THE CYCLE-DEPENDENT CHOICE OF PARTNER IN WOMEN

Fuat Sarp Olcay, BA

Matriculation Number: 11933989
Study Code: 033 551
E-Mail Address: a11933989@unet.univie.ac.at
Supervisor: Dipl.-Ing. Dr. Theresa Scharl-Hirsch

Institute for Statistics
University of Vienna

CONTENTS

List of Figures.

1. Introduction. Some studies have already shown that women prefer more masculine faces in the fertile phase of their menstruation cycle (Klusmann and Berner, 2011). These are often linked to good genes, high reproductive probability and immune competence. Outside of the fertile period, more feminine facial features are of interest, and these are often associated with good character traits, which in turn can be beneficial for raising offspring. A clear connection was shown in men between the 2D:4D ("index finger length"/"ring finger length") ratio and the morphological features of the face. Higher 2D:4D values indicate lower testosterone levels during development and more feminine facial features, lower 2D:4D values indicate higher testosterone levels during development and more masculine facial features (Fink et al., 2005). The hypotheses, which are going to be tested in this study are;

- Women prefer more masculine faces and are more likely to look for physical characteristics while choosing their partner during their fertile phase.
- Women outside the fertile phase prefer more feminine faces and tend to look for character or material traits while choosing their partner.

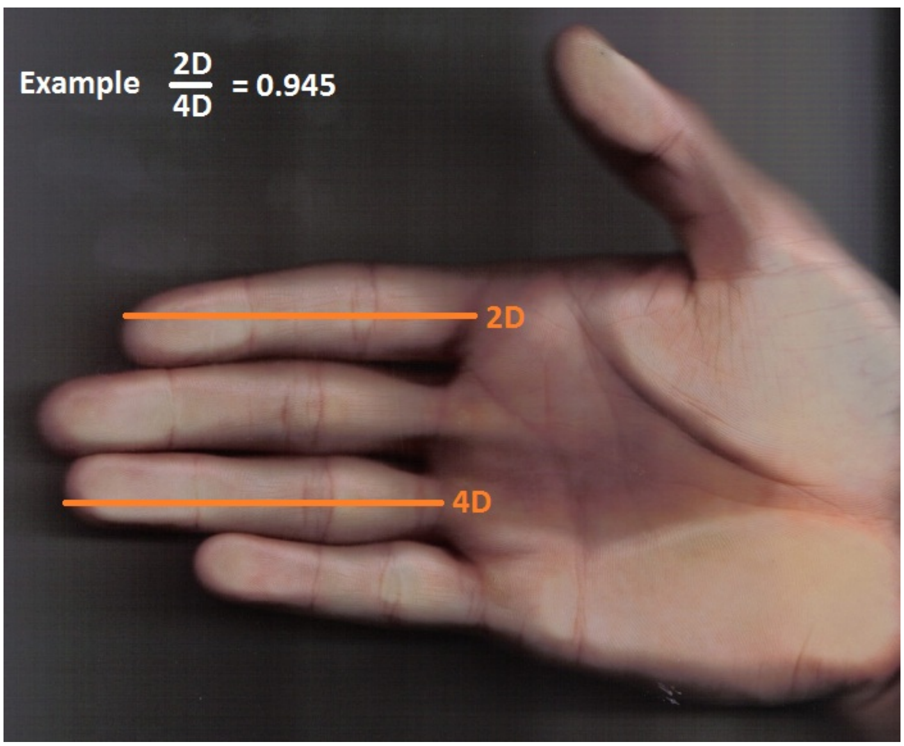

Fig 1. *An example of how the 2D:4D ratio is measured (Wikipedia contributors, 2022).*

2. Data. 63 women between the ages of 18 and 35 were interviewed. The group consisted of women who were not using hormonal birth control and who were in the fertile phase, women who were not using hormonal birth control and who were just outside the fertile phase, and women who were using hormonal birth control. First, the subjects were asked to rate 25 different male faces on a scale from 1 to 7 according to the criteria "attractiveness" (unattractive to attractive), "dominance" (submissive to dominant) and "friendliness" (unfriendly to friendly). The subjects were then asked to write a contact ad, in which they first described themselves and then the partner they were looking for. Finally, the demographic data "age", "relationship status" (single or in a relationship), "partner search" and "desired relationship type" (long-term relationship, short-term relationship, infidelity, "one-night stand") were collected on the one hand, and the necessary parameters on the other to calculate fertility (the first day of the last menstrual period, the average length of the cycle and the type of contraception). For each of the 25 facial exposures, there was an associated exposure of a person's hand. This made it possible to measure the index finger and ring finger and calculate the finger length ratio. The ratio obtained in this way (2D:4D) allows conclusions to be drawn about the amount of prenatal testosterone in the development phase of the 25 men shown. The study resulted with two data sets, which are: Face rating data set and Evaluation questionnaire data set.

2.1. *Face rating data set.* This data set contains 8 variables with 1575 observations. The variables are;

- ID: The identification number of the woman.
- Hormonal contraception: A binary variable, which indicates if the woman is using contraception(=1) or not (=0).
- Fertile: A binary variable, which indicates if the woman is fertile (=1) or not (=0).
- Face number: The identification number of the face, which is being rated by the women.
- 2D: The length of the index finger of the man, whose face is being rated by the women.
- 4D: The length of the ring finger of the man, whose face is being rated by the women.
- Ratio: The ratio between the length of the index finger and ring finger. It is being obtained by dividing the 2D and 4D values.
- Rating: The women's rating for the corresponding face. The variable is between 1 to 7 with 7 being the best rating possible.

2.2. *Evaluation questionnaire data set.* This data set contains 243 variables with 63 observations. The data set contains demographic data of the women, their preferred type of relationship, 104 binary variables about their own personality and 98 binary variables about the desired partner. As this study is about the differences of partner desires of the fertile and non-fertile women, the binary variable "fertility" and the other 98 binary variables about the desired partner are going to be used in the analysis.

3. Theory.

3.1. *Two-way ANOVA.* The two-way ANOVA is one of the most popular model in the *factorial experiments*. The definition of a factorial experiment is: An experiment that utilizes every combination of factor levels as treatments is called a factorial experiment (Cam and Neyman, 1967). In a factorial experiment with factor **A** at **a** levels and factor **B** at **b** levels, the model for the general model can be written as,

$$Y = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{3.1}$$

for $i = 1, 2,..., a$, $j = 1, 2,..., b$ and $k = 1, 2,..., r$, where Y is a continuous target variable, μ is the overall mean response, τ_i is the effect due to the i-th level of factor **A**, β_j is the effect due to the j-th level of factor **B** and γ_{ij} is the effect due to any interaction between the i-th level of **A** and the j-th level of **B**.
At this point, consider the levels of factor **A** and of factor **B** chosen for the experiment to be the only levels of interest to the experimenter. The factors **A** and **B** are said to be *fixed factors* and the model is a *fixed-effects model*.
When an $a \times b$ factorial experiment is conducted with an equal number of observations per treatment combination, the total (corrected) sum of squares is partitioned as:

$$SS(total) = SS(A) + SS(B) + SS(AB) + SSE, \tag{3.2}$$

where AB represents the interaction between **A** and **B**. For reference, the formulas for the sums of squares are:

$$SS(A) = rb\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \tag{3.3}$$

$$SS(B) = ra\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \tag{3.4}$$

$$SS(AB) = r\sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2 \tag{3.5}$$

$$SSE = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{ij\cdot})^2 \tag{3.6}$$

$$SS(TOTAL) = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2 \tag{3.7}$$

(Fujikoshi, 1993). The resulting ANOVA table for an $a \times b$ factorial experiment is:

Source	SS	df	Mean square
Factor A	SS(A)	(a - 1)	SS(A)/(a - 1)
Factor B	SS(B)	(b - 1)	SS(B) / (b - 1)
Interaction	SS(AB)	(a - 1) (b - 1)	SS(AB) / ((a - 1) (b - 1))
Error	SSE	(N - ab)	SSE / (N - ab)
Total(Corrected)	SS(Total)	(N - 1)	

3.2. *K-Means Clustering Algorithm.* K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori (Oti et al., 2021). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step (Bock, 2008). After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i - v_j\|)^2, \tag{3.8}$$

where x_i denotes the i-th data point of the set $X = x_1, ..., x_n$ and let v be the set of centers with $V = v_1, ..., v_c$. $\|x_i - v_j\|$ is the Euclidian distance between x_i and v_j, c_i is the number of data points in i-th cluster and c is the number of cluster centers (Ahmed, Seraj and Islam, 2020).

The k-means clustering algorithm functions as follows, (Cheung, 2003)

1. Randomly select c cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster center using (3.9)
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3).

$$v_i = (\frac{1}{c_i})\sum_{j=1}^{c_i} x_i \tag{3.9}$$

4. Methods. The software R was used for all computations and graphics (R Core Team, 2021a).

4.1. *Face rating data set.* With the command `read.csv()` (Wickham, Hester and Bryan, 2022) the data was read into R. Since the values in the columns 2D, 4D and ratio were separated by a comma instead of a dot, the function `gsub()` was used in order to replace the commas. The values were converted into type numeric with the help of the function `as.numeric()`. The column names were adjusted with the function `colnames()`.
First, the data was fit into a simple linear model with the interaction effect between *ratio* and *fertile*. To do this, the command `lm()` was used. Then the interaction plot was created using the function `interaction.plot()`.
Since the variable ratio was type numeric, it should have been converted into a character type variable in order to be able to use ANOVA. To do this, the variable ratio was converted into a character type variable named *ratio_group* with 3 levels; *low, medium* and *high*. Then the data was fit into an ANOVA model with the interaction effect above using the command `aov()`. Another interaction plot between the grouped variable and the other 2 variables were plotted using the command `interaction.plot()`.

4.2. *Evaluation questionnaire data set.* With the command `read.csv()` the data was read into R. Since the study is about the partner desires of fertile and non-fertile women, the binary variables for the preferred type of partner were separated with the identification number of the women into another data set. Since the column names were originally in german, R was not able to read all column names properly. With the command `setnames()` from the package *data.table* (Dowle and Srinivasan, 2021) the column names were corrected. With the help of the other data set in the previous section, the variable *fertile* was merged into the Evaluation Questionnaire set, so that the data set contains the "targeted" variable.
Some variables in the data set contained low amount of 1's and since this can be redundant for the analysis, the variables with less than 5 entries were dropped out of the data set. The sums were calculated with the command `colSums()`. The resulting data set contained 27 variables.

4.3. *K-Means Clustering.* Before the data was clustered, the *the random number generation* of R should be set to a seed, in order to be able to reproduce the same results each time the code runs. So, with the command `set.seed(123)` the seed was set to *123*. K-Means Clustering can be done in R using the command `kcca()` from the package *flexClust* (Leisch, 2006). The function `kcca()` requires two arguments: the data to be clustered and the number of clusters. Since the study is about clustering the fertile and non-fertile women, there should've been only 2 clusters. In order to set the number of clusters within the function `kcca()`, the argument `k` was set to 2. Then with the command `barplot()` (R Core Team, 2021b), the resulting clusters were plotted.

5. Results.

5.1. *Face rating data set.* The output of the first linear regression fit to the data set is displayed on the table below. The table was printed using the R library *stargazeR* (Hlavac, 2022).

TABLE 1

	Dependent variable:
	bewertung
ratio	3.399**
	(1.397)
fertile	−0.265
	(2.508)
ratio:fertile	−0.131
	(2.577)
Constant	−0.398
	(1.359)
Observations	1,290
R^2	0.021
Adjusted R^2	0.018
Residual Std. Error	1.474 (df = 1286)
F Statistic	9.070*** (df = 3; 1286)
Note:	*p<0.1; **p<0.05; ***p<0.01

Since the interaction effect between the variables *ratio* and *fertile* is not significant, one should observe the interaction plot between the two variables.

The interaction plot above indicates, that there is no evidence, so that the variables *ratio* and *fertile* have a significant interaction effect. The lines are not crossing each other often and are mostly parallel.

Then, after the variable *ratio* was grouped and the data with the new variable *ratio_group* was fit into an ANOVA-model, the resulting summary of the model is displayed below.

Variable	Df	Sum Sq.	Mean square	F-value	P-value
ratio_group	2	83.3	41.64	(4.11E-09)	
fertile	1	40.9	40.93	19.266	(1.23E-05)
ratio_group:fertile	2	1.4	0.72	0.337	0.714
Residuals	1284	2727.5	2.12		

Since the interaction effect between the variables *ratio_group* and *fertile* is not significant again, one can assume that there is no interaction effect between them. Creating the interaction plot yields the resulting Figure 3.

As the lines do not cross and seem to be parallel as the interaction plot in Figure 2, the conclusion can be drawn that fertility has no effect on the partner preferences of the women, when it comes to facial appearances.

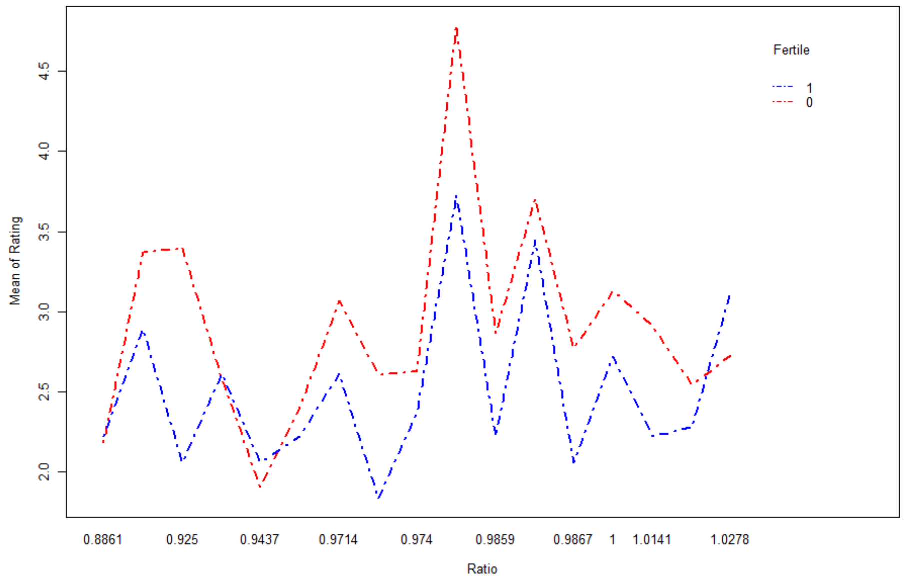

FIG 2. *The resulting interaction plot of variables ratio and fertile.*

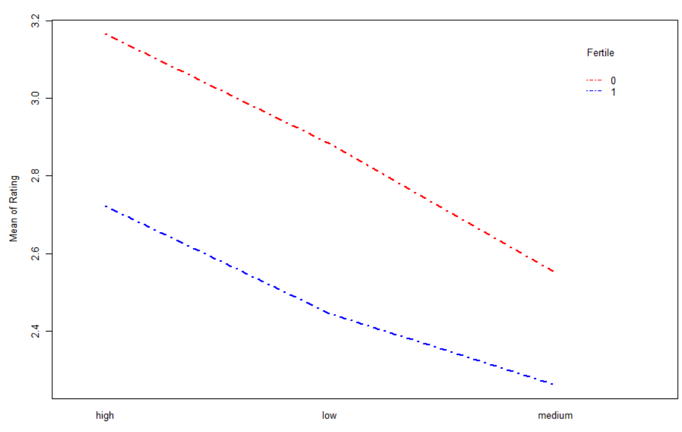

FIG 3. *The resulting interaction plot of the grouped variable* ratio_group *and fertile.*

5.2. *Evaluation questionnaire data set.* The Figures 4 and 5 show the resulting bar plots, when the K-Means Clustering Algorithm was applied to the data set.

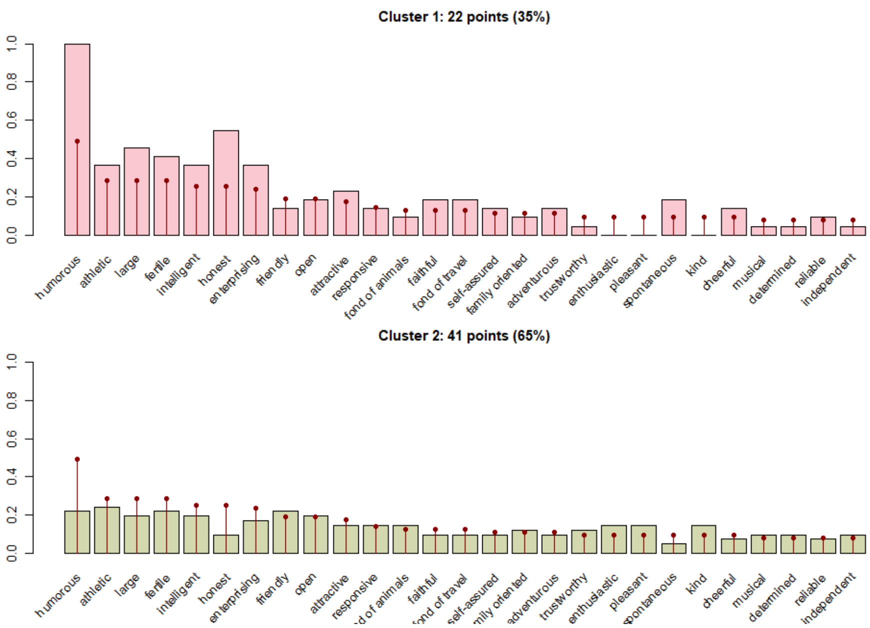

Fig 4. *Result of the K-Means Algorithm with 27 variables and 2 clusters.*

From the Figure 4, one can not derive any clear conclusion, because there are too many variables, which disturb the cluster structures. The problem of the data set is that there are too many character related variables than physical or appearance related variables. The solution to this problem is "balancing" the data, while keeping only 3 variables from both types in the data set. The variables with the highest number of 1's are remaining in the new data set.

The Figure 5 displays the clusters only with 7 variables, namely: *intelligent, humorous, honest, large, athletic, attractive* and *fertile*. The red dots in the middle of each bar, represent the global average of the variable. The interpretation can be that if the bar is above the red dot in the middle, then the variable is more dominant in the cluster and vice versa.

It can be observed on the Figure 5, that the first cluster is being dominated by fertile women. The variables *large* and *athletic* in this are significantly higher than their global means and the variable *attractive* has also a higher weight in the cluster compared to the second one. The second cluster however, is being dominated by the non-fertile women and the importance of the "physical" variables, which were dominant in the first cluster, is now much lower. The variables *intelligent, humorous* and *honest* dominate this cluster.

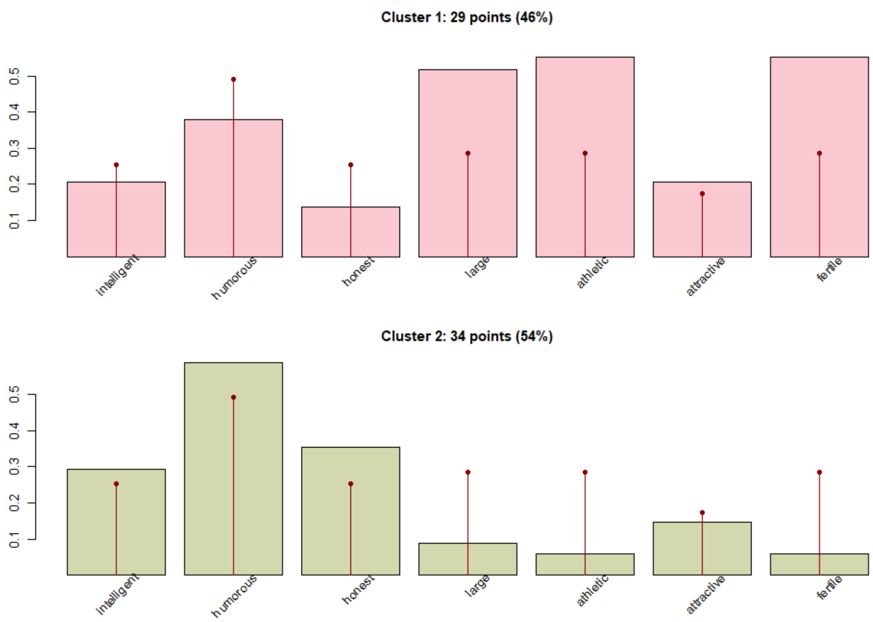

FIG 5. *Result of the K-Means Algorithm with 7 variables and 2 clusters.*

6. Discussion. The results above do not indicate, that the hypotheses,

- Women prefer more masculine faces and are more likely to look for physical characteristics while choosing their partner during their fertile phase.
- Women outside the fertile phase prefer more feminine faces and tend to look for character or material traits while choosing their partner.

should be accepted directly. The results of the first data set, indicate no clear connection between fertility and partner desires of women. One can conclude that there is no interaction effect between them and the first parts of the hypotheses should be rejected. However, the resulting clusters of the second data set indicate that the second parts of the hypotheses, that fertile women pay more attention to the physical traits of the desired partner and vise versa, should be accepted. The clusters in the Figure 5 show a clear distinction between fertile and non-fertile women.

References.

Ahmed, M., Seraj, R. and Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics* **9**.

Bock, H.-H. (2008). Origins and extensions of the k-means algorithm in cluster analysis. Accessed: 2022-8-29.

Cam, L. M. L. and Neyman, J. (1967). *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: Weather modification*. University of California Press, Berkeley, CA.

Cheung, Y.-M. (2003). -Means: A new generalized k-means clustering algorithm. *Pattern Recognit. Lett.* **24** 2883–2893.

Wikipedia contributors (2022). Digit ratio. https://en.wikipedia.org/w/index.php?title=Digit_ratio&oldid=1106080972.

Dowle, M. and Srinivasan, A. (2021). data.table: Extension of 'data.frame' R package version 1.14.2.

Fink, B., Grammer, K., Mitteroecker, P., Gunz, P., Schaefer, K., Bookstein, F. L. and Manning, J. T. (2005). Second to fourth digit ratio and face shape. *Proc. Biol. Sci.* **272** 1995–2001.

Fujikoshi, Y. (1993). Two-way ANOVA models with unbalanced data. *Discrete Math.* **116** 315–334.

Hlavac, M. (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables Social Policy Institute, Bratislava, Slovakia R package version 5.2.3.

Klusmann, D. and Berner, W. (2011). Veränderungen weiblicher Partnerpräferenzen im Menstruationszyklus. *Z. Sex. Forsch.* **24** 170–186.

Leisch, F. (2006). A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis* **51** 526–544.

Oti, E., Olusola, M., Eze, F. and Enogwe, S. (2021). Comprehensive Review of K-Means Clustering Algorithms. *International Journal of Advances in Scientific Research and Engineering* **07** 64-69.

R Core Team (2021a). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2021b). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H., Hester, J. and Bryan, J. (2022). readr: Read Rectangular Text Data R package

R-Codes

```
library(readr)
library(data.table)
library(dplyr)
library(flexclust)
```

Import the data sets and adjustments

```
Gesichterbewertung_1_ <- read.csv("Gesichterbewertung (1).csv", sep = ";")
gesicht <- Gesichterbewertung_1_
ges <- gesicht
ges$X2D <- as.numeric(gsub(",", ".", gsub("\\.", "", ges$X2D)))
ges$X4D <- as.numeric(gsub(",", ".", gsub("\\.", "", ges$X4D)))
ges$ratio <- as.numeric(gsub(",", ".", gsub("\\.", "", ges$ratio)))
ges$ratio <- as.numeric(ges$ratio)
colnames(ges) <- c("ID", "verhuetung", "fruchtbar", "Ges.ID", "2D", "4D", "ratio", "bewertung")
ges.final <- na.omit(ges)
```

create groups for ratio

```
int <- 1.0278 - 0.8861
intlength <- int/3
low <- 0.8861 + intlength
medium <- low + intlength
high <- medium + intlength

ges.final$ratio_group <- ges.final$ratio

ges.final$ratio_group[ges.final$ratio_group < low] <- "low"
ges.final$ratio_group[ as.numeric(ges.final$ratio_group) < medium &
                         low < as.numeric(ges.final$ratio_group)  ] <- "medium"
ges.final$ratio_group[as.numeric(ges.final$ratio_group) > medium] <- "high"
```

LM, ANOVA and interaction plots

```
lm_int <- lm(bewertung ~ ratio * fruchtbar, data = ges.final)
summary(lm_int)
interaction.plot(x.factor = ges.final$ratio,
                 trace.factor = ges.final$fruchtbar,
                 response = ges.final$bewertung,
                 fun = mean,
                 xlab = "Ratio",
                 lty = 4,
                 col = c("red", "blue"),
```

```
                 lwd = 2.5,
                 ylab = "Mean of Rating",
                 trace.label = "Fertile")


aov_group <- aov(bewertung ~ ratio_group * fruchtbar, data =ges.final)
summary(aov_group)

interaction.plot(x.factor = ges.final$ratio_group,
                 trace.factor = ges.final$fruchtbar,
                 response = ges.final$bewertung,
                 fun = mean,
                 xlab = "Ratio",
                 lty = 4,
                 col = c("red", "blue"),
                 lwd = 2.5,
                 ylab = "Mean of Rating",
                 trace.label = "Fertile")
```

Evaluation questionnaire data set

```
Auswertung_Fragebogen_2_ <- read.csv("Auswertung Fragebogen (2).csv", sep = ";")
frage <- Auswertung_Fragebogen_2_

frage_part <- frage[, c(1, 12:16, 137:234)]

colnames(frage_part) <- gsub(".1", "", colnames(frage_part))

setnames(frage_part, old = c("vertrauenswÃ.rdig", "begeisterungsfÃ.hig", "grÃ.ne.Augen",
                             "glÃ.cklich",
                             "guter.ZuhÃ.rer", "verstÃ.ndnisvoll", "verlÃ.sslich",
                             "verrÃ.ckt", "hÃ.flich",
                             "mitreiÃŸend", "fleiÃŸig", "zÃ.rtlich" , "selbststÃ.ndig",
                             "mÃ.nnlich" ,
                             "rÃ.cksichtsvoll" ,
                             "kompromissfÃ.hig", "unterstÃ.tzend", "muskulÃ.s" , "nicht.eifersÃ.chtig",
                             "natÃ.rlich" , "groÃŸ", "ï..ID"),
         new = c("vertrauenswuerdig", "begeisterungsfaehig", "gruene.Augen", "gluecklich",
                 "guter.Zuhoerer", "verstaendnisvoll", "verlaesslich",
                 "verrueckt", "hoeflich", "mitreissend", "fleissig", "zaertlich",
                 "selbststaendig", "maennlich", "ruecksichtsvoll", "kompromissfaehig",
                 "unterstuetzend", "muskuloes", "nicht.eifersuechtig",
                 "natuerlich", "gross","ID" ))

ges_temp <- ges[, c("ID", "fruchtbar")]

ges_uniq <- unique(ges_temp)
frage_part[frage_part$ID == "M23",]$ID <- "MS3"


match(ges_uniq$ID, frage_part$ID) #Order is correct

frage.final <- data.frame(frage_part, ges_uniq$fruchtbar)
```

```r
colnames(frage.final)[colnames(frage.final)=="ges_uniq.fruchtbar"] <- "fruchtbar"
frage.final.neu <- frage.final[,c( -25,-21, -30, -33, -64, -75, -76, -80, -81, -82, -83, -84, -97)]
```

Translations and eliminations

```r
rownames(frage.final.neu) <- frage.final.neu[,1]
frage.final.neu <- frage.final.neu[,-1]
frage.final.neu <- frage.final.neu[,-1:-4]
frage.final.neu
frage.final.neu[,85] <- as.numeric(frage.final.neu[,85])
frage.final.neu[1,85] <- 0
frage.final.neu <- frage.final.neu[,-1]

fragsums <- sort(colSums(frage.final.neu), decreasing = TRUE)
namen <- names(fragsums)[1:27] #Eliminate variables with low entries






namen_neu <- c("intelligent", "humorvoll", "ehrlich", "gross", "sportlich", "attraktiv", "fruchtbar")
frage_data <- frage.final.neu[, namen]

colnames(frage_data) <- c("humorous", "athletic",
                          "large", "fertile", "intelligent",
                          "honest", "enterprising",
                          "friendly", "open",
                          "attractive", "responsive",
                          "fond of animals", "faithful",
                          "fond of travel", "self-assured",
                          "family oriented", "adventurous",
                          "trustworthy", "enthusiastic",
                          "pleasant", "spontaneous",
                          "kind", "cheerful", "musical",
                         "determined",  "reliable", "independent") #27 Variables at the beginning

frage_data1 <- frage.final.neu[, namen_neu] # 7 Variables at the end
colnames(frage_data1) <- c("intelligent", "humorous", "honest", "large", "athletic",
                           "attractive", "fertile")
```

K-means Clustering

```r
set.seed(123)
res <- kcca(frage_data,k=2)
barplot(res)#the one with 27 Variables
set.seed(123)
res1 <- kcca(frage_data1 ,k=2) # With 7 Variables
barplot(res1)
```