The dataset was obtained from our defined questionnaire where 205 individuals participated. The questionnaire is now closed but can be found on the following website: questionnaire

For confidentiality, the emails of the participants have been removed from the original dataset. These emails were provided by the participants own will and it was provided from those who wanted to be informed about the result of the research.

Before starting, a few statistics is provided to get a brief overview of the dataset:

1- The data frame had 205 participants; out of which 161 represented themselves as female and 42 represented themselves as male. 2 individuals did not define their sex.

2- 148 participants have identified themselves as female gender. 41 had the gender male. For the genders non-binary, genderqueer, bigender, and agender, there were 9, 2, 1, and 1 participant respectively. 3 people preferred not to say their gender.

3- most participants are from 17-24 years of age with only three exceptions: one having 16 years of age, another 25 and the last 29 years old.

## ○ Data Wrangling

The name of each column of the data frame represents the question asked and each row is representative of each participant. The values are the answers of each participant to the given questions. The first step will be to rename the columns as they have very lengthy names: (it is worth mentioning that not all variables are used for analysis, let alone we will tidy all the data frame column names for any further analysis in the future). The column names have changed as following:

These variables are the general variables.

timestamp = Tijdstempel,
age = How.old.are.you.,
residence = Where.do.you.live.,
duration = How.long.have.you.been.living.in.the.Netherlands.,
frequency.of.nighout = How.often.do.you.go.out.at.night...to.the.city.center.to.party.or.to.bars.or.pubs.etc..,
sex = What.sex.were.you.assigned.at.birth.,
gender = What.gender.do.you.identify.as.,
gender.sex.equality = Does.your.gender.identity.align.with.your.sex.assigned.at.birth.,

These variables define the frequency of harassment.

total = How.many.times.have.you.been.verbally.harassed.in.the.last.year..This.includes.being.catcalled...whistled.at..being.shouted...cursed.at..people.making.insulting.and.hurtful.comments...being.laughed.at.,
catcall = How.many.times.in.the.last.year.have.you.experienced.the.following..been.catcalled.and.whistled.at.,
shouted = How.many.times.in.the.last.year.have.you.experienced.the.following..being.shouted.and.cursed.at.,

insult = How.many.times.in.the.last.year.have.you.experienced.the.following..people.making.insulting..hurtful.and.derogatory.comments...being.laughed.at.,

These variables define the type of harassment each participant has experienced.

type = Which.of.the.following.types.of.verbal.harassment.have.you.experienced.in.the.last.year.,
most_type = Which.of.the.following.have.you.experienced.the.most.,
duration = How.long.have.you.been.living.in.the.Netherlands.,
night.out = How.often.do.you.go.out.at.night...to.the.city.center.to.party.or.to.bars.or.pubs.etc..,

The below values represent female traits. In other words, it is expected that females have a strong score regarding the following variables.

skirts = I.like.to.wear.dresses.and.skirts.,
loose = I.like.to.wear.loose.clothes.,
jewellry = I.like.to.wear.jewellery.,
makeup = I.like.to.wear.visible.make.up.,
colourful = I.like.to.wear.clothes.with.bright.colours.and.patterns.,
longHair = I.have.hair.until.over.my.shoulders.or.longer.,
graceful = I.have.a.graceful.appearance.,
docile = I.am.docile.compliant.,
friendly = I.have.a.friendly.appearance.,
soft = I.talk.softly.,
follower = I.am.a.follower.,
nurturing = I.am.a.nurturing.person.,
empathy = I.show.empathy.,
emotions = I.show.emotions.,
feminine = I.consider.myself.as....,
idealFeminine = Ideally..I.would.like.to.be....,
interest = Traditionally..my.interests.would.be.considered.as....,
belief = Traditionally..my.attitudes.and.beliefs.would.be.considered.as....,
behaviour = Traditionally..my.behaviour.would.be.considered.as....,
appearance = Traditionally..my.outer.appearance.would.be.considered.as....,

The below values represent male traits. In other words, it is expected that males have a strong score regarding the following variables.

facialHair = I.have.a.visible.amount.of.facial.hair.,
bodyHair = I.have.visible.body.hair.,
lowPitch = I.have.a.low.pitched.voice.,
strong = I.am.physically.strong.,
initiative = I.take.initiative.,
cursing = I.often.use.curse.words.,
noEmotions = I.don.t.show.my.emotions.,
dominance = I.assert.dominance.,
agressive = I.often.behave.in.an.aggressive.way.,

insensitive = I.am.insensitive.,
confident = I.am.confident.,

These variables explicitly ask whether the participant conforms to their gender norm or not regarding their appearance.

conform1 = I.try.not.dress.too.much.like.people.from.the.opposite.sex.,
conform2 = I.put.a.lot.of.effort.into.making.myself.look.like.others.who.were.assigned.the.same.sex.at.birth.,

These variables define the post effect of the harassment on individuals.

trauma = I.experienced.an.incident.that.often.brings.up.thoughts.that.remind.me.of.the.unpleasant.situation.,
scared = I.experienced.an.incident.after.which.I.was.scared.to.walk.alone.on.the.street.at.night.,
upset = I.experienced.an.incident.that.made.me.feel.agitated.and.upset..,
avoidance = I.experienced.an.incident.that.I.avoid.thinking.or.talking.about.,
dream = I.experienced.an.incident.that.I.dream.about.a.lot.,
bad.feeling = I.experienced.an.incident.that.makes.me.feel.bad.when.I.am.reminded.of.it.,
bad.mood = I.experienced.an.incident.that.to.this.day.puts.me.in.a.bad.mood.,
alert = I.experienced.an.incident.after.which.anything.puts.me.on.alert.and.scares.me.,

In this section of the data analysis, for simplicity we will be referring to these variables as the given names instead of the previous lengthy names.

I have excluded the column with the question: "How frequenty have you been harassed in the last year." given that it is very similar to the question: "How many times have you been verbally harassed in the last year. This. includes. being. catcalled... whistled at… being shouted... cursed at… people making insulting and hurtful comments... being laughed at." Furthermore, although the names of the columns are informative, you should still see the questionnaire and the answers for further clarification.

The main analysis done in this part will be analyzing the correlation between the "type of harassment" and "conformity to gender norms" per group of sex (male and female). In other words, I will analyze whether there is any correlation between these variables or not and whether the correlation is significant. I will also see that whether the confounding variables "frequency of going out at night" and "type of residency area" have any effects on the frequency of type of harassment or not.

Therefore, the following variables are not needed for this part of the analysis and will be excluded from the data frame:
**timestamp, age, duration, gender.sex.equality, gender, total, upset, scared, trauma, avoidance, bad.feeling, dream, bad.mood, alert, conform1, and conform2.**

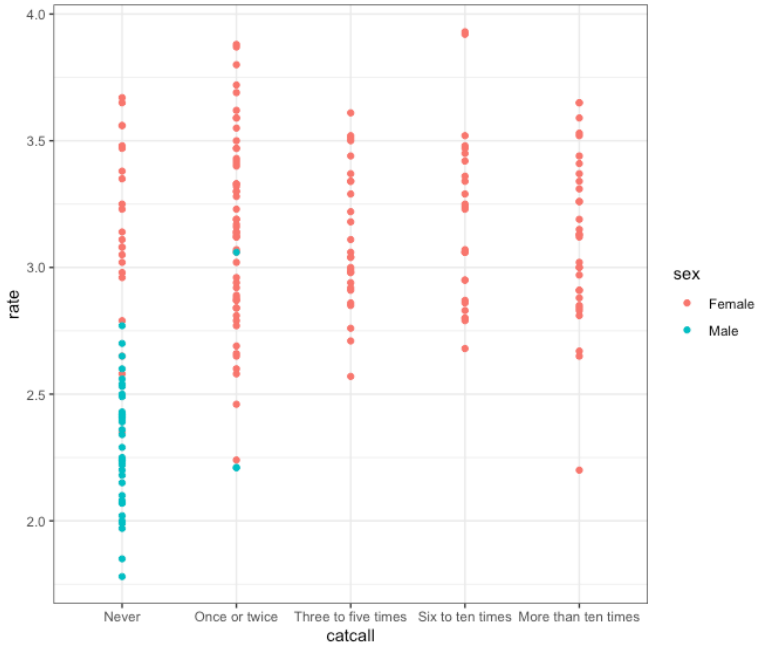
It would become clear as reading through the analysis why these variables are not of importance for this part of inferential statistics. It should be noted that one mistake was made in the creation

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, I have re-scaled all the variables to be from 1 - 5. Subsequently, because in the female-trait variables, female have a higher score whereas in the male-trait variables male have a higher score, I wanted to re-scale the variables that for each participant the variables can be interpreted in the same way regardless of their sex. Therefore, for the male-trait variables (all in scale 1- 5) I have re-calculated them by subtracting each from 5. This way for any variable representing a trait, the female should have a high score and the male is expected to have a low score.

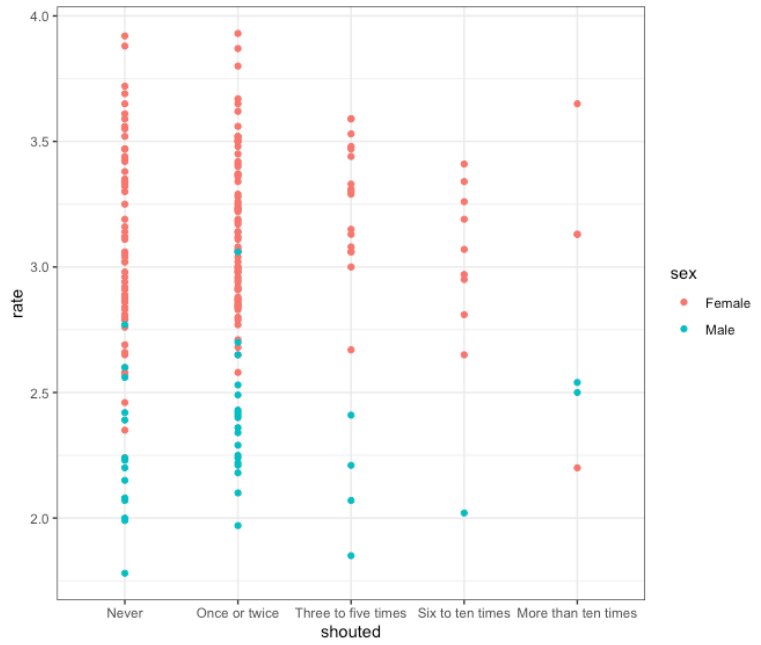
As the next step, I have added all the variables representing a trait and calculated the mean for each participant. This calculated mean is given as the name "rate" and represents the rate of how much an individual behaves and acts as a female (or conforms to the female gender). This number is from 1 - 5 and the higher the number the more the individual behaves more like "female" norms. We further exclude all the trait values and just use the "rate" number for further analysis.

- ## Plotting

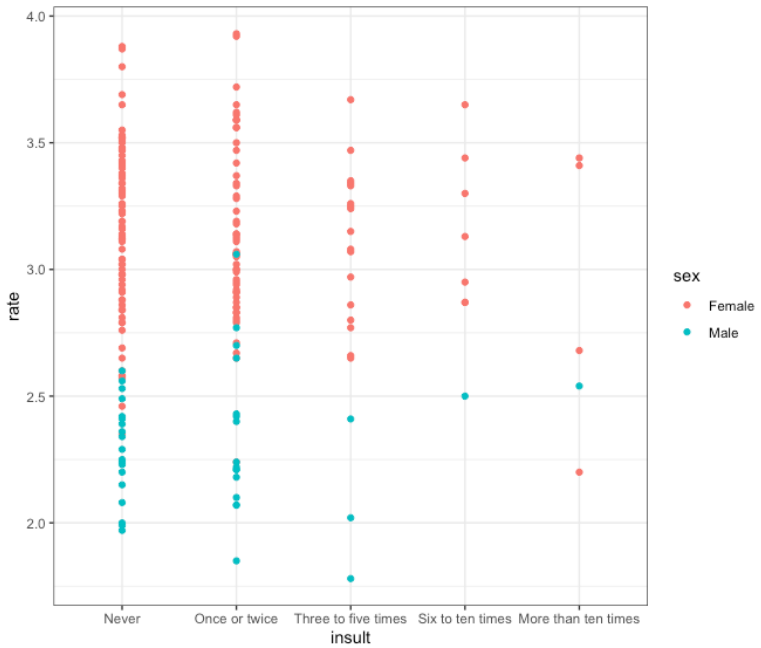
Let’s have a brief overview of how the data looks like with some plots before having run the statistical tests:
First, we will plot the frequency of each type of harassment against the rate of everyone. It should be noted, that in the below graphs, the individuals who have not declared anything have been exempted. The main reason is that there was an option "never" for each question and not indicating any answer cannot be counted as any of the categories. Moreover, the values of catcalling, insult, shouted, and night.out need to be re-leveled so that they appear in order in the plot - from the least frequent to the most frequent - from never to more than ten times.



As it can be seen, those who have the lowest rates tend to have never been catcalled. Specifically, roughly no individual with a rate lower than 2.25 has been catcalled. This shows that the less the "rate" (which shows that the individual conforms to the male gender norm more), the less the catcalling. Moreover, those who were assigned the "male" sex in birth, are shown with blue points. These individuals have mainly never been catcalled. However, is this enough to conclude that male is less catcalled than female? No, whether this is true or not will be shown in the following sections by running inferential statistics. As a final note, just for clarification if there is any confusion, the blue points show individuals who were born as male, the green ones are individuals who were born as female. The rate is the scale of the individual adhering to the female behavior; the lower the number the less the individual conforms to the female gender norm (in other words, they have more male-like attitude).

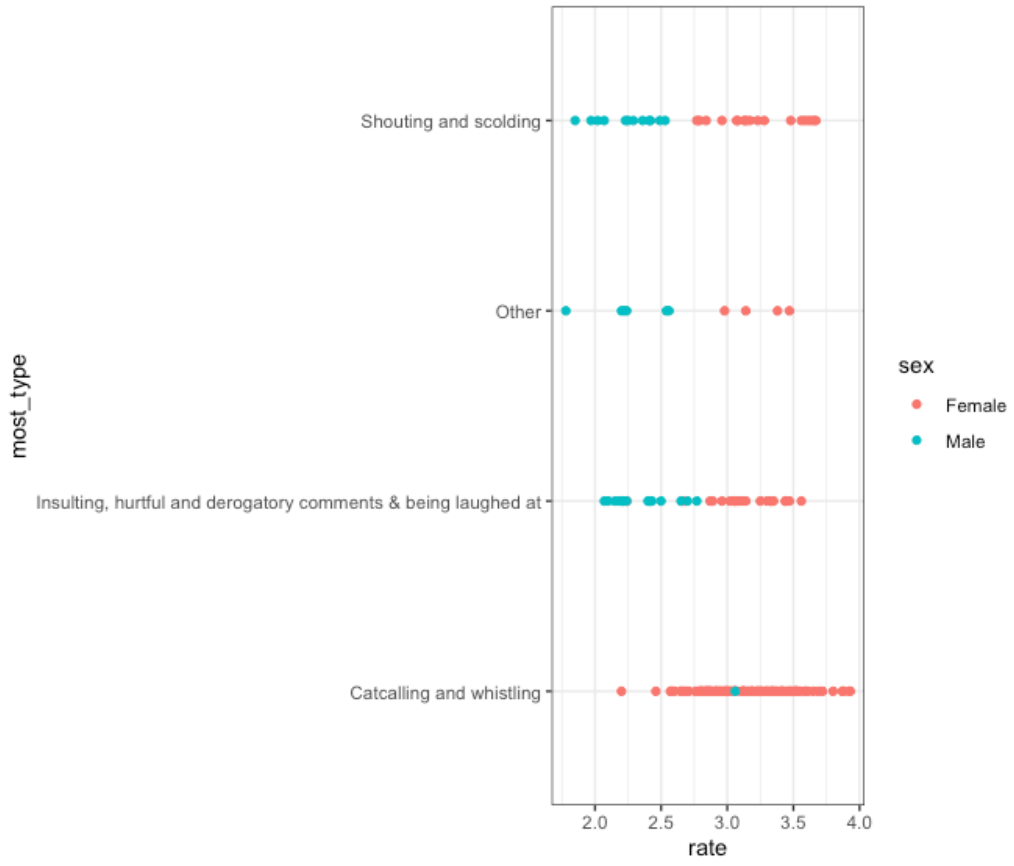



However, as seen in the above plot, for shouting there certainly is no clear pattern for either sex or rate.



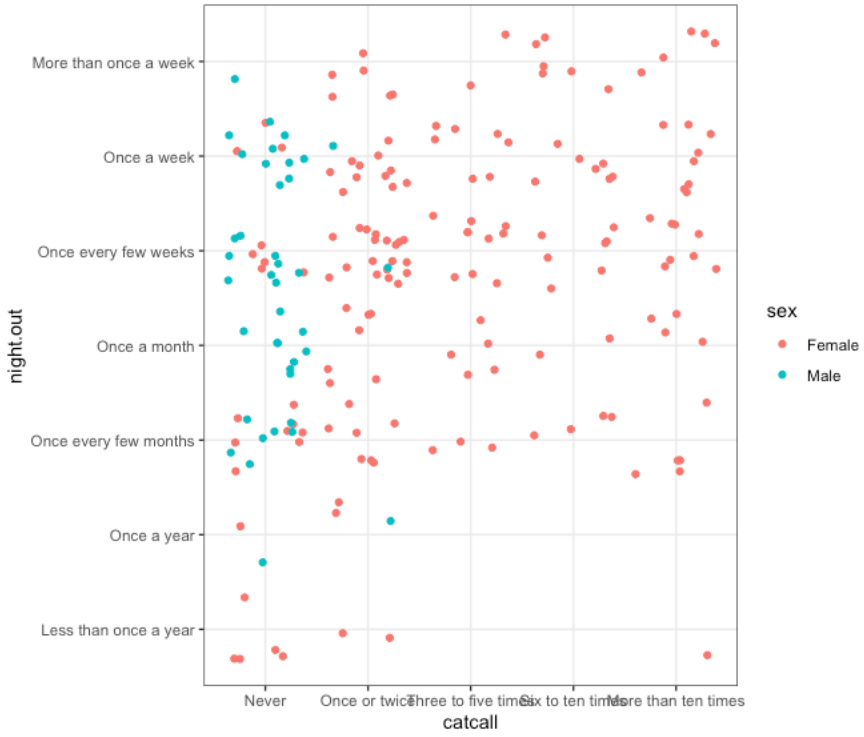
For insult, there is also no clear pattern as well. As a final note, as it can be seen in the above graphs, where the color of a point is blue, the rate is lower. Likewise, where the color is green, the rate is higher. This shows that most participants conform to their gender norms.

To recap, only for catcalling there could be seen that there is a clear pattern between male and female. In the following I have done the same, but this time, I have plotted the most type of harassment a participant has noted against their rate.



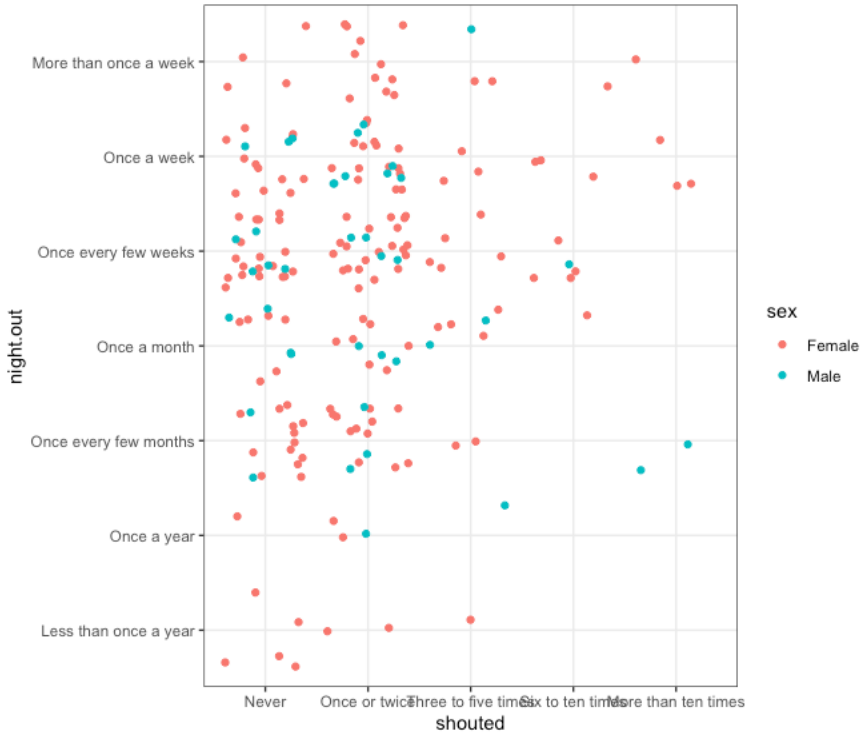

As it can be seen, the most type of harassment a participant has experienced, no male participant has noted catcalling whereas most females have said catcalling. Again, in the above graph, the individuals who have not declared anything have been exempted. The main reason is that there was an option "other" for each question and not indicating any answer cannot be taken into account.

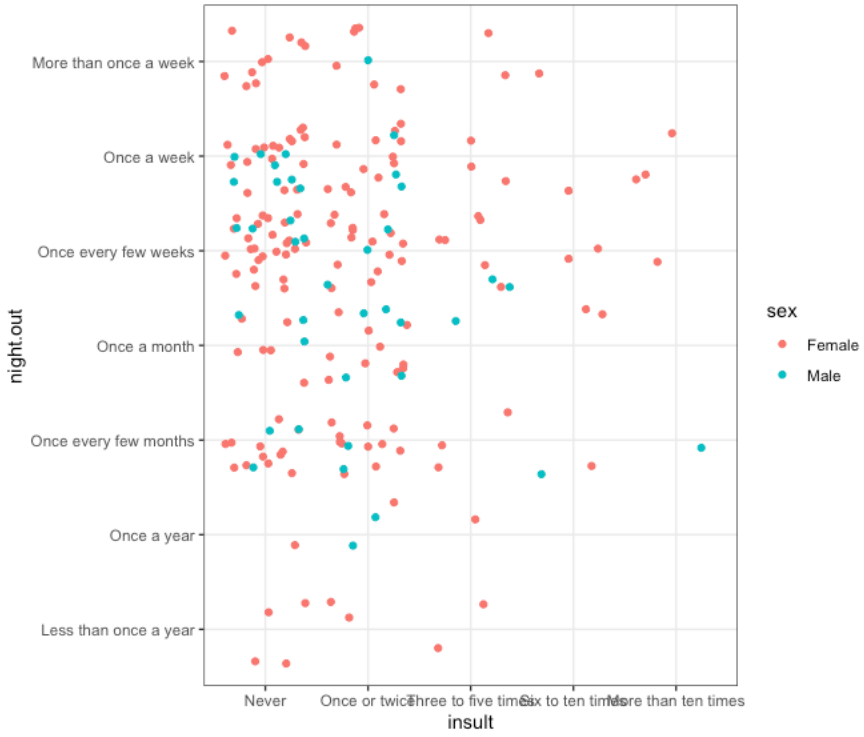
In the below plot, I have plotted the frequency of going out at night and being catcalled. I have used a jitter plot in this situation given that both variables are categorical and plotting them will make all points fall on the same location.




As seen above, again there is no clear indication that whether going out more leads to more catcalling.

Same as above we repeat the process for Shouted and Insulted type of harassment.






Again, there is no clear pattern for any of the above variables. To sum up, going out to the bar or generally going out more at night from the plots does not have a clear correlation with the increase of harassment. This statement stands true for any type of harassment - insult, shouted, and catcalled.

As a final plot, there was this one question that asked participants which kind of harassment they have experienced. Each participant could select multiple from the following options: 1- Insulting, hurtful and derogatory comments & being laughed at 2- Shouting & cursing 3- Catcalling & whistling 4- other. We want to see whether there is any pattern here. Take note that for analysis of this parameter we parsed each value of the variable "type of harassment" and had to check that whether it consisted of any of the above strings or not. Therefore, the process was repeated four times. From the results, 131 females have recorded they have experienced catcalled compared to only 3 males whereas one who noted they have experienced catcalling did not identify their sex of birth. 67 females have noted being shouted at compared to 15 males. 60 females said they have been insulted and laughed at whereas 20 males said so. Finally, 12 Female said they have experienced other kind of harassment whereas 15 males have said so.

- ## Checking conformity to gender norm

In this section, we will see how many of our participants have conformed to their gender norms. Specifically, their rate and their sex of birth will be compared. First, 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 female participants tend to have a higher rate and the more the rate, it shows that the participant holds more female traits (refer to the definition of rate). Therefore, for the female 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 male 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, we combine the two datasets female and male again into one dataset. Take note that two different formula was used for the participants based on whether they were male or female at birth.

Doing the above process, we now have a variable in our dataset that indicates whether a participant conforms to gender norms or not. This variable is named "conformity" and is 1 if the participant is devoted to conforming and is 0 otherwise. The final part of this section shows that there are 23 females not conforming to gender norms. 138 females conform to gender norms. 6 males do not conform to gender norms whereas 36 males do conform to gender norms.

## ○ Hypothesis test

Up till now we have wrangled, tidied, and modified the dataset. Moreover, we have explored plots and have got an overview of the relationships between the variables. In the final step, we will run hypothesis tests, specifically Chi-square test and Anova test 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 to gender norms or not conforming. 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)**

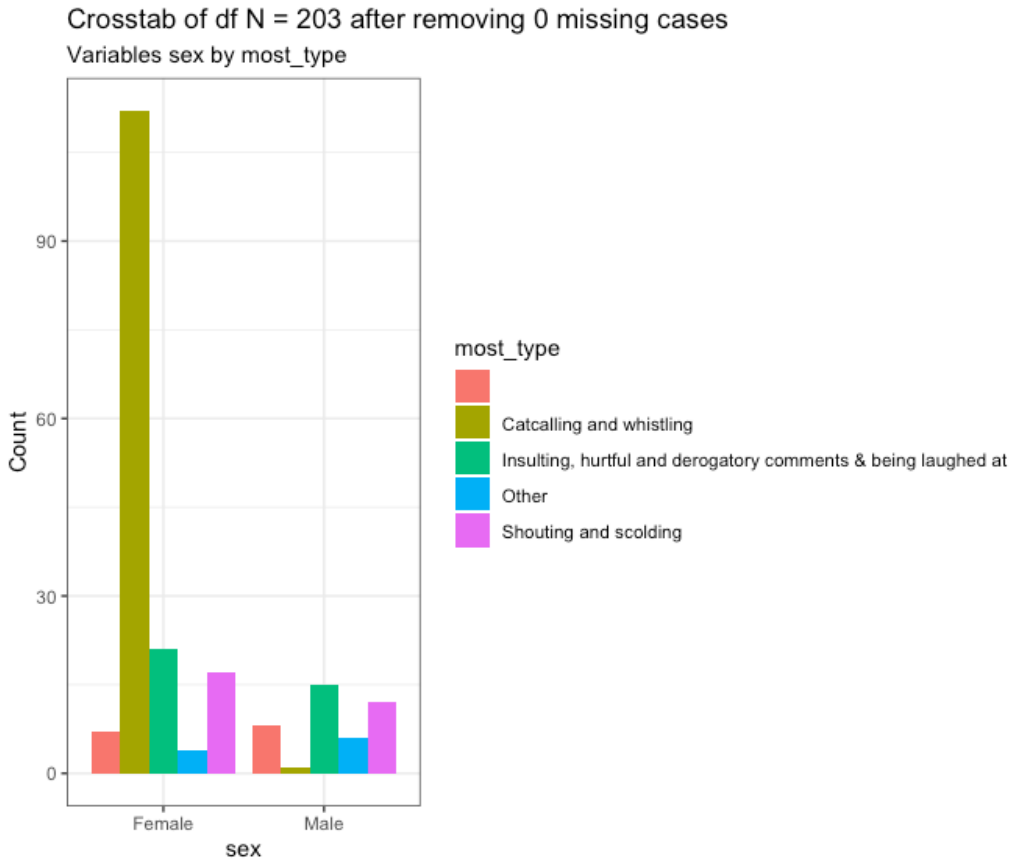
The same test will be run between sex at birth and most_type.
null hypothesis = there is no correlation between the most type of harassment experienced and the given sex at birth.
alternative hypothesis = there is a correlation between the most type of harassment experienced and the given sex at birth.

| 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 females at birth said "nothing", "catcalling", "insulting", "other", and "shouting" are the most type of harassment experienced. Whereas these numbers for male at birth 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.

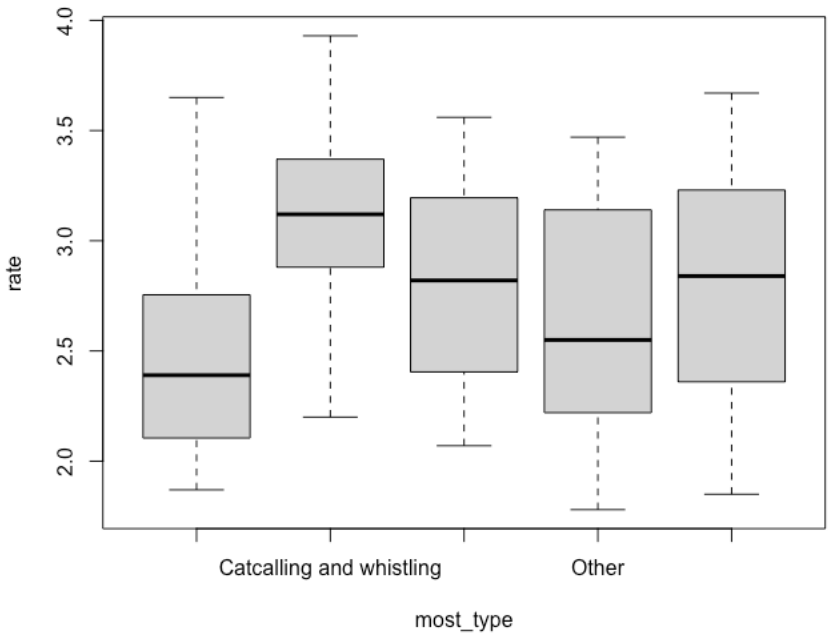
In a more specific statement, female at birth have experienced catcalling more than male at birth. Particularly, 55% of the total population are female indicating catcalling their most type of harassment whereas only 0.04% of the population are male indicating catcalling as their main type of harassment. Moreover, as the table shows, 69% of the female indicated catcalling as their main type of harassment and 35% of the male 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 (the rate to conforming to female traits on a scale of 1 to 5) 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.

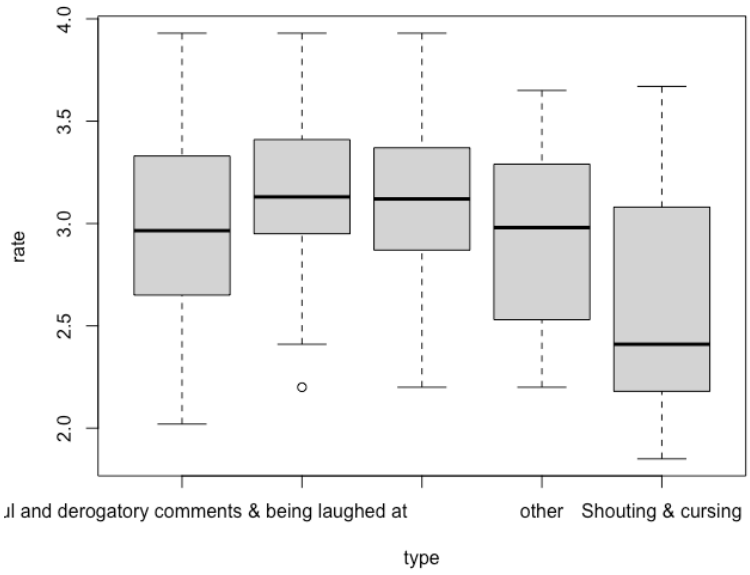
2- Also when setting a threshold, valuable data will be discarded, so, our test has lower precision in the chi-square test.

**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.