Undergraduates Performance on General Education Prediction

Xiangman Zhao
xiangmaz@andrew.cmu.edu
November 2021

Abstract

Four related questions about finding appropriate variables to predict Carnegie Mellon University undergraduates' performances on statistics papers are examined in this paper. The data used is gathered by Dietrich College and includes 91 project papers — referred to as "artifacts"—were randomly sampled from a Fall and Spring section of Freshman Statistics. The model shows that rating is largely affected by Rater, Semester and Rubric. The dataset is split into 7 subsets for each rubric. The intercept only linear mixed effect model is fit on each subset and ICC is used to measure the agreement among raters on each rubric. The same model is also implemented on the full dataset, and ANOVA test is used to evaluate significant predictors of the model. The final model includes some important fixed effects like Rater, Semester and Rubric as well as some added interaction terms with Rubric. There is still room for model improvement if additional data on students' performances could be added.

1. Introduction

Rating is an important metric to evaluate how successful the new "General Education" program is and whether it will help students improve their performances at CMU. Dietrich College can investigate how rating in Freshman Statistics is affected by any possible factors.

This question is critical for the school to solve to find the relationship between ratings and other variables associated with the students' performances. Solving the problem also helps to improve students' academic performances, and the main aim of the paper is to build an optimal model to predict students' ratings and investigate if there are any additional information needed to better answer the question.

Four questions related to the per capita income are presented below:

1. Is the distribution of ratings for each rubric or given by each rater indistinguishable from others?
2. For each rubric, do the raters generally agree on their scores?
3. How are various factors in the experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings?
4. Is there anything else interesting to say about the data?

2. Data

The data for this study provides 91 project papers — referred to as "artifacts"—were randomly sampled from a Fall and Spring section of Freshman Statistics. Three raters from three different departments were asked to rate these artifacts on seven rubrics, as shown in Table 1. The rating

scale for all rubrics is shown in Table 2. Thirteen artifacts are viewed by all three raters, while the remaining 78 artifacts are viewed by only one rater. Table 3 shows the definitions for all the variables available for the study.

Rubric short name	Rubric full name	Description
RsrchQ	Research Question	Given a scenario, the student generates, critiques, or evaluates a relevant empirical research question.
CritDes	Critique Design	Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answer that question.
InitEDA	Initial EDA	Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis.
SelMeth	Select Method(s)	Given a data set and a research question, the student selects appropriate method(s) to analyze the data.
InterpRes	Interpret Results	The student appropriately interprets the results of the selected method(s).
VisOrg	Visual Organization	The student communicates in an organized, coherent, and effective fashion with visual elements (charts, graphs, tables, etc.).
TxtOrg	Text Organization	The student communicates in an organized, coherent, and effective fashion with text elements (words, sentences, paragraphs, section, and subsection titles, etc.).

Table 1: Rubrics for rating Freshman Statistics projects

Rating	Meaning
1	Student does not generate any relevant evidence.
2	Student generates evidence with significant flaws.
3	Student generates competent evidence; no flaws, or only minor ones.
4	Student generates outstanding evidence; comprehensive and sophisticated.

Table 2: Rating scale used for all rubrics

Variable Name	Values	Description
(X)	1, 2, 3, …	Row number in the data set
Rater	1, 2 or 3	Which of the three raters gave a rating
(Sample)	1, 2, 3, …	Sample number
(Overlap)	1, 2, …, 13	Unique identifier for artifact seen by all 3 raters
Semester	Fall or Spring	Which semester the artifact came from
Sex	M or F	Sex or gender of student who created the artifact
RsrchQ	1, 2, 3 or 4	Rating on Research Question
CritDes	1, 2, 3 or 4	Rating on Critique Design
InitEDA	1, 2, 3 or 4	Rating on Initial EDA
SelMeth	1, 2, 3 or 4	Rating on Select Method(s)
InterpRes	1, 2, 3 or 4	Rating on Interpret Results
VisOrg	1, 2, 3 or 4	Rating on Visual Organization
TxtOrg	1, 2, 3 or 4	Rating on Text Organization
Artificial	(text table)	Unique identifier for each artifact
Repeated	0 or 1	1 = this is one of the 13 artifacts seen by all 3 raters

Table 3: Variables Definitions

Histograms are plotted for seven rubrics for further analysis in Figure 1:

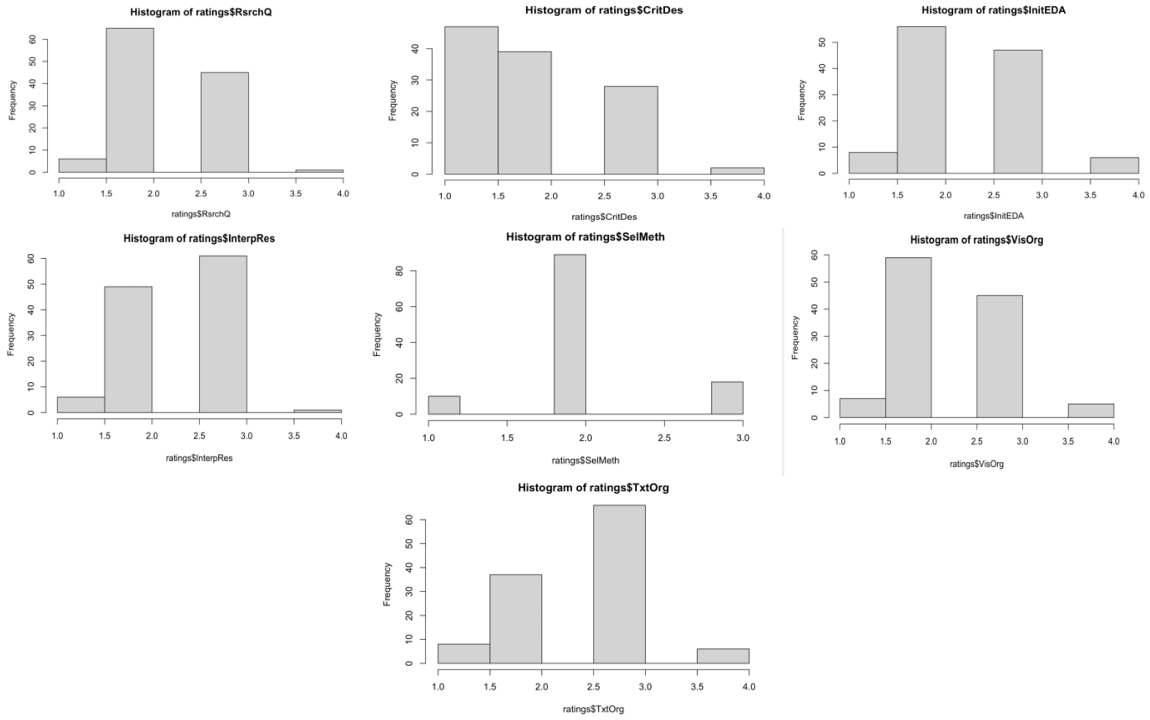


Figure 1: Histograms of Ratings for each Rubric

Counts of four ratings for all artifacts are summarized in Table 4.

	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg
Rating 1	47	8	6	6	10	8	7
Rating 2	39	56	49	65	89	37	59
Rating 3	28	47	61	45	18	66	45
Rating 4	2	6	1	1	0	6	5
NA	1	0	0	0	0	0	1

Table 4: Counts of ratings for all artifacts

3. Methods

3.1 The distribution of ratings for each rubric and rater

For the first research question, the full dataset, and the dataset only including 13 artifacts viewed by all raters are both used to solve the question. The distribution of ratings for each rubric and rater is plotted and compared using bar plots and counts tables.

3.2 If raters agree on their ratings for each rubric

For the second research question, the linear mixed effects model that only include intercept is fitted for each of the seven rubrics using both full dataset and the dataset only including 13 artifacts viewed by all raters. ICC for each model is calculated to investigate the correlation between raters on the same artifact. On the dataset including 13 artifacts viewed by all raters, the exact agreement rate is also calculated between raters is computed to cross classify the rating that each pair of raters gives for each rubric.

3.3 Finding significant factors related to the ratings

For the third research question, ANOVA test is mainly used to find significant variables as fixed effects and fitLMER.fnc is used to evaluate random effects. There are mainly four steps to solve the question:

1. Using the dataset containing 13 artifacts viewed by all raters, fixed effects are added to seven rubric-specific models one at each time using ANOVA test.
2. Using the full dataset containing all 91 artifacts, fixed effects are added to seven rubric-specific models one at each time using ANOVA test.
3. Continuing with the seven models fitted on the full dataset, interactions between significant fixed effects are added, and ANOVA test is used to find the significant interaction terms. Adding all significant fixed effects as random effects, fitLMER.fnc is used to find significant random effects.
4. All significant fixed effects, interactions and random effects from previous steps are added to the "combined model" Rating ~ 1 + (1 | Artifact), using the full dataset.

3.4 Investigating additional interesting features of the dataset

For the fourth research question, additional EDAs are tried to explore more on the data and understand the difference between models fitted on the dataset containing only 13 artifacts and the full dataset containing 91 artifacts.

4. Results

4.1 The distribution of ratings for each rubric and rater

To answer the first research question, the distribution of ratings for each rubric on the full dataset and the dataset containing only 13 artifacts are shown in Figure 2 and Figure 3. The distribution of ratings for each rater on the full dataset and the dataset containing only 13 artifacts are shown in Figure 4 and Figure 5, and it is found that:

- Focusing on the full dataset, rubrics RsrchQ, InitEDA, InterpRes, VisOrg and TxtOrg have similar normal distributions of ratings. Most of ratings are 2 and 3, and they all have very skinny tails, which means very few 1 and 4. Rubrics SelMeth and CritDes have a distribution that is more skewed to the right.
- Focusing on the dataset only containing 13 artifacts, rubrics RsrchQ, InitEDA, SelMeth, InterpRes and VisOrg have a nearly normal distribution of ratings. TxtOrg has a distribution skewing to the left, while CritDes has a distribution skewing to the right.
- Focusing on both the full dataset and the dataset only containing 13 artifacts, distributions of ratings for each rater show that each rater's ratings vary a lot between each rubric. Each rater gives consistent ratings for each rubric.

More information can be found in EDA section (page 1 to 6) of the code appendix.

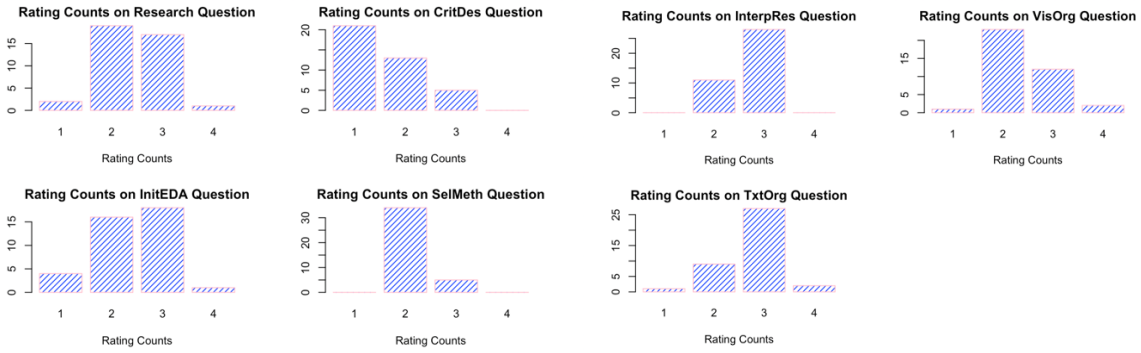
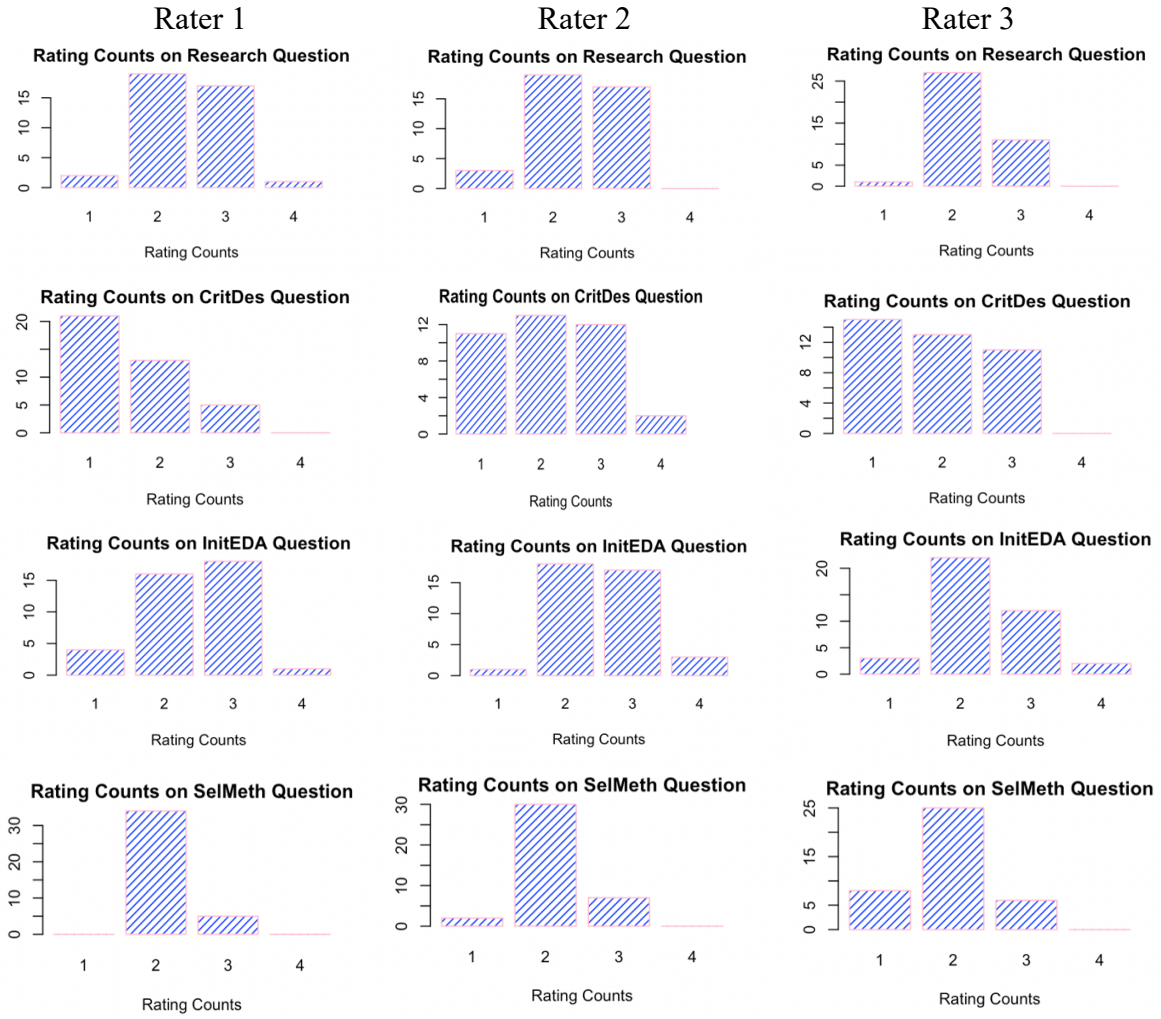


Figure 2: Bar plots of ratings for each rubric for the full dataset

Figure 3: Bar plots of ratings for each rubric for the dataset containing only 13 artifacts



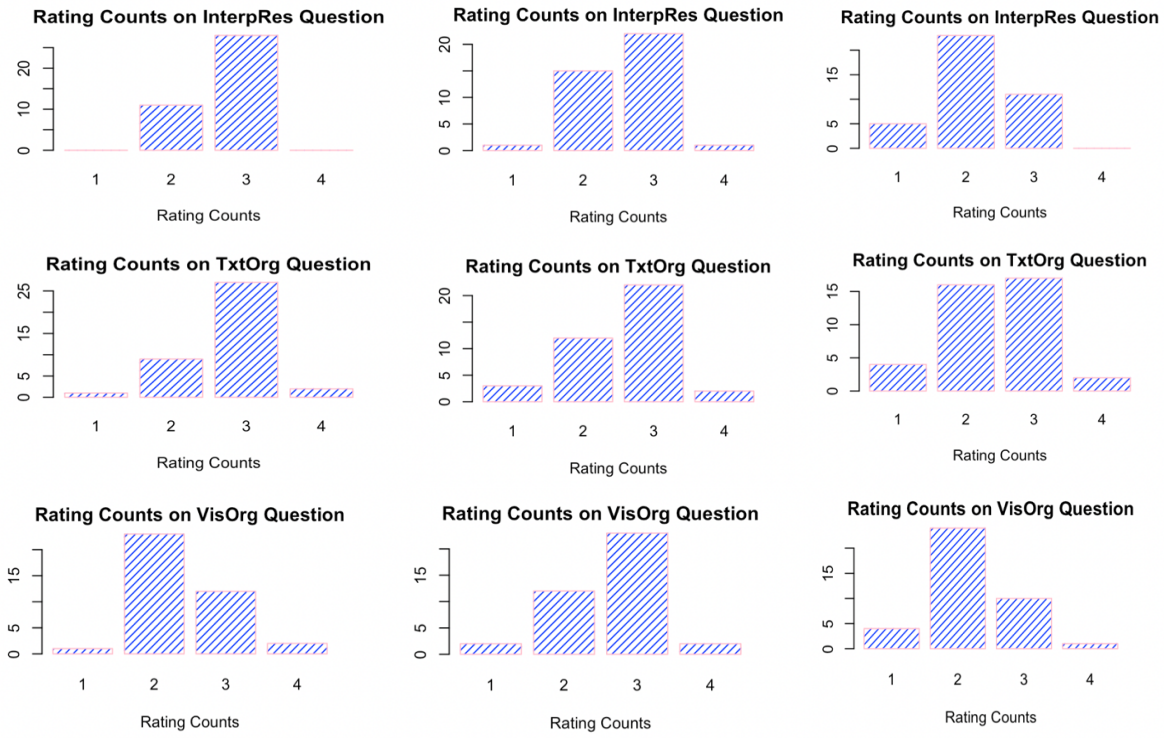
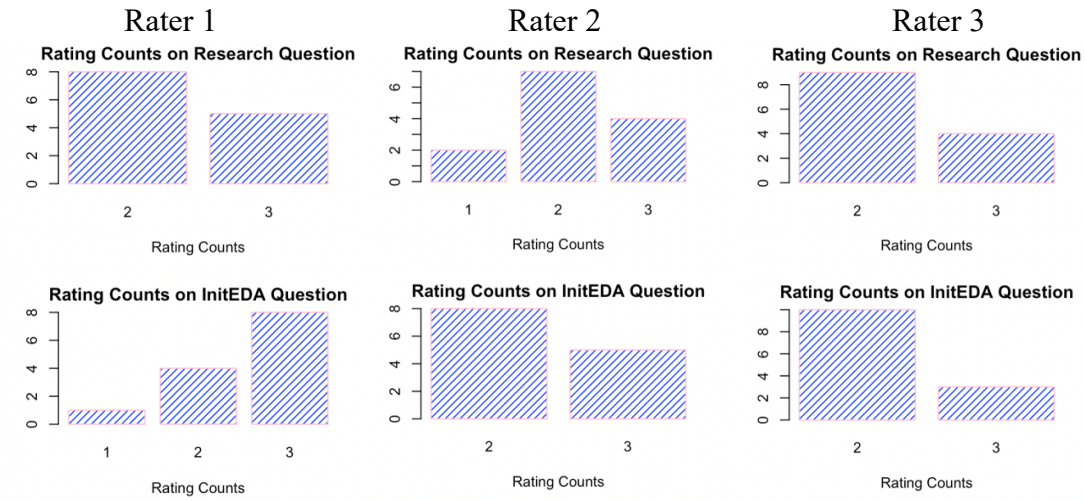


Figure 4: Bar plots of ratings for each rater for the full dataset



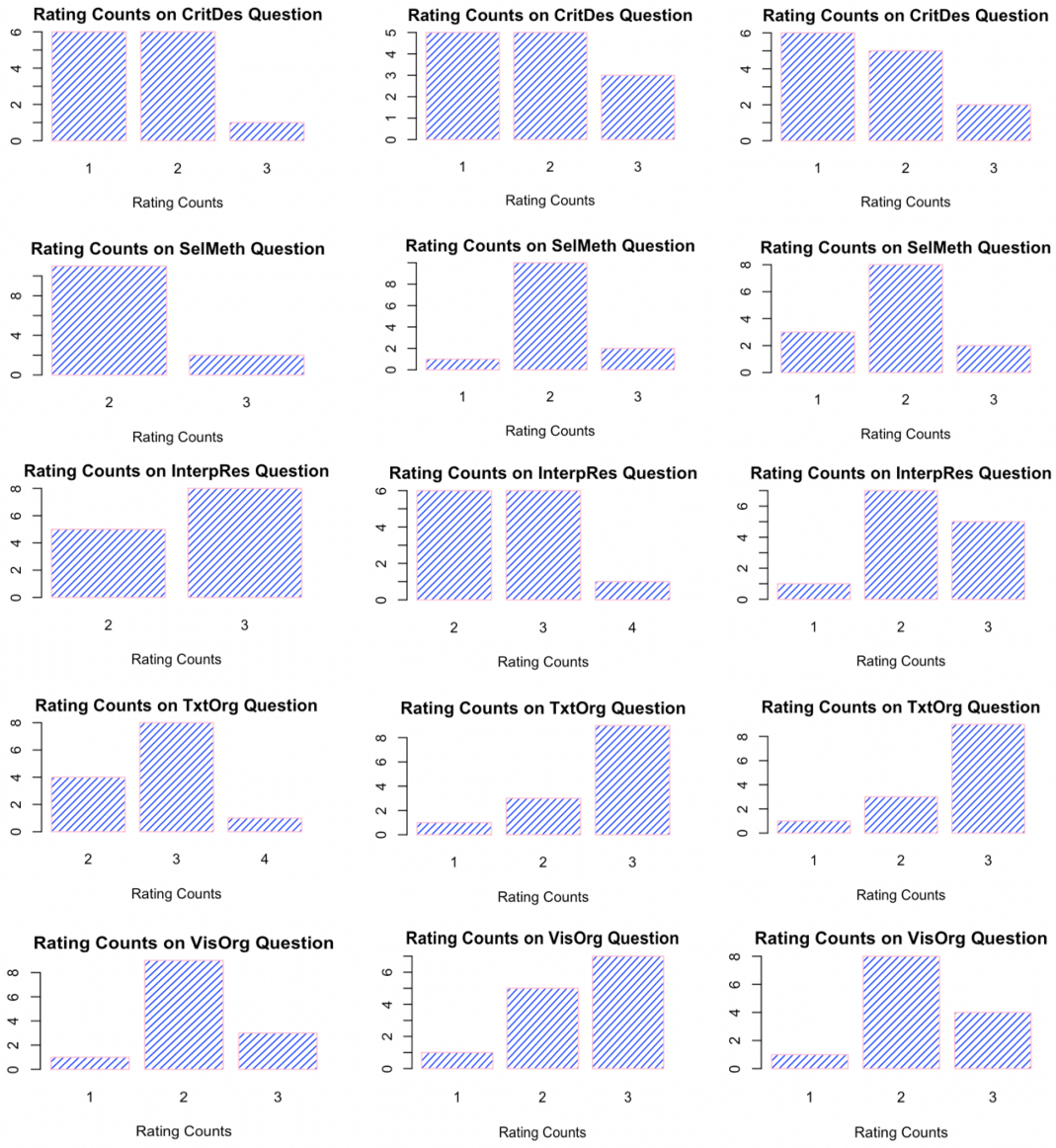


Figure 5: Bar plots of ratings for each rater for the dataset containing only 13 artifacts

4.2 If raters agree on their ratings for each rubric

For the second research question on whether raters agree on their scores, some findings from comparisons between each model's ICC and exact agreement rates between raters in Table 5 are shown below:

1. From columns ICC_fulldata and ICC_common, it is found that the correlation between raters on the same artifact for rubrics InitEDA, CritDes, SelMeth and VisOrg is high, which implies that raters' ratings vary a lot for different artifacts for these rubrics.
2. From exact agreement rates between raters, raters tend to have consistent ratings on most rubrics, which supports the result from ICC that raters are consistent with each other in how they rate. For RsrchQ, the exact agreement rate between rater 1 and rater 2 is the

lowest, which means that they don't agree with each other very often. Interestingly, for SelMeth, the exact agreement rate between rater 1 and rater 2 is the highest.

	ICC_fulldata	ICC_common	a12	a23	a13
RsrchQ	0.21	0.19	0.38	0.54	0.77
InitEDA	0.69	0.49	0.69	0.85	0.54
InterpRes	0.22	0.23	0.62	0.62	0.54
CritDes	0.67	0.57	0.54	0.69	0.62
SelMeth	0.47	0.52	0.92	0.69	0.62
TxtOrg	0.19	0.14	0.69	0.54	0.62
VisOrg	0.66	0.59	0.54	0.77	0.77

Table 5: Summary table of ICCs and exact agreement rates

More details can be found in crimes section (page 6 to 11) of the code appendix.

4.3 Finding significant factors related to the ratings

The third research question is trying to find significant factors relating to the ratings. Using only ANOVA test, the significant fixed effects for each rubric are shown below:

1. RsrchQ*: fix_RsrchQ = lmer (Rating ~ 1 + (1|Artifact), data=RsrchQ.full)*
2. CritDes: *fix_CritDes = lmer (Rating ~ 1 + (1|Artifact), data=CritDes.full)*
3. InitEDA: *fix_InitEDA = lmer (Rating ~ 1 + (1|Artifact), data=InitEDA.full)*
4. SelMeth: *fix_SelMeth = lmer (Rating ~ 1 + Rater + Sex + Semester + (1|Artifact), data=SelMeth.full)*
5. InterpRes: *fix_InterpRes = lmer (Rating ~ 1 + Rater + (1|Artifact), data=InterpRes.full)*
6. VisOrg: *fixed_VisOrg = lmer (Rating ~ 1 + (1|Artifact), data=VisOrg.full)*
7. TxtOrg: *fixed_TxtOrg = lmer (Rating ~ 1 + Rater + (1|Artifact), data=TxtOrg.full)*

When finding the significant fixed effects by backwards-elimination using fitLMER.fnc, the result is different as the intercept-only model is selected for all rubrics. The result given by fitLMER.fnc is more reliable as fitLMER.fnc uses t-tests with a threshold of 2 using ML instead of REML, which returns the mostly correct result. Therefore, for the dataset containing 13 artifacts viewed by all raters, there is no significant fixed effect to predict the rating for each rubric.
When finding of the significant fixed effects by using fitLMER.fnc on the full data, seven models are shown below:

1. RsrchQ*: fix_RsrchQ = lmer (Rating ~ (1|Artifact), data=RsrchQ.full)*
2. CritDes: *fix_CritDes = lmer (Rating ~ Rater + (1|Artifact) - 1, data=CritDes.full)*
3. InitEDA: *fix_InitEDA = lmer (Rating ~ (1|Artifact), data=InitEDA.full)*
4. SelMeth: *fix_SelMeth = lmer (Rating ~ Rater + Semester + (1|Artifact) - 1, data=SelMeth.full)*

5. InterpRes: *fix_InterpRes = lmer (Rating ~ Rater + (1|Artifact) - 1, data=InterpRes.full)*
6. VisOrg: *fixed_VisOrg = lmer (Rating ~ Rater + (1|Artifact) - 1, data=VisOrg.full)*
7. TxtOrg: *fixed_TxtOrg = lmer (Rating ~ (1|Artifact), data=TxtOrg.full)*

Lastly, instead of treating each rubric separately, a combined model is fitted as below:

finalmod = lmer (Rating ~ 1 + Rater + Semester + Rubric + Rater:Rubric + (0+ Rubric|Artifact)

Figure 6 shows the summary table of the final model:

```
Fixed effects:
                          Estimate Std. Error t value
(Intercept)                1.75945    0.11785  14.929
as.factor(Rater)2          0.22093    0.11780   1.875
as.factor(Rater)3         -0.11959    0.11834  -1.011
SemesterS19               -0.17780    0.08228  -2.161
RubricInitEDA              0.74625    0.13676   5.457
RubricInterpRes            1.01453    0.13478   7.527
RubricRsrchQ               0.74926    0.12419   6.033
RubricSelMeth              0.42671    0.13040   3.272
RubricTxtOrg               1.04967    0.13551   7.746
RubricVisOrg               0.68353    0.13947   4.901
RubricCritDes:Rater2       0.14443    0.17442   0.828
RubricInitEDA:Rater2      -0.16400    0.15528  -1.056
RubricInterpRes:Rater2    -0.39230    0.15193  -2.582
RubricRsrchQ:Rater2       -0.35713    0.15697  -2.275
RubricSelMeth:Rater2      -0.25158    0.15659  -1.607
RubricTxtOrg:Rater2       -0.43936    0.15065  -2.916
RubricCritDes:Rater3       0.33380    0.17481   1.910
RubricInitEDA:Rater3       0.03858    0.15615   0.247
RubricInterpRes:Rater3    -0.41867    0.15280  -2.740
RubricRsrchQ:Rater3       -0.03687    0.15782  -0.234
RubricSelMeth:Rater3      -0.07942    0.15745  -0.504
RubricTxtOrg:Rater3       -0.15268    0.15153  -1.008
```

Figure 6: Coefficients of Fixed Effects

Table 6 is the interpretation of the final model:

Predictor	Interpretation
Rater 2	Compared to Rater 1, Rater 2 will give 0.22093 higher ratings
Rater 3	Compared to Rater 1, Rater 3 will give 0.11959 lower ratings
SemesterS19	Compared to F19, ratings given during semester S19 will be 0.1778 lower
RubricInitEDA	Compared to RubricCritDes, RubricInitEDA's rating will be 0.74625 higher
RubricInterpRes	Compared to RubricCritDes, RubricInterpRes's rating will be 1.01453 higher
RubricRsrchQ	Compared to RubricCritDes, RubricRsrchQ's rating will be 0.74926 higher
RubricSelMeth	Compared to RubricCritDes, RubricSelMeth's rating will be 04267 higher
RubricTxtOrg	Compared to RubricCritDes, RubricTxtOrg's rating will be 1.04967 higher

RubricVisOrg	Compared to RubricCritDes, RubricVisOrg's rating will be 0.68353 higher
Interaction terms with rater	Compared to Rater 1, Rater 2 tends to give lower ratings on most rubrics except for RubricCritDes; while Rater 3 tends to give higher ratings on RubricCritDes and RubricInitEDA, and lower ratings to the rest rubrics.

Table 5: Interpretations of the final model

More information can be found in model fitting section (page 11 to 18) of the code appendix.

4.4 Investigating additional interesting features of the dataset

For the fourth research question, Bar plots showing the distributions of ratings on S19 and F19 are plotted in Figure 7. The EDA shows that the ratings distribution between these two semesters are not that different, but it is a significant variable for the model, as well as the interaction term between semester and rubric. One reason is that the dataset is not evenly split. The data of fall semester is much more than that of the spring semester, which may cause the difference between ratings in these two semesters. More information can be found in the cross-validation section (page 18 to 27) of the code appendix.

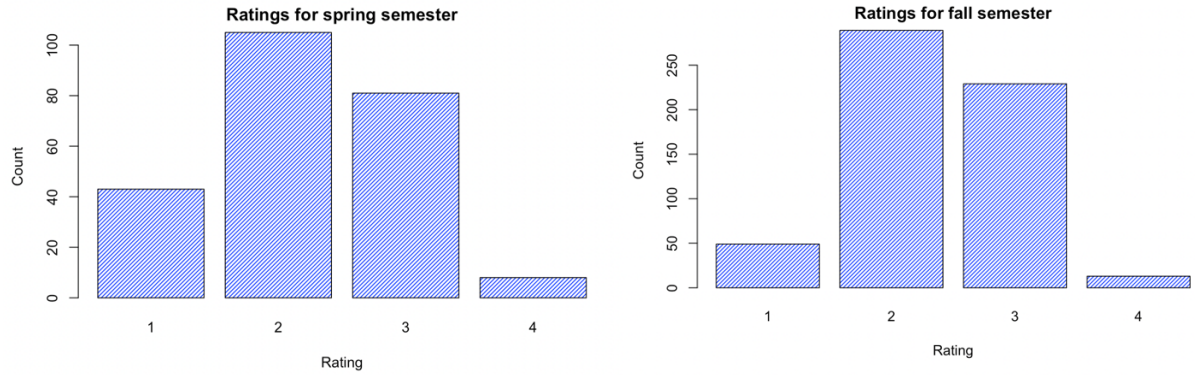


Figure 7: Bar plots comparing ratings between spring and fall semesters

5. Discussions

5.1 The distribution of ratings for each rubric and rater

The distributions of ratings between different rubrics and raters are not largely different. Comparing between the full dataset and the dataset containing 13 artifacts viewed by all raters, the distributions of ratings for each rubric are also similar. Although the sizes of datasets are different and the distribution on the full dataset tends to be more skewed due to the larger data size, the means and medians are close between these two datasets. Focusing on the distribution

of ratings for each rater, each rater's rating varies a lot between rubrics, and there is no obvious pattern in their distributions of ratings.

5.2 If raters agree on their ratings for each rubric

For the second research question, it is found that all raters agree with each other for most of rubrics. Especially for rubrics InitEDA, CritDes, SelMeth and VisOrg, the correlation between raters' ratings on the same artifact is high, which means that raters are consistent in their ratings for these rubrics. From exact agreement rates between raters, raters tend to have consistent ratings on most rubrics, which supports the result from ICC that raters are consistent with each other in how they rate. The result for the second question corresponds with the result in the first research question that ratings largely depend on the rubric instead of raters.

5.3 Finding significant factors related to the ratings

When focusing on the small dataset containing only 13 artifacts viewed by all raters, intercept-only models with not fixed effect are chosen. The result shows that the ratings are not affected by raters, sex, or other variables. When using the full dataset, some rubrics have raters and semester as fixed effects, while rubrics RsrchQ, InitEDA and TxtOrg still choose intercept-only models. For other rubrics, rater and semester will affect the rating, which indicates problematic gradings and needs further improvements. For the final model on the full dataset, ANOVA test and fitLMER.fnc both choose rater, semester and rubric as the significant fixed effects, and the interaction between rubric and rater is also significant as an interaction term in the final model. It is expected that rubric is chosen as a fixed effect as some rubrics are straightforward and tend to have higher ratings like InterpRes and TxtOrg. It is problematic that semester and rater are fixed effects, which shows that the ratings are not consistent between different raters and semesters. One possible explanation is that raters come from different departments and probably focus on different perspectives in their ratings on each artifact.

5.4 Investigating additional interesting features of the dataset

For the last question, two bar plots show that distribution of ratings on two semesters are very similar to each other. Although semester is a fixed effect in the final model, there is not an obvious difference in the distribution of ratings between Fall 2019 and Spring 2019. The data of the fall semester is much more than that of the spring semester, and it might be better to collect more data to have an evenly split dataset between two semesters to have a more accurate result.

5.5 Limitations and future improvements

Some strengths of the modeling part are that both full dataset and the dataset containing only 13 artifacts viewed by all raters are taken into considerations, which shows mores features of the dataset. There are still some limitations with the final model. First, the dataset is small and contains some missing values including the missing input values for sex. When fitting the models, these missing values are just removed. It might be better to consider a more comprehensive way to deal with these missing values. Second, random effects are not considered in the modeling process, it might be a good idea to consider them later. Lastly, the paper does not

explain the difference between fitting models on the dataset containing 13 artifacts viewed by all raters and the full dataset. More interpretations and EDAs can be included to better understand the difference between these two datasets. Some other models can be fitted to see if the interpretation will be any different. For example, how about treating the response variable as a categorical variable and fitting a classification model to if the significant variables will be different. These future steps can be taken to understand the dataset better.

R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplotify)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(arm)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## arm (Version 1.12-2, built: 2021-10-15)
```

```
## Working directory is /Users/zhaoxiangman/Desktop/36-617
```

```
library(performance)
```

```
##
## Attaching package: 'performance'
```

```
## The following object is masked from 'package:arm':
##
##     display
```

```
cdi <- read.table("/Users/zhaoxiangman/Desktop/36-617/cdi.dat", header = TRUE)
```

```
cdi$pci <- rescale(2*cdi$per.cap.income)
cdi$phg <- rescale(2*cdi$pct.hs.grad)
```

```
cdi$state = as.factor(cdi$state)
lmer.1 <- lmer(pci ~ 1 + phg + (1|state) +(0 + phg|state),data=cdi)
```

```
str(fixef(lmer.1))
```

```
##  Named num [1:2] -0.058 0.536
##  - attr(*, "names")= chr [1:2] "(Intercept)" "phg"
```

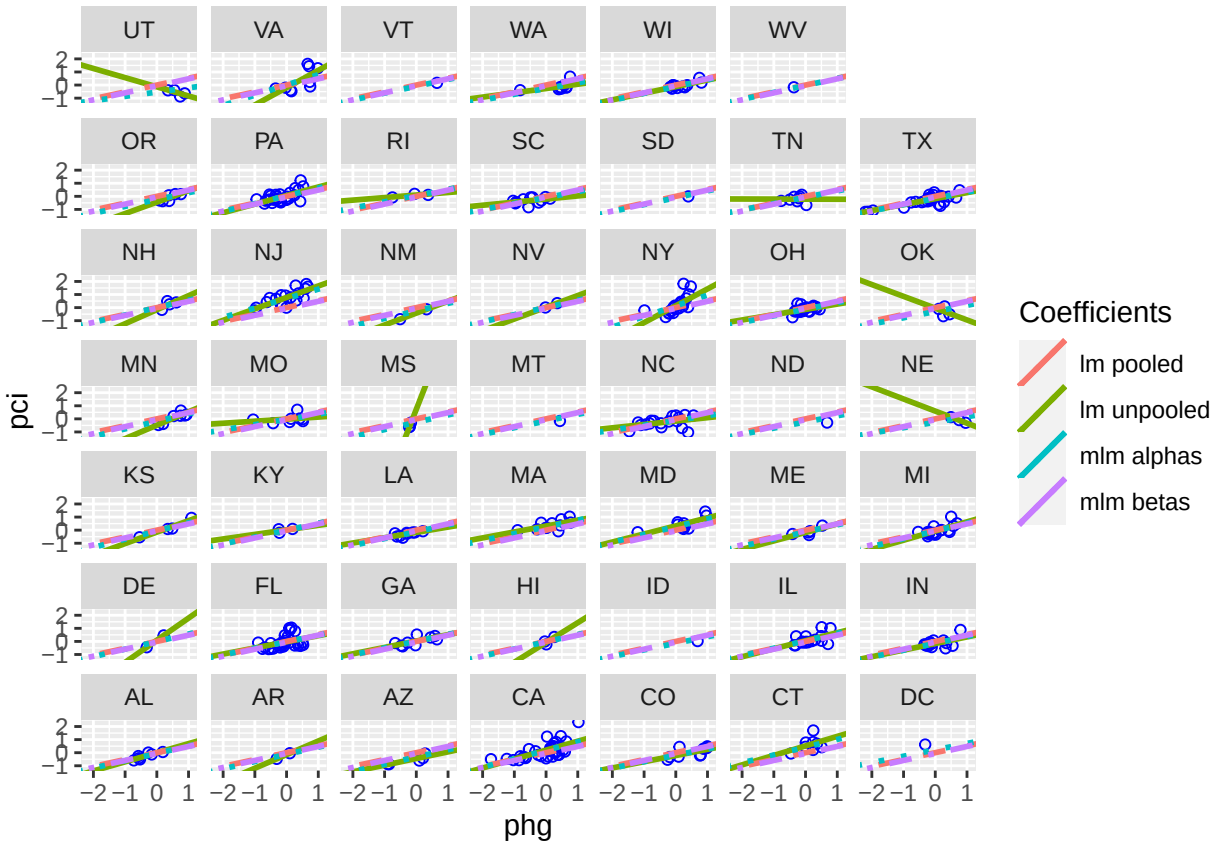
```
beta0 <- fixef(lmer.1)[1]
beta1 <- fixef(lmer.1)[2]

str(ranef(lmer.1))
```

```
## List of 1
##  $ state:'data.frame':   48 obs. of  2 variables:
##   ..$ (Intercept): num [1:48] 0.00227 -0.07002 -0.26943 0.23747 -0.13598 ...
##   ..$ phg        : num [1:48] 0.0218 0.0152 0.0174 0.1109 -0.0444 ...
##   ..- attr(*, "postVar")=List of 2
##   .. ..$ (Intercept): num [1, 1, 1:48] 0.01573 0.03021 0.01792 0.00338 0.01378 ...
##   .. ..$ phg        : num [1, 1, 1:48] 0.02671 0.03182 0.02331 0.00758 0.0196 ...
##  - attr(*, "class")= chr "ranef.mer"
```

```
eta <- ranef(lmer.1)$state

alpha <- matrix(NA,nrow=dim(eta)[1],ncol=dim(eta)[2])

alpha[,1] <- beta0 + eta[,1]   # alpha0
alpha[,2] <- beta1 + eta[,2]   # alpha1

lm.0 <- lm(pci ~ phg,data=cdi)

lm.1 <- lm(pci ~ state*phg - phg - 1,data=cdi)

g <- ggplot(cdi,aes(x=phg,y=pci)) +
  facet_wrap( ~ state, as.table=F) +
    geom_point(pch=1,color="blue")

int1 <- rep(NA,length(unique(cdi$state)))
int2 <- int1
int3 <- int1
int4 <- int1
slo1 <- int1
slo2 <- int1
slo3 <- int1
slo4 <- int1

J <- length(unique(cdi$state))
for (j in 1:J) {
    int1[j] <- coef(lm.1)[j]
    slo1[j] <- coef(lm.1)[J+j]
    int2[j] <- coef(lm.0)[1]
    slo2[j] <- coef(lm.0)[2]
    int3[j] <- alpha[j,1]
    slo3[j] <- alpha[j,2]
    int4[j] <- beta0
```

```
    slo4[j] <- beta1
}

params <- ddply(cdi, .(state), summarize,
                int1 <- int1[state[1]],
                slo1 <- slo1[state[1]],
                int2 <- int2[state[1]],
                slo2 <- slo2[state[1]],
                int3 <- int3[state[1]],
                slo3 <- slo3[state[1]],
                int4 <- int4[state[1]],
                slo4 <- slo4[state[1]])

names(params) <- c("state",
                   "int1","slo1",
                   "int2","slo2",
                   "int3","slo3",
                   "int4","slo4")

g + geom_abline(data=params,
                    aes(intercept=int1,slope=slo1,
                        color="lm unpooled"),lty=1,size=1) +
        geom_abline(data=params,
                    aes(intercept=int2,slope=slo2,
                        color="lm pooled"),lty=2,size=1) +
        geom_abline(data=params,
                    aes(intercept=int3,slope=slo3,
                        color="mlm alphas"),lty=3,size=1) +
        geom_abline(data=params,
                    aes(intercept=int4,slope=slo4,
                        color="mlm betas"),lty=4,size=1) +
        labs(color="Coefficients")
```

```
## Warning: Removed 7 rows containing missing values (geom_abline).
```




```
source("/Users/zhaoxiangman/Desktop/36-617/residual-functions.r")

str(fixef(lmer.1))
```

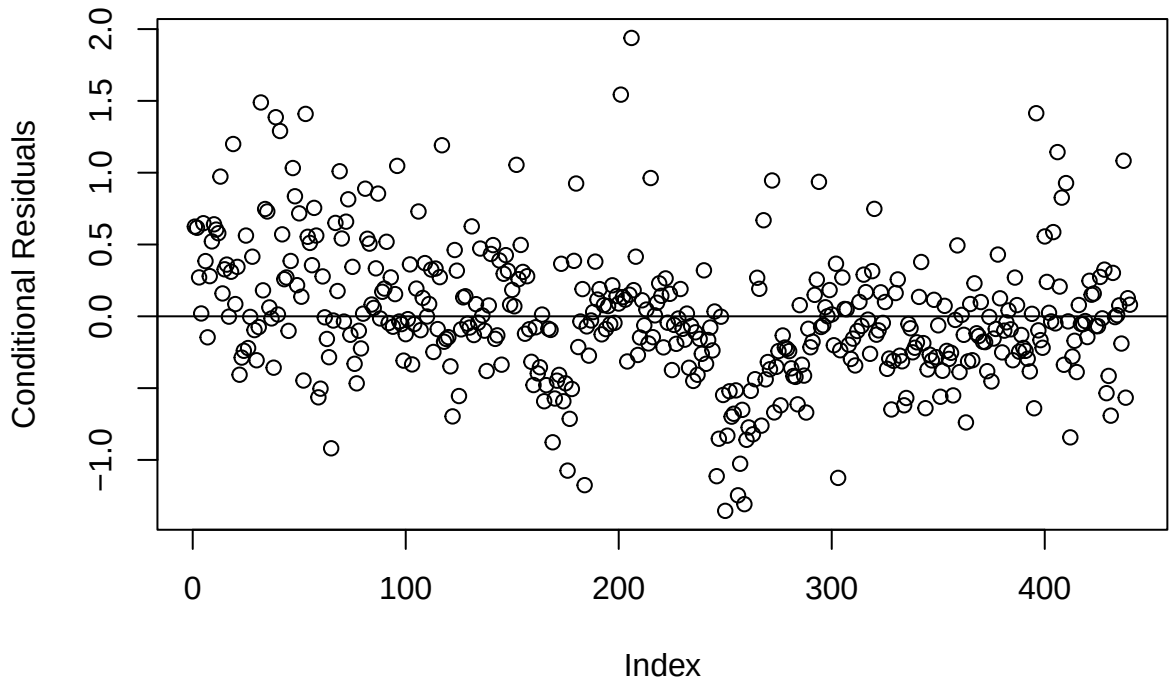
```
##  Named num [1:2] -0.058 0.536
##  - attr(*, "names")= chr [1:2] "(Intercept)" "phg"
```

```
beta0 <- fixef(lmer.1)[1]
beta1 <- fixef(lmer.1)[2]

str(ranef(lmer.1))
```

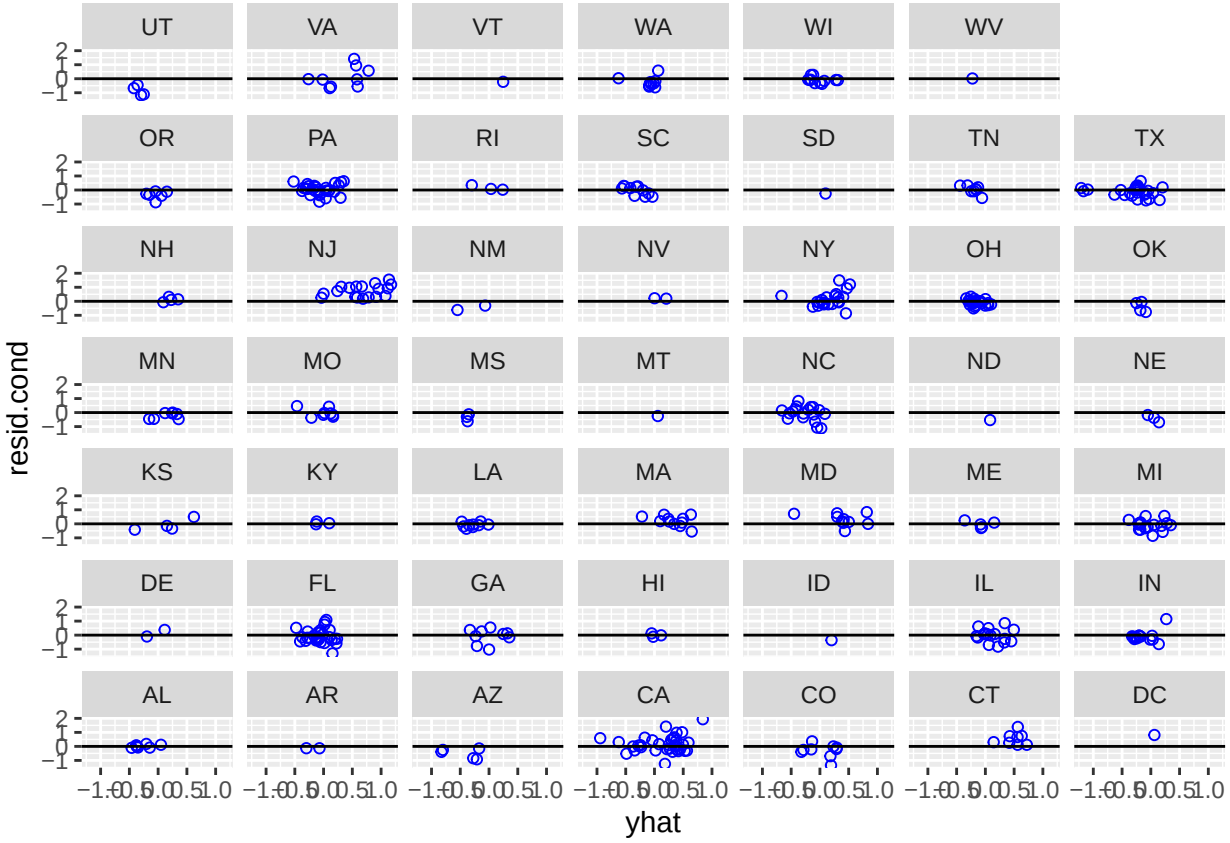
```
## List of 1
##  $ state:'data.frame':   48 obs. of  2 variables:
##   ..$ (Intercept): num [1:48] 0.00227 -0.07002 -0.26943 0.23747 -0.13598 ...
##   ..$ phg        : num [1:48] 0.0218 0.0152 0.0174 0.1109 -0.0444 ...
##   ..- attr(*, "postVar")=List of 2
##   .. ..$ (Intercept): num [1, 1, 1:48] 0.01573 0.03021 0.01792 0.00338 0.01378 ...
##   .. ..$ phg        : num [1, 1, 1:48] 0.02671 0.03182 0.02331 0.00758 0.0196 ...
##  - attr(*, "class")= chr "ranef.mer"
```

```
eta <- ranef(lmer.1)$state

alpha <- matrix(NA,nrow=dim(eta)[1],ncol=dim(eta)[2])

alpha[,1] <- beta0 + eta[,1]    # alpha0
alpha[,2] <- beta1 + eta[,2]    # alpha1

attach(cdi)
```

```r
X <- cbind(1,phg)

blocks <- lapply(split(X,state),function(x){matrix(x,ncol=2)})
J <- length(blocks)
n <- dim(cdi)[1]

Z <- matrix(0,nrow=n,ncol=J*2)

row <- 1
for (j in 1:J) {
  col <- 2*j
  nj <- dim(blocks[[j]])[1]
  Z[row:(row+nj-1),c(col-1,col)] <- blocks[[j]]
  row <- row + nj
}

beta <- rbind(beta0,beta1) # so beta is a column vector

eta <- c(t(eta))           # so eta is a column vector

resid.marg <- pci - X%*%beta

resid.cond <- pci - X%*%beta - Z%*%eta

resid.reff <- Z%*%eta

yhat <- yhat.cond(lmer.1)

plot(resid.cond,xlab="Index",ylab="Conditional Residuals")
abline(0,0)
```



```
sch <- as.numeric(state)
index <- sch
for (j in unique(sch)) {
  len <- sum(sch==j)
  index[sch==j] <- 1:len
}

new.data <- data.frame(yhat,resid.cond,state)
names(new.data) <- c("yhat","resid.cond","state")
ggplot(new.data,aes(x=yhat,y=resid.cond)) +
  facet_wrap( ~ state, as.table=F) +
  geom_point(pch=1,color="Blue") +
  geom_hline(yintercept=0)
```




```
detach()
```

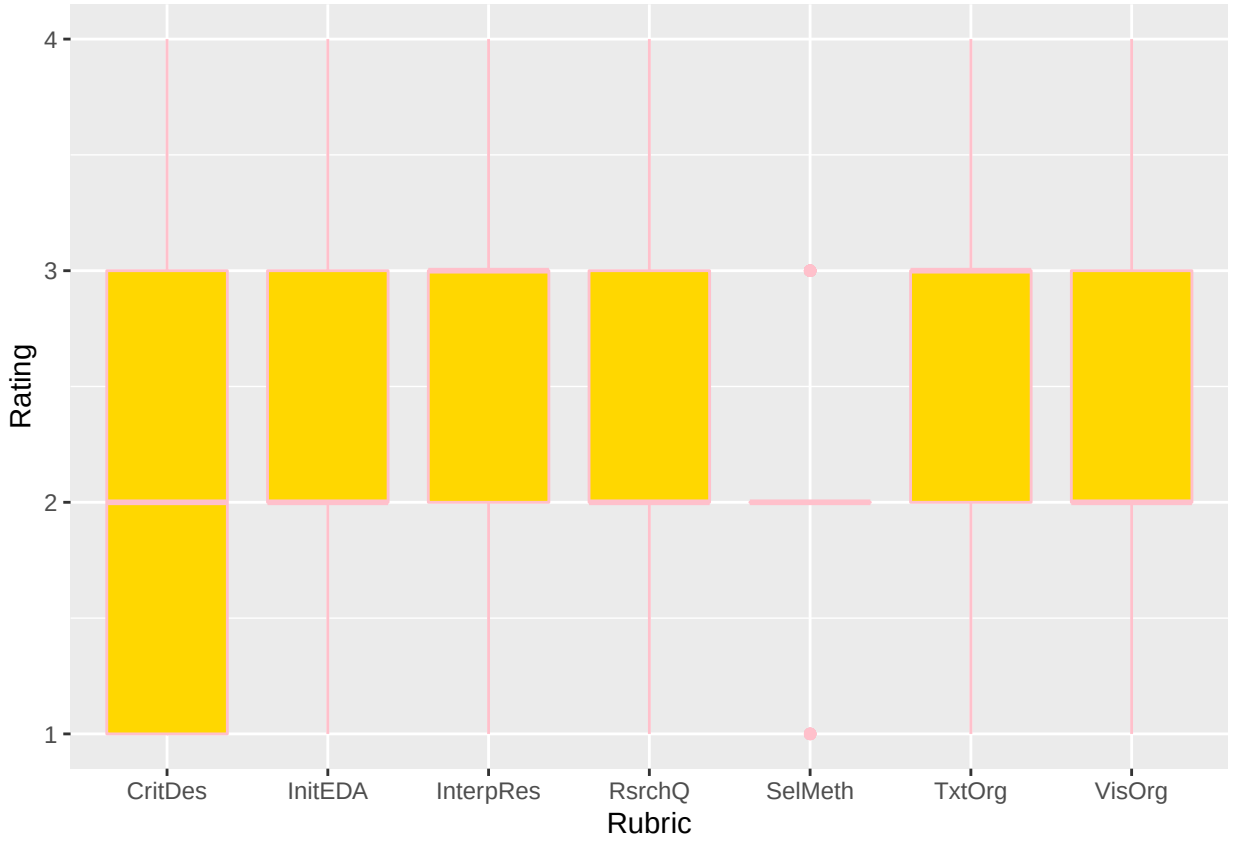
2

1

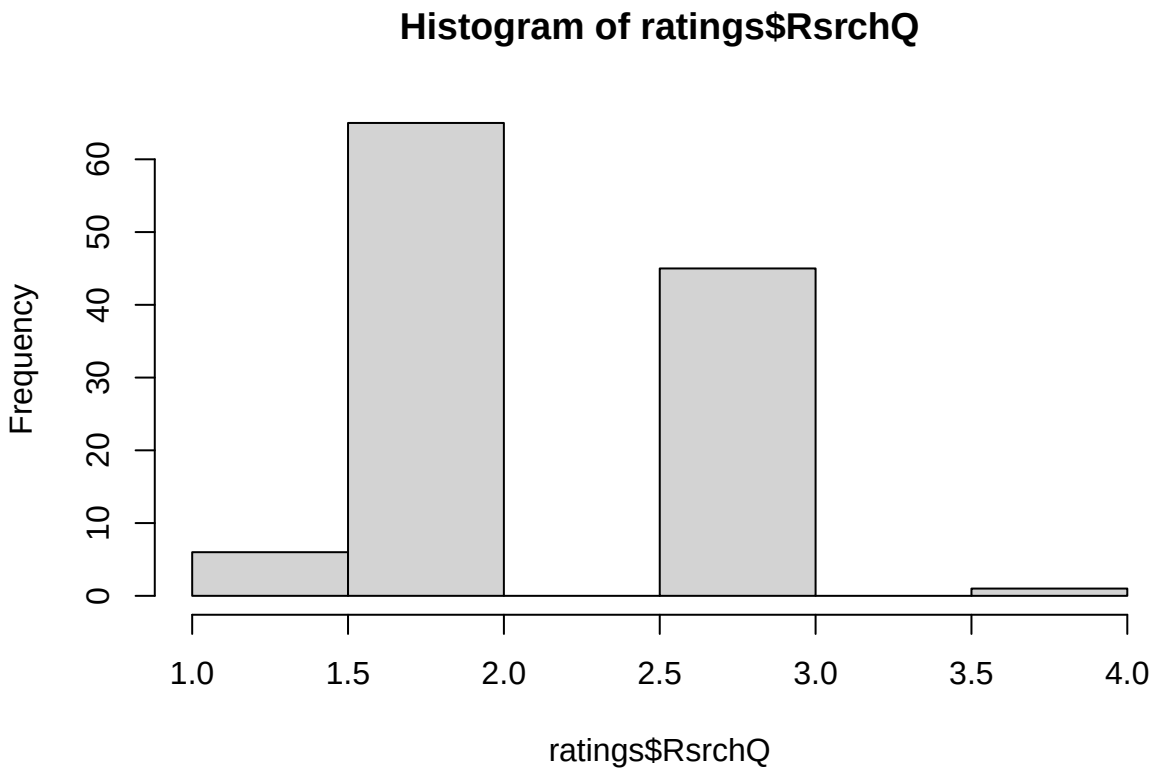
```
ratings <- read.csv("/Users/zhaoxiangman/Desktop/36-617/ratings.csv")
tall <- read.csv("/Users/zhaoxiangman/Desktop/36-617/tall.csv")
```

```
ggplot(tall, aes(x = Rubric, y = Rating)) +
        geom_boxplot(fill = "gold1", color = "pink")
```

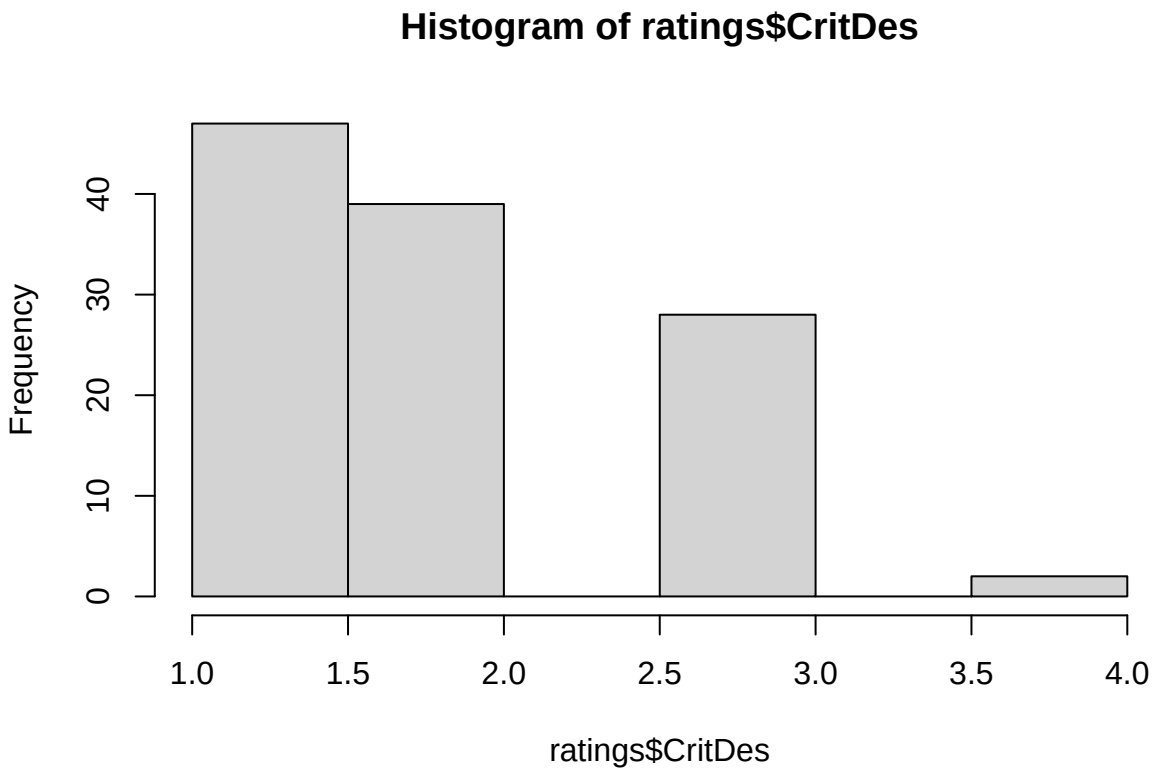
```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



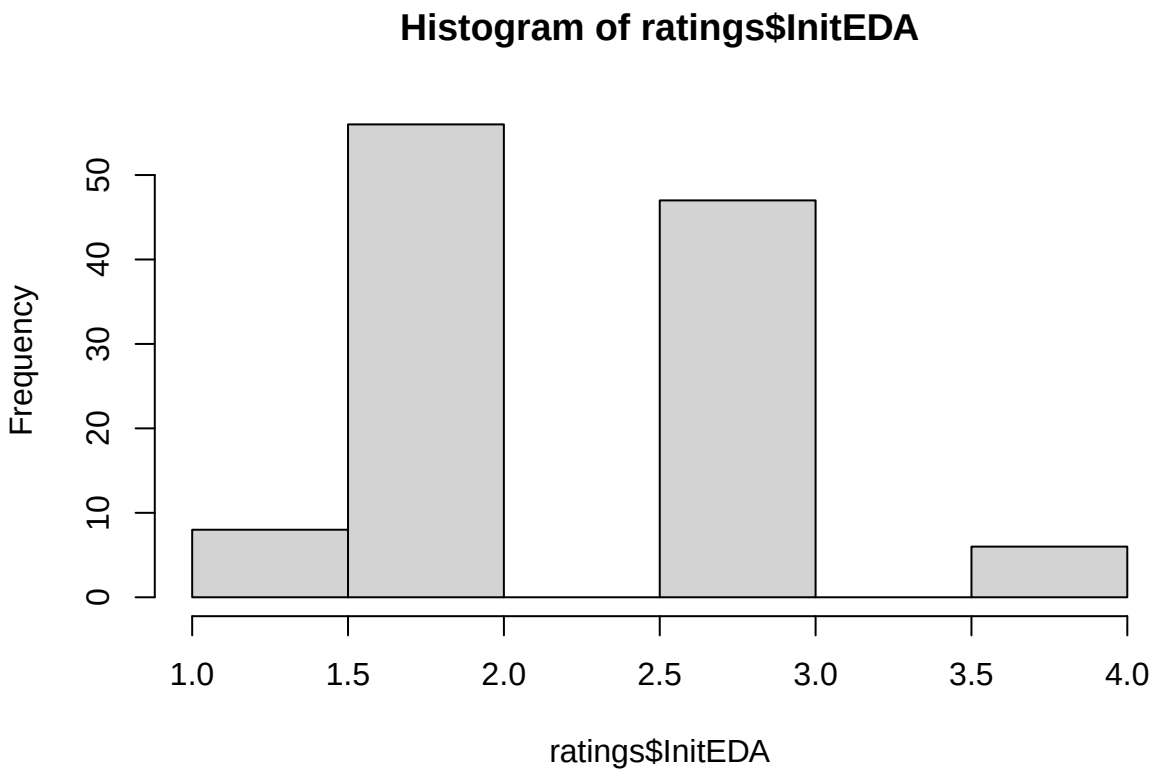

```
hist(ratings$RsrchQ)
```



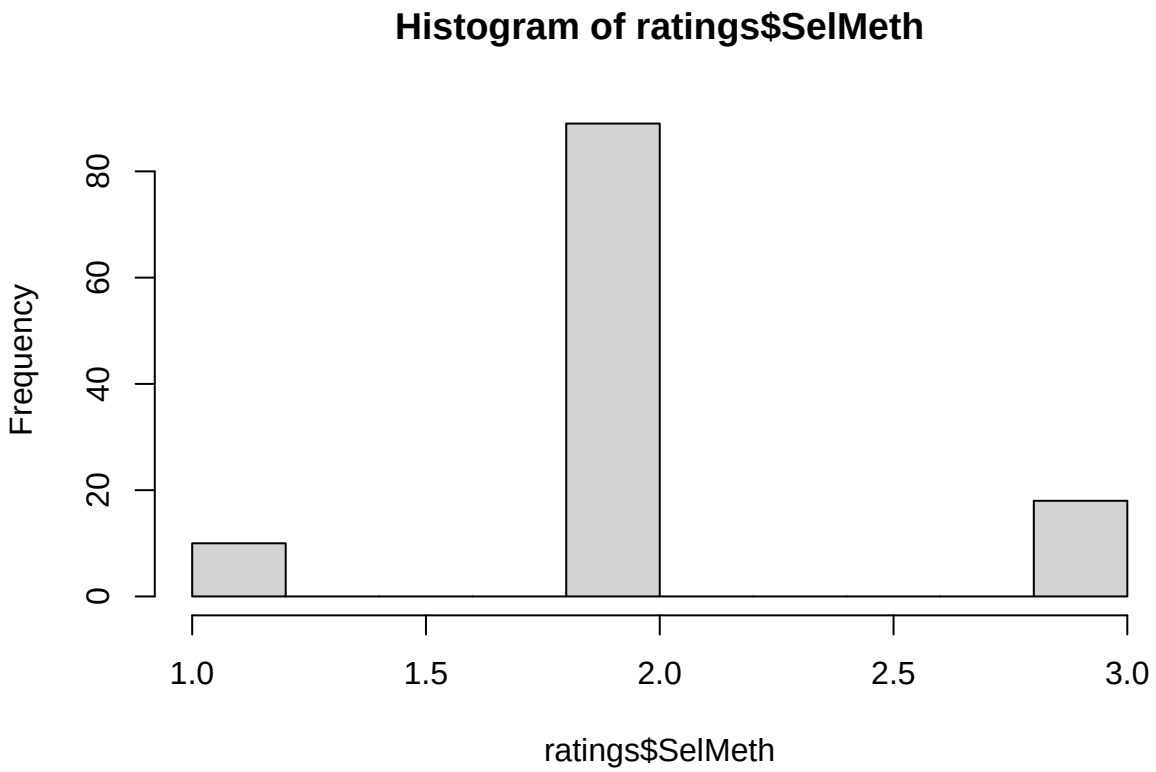
```
hist(ratings$CritDes)
```



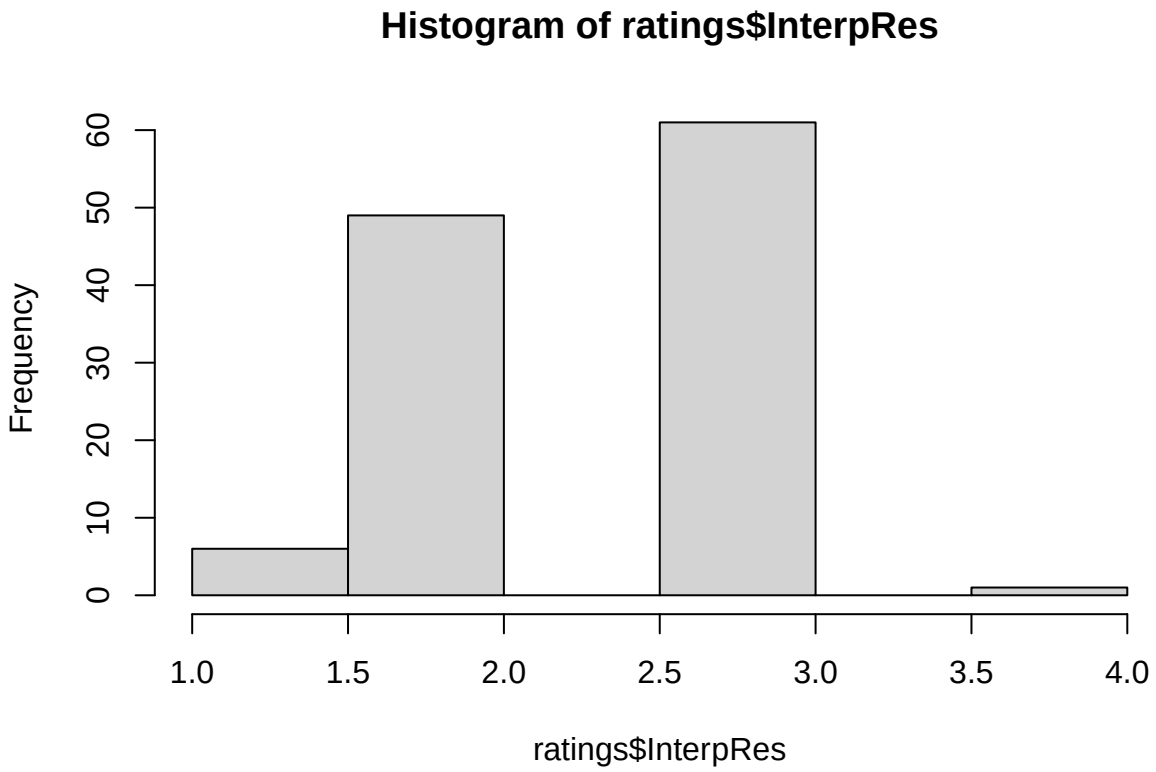

```
hist(ratings$InitEDA)
```



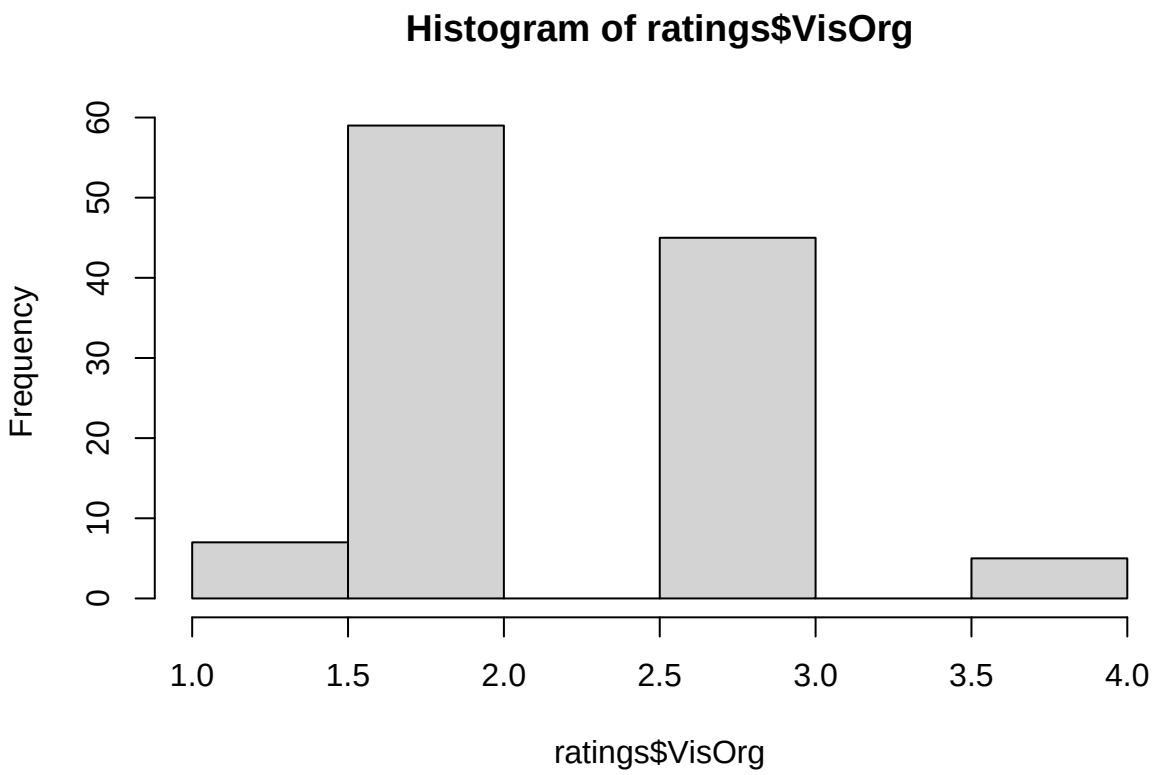
```
hist(ratings$SelMeth)
```



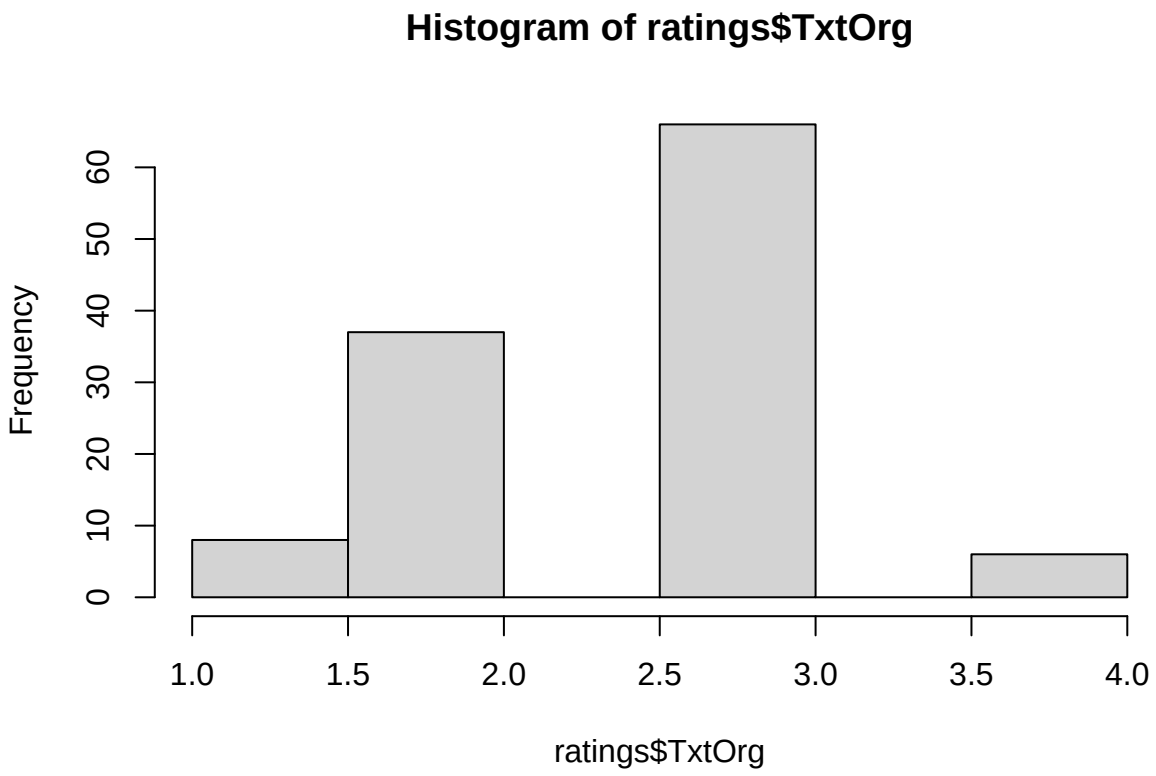

```
hist(ratings$InterpRes)
```



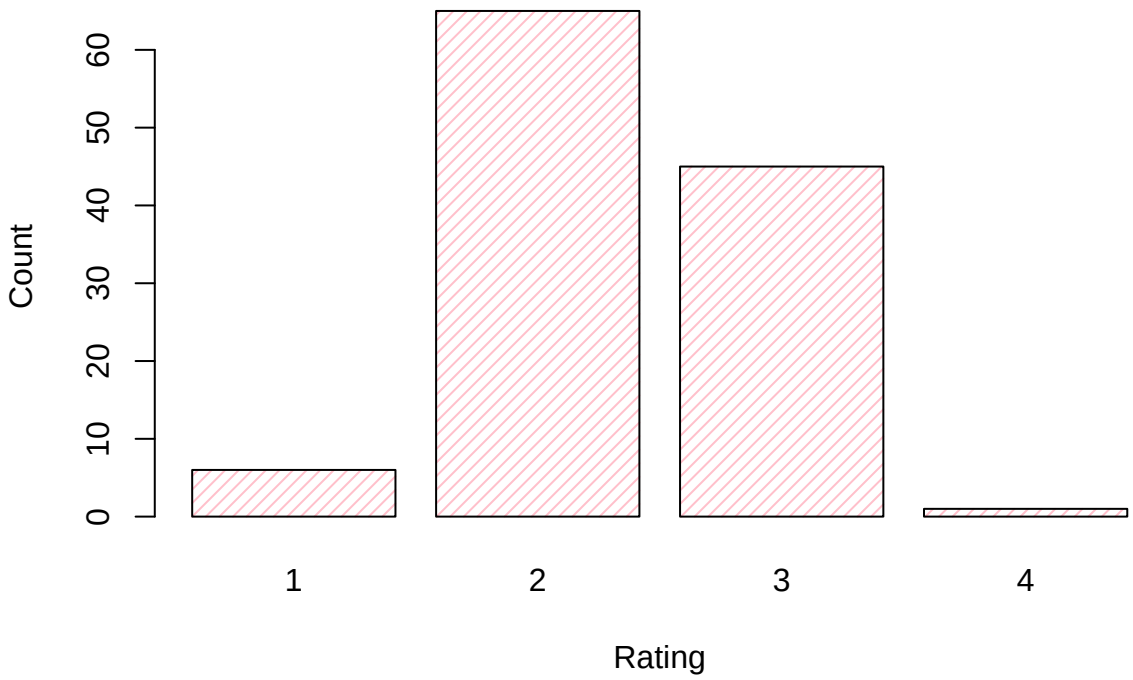
```
hist(ratings$VisOrg)
```



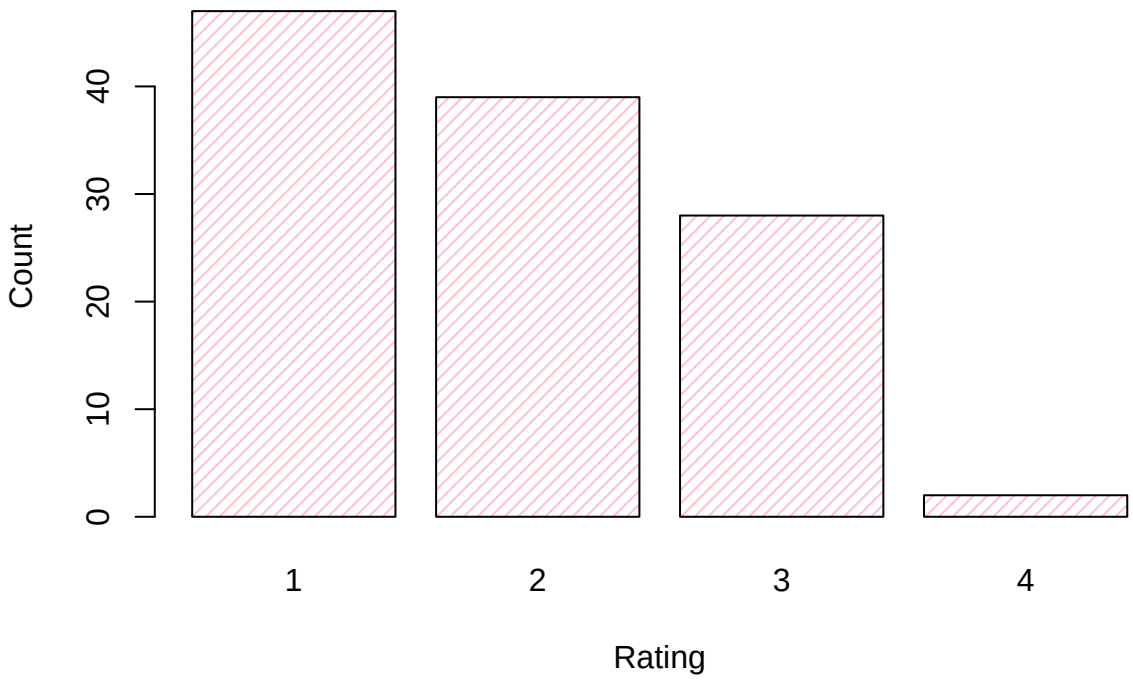

```
hist(ratings$TxtOrg)
```



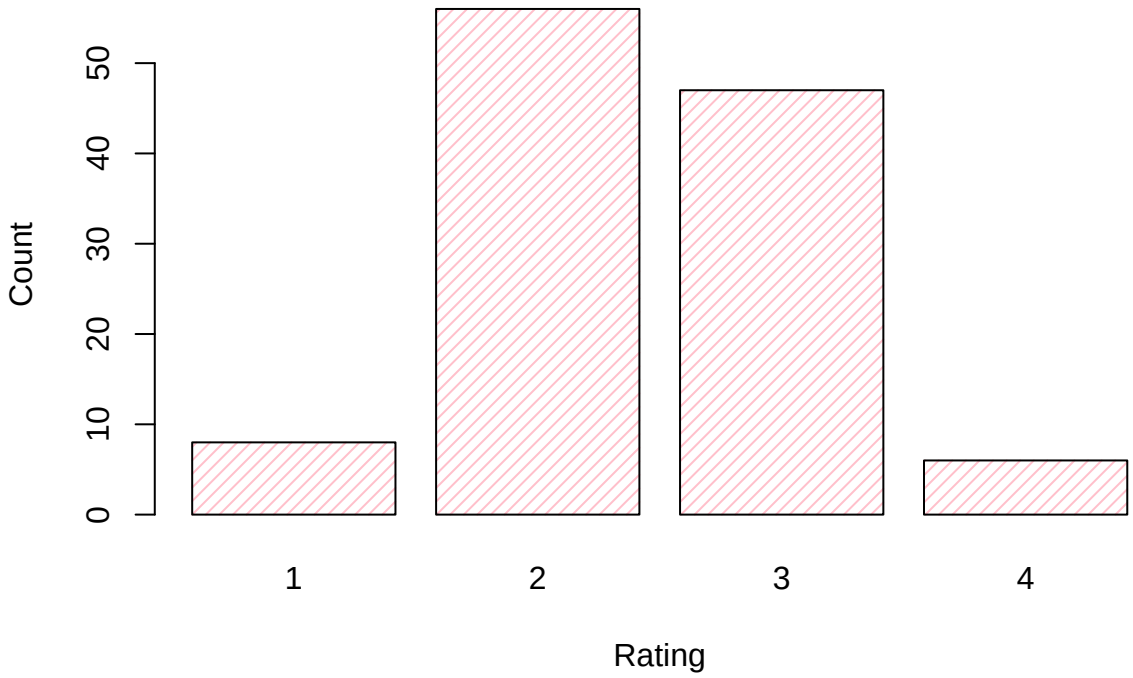
```
barplot(table(ratings$RsrchQ), main = "Ratings for Rubric RsrchQ", xlab = "Rating", ylab = "Count", col=
```



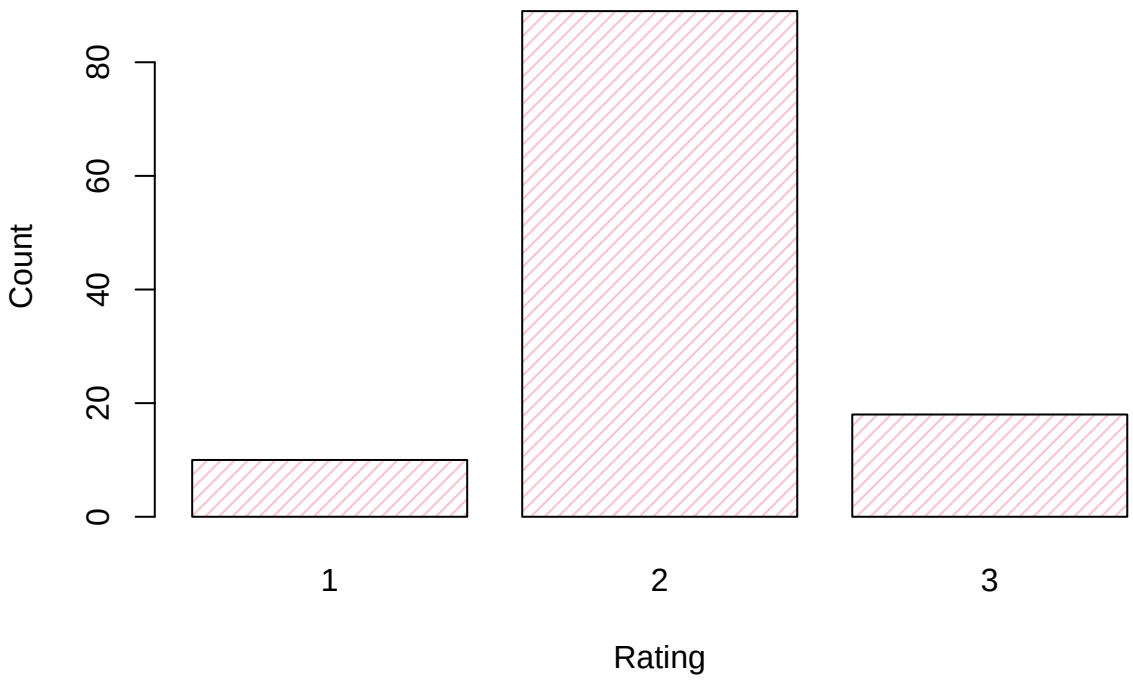

```
barplot(table(ratings$CritDes), main = "Ratings for Rubric CritDes", xlab = "Rating", ylab = "Count", co
```



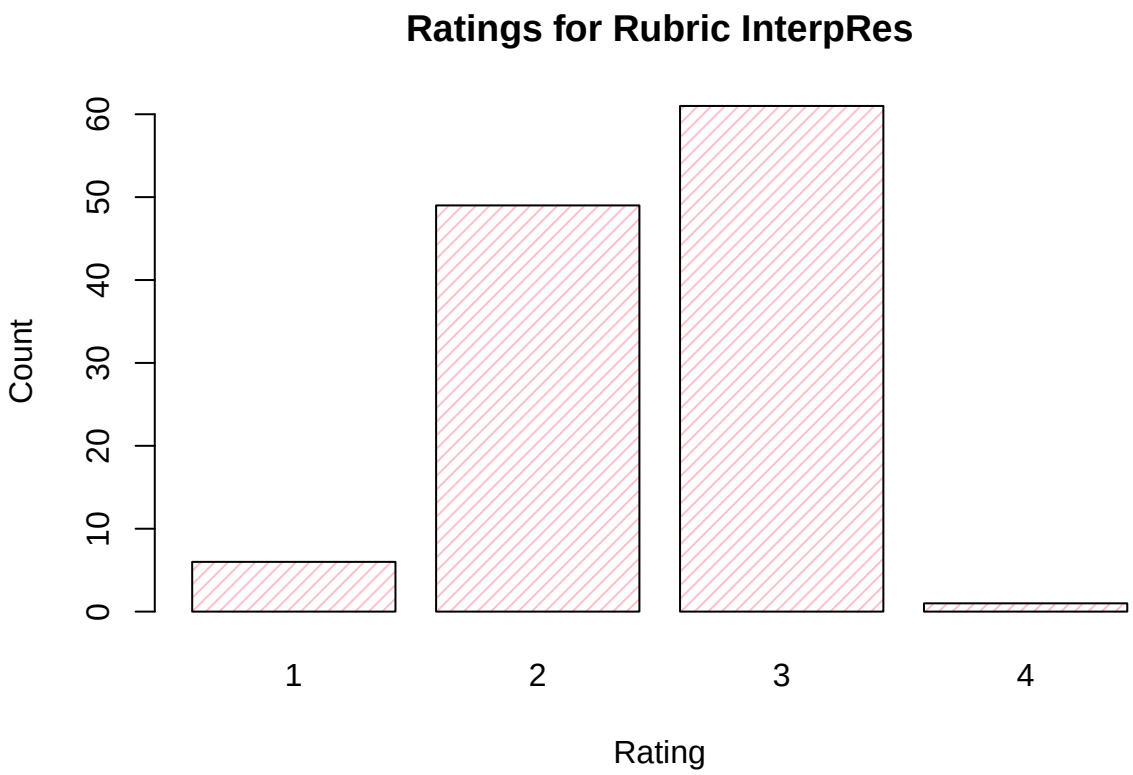
```
barplot(table(ratings$InitEDA), main = "Ratings for Rubric InitEDA", xlab = "Rating", ylab = "Count", co
```



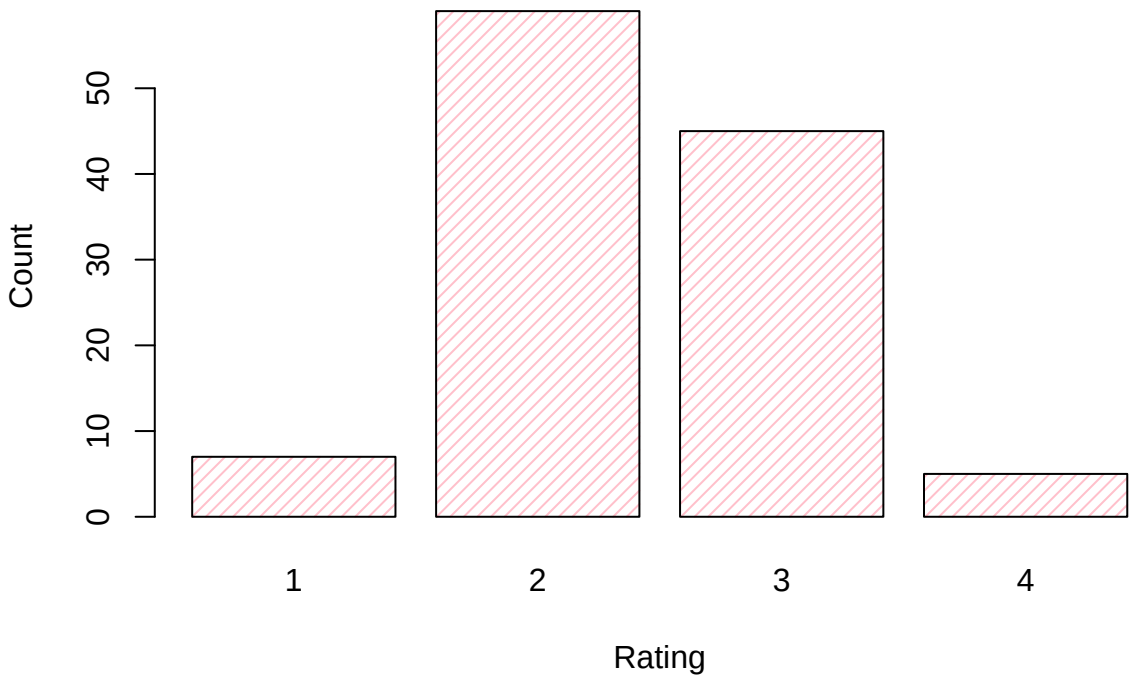

```
barplot(table(ratings$SelMeth), main = "Ratings for Rubric SelMeth", xlab = "Rating", ylab = "Count", co
```



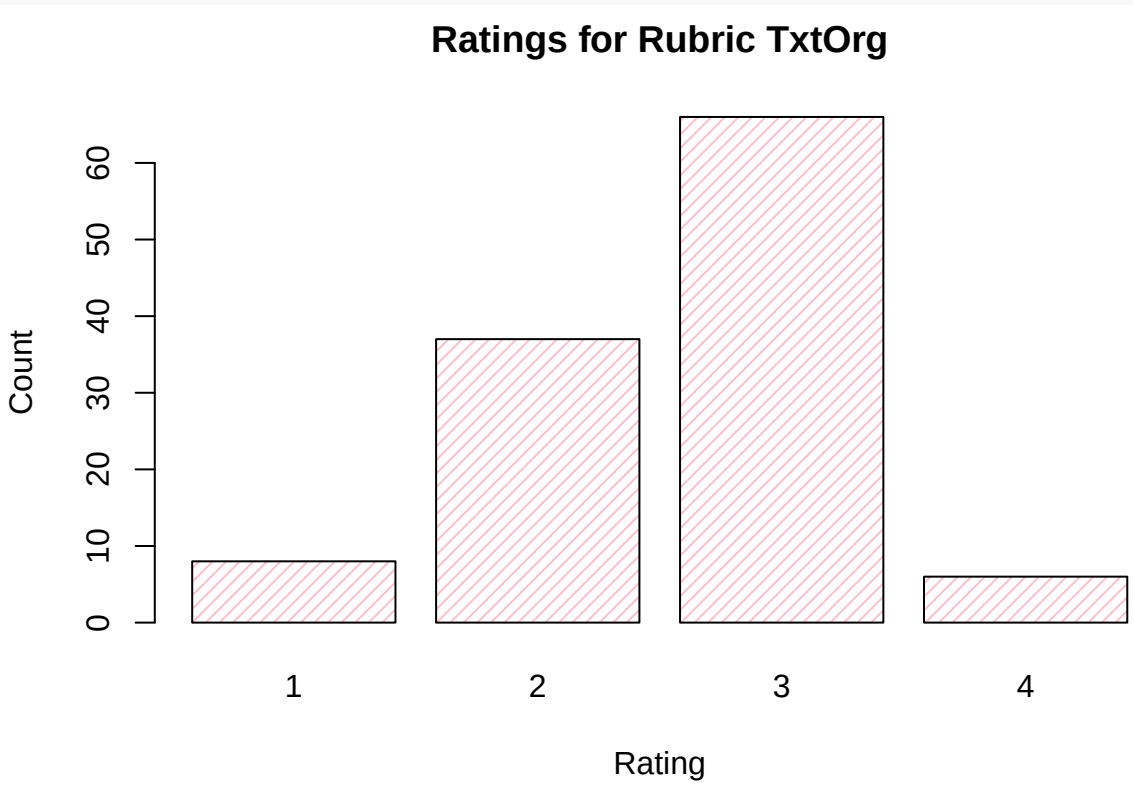
```
barplot(table(ratings$InterpRes), main = "Ratings for Rubric InterpRes", xlab = "Rating", ylab = "Count
```




```
barplot(table(ratings$VisOrg), main = "Ratings for Rubric VisOrg", xlab = "Rating", ylab = "Count", col=
```

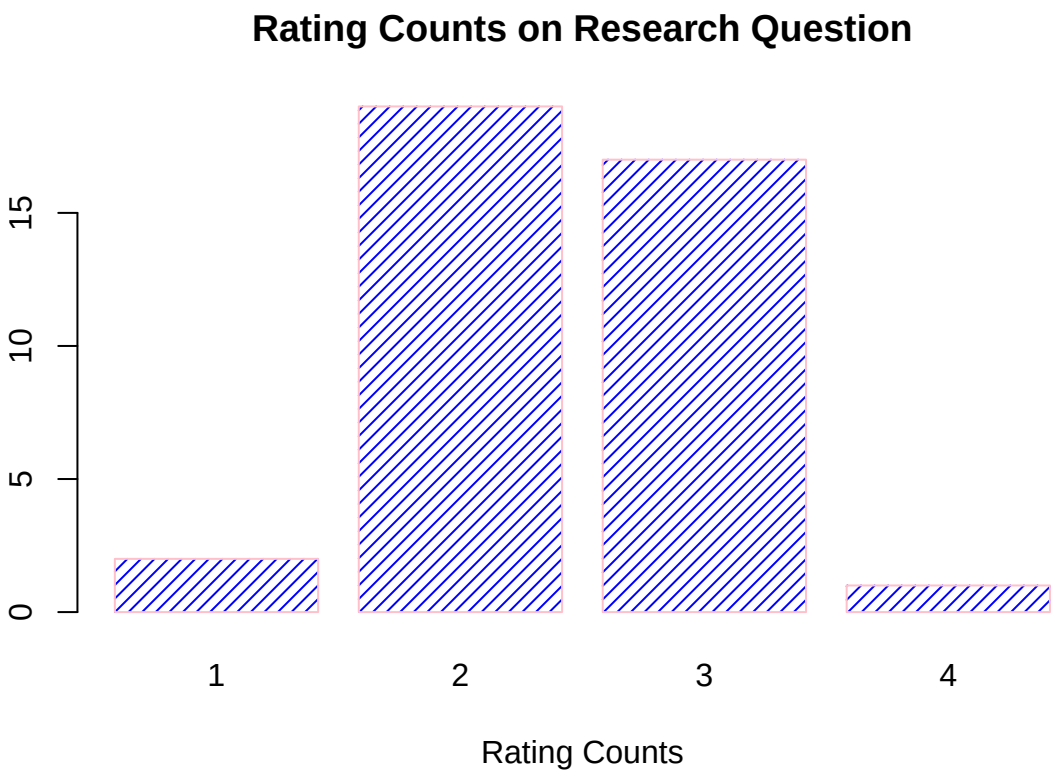


```
barplot(table(ratings$TxtOrg), main = "Ratings for Rubric TxtOrg", xlab = "Rating", ylab = "Count", col=
```

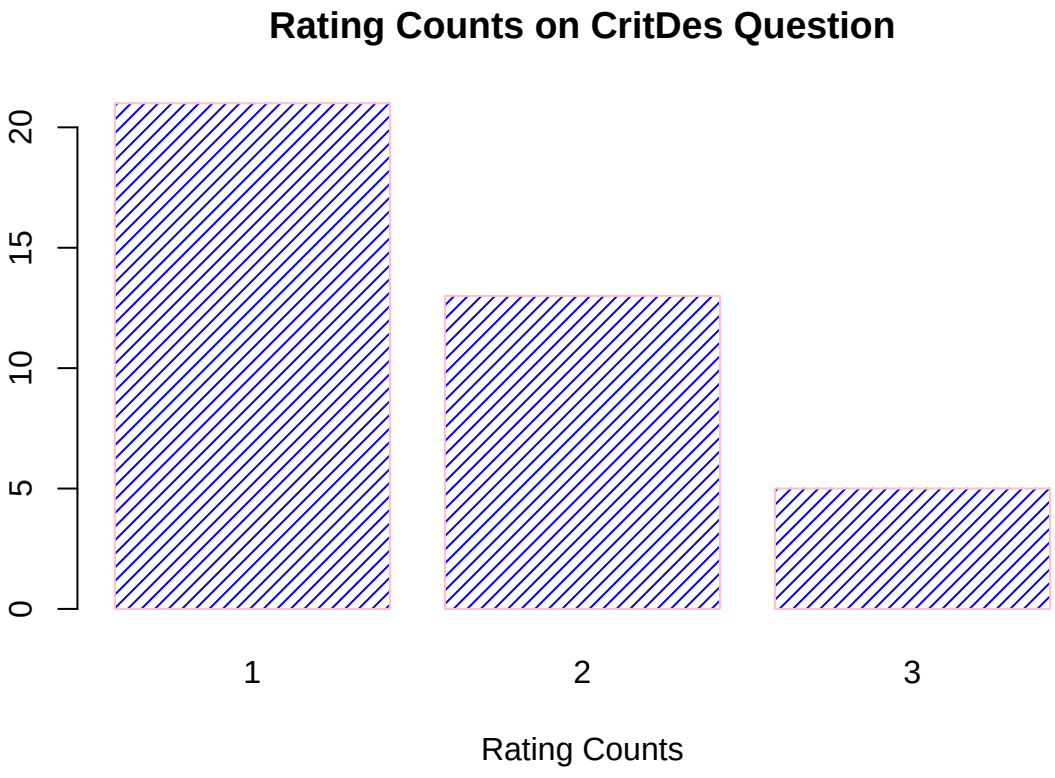


Histograms shows that each Rurbic tends to have different distribution. RsrchQ, InitEDA, InterpRes, VisOrg and TxtOrg have rather similar distributions. Most of the ratings are around 2 for SelMeth, The distribution of CritDes is skewed to the right as there are much more lower ratings for CritDes.

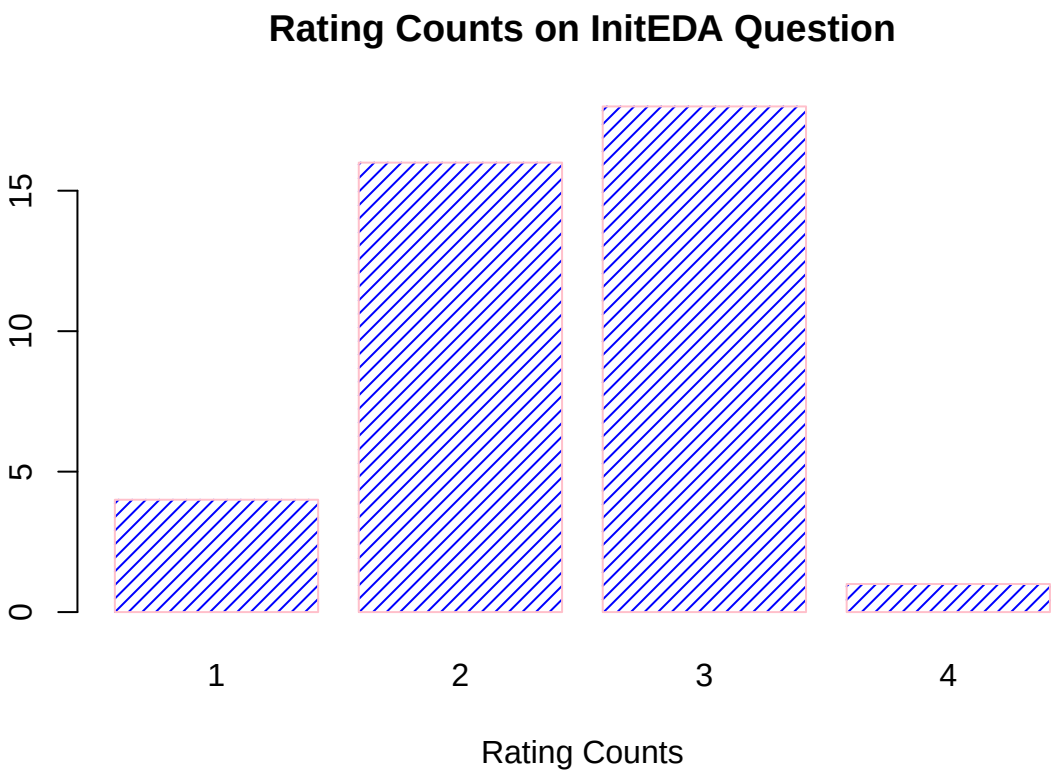
```
rating1 <- ratings %>% filter(ratings$Rater==1)
barplot(table(rating1$RsrchQ), main = "Rating Counts on Research Question", xlab = "Rating Counts", bor
```



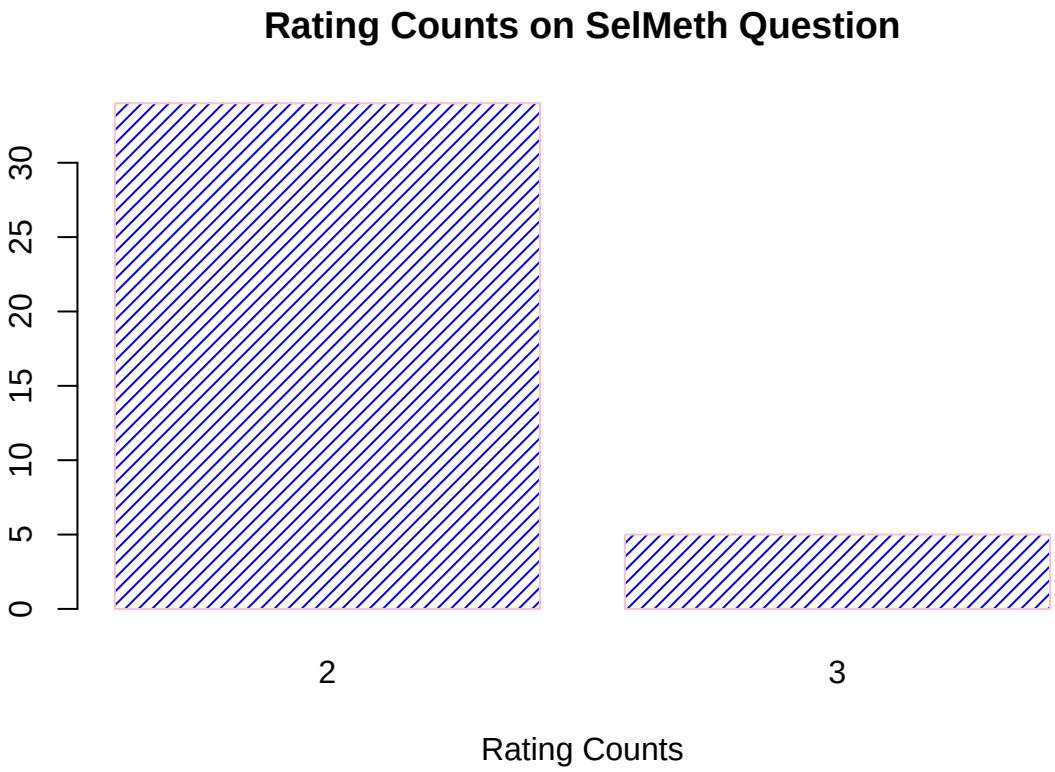

```
barplot(table(rating1$CritDes), main = "Rating Counts on CritDes Question", xlab = "Rating Counts", bor
```



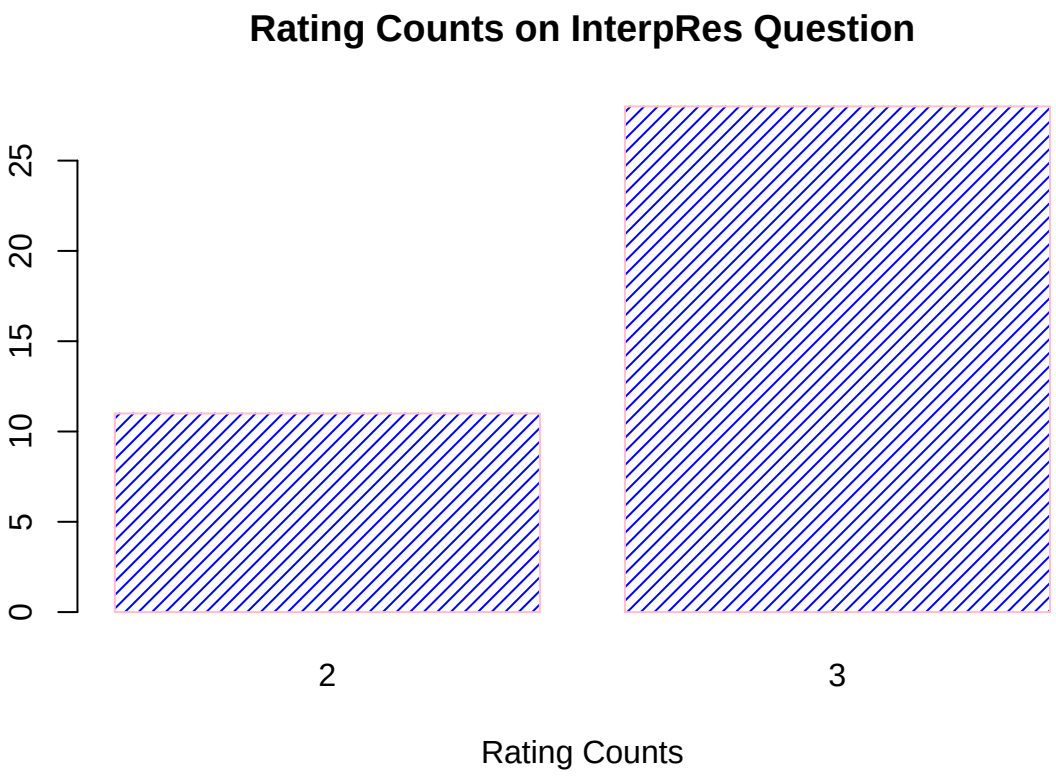

```
barplot(table(rating1$InitEDA), main = "Rating Counts on InitEDA Question", xlab = "Rating Counts", bor
```



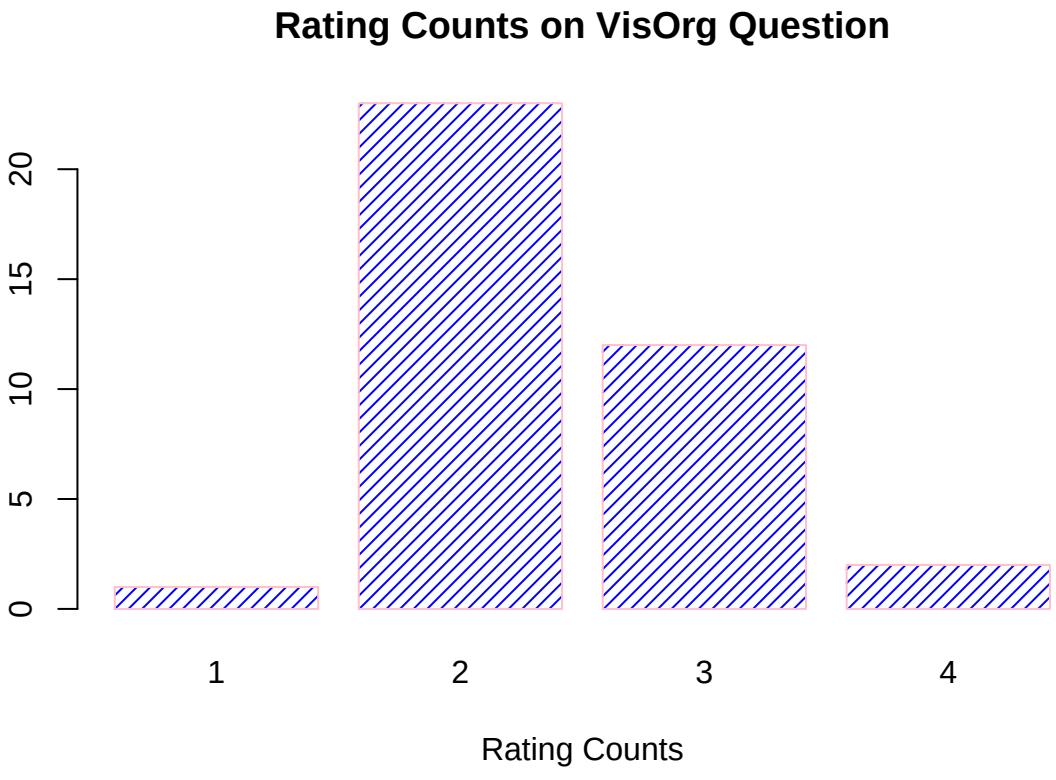

```
barplot(table(rating1$SelMeth), main = "Rating Counts on SelMeth Question", xlab = "Rating Counts", bor
```



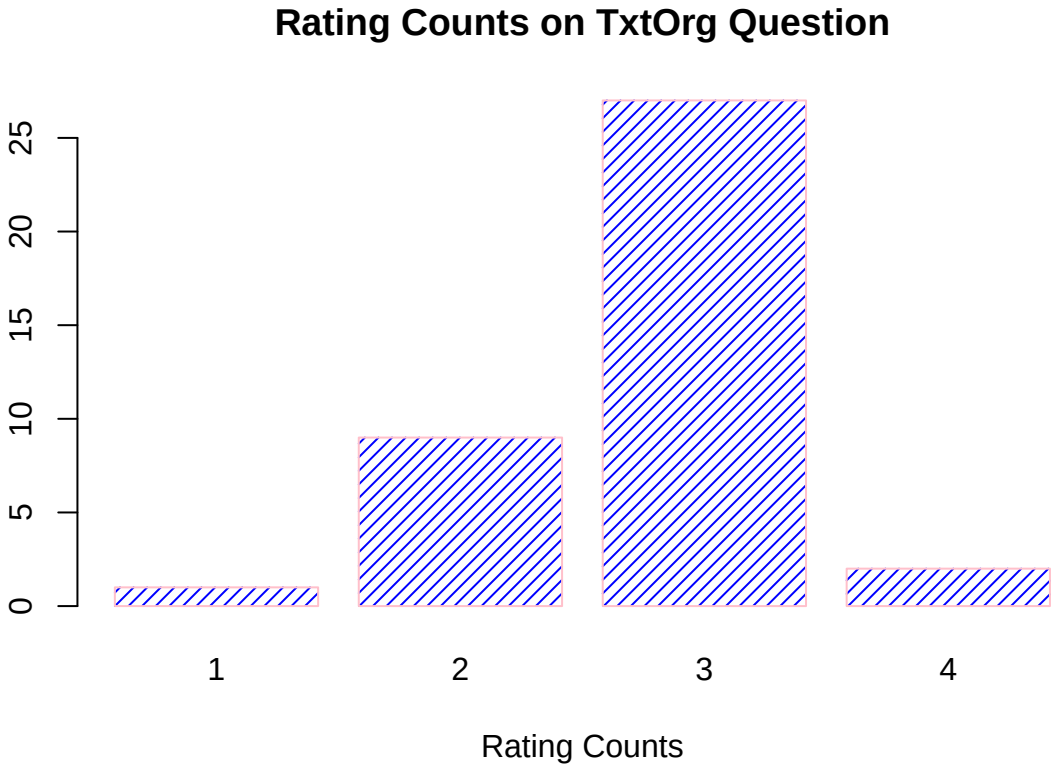

```
barplot(table(rating1$InterpRes), main = "Rating Counts on InterpRes Question", xlab = "Rating Counts",
```




```
barplot(table(rating1$VisOrg), main = "Rating Counts on VisOrg Question", xlab = "Rating Counts", borde
```

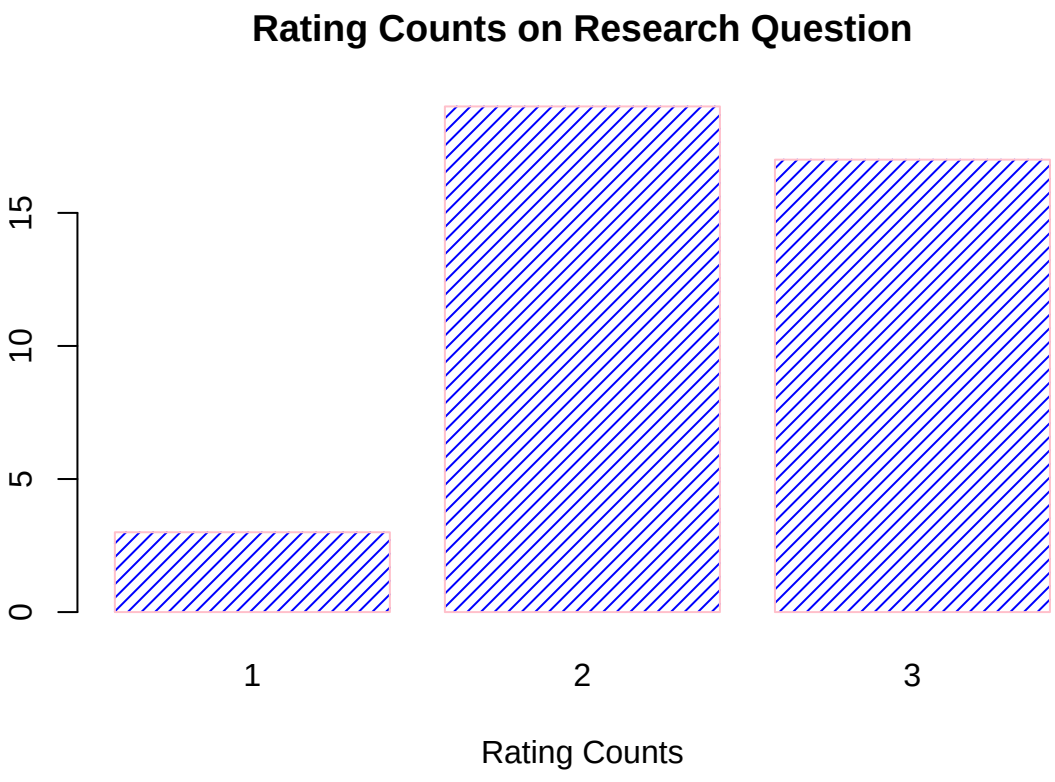



```
barplot(table(rating1$TxtOrg), main = "Rating Counts on TxtOrg Question", xlab = "Rating Counts", borde
```

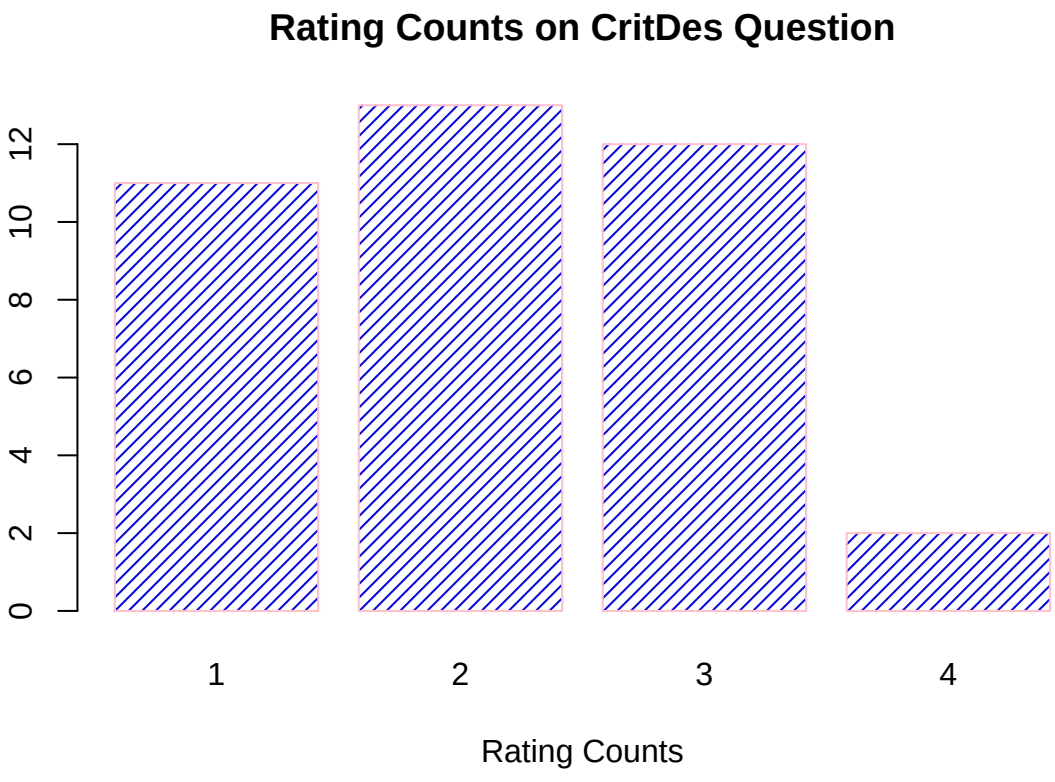


For rater 1, the distribution of ratings on different rubrics are different. Rubrics Research question, InitEDA, VisOrg and TxtOrg all have four ratings and most of the ratings have 2 and 3s. Rater 1's ratings on CritDes question is mostly 1s and the second most ratings are 2, and the least one is 3. For rubric SelMeth, there are only two ratings 2 and 3. There are much more 2s than 3s. The rubric InterpRes also only has two ratings 2 and 3. There are more 3s than 2s.

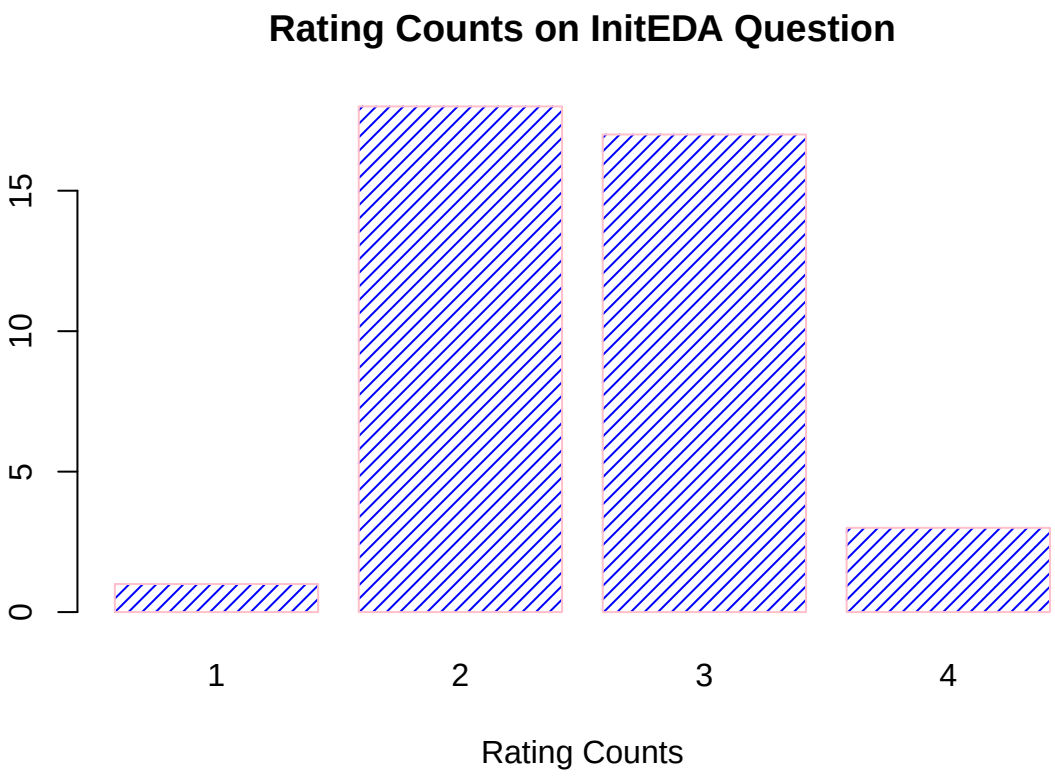
```
rating2 <- ratings %>% filter(ratings$Rater==2)
barplot(table(rating2$RsrchQ), main = "Rating Counts on Research Question", xlab = "Rating Counts", bor
```



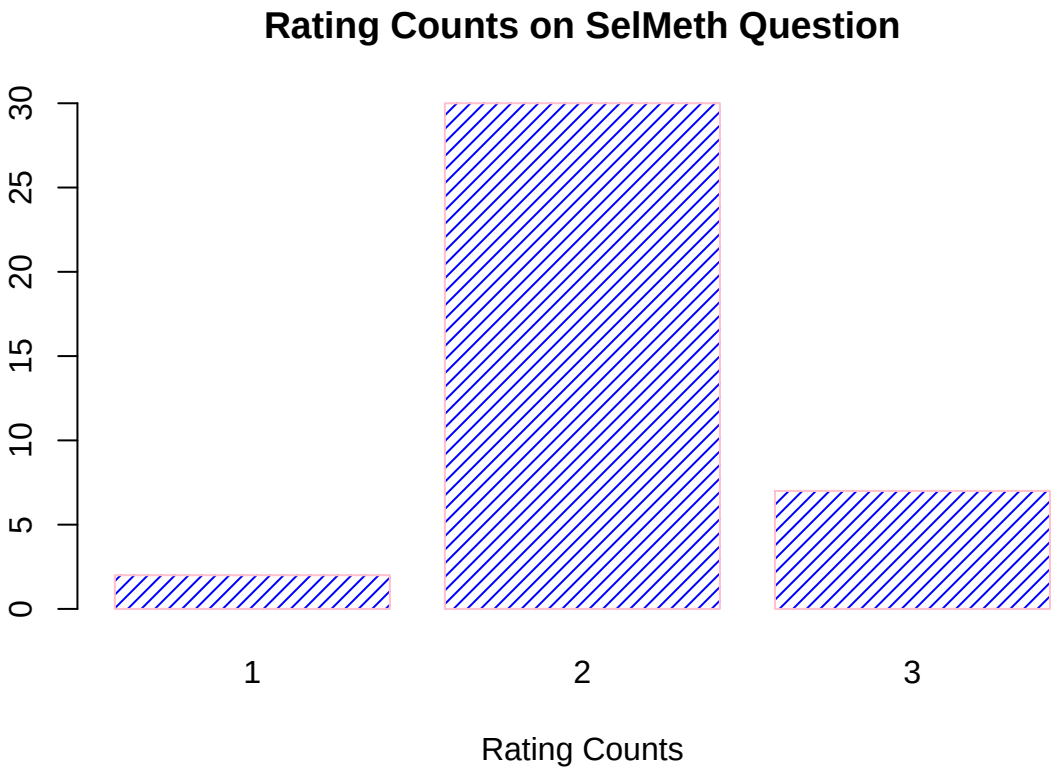

```
barplot(table(rating2$CritDes), main = "Rating Counts on CritDes Question", xlab = "Rating Counts", bor
```



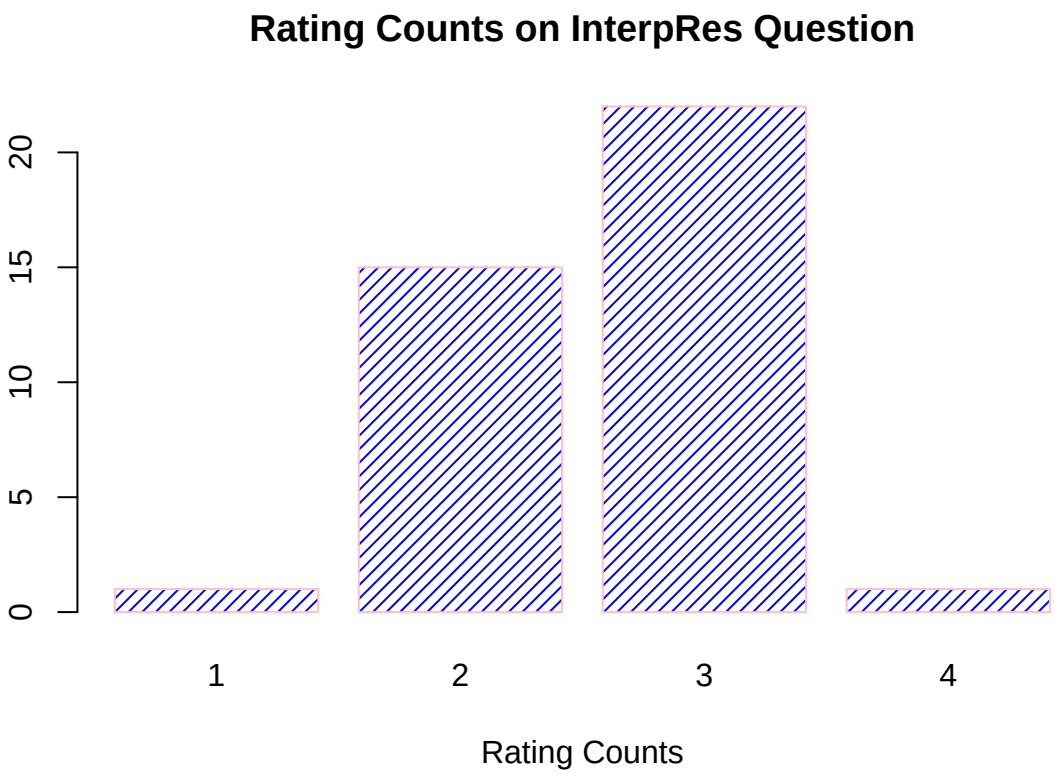

```
barplot(table(rating2$InitEDA), main = "Rating Counts on InitEDA Question", xlab = "Rating Counts", bor
```



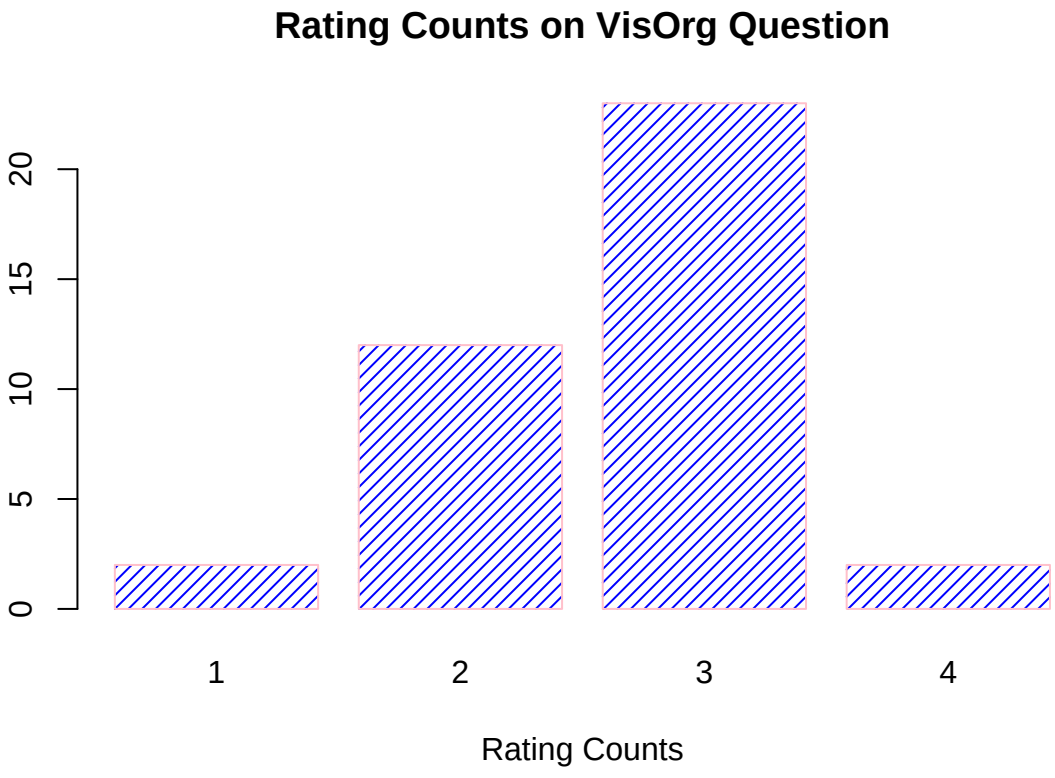

```
barplot(table(rating2$SelMeth), main = "Rating Counts on SelMeth Question", xlab = "Rating Counts", bor
```



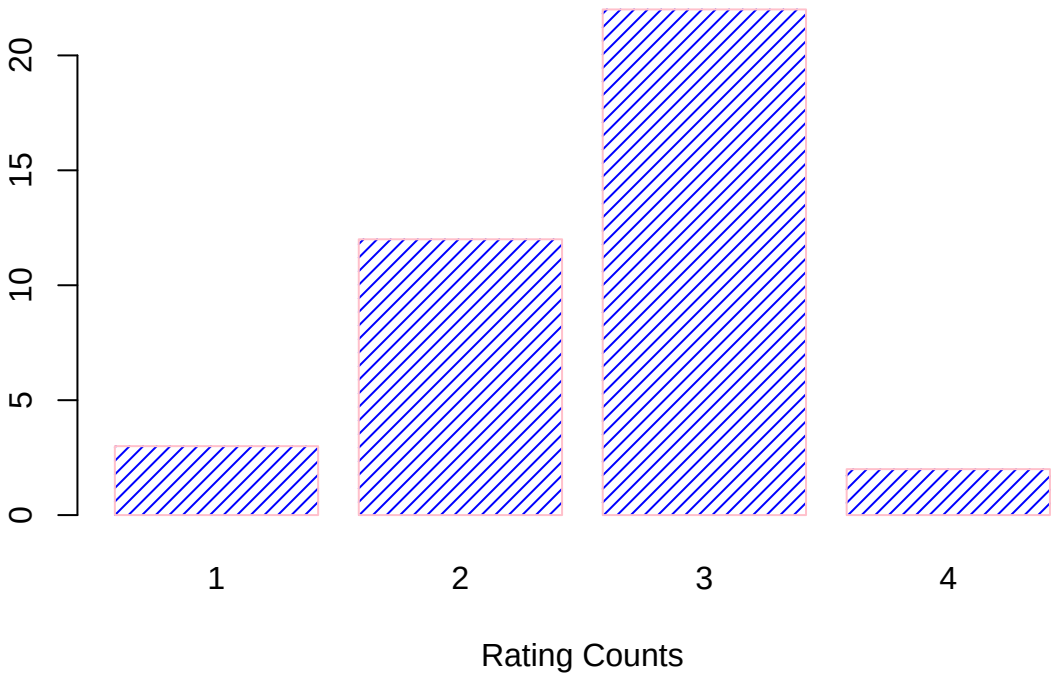

```
barplot(table(rating2$InterpRes), main = "Rating Counts on InterpRes Question", xlab = "Rating Counts",
```




```
barplot(table(rating2$VisOrg), main = "Rating Counts on VisOrg Question", xlab = "Rating Counts", borde
```

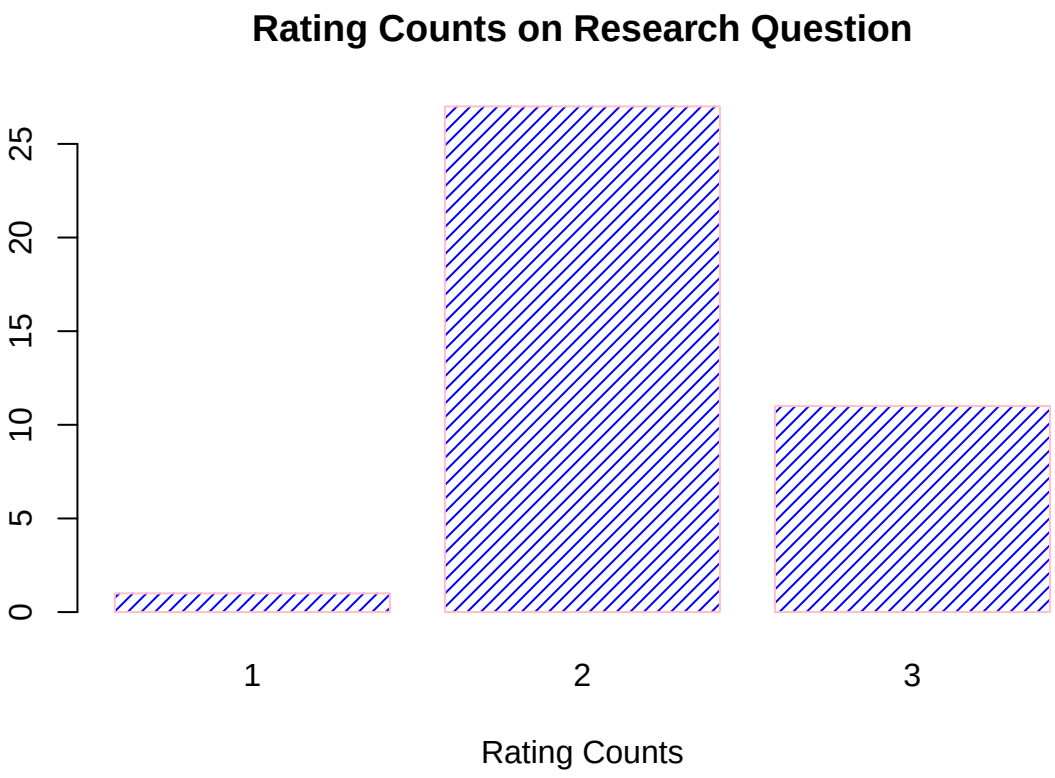



```
barplot(table(rating2$TxtOrg), main = "Rating Counts on TxtOrg Question", xlab = "Rating Counts", borde
```

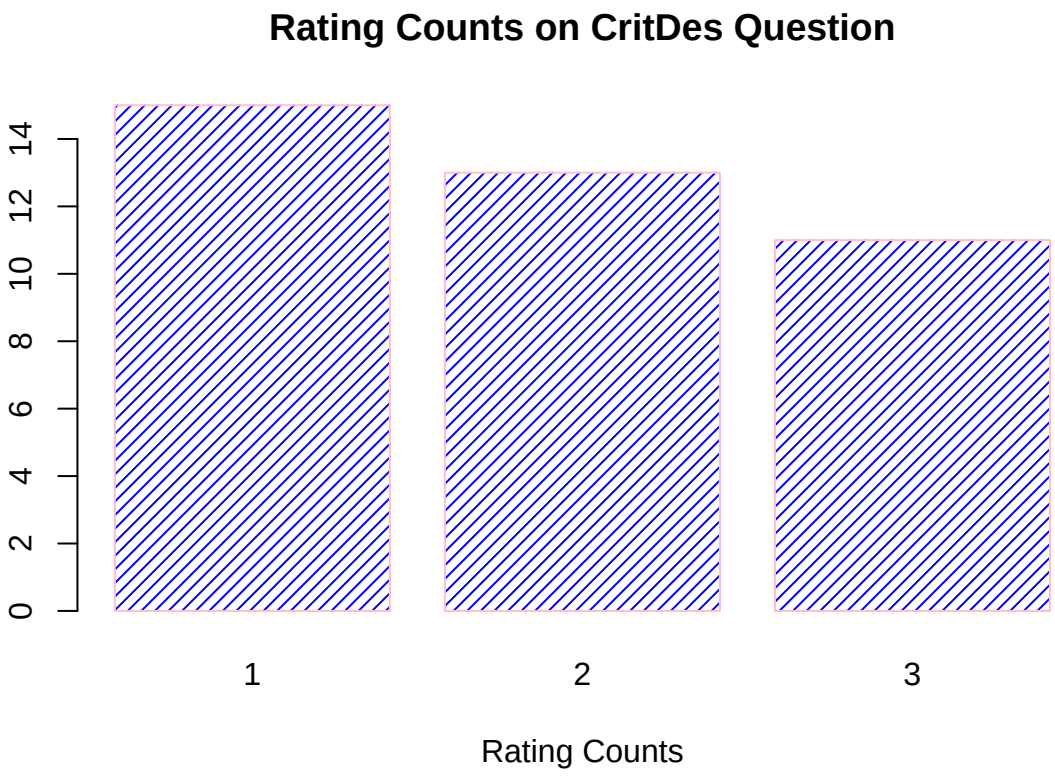


For rater 2, rubric research question has most ratings 2 and the next most ratings is 3 and the least rating is 1. For rubric CritDes, most ratings are 1, 2, and 3, and 4 is the least. For rubric InitEDA, most ratings are 2 and 3, and there are only 1 and 4. For rubric SelMeth, most of the ratings are 2 and there are much less 1s and 3s. For InterpRes, VisOrg and Txt the most rating is 3 and 2, there are only very few 1 and 4.

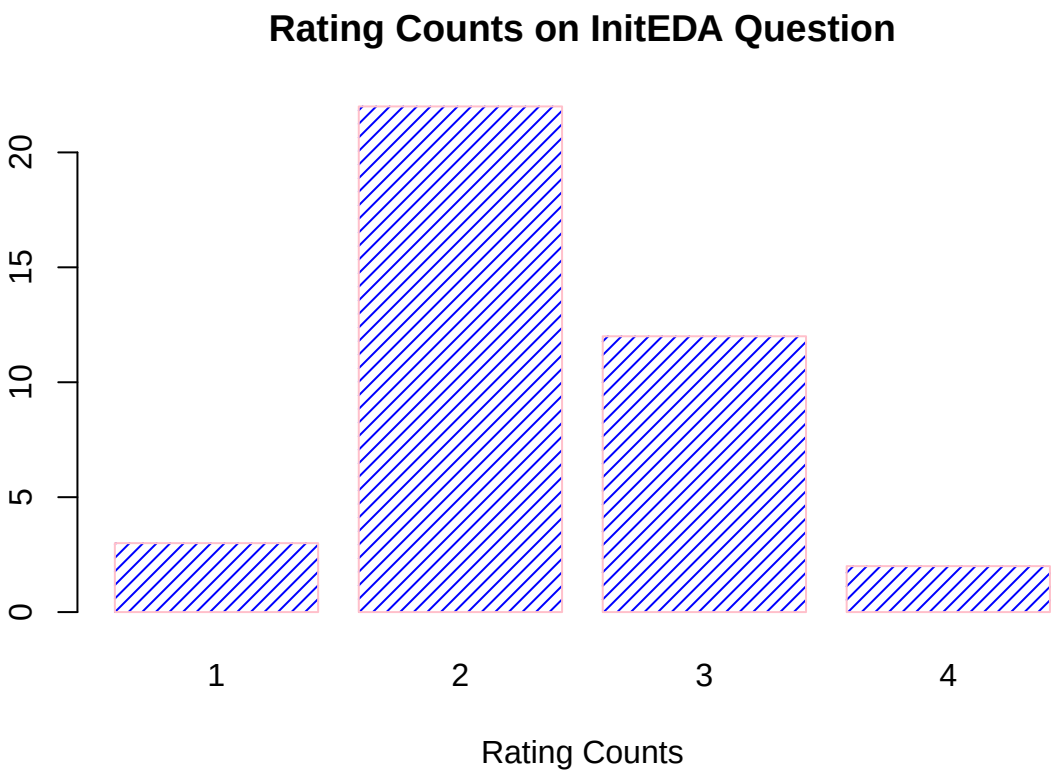
```
rating3 <- ratings %>% filter(ratings$Rater==3)
barplot(table(rating3$RsrchQ), main = "Rating Counts on Research Question", xlab = "Rating Counts", bor
```



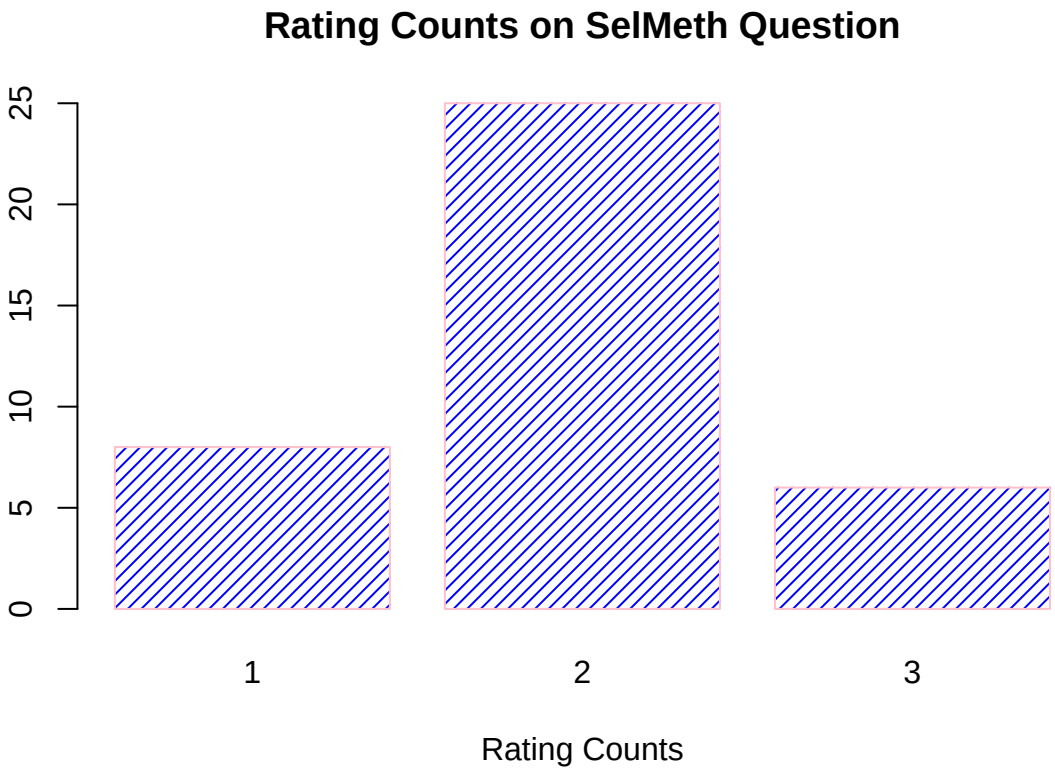

```
barplot(table(rating3$CritDes), main = "Rating Counts on CritDes Question", xlab = "Rating Counts", bor
```



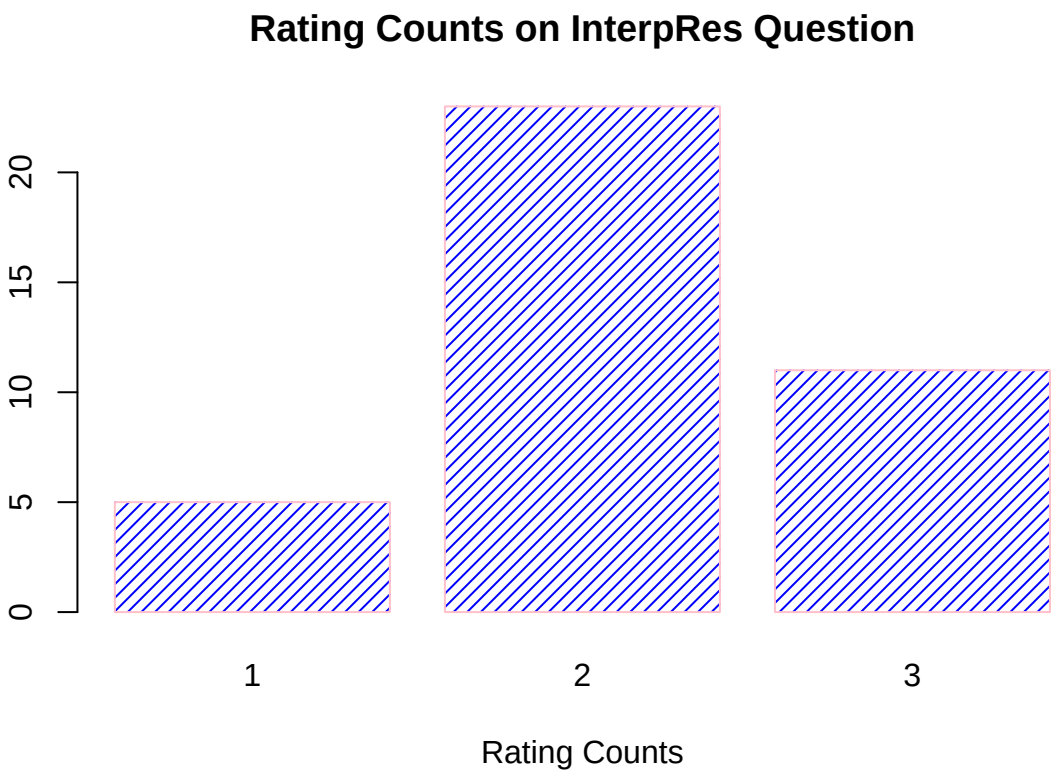

```
barplot(table(rating3$InitEDA), main = "Rating Counts on InitEDA Question", xlab = "Rating Counts", bor
```



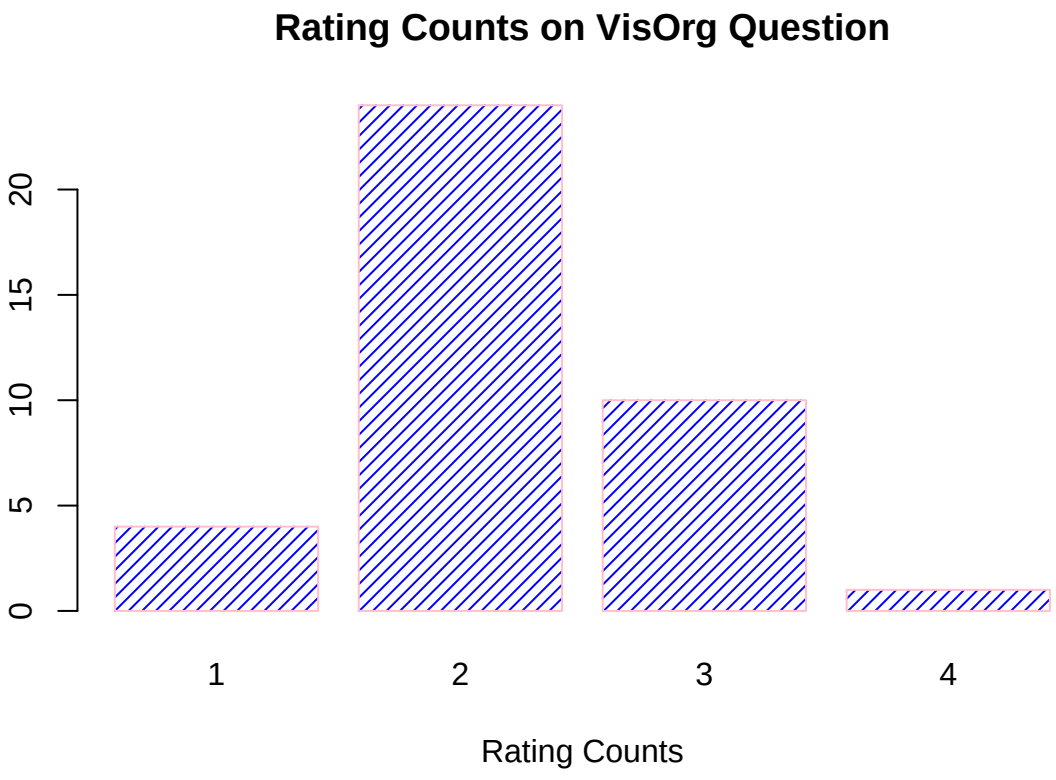

```
barplot(table(rating3$SelMeth), main = "Rating Counts on SelMeth Question", xlab = "Rating Counts", bor
```



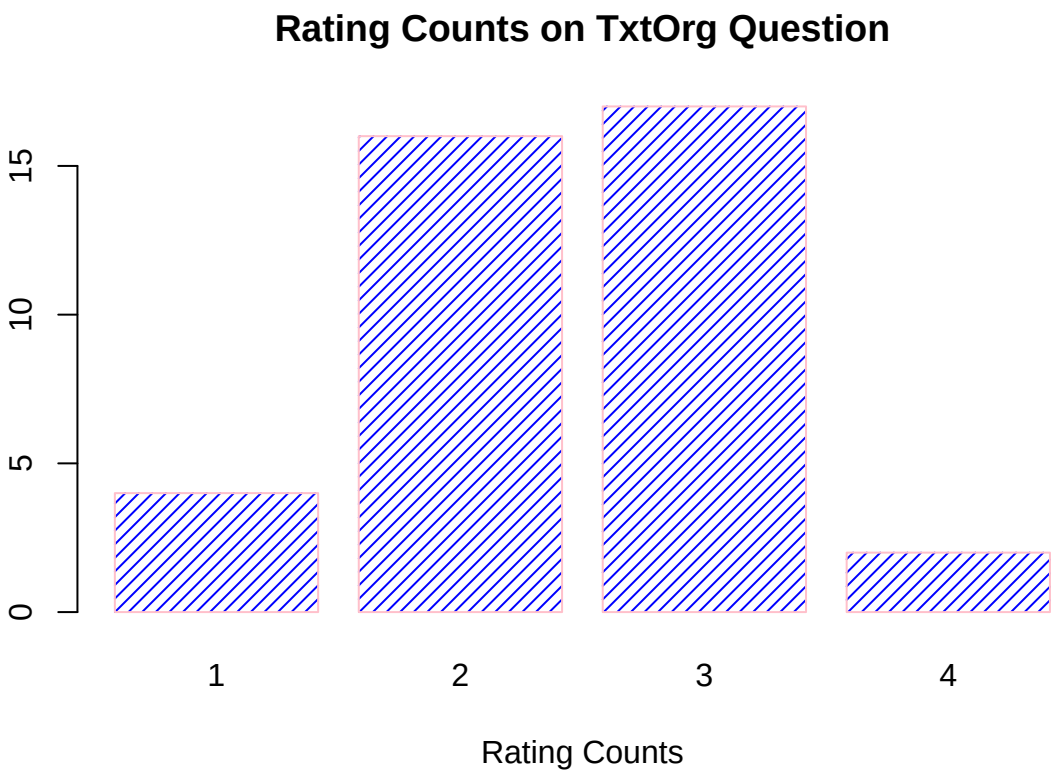

```
barplot(table(rating3$InterpRes), main = "Rating Counts on InterpRes Question", xlab = "Rating Counts",
```



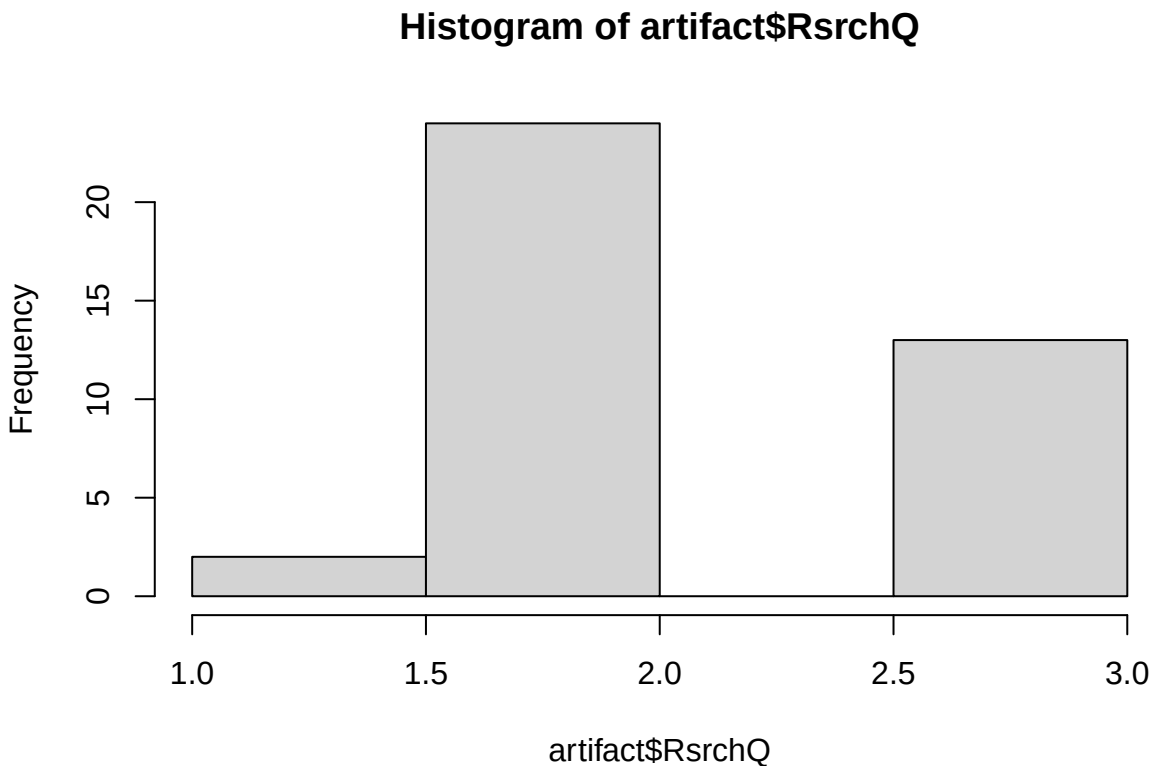

```
barplot(table(rating3$VisOrg), main = "Rating Counts on VisOrg Question", xlab = "Rating Counts", borde
```



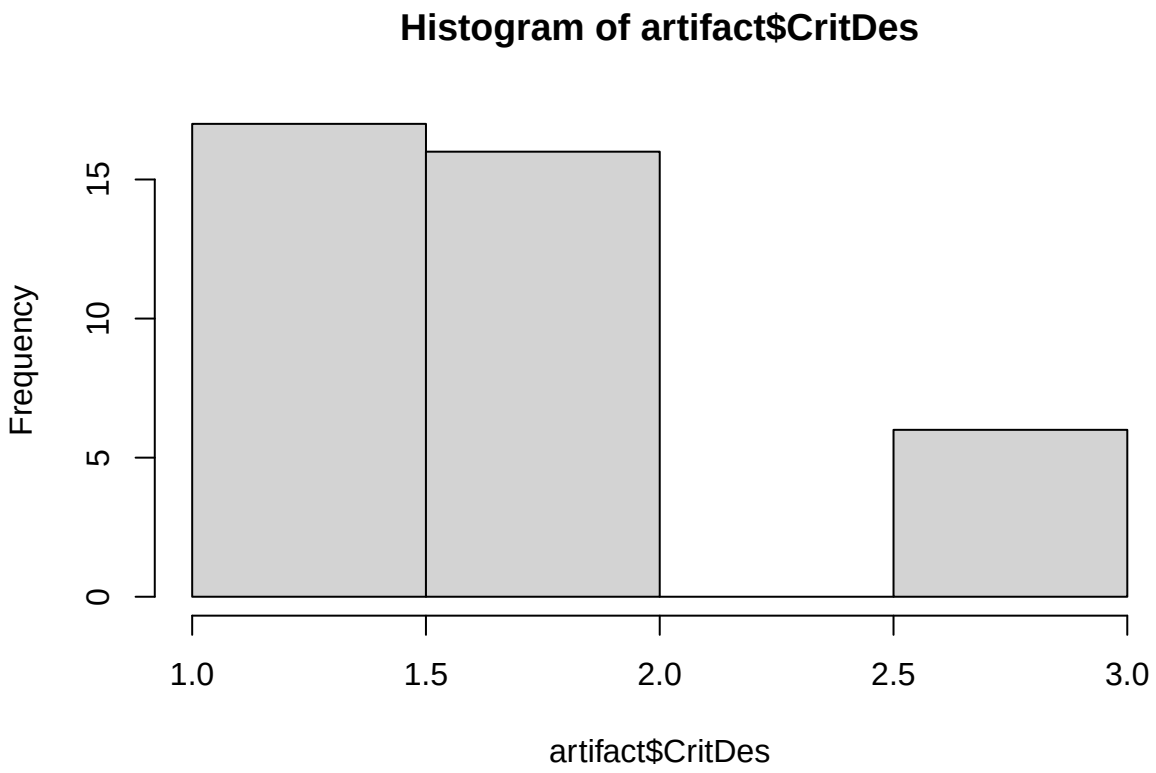

```
barplot(table(rating3$TxtOrg), main = "Rating Counts on TxtOrg Question", xlab = "Rating Counts", borde
```



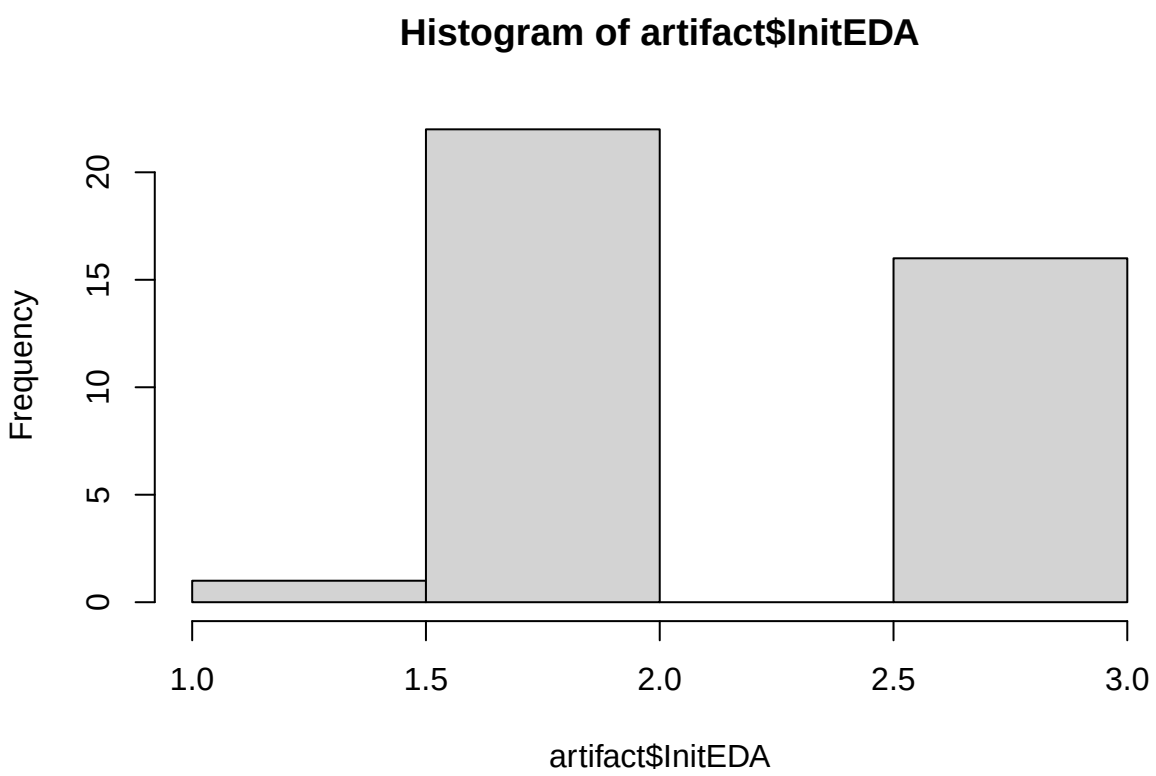

```
artifact <- ratings[grep("O",ratings$Artifact),]
hist(artifact$RsrchQ)
```



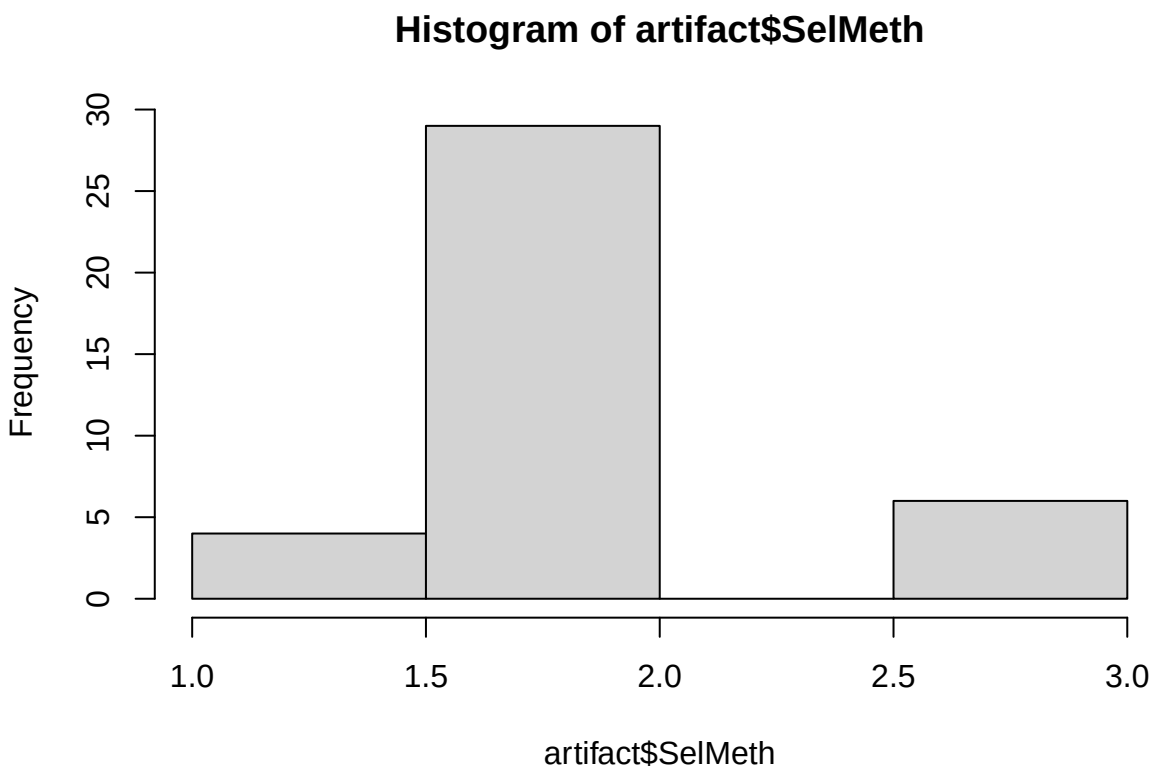
```
hist(artifact$CritDes)
```



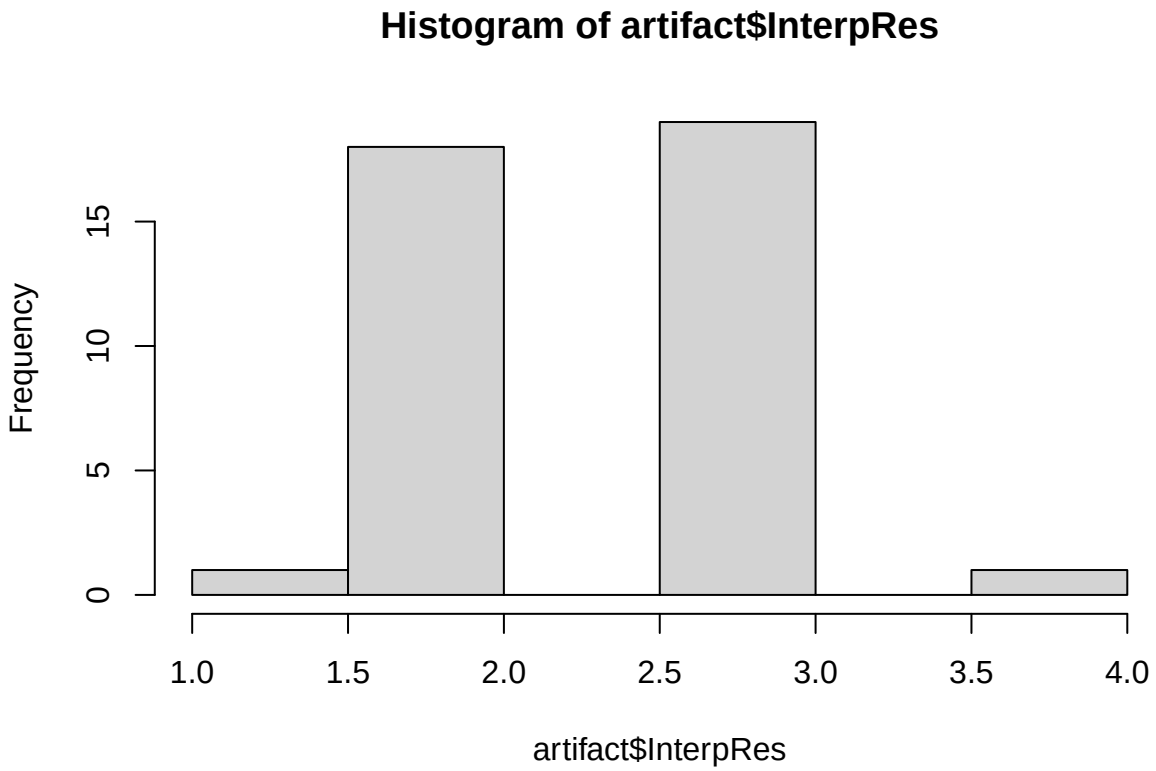

```
hist(artifact$InitEDA)
```



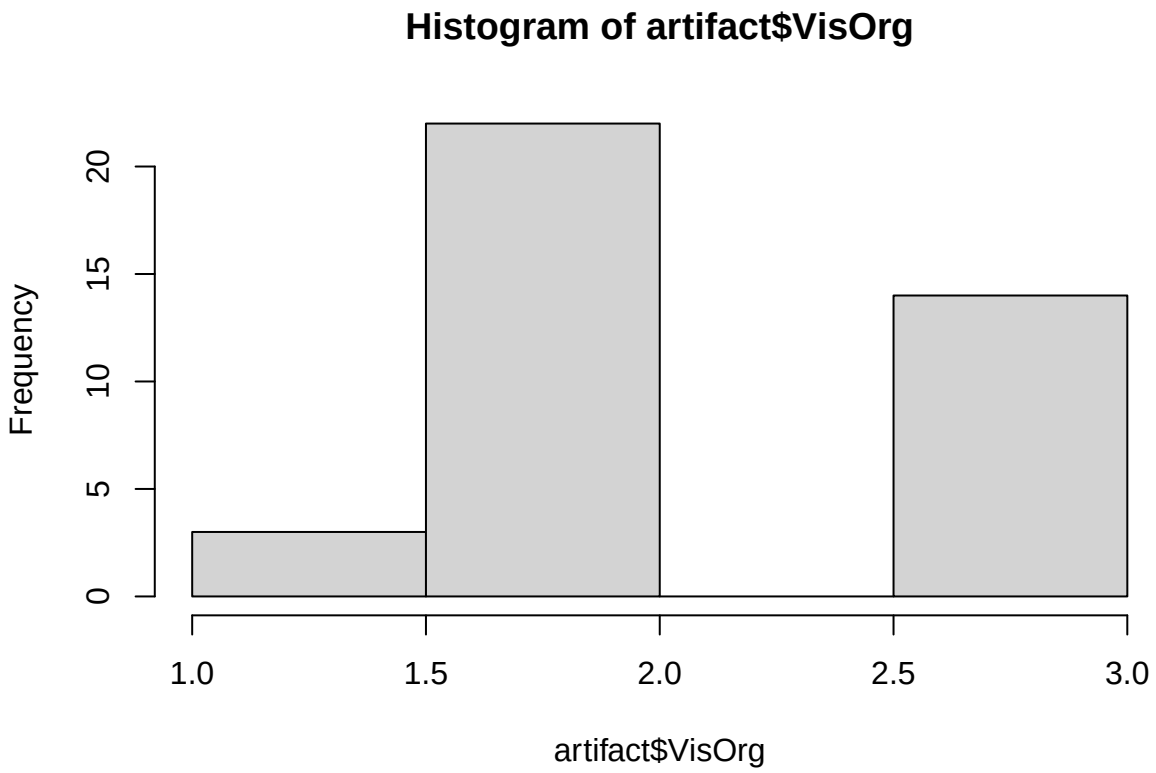
```
hist(artifact$SelMeth)
```



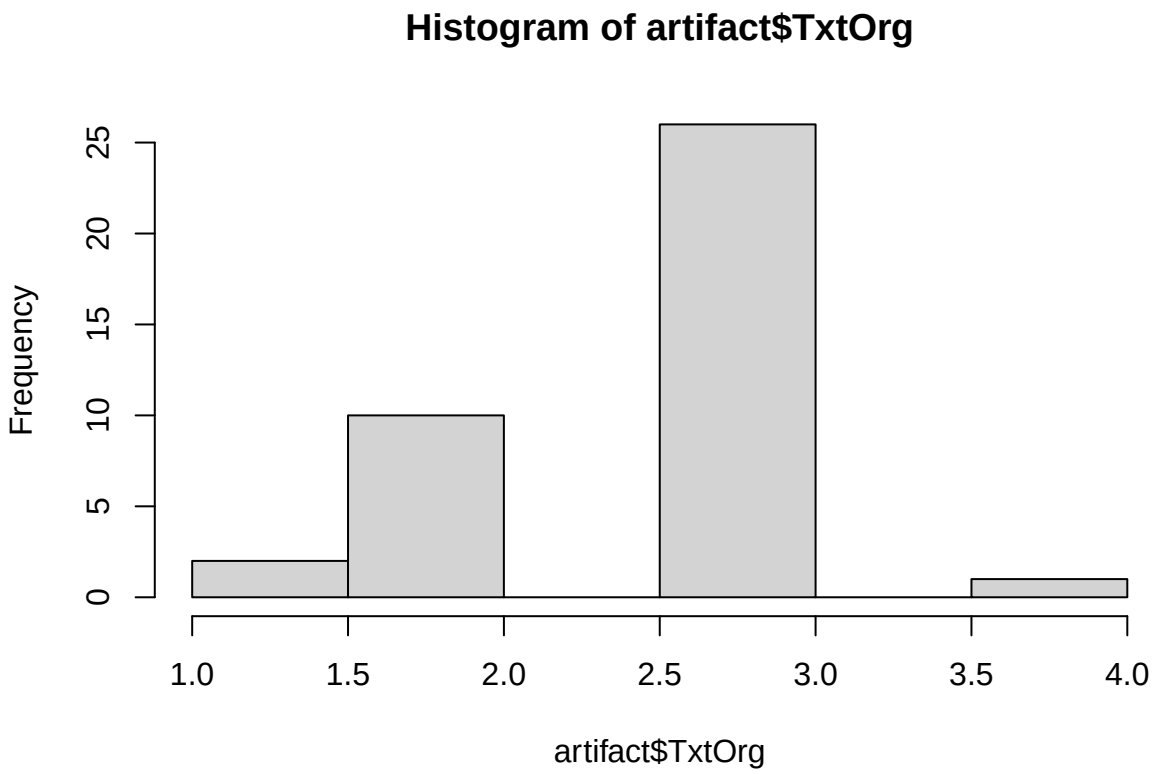

```
hist(artifact$InterpRes)
```



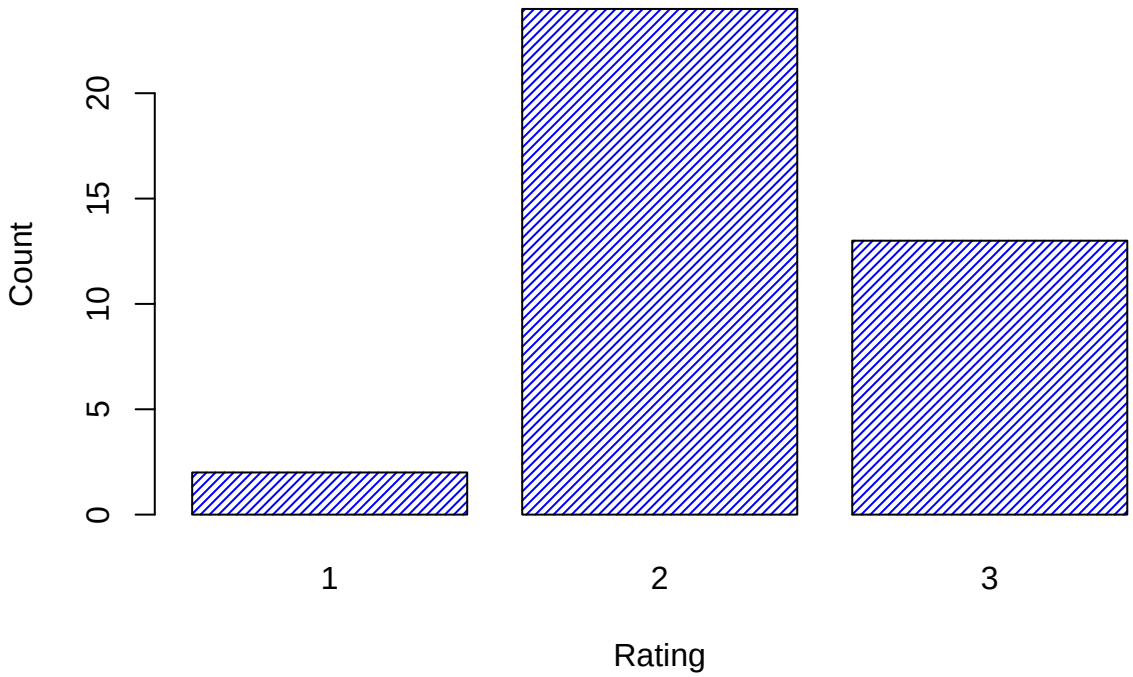
```
hist(artifact$VisOrg)
```



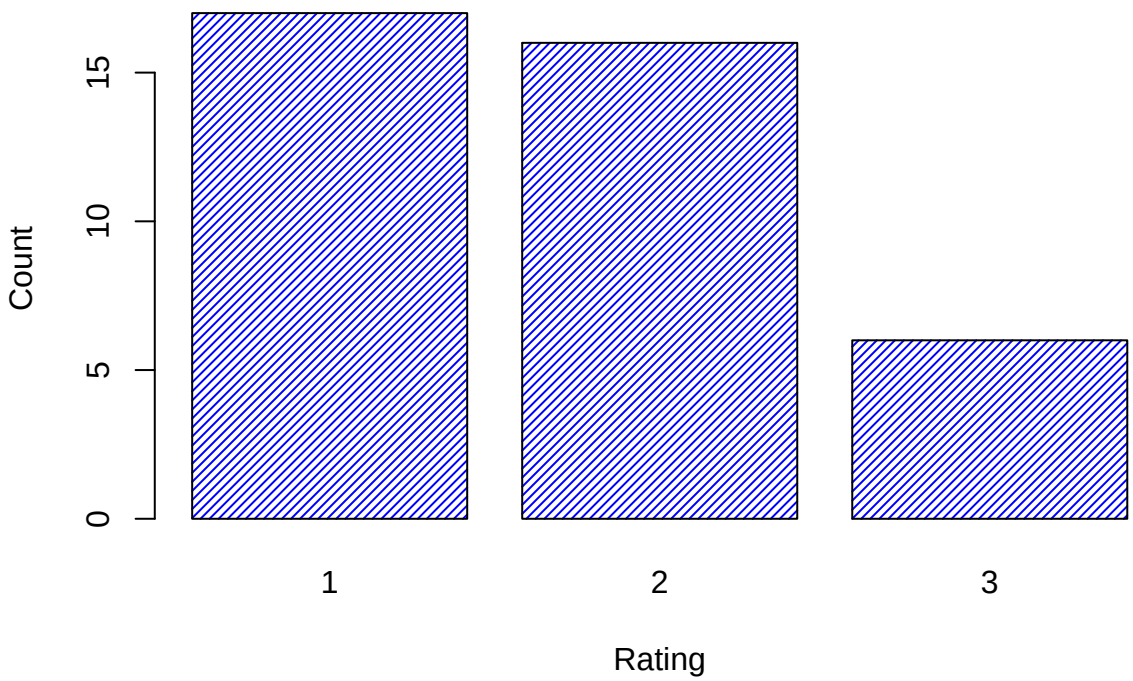

```
hist(artifact$TxtOrg)
```



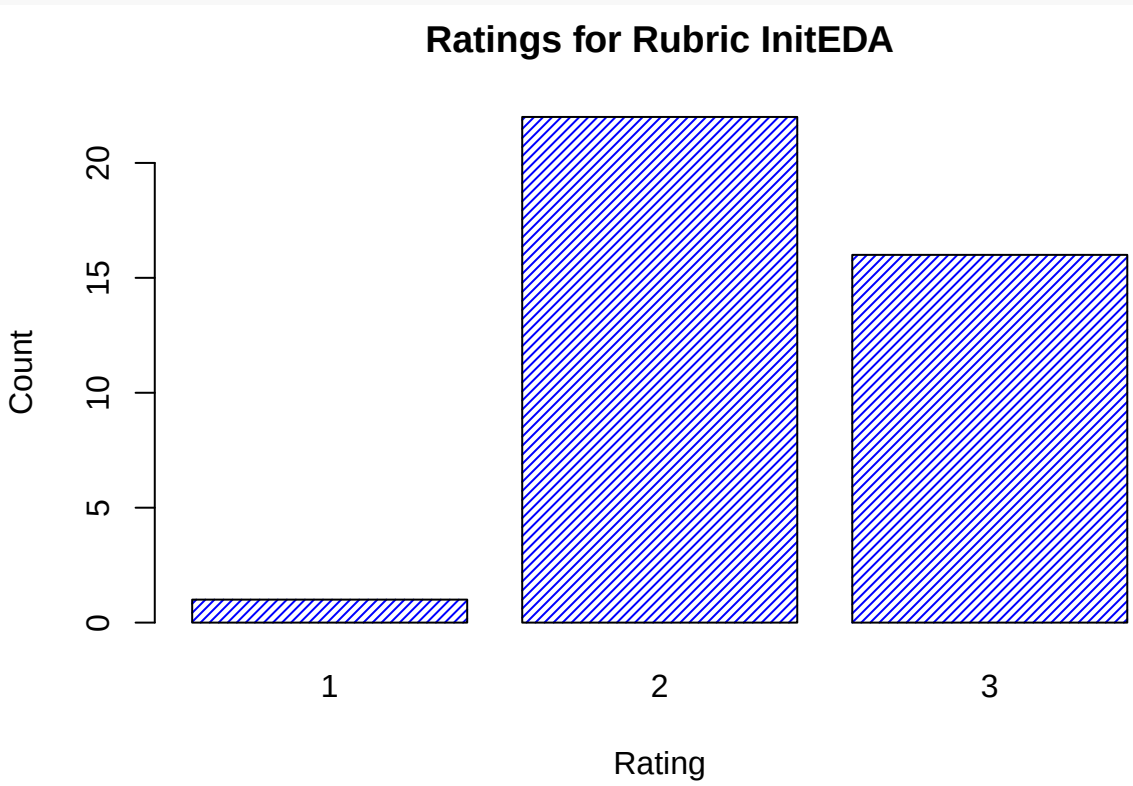
```
barplot(table(artifact$RsrchQ), main = "Ratings for Rubric RsrchQ", xlab = "Rating", ylab = "Count", col
```



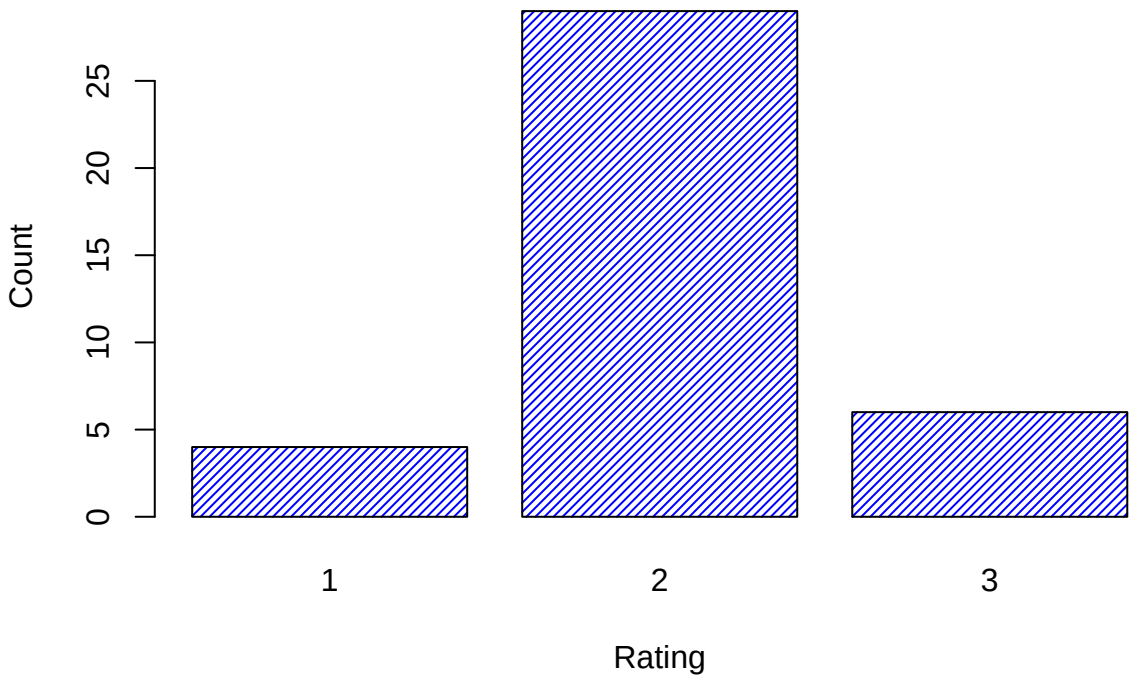

```
barplot(table(artifact$CritDes), main = "Ratings for Rubric CritDes", xlab = "Rating", ylab = "Count", c
```



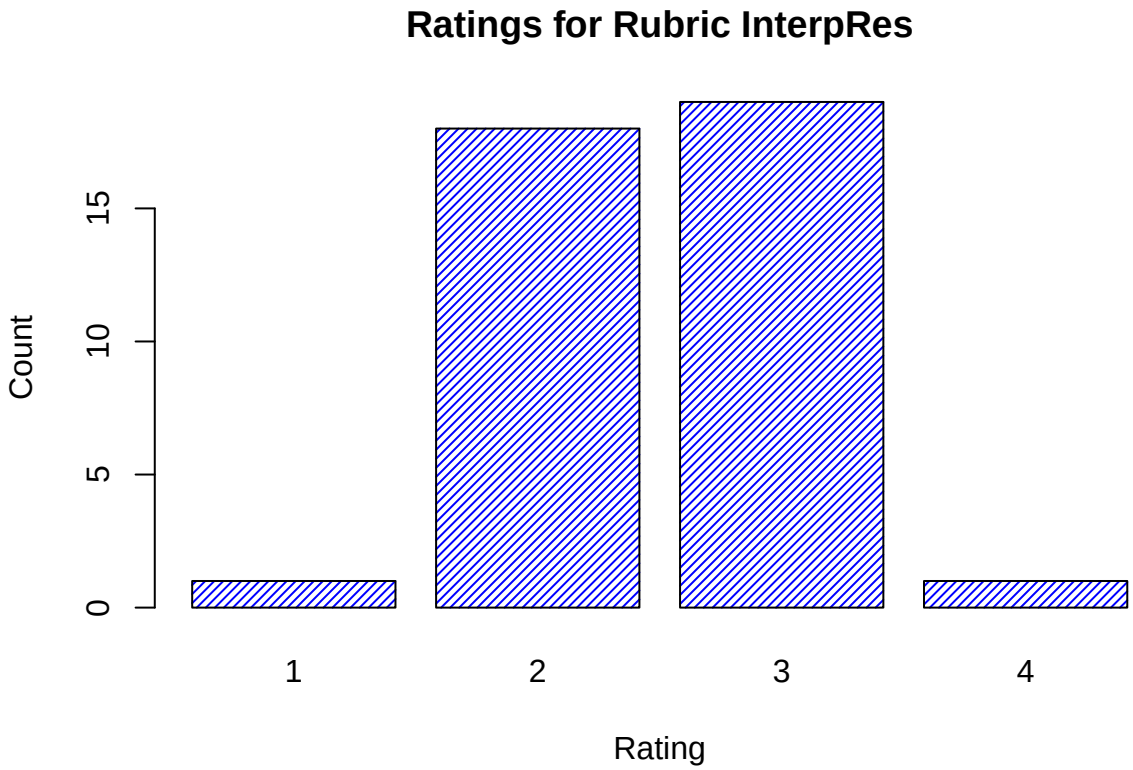
```
barplot(table(artifact$InitEDA), main = "Ratings for Rubric InitEDA", xlab = "Rating", ylab = "Count",
```



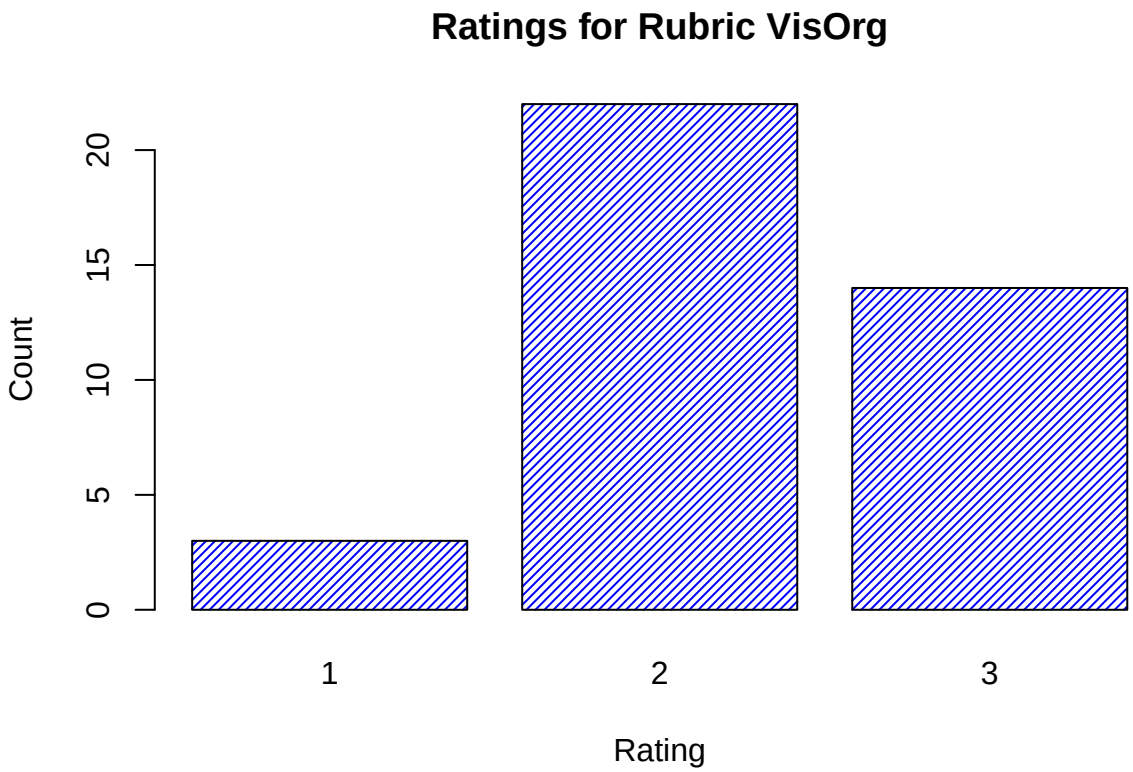

```
barplot(table(artifact$SelMeth), main = "Ratings for Rubric SelMeth", xlab = "Rating", ylab = "Count",
```



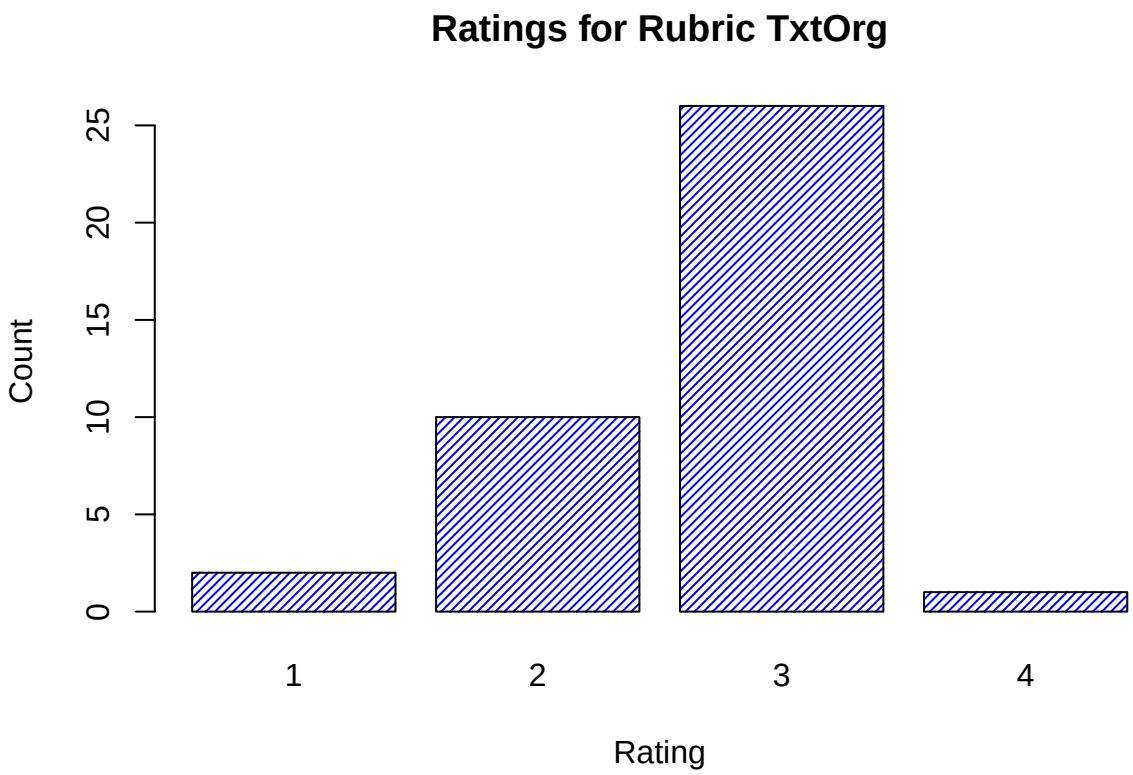
```
barplot(table(artifact$InterpRes), main = "Ratings for Rubric InterpRes", xlab = "Rating", ylab = "Count
```




```
barplot(table(artifact$VisOrg), main = "Ratings for Rubric VisOrg", xlab = "Rating", ylab = "Count", col
```

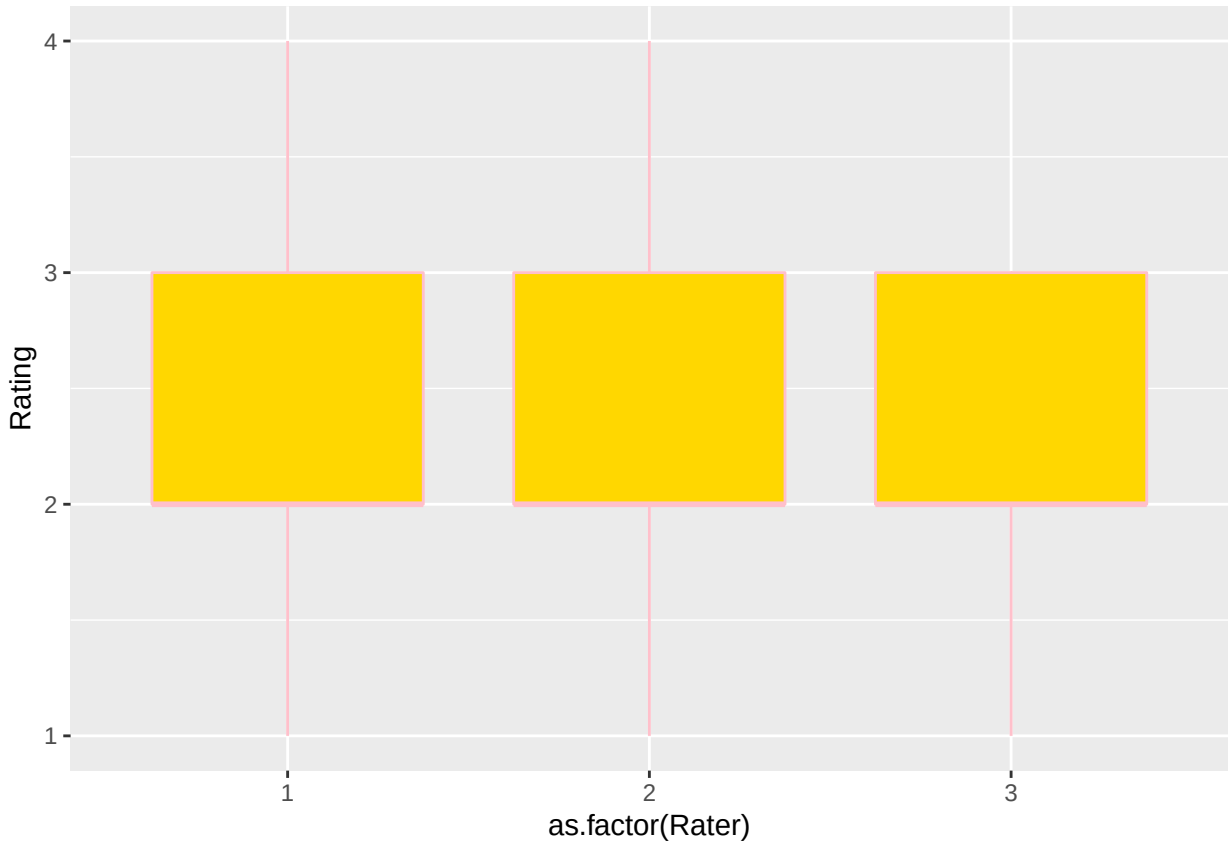


```
barplot(table(artifact$TxtOrg), main = "Ratings for Rubric TxtOrg", xlab = "Rating", ylab = "Count", col
```



Now I will take a closer look at the subset of 13 artifacts that are viewed by all raters. The distribution of ratings is very different from that on the entire dataset. Each rater is very different from their ratings. Ratings for rubric rsrchQ, ratings for Rubric InitEDA and ratinsg for rubric Visorg are very similar to each other, as rating 2 has the most counts and 3 has less count, and there are only a few 1s. Rubric CritDes has lower ratings as there are more 1s and 2s. Rubric InterpRes and TxtOrg both have 4 ratings and most are centered aroung 2 and 3.

```
common <- tall[grep("O",tall$Artifact),]
ggplot(common, aes(x = as.factor(Rater), y = Rating)) +
        geom_boxplot(fill = "gold1", color = "pink")
```



2

```
tall <- tall %>% na.omit()
summary(tall)
```

```
##        X           Rater         Artifact            Repeated
##  Min.   :  1   Min.   :1.000   Length:817         Min.   :0.0000
##  1st Qu.:206   1st Qu.:1.000   Class :character   1st Qu.:0.0000
##  Median :410   Median :2.000   Mode  :character   Median :0.0000
##  Mean   :410   Mean   :2.001                      Mean   :0.3341
##  3rd Qu.:614   3rd Qu.:3.000                      3rd Qu.:1.0000
##  Max.   :819   Max.   :3.000                      Max.   :1.0000
##    Semester             Sex              Rubric             Rating
##  Length:817         Length:817         Length:817         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :2.318
##                                                           3rd Qu.:3.000
##                                                           Max.   :4.000
```

```
head(common)
```

```
##     X Rater Artifact Repeated Semester Sex Rubric Rating
## 1   1     3       O5        1      F19   M RsrchQ      3
## 2   2     3       O7        1      F19   F RsrchQ      3
## 3   3     3       O9        1      S19   F RsrchQ      2
## 4   4     3       O8        1      S19   M RsrchQ      2
```

```
## 10 10     3      O10        1      F19   F RsrchQ      2
## 11 11     3      O13        1      F19   M RsrchQ      2
```

```
dim(common)
```

```
## [1] 273   8
```

```
RsrchQ.ratings <- common[common$Rubric=="RsrchQ",]
CritDes.ratings <- common[common$Rubric == "CritDes",]
InitEDA.ratings <- common[common$Rubric == "InitEDA",]
SelMeth.ratings <- common[common$Rubric == "SelMeth",]
InterpRes.ratings <- common[common$Rubric == "InterpRes",]
VisOrg.ratings <- common[common$Rubric == "VisOrg",]
TxtOrg.ratings <- common[common$Rubric == "TxtOrg",]
```

```
summary(tall)
```

```
##        X           Rater         Artifact            Repeated
##  Min.   :  1   Min.   :1.000   Length:817         Min.   :0.0000
##  1st Qu.:206   1st Qu.:1.000   Class :character   1st Qu.:0.0000
##  Median :410   Median :2.000   Mode  :character   Median :0.0000
##  Mean   :410   Mean   :2.001                      Mean   :0.3341
##  3rd Qu.:614   3rd Qu.:3.000                      3rd Qu.:1.0000
##  Max.   :819   Max.   :3.000                      Max.   :1.0000
##    Semester             Sex               Rubric             Rating
##  Length:817         Length:817         Length:817         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :2.318
##                                                           3rd Qu.:3.000
##                                                           Max.   :4.000
```

```
mod1 <- lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ.ratings)
mod2 <- lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)
mod3 <- lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)
mod4 <- lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)
mod5 <- lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)
mod6 <- lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)
mod7 <- lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.ratings)
```

```
RsrchQ.full <- tall[tall$Rubric=="RsrchQ",]
CritDes.full <- tall[tall$Rubric == "CritDes",]
InitEDA.full <- tall[tall$Rubric == "InitEDA",]
SelMeth.full <- tall[tall$Rubric == "SelMeth",]
InterpRes.full <- tall[tall$Rubric == "InterpRes",]
VisOrg.full <- tall[tall$Rubric == "VisOrg",]
TxtOrg.full <- tall[tall$Rubric == "TxtOrg",]
```

```
mod1_1 <- lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ.full)
mod2_1 <- lmer(Rating ~ 1 + (1|Artifact), data=CritDes.full)
mod3_1 <- lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.full)
mod4_1 <- lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.full)
mod5_1 <- lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.full)
mod6_1 <- lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.full)
mod7_1 <- lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.full)
```

```
icc(mod1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.189
##   Conditional ICC: 0.189
```

```
icc(mod2)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.573
##   Conditional ICC: 0.573
```

```
icc(mod3)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.493
##   Conditional ICC: 0.493
```

```
icc(mod4)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.521
##   Conditional ICC: 0.521
```

```
icc(mod5)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.230
##   Conditional ICC: 0.230
```

```
icc(mod6)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.592
##   Conditional ICC: 0.592
```

```
icc(mod7)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.143
##   Conditional ICC: 0.143
```

Just looking at the subset of artifact viewed by all raters, and looking at each rubric separately, Rubric CritDes, InitEDA, SelMeth and VisOrg have high ICC which means high correlation between raters and ratings on the same artifact, which means that the raters are consistent with one another in how they rate. Rubric RsrchQ, InterpRes and TxtOrg all have lower ICC which means low correlation between raters and ratings on the same artifact, which means that the raters are not consistent with one another in how they rate.

```
icc(mod1_1)
```

```
## # Intraclass Correlation Coefficient
##
```

```
##      Adjusted ICC: 0.210
##   Conditional ICC: 0.210
```

```
icc(mod2_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.673
##   Conditional ICC: 0.673
```

```
icc(mod3_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.687
##   Conditional ICC: 0.687
```

```
icc(mod4_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.472
##   Conditional ICC: 0.472
```

```
icc(mod5_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.220
##   Conditional ICC: 0.220
```

```
icc(mod6_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.661
##   Conditional ICC: 0.661
```

```
icc(mod7_1)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.188
##   Conditional ICC: 0.188
```

Just looking at the all artifacts, and looking at each rubric separately, Rubric CritDes, InitEDA, SelMeth and VisOrg have high ICC which means high correlation between raters and ratings on the same artifact, which means that the raters are consistent with one another in how they rate. Rubric RsrchQ, InterpRes and TxtOrg all have lower ICC which means low correlation between raters and ratings on the same artifact, which means that the raters are not consistent with one another in how they rate.

The ICC from the full dataset is consistent with the subset dataset ICC, so the raters's ratings do not vary across different artifacts.

RsrchQ

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_RsrchQ <- data.frame(r1=repeated$RsrchQ[repeated$Rater==1],r2=repeated$RsrchQ[repeate
```

```
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_RsrchQ,table(r1,r2))
```

```
##    r2
## r1  1 2 3
##   2 1 4 3
##   3 1 3 1
```

```
r1 <- factor(raters_1_and_2_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_RsrchQ$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 4 3 0
##   3 1 3 1 0
##   4 0 0 0 0
```

```
5/13
```

```
## [1] 0.3846154
```

Raters group 1 and group 2 only have 38.46% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_RsrchQ <- data.frame(r2=repeated$RsrchQ[repeated$Rater==2],r3=repeated$RsrchQ[repeate
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_RsrchQ,table(r2,r3))
```

```
##    r3
## r2  2 3
##   1 2 0
##   2 5 2
##   3 2 2
```

```
r2 <- factor(raters_2_and_3_on_RsrchQ$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_RsrchQ$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 2 0 0
##   2 0 5 2 0
##   3 0 2 2 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 2 and group 3 have 53.85% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_RsrchQ <- data.frame(r1=repeated$RsrchQ[repeated$Rater==1],r3=repeated$RsrchQ[repeate
```

```
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_RsrchQ,table(r1,r3))
```

```
##    r3
## r1  2 3
##   2 7 1
##   3 2 3
```

```
r1 <- factor(raters_1_and_3_on_RsrchQ$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_RsrchQ$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 0 7 1 0
##   3 0 2 3 0
##   4 0 0 0 0
```

```
10/13
```

```
## [1] 0.7692308
```

Raters group 1 and group 3 have 76.92% of times when they give the same ratings.

CritDes

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],r2=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_CritDes,table(r1,r2))
```

```
##    r2
## r1  1 2 3
##   1 3 2 1
##   2 2 3 1
##   3 0 0 1
```

```
r1 <- factor(raters_1_and_2_on_CritDes$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_CritDes$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 3 2 1 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 1 and group 2 have 53.85% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_CritDes <- data.frame(r2=repeated$CritDes[repeated$Rater==2],r3=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_CritDes,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   1 5 0 0
##   2 1 3 1
##   3 0 2 1
```

```
r2 <- factor(raters_2_and_3_on_CritDes$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_CritDes$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 5 0 0 0
##   2 1 3 1 0
##   3 0 2 1 0
##   4 0 0 0 0
```

```
9/13
```

```
## [1] 0.6923077
```

Raters group 2 and group 3 have 69.23% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],r3=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_CritDes,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   1 4 2 0
##   2 2 3 1
##   3 0 0 1
```

```
r1 <- factor(raters_1_and_3_on_CritDes$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_CritDes$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 4 2 0 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 1 and group 3 have 61.53% of times when they give the same ratings.

CritDes

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],r2=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_CritDes,table(r1,r2))
```

```
##    r2
## r1  1 2 3
##   1 3 2 1
##   2 2 3 1
##   3 0 0 1
```

```
r1 <- factor(raters_1_and_2_on_CritDes$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_CritDes$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 3 2 1 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 1 and group 2 have 53.85% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_CritDes <- data.frame(r2=repeated$CritDes[repeated$Rater==2],r3=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_CritDes,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   1 5 0 0
##   2 1 3 1
##   3 0 2 1
```

```
r2 <- factor(raters_2_and_3_on_CritDes$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_CritDes$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 5 0 0 0
##   2 1 3 1 0
##   3 0 2 1 0
##   4 0 0 0 0
```

```
9/13
```

```
## [1] 0.6923077
```

Raters group 2 and group 3 have 69.23% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],r3=repeated$CritDes[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_CritDes,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   1 4 2 0
##   2 2 3 1
##   3 0 0 1
```

```
r1 <- factor(raters_1_and_3_on_CritDes$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_CritDes$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 4 2 0 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 1 and group 3 have 61.54% of times when they give the same ratings.

InitEDA

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_InitEDA <- data.frame(r1=repeated$InitEDA[repeated$Rater==1],r2=repeated$InitEDA[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_InitEDA,table(r1,r2))
```

```
##    r2
## r1  2 3
##   1 1 0
##   2 4 0
##   3 3 5
```

```
r1 <- factor(raters_1_and_2_on_InitEDA$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_InitEDA$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 1 0 0
##   2 0 4 0 0
##   3 0 3 5 0
##   4 0 0 0 0
```

```
9/13
```

```
## [1] 0.6923077
```

Raters group 1 and group 2 have 69.23% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_InitEDA <- data.frame(r2=repeated$InitEDA[repeated$Rater==2],r3=repeated$InitEDA[repea
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_InitEDA,table(r2,r3))
```

```
##    r3
## r2  2 3
##   2 8 0
##   3 2 3
```

```
r2 <- factor(raters_2_and_3_on_InitEDA$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_InitEDA$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 0 0 0
##   2 0 8 0 0
##   3 0 2 3 0
##   4 0 0 0 0
```

```
11/13
```

```
## [1] 0.8461538
```

Raters group 2 and group 3 have 84.62% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_InitEDA <- data.frame(r1=repeated$InitEDA[repeated$Rater==1],r3=repeated$InitEDA[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_InitEDA,table(r1,r3))
```

```
##    r3
## r1  2 3
##   1 1 0
##   2 4 0
##   3 5 3
```

```
r1 <- factor(raters_1_and_3_on_InitEDA$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_InitEDA$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 1 0 0
##   2 0 4 0 0
##   3 0 5 3 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 1 and group 3 have 53.85% of times when they give the same ratings.

SelMeth

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_SelMeth <- data.frame(r1=repeated$SelMeth[repeated$Rater==1],r2=repeated$SelMeth[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_SelMeth,table(r1,r2))
```

```
##    r2
## r1   1  2  3
##   2  1 10  0
##   3  0  0  2
```

```
r1 <- factor(raters_1_and_2_on_SelMeth$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_SelMeth$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1   1  2  3  4
##   1  0  0  0  0
##   2  1 10  0  0
##   3  0  0  2  0
##   4  0  0  0  0
```

```
12/13
```

```
## [1] 0.9230769
```

Raters group 1 and group 2 have 92.31% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_SelMeth <- data.frame(r2=repeated$SelMeth[repeated$Rater==2],r3=repeated$SelMeth[repea
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_SelMeth,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   1 1 0 0
##   2 2 7 1
##   3 0 1 1
```

```
r2 <- factor(raters_2_and_3_on_SelMeth$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_SelMeth$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 1 0 0 0
##   2 2 7 1 0
```

```
##   3 0 1 1 0
##   4 0 0 0 0
```

```
9/13
```

```
## [1] 0.6923077
```

Raters group 2 and group 3 have 69.23% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_SelMeth <- data.frame(r1=repeated$SelMeth[repeated$Rater==1],r3=repeated$SelMeth[repea
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_SelMeth,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   2 3 7 1
##   3 0 1 1
```

```
r1 <- factor(raters_1_and_3_on_SelMeth$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_SelMeth$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 3 7 1 0
##   3 0 1 1 0
##   4 0 0 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 1 and group 3 have 61.54% of times when they give the same ratings.

InterpRes

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_InterpRes <- data.frame(r1=repeated$InterpRes[repeated$Rater==1],r2=repeated$InterpRes
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_InterpRes,table(r1,r2))
```

```
##    r2
## r1  2 3 4
##   2 3 1 1
##   3 3 5 0
```

```
r1 <- factor(raters_1_and_2_on_InterpRes$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_InterpRes$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
```

```
##   2 0 3 1 1
##   3 0 3 5 0
##   4 0 0 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 1 and group 2 have 61.54% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_InterpRes <- data.frame(r2=repeated$InterpRes[repeated$Rater==2],r3=repeated$InterpRes
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_InterpRes,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   2 1 4 1
##   3 0 2 4
##   4 0 1 0
```

```
r2 <- factor(raters_2_and_3_on_InterpRes$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_InterpRes$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 0 0 0
##   2 1 4 1 0
##   3 0 2 4 0
##   4 0 1 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 2 and group 3 have 61.54% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_InterpRes <- data.frame(r1=repeated$InterpRes[repeated$Rater==1],r3=repeated$InterpRes
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_InterpRes,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   2 1 3 1
##   3 0 4 4
```

```
r1 <- factor(raters_1_and_3_on_InterpRes$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_InterpRes$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 3 1 0
```

```
##   3 0 4 4 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 1 and group 3 have 53.85% of times when they give the same ratings.

VisOrg

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_VisOrg <- data.frame(r1=repeated$VisOrg[repeated$Rater==1],r2=repeated$VisOrg[repeate
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_VisOrg,table(r1,r2))
```

```
##    r2
## r1  1 2 3
##   1 1 0 0
##   2 0 4 5
##   3 0 1 2
```

```
r1 <- factor(raters_1_and_2_on_VisOrg$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_VisOrg$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 1 0 0 0
##   2 0 4 5 0
##   3 0 1 2 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 1 and group 2 have 53.85% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_VisOrg <- data.frame(r2=repeated$VisOrg[repeated$Rater==2],r3=repeated$VisOrg[repeate
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_VisOrg,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   1 1 0 0
##   2 0 5 0
##   3 0 3 4
```

```
r2 <- factor(raters_2_and_3_on_VisOrg$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_VisOrg$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
```

```
## r2  1 2 3 4
##   1 1 0 0 0
##   2 0 5 0 0
##   3 0 3 4 0
##   4 0 0 0 0
```

```
10/13
```

```
## [1] 0.7692308
```

Raters group 2 and group 3 have 76.92% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_VisOrg <- data.frame(r1=repeated$VisOrg[repeated$Rater==1],r3=repeated$VisOrg[repeate
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_VisOrg,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   1 1 0 0
##   2 0 7 2
##   3 0 1 2
```

```
r1 <- factor(raters_1_and_3_on_VisOrg$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_VisOrg$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 1 0 0 0
##   2 0 7 2 0
##   3 0 1 2 0
##   4 0 0 0 0
```

```
10/13
```

```
## [1] 0.7692308
```

Raters group 1 and group 3 have 76.92% of times when they give the same ratings.

TxtOrg

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_2_on_TxtOrg <- data.frame(r1=repeated$TxtOrg[repeated$Rater==1],r2=repeated$TxtOrg[repeate
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==2]
)
with(raters_1_and_2_on_VisOrg,table(r1,r2))
```

```
##    r2
## r1  1 2 3
##   1 1 0 0
##   2 0 4 5
##   3 0 1 2
```

```
r1 <- factor(raters_1_and_2_on_TxtOrg$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_TxtOrg$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
##   2 0 2 2 0
##   3 0 1 7 0
##   4 1 0 0 0
```

```
9/13
```

```
## [1] 0.6923077
```

Raters group 1 and group 2 have 69.23% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_2_and_3_on_TxtOrg <- data.frame(r2=repeated$TxtOrg[repeated$Rater==2],r3=repeated$TxtOrg[repeate
a1=repeated$Artifact[repeated$Rater==2],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_2_and_3_on_TxtOrg,table(r2,r3))
```

```
##    r3
## r2  1 2 3
##   1 0 1 0
##   2 1 0 2
##   3 0 2 7
```

```
r2 <- factor(raters_2_and_3_on_TxtOrg$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_TxtOrg$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 1 0 0
##   2 1 0 2 0
##   3 0 2 7 0
##   4 0 0 0 0
```

```
7/13
```

```
## [1] 0.5384615
```

Raters group 2 and group 3 have 53.85% of times when they give the same ratings.

```
repeated <- ratings[ratings$Repeated==1,]
raters_1_and_3_on_TxtOrg <- data.frame(r1=repeated$TxtOrg[repeated$Rater==1],r3=repeated$TxtOrg[repeate
a1=repeated$Artifact[repeated$Rater==1],
a2=repeated$Artifact[repeated$Rater==3]
)
with(raters_1_and_3_on_TxtOrg,table(r1,r3))
```

```
##    r3
## r1  1 2 3
##   2 1 1 2
##   3 0 1 7
##   4 0 1 0
```

```
r1 <- factor(raters_1_and_3_on_TxtOrg$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_TxtOrg$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 1 2 0
##   3 0 1 7 0
##   4 0 1 0 0
```

```
8/13
```

```
## [1] 0.6153846
```

Raters group 1 and group 3 have 61.53% of times when they give the same ratings.

```
fullmod1 <- lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ.full)
fullmod2 <- lmer(Rating ~ 1 + (1|Artifact), data=CritDes.full)
fullmod3 <- lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.full)
fullmod4 <- lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.full)
fullmod5 <- lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.full)
fullmod6 <- lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.full)
fullmod7 <- lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.full)
```

```
all.mod <- lmer(Rating ~ 1 + Sex + Semester + (1|Artifact), data=tall)
icc(all.mod)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.260
##   Conditional ICC: 0.257
```

3

RsrchQ

```
fullmod1.1 <- update(fullmod1, .~. + Rater)
anova(fullmod1, fullmod1.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: RsrchQ.full
## Models:
## fullmod1: Rating ~ 1 + (1 | Artifact)
## fullmod1.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## fullmod1      3 213.19 221.48 -103.6   207.19
## fullmod1.1    4 213.39 224.44 -102.7   205.39 1.8008  1     0.1796
```

Rater is not significant for Rubbric RsrchQ suggested by ANOVA test as ANOVA test perfers fullmod1.

```
fullmod1.2 <- update(fullmod1, .~. + Semester)
anova(fullmod1, fullmod1.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: RsrchQ.full
## Models:
## fullmod1: Rating ~ 1 + (1 | Artifact)
## fullmod1.2: Rating ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod1      3 213.19 221.48 -103.60   207.19
## fullmod1.2    4 214.57 225.62 -103.28   206.57 0.6253  1     0.4291
```

Semester is not significant for Rubbric RsrchQ suggested by ANOVA test as ANOVA test perfers fullmod1.

```
fullmod1.3 <- update(fullmod1, .~. + Sex)
anova(fullmod1, fullmod1.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: RsrchQ.full
## Models:
## fullmod1: Rating ~ 1 + (1 | Artifact)
## fullmod1.3: Rating ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod1      3 213.19 221.48 -103.60   207.19
## fullmod1.3    5 215.37 229.18 -102.68   205.37 1.8253  2     0.4015
```

Sex is not significant for Rubbric RsrchQ suggested by ANOVA test as ANOVA test perfers fullmod1.

```
fullmod1.4 <- update(fullmod1, .~. + Repeated)
anova(fullmod1, fullmod1.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: RsrchQ.full
## Models:
## fullmod1: Rating ~ 1 + (1 | Artifact)
## fullmod1.4: Rating ~ (1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## fullmod1      3 213.19 221.48 -103.60   207.19
## fullmod1.4    4 214.57 225.62 -103.28   206.57 0.627  1     0.4285
```

Repeated is not significant for Rubbric RsrchQ suggested by ANOVA test as ANOVA test perfers fullmod1.

```
fix_RsrchQ <- lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ.full)
```

CritDes

```
fullmod2.1 <- update(fullmod2, .~. + Rater)
anova(fullmod2, fullmod2.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: CritDes.full
## Models:
## fullmod2: Rating ~ 1 + (1 | Artifact)
## fullmod2.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod2      3 280.86 289.12 -137.43   274.86
## fullmod2.1    4 280.76 291.77 -136.38   272.76 2.0985  1     0.1474
```

Rater is not significant for Rubbric CritDes suggested by ANOVA test as ANOVA test perfers fullmod2.

```
fullmod2.2 <- update(fullmod2, .~. + Semester)
anova(fullmod2, fullmod2.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: CritDes.full
## Models:
## fullmod2: Rating ~ 1 + (1 | Artifact)
## fullmod2.2: Rating ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod2      3 280.86 289.12 -137.43   274.86
## fullmod2.2    4 282.58 293.60 -137.29   274.58 0.2751  1     0.5999
```

Semester is not significant for Rubbric CritDes suggested by ANOVA test as ANOVA test perfers fullmod2.

```
fullmod2.3 <- update(fullmod2, .~. + Sex)
anova(fullmod2, fullmod2.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: CritDes.full
## Models:
## fullmod2: Rating ~ 1 + (1 | Artifact)
## fullmod2.3: Rating ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod2      3 280.86 289.12 -137.43   274.86
## fullmod2.3    5 282.65 296.42 -136.33   272.65 2.2017  2     0.3326
```

Sex is not significant for Rubbric CritDes suggested by ANOVA test as ANOVA test perfers fullmod2.

```
fullmod2.4 <- update(fullmod2, .~. + Repeated)
anova(fullmod2, fullmod2.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: CritDes.full
## Models:
## fullmod2: Rating ~ 1 + (1 | Artifact)
## fullmod2.4: Rating ~ (1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod2      3 280.86 289.12 -137.43   274.86
## fullmod2.4    4 281.85 292.87 -136.93   273.85 1.0045  1     0.3162
```

Repeated is not significant for Rubbric CritDes suggested by ANOVA test as ANOVA test perfers fullmod2.

```
fix_CritDes <- lmer(Rating ~ 1 + (1|Artifact), data=CritDes.full)
```

InitEDA

```
fullmod3.1 <- update(fullmod3, .~. + Rater)
anova(fullmod3, fullmod3.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InitEDA.full
## Models:
## fullmod3: Rating ~ 1 + (1 | Artifact)
## fullmod3.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod3      3 243.42 251.71 -118.71   237.42
## fullmod3.1    4 243.26 254.31 -117.63   235.26 2.1635  1     0.1413
```

Rater is not significant for Rubbric InitEDA suggested by ANOVA test as ANOVA test perfers fullmod3.

```
fullmod3.2 <- update(fullmod3, .~. + Semester)
anova(fullmod3, fullmod3.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InitEDA.full
## Models:
## fullmod3: Rating ~ 1 + (1 | Artifact)
## fullmod3.2: Rating ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod3      3 243.42 251.71 -118.71   237.42
## fullmod3.2    4 245.38 256.43 -118.69   237.38 0.0391  1     0.8432
```

Semester is not significant for Rubbric InitEDA suggested by ANOVA test as ANOVA test perfers fullmod3.

```
fullmod3.3 <- update(fullmod3, .~. + Sex)
anova(fullmod3, fullmod3.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InitEDA.full
## Models:
## fullmod3: Rating ~ 1 + (1 | Artifact)
## fullmod3.3: Rating ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod3      3 243.42 251.71 -118.71   237.42
## fullmod3.3    5 246.75 260.56 -118.38   236.75 0.6718  2     0.7147
```

Sex is not significant for Rubbric InitEDA suggested by ANOVA test as ANOVA test perfers fullmod3.

```
fullmod3.4 <- update(fullmod3, .~. + Repeated)
anova(fullmod3, fullmod3.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InitEDA.full
## Models:
## fullmod3: Rating ~ 1 + (1 | Artifact)
## fullmod3.4: Rating ~ (1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod3      3 243.42 251.71 -118.71   237.42
## fullmod3.4    4 245.27 256.32 -118.63   237.27 0.1544  1     0.6944
```

Repeated is not significant for Rubbric InitEDA suggested by ANOVA test as ANOVA test perfers fullmod3.

```
fix_InitEDA <- lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.full)
```

SelMeth

```
fullmod4.1 <- update(fullmod4, .~. + Rater)
anova(fullmod4, fullmod4.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: SelMeth.full
## Models:
## fullmod4: Rating ~ 1 + (1 | Artifact)
## fullmod4.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod4      3 159.53 167.82 -76.768   153.53
## fullmod4.1    4 157.43 168.48 -74.714   149.43 4.1064  1    0.04272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rater is significant for Rubbric SelMeth suggested by ANOVA test as ANOVA test perfers fullmod4.1.

```
fullmod4.2 <- update(fullmod4.1, .~. + Semester)
anova(fullmod4.1, fullmod4.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: SelMeth.full
## Models:
## fullmod4.1: Rating ~ (1 | Artifact) + Rater
## fullmod4.2: Rating ~ (1 | Artifact) + Rater + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod4.1    4 157.43 168.48 -74.714   149.43
## fullmod4.2    5 145.86 159.67 -67.928   135.86 13.572  1  0.0002296 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Semester is significant for Rubbric SelMeth suggested by ANOVA test as ANOVA test perfers fullmod4.2.

```
fullmod4.3 <- update(fullmod4.2, .~. + Sex)
anova(fullmod4.2, fullmod4.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: SelMeth.full
## Models:
## fullmod4.2: Rating ~ (1 | Artifact) + Rater + Semester
## fullmod4.3: Rating ~ (1 | Artifact) + Rater + Semester + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod4.2    5 145.86 159.67 -67.928   135.86
## fullmod4.3    7 143.74 163.08 -64.872   129.74 6.1128  2    0.04706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sex is significant for Rubbric SelMeth suggested by ANOVA test as ANOVA test perfers fullmod4.3.

```
fullmod4.4 <- update(fullmod4.3, .~. + Repeated)
anova(fullmod4.3, fullmod4.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: SelMeth.full
## Models:
## fullmod4.3: Rating ~ (1 | Artifact) + Rater + Semester + Sex
## fullmod4.4: Rating ~ (1 | Artifact) + Rater + Semester + Sex + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod4.3    7 143.74 163.08 -64.872   129.74
## fullmod4.4    8 145.57 167.67 -64.785   129.57 0.1727  1     0.6777
```

Repeated is not significant for Rubbric SelMeth suggested by ANOVA test as ANOVA test perfers fullmod4.3.

```
fix_SelMeth <- lmer(Rating ~ 1 + Rater + Sex + Semester + (1|Artifact), data=SelMeth.full)
```

InterpRes

```
fullmod5.1 <- update(fullmod5, .~. + Rater)
anova(fullmod5, fullmod5.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InterpRes.full
## Models:
## fullmod5: Rating ~ 1 + (1 | Artifact)
## fullmod5.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## fullmod5      3 220.09 228.38 -107.048   214.09
## fullmod5.1    4 203.79 214.84  -97.897   195.79 18.302  1  1.885e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rater is significant for Rubbric InterpRes suggested by ANOVA test as ANOVA test perfers fullmod5.1.

```
fullmod5.2 <- update(fullmod5.1, .~. + Semester)
anova(fullmod5.1, fullmod5.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InterpRes.full
## Models:
## fullmod5.1: Rating ~ (1 | Artifact) + Rater
## fullmod5.2: Rating ~ (1 | Artifact) + Rater + Semester
##            npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## fullmod5.1    4 203.79 214.84 -97.897   195.79
## fullmod5.2    5 205.26 219.07 -97.630   195.26 0.533  1     0.4653
```

Semester is not significant for Rubbric InterpRes suggested by ANOVA test as ANOVA test perfers fullmod5.1.

```
fullmod5.3 <- update(fullmod5.1, .~. + Sex)
anova(fullmod5.1, fullmod5.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InterpRes.full
## Models:
## fullmod5.1: Rating ~ (1 | Artifact) + Rater
## fullmod5.3: Rating ~ (1 | Artifact) + Rater + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod5.1    4 203.79 214.84 -97.897   195.79
## fullmod5.3    6 205.68 222.26 -96.842   193.68 2.1086  2     0.3484
```

Sex is not significant for Rubbric InterpRes suggested by ANOVA test as ANOVA test perfers fullmod5.1.

```
fullmod5.4 <- update(fullmod5.1, .~. + Repeated)
anova(fullmod5.1, fullmod5.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: InterpRes.full
## Models:
## fullmod5.1: Rating ~ (1 | Artifact) + Rater
## fullmod5.4: Rating ~ (1 | Artifact) + Rater + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod5.1    4 203.79 214.84 -97.897   195.79
## fullmod5.4    5 205.70 219.51 -97.848   195.70 0.0967  1     0.7559
```

Repeated is not significant for Rubric InterpRes suggested by ANOVA test as ANOVA test perfers fullmod5.1.

```
fix_InterpRes <- lmer(Rating ~ 1 + Rater + (1|Artifact), data=InterpRes.full)
```

VisOrg

```
fullmod6.1 <- update(fullmod6, .~. + Rater)
anova(fullmod6, fullmod6.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: VisOrg.full
## Models:
## fullmod6: Rating ~ 1 + (1 | Artifact)
## fullmod6.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod6      3 228.95 237.21 -111.47   222.95
## fullmod6.1    4 230.40 241.42 -111.20   222.40 0.5461  1     0.4599
```

Rater is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod6.

```
fullmod6.2 <- update(fullmod6, .~. + Semester)
anova(fullmod6, fullmod6.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: VisOrg.full
## Models:
## fullmod6: Rating ~ 1 + (1 | Artifact)
## fullmod6.2: Rating ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod6      3 228.95 237.21 -111.47   222.95
## fullmod6.2    4 229.33 240.34 -110.67   221.33 1.6196  1     0.2031
```

Semester is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod6.

```
fullmod6.3 <- update(fullmod6, .~. + Sex)
anova(fullmod6, fullmod6.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: VisOrg.full
## Models:
## fullmod6: Rating ~ 1 + (1 | Artifact)
## fullmod6.3: Rating ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod6      3 228.95 237.21 -111.47   222.95
## fullmod6.3    5 231.47 245.23 -110.73   221.47 1.4831  2     0.4764
```

Sex is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod6.

```
fullmod6.4 <- update(fullmod6, .~. + Repeated)
anova(fullmod6, fullmod6.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: VisOrg.full
## Models:
## fullmod6: Rating ~ 1 + (1 | Artifact)
## fullmod6.4: Rating ~ (1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod6      3 228.95 237.21 -111.47   222.95
## fullmod6.4    4 229.76 240.77 -110.88   221.76 1.1894  1     0.2754
```

Repeated is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod6.

```
fixed_VisOrg <- lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.full)
```

TxtOrg

```
fullmod7.1 <- update(fullmod7, .~. + Rater)
anova(fullmod7, fullmod7.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: TxtOrg.full
## Models:
## fullmod7: Rating ~ 1 + (1 | Artifact)
## fullmod7.1: Rating ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod7      3 251.45 259.74 -122.73   245.45
## fullmod7.1    4 248.88 259.93 -120.44   240.88 4.5725  1    0.03249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rater is significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod7.1.

```
fullmod7.2 <- update(fullmod7.1, .~. + Semester)
anova(fullmod7.1, fullmod7.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: TxtOrg.full
## Models:
## fullmod7.1: Rating ~ (1 | Artifact) + Rater
## fullmod7.2: Rating ~ (1 | Artifact) + Rater + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod7.1    4 248.88 259.93 -120.44   240.88
## fullmod7.2    5 249.15 262.95 -119.57   239.15 1.7366  1     0.1876
```

Semester is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod7.1.

```
fullmod7.3 <- update(fullmod7.1, .~. + Sex)
anova(fullmod7.1, fullmod7.3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: TxtOrg.full
## Models:
## fullmod7.1: Rating ~ (1 | Artifact) + Rater
## fullmod7.3: Rating ~ (1 | Artifact) + Rater + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod7.1    4 248.88 259.93 -120.44   240.88
## fullmod7.3    6 252.11 268.68 -120.05   240.11 0.7737  2     0.6792
```

Sex is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod7.1.

```
fullmod7.4 <- update(fullmod7.1, .~. + Repeated)
anova(fullmod7.1, fullmod7.4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: TxtOrg.full
## Models:
## fullmod7.1: Rating ~ (1 | Artifact) + Rater
## fullmod7.4: Rating ~ (1 | Artifact) + Rater + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fullmod7.1    4 248.88 259.93 -120.44   240.88
## fullmod7.4    5 250.39 264.20 -120.19   240.39 0.4922  1     0.4829
```

Repeated is not significant for Rubric VisOrg suggested by ANOVA test as ANOVA test perfers fullmod7.1.

```
fixed_TxtOrg <- lmer(Rating ~ 1 + Rater + (1|Artifact), data=TxtOrg.full)
```

```
library(RLRsim)
# model1 <- lmer(Rating ~ 1 + Rate + Rubric + Rate * Rubric + (0+Rubric|Artifact), data = tall)
model0 <- lmer(Rating ~ 1 + (0+Rubric|Artifact), data = tall)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00260717 (tol = 0.002, component 1)
```

```
model1  <- update(model0, .~. + Rater)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

```
anova(model0, model1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model0: Rating ~ 1 + (0 + Rubric | Artifact)
## model1: Rating ~ (0 + Rubric | Artifact) + Rater
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model0   30 1537.2 1678.3 -738.58   1477.2
## model1   31 1530.9 1676.8 -734.45   1468.9 8.2508  1   0.004073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable Rater is significant as AIC/BIC prefer model1.

```
model2 <- update(model1, .~.+Sex)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00205461 (tol = 0.002, component 1)
```

```
anova(model1, model2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model1: Rating ~ (0 + Rubric | Artifact) + Rater
## model2: Rating ~ (0 + Rubric | Artifact) + Rater + Sex
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   31 1530.9 1676.8 -734.45   1468.9
## model2   33 1529.7 1685.0 -731.83   1463.7 5.2399  2     0.0728 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable Sex is not significant as AIC/BIC perfer model1

```
model3 <- update(model1, .~. +Semester)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00487549 (tol = 0.002, component 1)
```

```
anova(model1, model3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model1: Rating ~ (0 + Rubric | Artifact) + Rater
## model3: Rating ~ (0 + Rubric | Artifact) + Rater + Semester
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   31 1530.9 1676.8 -734.45   1468.9
## model3   32 1528.6 1679.1 -732.28   1464.6 4.3453  1    0.03711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable Semester is significant as AIC/BIC prefer model3

```
model4 <- update(model3, .~. + Repeated)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model3, model4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model3: Rating ~ (0 + Rubric | Artifact) + Rater + Semester
## model4: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Repeated
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3   32 1528.6 1679.1 -732.28   1464.6
## model4   33 1529.2 1684.5 -731.62   1463.2 1.3198  1     0.2506
```

Variable Repeated is not significant as AIC/BIC prefer model3

```
model5 <- update(model3, .~. + Rater:Semester)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0963624 (tol = 0.002, component 1)
```

```
anova(model3, model5)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model3: Rating ~ (0 + Rubric | Artifact) + Rater + Semester
## model5: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rater:Semester
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3   32 1528.6 1679.1 -732.28   1464.6
## model5   33 1529.5 1684.8 -731.73   1463.5 1.0939  1     0.2956
```

Interaction term between Rater and Semester is not significant as AIC/BIC prefer model3.

```
model6 <- update(model3, .~. +Rubric)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model3, model6)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model3: Rating ~ (0 + Rubric | Artifact) + Rater + Semester
## model6: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3   32 1528.6 1679.1 -732.28   1464.6
## model6   38 1476.2 1655.0 -700.09   1400.2 64.374  6  5.788e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable Rubric is significant as AIC/BIC prefers model6.

```
model7 <- update(model6, .~. + Rater:Rubric)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model6, model7)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model6: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric
## model7: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model6   38 1476.2 1655.0 -700.09   1400.2
## model7   44 1467.4 1674.5 -689.71   1379.4 20.765  6   0.002022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction term between Rater and Rubric is significant as AIC/BIC prefer model7.

```
model8 <- update(model7, .~. + Semester:Rubric)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model7, model8)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall
## Models:
## model7: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
## model8: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric + Semester:Rubri
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model7   44 1467.4 1674.5 -689.71   1379.4
## model8   50 1467.7 1703.0 -683.87   1367.7 11.686  6    0.06935 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction term between Semester and Rubric is not significant as AIC/BIC prefer model7.

```
fixed_all <- lmer(Rating ~ 1 + Rater + Semester + Rubric + Rater:Rubric + (0+Rubric|Artifact), data = ta
```

```
## boundary (singular) fit: see ?isSingular
```

```
library(LMERConvenienceFunctions)
test_model <- lmer(Rating ~ 1 + Rater + Semester + Rubric + Sex + Repeated + (0+Rubric|Artifact), data =
test_model1 <- fitLMER.fnc(test_model, ran.effects=c("(Rater|Artifact)", "(Semester|Artifact)",  "(Sex|A
```

```
## ===========================================================
## ===              backfitting fixed effects            ===
## ===========================================================
## setting REML to FALSE
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Repeated" = 0.4811 >= 0.05
##     not part of higher-order interaction
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00597926 (tol = 0.002, component 1)
```

```
##     BIC simple = 1664; BIC complex = 1670; decrease = -6 < 5
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1173 >= 0.05
##     not part of higher-order interaction
```

```
## boundary (singular) fit: see ?isSingular
```

```
##     BIC simple = 1655; BIC complex = 1664; decrease = -9 < 5
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ======================================================
## ===            forwardfitting random effects       ===
## ======================================================
## evaluating addition of (Rater|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.0005778625
##  adding (Rater|Artifact) to model
## evaluating addition of (Semester|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9229156
##  not adding (Semester|Artifact) to model
## evaluating addition of (Sex|Artifact) to model

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 2 negative eigenvalues

##  log-likelihood ratio test p-value = 0.4954787
##  not adding (Sex|Artifact) to model
## evaluating addition of (Repeated|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9016879
##  not adding (Repeated|Artifact) to model
## ======================================================
## ===            re-backfitting fixed effects        ===
## ======================================================
## setting REML to FALSE

## boundary (singular) fit: see ?isSingular

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced
```

```
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.0696 >= 0.05
##     not part of higher-order interaction
```

```
## boundary (singular) fit: see ?isSingular
```

```
##     BIC simple = 1654; BIC complex = 1658; decrease = -3 < 5
##     removing term
```

```
## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced
```

```
## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced
```

```
## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced
```

```
## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced
```

```
## resetting REML to TRUE
```

```
## boundary (singular) fit: see ?isSingular
```

```
## pruning random effects structure ...
##   nothing to prune
## log file is mylogfile.txt
```

```
finalmod <- lmer(Rating ~ 1 + Rater + Semester + Rubric + Rater:Rubric + (0+ Rater + Rubric|Artifact), c
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00998686 (tol = 0.002, component 1)
```

```
summary(finalmod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + Rater + Semester + Rubric + Rater:Rubric + (0 +
##     Rater + Rubric | Artifact)
##    Data: tall
##
## REML criterion at convergence: 1413.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9793 -0.4778 -0.0650  0.4944  3.3656
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact Rater           0.02836  0.1684
##           RubricCritDes   0.60143  0.7755   -0.35
##           RubricInitEDA   0.28371  0.5326   -0.12  0.46
##           RubricInterpRes 0.11400  0.3376   -0.38  0.43  0.72
##           RubricRsrchQ    0.28146  0.5305   -0.65  0.66  0.40  0.81
##           RubricSelMeth   0.02096  0.1448   -0.56  0.72  0.33  0.62  0.73
##           RubricTxtOrg    0.15801  0.3975   -0.09  0.39  0.43  0.38  0.51  0.00
##           RubricVisOrg    0.23469  0.4844   -0.24  0.41  0.62  0.59  0.53  0.06
```

```
##  Residual                 0.16984  0.4121
##
##
##
##
##
##
##
##
##   0.59
##
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)             1.75913    0.16482  10.673
## Rater                   0.09638    0.06997   1.377
## SemesterS19            -0.12025    0.07575  -1.587
## RubricInitEDA           0.83624    0.19165   4.363
## RubricInterpRes         1.31227    0.18956   6.923
## RubricRsrchQ            0.81056    0.17768   4.562
## RubricSelMeth           0.52326    0.18577   2.817
## RubricTxtOrg            1.15502    0.19321   5.978
## RubricVisOrg            0.84461    0.19444   4.344
## Rater:RubricInitEDA    -0.14813    0.08357  -1.772
## Rater:RubricInterpRes  -0.36493    0.08248  -4.424
## Rater:RubricRsrchQ     -0.17827    0.07860  -2.268
## Rater:RubricSelMeth    -0.18396    0.08040  -2.288
## Rater:RubricTxtOrg     -0.23606    0.08399  -2.810
## Rater:RubricVisOrg     -0.15948    0.08443  -1.889
##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00998686 (tol = 0.002, component 1)
```

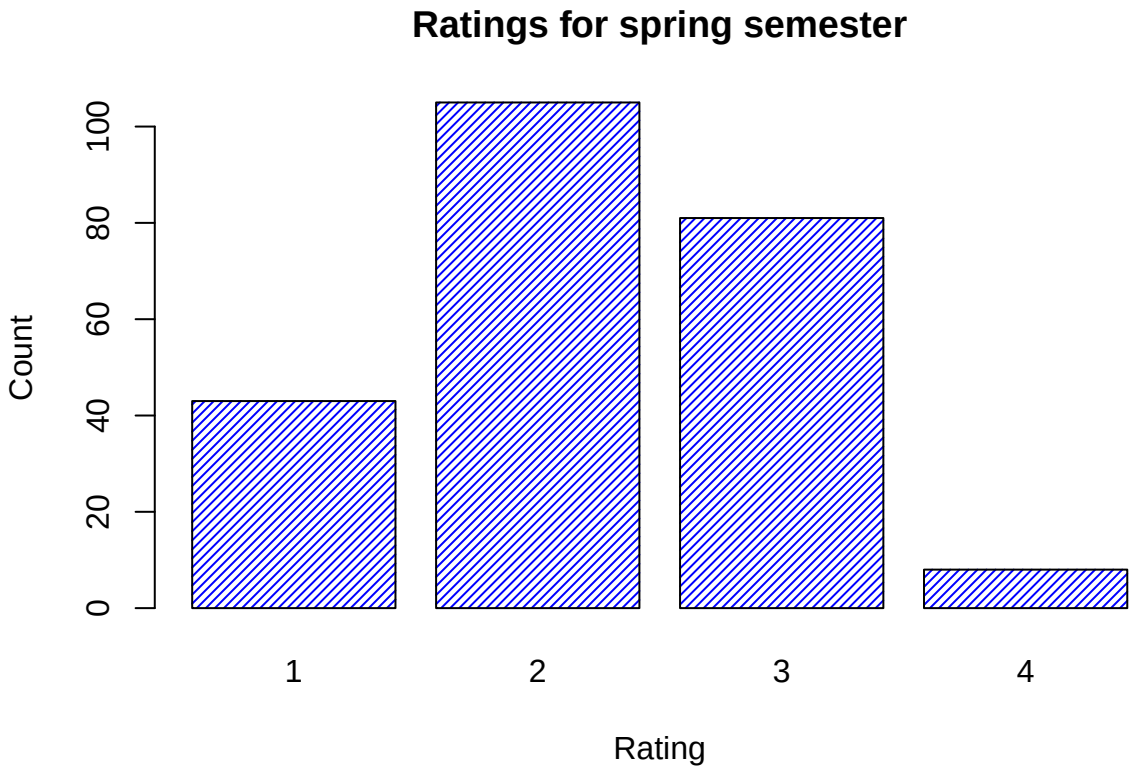
For fixed effect, the ANOVA test agrees with the result in fitLMER.fnc(), which both show that variables Rater, Semester, Rubric are significant predictors. The ANOVA test also picked the interaction term between Rubric and Rater as a fixed effect. For random effect, Rater and Rubric are added based on the result in fitLMER.fnc(). All in all, variables Rater and Rubric are both significant in fixed and random effects, while Semester is significant as fixed effect, the interaction term between Rater and Rubric is also significant.

4

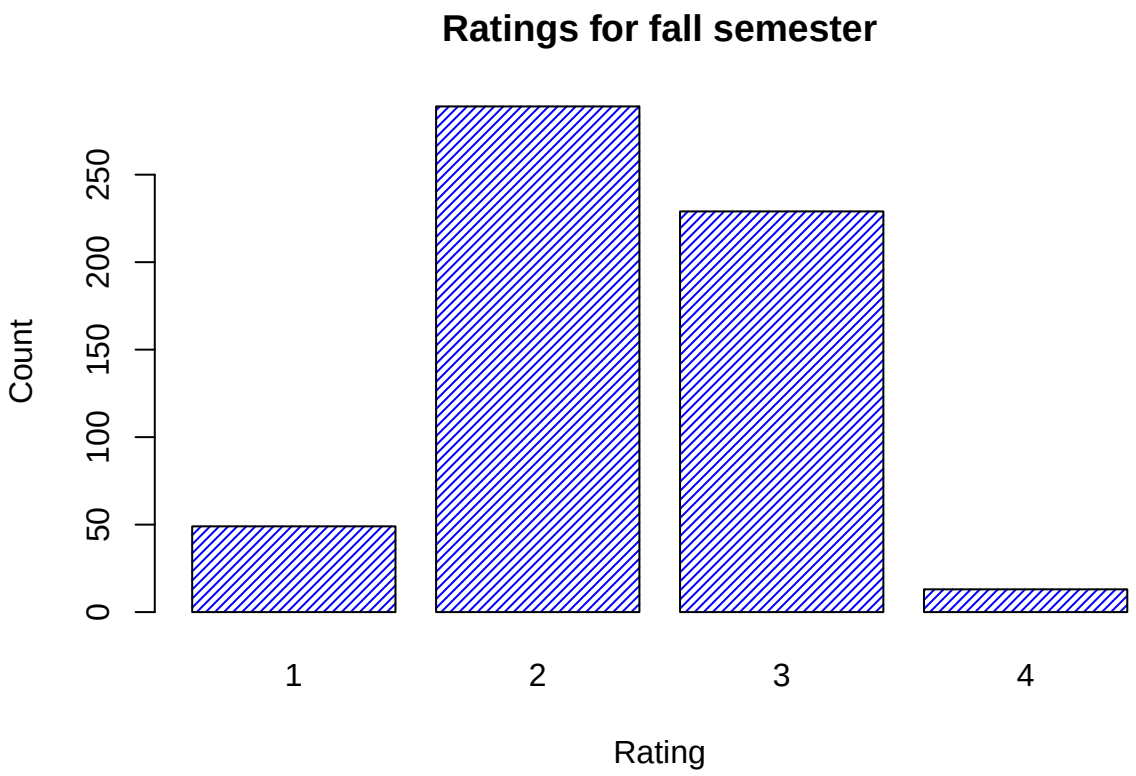
It is interesting to see that significant factors in different rubric specific dataset are different, which means that raters do grade each rubric differently. One other interesting finding is that semester is a significant factor not only in some rubric specific datasets but also in the tall dataset. It will be interesting to see why different raters tend to give different ratings to different semesters. The EDA shows that the ratings distribution between these two semesters are not that different, but it is a significant variable for the model as long as the interaction term between semester and rubric. One reason is that the dataset is not evenly split. The data in fall semester is much more

than that in the spring semester, which may cause the difference between ratings in these two semesters.

```
spring <- tall[tall$Semester=="S19",]
fall <- tall[tall$Semester=="F19",]
barplot(table(spring$Rating), main = "Ratings for spring semester", xlab = "Rating", ylab = "Count", co
```




```
barplot(table(fall$Rating), main = "Ratings for fall semester", xlab = "Rating", ylab = "Count", col="bl
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.