In this part four main analysis will be done:

1- Whether the “most type of harassment experienced” is correlated to “conforming to gender norms” will be tested using a “chi-square” test.
2- The significance of the relationship between “most type of harassment” and “sex at birth” will be tested with a “chi-square” test.
3- A one-way Anova test between the different categories in the "most type of harassment experienced" and the "rate" variables will be applied.
4- As the final test, an Anova test between the "type of harassment" variable and "rate" variable will be applied.

In the above statements the following assumptions have been made:

The “most type of harassment experienced” is the response of the participants to the question “Which of the following have you experienced the most?”. They could only choose one option among “Catcalling and whistling”, “Shouting and scolding”, “Insulting, hurtful and derogatory comments & being laughed at”, and “other”. The “type of harassment” variable is the response of the participants to the question “Which of the following types of verbal harassment have you experienced in the last year?”. The participants could choose any options among “Catcalling and whistling”, “Shouting and cursing”, “Insulting, hurtful and derogatory comments & being laughed at”, and “other”. In other words, multiple answers were accepted. Furthermore, this variable does not refer to the “most” type of harassment but any kind of harassment that an individual has encountered. Finally, one mistake has been done and “Shouting and cursing” and “Shouting and scolding” represent the same category in the two questions.

The “rate” variable is a floating number between one to five (inclusive). This number is calculated for each participant through the following process: First, it should be noted that one mistake was made in the creation of the questionnaire which was only recognized after publishing. The issue is that some of the variables are from a scale of 1 - 5; however, other values are from 1 - 7. There is no possible way to diminish the impact this might have on our research given that when there are 7 options participants might not choose the higher or lower numbers compared to when there are only 5 options. Having said all that, to at least fade away this impact, all the variables representing a trait have been re-scaled to be from 1 - 5. Accordingly, in the female-trait variables, it is expected that AFAB have a higher score whereas in the male-trait variables AMAB are expected to have a higher score. Therefore, the values have been further modified that for each participant the variables can be interpreted in the same way regardless of their sex. In other words, for the male-trait variables (all in scale 1-5) a re-calculation has been done by subtracting each from 5. This way for any variable representing a trait, the AFAB should have a “high score” and the AMAB is expected to have a “low score.”

As the next step, the mean of all the variables representing a trait has been calculated per participant. This calculated mean is given as the name "rate" and represents the rate of how much an individual behaves and acts as a female (or conforms to the female gender). This number is from 1 - 5 and the higher the number the more the individual behaves more like "female" norms. It is expected that AFAB have a higher “rate” compared to AMAB. We further exclude all the trait values and just use the "rate" number for further analysis.

Finally, the “conforming to gender norms” variable is calculated through the following process: The dataset will be divided into two separate datasets; the division will be done based on the "sex by birth" variable. One dataset will be containing only the “female” and the other will have only the “male” participants. It is known that AFAB participants tend to have a higher “rate” and the more the rate. Therefore, for the AFAB dataset, we will say that the participants that have a rate less than (mean of rate - standard deviation of rate) do not conform (or in a more general sense, less conform) to their gender norms. Furthermore, as the rate variable refers to female traits, for the AMAB dataset we have said that if the rate is higher than (mean of rate + standard deviation of rate) they do not conform to their gender norms (or less conform to their gender norms). In the end, the two datasets AFAB and AMAB are combined to one dataset. Take note that two different formula was used for the participants based on whether they were male or female at birth. Doing this process, we now have a variable in our dataset that indicates whether a participant conforms to “their own gender norm” or not. This variable is named “conformity” and is 1 if the participant is devoted to conforming and is 0 otherwise. Analysis indicates that there are 23 AFAB not conforming to gender norms. 138 AFAB conform to gender norms. 6 AMAB do not conform to gender norms whereas 36 AMAB do conform to gender norms.

After introducing the tests and the variables used in the tests, in the final step, we will run the four mentioned hypothesis tests for further clarification.

4.1)

In the first subsection, we will test whether the most type of harassment experienced is correlated to conforming to gender norms. Each variable is a categorical variable; thus, a chi-square test can be used. The null hypothesis is the type of harassment mostly experienced does not have any correlation with conforming or not conforming to gender norms. The alternative hypothesis is the type of harassment mostly experienced differs between people conforming to gender norms and those not conforming to gender norms. In the following we have first created the contingency table:

most_type	| Catcalling and whistling	Insulting, hurtful and derogatory comments & being laughed at	Other	Shouting and scolding
conforming				
FALSE	4	16	6	1	2
TRUE	11	97	30	9	27

As can be seen, 15 people in total have not indicated their most type of harassment; 4 of whom do not conform to their gender norms and the rest do conform to their gender norm. 113 said catcalling was the most type of harassment; out of whom 16 do not conform to their gender norm. 36 said hurtful comments and insulting, where 6 did not conform to their gender norm. Furthermore, 10 said other reasons were the most type of harassment they experienced; out of whom 1 did not conform to their gender norm. Finally, 29 people said shouting and scolding, where 2 did not conform to their gender norm.

```
Number of cases in table: 203
Number of factors: 2
Test for independence of all factors:
        Chisq = 3.489, df = 4, p-value = 0.4795
```

Running the chi-square test shows that there are 203 observations and 2 factors. The chi-square value is 3.489, "degree of freedom" is 4 and the p-value is 0.4795. Moreover, the cramersV value is interpreted as a measure of the relative associativity between two variables. The value is from 0 - 1 and in practice, a Cramer's V of 0.10 normally provides a good minimum threshold for suggesting there is a relationship between the two variables. As seen this value is 0.13. All in all, as the p-value suggests the result is not significant with a significance level of 0.05 and thus, the null hypothesis cannot be rejected. In other words, there is no clear significance that the most type of harassment experienced has a correlation on conforming or not conforming to gender norm.

4.2)

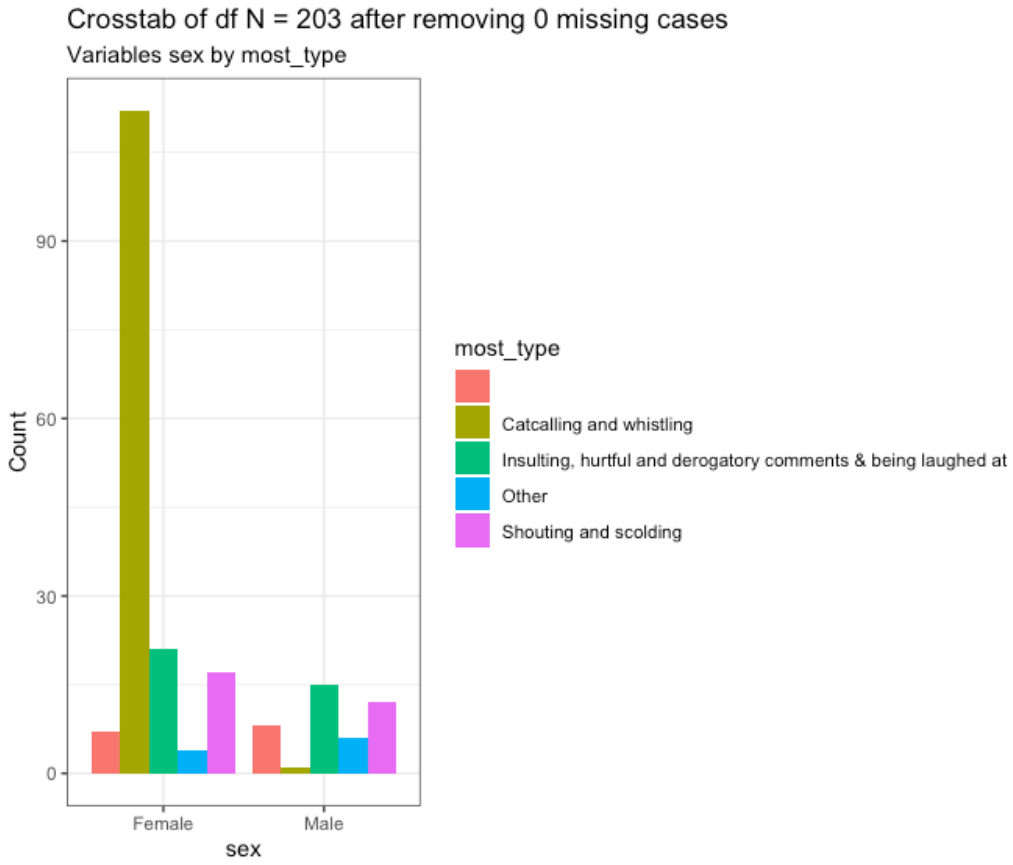
The same test will be run between sex at birth and most_type.
null hypothesis = there is no correlation between the most type of harassment experienced and the given sex at birth.
alternative hypothesis = there is a correlation between the most type of harassment experienced and the given sex at birth.

most_type	Catcalling and whistling	Insulting, hurtful and derogatory comments & being laughed at	Other	Shouting and scolding
sex				
Female	7	112	21	4	17
Male	8	1	15	6	12

```
Number of cases in table: 203
Number of factors: 2
Test for independence of all factors:
        Chisq = 63.39, df = 4, p-value = 5.622e-13
```

As it can be seen, respectively, 7, 112, 21, 4, and 17 AFAB said "nothing", "catcalling", "insulting", "other", and "shouting" are the most type of harassment experienced. Whereas these numbers for AMAB are 8, 1, 15, 6, 12 in order. Furthermore, the chi square score is 63.69, the degree of freedom is 4 and the p-value is 5.622e-13. And the Cramer's V is 0.55; therefore, there is a significant result between the variables. Specifically, the result is significant at $p < .05$. Therefore, we can reject the null hypothesis and conclude that there is a relationship between the sex given at birth and the most type of harassment experienced.

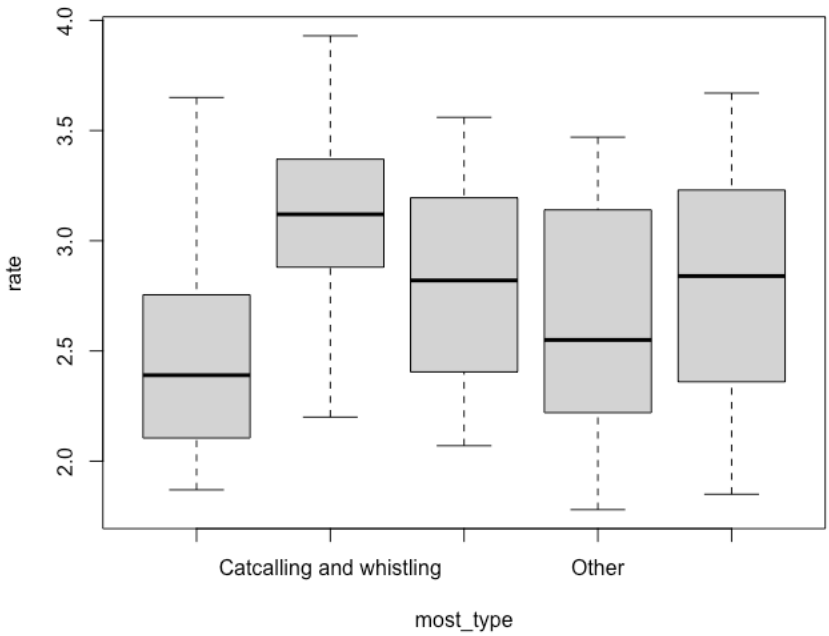
In a more specific statement, AFAB have experienced catcalling more than AMAB. Particularly, 55% of the total population are AFAB indicating catcalling their most type of harassment whereas only 0.04% of the population are AMAB indicating catcalling as their main type of harassment. Moreover, as the table shows, 69% of the AFAB indicated catcalling as their main type of harassment and 35% of the AMAB indicated insulting as their main type of harassment. The plot below is a nice summary of the correlation between most type of harassment experienced and sex at birth.



4.3)

We will now run a one-way Anova test between the different categories in the "most type of harassment experienced" and the "rate" column. Specifically, we will test whether the means of the rate variable differ between the categories or not.
null hypothesis = There is no difference in the means of the rate between the categories of most type of harassment experienced.
alternative hypothesis = There is a difference in at least the means of two categories.

The following boxplots will give us a decent overview before starting the test. Seeing the boxplots, we expect to reject the null hypothesis and say that there is a difference, but we will run the test to ensure this prediction.



Before running the One-way Anova test, the sample should have some underlying assumptions:

1- Variable type: There is one qualitative variable (most_type) and one quantitative variable (rate); so, this requirement is fulfilled.
2- Independence: In the design of the experiment, "snowball sampling" was used. This decreases the chances that the data is random. Therefore, the independency of the observations is not clear but for testing the data, it will be assumed that the sample is a decent reflection of the population.
3- Outliers: Looking at the box plots, there are no significant outliers.
4- Equality of variances: We must check the equality of the variances of the categories. This can informally be seen from the above box plots. However, formally a Levene test will be used to ensure the homogeneity of the variances. The variances of the different groups should be equal if we want to run an Anova test. This is called homogeneity of the variances (homoscedasticity) as opposed to heteroscedasticity where the variances differ across groups.
   In the "Levene" test, the Null Hypothesis is: “All populations variances are equal.” The Alternative Hypothesis is “At least two of them differ.”

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   4  7.1447 2.153e-05
      198
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Running the Levene test, based on the results, the p-value (2.153e-05) is lower than the significance value (0.05). Thus, it can be concluded that the null hypothesis can be rejected and that the variances are not homogeneous.

5- Normality: The number of participants in each group is not higher than 30, in fact 15 people did not declare their most type of harassment. 10 said other and 29 said shouting. The only two categories with above 30 participants were catcalling with 113 participants declaring this as their most type of harassment and insulting with 36. Hence, it cannot be assumed that the data has normality, i.e., the residuals follow a normal distribution. Therefore, the normality will be tested using a Shapiro-Wilk test. In this test the null hypothesis is that the data has a normal distribution, and the alternative hypothesis is that it does not have normality.

```
        Shapiro-Wilk normality test

data:  df$rate
W = 0.97762, p-value = 0.002511
```

As seen in the results the p-value (0.002511) is lower than the significance level (0.05) and thus the null hypothesis that the data has normality is rejected. In other words, the data does not have a normal distribution with regards to the rate. Therefore, the data is nor normal nor has a homogeneous variance, thus, a one-way Anova test cannot be used. Instead, a Kruskal-Wallis test will be used which is a nonparametric test, so the normality assumption nor the homogeneity requirement is not required. However, the independence assumption must still hold. This method uses sample medians instead of sample means to compare groups.

```
        Kruskal-Wallis rank sum test

data:  rate by most_type
Kruskal-Wallis chi-squared = 28.39, df = 4, p-value = 1.04e-05
```

As derived from the test, Kruskal-Wallis chi-squared is 28.39 and the p-value is 1.04e-05. Therefore, the null hypothesis can be rejected, and it can be concluded that the rate variable (which represents a value between 1 - 5 and the higher the number the more the individual hold female traits) is correlated to which most-type of harassment the participant has declared. It is interesting that the Kruskal-Wallis test has shown correlation. However, doing the chi-square test on the variables "most-type" and "conformity" showed no correlation even though conformity was derived from "rate" by defining a threshold. The main reason for this difference is:

1- These two tests in the first sight show a test between two similar variables, but that is not the case. In section 4.1 a chi-square test was used between two variables, namely, "conformity" - which was given 1 if the individual conformed to their own gender and was given 0 otherwise - & most type of "harassment". In the Kruskal-Wallis test the first variable is the most type of harassment but the second is the scale of which the individual has female traits or conforms to female norms regardless of their own gender.

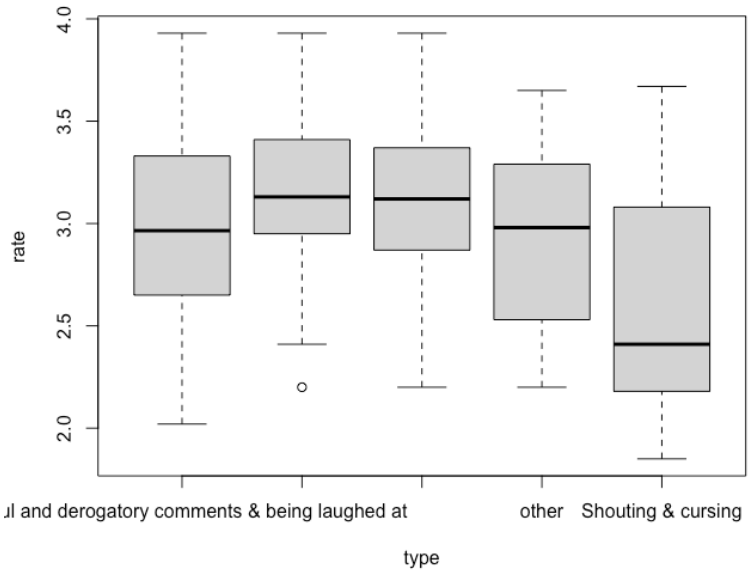
2- Also when setting a threshold, valuable data will be discarded, so, our test has lower precision in the chi-square test.

4.4)

As the final test, an Anova test between the "type" variable and "rate" variable will be applied. This is just as same as the "most-type" variable with the difference that here participants could have chosen “multiple” types of harassment they have experienced and not necessarily the most type of harassment they have experienced. Therefore, each participant might be calculated multiple times in the sum.
null hypothesis = the means of rate is the same for the all the categories of the types of harassment.
alternative hypothesis = there is a difference in the means of rate between at least two different categories of the types of harassment.

The first thing we must do is to tidy the data again so that each value in the "type" column only represents one type of harassment.

The mean and standard deviation of the rates for "insulting" are 2.94 and 0.48 in order. The mean and standard deviation of the rates for "catcalling" are 3.12 and 0.35 in order. The mean and standard deviation of the rates for "shouting" are 3.02 and 0.48 in order. The mean and standard deviation of the rates for "other" are 2.6 and 0.53 in order. The null hypothesis will be that the means are equal, and the alternative will be that there is a difference in the means of at least two different groups.



Again, there is one qualitative variable (type) and one quantitative variable (rate). Moreover, in the design of our experiment we used "snowball sampling" so therefore, this decreases our chance that the data was random. Therefore, the independency of our observations is not clear but for using Anova test we will assume that our sample is a decent reflection of our population. Looking at the box plots, there is only one significant outlier which can be neglected. Furthermore, we must check the variance of the categories. This can informally be seen from

the above box plots. However, formally we will again run the Levene test to ensure the homogeneity of the variances.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   4  6.4807 5.102e-05
      308
---
```

Running the Levene test, based on the results, we conclude that the null hypothesis can be rejected and that the variances are not homogeneous because the p-value (5.102e-05) is lower than the significance level 0.05. We again use the Kruskal-Wallis test.

```
        Kruskal-Wallis rank sum test

data:  rate by type
Kruskal-Wallis chi-squared = 48.415, df = 20, p-value = 0.0003719
```

As derived from the test, Kruskal-Wallis chi-squared is 48.415, the p-value is 0.0003719, therefore, we can reject our null hypothesis on a significance level of 0.05 and conclude that the rate variable is correlated to the type of harassment the participant has declared. In other words, there is at least two types which have a different means.

Finally, if you wish to further analyze the data, be careful of the values. The values were entered by the participants and not all have been tidied. Just to mentioned a few: To the question "What.gender.do.you.identify.as" the responses "Genderqueer" and "Gender queer" both can be seen whereas they represent the same thing. They are also people who have not given their gender. Moreover, for the age of the participants, there is a response "18 years" which is the same as "18". These variables have not been used for the current analysis and thus, their values have not been tidied.