Analysis

In this document, we do the analysis presented in the paper.

Currently, the analysis uses fake data.

Setup

```
library(testthat)
```

Reading the data

```
ratings <- readr::read_csv("ratings.csv", show_col_types = FALSE)
n_ratings <- nrow(ratings)
```

There are 1000 ratings.

Analysis

Connecting the ratings to the formations:

```
songs <- dplyr::select(heyahmama::get_songs(), cd_title, song_title)
n_songs <- nrow(songs)
```

There are 270 songs.

```
cds <- dplyr::select(heyahmama::get_cds(), cd_title, formation)
n_cds <- nrow(cds)
n_formations <- length(unique(cds$formation))
```

There are 22 CDs.

```
songs_per_formation <- dplyr::select(merge(songs, cds), song_title, formation)
testthat::expect_equal(n_songs, nrow(songs_per_formation))
knitr::kable(head(songs_per_formation))
```

song_title	formation
10.000 luchtballonnen	3
Kusjessoldaten	3
Als het binnen regent	3
Jodelee	3
Kus van de juf	3
Jij bent de bom!	3

Add the formations to the ratings:

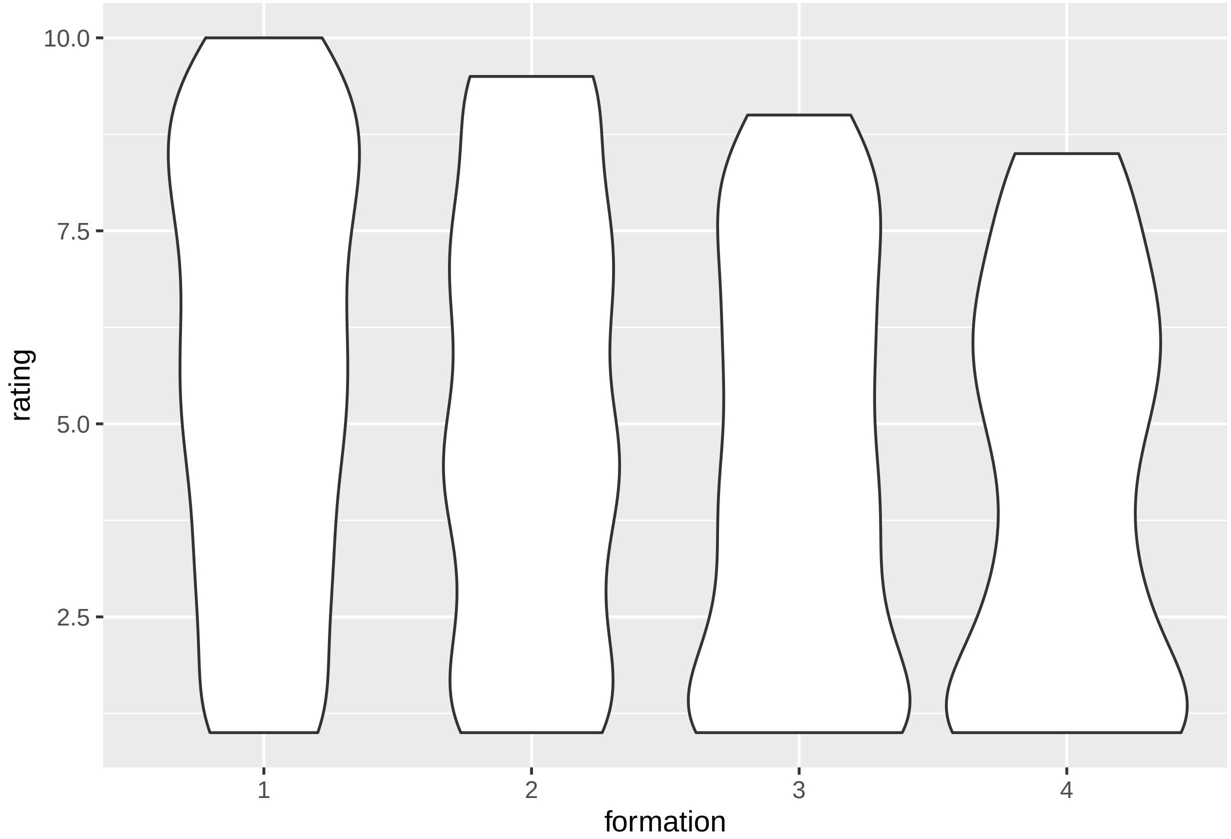
```
ratings_per_formation <- dplyr::select(merge(ratings, songs_per_formation), formation, rating)
testthat::expect_equal(n_ratings, nrow(ratings_per_formation))
```

```
ratings_per_formation$formation <- as.factor(ratings_per_formation$formation)
knitr::kable(head(ratings_per_formation))
```

formation	rating
1	9
1	1
1	7
1	2
3	1
3	1

Plot:

```
ggplot2::ggplot(
  ratings_per_formation,
  ggplot2::aes(x = formation, y = rating)
) + ggplot2::geom_violin()
```



Order formations by ratings:

```
average_rating_per_formation <-
  ratings_per_formation |>
  dplyr::group_by(formation) |>
  dplyr::summarise(average_rating = mean(rating))
testthat::expect_equal(n_formations, nrow(average_rating_per_formation))
```

```
ordered_average_rating_per_formation <-
  average_rating_per_formation |>
  dplyr::arrange(dplyr::desc(average_rating))
testthat::expect_equal(n_formations, nrow(ordered_average_rating_per_formation))

knitr::kable(ordered_average_rating_per_formation)
```

formation	average_rating
1	5.725441
2	5.169312
3	4.473684
4	4.192568

Statistics

Do the formations have different ratings?

```
n_combinations <- factorial(n_formations - 1)
```

There will be 6 comparisons.

```
alpha <- 0.05 / n_combinations
```

Due to 6 comparisons, the alpha value is (`0.05` divided by 6 equals) 0.0083333.

```
p_values_table <- tibble::tibble(
  a = rep(NA, n_combinations),
  b = NA,
  p_value = NA
)

i <- 1
for (lhs in seq(1, n_formations - 1)) {
  ratings_lhs <- ratings_per_formation[ratings_per_formation$formation == lhs, ]$rating
  for (rhs in seq(lhs + 1, n_formations)) {
    ratings_rhs <- ratings_per_formation[ratings_per_formation$formation == rhs, ]$rating
    p_value <- ks.test(ratings_lhs, ratings_rhs, alternative = "two.sided")$p.value
    testthat::expect_true(i >= 1)
    testthat::expect_true(i <= nrow(p_values_table))
    p_values_table$a[i] <- lhs
    p_values_table$b[i] <- rhs
    p_values_table$p_value[i] <- p_value
    i <- i + 1
  }
}
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
#> Warning in ks.test.default(ratings_lhs, ratings_rhs, alternative =
#> "two.sided"): p-value will be approximate in the presence of ties
p_values_table$is_the_same <- p_values_table$p_value > alpha
knitr::kable(p_values_table)
```

a	b	p_value	is_the_same
1	2	0.0033984	FALSE
1	3	0.0001015	FALSE
1	4	0.0000012	FALSE
2	3	0.0043610	FALSE
2	4	0.0885609	TRUE
3	4	0.1997498	TRUE