Factor analysis: a way to take a mass of data and shrink it into a smaller, more manageable, and more understandable set. It's a way to find hidden patterns, show how they overlap, and show what characteristics are seen in multiple patterns. It is also used to create a set of variables for similar items in the set (these sets of variables are called dimensions).

Multiple Factor Analysis: This subset of factor analysis is used when your variables are organized into groups of variables. For example, you might have a student health questionnaire with several items like sleep patterns, addictions, psychological health, or learning disabilities.

There are three types of analysis.
PCA (Principal Component Analysis) analysis: only all variables are continuous variables.

- Example: data mining analysis (e.g., line of code)

MCA (Multiple Correspondence Analysis) analysis: all variables are categorical variables or continuous variables (qualitative variable---term used in statistics)

- Example: survey analysis

MFA (Multiple Factor Analysis) is a multivariate data analysis method for summarizing and visualizing a complex data table in which individuals are described by several sets of quantitative and/or qualitative data structured into ***groups***.

- Example: survey analysis + mining analysis (e.g., line of code)

In this study, we only focus on MCA and MFA since surveys are very popular in our group. In the sample spreadsheet (SOB.CSV), we have 19 columns:

Demographic info:
Gender: Man, Non-man
Education: high school, undergraduate, master and above
Residency: North America, Non-North America
Seniority: less than 3 years, 3-5 years, over 5 years
English_confident: yes, no
compensation: paid, unpaid
Challenges (how often do you face the following challenges in OSS):
Process.getting_started_on: Never, Rarely, Sometimes, Often
Process.navigating_contribution_process: Never, Rarely, Sometimes, Often
Process.reception_issues: Never, Rarely, Sometimes, Often
Process. licenses: Never, Rarely, Sometimes, Often
Social.communication_styles: Never, Rarely, Sometimes, Often
Social.fear_of_making_mistakes: Never, Rarely, Sometimes, Often
Social.lack_of_recognition: Never, Rarely, Sometimes, Often
Social.unwelcoming_environment: Never, Rarely, Sometimes, Often
Social. nationality: Never, Rarely, Sometimes, Often
Social.cultural_differences: Never, Rarely, Sometimes, Often
Project. documentation: Never, Rarely, Sometimes, Often
Project.technical_Hurdles: Never, Rarely, Sometimes, Often
SOB (I feel I belong to OSS): Agree, neutral, Disagree

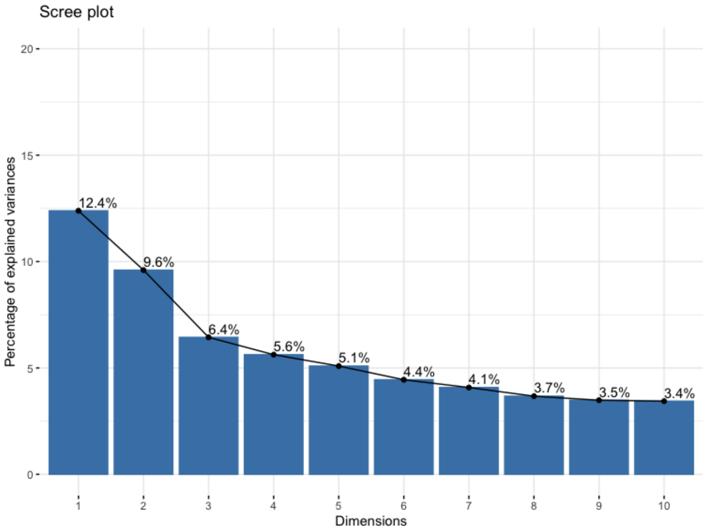
MCA analysis:

```
library("FactoMineR")
library("factoextra")
data(SOB)
#summary of the data:
Summary(SOB)

MCA(data, ncp = 5, graph = TRUE)
#X: a data frame with n rows (individuals) and p columns (categorical
#variables)
#ncp: number of dimensions kept in the final results
#graph: a logical value If true, a graph is displayed.
```

Let us start to visualize and interpret:

#To visualize the percentages of inertia explained by each MCA dimension, use the functions `fviz_eig()` or `fviz_screeplot()`:

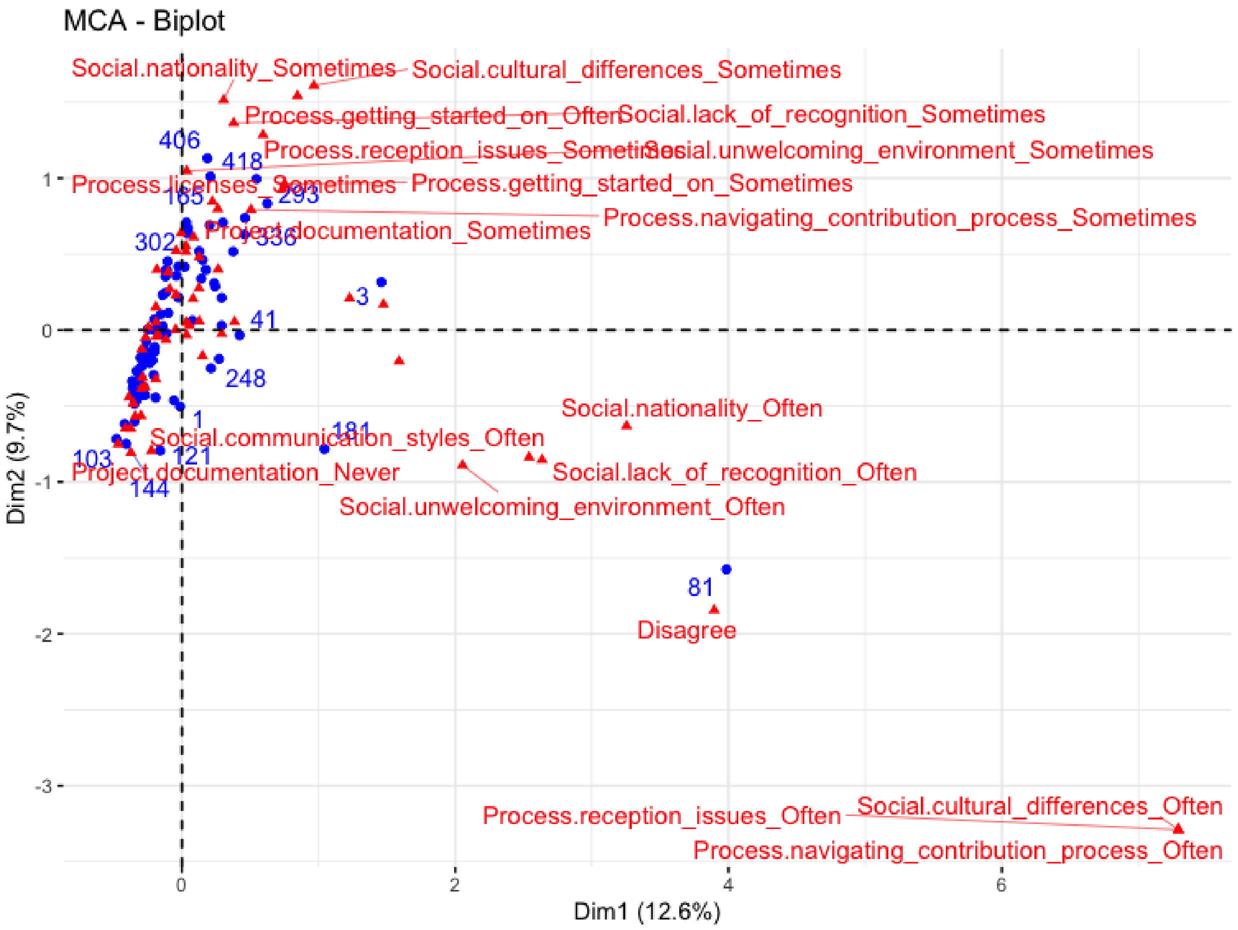


Scree plot: ([how to read scree plot]())

Rule of thumb: The point where the slope of the curve is leveling off (the “elbow”) indicates the number of factors that should be generated by the analysis. In this case, we see that after dimension 2, the trend tends to be flat. Thus we will use the first two dimensions as a starting point.

The next three plots will help us find variables we can use to train models in the future (but these plots are only used for us to do a preliminary analysis and do not include in the paper or report).

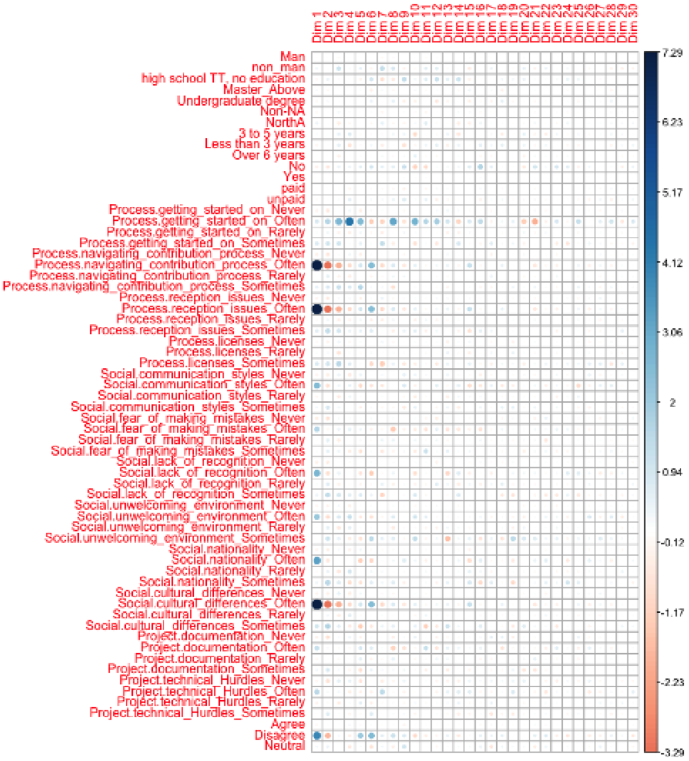
```
Biplot
#fviz_mca_biplot (res. mca, repel = TRUE, # Avoid text overlapping (slow if
many point) ggtheme = theme_minimal())
```



The plot above shows a global pattern within the data. Rows (individuals) are represented by blue points, and columns (variable categories) by red triangles. The distance between row or column points measures their similarity (or dissimilarity). Row points with similar profiles are closed on the factor map. The same holds for column points. The 1st dimension is strongly correlated with **Disagree SOB**; **Social. Cultural.Differences.often**; **Process.reception_issue.often**; **Process.navigating_contribution.often**. The 2nd dimension is strongly correlated with **Socia.Cultural.differences**; **Process.reception_issue**; **Process.navigating_contribution**.

We could also check it through CorrPlot and : corrplot(res.mca$var$coord, is.corr=FALSE)

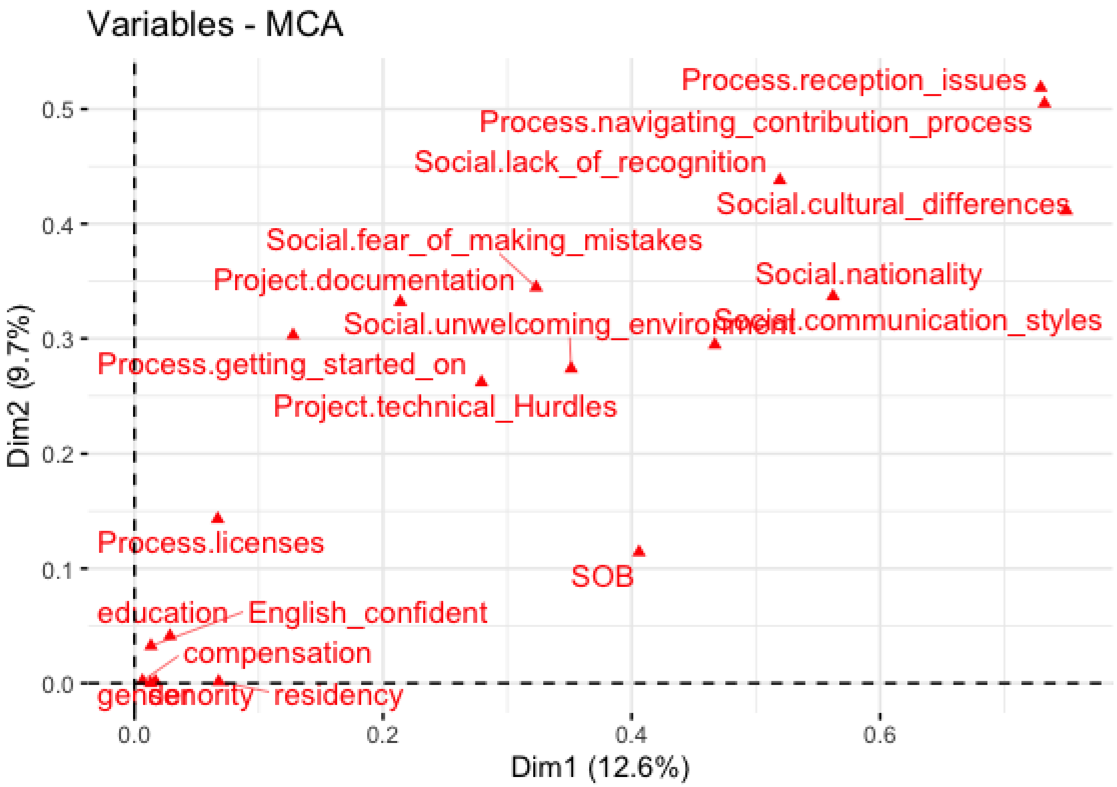
From the first dimension: Process.navigating_contribution.often; Process.reception_issues.often; Social.cultural.difference.often; Disagree with SOB



To visualize the correlation between variables and MCA principal dimensions:

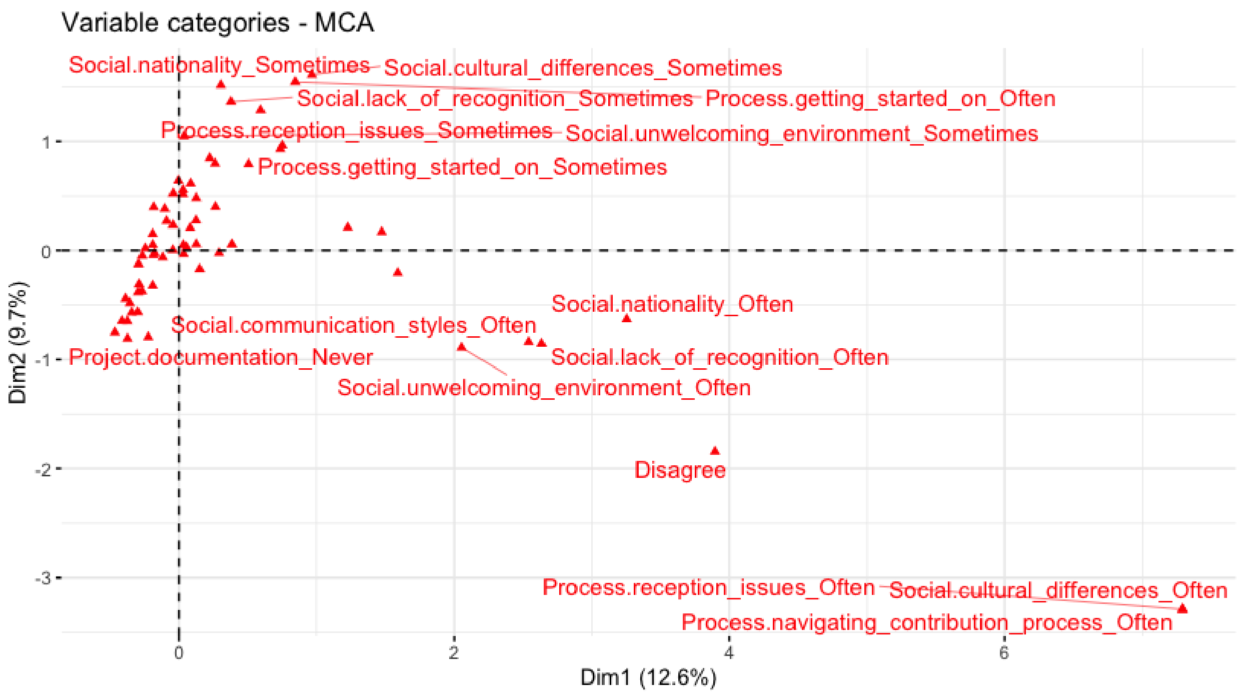
#Correlation between variables and principal dimensions

#fviz_mca_var(res.mca, choice = "mca.cor", repel = TRUE, ggtheme = theme_minimal())



- The plot above helps to identify variables that are the most correlated with each dimension. The squared correlations between variables and the dimensions are used as coordinates.
- It can be seen that the variables process.reception_issues, process.navigating_contribution_process, social.cultural_differences, social. nationality, social.communication_styles are the most correlated with dimensions 1 and 2.

```
fviz_mca_var(res.mca,
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
```



- Variable categories with a similar profile are grouped together.
- Negatively correlated variable categories are positioned opposite sides of the plot origin (opposed quadrants).
- The distance between category points and the origin measures the quality of the variable category on the factor map. Category points away from the origin are well represented on the factor map.

How to interpret: (image draw a line from the point to coordinate (0,0))

Length of the origin
Far => row is highly associated with some column

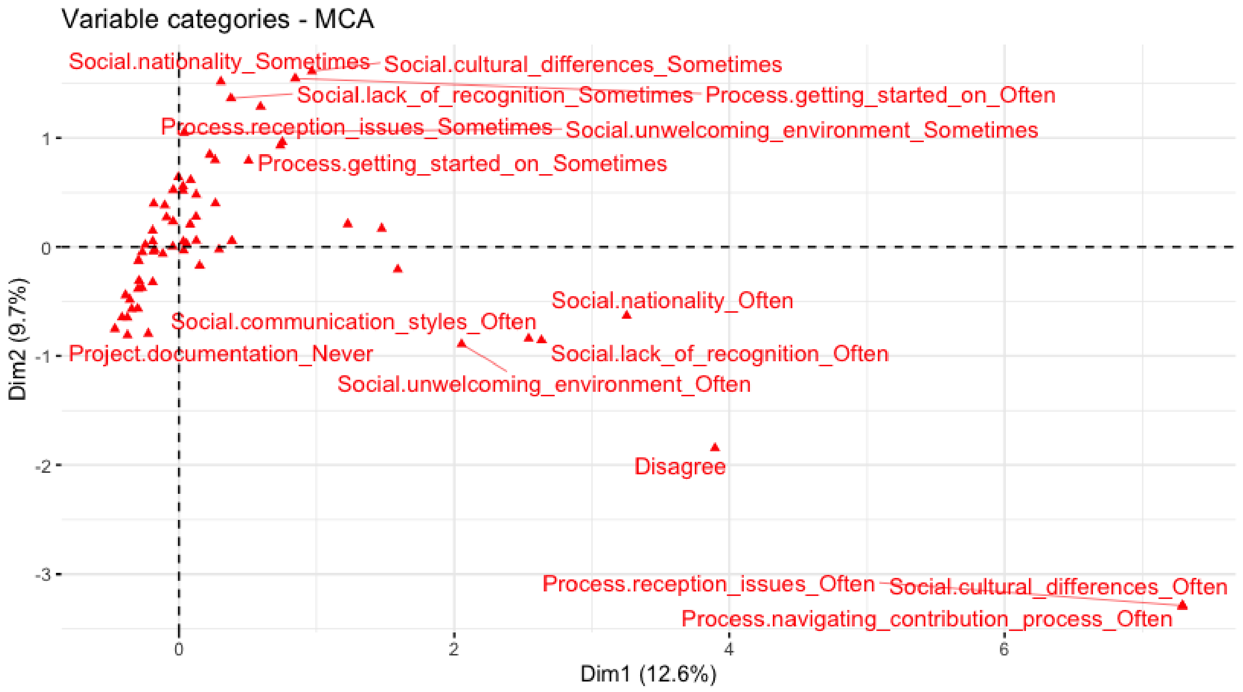
Angle between two variables:
Small angles -> association,
E.g., Process.reception_issue_often is highly associated with social.cultural_dufferences_often and process.navigating_contribution_process_Often

90-degree angles indicate no relationship
Nearly 180 degrees => negative association

In conclusion,

In our original datasets, we have 19 columns, after we applied MCA analysis, we noticed that these variables often pair for future analysis:

SOB (contributor who disagree with “I belong to the OSS”)
Challenges: Process.reception_issue_Often
Social.cultural_difference_Often
PRocess.navigating_contribution_process_Often
Social.cultural_differences
social.nationality



Note: this dataset is not designed to show the performance of MCA/MFA analysis; for more examples, please go to the following resources (they used a designed dataset to show the “ideal” case)

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/

https://www.youtube.com/watch?v=3znbLwWlzao&t=387s

https://www.youtube.com/watch?v=0z30RDikYaw&t=800s

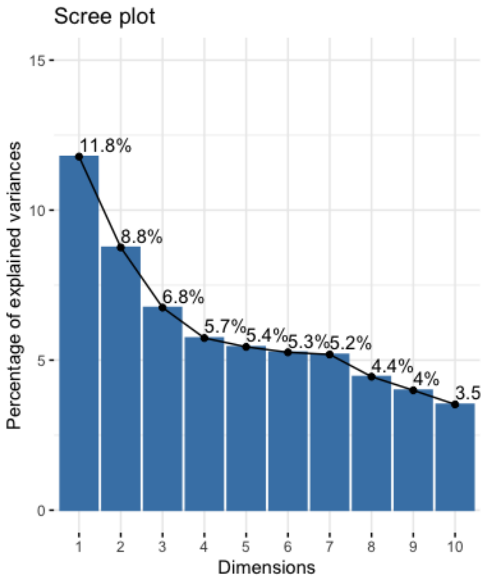
MFA analysis:

sample spreadsheet (SOB.CSV), we have 19 columns:

Demographic info:
Gender: Man, Non-man
Education: high school, undergraduate, master and above
Residency: North America, Non-North America
Seniority: less than 3 years, 3-5 years, over 5 years
English_confident: yes, no
compensation: paid, unpaid
Challenges (how often do you face the following challenges in OSS):
Process.getting_started_on: Never, Rarely, Sometimes, Often
Process.navigating_contribution_process: Never, Rarely, Sometimes, Often
Process.reception_issues: Never, Rarely, Sometimes, Often
Process. licenses: Never, Rarely, Sometimes, Often
Social.communication_styles: Never, Rarely, Sometimes, Often
Social.fear_of_making_mistakes: Never, Rarely, Sometimes, Often
Social.lack_of_recognition: Never, Rarely, Sometimes, Often
Social.unwelcoming_environment: Never, Rarely, Sometimes, Often
Social. nationality: Never, Rarely, Sometimes, Often
Social.cultural_differences: Never, Rarely, Sometimes, Often
Project. documentation: Never, Rarely, Sometimes, Often
Project.technical_Hurdles: Never, Rarely, Sometimes, Often
SOB (I feel I belong to OSS): Agree, neutral, Disagree

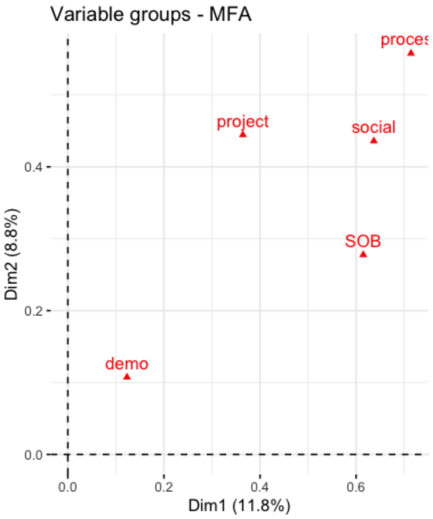
We split the information into 5 groups: demographic, challenge.process, challenges.social, challenges.project, and SOB.

```
res.mfa <- MFA(data, group = c(6,4,6,2,1), type = c("n",rep("n",9)),ncp=5,
name.group = c("demo", "process",  "social", "project", "SOB"),
num.group.sup = c(10), graph = FALSE)
```



Scree plot: (how to read scree plot)
Rule of thumb: The point where the slope of the curve is leveling off (the "elbow") indicates the number of factors that should be generated by the analysis. In this case, we see that after dimension 2, the trend tends to be flat. Thus we will use the first two dimensions as a starting point.

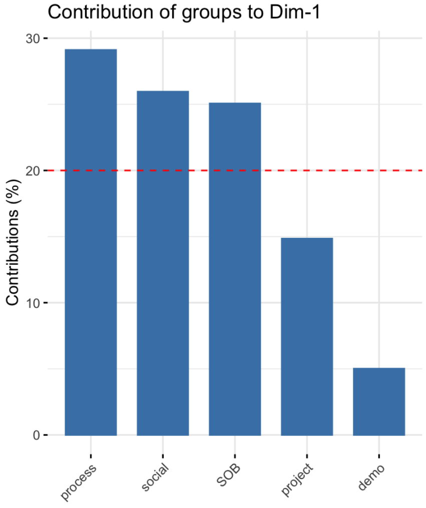
The plot shows the groups of variables:
#fviz_mfa_var(res.mfa, "group")

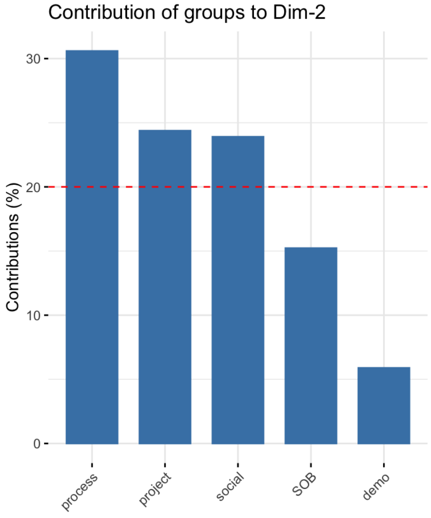


The plot above illustrates the correlation between groups and dimensions. The coordinates of the three groups on the first dimension are almost identical (process challenges, social challenges, and SOB). This means that they contribute similarly to the first dimension. Concerning the second dimension, the three groups: project challenges, social challenges, and process challenges.

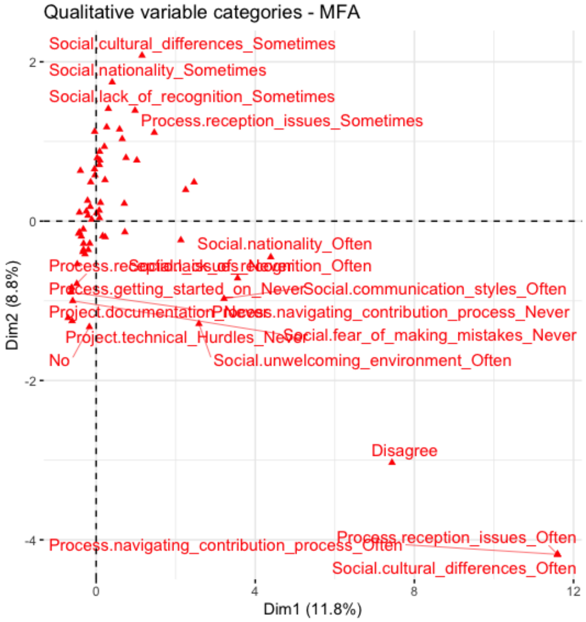
We could also see similar trends in contribution plots for dimensions 1 and 2

```
#fviz_contrib(res.mfa, "group", axes = 1)
#fviz_contrib(res.mfa, "group", axes = 1)
```





Thus, process challenges, social challenges, project challenges, and SOB are the primary groups. We will plot the same variable biplot as MCA, we see the similar trends from MCA analysis.



For more information please refer to this link: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/116-mfa-multiple-factor-analysis-in-r-essentials/