# Bonus Questions – part 1

Lavi.lazarovitz

Rani.Sharim

1. Data preparation on data segments – please see notebook Bonus-1 for the data processing.

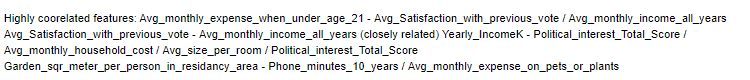
The first step into understanding the effect of working on segments of the data, is to actually run the analysis we have done on all the data, only on the training data. The process included the following steps:

* Understanding class dependency of features - we found the following features (27) to be somewhat dependent on the target class:



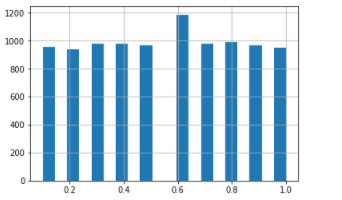
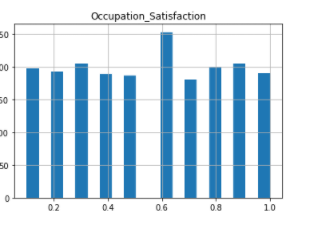
As opposed to running on the whole data which resulted in fewer features (18).

* Understanding features pairwise correlation – we have found the following features to be correlated:



* Each row resembles different cluster of correlation. For example the 2nd row means that previous\_vote, avg\_monthly\_income and the following features are correlated.

No big change identified here when comparing to the full analysis.

* Filling gaps – based on the above data, we filled each feature NaNs based on other correlated features or class dependency. Even though we based the filling strategy based on the insights above (class dependency and correlation), as we filled and filtered conservatively, we couldn’t notice any major / significant degradation. The distribution of the features remained pretty much the same as before:
*  

For example – occupation satisfaction distribution after filling it based on all the data (on the left) and based on training data on the right.

1. Hybrid strategy – We ran mutual information, variance and backwards elimination. This is the union set of the selected features (execution in the notebook line 247):
   1. '%Of\_Household\_Income',
   2. '%\_satisfaction\_financial\_policy',
   3. 'AVG\_lottary\_expanses',
   4. 'Avg\_Residancy\_Altitude',
   5. 'Avg\_Satisfaction\_with\_previous\_vote',
   6. 'Avg\_environmental\_importance',
   7. 'Avg\_monthly\_expense\_when\_under\_age\_21',
   8. 'Avg\_monthly\_household\_cost',
   9. 'Avg\_monthly\_income\_all\_years',
   10. 'Financial\_agenda\_matters\_int',
   11. 'Garden\_sqr\_meter\_per\_person\_in\_residancy\_area',
   12. 'Last\_school\_grades',
   13. 'Looking\_at\_poles\_results\_int',
   14. 'Main\_transportation\_int',
   15. 'Married\_int',
   16. 'Most\_Important\_Issue\_int',
   17. 'Number\_of\_valued\_Kneset\_members',
   18. 'Occupation\_Satisfaction',
   19. 'Occupation\_int',
   20. 'Overall\_happiness\_score',
   21. 'Phone\_minutes\_10\_years',
   22. 'Political\_interest\_Total\_Score',
   23. 'Voting\_Time\_int',
   24. 'Weighted\_education\_rank',
   25. 'Will\_vote\_only\_large\_party\_int',
   26. 'Yearly\_IncomeK'
2. The roles of the attributes can be identified by looking at the crosstab plots. The plots describe how each feature might be used to identify the associated class (line 10 in the notebook):
   1. 'Occupation\_Satisfaction' – distinguishes Greys (4.9) well
   2. Avg\_monthly\_expense\_when\_under\_age\_21 – distinguishes Blues(1502) well
   3. AVG\_lottary\_expanses – distinguishes Blues well
   4. Most\_Important\_Issue - distinguishes between Blues-Browns-Greens-Pinks-Purples-Yellows and Reds-Oranges-Greys (roughly)
   5. Avg\_Satisfaction\_with\_previous\_vote – distinguishes the Blues well
   6. Looking\_at\_poles\_results – distinguishes Yellows, Browns and Purples well.
   7. Garden\_sqr\_meter\_per\_person\_in\_residancy\_area – distinguishes between all the different classes well.
   8. Voting\_Time - distinguishes between all the different classes well.
   9. %Of\_Household\_Income - distinguishes Blues well
   10. Avg\_environmental\_importance – distinguishes purples well
   11. Yearly\_ExpensesK – distinguishes Browns well
   12. 'Yearly\_IncomeK' – distinguishes between all the different classes well
   13. Avg\_monthly\_household\_cost - distinguishes between all the different classes well
   14. Will\_vote\_only\_large\_party - distinguishes between all the different classes well
   15. Phone\_minutes\_10\_years – distinguishes the Green class well
   16. Weighted\_education\_rank - distinguishes between all the different classes well
   17. %\_satisfaction\_financial\_policy – distinguishes the Reds well
   18. Last\_school\_grades – Greys and Oranges well (34,39)
   19. Main\_transportation – distinguishes between all different classes well (roughly)
   20. Occupation - distinguishes roughly between all different classes
   21. Overall\_happiness\_score - distinguishes roughly between all different classes
   22. Financial\_agenda\_matters - distinguishes Purples and Greens well
   23. Political\_interest\_Total\_Score – distinguishes Greens and Blues well
   24. Number\_of\_valued\_Kneset\_members – distinguishes between Blues-Greens-Purples-Yellows and the rest well
   25. Married – distinguishes between all different class\
   26. 'Avg\_Residancy\_Altitude' – Distinguished the Reds and Browns well.

* The different attributes have some information about all or few of the different classes