnsers were NA.
- Handling and imputingmissing values: I first removed rows from vaalikone_questions that were all NA (excluding ID,) also rows that were at least 50% NA. After that I imputed all the left missing values, I did it separatly to numeric and char colums.
- Converting columns starting with 'Q' to factors, starting with 'W' to ordered factors.

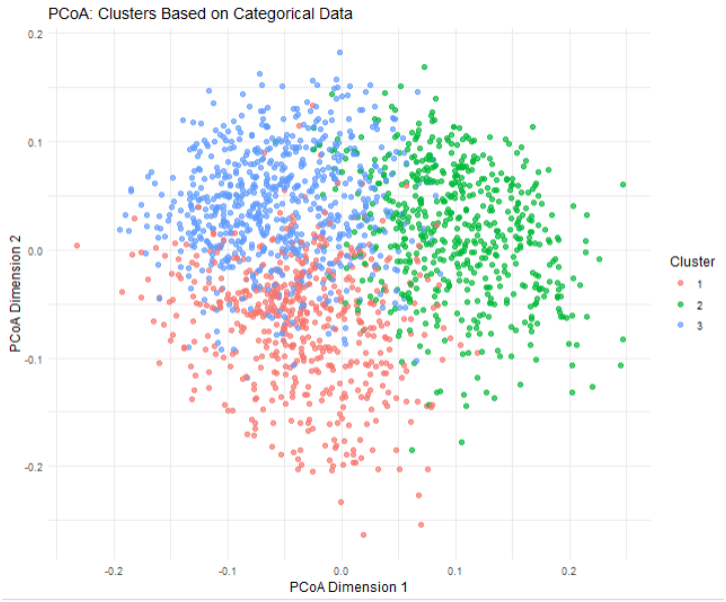
2. Clustering

- Computing Gower distance
- Finding out the optimal number of clusters with Silhoutte width: by the plot, optimal k is 2-3
- Performing PAM clustering with k=3
- Vizualizing with ggplot2
- Adding cluster results to vaalikone_questions
- Merging vaalikone_questions with vaalikone_profiles by ID
- Visualizing different variables accross clusters

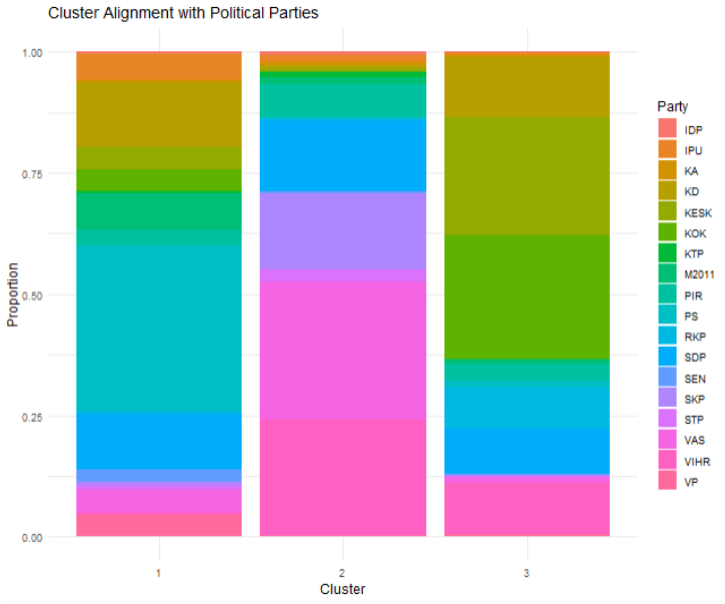
I ended up choosing Gower distance, because it is designed for datasets with mixed datatypes, and my cleaned dataset consisted of factor or ordered factor columns.

I used PAM clustering, because it directly utilizes Gower distance and it is good technique for mixed data types. Silhoutte width suggested that the optimal number of cluster could be 2 or 3.

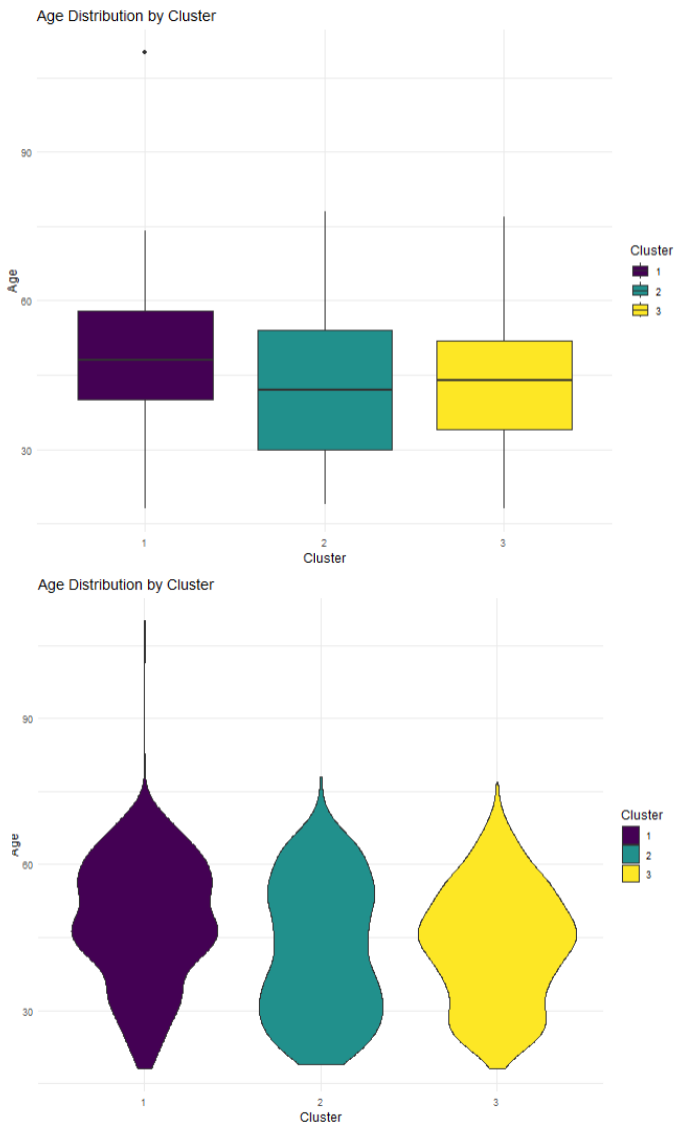
3. Results, key findings & insights



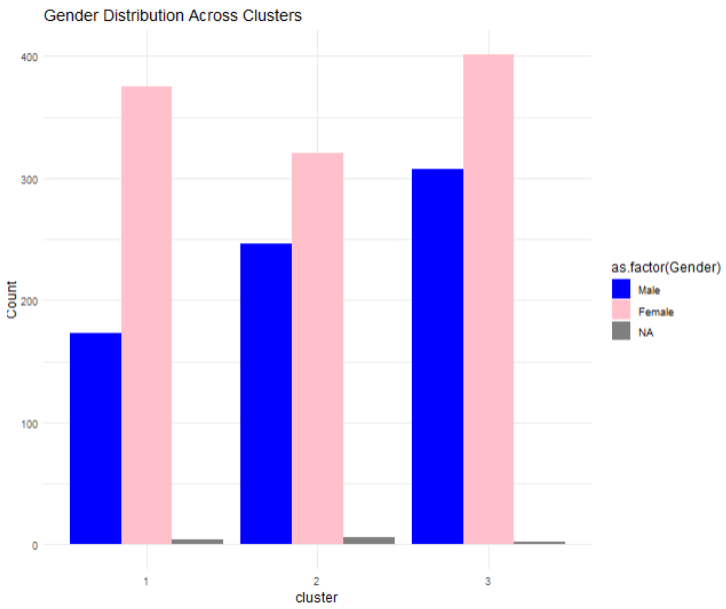
I used 3 as the number of clusters. PCoA visualization of the clusters shows, that each cluster appears to group individual data points with similar characteristics. There is a noticable separation between the 3 clusters. In the visualization, cluster 2 (blue dots) overlaps partially with both clusters 1 and 3, which suggests that it can share traits with both, or be a transitive group.



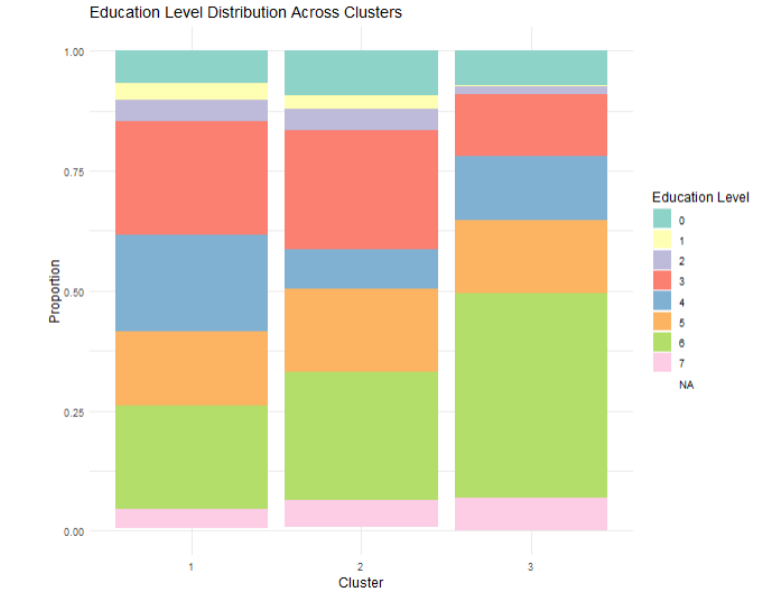
After merging the datasets, I visualized the cluster alignment with political parties. The bar chart shows, that cluster 1 is predominantly represented by party PS, and also other blueish colours. Cluster 2 consists of dominant segments, which are the two parties marked in pink: VIHR, VAS. Cluster 3 is heavily dominated by parties marked in green and brown which are KD, KESK and KOK.

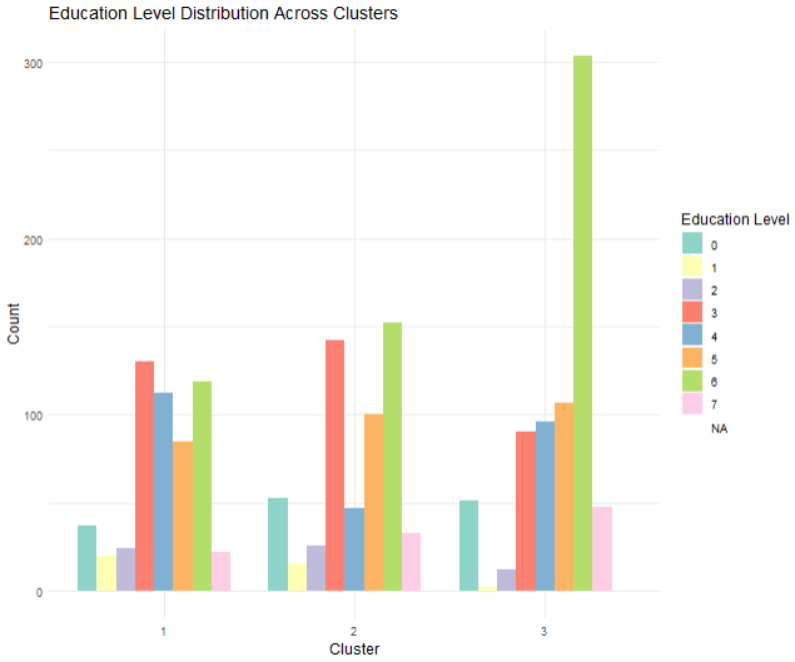


Ggplot and geom_violin visualization of cluster alignment with age shows, that all clusters have fairly evenly distriputed age, but cluster 1 seems to have fewer individuals over 30. Also cluster 2 has most variety in age. Geom_violin shows more detailed distripution, and there it can be seen that cluster 2 has the highest concenration of young candidates, when cluster 1 has lowest.

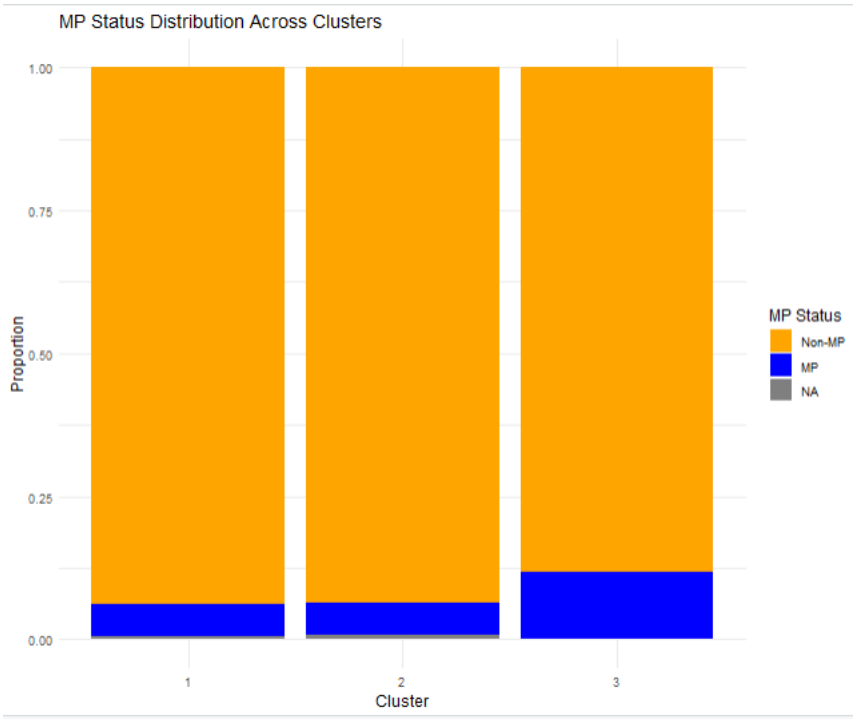


I visualized gender distripution accross clusters with side by side barblot (geom_bar), and it can be seen, that each cluster has higher number of female candidates. The highest number of females is in cluster 3, and the largest difference between two genders was in cluster 1.





When visualizing distripution on education levels across clusters using side- by-side bar plot, it shows that cluster 3 has twise as many candidates with education level 6 compared to cluster 2. However, cluster 2 has the second-highest number of level 6 candidates. Cluster 3 also has the highest number of candidates with highest education level 7. All of the clusters show tendency towards higher education.



I checked MP status distribution across every cluster, and it showed, that cluster 1 has least MP members, cluster 2 has a bit more compared to cluster 1. Cluster 3 had the most amount of MP members, almost twice as many MP candidates compared to cluster 1 and 2.

Analyzing the meaning of clusters:

Cluster 1: Right-Wing Nationalism, Social Conservatism, and Social Democracy

Cluster 1 is predominantly aligned with PS (Perus-Suomalaiset), which is typically associated with right-wing or nationalist ideologies. It is also represented second most by party SDP which traditionally advocates for social welfare and equality, and also KD, which focus on socially conservative and Christian democratic values. In short, cluster 1 reflects a blend of right-wing nationalism, social democracy, and Christian conservatism.

Cluster alignment with age showed, that cluster 1 has largest number of older candidates, and lowest number of younger candidates. Cluster 1 seems to lean towards older ages more than other clusters.

Cluster alignment with gender showed, that there is biggest difference between two genders in cluster 1: there is about two times more women than men in cluster 1.

Education level distripution across clusters showed, that there is relatively even distripution accross different education levels. Still, higher education (3, 4, 5 & 6) are dominant.

Cluster 2: Left-Wing, Progressive, and Environmental Focus

Cluster 2 is largely presented by more left-wing parties, notable VIHR and VAS. Cluster 2 is also almost entirely devoid of representation from KD, KESK, KOK, KTP, M2011, PS and KA. From this, it can be concluded that cluster 2 is not significantly connected to right-wign or traditionally center-right parties. This suggests, that cluster 2 is likely more strongly associatedwith left-wign and progressive parties, such as VIHR and VAS, which focus on enviromental protection and social justice.

Cluster alignment with age showed, that cluster 2 has highest number of younger candidates, but a peak around the middle age.
Cluster alignment with gender showed, that in cluster 2, there is the smallest difference between men and women.
Education level distripution across clusters showed a similar pattern to cluster 1, but has a bit higher count in level 3, 5 and 6.

Cluster 3: Center-Right to Right-Wing Conservatism and Academic Excellence

Cluster 3 is heavily dominated by parties KD, KESK and KOK. These parties are generally associated with center-right to right-wing ideologies. Thus, cluster 3 likely represents a center-right to conservative political orientation, with a focus on traditional values and moderate right-wing policies. Cluster alignment with age showed, that cluster 3 has a younger age distripution, with most individuals in their 30s or 40s.
Cluster alignment with gender showed, that cluster 3 has the highest numer of women accross other clusters.
Education level distripution across clusters showed, that cluster 3 has the highest number of education level 6 candidates. The proportion of level 6 compared to other levels is large, and the proportion of level 6 education level is more than twice as high compared to the cluster 2, where level 6 is the second most common. In cluster 3 there is also the highest amount of the highest levet 7. This may suggest that cluster 3 primarily consists of individuals with a high level of education. The high proportion of higher education, particularly level 6, may also indicate that this cluster likely contains more individuals with an academic background or specialized skills.
Cluster 3 also has high representation of MP candidates compared to other clusters. This high representation suggests that Cluster 3 includes individuals with stronger political engagement or leadership roles, possibly due to their high education.