# QML - Formative Assessment 1

YOUR EXAM NUMBER

2023-10-11

## 1 Instructions

**PLEASE READ CAREFULLY**

**DUE Week 5 - Thu 19 October at noon**

You must include your **exam number as the author** in the document preamble above.

You'll notice that line 10 of the preamble says `mainfont: DejaVu Sans`. We would appreciate it if you'd use this font (or at least some other sans-serif font). **Try to render this Rmd file now, before making any changes, to see if you have this font installed.**

- If you get an error message saying that DejaVu Sans could not be found, you can download it here: http://sourceforge.net/projects/dejavu/files/dejavu/2.37/dejavu-fonts-ttf-2.37.zip.
- Then, install it appropriately for your operating system.
  - Here's a guide for Windows: https://www.digitaltrends.com/computing/how-to-install-fonts-in-windows-10/.
  - Here's a guide for Mac: https://support.apple.com/en-gb/HT201749.

**Do each exercise** by completing tasks, answering questions and/or providing code if required. Please **keep your written answers as concise as possible.**

Feel free to **add as many code chunks as you want** throughout.

**When you are ready to submit**:

1. Render the Rmd file to **PDF**.
2. **Rename** the PDF to your exam number only.
3. **Upload** the PDF file to Learn.

# 2 Exercises

## 2.1 Exercise 1: Creating plots

The next three exercises below require you to read in a particular file, filter/transform the data as needed, and create one or more plots that appropriately illustrate the described aspects of the data. Please also include a concise written description of the patterns you notice in each plot you make.

### 2.1.1 Exercise 1.1

Based on `data_e1_1.csv`: Plot speech rate against midpoint f0, colouring by condition and faceting by vowel. Describe what you see.

### 2.1.2 Exercise 1.2

Based on `data_e1_2.csv`: Create a single plot with logged reaction times by language, by environment, and by age (use any means). Describe what you see.

### 2.1.3 Exercise 1.3

Based on `data_e1_3.csv`: Plot the proportion of correct responses for each trial in the easy and difficult conditions, faceting by priming setting. Describe what you see.

## 2.2 Exercise 2: Critiquing and correcting plots

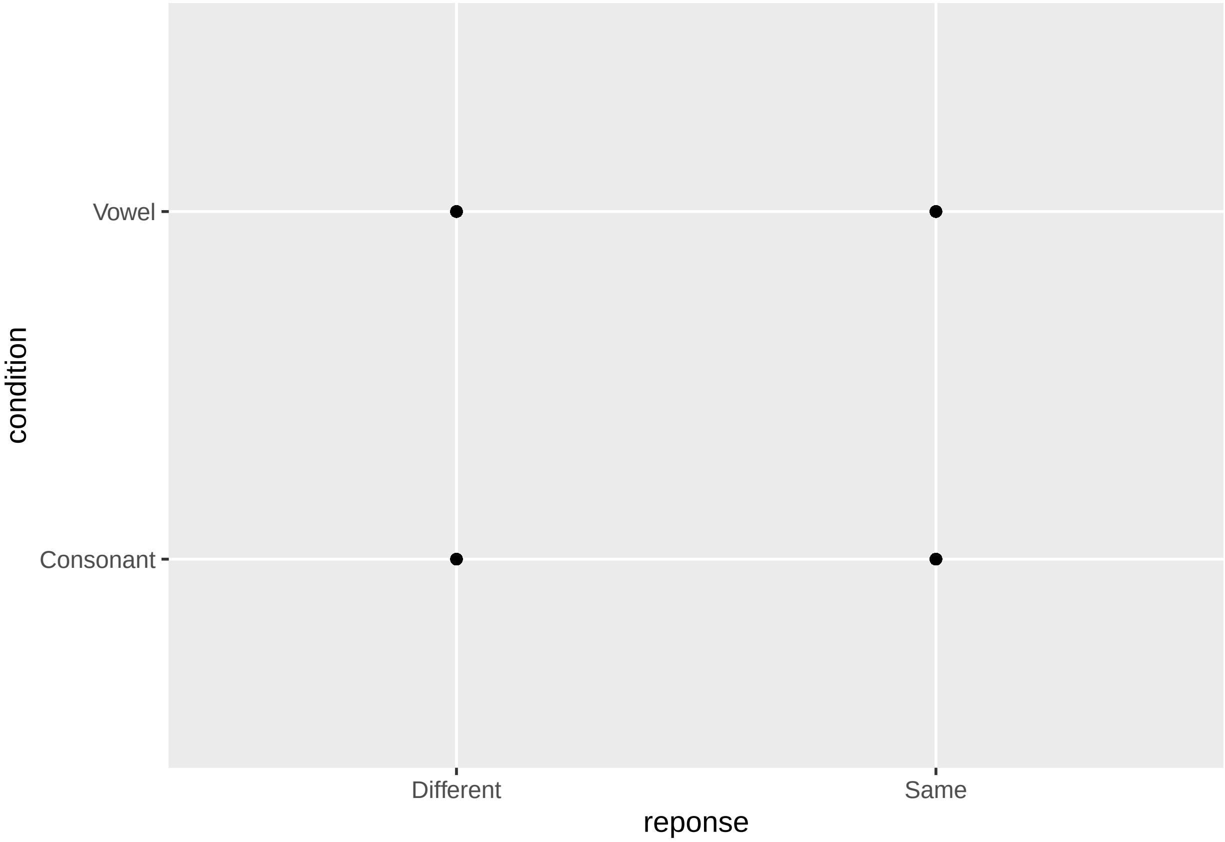
The following two plots are not appropriate for the type of data they show. Briefly describe what is wrong with each plot, try to figure out what the plots might be aiming to visualise, and write your own code to create a more appropriate plot for the exact same data. (If you're unsure about the kind of data you're dealing with, having a look at the data frame might help.)

### 2.2.1 Exercise 2.1

```
data_e2_1 <- read_csv("data/data_e2_1.csv")
```

```
## Rows: 600 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): reponse, condition
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_e2_1 %>%
  ggplot(aes(reponse, condition)) +
  geom_point()
```

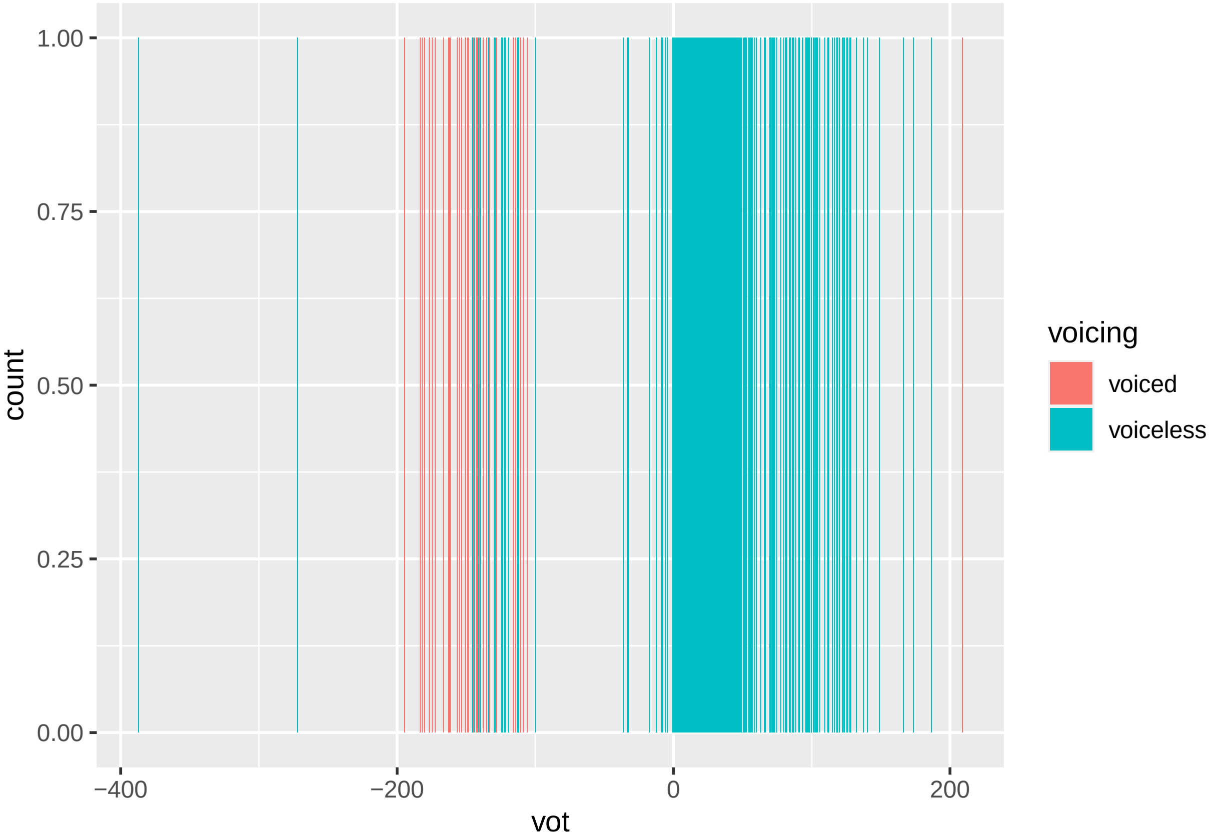


### 2.2.2 Exercise 2.2

```
data_e2_2 <- read_csv("data/data_e2_2.csv")
```

```
## Rows: 1035 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): voicing
## dbl (1): vot
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_e2_2 %>%
  ggplot(aes(vot, fill = voicing)) +
  geom_bar(width=0.5)
```



### 2.3 Exercise 3: Choosing appropriate summary measures

- Read in `data_e3.csv`
- Obtain summary measures (central tendency: mean, median, or mode; dispersion: standard deviation or range):
  - For each variable on their own.
  - For each variable in each group/accuracy combination.
- Make sure to pick the correct measure(s) for the respective variable type.
- Report all the measures in writing as you would in a paper.

## 2.4 Exercise 4: Identifying probability distributions

**This is a challenge exercise!** So far we have only looked at Gaussian distributions, but if you feel like you can learn about the log-normal and Bernoulli distributions and try to complete the exercise. The first summative assessment will have an exercise like this one (but we will have seen all the relevant probabilities by then).

For each variable in the table below, specify in the "Probability distribution" column whether it's (in principle) distributed according to a Gaussian, a log-normal, or a Bernoulli distribution, or according to some different one (put "other" in this case).

Alternatively, just try to identify variables that could be Gaussian and use "other" for the rest.

If you have doubts about any of the variables, you can write about it briefly below the table.

| | Variable | Probability distribution |
|---|---|---|
| 1 | Vowel duration (ms) | |
| 2 | Formant values (hz) | |
| 3 | Accuracy (binary) | |
| 4 | Readability (0-100) | |
| 5 | Reaction times (ms) | |
| 6 | Number of relative clauses | |
| 7 | Scots vs English | |
| 8 | Counts of infant gestures | |
| 9 | Logged reaction times | |
| 10 | Ratio of 1st vs non-1st pronouns | |

## 2.5 Exercise 5: Running a Bayesian linear model

Please, choose either 5a or 5b (but you are free to do both if you wish!).

### 2.5.1 Exercise 5a

The file `data_e5a.csv` contains a list of words from eight songs by Nina Simone their sentiment score (from -5 to +5, from most negative to most positive).

From Wikipedia, “Nina Simone was an American singer, songwriter, pianist, composer, arranger and civil rights activist. Her music spanned styles including classical, folk, gospel, blues, jazz, R&B, and pop.”

Read in the data, then fit a model with `brm()` with `value` as the outcome variable and answer this question: *What is the average sentiment in Nina Simone’s songs? Positive, negative or neutral?*

Is the data in a good shape to answer that question or do you think further processing like filtering might be needed?

Feel free to also create plots to explore the data.

**NOTE**: Warning messages from the MCMC sampling might arise, do not worry about these for the time being. They are in most cases innocuous.

### 2.5.2 Exercise 5b

**WARNING**: This exercise asks you to model data from the Harry Potter saga. We in no way support the transgender-phobic views of the saga’s author (for background, check this: https://www.vox.com/culture/22254435/harry-potter-tv-series-hbo-jk-rowling-transphobic).

The file `data_e5b.csv` contains a list of words from each book of the Harry Potter saga with their sentiment score (from -5 to +5, from most negative to most positive).

Read in the data and then create a new tibble containing only the words from Book 7.

Then fit a model with `brm()` with `value` as the outcome variable and answer this question: *What is the average sentiment in Book 7 of the HP saga? Positive, negative or neutral?*

Feel free to also create plots to explore the data (in its entirety and/or just for Book 7).

**NOTE**: Warning messages from the MCMC sampling might arise, do not worry about these for the time being. They are in most cases innocuous.