*Lev Tsipes MASNA’23*

Multivariate Data Analysis

**Intro**

In this project I decided to dig deeply into human values of Italians. So, in this dataset throughout all the project I will use variables of the Human values sector. The purpose of this research, split into three parts (data exploration, clasterization and factor analysis) is to understand what the values of the Italians are, how are they distributed within the society, and what are their drivers. Theoretically, this project could be further broadened by doing the same thing but with different countries and develop some sort of HW (here and after – human values) map and check the Europe’s value consistency. Also, the factors generated in this research can be further used in regressions to build models, for examples, for predicting political views. For this project I used Python and Stata.

**Project One – Data Exploration**

In this research I use the ESS dataset of 2016. As stated on the website, “the survey involves strict random probability sampling, a minimum target response rate of 70% and rigorous translation protocols. The hour-long face-to-face interview includes questions on a variety of core topics.”

From this dataset I took the following variables:

Controlling variables:

1. gndr – gender (1 – man, 2 – woman)
2. Tporgwk – sector of economy the respondent works in (1 -Central or local government; 2 - Other public sector (such as education and health); 3 - A state owned enterprise; 4 - A private firm; 5 - Self employed
3. Agea – age
4. edlvdit – education level. The bigger the number, the higher is the step obtained.

HW variables (thematic):

1. Impdiff (important to try new things),
2. ipgdtim (important to have good time),
3. ipadvnt (important to seek adventures),
4. impfun important to seek fun)
5. imprich (important to be rich),
6. ipshabt (important to be admired),
7. ipsuces (important to be successful),
8. iprspot (important to get respect from others)
9. impsafe (important to live safe life),
10. ipfrule (important to follow the rules),
11. ipstrgv (important to have a protecting government),
12. ipbhprp (important to behave properly)
13. ipudrst (important to understand people),
14. iphlppl (important to care for others),
15. iplylfr (important to be loyal to friends)

Because the number of vars is big and it is hard to name each one with one short word, the names were not changed. Please refer to the translation presented above hereafter.

All Hw values are on scale 1-6, where 1 = Very much like me and 6 = Not like me at all.

All the rows with missing values or answers with refusal or ‘other’ were erased, resulting in 1773 rows out of 2626 intial. All HW are ordinal likert scales and we treated as numeric assuming equal distance between each answer.

For age we see something more or less like normal distribution:

Chart, histogram

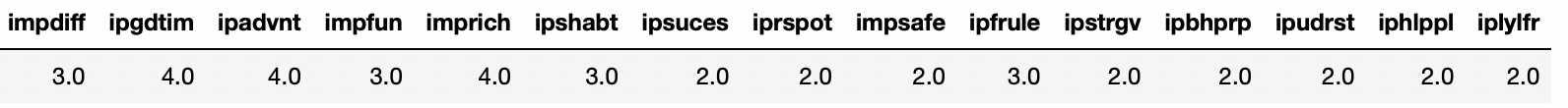
Description automatically generated

Whereas for all the thematic variables distribution is skewed: the right like for imprich or to the left like impsafe

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

So, due to skewness we would like to explore medians rather than means which are:

So, we can say that Italians generally less care about having fun and adventures and being rich whereas they care much more about feeling safe, taking care of fellows and being respected. Interestingly, results for success and richness variables are quite opposite, meaning that Italians, probably, do not connect these concepts directly.

Then, a Pearson correlation matrix was constructed and visualized as a heatmap. There are obviously two overlapping clusters of variables: impdiff – impsucess and iprspot – iplylfr. The first set of variables is about being admired, respected and have fun and the second one is about caring one another and be safe and sound. Interestingly, having respect is in the second cluster. Because the question sounds like “It is important to her/him to get respect from others. She/he wants people to do what she/he says” it may be precepted as a question about following orders and subordination.

A picture containing square

Description automatically generated

Based on the questions and the correlation table there were the following composite variables constructed:

1. Hedonism (mean of 'impdiff','ipgdtim','ipadvnt','impfun')
2. [Society] Status (mean of 'imprich','ipshabt','ipsuces')
3. Order (mean of 'iprspot', 'impsafe','ipfrule','ipstrgv','ipbhprp')
4. Care (mean of 'ipudrst','iphlppl','iplylfr')

For these composite variables we got such correlations:

Chart

Description automatically generated

So, we see that Italians preferring safer life also tend to care for their mates themselves (not egoistic people!) and people who care about having life full of excitements tend to demand for being respected. These correlations are positive moderate. Respect is also lowly correlated with security. If we create only two groups, as correlation matric tells us, it would be:

A picture containing text, electronics, screenshot

Description automatically generated

Full\_headonism is the mean of Hedonism and Status, Care\_and\_cared is the mean of Order and Care. We see that the correlation between them is low, which is good if the variables were needed for regressions. For our purpose, we now know drivers of Italian HW that can be summarized in two megavalues: living hedonistically and living safely.

Interestingly, the difference in means between tested by t-test is significant only for the megavalue care\_and\_cared, which for women is 2.2991 and 2.415.

When splitted by education, Fisher exact one-way test set showed difference in distribuition of care\_and\_cared and full\_hedonism as well.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

We can see that with education fun becomes less and less relevant. Interestingly, people with education level == 10, which is ‘Diploma universitario vecchio ordinamento’, which is Italian Masters of the pre-Bologna system. So, these people got their diploma long time ago and are 60 years old on average. However, this type of diploma got only 6 people in our group. For people with the highest education levels (15+) there are less than 10 people in each, so we will not consider the upgoing trend in end. So, neglecting all low-count groups, we see that need for care is more or less stable and independent of education, whereas more educated = more hedonistic the person is. Probably, people with higher education enjoy higher society status and are richer than people with lower education levels, who probably have to work more hours and at the same time get less riches.

So, on this stage we had a glimpse over the dataset we are going to work with later. We saw that all the thematic variables are ordinal and skewed, which may probably be needed to be accounted for later. We have generated already two proposed factors: 4 based on theory alone and two based on Pearson correlations analysis. It was seen that Italians are less prone to having fun and seeking for joy rather than being in need for safety and caring about others (my ex is Italian, so I know what I’m talking about). We also found out that the more educated the person is the more they are in need of dopamine and oxytocin. Probably, since higher education in Italy is not free, only people from rich families can afford it. However, the most interesting thing for me is that importance of being ‘rich’ and ‘successful’ are correlated on insignificant level (0.26) and being rich is much less important than being successful (2<4). So, what is ‘success’ for Italians? I want to get back to this question in the third part of the research

**Project Two – Clusters and PCA**

In this part of the work, we will, firstly, try to divide Italians into groups based on their human values and, secondly, and create components for further analysis that may be furtherly used, even though it is not needed for the main goal of our research.

So, let’s start off with hierarchical clustering. Unluckily, I am not good with R, I use python instead, so I cannot redo the R code. This is even more tragic, considering that there is no library computing Ward’s criterion function. So, instead of it I used silhouette score for cluster assessment. Hopefully, it is not a big change.

So, for hierarchical clustering I tried Ward, Average and Complete links with Euclidean distance each. The best score totaled for Ward in 0.07 for two clusters. For kmeans I got the same number for two clusters as well. I built centroid for them; however, they are not worth considering due to clusters being not informative (Ward too close to 0).

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

So, for clustering I tried clustering using four composite variables (Hedonism, Status, Order and Care) made in the first part. For K-means I got 0.275 and 0.282 scores for Ward and K-Means accordingly. This is better, but still the clusters are too overlapping. As I know, silhouette score of 0.6+ is considered satisfactory. However, let’s check this clusters.

Table

Description automatically generatedTable

Description automatically generated

Above are the centroids for K-Means on the left and hierarchical clustering on the right. In both cauterizations we see visually significant difference in centroids mainly between Hedonistic and Respect variables. On the one side there’s a group of Italians who praise living joyfully and being respected, whereas others cherish security and care more. Even though the numbers in the cauterizations are similar, the difference in distribution between clusters is dramatic. For KMeans the split is symmetric (841 people for 0 cluster and 928 for 1 cluster), whereas for hierarchical it is not (1308 people for 0 cluster and 461 for the first one).

Moving forwards to PCA, the appropriateness was checked with Bartlett test and p-value of 3.612151411523583e-176 was gotten, proving the sensibility of the method. Since PCA is aimed to save as much of variance as possible, my choice is based on cumulative variance explained. So, even though the most informative components are only the first five, I chose 8 componential model to retain more than 80% of variance. To retain all the variance 15 components (total number of thematic variables) would be needed. The most informative component is the first one, explaining about 30% of total variance.

Chart, line chart

Description automatically generatedChart

Description automatically generated

Since there are many components and it would be too much to interpret them all, I would attempt to do the first two informative ones, that explain about 50% of all the variance together. Unluckily, there’s no strict division between the variables and it’s hard to interpret the components. I can only say that PC1 is associated with ipgdtim, impfun and impdiff (a.k.a. having good time, fun and trying different things), meaning this component is close to our ‘Hedonism’ composite variable from the previous project. The biplot is presented below.

Chart, scatter chart

Description automatically generated

Overall, we can conclude that clustering and PCA are not informative methods in terms of explanatory data analysis. We can even say that the population cannot be successfully clustered (at least with the methods tested). Actually, this is a positive finding, meaning that there is no distinct separation within Italian society, and it is homogeneous in teams of HW. Otherwise, it would be full of conflicts and misunderstandings. Regarding the PCA method, it works for the data, cutting the number of variables in half, yet the components created are not senseful. Hence, for better components exploration Factor analysis is needed.

**Project Three – Factor Analysis**

The final stage of this research is Factor Analysis to explore what drivers Italian human values have. We will derive factors and analyze, how the HW are connected to one another and what broad ones like ‘success’ mean. The ‘factorability’ can be checked with Bartlett’s Test. Bartlett’s test of sphericity checks whether the observed variables intercorrelate at all using the observed correlation matrix against the identity matrix. If the test found statistically insignificant, you should not employ a factor analysis. P-val for the dataset is equal to zero, so we can apply factor analysis. Since the dataset consists of categorical ordinal variables, polychoric correlation matrix should be used.

A picture containing table

Description automatically generated

Recommended method for choosing the number of factors is selecting as many as have eigenvalue of more than one, which is two for our case. Actually, this is consistent with the correlation and theoretical assumptions from the first project. In total, the two factors together explain 45% of variance.

We can see quite appropriate uniqueness scores before any rotation for all the variables except for ipfrule (a.k.a “Important to do what is told and follow rules”) which has 0.82 uniqueness. It looks like this variable does not have strong connection to any of the two drivers. This is an interesting observation, meaning the needs for having secure government, staying safe and behaving properly have a different driver. So, probably, properness of behavior does not suppose following the rules, meaning that Italian society is regulated in an informal and self-regulatory fashion. The reason why wanting a secure government and willing to follow the rules can be explained this way: Italians do want a government that can dominate “the bad guys” but do not want to follow the rules themselves and by the rules they may understand paying taxes, parking or other “annoying” and “unnecessary” rules that do not cause threat to any people.

Table

Description automatically generated

After I apply orthogonal varimax rotation to get uncorrelated factors for better interpretability. I also tried oblique rotations, but it did not cause any improvements, so I stayed with the orthogonal one.

In the outlook I hid the loadings that are lower than 0.3. So, we see that the vast majority of variables are tied with only one of the two factors. The problems are tied with three variables: impdiff (Important to try new and different things in life), ipshabt (Important to show abilities and be admired) and ipsuces (Important to be successful and that people recognize achievements). Let’s get back to them later and interpret the factors ignoring these three variables.

The first factor including ipgdtim (good time), ipadvnt (adventures), impfun (fun) and imprich (rich) can be named as ‘Hedonistic’ factor. The second factor including impsafe (safe), ipstrgv (strong goverment), ipbhprp (proper bahavior), iprspot (getting respect), ipudrst (understand people), iphlppl (help people) and iplylfr (loyal to friends) I would name as ‘Community’ factor. So, Italians have two main HW drivers: 1) have good time and enjoy life and 2) belong to the community where a person feels protected and knows hir or her place in the social hierarchy.

Table

Description automatically generated with low confidence

Such interpretation helps explain two of the cross-loadings: ipshabt and ipsuces are products of both of these HW drivers. Success involves both joy of life with money and adventures and feeling involved in the community getting respect and friends. Showing abilities and getting admired also can be caused by getting, on the one hand, good emotions and, on the other hand respect and recognition.

As for impdiff (Important to try new and different things in life) it is a little bit trickier, yet I would suggest the following explanation. Probably, new and different things in life are associated with friends for Italians (as they are well-known to be an extraverted nation), so it means that the will for being a part of a community can be driven not only by the will for recognition, safety and engagement but also by the will to experience things with others, which is reflected in high loadings of impdiff with both factors. Below is presented the loading graph for better visualization. We can clearly see to distinctive lusters with impdiff, ipshabt and ipsuces in between them.

Chart, scatter chart

Description automatically generated

**Final Thoughts**

Human values are variables of an interesting nature. We have discussed that there are two basic values: having fun and be safe & cared. This is proven by both correlation and factor analyses. HV of being successful, living a different life and living a different live, though, do not suite well any of the two factors. This means, that these feeling/values are not strongly connected to fun&money and care&safety. I suppose, these values have the broadest meanings out of all other HWs in the research, so they contain other feelings that are not captured well with the paradigma of two meagvalues.

Also, as FA suggests, usually HV variables have not very small uniquenesses, meaning they are not strongly connected with one another. Hence, if one would like to use HW for their regression models, FA is not the best choice. PCA would get high variance explained but not make sensible components. Hence, if making a model, it is better to use the variables untouched or reduce their number by excluding some from the research.