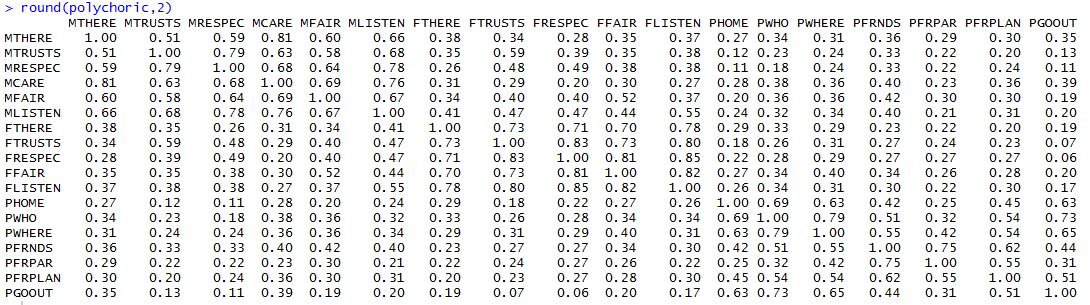
**Exam 3**

Luyao Zhang (NetID: lzhang94)

1. **EFA**
2. Since participants responded to the survey on a Likert scale from 0-4, it’s more appropriate to treat the 18 variables as ordinal. Therefore, for this dataset, we obtained a polychoric correlation matrix instead of the observed sample correlation matrix (Table 1).

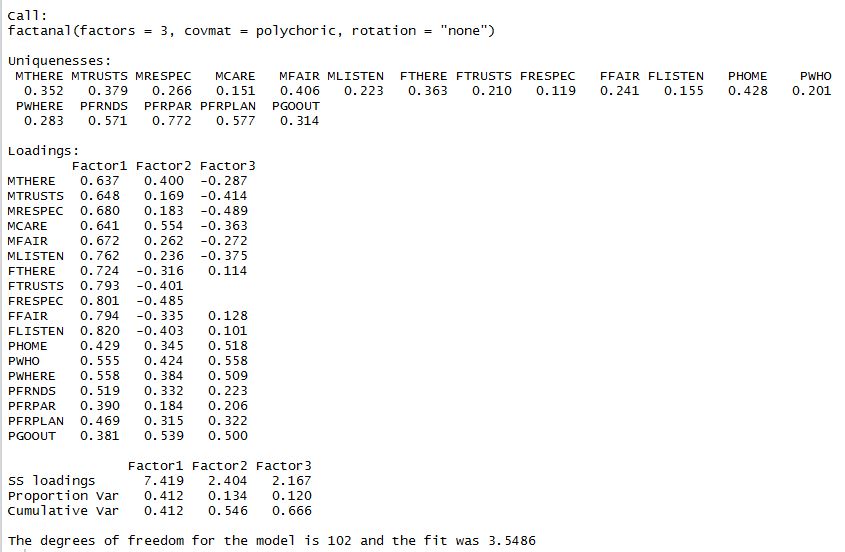
Table 1

*The polychoric correlation matrix for all 18 variables*



After looking through the data, I came up with a k equal to 3, because I think the 18 items on the survey asked participants about their relationships with (1) mothers, (2) fathers, and (3) parents as a whole.

Therefore, a 3-factor model was fitted with no rotation, and Figure 1 shows the results of the model. However, the loadings are not very satisfying, because most items have high loadings (> 0.3) on more than one factor, indicating that a 3-factor model is pretty far from the simple structure.



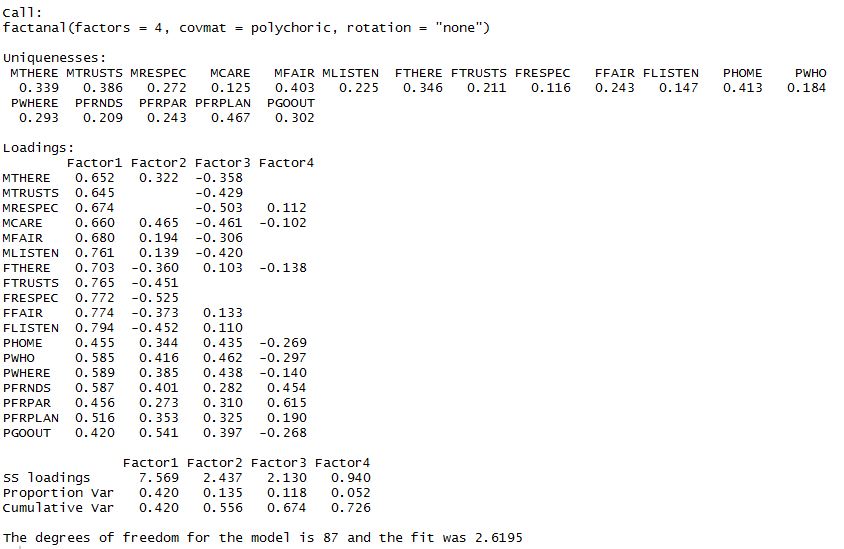
*Figure 1.* Results of fitting a 3-factor model to the dataset

To determine the value of k based on the eigenvalue > 1 rule, I first obtained the eigenvalues of all possible values of k (1 to 18), which can be found in Figure 2, indicating that a 4-factor model will be more appropriate in terms of having eigenvalues larger than 1.

1_a_2_eigenval.JPG

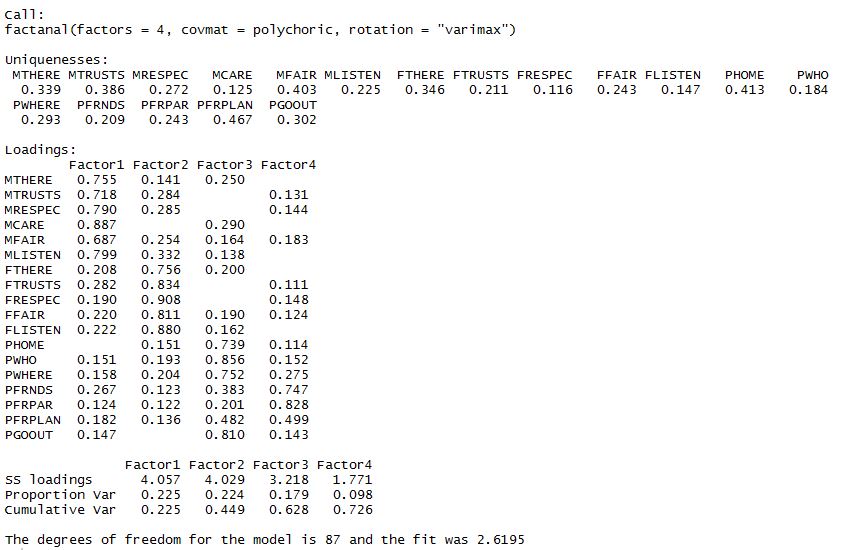
*Figure 2.* Eigenvalues for all values of k

Therefore, we fit a 4-factor model with no rotation, and the results (Figure 3) looked better than a 3-factor model. 4 factors together accounted for 72.6% of total variance, which is better than the 66.6% variance explained by a 3-factor model. However, because there was no rotation, the results are still quite different from a simple structure, where each variable has a very large loading on just one factor, and very small loadings on all of the other factors.



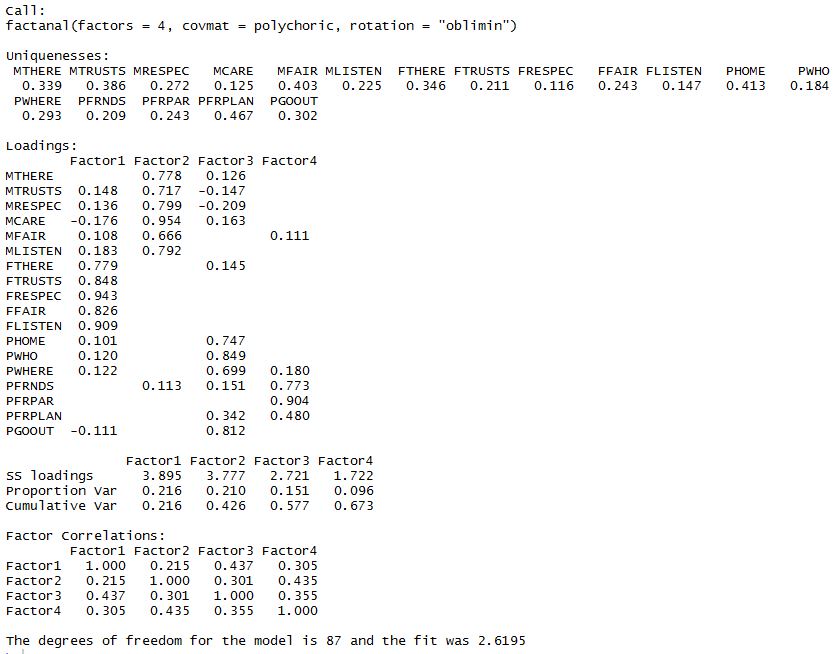
*Figure 3.* Results of fitting a 4-factor model to the dataset

1. Based on the apparent improvement after fitting a 4-factor model, I think a 4-factor model is more appropriate for the dataset and worth further exploration. The communalities can be computed by “1 – uniqueness”, and therefore MCARE, MLISTEN, FTRUSTS, FRESPEC, FLISTEN, PWHO, PFRNDS are the variables having most of their variance explained by the common factors, with their communalities being 0.875, 0.775, 0.789, 0.884, 0.853, 0.816, and 0.791, respectively.
2. Fitting a 4-factor model with varimax rotations returns a much better structure compared with when there was no rotation. Figure 4 shows the results. It looks that all mother-related variables fall under Factor1, and all father-related variables fall under Factor2. PHOME, PWHO, PWHERE, PGOOUT all fall under Factor3, and PFRNDS, PFRPAR fall under Factor4. However, it’s not clear which factor PFRPLAN belongs to, because it loads equally highly on Factor3 and Factor4.



*Figure 4.* Results of fitting the dataset with a 4-factor model using varimax rotation

Next, we fit a 4-factor model using oblimin rotation. Not surprisingly, the model yielded a similar pattern with a better structure than when a varimax rotation was used. Figure 5 shows the result. The structure is simpler than obtained using varimax rotation. Even for PFRPLAN, it is more clear that this variable belongs to Factor4 more than Factor3.



*Figure 5.* Results of fitting the dataset with a 4-factor model using oblimin rotation

According to the structure obtained with varimax rotation and oblique rotation, it’s obvious that oblique rotation describes the dataset better, in that it takes into consideration that factors are correlated with each other.

It’s pretty intuitive that relationship with mother, father, parents as a whole are associated with each other. It looks seems that Factor1 stands for relationship with mother, Factor2 stands for relationship with father, Factor3 stands for parents’ knowledge about where their children are, and Factor4 has to do with parents’ knowledge of their children’s friends.

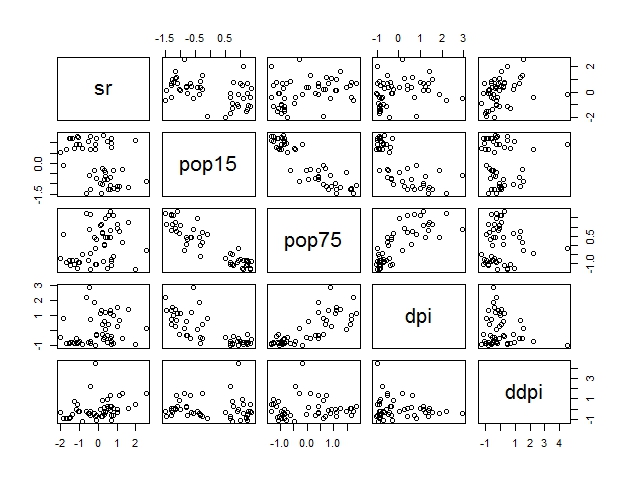
It’s not hard to understand why the loadings on Factor3 and Factor4 are not as different for PFRPLAN. Plans with friends have to do with both social life (who the child is hanging out with) and location (where the child hangs out with friends), which makes it harder for PFRPLAN to have overwhelmingly large loading on just one factor. However, it seems that PFRPLAN has more to do with parents’ knowledge of their children’s friends/friendships than with their children’s location.

1. **Cluster analysis**
2. To decide if the data needs to be standardized, first we obtained the column standard deviation (Figure 6). Apparently, columns have very different standard deviation, meaning that the data needs to be standardized before conducting the cluster analysis.

2_a_1_sd.JPG

*Figure 6.* Column standard deviation

To have a better idea of how many clusters the dataset contains, we obtained pair plots of the dataset (Figure 7). Although the pattern is somewhat ambiguous, it seems that 3 groups may be appropriate. Also, the 50 countries can be briefly categorized into the most developed (e.g., Switzerland, the U.S.), less developed (e.g., Brazil, Chile), and the least developed (e.g., Zambia).

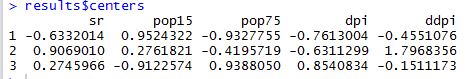


*Figure 7.* Plots by pairs

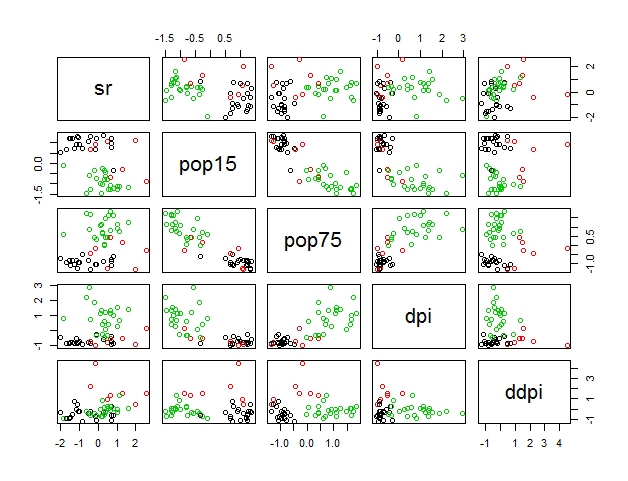
Therefore, we carried out a K-means cluster analysis with k = 3. Figure 8 and Figure 9 show the result. The pair plots can be found in Figure 10.



*Figure 8.* Cluster analysis (k = 3)

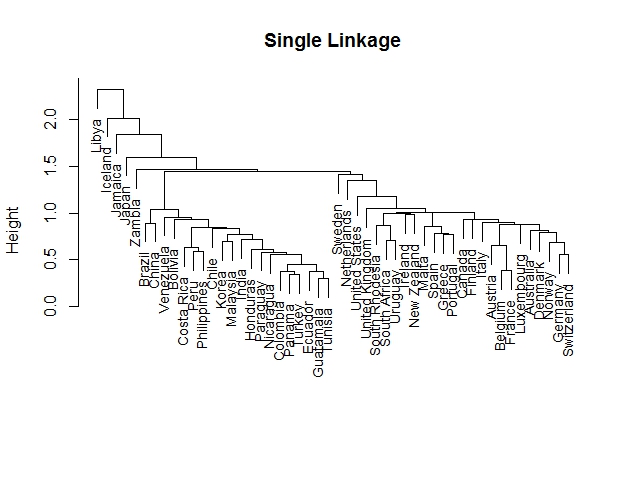


*Figure 9.* Centers (k = 3)

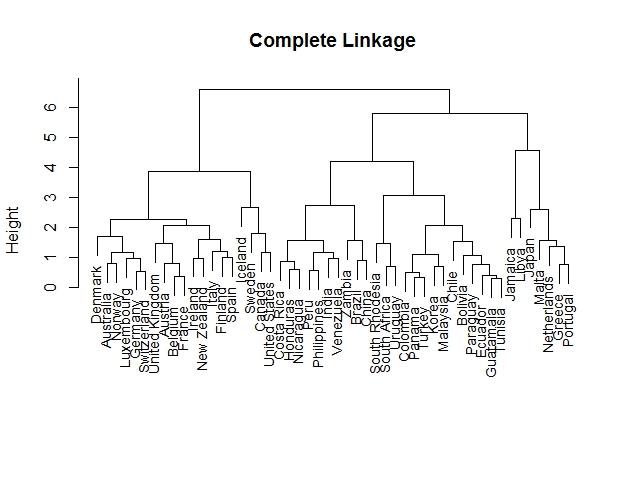


*Figure 10.* Pair plots (colored by cluster)

1. Next, we performed hierarchical cluster analysis using the standardized dataset. Figure 11 shows the results of using single linkage, and Figure 12 complete linkage.

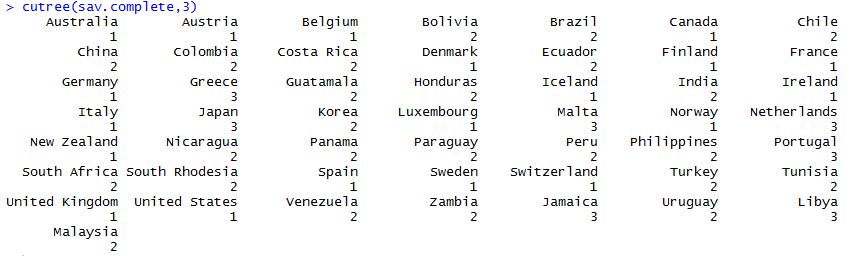


*Figure 11.* HCA with single linkage



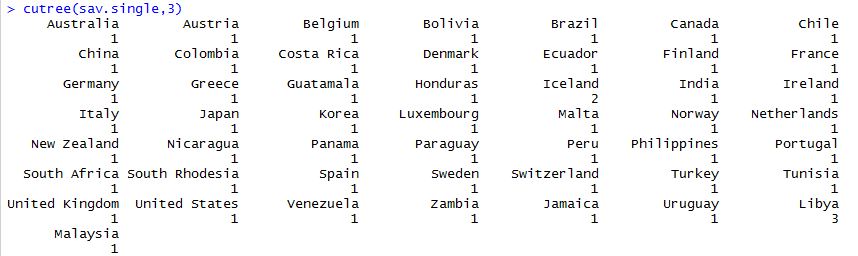
*Figure 12.* HCA with complete linkage

The complete linkage tree shows a somewhat clear pattern of 3 clusters, so we cut the tree into 3 clusters. Figure 13 shows the clusters.



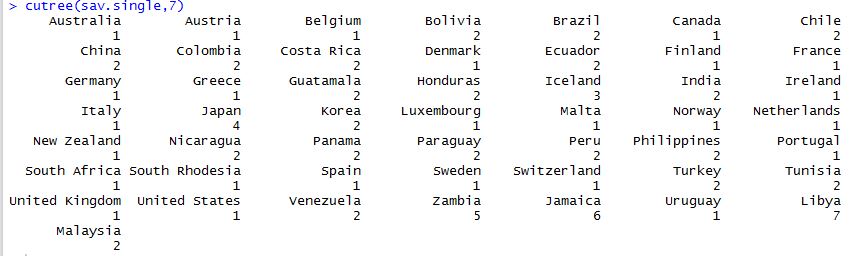
*Figure 13.* HCA with complete linkage (3 clusters)

We also tried cutting the tree at 3 clusters with single linkage, but the results grouped Iceland and Libya alone under Cluster2 and Cluster3, respectively, and all the other countries under Cluster1. Figure 14 shows the clusters.



*Figure 14.* HCA with single linkage (3 clusters)

Then we tried 4, 5, and 6 clusters, but obtained similar pattern for all 3 conditions (i.e., most countered under Cluster1 and all the other countries belonging to their own clusters). Until we tried cluster = 7, a more sensible pattern emerged (Figure 15). We still had some singletons, but at least we now have 2 clear clusters. We can tell from the clusters that the singletons are the countries on the very left of the single linkage tree plot (Figure 11), and the other two clusters are the bunches to the right.



*Figure 15.* HCA with single linkage (7 clusters)

**Comments:**

It seems that K-means and complete linkage cluster analyses yielded similar patterns, while single linkage yielded somewhat different groupings.

According to K-means and complete linkage cluster analyses, it is appropriate to say that the dataset roughly have 3 clusters of countries when taking all 5 variables into consideration. It seems that geography is related to the clustering, in that countries in Europe, North America, and Oceania. Also, these are traditionally more developed countries, and therefore, they have similar patterns in term of savings.

South American/Central American countries tend do belong to another cluster, with a few exceptions (e.g., Jamaica). Belonging to the same cluster are some Asian countries like Malaysia, Turkey, Philippines and Korea, as well as African countries such as Tunisia and South Rhodesia.

However, the pattern does not apply to all 50 countries. For example, grouping of China, Netherlands, and South Africa are inconsistent between the two approaches. Japan, and Eastern Asian country, doesn’t seem to be grouped with its fellow Asian countries, and Libya, an African country, does not align with other African countries, either.

In general, from the clustering, we can still tell that geography can help us predict to some extent which cluster a specific country belongs to.