# Read Me

**Import of file**
We import the data (raw_data.xlsx) to R. At this point, we have 353 observations and 31 variables in the dataset. The non-binary test results are in the file (Mind_the_gap_non_binary.xlsx). We have already removed the noise entries from the file.

**Data Cleaning**
In the second stage, we clean the data

- **Education:** Since there are different ways of how "Associates Degree" is spelled, we first make all entries related to education consistent.
- **Age:** Since there are not enough observation points for age, we create age clusters. For that, we merge the age groups 46-55 and 56-65 to one group (46-65) and name the variable "agecluster".
- **Involvement in software:** Since there are inconsistencies in the data about the time frame of the involvement in software, we make the data more consistent by using the time frame 2-5 years in software instead of 3-5 years.

**Data Transformation**
We transform the Likert scale data in numbers:

- **Eye contact:** We create a new column (eye) for the transformed variable and convert the responses in an ordinal scale (Never = -2; Rarely = -1; Sometimes= 0; Very Often = 1; Always= 2)
- **Managing:** We create a new column (Manage) for the transformed variable and convert the responses in a binary scale (False = 0; True = 1).
- **Meeting interrupted:** We create a new column (meeting) for the transformed variable and convert the responses in an ordinal scale (Never = -2; Rarely = -1; Sometimes= 0; Very Often = 1; Always= 2)
- **Asked_question_to_another:** We create a new column (question) for the transformed variable and convert the responses in an ordinal scale (Never = -2; Rarely = -1; Sometimes= 0; Very Often = 1; Always= 2)
- **Presented_idea:** We create a new column (idea) for the transformed variable and convert the responses in an ordinal scale (Never = -2; Rarely = -1; Sometimes= 0; Very Often = 1; Always= 2)
- **Menial_tasks:** We create a new column (tasks) for the transformed variable and convert the responses in an ordinal scale (Never = -2; Rarely = -1; Sometimes= 0; Very Often = 1; Always= 2)
- **Make_necessary_decisions:** We create a new column (decisions) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -2; Disagree = -1; Neutral = 0; Agree = 1; Strongly agree= 2)
- **Team_values:** We create a new column (decisions) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -2; Disagree = -1; Neutral = 0; Agree = 1; Strongly agree= 2)
- **Growth_opportunities:** We create a new column (growth) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -1; Disagree

= -1; Neutral = 0; Agree = 1; Strongly agree= 2). We merged the answer options “strongly disagree” and “disagree” because of the scarcity of date for this variable.

- **Fairly compensated:** We create a new column (compensate) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -2; Disagree = -1; Neutral = 0; Agree = 1; Strongly agree= 2)
- **Start_over:** We create a new column (start) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -2; Disagree = -1; Neutral = 0; Agree = 1; Strongly agree= 2)
- **Aggressive_or_bossy:** We create a new column (aggressive) for the transformed variable and convert the responses in an ordinal scale (No, never = 0; 1 or 2 times = 1; 3 or 4 times = 2; More than 4 times = 3)
- **Harassment:** We create a new column (Harassment1) for the transformed variable and convert the responses in an ordinal scale (I haven't experienced or witnessed this behaviour = 0; I have witnessed this behaviour = 1; I have experienced this behaviour = 2; I have experienced and witnessed this behaviour = 3)
- **Sexism:** We create a new column (Sexism1) for the transformed variable and convert the responses in an ordinal scale (I haven't experienced or witnessed this behaviour = 0; I have witnessed this behaviour = 1; I have experienced this behaviour = 2; I have experienced and witnessed this behaviour = 3)
- **Exluded_from_social:** We create a new column (excluded) for the transformed variable and convert the responses in an ordinal scale (No, never = 0; 1 or 2 times = 1; 3 or 4 times = 2; More than 4 times = 3)
- **More_challenging_gender:** We create a new column (challenge) for the transformed variable and convert the responses in an ordinal scale (Strongly disagree = -2; Disagree = -1; Neutral = 0; Agree = 1; Strongly agree= 2)
- **Gender:** We create a new column for gender (gender1) to be able to run the Mann-Whitney U Tests in the demographic section (Female = 0, Male = 1)
- **Years of experience in software:** We create a new column for years in the software industry (years) for the transformed variable and convert the responses in an ordinal scale (Less than 2 years= 1; 2-5 years= 2, 6-10 years = 3; 11-15 years = 4; 16-20 years= 5, 21+ years = 6)
- **Age:** We create a new column for age (age2) for the transformed variable and convert the responses in an ordinal scale (18-25= 1; 26-35 = 2; 36-45 = 3; 46-55 = 4; 56-65 = 5)
-

**Demographics**

- **Gender**: Check of share of each gender
- **Years of experiences in the software industry:** Check of share of each gender; Mann-Whitney-U- test to check if the influence of Gender on the years of experiences is significant.
- **Age:** Check of proportions per age, check of share per gender in each age group, Mann-Whitney-U- test to check if the influence of Gender on age is significant.
- **Country:** Check of proportions per age, check of share per gender for each country, Chi-Squared Test of Independence for country and gender

- **Education:** check of share of education per gender, Chi-Squared Test of Independence for gender and Education
- **Employment Status:** check of share of employment status
- **Role:** check of share of role per gender, Chi-Squared Test of Independence for gender and role

**Statistical Analysis:**

**Assumptions for ordinal logistic regression**

- The dependent variable is ordered.
- One or more of the independent variables are either continuous, categorical, or ordinal.
- No multi-collinearity.
- Proportional odds

**Research Question 1 Microinequities**

- **Q11:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q12:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q13:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q14:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since the age variable violates the partial odds assumption model, we apply the Partial Proportional Odds Model with reversed=TRUE to make it comparable to polr.
- **Q15:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since the independent variables violate the partial odds assumption model, we apply the Cumulative Logit Model with reversed=TRUE to make it comparable to polr.
- **Q19:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q20:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q16:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). However, since this variable is a binary variable, we cannot use an ordinal logistic regression. Therefore, we apply a general binomial logistic regression.

- **Q17:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q18:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q23:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q24:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since the independent variables violate the partial odds assumption model, we apply the Cumulative Logit Model with reversed=TRUE to make it comparable to polr.
- **Q25:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q26:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since the independent variables violate the partial odds assumption model, we apply the Cumulative Logit Model with reversed=TRUE to make it comparable to polr.
- **Q27:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since both are fulfilled, we apply the ordinal logistic regression.
- **Q28:** We get the share for each response of the variable. Then we test the assumptions of the ordinal logistic regression (no multi-collinearity and proportional odds). Since the independent variables violate the partial odds assumption model, we apply the Cumulative Logit Model with reversed=TRUE to make it comparable to polr.