The Success and Fairness of CMU Dietrich College's Newly Implemented "General Education" Program

Emily Zeng | emilyzen@andrew.cmu.edu

1. ABSTRACT

Carnegie Mellon University's Dietrich College is interested in determining whether their newly implemented "General Education" program for undergraduates is successful, specifically by predicting scores via various factors associated with a student's project. The data consists of rubric items, demographic information, and the score that raters gave each student for 91 project papers for a Freshman Statistics course. To answer the research questions presented, we use exploratory data analysis methods, model building, and model selection methods. We determine that ratings for rubric items and for each rater differs are not indistinguishable from another, and that Rater and Rubric are important factors related to Rating. Overall, there is potential success in the new "General Education" program, but there still needs to a focus on ensuring that grades are fair for all the students. Future work could be done analyze the success of the "General Education" program through a different course, and further investigation could be done to determine how Sex and Semester affect Rating.

2. INTRODUCTION

Dietrich College of Humanities and Social Sciences at Carnegie Mellon University is interested in creating a new "General Education" program for undergraduates, in which students are required to take a certain set of courses. In order to determine whether this new program is considered successful, Dietrich College wants to rate the student work in some of these courses offered in the program. Specifically, an experiment was done to rate student work in the Freshman Statistics course. If this experiment demonstrates that the "General Education" program is successful, it would be a valuable experience for all incoming Carnegie Mellon students to have, as having a well-rounded, interdisciplinary education is crucial for scholarly growth. Below, we list the main guiding research questions that are the basis to our study and analysis.

The 4 main research questions of this study are as follows:

1. Is the distribution of ratings for each rubric pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

2. For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

3. More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

4. Is there anything else interesting to say about this data?

3. DATA

The data used in this study come from the ratings for seven rubric items for the sample of 91 project papers for the Freshman Statistics course (Junker 2021). Three different raters rated the 91 papers, or "artifacts", without knowing what class or which student the artifacts were from. 13 of the artifacts were rated by all three raters, while the remaining 78 were rated only by one rater each. We were provided two different datasets, with identical data just formatted in different ways: `ratings.csv` has data with the variables and their definitions shown in *Table 1* (page 2). In terms of analysis and modeling, we do not expect `X`, `Sample`, and `Overlap` to be useful variables, so we have indicated this in *Table 1* (page 2) with an asterisk. The other dataset, `tall.csv`, has a row for each rating, shown in the column Rating and the rubric for that rating in the column Rubric. *Table 2* (page 2) shows the seven rubric items that the three raters rated the artifacts on, while *Table 3* (page 2) shows the rating scale for the rubric items.

Numeric summaries for each rubric are shown in *Table 4* (page 3). Additionally, numerical summaries for each rater are shown in *Tables 5 – 7* (pages 3-4).

Variable Name	Description
X*	Row number in the dataset
Rater	Which of the 3 raters gave a rating
Sample*	Sample number
Overlap*	Unique identifier for artifact seen by all 3 raters
Semester	Spring 19 or Fall 19 – which semester the artifact came from
Sex	Sex of student who created artifact
RsrchQ	Rating on research question
CritDes	Rating on critique design
InitEDA	Rating on initial EDA
SelMeth	Rating on selection method(s)
InterpRes	Rating on interpret results
VisOrg	Rating on visual organization
TxtOrg	Rating on text organization
Artifact	Unique identifier for each artifact
Repeated	Zero (0) or one (1), where 1 means artifact was rated by all 3 raters

Table 1: Variables and their definitions in `ratings.csv`*. Variables not expected to be useful for analysis have an asterisk next to them.*

Short Name	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 2: Numerical summary for each rubric for Rater 3.

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 3: Rating scale for each rubric item.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.35	0.59	1	4	3
CritDes	1.87	0.84	1	4	3
InitEDA	2.44	0.70	1	4	3
SelMeth	2.07	0.49	1	3	2
InterpRes	2.49	0.61	1	4	3
VisOrg	2.41	0.67	1	4	3
TxtOrg	2.60	0.70	1	4	3

Table 4: Numerical summary for each rubric.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.44	0.64	1	4	3
CritDes	1.59	0.72	1	3	2
InitEDA	2.41	0.72	1	4	3
SelMeth	2.13	0.34	2	3	1
InterpRes	2.72	0.46	2	3	1
VisOrg	2.39	0.64	1	4	3
TxtOrg	2.77	0.58	1	4	3

Table 5: Numerical summary for each rubric for Rater 1.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.36	0.63	1	3	2
CritDes	2.13	0.91	1	4	3
InitEDA	2.56	0.68	1	4	3
SelMeth	2.13	0.47	1	3	2
InterpRes	2.59	0.59	1	4	3
VisOrg	2.64	0.67	1	4	3
TxtOrg	2.59	0.72	1	4	3

Table 6: Numerical summary for each rubric for Rater 2.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.44	0.64	1	4	3
CritDes	1.59	0.72	1	3	2
InitEDA	2.41	0.72	1	4	3
SelMeth	2.13	0.34	2	3	1
InterpRes	2.72	0.46	2	3	1
VisOrg	2.39	0.64	1	4	3
TxtOrg	2.77	0.58	1	4	3

Table 7: Rubric items for Freshman Statistics projects

4. METHODS

Below is a reminder of the four research questions we are aiming to answer:

1. Is the distribution of ratings for each rubric pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

2. For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

3. More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

4. Is there anything else interesting to say about this data?

For the first research question, we look at numerical summaries and bar plots to determine the distribution of ratings for each rubric and to determine the distribution of ratings given by each rater. Additionally, we look at the distribution of ratings mentioned above for the 13 artifacts that were rated by all three raters compared to the distribution of ratings mentioned above for the remaining 91 artifacts that were only rated by one rater.

For the second research question, we initially focus only on the 13 artifacts that were rated by all three raters to determine whether the raters agree on their scores. We quantify the level of agreement between the raters by comparing intraclass correlations (ICC), which is calculated from seven random-intercept models (one for each rubric). Additionally, to identify exactly which rater is contributing to disagreement, we create contingency tables for the ratings between each pair of raters for each rubric: in total, we create 21 contingency tables to show the counts of ratings given by each pair of raters. Then, we calculate the exact percentage of agreement for each pair of raters for each rubric item. Lastly, we repeat the process of calculating the ICCs with the full dataset `tall.csv` and compare these ICCs with the ICCs from the 13 artifacts subset.

For the third research question, our goal is to fit a linear mixed effects model. Here, we use the `tall.csv` dataset to create our initial model that only includes `Rubric` as a random effect. Then, we add in fixed effects for all the variables, which includes `Rater`, `Semester`, `Sex`, `Repeated`, and `Rubric`. After adding in the fixed effects that are important to our model, we add in random effects from the same five variables. Lastly, we explore interactions between the five variables and add the meaningful interactions to the model. To determine whether the model with interactions performs better than the model without interactions, we perform model selection using an ANOVA test. The final mixed effects model is created using automatic backward selection on fixed effects, forward selection in random effects, and then backward selection again on fixed effects.

For the fourth research question, we do further exploratory data analysis to see what insights may need further investigation by looking at numerical summaries and bar plots for ratings by `Sex` and `Semester`. Additionally, we look at the entire dataset to see if there are missing values, and then determine what is the best way to go about filling in those missing data.

5. RESULTS

Our first research question asks whether the ratings distributions for the rubrics are indistinguishable from another, as well as whether the ratings given by each rater is indistinguishable from one another. Firstly, to determine whether there is a difference between each rubric's ratings, we look at numerical summaries, histograms, and bar plots for each rubric's ratings (pages 14-16 in Technical Appendix). Looking at the distributions of the scores for each of the seven rubrics in *Figure 1* (page 8), it seems like `CritDes` was rated the lowest (right skewed and has the greatest count of 1s compared to the rest of the rubrics), while `RsrchQ`, `InitEDA`, and `VisOrg` scored lower (all are right skewed). `SelMeth` seemed to be scored very fairly (nearly uniform distribution). Both `InterpRes` and `TxtOrg` scored the highest compared to the other rubrics, since they are both left skewed. However, `TxtOrg` scored best, with the highest mean of 2.598, as shown in the numerical summaries for each rubric. Overall, the distribution of ratings for each rubric does not seem to be indistinguishable from one another.

Lastly, we see the distributions of ratings by each rater in *Figure 2* (page 8). Upon initial investigation, it seems like rater 3 on the far right in *Figure 2* tends to give lower scores than raters 1 and 2. Raters 1 and 2 have very similar rating distributions, indicating that their ratings agree more with one another.

Secondly, to determine whether there is a difference between each rater's ratings we look at numerical summaries and bar plots for each rater's ratings (pages 18-24 in Technical Appendix). When we look at the distributions of each rater's ratings for each rubric (*Figures 3*, *4*, and *5* on pages 9-10), it looks like rater 3 is harsher than the other 2 raters. Most of the distributions rater 3's ratings for each rubric are somewhat right skewed. Rater 1 is the only rater that sometimes gives binary ratings, meaning only rating 2 values, as opposed to 3 or 4 ratings. These findings above are confirmed by the bar plot of each rater's ratings, again, in *Figure 2* (page 8). Rater 3 has a right skewed distribution of ratings, meaning they tend to give lower scores of 1's and 2's, as opposed to more 3's and 4's. Rater 1 and Rater 2, on the other hand, have similar distributions of ratings. Based on the above findings, it does not seem like the rater's ratings are indistinguishable from one another: rater 3 is a harsher grader overall and tends to give lower ratings.

Our second research question asks whether the raters agree on their scores, and if not, which rater disagrees with the others. As mentioned in the Methods section, we determine that ICC is a good measure of interrater agreement. In *Table 8* (page 10), we see the ICC values for each rubric for the 13 artifacts seen by all three raters (page 28 in Technical Appendix). The ICC values for `CritDes`, `InitEDA`, `SelMeth`, and `VisOrg` are the highest amongst the seven rubrics, meaning that the three raters agreed the most on these four amentioned rubrics. On the other hand, the lower the ICC value, the less the raters agreed

on rubric items. It looks like they disagreed the most on TxtOrg. Looking at ICC values only gives a broad view on whether the raters are in general agreement or disagreement, but they do not provide information on which rater is contributing to disagreement.

To combat this issue of lack of specificity in which rater is contributing to disagreement, we look at contingency tables between pairs of raters to determine the percentage of agreement for each rubric (pages 29-40 in Technical Appendix). In *Table 8* (page 10), we see the agreement rates between each pair of raters for each rubric. Below are the agreement rates and results for each rubric item.

- For RsrchQ, raters 1 and 3 agree 77% of the time. However, rater 2 is the one that disagrees more, especially when compared to rater 1.
- For CritDes, rater 2 seems to disagree more.
- For InitEDA, this time rater 3 is the one that disagrees more. Surprisingly, raters 1 and 2 have a relatively high agreement rate for InitEDA.
- For SelMeth, the agreement rates are relatively high between all 3 raters.
- For InterpRes, relatively the same agreement rates across all 3 raters.
- For TxtOrg, relatively the same agreement rate across all 3 raters.

Table 8 (page 10) also shows the ICC values for the full dataset, and we see that CritDes, InitEDA, VisOrg, and TxtOrg have the highest ICCs. This means the raters agree the most for these four rubrics. When comparing to the subset of 13 artifacts, the ICCs are not the same, especially for TxtOrg – its ICC value is much higher for the full dataset. Otherwise, the ICCs are relatively similar.

Our third research question asks which factors out of the five variables (Rater, Semester, Sex, Repeated, and Rubric) are related to Rating, and if there are any interactions between the variables that can predict Rating. Our final model (*Model 1.1*) to predict Rating including fixed effects of Rater and Rubric, random effects of Rater and Rubric, and an interaction term between Rubric and Rater is as follows:

$$Rating \sim Rater + Rubric + Rater*Rubric + (0 + Rater + Rubric \mid Artifact) \qquad (1.1).$$

Table 9 (page 11) shows the estimated coefficients for Model 1.1. An interpretation of the final model (*Model 1.1*) is as follows:

- Rater is a fixed effect, meaning that each rater has the same variance in ratings as the other raters.
- Rubric is a fixed effect, meaning that the variance in ratings for each rubric is relatively constant.
- Rater is also a random effect, meaning that we can estimate both the mean and variance of each rater and make predictions about raters that were not included in this study. We can make broad inferences about raters that are not dependent on

other factors such as being hired by Carnegie Mellon. Since `Rater` is both a fixed and random effect, this means that each rater's ratings differ from artifact to artifact, based on the random effect that depends on `Artifact`.

- `Rubric` is also a random effect, meaning that we can make broad generalizations about rubrics not just in the context of these 91 students' artifacts, or a certain professor's rubric scale. Because `Rubric` is both a fixed and random effect, this indicates that there are different average scores for each rubric, but there is also variation in rubric averages from one artifact to another, based on the random effect that depends on `Artifact`.
- There is interaction between `Rater` and `Rubric`, indicating that each rater tends to rate each rubric differently. This means that the rating for each rubric differs from rater to rater.

Our fourth question asks whether there are any other interesting insights that should be mentioned to the Dean. After conducting more EDA on `Sex` and `Semester`, we determine that there does seem to be subtle differences in ratings depending on `Sex` and Semester.

Looking at the bar plots by `Sex` in *Figure 6* (page 11), both female and male ratings have similar distribution shapes. Both are right skewed, meaning that typically, everyone is being rated on the lower end. When looking at the means for each rubric by `Sex` in *Table 10* (page 12), we can see that for five out of seven rubrics, on average, males scored higher. The two rubrics that females scored higher than males did on average are `VisOrg` and `TxtOrg`.

Looking the bar plots by `Semester` in *Figure 7* (page 12), both fall and spring semester ratings are right skewed. Again, this means that most students are getting low ratings in the 2s and 3s. Additionally, we can see in *Table 11* (page 12) that the Spring 19 mean for ratings is the lowest, when compared to the Fall 19 and Overall means.

Lastly, we did see that there were two instances of missing data: one for `Semester` and one for `Sex` (page 14 in Technical Appendix). Typically, in the case of missing data, you can either remove the entire row entirely or replace the missing value with a summary statistic (e.g. mean, mode, or median) for that column. In our case, we decided to replace the missing values with the mode of `Sex` and `Semester`, which were Female and Fall respectively.

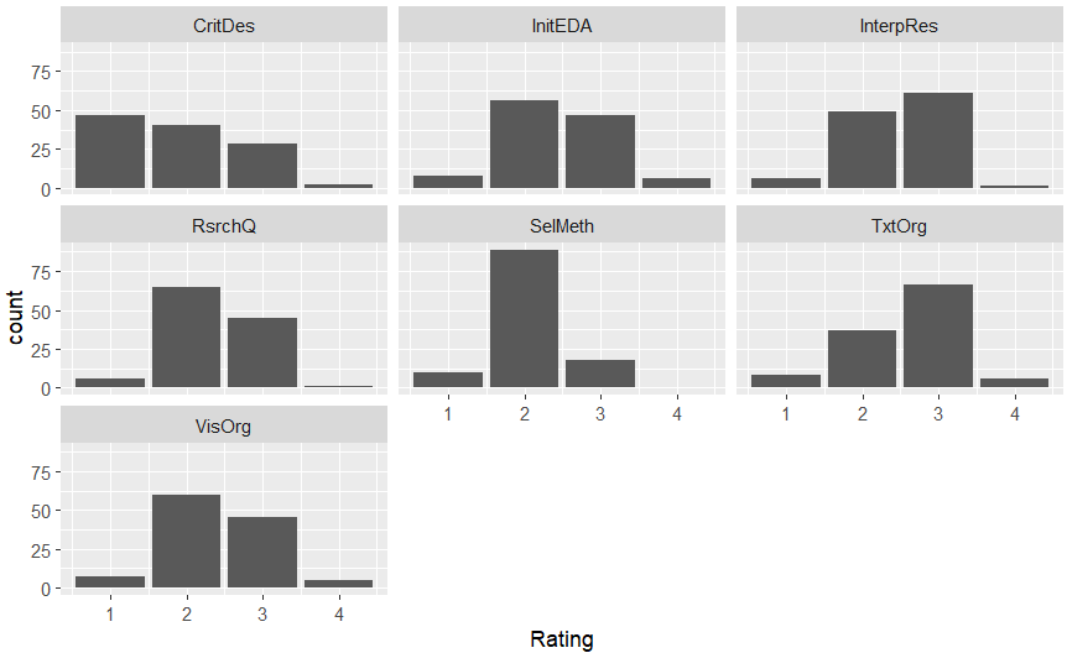


Figure 1: Bar plots for each rubric's ratings.

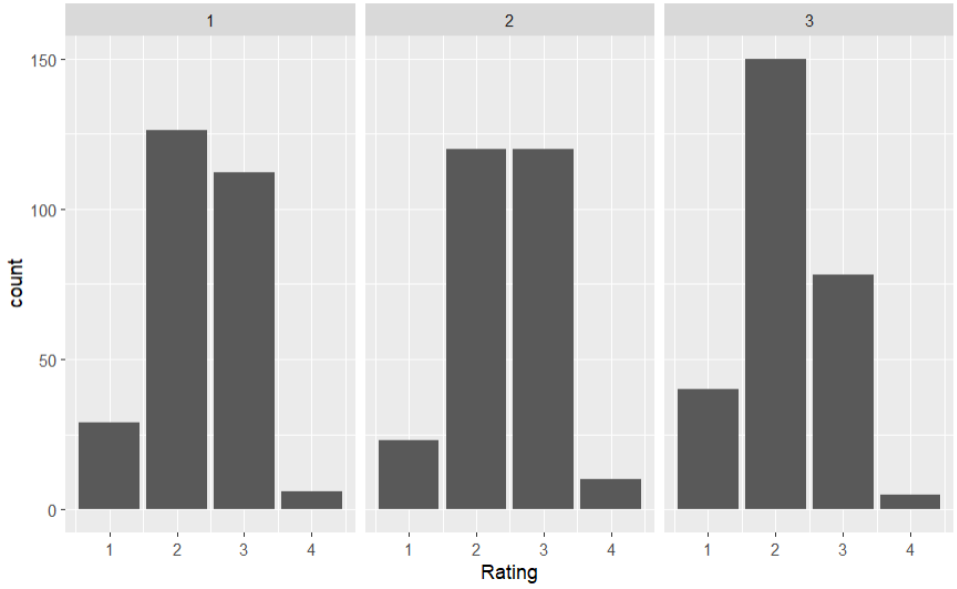


Figure 2: Bar plots of each rater's ratings.

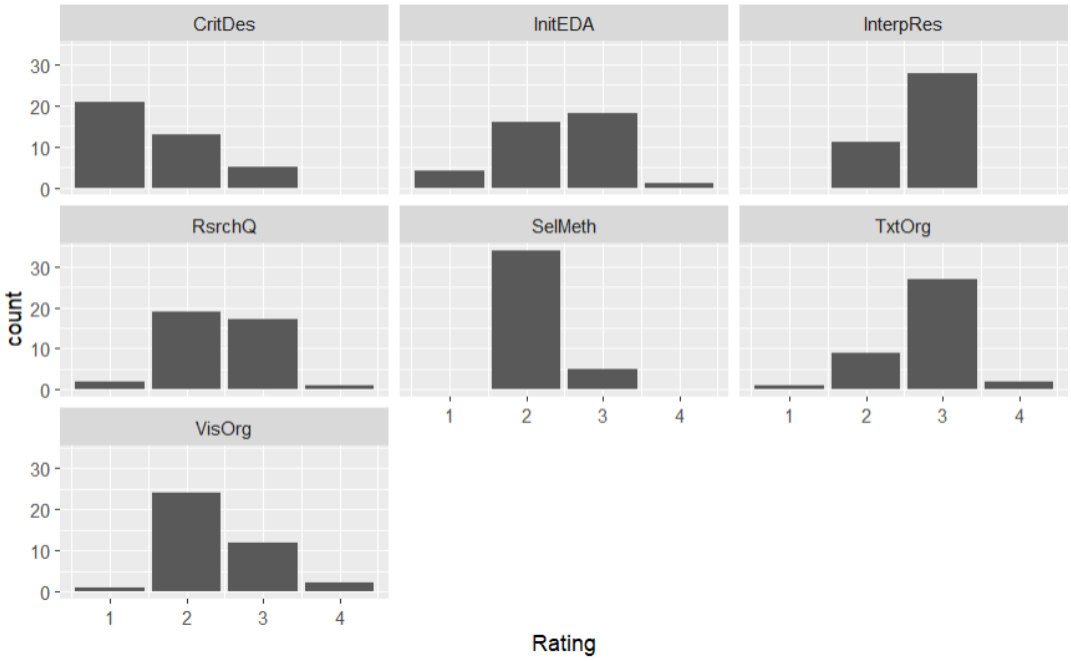


Figure 3: Bar plots of rater 1's ratings for each rubric.

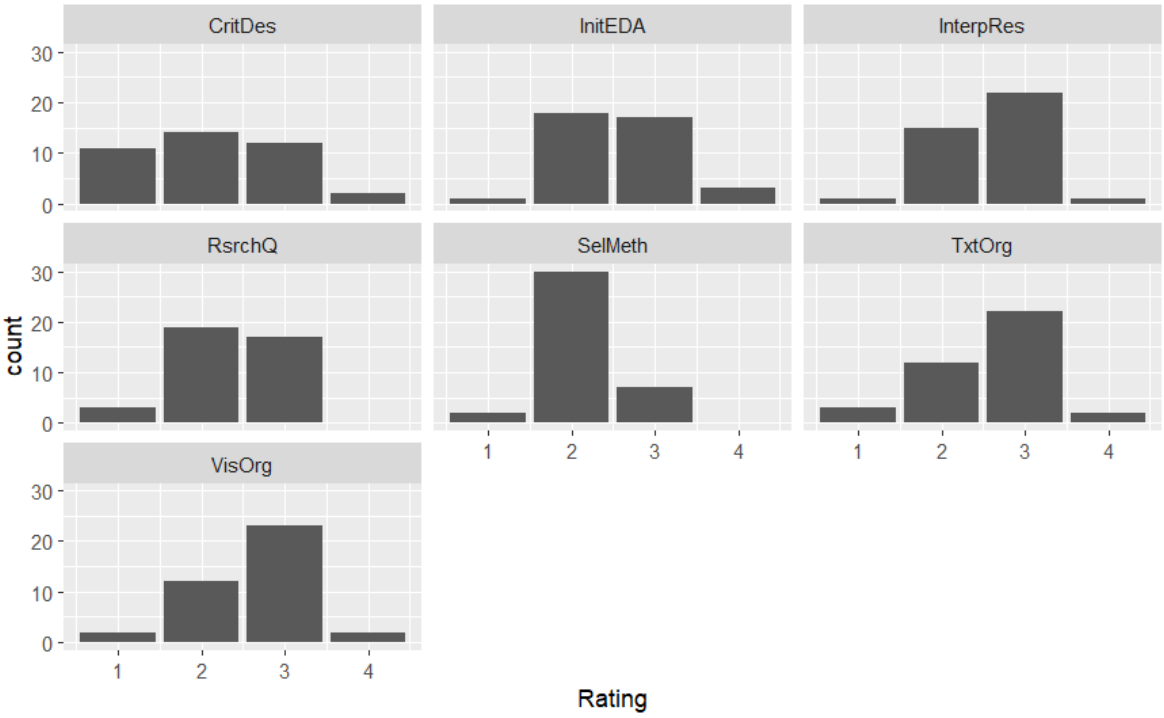


Figure 4: Bar plots of rater 2's ratings for each rubric.

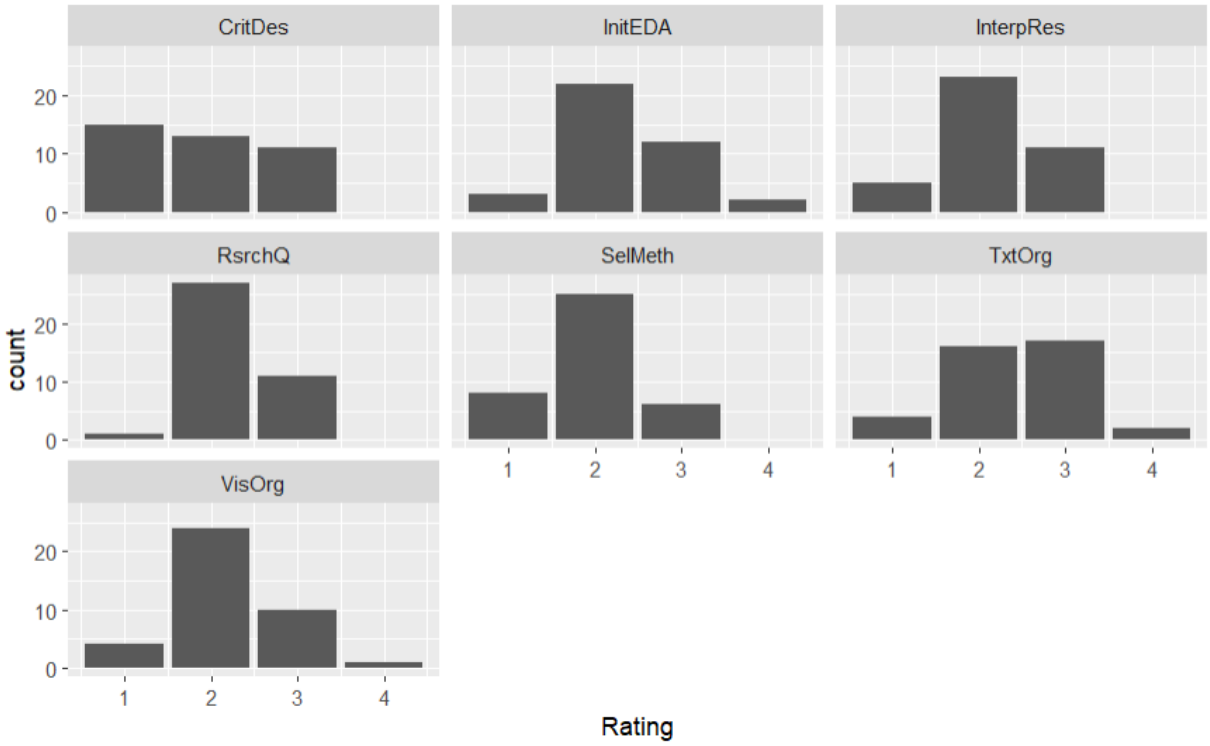


Figure 5: Bar plots of rater 3's ratings for each rubric.

Rubric	ICC for 13 Artifacts	ICC for Full Dataset	Percent Agreement for Rater 1 and 2	Percent Agreement for Rater 1 and 3	Percent Agreement for Rater 2 and 3
RsrchQ	0.19	0.21	0.38	0.77	0.54
CritDes	0.57	0.67	0.54	0.62	0.69
InitEDA	0.49	0.69	0.69	0.54	0.85
SelMeth	0.52	0.47	0.92	0.62	0.69
InterpRes	0.23	0.22	0.62	0.54	0.62
VisOrg	0.59	0.66	0.54	0.77	0.77
TxtOrg	0.14	0.67	0.69	0.62	0.54

Table 8: Intraclass correlations for each rubric for 13 artifacts, full dataset, and percent agreement for each rubric between each pair of raters – all rounded to 2 decimal places. Eg: Raters 1 and 2 agree on ratings for `RsrchQ` 38% of the time.

Variable	Estimate	Standard Error	T Value
Intercept	1.73	0.16	10.59
Rater	0.09	0.07	1.33
RubricInitEDA	0.83	0.19	4.35
RubricInterpRes	1.31	0.19	6.89
RubricRsrchQ	0.81	0.18	4.55
RubricSelMeth	0.51	0.19	2.77
RubricTxtOrg	1.15	0.19	5.97
RubricVisOrg	0.83	0.19	4.30
Rater:RubricInitEDA	-0.15	0.08	-1.77
Rater:RubricInterpRes	-0.36	0.08	-4.40
Rater:RubricRsrchQ	-0.18	0.08	-2.26
Rater:RubricSelMeth	-0.18	0.08	-2.24
Rater:RubricTxtOrg	-0.23	0.08	-2.80
Rater:RubricVisOrg	-0.16	0.08	-1.85

Table 9: Coefficients for Model 1.1.

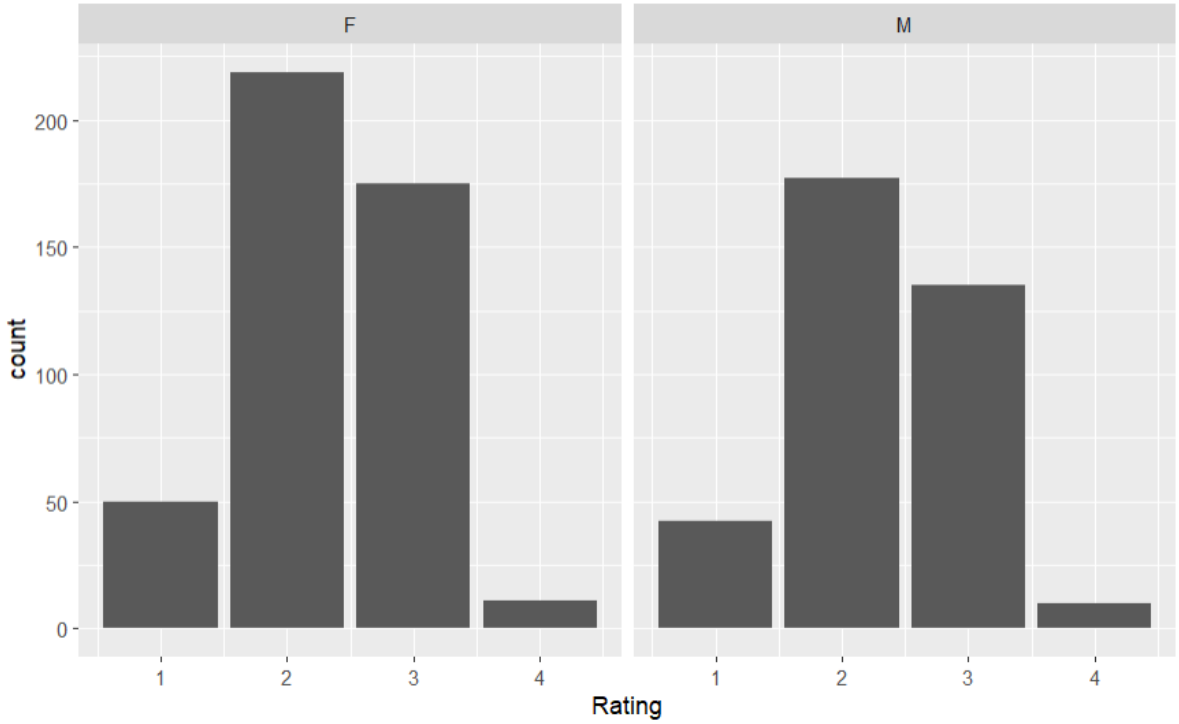


Figure 6: Bar plot of ratings for Female and Male.

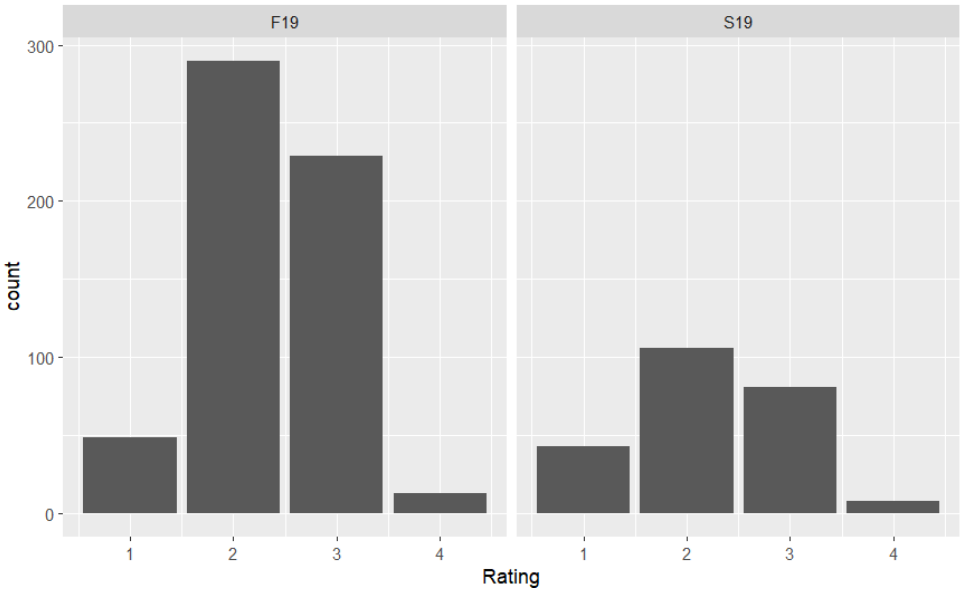


Figure 7: Bar plot of ratings for Fall and Spring semester.

	Female Mean	Male Mean
RsrchQ	2.40	2.29
CritDes	1.80	1.96
InitEDA	2.42	2.46
SelMeth	2.00	2.15
InterpRes	2.52	2.44
VisOrg	2.50	2.31
TxtOrg	2.63	2.56

Table 10: Mean for each rubric's rating, Female vs Male.

Overall Mean	Fall Semester Mean	Spring Semester Mean
2.32	2.35	2.23

Table 11: Difference between means for Fall, Spring, and Overall.

6. DISCUSSION

As a reminder, we were trying to determine the success and fairness of Dietrich College's new "General Education" undergraduate program. Overall, we have concluded that raters are not rating the artifacts the same way, which was confirmed by our mixed model results. Specifically, our analyses and statistical methods all aim to answer the 4 research questions that were presented in the Introduction.

For the first question, we looked at distributions of ratings for each rubric as well as ratings for each rater. This answers the question of whether these distributions differ from rubric

to rubric. We determined that the ratings are not indistinguishable for each rubric, and that the rater's ratings were also not indistinguishable from each other.

For the second question, we looked at exactly how much each rater agreed with one another by calculating intraclass correlations, as well as exact percentage agreement rates between the raters for each rubric. This answers the question of whether the raters disagree, and who disagrees with the others.

For the third question, we built a model that predicts `Rating`, which included fixed effects, random effects, and an interaction term. This answers the question of what factors from this experiment are related to `Rating`. What is concerning about these results is that there is an interaction term between Rubric and Rater in Model 1.1, indicating that the raters are potentially not interpreting the rubrics in the same way. This led to variation in ratings for the rubrics, depending on the rater. Hence, it would be best to train the raters to interpret the rubric in the same way and to grade fairly – this would ensure that the artifacts are being rated more similarly.

Lastly, for the fourth question, we looked at additional EDA to determine what further insights would be interesting to bring forth to the Dean. This answers the question because by being creative and thinking about future steps, we were able to think about what would be both interesting and relevant to discuss with the Dean. We discovered that it would be best to recreate this experiment for `Sex` and `Semester` with equal number of artifacts. Additionally, our results show that the Fall semester had a higher mean in ratings than for both Overall and Spring means. This indicates that as the "General Education" program progressed from Spring 19 to Fall 19 semester, the instructors and/or directors realized that perhaps, a) the difficulty in the courses was too high or that b) the instructors adjusted and improved their teaching styles that allowed for students to perform better on projects. Another thing to note is that we saw that Females had higher ratings, on average, for `TxtOrg` and `VisOrg`, which are both organizational skills. The Dean may consider holding workshops to encourage students to improve upon both their textual and visual organization skills.

Every study has strengths and weaknesses, and specifically with this study, it suffers from several limitations. There was only one method of variable selection for the model that answered question three, so a potentially better model could be produced if other variable selection methods were employed. Additionally, there were some missing values in the dataset that had to be filled in with educated guesses. The missing data occurred in the `Rating` and `Sex` columns in the tall.csv dataset, as mentioned in the Results section for research question four. It could be possible that the way in which we handled missing data may have produced inaccurate analyses and results.

Future work could be done in terms of analyzing `Sex` and `Semester`, as mentioned in the last part of the Discussion section. If there were equal numbers for Female / Male, or Fall / Spring semester artifacts, then comparing the distributions of ratings would provide better results in determining whether there is a difference in ratings (depending on `Sex` and/or `Semester`). Additionally, it might be interesting to look at other courses in the "General Education" program that is not Freshman Statistics, as statistics is generally a difficult

course that may lead to grade deflation and thus, an inaccurate representation of the actual grade distribution and success of the new program.

7. REFERENCES

Junker, B. W. (2021). *Project 02 assignment sheet and data for 36-617: Applied Regression Analysis.* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh PA. Accessed Nov 29, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02.

Midway, S. (2021), *Data Analysis in R*. Available at *https://bookdown.org/steve_midway/DAR/random-effects.html*.

Sheather, S. J. (2009), *A Modern Approach to Regression with R, Springer eBooks.*

8. TECHNICAL APPENDIX

Table of Contents

Appendix A

Check to see where missing data is. Replace missing data with mode of sex and semester. Create histograms, barplots, and numerical summaries for each rubric. Then, repeat for each rater.

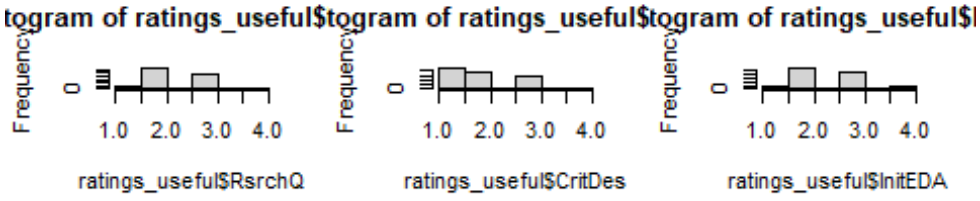
```
ratings_useful <- ratings[,-c(1,3,4)]
## x, sample, overlap are useless vars - remove them from data

ratings_useful_13 <- ratings_useful %>%
  filter(Repeated == 1) ## subset of data with 13 artifacts that had all 3 ra
ters rate them
tall_13 <- tall %>% filter(Repeated == 1)

ratings_useful_91 <- ratings_useful %>%
  filter(Repeated == 0)
tall_91 <- tall %>% filter(Repeated == 0)

which(tall$Sex == "") ## indices for missing data in sex ## set missing data
to female (mode)

idx <- as.vector(which(tall$Sex == ""))
  for (i in idx) {
```

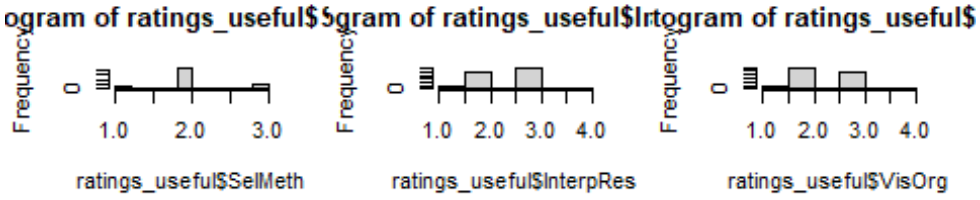
```
    tall$Sex[i] <- "F"
}

which(is.na(tall$Rating)) ## indices for na for ratings
## set missing data (na) to mode of rating = 2

tall$Rating[161] <- 2
tall$Rating[684] <- 2

## set missing values in ratings dataset to female and 2 (mode)
ratings$Sex[5] <- "F"
ratings$CritDes[44] <- 2

## distributions and numeric summaries of each rubric
par(mfrow=c(3,3))
hist(ratings_useful$RsrchQ)
summary(ratings_useful$RsrchQ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    2.00    2.35    3.00    4.00

hist(ratings_useful$CritDes)
summary(ratings_useful$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   1.871   3.000   4.000       1

hist(ratings_useful$InitEDA)
summary(ratings_useful$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.436   3.000   4.000

hist(ratings_useful$SelMeth)
summary(ratings_useful$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.068   2.000   3.000

hist(ratings_useful$InterpRes)
summary(ratings_useful$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.487   3.000   4.000

hist(ratings_useful$VisOrg)
summary(ratings_useful$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   2.000   2.414   3.000   4.000       1

hist(ratings_useful$TxtOrg)
summary(ratings_useful$TxtOrg)
```

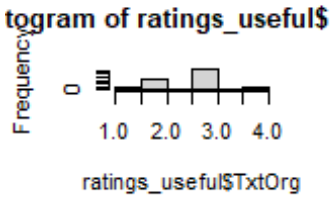
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.598   3.000   4.000
```

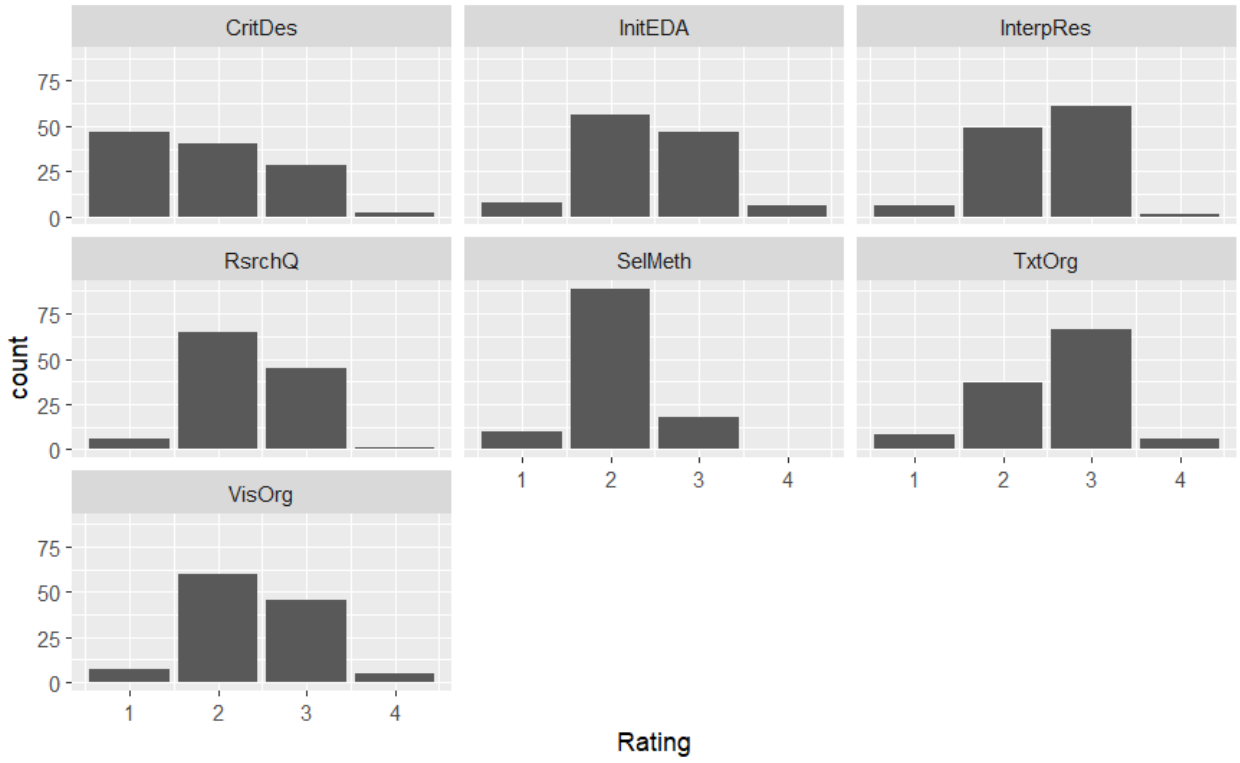
```
ggplot(tall,aes(x = Rating)) + facet_wrap( ~ Rubric) + geom_bar()
```

```
whole <- describe(ratings_useful[ , c('RsrchQ', 'CritDes','InitEDA','SelMeth'
,'InterpRes','VisOrg','TxtOrg')], fast=TRUE)
names(whole) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum","
Maximum","Range","Standard Error")

by_rater <- describeBy(ratings_useful[,c('RsrchQ', 'CritDes','InitEDA','SelMe
th','InterpRes','VisOrg','TxtOrg')], group=ratings_useful$Rater, fast=TRUE)

rater1 <- by_rater$`1`
names(rater1) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")
rater2 <- by_rater$`2`
names(rater2) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")
rater3 <- by_rater$`3`
names(rater3) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")
```





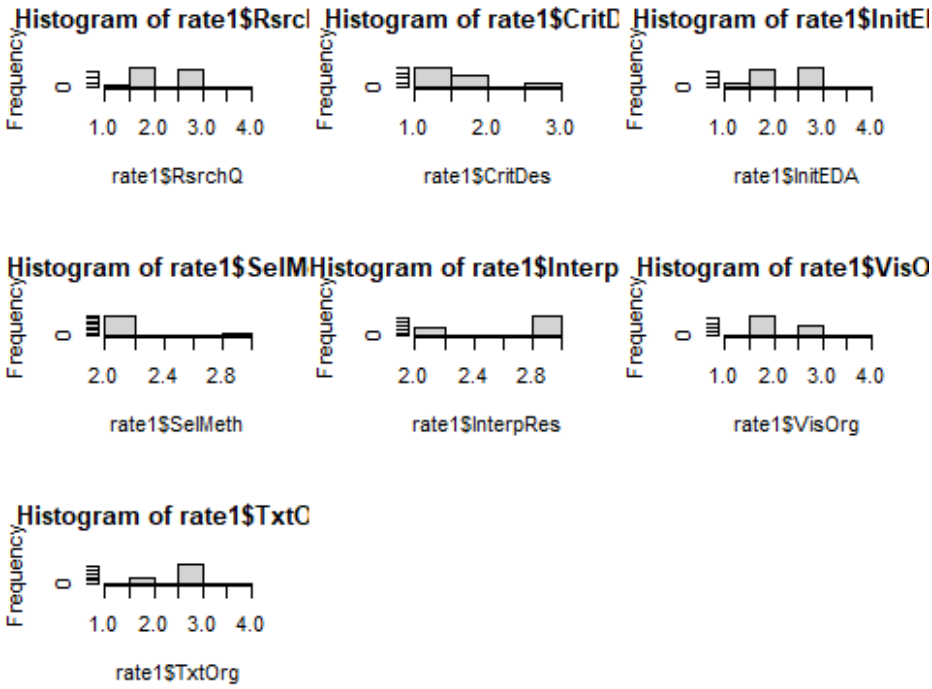


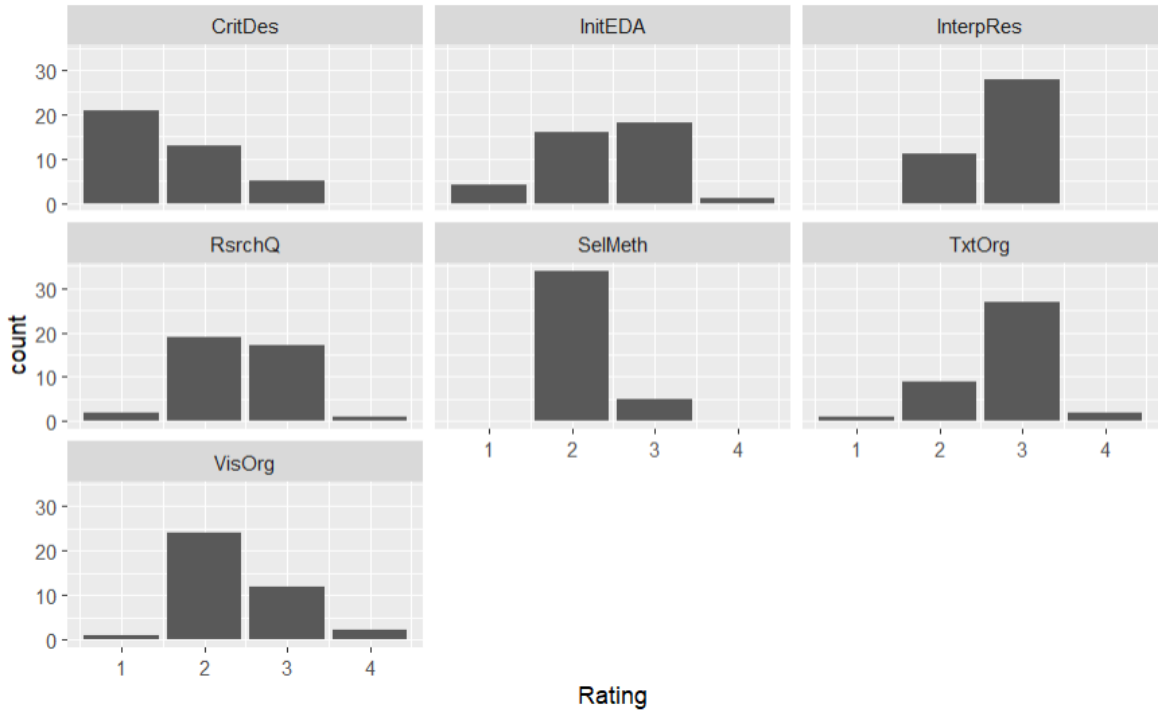


looking at the distributions of the scores for each of the 7 rubrics, it looks like critique design scored lowest (extremely right skewed), while research question, initial eda, and visual organization scored lower (right skewed). selection method seemed to be scored very fairly (almost uniform distribution). text organization scored slightly better than all 7 rubrics, with the highest mean of 2.598.

```
## subset data for each rater
rate1 <- ratings_useful %>%
  filter(Rater == 1)
rate1.tall <- tall %>% filter(Rater == 1)

rate2 <- ratings_useful %>%
  filter(Rater == 2)
rate2.tall <- tall %>% filter(Rater == 2)

rate3 <- ratings_useful %>%
  filter(Rater == 3)
rate3.tall <- tall %>% filter(Rater == 3)

## rater 1 distributions
par(mfrow=c(3,3))
hist(rate1$RsrchQ)
summary(rate1$RsrchQ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.436   3.000   4.000
```

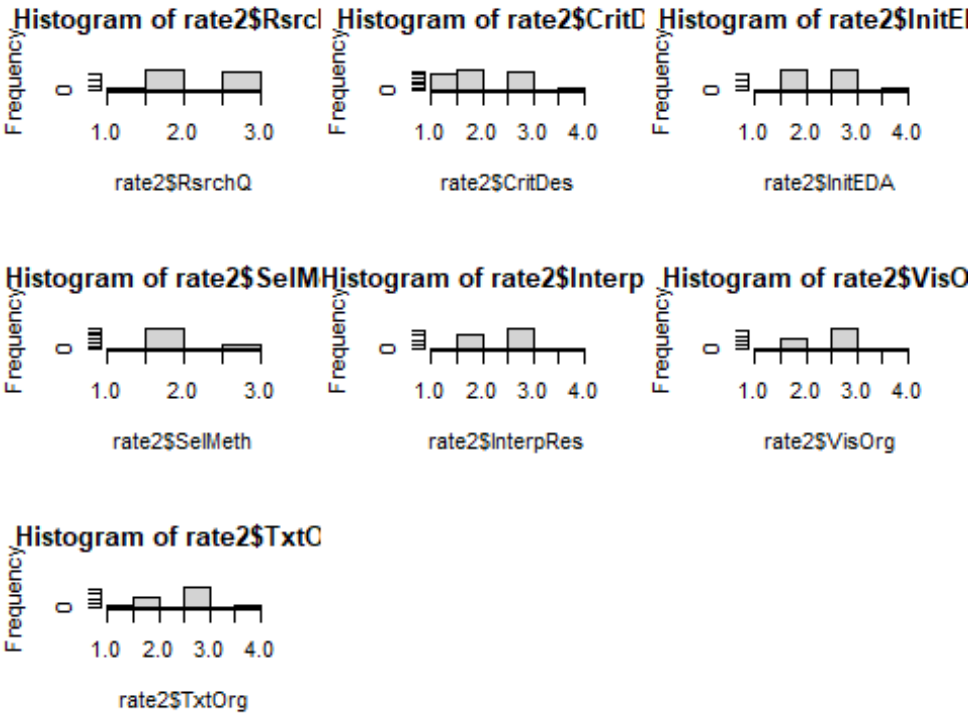
```
hist(rate1$CritDes)
summary(rate1$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    1.00    1.59    2.00    3.00

hist(rate1$InitEDA)
summary(rate1$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    2.00    2.41    3.00    4.00

hist(rate1$SelMeth)
summary(rate1$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   2.000   2.128   2.000   3.000

hist(rate1$InterpRes)
summary(rate1$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   3.000   2.718   3.000   3.000

hist(rate1$VisOrg)
summary(rate1$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   2.000   2.395   3.000   4.000       1

hist(rate1$TxtOrg)
summary(rate1$TxtOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.500   3.000   2.769   3.000   4.000

ggplot(rate1.tall,aes(x = Rating)) + facet_wrap( ~ Rubric) + geom_bar()
```

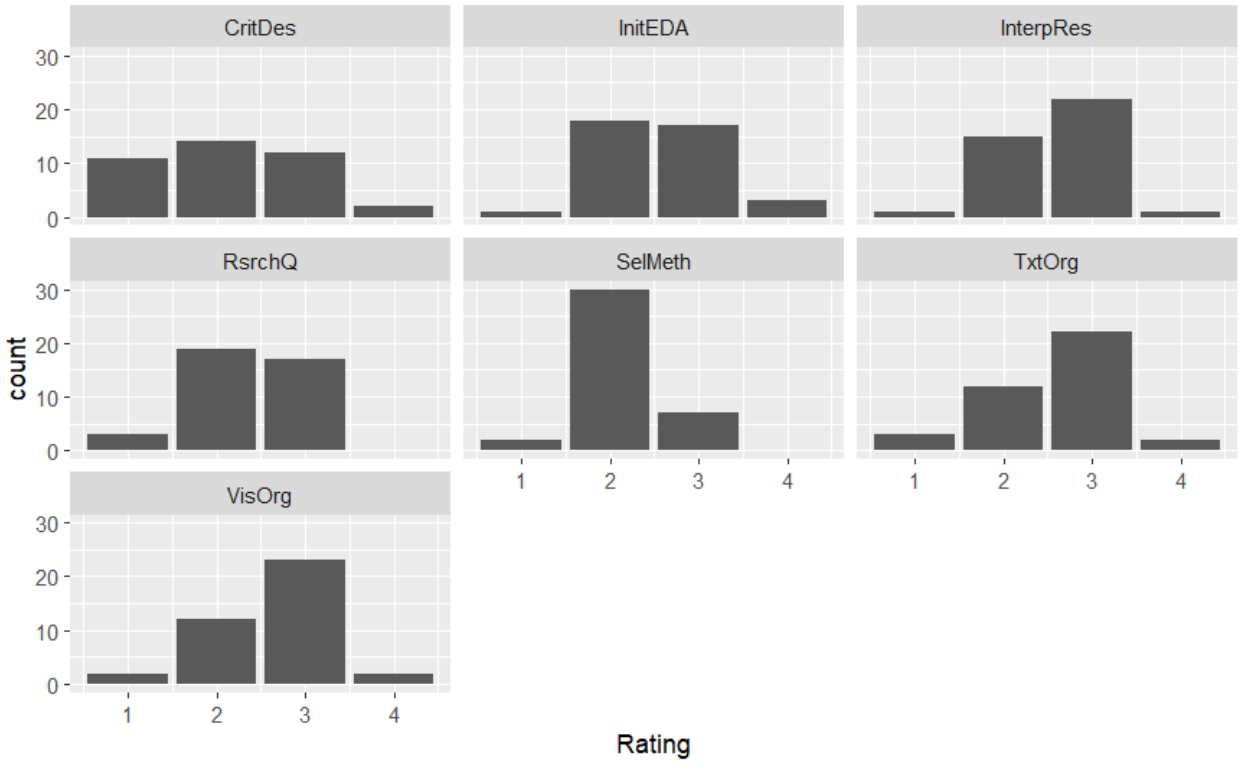





```
## rater 2 distributions
par(mfrow=c(3,3))
hist(rate2$RsrchQ)
summary(rate2$RsrchQ)
```

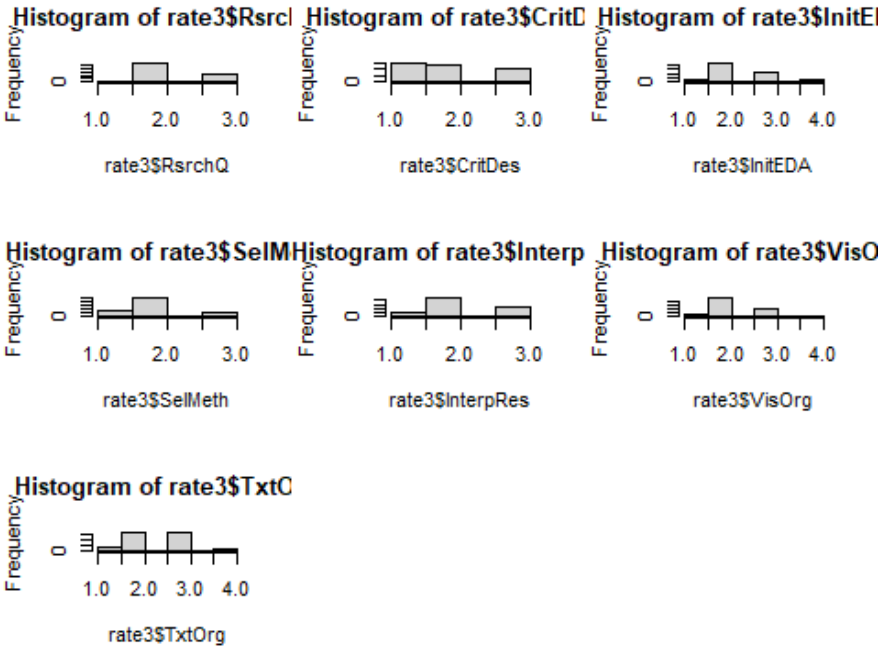
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.359   3.000   3.000

hist(rate2$CritDes)
summary(rate2$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   2.132   3.000   4.000       1

hist(rate2$InitEDA)
summary(rate2$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.564   3.000   4.000

hist(rate2$SelMeth)
summary(rate2$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.128   2.000   3.000

hist(rate2$InterpRes)
summary(rate2$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00    2.59    3.00    4.00

hist(rate2$VisOrg)
summary(rate2$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.641   3.000   4.000

hist(rate2$TxtOrg)
summary(rate2$TxtOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00    2.59    3.00    4.00

ggplot(rate2.tall,aes(x = Rating)) + facet_wrap( ~ Rubric) + geom_bar()
```

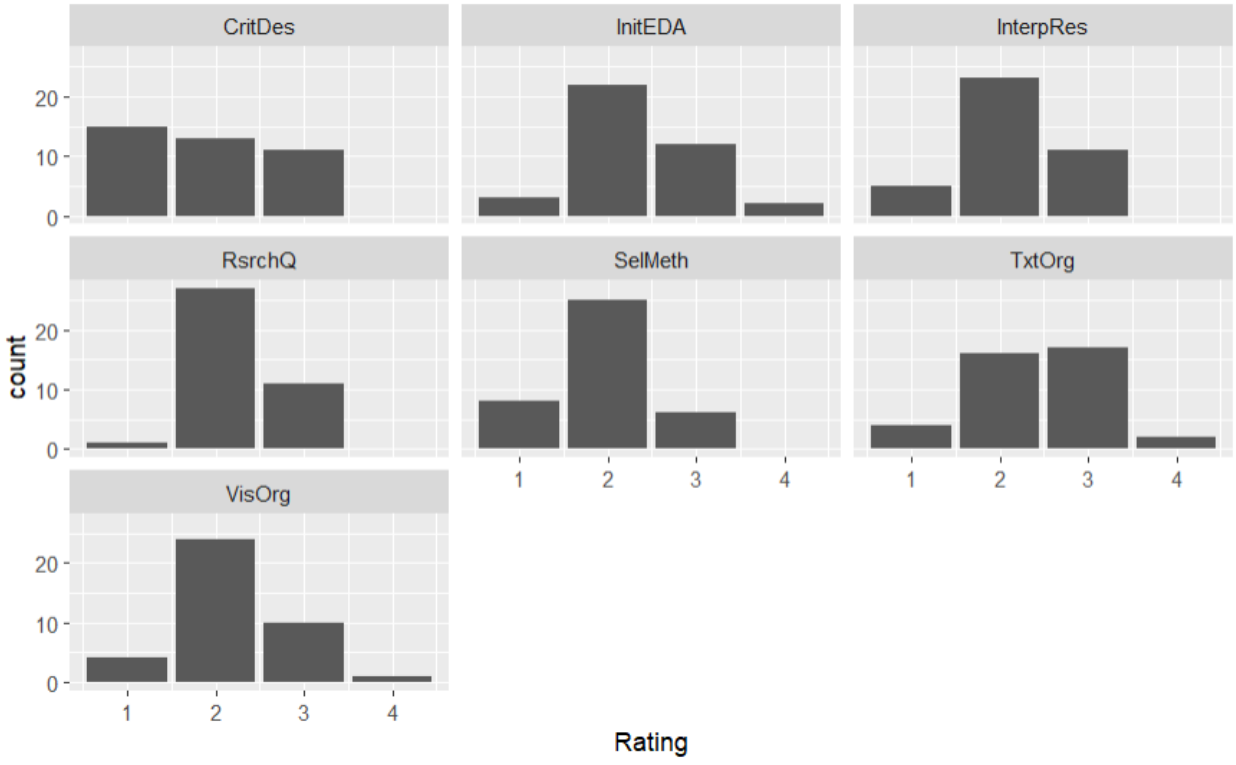


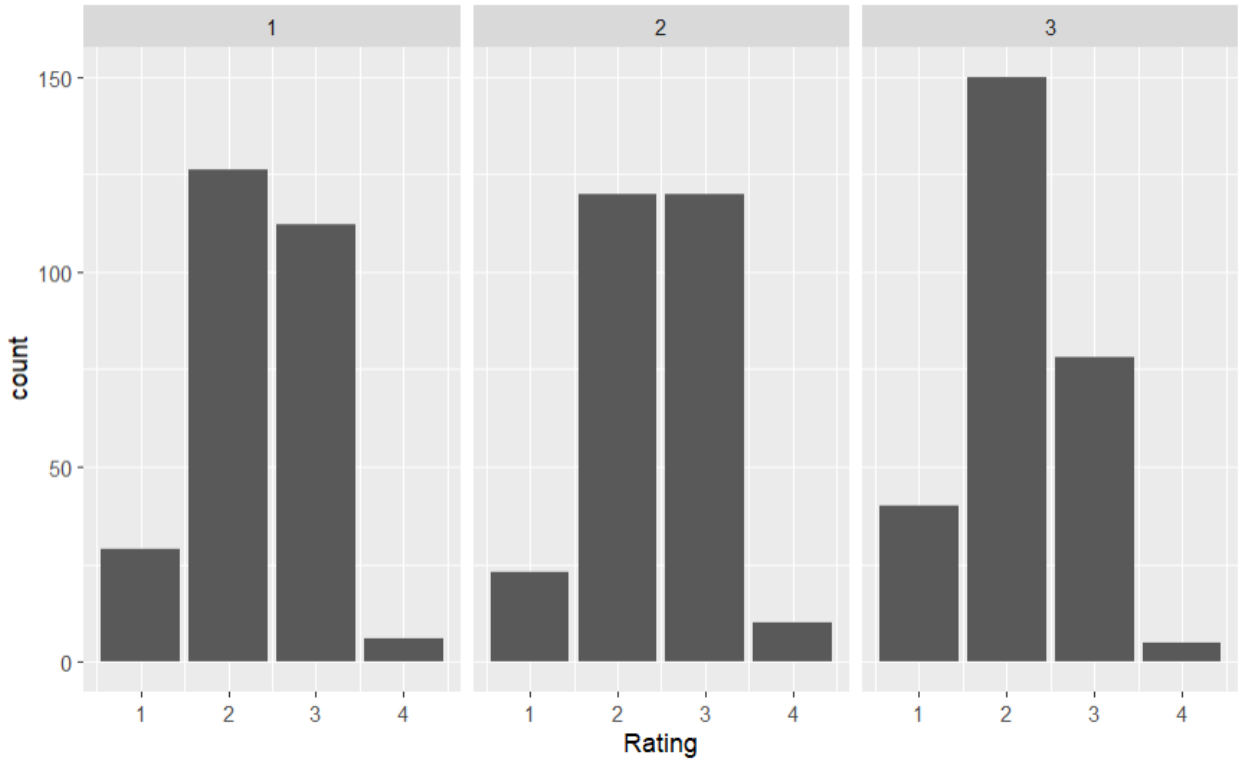



```
## rater 3 distributions
par(mfrow=c(3,3))
```

```
hist(rate3$RsrchQ)
summary(rate3$RsrchQ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.256   3.000   3.000

hist(rate3$CritDes)
summary(rate3$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.897   3.000   3.000

hist(rate3$InitEDA)
summary(rate3$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.333   3.000   4.000

hist(rate3$SelMeth)
summary(rate3$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.949   2.000   3.000

hist(rate3$InterpRes)
summary(rate3$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.154   3.000   3.000

hist(rate3$VisOrg)
summary(rate3$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.205   3.000   4.000

hist(rate3$TxtOrg)
summary(rate3$TxtOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.436   3.000   4.000

ggplot(rate3.tall,aes(x = Rating)) + facet_wrap( ~ Rubric) + geom_bar()

ggplot(tall,aes(x = Rating)) + facet_wrap( ~ Rater) + geom_bar()
```


Histogram of rate3$RsrcI
Frequency
0
1.0 2.0 3.0
rate3$RsrchQ
Histogram of rate3$CritD
Frequency
0
1.0 2.0 3.0
rate3$CritDes
Histogram of rate3$InitEI
Frequency
0
1.0 2.0 3.0 4.0
rate3$InitEDA
Histogram of rate3$SelM
Frequency
0
1.0 2.0 3.0
rate3$SelMeth
Histogram of rate3$Interp
Frequency
0
1.0 2.0 3.0
rate3$InterpRes
Histogram of rate3$VisO
Frequency
0
1.0 2.0 3.0 4.0
rate3$VisOrg
Histogram of rate3$TxtO
Frequency
0
1.0 2.0 3.0 4.0
rate3$TxtOrg


CritDes
InitEDA
InterpRes
RsrchQ
SelMeth
TxtOrg
VisOrg
count
0
10
20
1
2
3
4
Rating

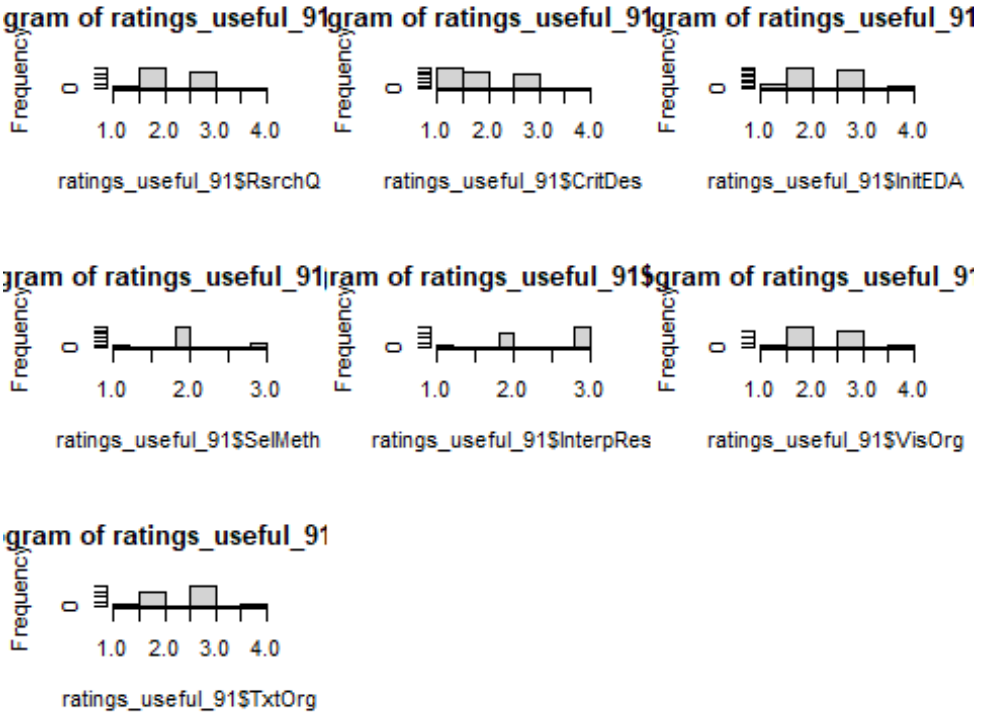


it looks like rater 3 is a bit harsher than the other 2 raters. most of the distributions for the rubrics for rater 3 are closer to right skewed. rater 1 is the only rater that sometimes gives binary ratings, meaning only rating 2 values, as opposed to 3 or 4 ratings. this is confirmed by the bar plot of each rater's ratings. rater 3 has a right skewed distribution of ratings, meaning they tend to give lower scores (1 and 2).

Repeat same thing (barplots, histograms, rubrics for each rubric) for 91 artifacts that were rated only by 1 rater.

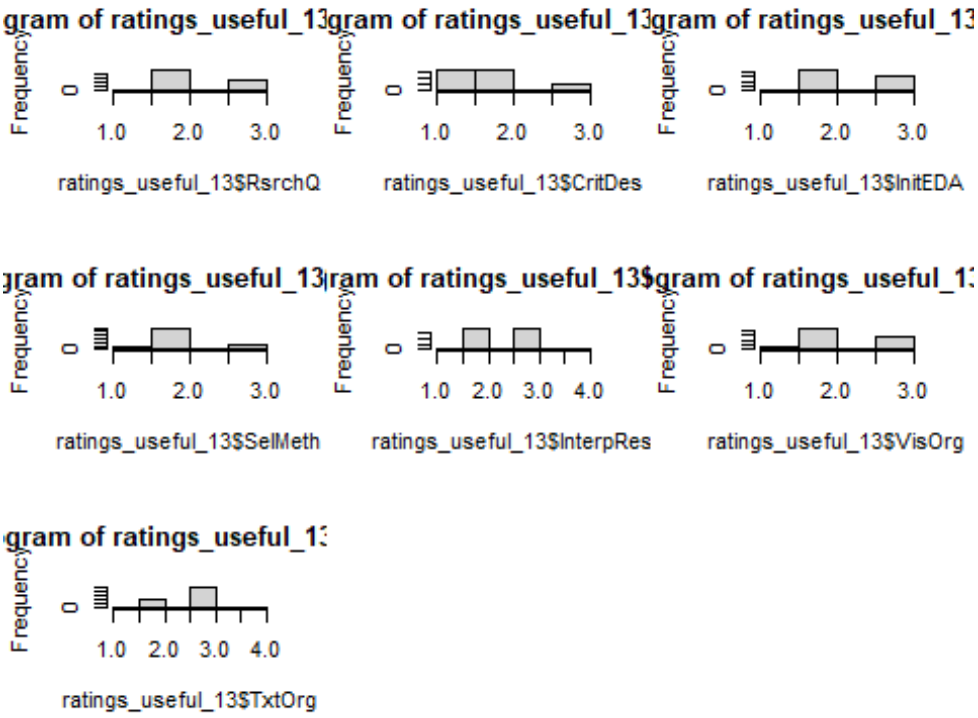
```
## distributions and summaries for 91 artifacts
par(mfrow=c(3,3))
hist(ratings_useful_91$RsrchQ)
summary(ratings_useful_91$RsrchQ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.385   3.000   4.000

hist(ratings_useful_91$CritDes)
summary(ratings_useful_91$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   1.948   3.000   4.000       1

hist(ratings_useful_91$InitEDA)
summary(ratings_useful_91$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.462   3.000   4.000
```

```
hist(ratings_useful_91$SelMeth)
summary(ratings_useful_91$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.077   2.000   3.000

hist(ratings_useful_91$InterpRes)
summary(ratings_useful_91$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.474   3.000   3.000

hist(ratings_useful_91$VisOrg)
summary(ratings_useful_91$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   2.000   2.481   3.000   4.000       1

hist(ratings_useful_91$TxtOrg)
summary(ratings_useful_91$TxtOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.564   3.000   4.000
```



Repeat same thing (histogram, barplot, summaries) for each rating for subset of 13 artifacts that was rated by all 3 raters.

```
## distributions for subset of 13 artifacts rated by all 3 raters
par(mfrow=c(3,3))
hist(ratings_useful_13$RsrchQ)
summary(ratings_useful_13$RsrchQ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.282   3.000   3.000

hist(ratings_useful_13$CritDes)
summary(ratings_useful_13$CritDes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.718   2.000   3.000

hist(ratings_useful_13$InitEDA)
summary(ratings_useful_13$InitEDA)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.385   3.000   3.000

hist(ratings_useful_13$SelMeth)
summary(ratings_useful_13$SelMeth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.051   2.000   3.000

hist(ratings_useful_13$InterpRes)
summary(ratings_useful_13$InterpRes)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.513   3.000   4.000

hist(ratings_useful_13$VisOrg)
summary(ratings_useful_13$VisOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.282   3.000   3.000

hist(ratings_useful_13$TxtOrg)
summary(ratings_useful_13$TxtOrg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.667   3.000   4.000
```



comparing the distributions for each rubric between the subset of 91 artifacts vs the subset of the 13 artifacts, the distributions for each rubric actually look quite similar between the 2 different datasets. this means that the subset of data could actually be representative of the entire set of 91 artifacts.

Appendix B

Focus first only on 13 artifacts that were rated by all 3 raters. Generate 7 lmer models, one for each rubric, with artifact as the random effect because then you have 13 groups in which you can check to see the correlation of ratings for each artifact by each rater.

```
## create 13 artifacts subset using tall data
tall_13 <- tall %>%
  filter(Repeated == 1) %>%
  select(-X)

## ratings for research question
## group is which artifact (13 groups) b/c then you can check to see correlat
ion between each rater's ratings for each artifact
## icc is calculated by sigma^2 / (sigma^2 + tao^2), where sigma^2 is artifac
t variance and tao^2 is residual variance
## can also use icc function from whatever function to make life easier witho
ut having to copy and paste so much
icc_sub <- c()

rsrchq.ratings <- tall_13[tall_13$Rubric=="RsrchQ",]
mod1 <- lmer(Rating ~ 1 + (1|Artifact), data=rsrchq.ratings)
```

```
summary(mod1)
icc_sub[1] <- icc(mod1)[[1]]

critdes.ratings <- tall_13[tall_13$Rubric=="CritDes",]
mod2 <- lmer(Rating ~ 1 + (1|Artifact), data=critdes.ratings)
summary(mod2)
icc_sub[2] <- icc(mod2)[[2]]

initeda.ratings <- tall_13[tall_13$Rubric=="InitEDA",]
mod3 <- lmer(Rating ~ 1 + (1|Artifact), data=initeda.ratings)
summary(mod3)
icc_sub[3] <- icc(mod3)[[1]]

selmeth.ratings <- tall_13[tall_13$Rubric=="SelMeth",]
mod4 <- lmer(Rating ~ 1 + (1|Artifact), data=selmeth.ratings)
summary(mod4)
icc_sub[4] <- icc(mod4)[[1]]

interpres.ratings <- tall_13[tall_13$Rubric=="InterpRes",]
mod5 <- lmer(Rating ~ 1 + (1|Artifact), data=interpres.ratings)
summary(mod5)
icc_sub[5] <- icc(mod5)[[1]]

visorg.ratings <- tall_13[tall_13$Rubric=="VisOrg",]
mod6 <- lmer(Rating ~ 1 + (1|Artifact), data=visorg.ratings)
summary(mod6)
icc_sub[6] <- icc(mod6)[[1]]

txtorg.ratings <- tall_13[tall_13$Rubric=="TxtOrg",]
mod7 <- lmer(Rating ~ 1 + (1|Artifact), data=txtorg.ratings)
summary(mod7)
icc_sub[7] <- icc(mod7)[[1]]

rubric = c(unique(tall$Rubric))

data.frame(rubric, icc_sub)

##      rubric   icc_sub
## 1    RsrchQ 0.1891892
## 2   CritDes 0.5725594
## 3   InitEDA 0.4929577
## 4   SelMeth 0.5212766
## 5 InterpRes 0.2295720
## 6    VisOrg 0.5924529
## 7    TxtOrg 0.1428571
```

icc values

researchq: 0.1891918 critdes: 0.5725134 initeda: 0.4930784 selmeth: 0.5212845 interpres: 0.2295821 visorg: 0.5924748 txtorg: 0.1428682 comparing the icc values for the

rubrics, critdes, initeda, selmeth, and visorg are the highest, meaning that the 3 raters agreed the most on these rubric items. the lower the icc value, the less the raters agreed on rubric items. it looks like they disagreed the most on txtorg.

Create vectors for exact percentages and append each pair of rater's exact agreement rates into respective vector.

Create table for each pair of raters that shows the number of ratings for each rubric. Then, calculate the exact percentage of agreement by dividing the sum of the diagonal by 13. Repeat this for each rubric.

```
agree_12 <- c()
agree_23 <- c()
agree_13 <- c()

## create table that shows the number of ratings for rater 1 and 2, with main
diagonal as the number where raters 1 and 2 agree with each other

## create data frame with rater 1 and rater 2 ratings for research q rubric
raters_1_and_2_on_RsrchQ <- data.frame(r1=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==1],
                                       r2=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==2],
                                       a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2])

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)
t12
agree_12[1] <- 5/13
```

```
##    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
```

rater 1 and 2 have a 5/13 = 38% agreement for rsrchq

```
## create table that shows the number of ratings for rater 1 and 3, with main
diagonal as the number where raters 1 and 3 agree with each other
raters_1_and_3_on_RsrchQ <- data.frame(r1=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==1],
                                       r3=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==3],
                                       a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
```

```
                                        a3=ratings_useful_13$Artifact[ratings_
useful_13$Rater==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)
t13
agree_13[1] <- 10/13


##    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
```

raters 1 and 3 have a 10/13 = 77% agreement for rsrchq

```
## create table that shows the number of ratings for rater 2 and 3, with main
diagonal as the number where raters 2 and 3 agree with each other
raters_2_and_3_on_RsrchQ <- data.frame(r2=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==2],
                                       r3=ratings_useful_13$RsrchQ[ratings_us
eful_13$Rater==3],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                       a3=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t23
agree_23[1] <- 7/13

##    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
```

raters 2 and 3 have a 7/13 = 54% agreement for rsrchq

for rsrchq, raters 1 and 3 agree 77% of the time. however, again, rater 2 is the one that disagrees more, especially when compared to rater 1.

```
## do the same for critdes
raters_1_and_2_on_CritDes <- data.frame(r1=ratings_useful_13$CritDes[ratings_
useful_13$Rater==1],
```

```
                                          r2=ratings_useful_13$CritDes[ratings_u
seful_13$Rater==2],
                                          a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                          a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2])

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)
t12
agree_12[2] <- 7/13
```

```
##    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
```

raters 1 and 2 have a 7/13 = 54% agreement for critdes

```
raters_1_and_3_on_CritDes <- data.frame(r1=ratings_useful_13$CritDes[ratings_
useful_13$Rater==1],
                                          r3=ratings_useful_13$CritDes[ratings_u
seful_13$Rater==3],
                                          a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                          a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t13
agree_13[2] <- 8/13
```

```
##    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
```

raters 1 and 3 have 8/13 = 62% agreement for critdes

```
raters_2_and_3_on_CritDes <- data.frame(r2=ratings_useful_13$CritDes[ratings_
useful_13$Rater==2],
                                          r3=ratings_useful_13$CritDes[ratings_u
seful_13$Rater==3],
                                          a2=ratings_useful_13$Artifact[ratings_
```

```
useful_13$Rater==2],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t23
agree_23[2] <- 9/13
```

```
##    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
```

raters 2 and 3 have a 9/13 = 69% agreement for critdes

for critdes, rater 2 seems to disagree more.

```
## do the same for initeda
raters_1_and_2_on_InitEDA <- data.frame(r1=ratings_useful_13$InitEDA[ratings_
useful_13$Rater==1],
                                        r2=ratings_useful_13$InitEDA[ratings_u
seful_13$Rater==2],
                                        a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2])

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)
t12
agree_12[3] <- 9/13
```

```
##    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
```

raters 1 and 2 have a 9/13 = 69% agreement for initeda

```
raters_1_and_3_on_InitEDA <- data.frame(r1=ratings_useful_13$InitEDA[ratings_
useful_13$Rater==1],
                                        r3=ratings_useful_13$InitEDA[ratings_u
seful_13$Rater==3],
                                        a1=ratings_useful_13$Artifact[ratings_
```

```
useful_13$Rater==1],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==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)
t13
agree_13[3] <- 7/13
```

```
##    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
```

raters 1 and 3 have a 7/13 = 54% agreement for initeda

```
raters_2_and_3_on_InitEDA <- data.frame(r2=ratings_useful_13$InitEDA[ratings_
useful_13$Rater==2],
                                        r3=ratings_useful_13$InitEDA[ratings_u
seful_13$Rater==3],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==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)
t23
agree_23[3] <- 11/13
```

```
##    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
```

raters 2 and 3 have a 11/13 = 85% agreement for initeda

for initeda, this time rater 3 is the one that disagrees more. surprisingly, rater 1 and 2 have a relatively high agreement rate for initeda.

```
## do the same for selmeth
raters_1_and_2_on_SelMeth <- data.frame(r1=ratings_useful_13$SelMeth[ratings_
useful_13$Rater==1],
                                        r2=ratings_useful_13$SelMeth[ratings_u
seful_13$Rater==2],
```

```
                                        a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==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)
t12
agree_12[4] <- 12/13
```

```
##    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
```

raters 1 and 2 have a 12/13 = 92% agreement for selmeth

```
raters_1_and_3_on_SelMeth<- data.frame(r1=ratings_useful_13$SelMeth[ratings_u
seful_13$Rater==1],
                                        r3=ratings_useful_13$SelMeth[ratings_u
seful_13$Rater==3],
                                        a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t13
agree_13[4] <- 8/13
```

```
##    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
```

raters 1 and 3 have a 8/13 = 62% agreement for selmeth

```
raters_2_and_3_on_SelMeth <- data.frame(r2=ratings_useful_13$SelMeth[ratings_
useful_13$Rater==2],
                                        r3=ratings_useful_13$SelMeth[ratings_u
seful_13$Rater==3],
                                        a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                        a2=ratings_useful_13$Artifact[ratings_
```

```
useful_13$Rater==3])

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)
t23
agree_23[4] <- 9/13
```

```
##    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
```

raters 2 and 3 have a 9/13 = 69% agreement on selmeth

for selmeth, the agreement rates are relatively high between all 3 raters.

```
## do the same for interpres
raters_1_and_2_on_InterpRes <- data.frame(r1=ratings_useful_13$InterpRes[rati
ngs_useful_13$Rater==1],
                                          r2=ratings_useful_13$InterpRes[ratings
_useful_13$Rater==2],
                                          a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                          a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2])

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)
t12
agree_12[5] <- 8/13
```

```
##    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
```

raters 1 and 2 have a 8/13 = 62% agreement for interpres

```
raters_1_and_3_on_InterpRes<- data.frame(r1=ratings_useful_13$InterpRes[ratin
gs_useful_13$Rater==1],
                                          r3=ratings_useful_13$InterpRes[ratings
_useful_13$Rater==3],
                                          a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                          a2=ratings_useful_13$Artifact[ratings_
```

```
useful_13$Rater==3])

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)
t13
agree_13[5] <- 7/13
```

```
##    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
```

raters 1 and 3 have a 7/13 = 54% agreement for interpres

```
raters_2_and_3_on_InterpRes <- data.frame(r2=ratings_useful_13$InterpRes[rati
ngs_useful_13$Rater==2],
                                          r3=ratings_useful_13$InterpRes[ratings
_useful_13$Rater==3],
                                          a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                          a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t23
agree_23[5] <- 8/13
```

```
##    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
```

raters 2 and 3 have a 8/13 = 62% agreement on interpres

for interpres, relatively the same agreement rates across all 3 raters.

```
## do the same for visorg
raters_1_and_2_on_VisOrg <- data.frame(r1=ratings_useful_13$VisOrg[ratings_us
eful_13$Rater==1],
                                       r2=ratings_useful_13$VisOrg[ratings_us
eful_13$Rater==2],
                                       a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                       a2=ratings_useful_13$Artifact[ratings_
```

```
useful_13$Rater==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)
t12
agree_12[6] <- 7/13
```

```
##    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
```

raters 1 and 2 have a 7/13 = 54% agreement on visorg

```
raters_1_and_3_on_VisOrg<- data.frame(r1=ratings_useful_13$VisOrg[ratings_use
ful_13$Rater==1],
                                      r3=ratings_useful_13$VisOrg[ratings_us
eful_13$Rater==3],
                                      a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                      a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t13
agree_13[6] <- 10/13
```

```
##    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
```

raters 1 and 3 have a 10/13 = 77% agreement for visorg

```
raters_2_and_3_on_VisOrg <- data.frame(r2=ratings_useful_13$VisOrg[ratings_us
eful_13$Rater==2],
                                       r3=ratings_useful_13$VisOrg[ratings_us
eful_13$Rater==3],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t23
agree_23[6] <- 10/13
```

```
##    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
```

raters 2 and 3 have a 10/13 = 77% agreement for visorg

not sure what to say about rater agreement for visorg?

```
## do the same for txtorg
raters_1_and_2_on_TxtOrg <- data.frame(r1=ratings_useful_13$TxtOrg[ratings_us
eful_13$Rater==1],
                                       r2=ratings_useful_13$TxtOrg[ratings_us
eful_13$Rater==2],
                                       a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==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)
t12
agree_12[7] <- 9/13
```

```
##    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
```

raters 1 and 2 have a 9/13 = 69% agreement on txtorg

```
raters_1_and_3_on_TxtOrg<- data.frame(r1=ratings_useful_13$TxtOrg[ratings_use
ful_13$Rater==1],
                                       r3=ratings_useful_13$TxtOrg[ratings_us
eful_13$Rater==3],
                                       a1=ratings_useful_13$Artifact[ratings_
useful_13$Rater==1],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t13
agree_13[7] <- 8/13
```

```
##    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
```

raters 1 and 3 have a 8/13 = 62% agreement for txtorg

```
raters_2_and_3_on_TxtOrg <- data.frame(r2=ratings_useful_13$TxtOrg[ratings_us
eful_13$Rater==2],
                                       r3=ratings_useful_13$TxtOrg[ratings_us
eful_13$Rater==3],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==2],
                                       a2=ratings_useful_13$Artifact[ratings_
useful_13$Rater==3])

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)
t23
agree_23[7] <- 7/13
```

```
##    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
```

raters 2 and 3 have a 7/13 = 54% agreement for txtorg

relatively the same agreement rate for txtorg.

```
## repeat icc for full dataset (178 rows)
icc_full <- c()

rsrchq.ratings <- tall[tall$Rubric=="RsrchQ",]
mlm1 <- lmer(Rating ~ 1 + (1|Artifact), data=rsrchq.ratings)
summary(mlm1)
icc_full[1] <- icc(mlm1)[[1]]

critdes.ratings <- tall[tall$Rubric=="CritDes",]
mlm2 <- lmer(Rating ~ 1 + (1|Artifact), data=critdes.ratings)
summary(mlm2)
```

```
icc_full[2] <- icc(mlm2)[[1]]

initeda.ratings <- tall[tall$Rubric=="InitEDA",]
mlm3 <- lmer(Rating ~ 1 + (1|Artifact), data=initeda.ratings)
summary(mlm3)
icc_full[3] <- icc(mlm3)[[1]]

selmeth.ratings <- tall[tall$Rubric=="SelMeth",]
mlm4 <- lmer(Rating ~ 1 + (1|Artifact), data=selmeth.ratings)
summary(mlm4)
icc_full[4] <- icc(mlm4)[[1]]

interpres.ratings <- tall[tall$Rubric=="InterpRes",]
mlm5 <- lmer(Rating ~ 1 + (1|Artifact), data=interpres.ratings)
summary(mlm5)
icc_full[5] <- icc(mlm5)[[1]]

visorg.ratings <- tall[tall$Rubric=="VisOrg",]
mlm6 <- lmer(Rating ~ 1 + (1|Artifact), data=visorg.ratings)
summary(mlm6)
icc_full[6] <- icc(mlm6)[[1]]

txtorg.ratings <- tall[tall$Rubric=="Txtorg",]
mlm7 <- lmer(Rating ~ 1 + (1|Artifact), data=critdes.ratings)
summary(mlm7)
icc_full[7] <- icc(mlm7)[[1]]

rubric = c(unique(tall$Rubric))

data.frame(rubric, icc_full, icc_sub)

##      rubric  icc_full   icc_sub
## 1    RsrchQ 0.2096214 0.1891892
## 2   CritDes 0.6699202 0.5725594
## 3   InitEDA 0.6867210 0.4929577
## 4   SelMeth 0.4719014 0.5212766
## 5 InterpRes 0.2200285 0.2295720
## 6    VisOrg 0.6586320 0.5924529
## 7    TxtOrg 0.6699202 0.1428571
```

icc's for rubrics rsrchq: 0.2096214 critdes: 0.6730647 initeda: 0.6867210 selmeth: 0.4719014 interpres: 0.2200285 visorg: 0.6607372 txtorg: 0.6730647

critdes, initeda, visorg, and txtorg have the highest icc's. this means the raters agree the most for these 4 rubrics. when comparing to the subset of 13 artifcats, the icc's are not the same, especially for txtorg - icc is much higher for full dataset. otherwise, the icc's are similar enough.

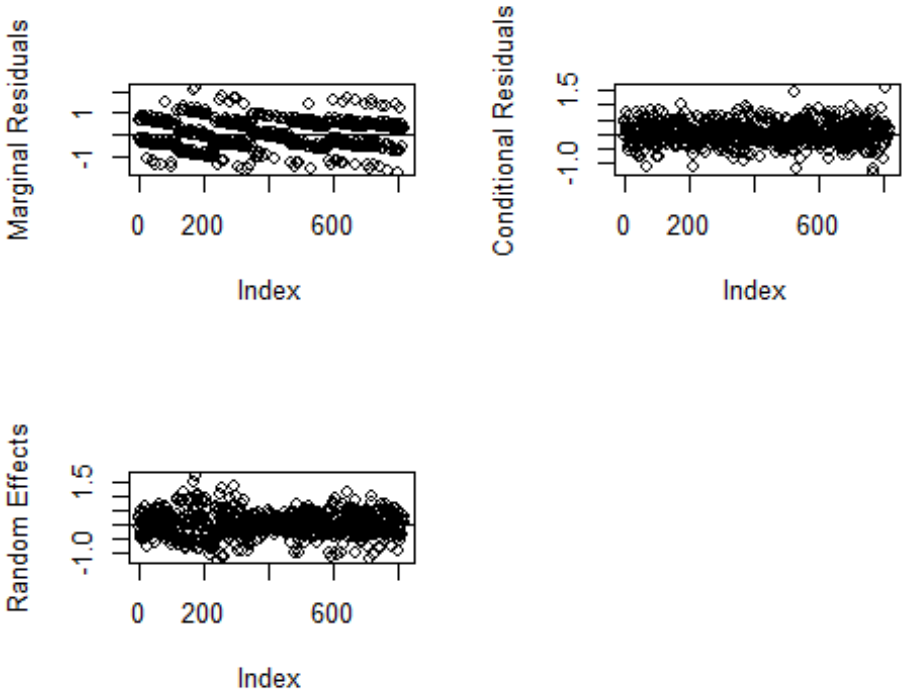
Appendix C

Initial model only had artifact as random effect. Below code is final model after manual forward selection.

```
# fm4 <- lmer(Rating ~ Rater + Semester + Sex + Repeated + (Rubric|Artifact),
tall)
fm5 <- update(fm2, .~. + Rubric)

## boundary (singular) fit: see ?isSingular

ss <- getME(fm5,c("theta","fixef"))
m4u<- update(fm5,start=ss, control=lmerControl(optimizer="bobyqa", optCtrl=li
st(maxfun=2e5)))

## boundary (singular) fit: see ?isSingular

fm5 <- m4u
summary(fm5)
mcp.fnc(fm5)

anova(fm2,fm5) ## anova, aic, bic chose fm5
```

after manual forward selection, it seems rater, semester, and rubric as fixed effects improved initial model.

```
par(mfrow=c(2,2))
plot(r.marg(fm5),xlab="Index",ylab="Marginal Residuals")
abline(0,0)
plot(r.cond(fm5),xlab="Index",ylab="Conditional Residuals")
abline(0,0)
plot(r.reff(fm5),xlab="Index",ylab="Random Effects")
abline(0,0)
```



the residuals looks pretty good for conditional residuals: uniform and looks homoskedastic. marginal residuals looks like have mean 0. random effects ard harder to interpret (look like mean zero for some reason).

Instead, now use automatic variable selection. Final model is produced below.

```
## automatic variable selection for fixed effects and random effects
fm6 <- lmer(Rating ~ Rubric + Sex + Repeated + Semester + Rater + (0+Rubric|A
rtifact), data = tall)
# summary(fm6)
fm7 <- fitLMER.fnc(fm6, ran.effects = c("(Rater|Artifact)", "(Semester|Artifa
ct)"))

## ===========================================================
## ===              backfitting fixed effects            ===
## ===========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.6532 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
##   iteration 2
##     p-value for term "Repeated" = 0.5368 >= 0.05
##     not part of higher-order interaction
```

```
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
## evaluating addition of (Rater|Artifact) to model

## boundary (singular) fit: see ?isSingular

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

##  log-likelihood ratio test p-value = 0.0004713454
##  adding (Rater|Artifact) to model
## evaluating addition of (Semester|Artifact) to model

## boundary (singular) fit: see ?isSingular
## refitting model(s) with ML (instead of REML)

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

##  log-likelihood ratio test p-value = 0.9880335
##  not adding (Semester|Artifact) to model
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================

## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.0587 >= 0.05
##     not part of higher-order interaction
```

final model chosen automatically by fitlmer is Rating = Rater + Rubric + (0+Rubric+Rater|Artifact).

now add interaction between only fixed effects left (rater and rubric) and compare to model without interaction term.

```
fm8 <- lmer(Rating ~ Rater + Rubric + Rater*Rubric + (0+Rubric+Rater|Artifact
), data = tall)

## boundary (singular) fit: see ?isSingular

anova(fm7,fm8) ## interaction model does better

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

## Data: tall
## Models:
## fm7: Rating ~ Rubric + Rater + (0 + Rubric | Artifact) + (Rater | Artifact
)
## fm8: Rating ~ Rater + Rubric + Rater * Rubric + (0 + Rubric + Rater | Arti
```

```
fact)
##     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fm7   40 1467.9 1656.3 -693.97   1387.9
## fm8   51 1462.5 1702.6 -680.23   1360.5 27.476 11   0.003892 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

final final model is Rating = Rater + Rubric + Rater*Rubric + (0+Rubric+Rater|Artifact). Model with interaction term performs better, via anova test.

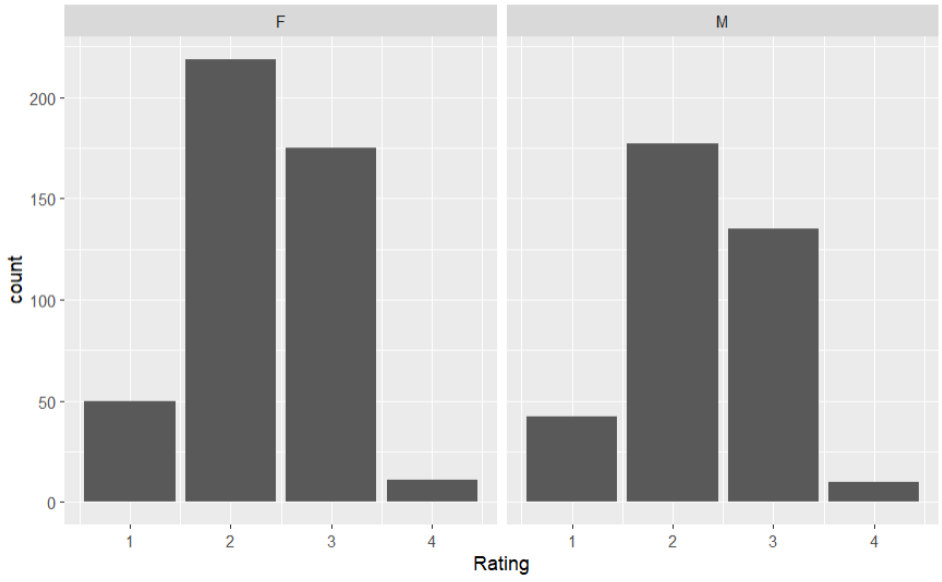
the factors that are correlated with ratings are rater and rubric, as fixed effects, and rubric and rater as random effects. rubric and rater interact in an interesting way, which makes sense because raters give different ratings for the rubric items.

Appendix D

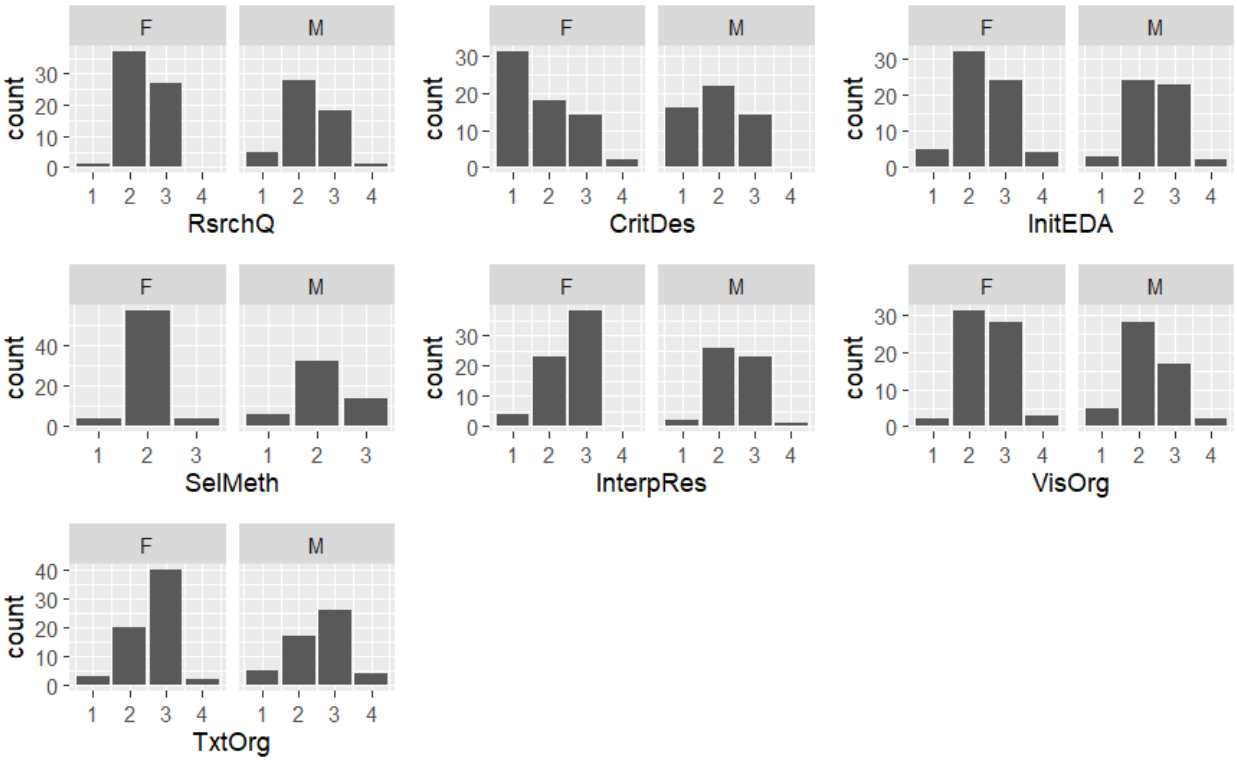
Look at barplot for ratings by sex (female and male) and by semester (f19 and s19). Also look at numeric summaries for ratings for each sex and each semester.

```
ggplot(tall,aes(x = Sex)) + facet_wrap( ~ Rubric) + geom_bar()

ggplot(tall,aes(x = Rating)) + facet_wrap( ~ Rubric) + geom_bar()

by_sex <- describeBy(ratings[,c('RsrchQ', 'CritDes','InitEDA','SelMeth','Inte
rpRes','VisOrg','TxtOrg')], group=ratings$Sex, fast=TRUE)
female <- by_sex$F
names(female) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")
male <- by_sex$M
names(rater2) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")

by_sem <- describeBy(ratings[,c('RsrchQ', 'CritDes','InitEDA','SelMeth','Inte
rpRes','VisOrg','TxtOrg')], group=ratings$Semester, fast=TRUE)
fall <- by_sem$Fall
names(fall) <- c("Variance",'Number',"Mean","Standard Deviation","Minimum","M
aximum","Range","Standard Error")
spring <- by_sex$Spring
names(spring) <- c("Variance","Number","Mean","Standard Deviation","Minimum",
"Maximum","Range","Standard Error")

sex.mean <- data.frame(c("RsrchQ","CritDes","InitEDA","SelMeth","InterpRes","
VisOrg","TxtOrg"),female$Mean, male$Mean)
names(sex.mean) <- c("","Female Mean", "Male Mean")

## grid of barplots for each rubric, by sex

a <- ggplot(ratings, aes(x = RsrchQ)) + facet_wrap( ~ Sex) + geom_bar()

b<-ggplot(ratings, aes(x = CritDes)) + facet_wrap( ~ Sex) + geom_bar()

c<-ggplot(ratings, aes(x = InitEDA)) + facet_wrap( ~ Sex) + geom_bar()
```

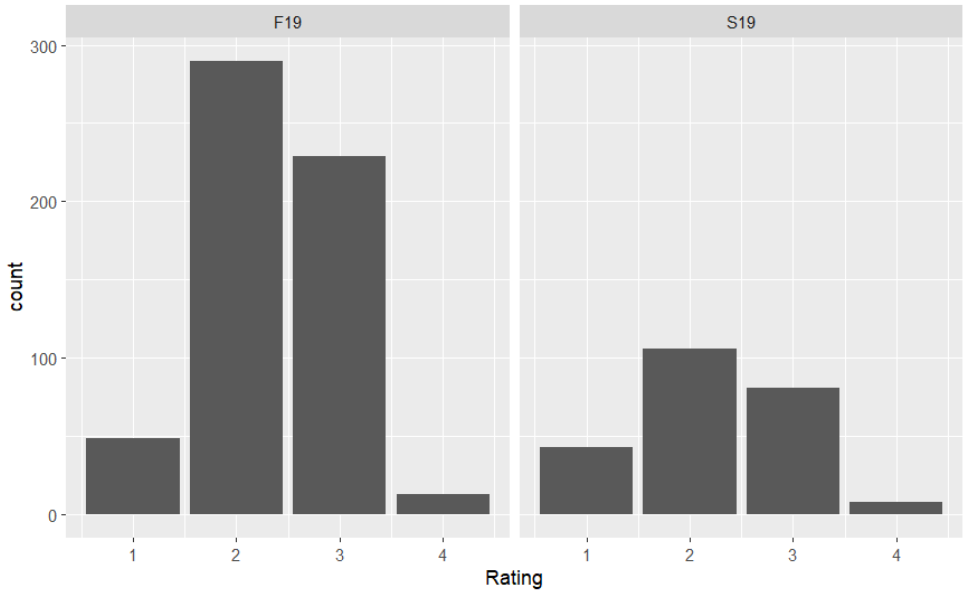
```
d<-ggplot(ratings, aes(x = SelMeth)) + facet_wrap( ~ Sex) + geom_bar()
e<-ggplot(ratings, aes(x = InterpRes)) + facet_wrap( ~ Sex) + geom_bar()
f<-ggplot(ratings, aes(x = VisOrg)) + facet_wrap( ~ Sex) + geom_bar()
g<-ggplot(ratings, aes(x = TxtOrg)) + facet_wrap( ~ Sex) + geom_bar()
plot_grid(a,b,c,d,e,f,g)
## grid of barplots for each rubric, by semester
a <- ggplot(ratings, aes(x = RsrchQ)) + facet_wrap( ~ Semester) + geom_bar(
b<-ggplot(ratings, aes(x = CritDes)) + facet_wrap( ~ Semester) + geom_bar()
c<-ggplot(ratings, aes(x = InitEDA)) + facet_wrap( ~ Semester) + geom_bar()
d<-ggplot(ratings, aes(x = SelMeth)) + facet_wrap( ~ Semester) + geom_bar()
e<-ggplot(ratings, aes(x = InterpRes)) + facet_wrap( ~ Semester) + geom_bar()
f<-ggplot(ratings, aes(x = VisOrg)) + facet_wrap( ~ Semester) + geom_bar()
g<-ggplot(ratings, aes(x = TxtOrg)) + facet_wrap( ~ Semester) + geom_bar()
plot_grid(a,b,c,d,e,f,g)
```



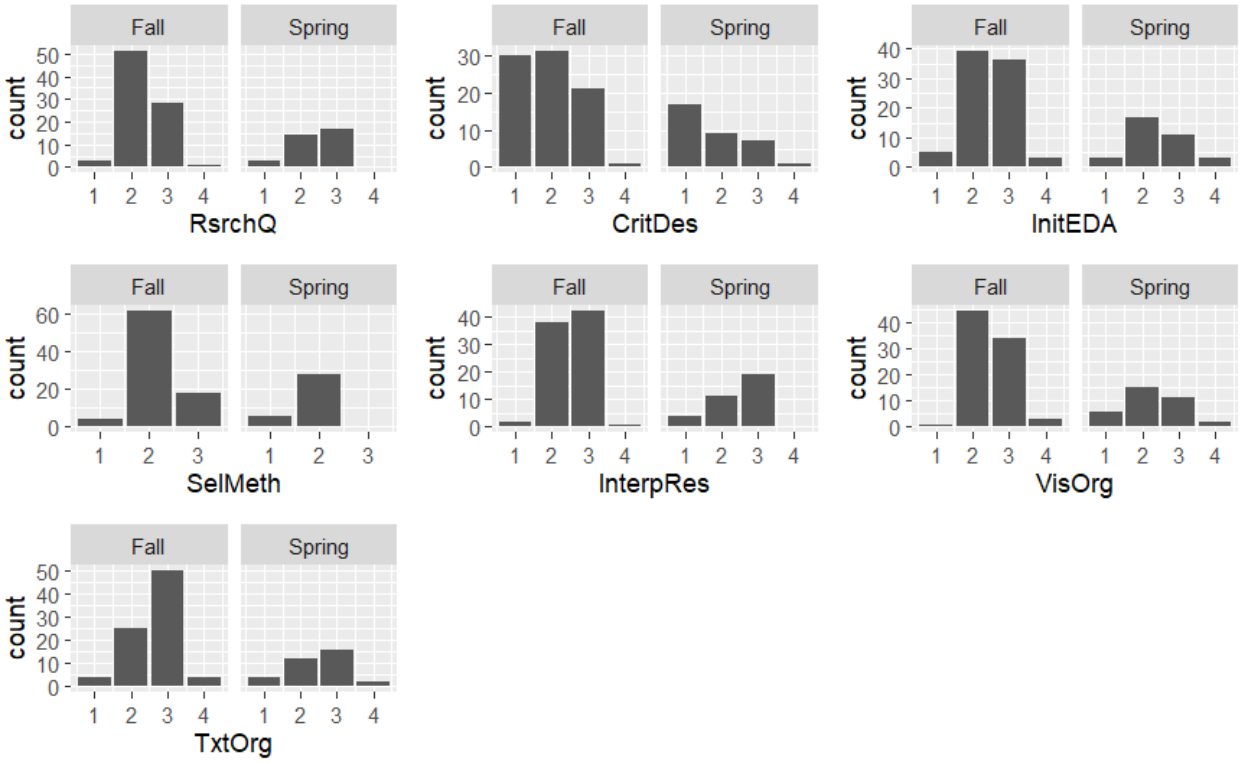
Bar plot for female and male.



Bar plot of ratings for each rubric, shown for each sex of artifact.



Bar plot by semester, fall 19 and spring 19.



Bar plot of each rubric, shown for each semester.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.40	0.52	1	3	2
CritDes	1.80	0.89	1	4	3
InitEDA	2.42	0.73	1	4	3
SelMeth	2.00	0.35	1	3	2
InterpRes	2.52	0.62	1	3	2
VisOrg	2.50	0.64	1	4	3
TxtOrg	2.63	0.63	1	4	3

Numerical summary of each rubric for female.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.29	0.67	1	4	3
CritDes	1.96	0.77	1	3	2
InitEDA	2.46	0.67	1	4	3
SelMeth	2.15	0.61	1	3	2
InterpRes	2.44	0.61	1	4	3
VisOrg	2.31	0.70	1	4	3
TxtOrg	2.56	0.78	1	4	3

Numerical summary of each rubric for male.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.33	0.57	1	4	3
CritDes	1.92	0.81	1	4	3
InitEDA	2.45	0.67	1	4	3
SelMeth	2.17	0.49	1	3	2
InterpRes	2.51	0.57	1	4	3
VisOrg	2.48	0.59	1	4	3
TxtOrg	2.65	0.65	1	4	3

Numerical summary for each rubric for fall semester.

	Mean	Standard Deviation	Minimum	Maximum	Range
RsrchQ	2.40	0.52	1	3	2
CritDes	1.80	0.89	1	4	3
InitEDA	2.42	0.73	1	4	3
SelMeth	2.00	0.35	1	3	2
InterpRes	2.52	0.62	1	3	2
VisOrg	2.50	0.64	1	4	3
TxtOrg	2.63	0.63	1	4	3

Numerical summary for each rubric for spring semester.

	Female Mean	Male Mean
RsrchQ	2.40	2.29
CritDes	1.80	1.96
InitEDA	2.42	2.46
SelMeth	2.00	2.15
InterpRes	2.52	2.44
VisOrg	2.50	2.31
TxtOrg	2.63	2.56

Means for each rubric, female vs male.

it's interesting to say that sex doesn't seem to affect the ratings, since usually gender is usually an apparent factor that leads to differences. i think it would also be interesting to conduct further analysis on whether the semester that this stat class was taken makes a difference in the grades are distributed. different professors have different guidelines and grading scales that could also lead to differences in the rating distributions.

Looking at the bar plots by sex, both female and male ratings have similar distribution shapes. Both are right skewed, meaning that typically, everyone is being rated on the lower

end. However, despite the number of male and females not being drastically different (65 female, and 52 male), females scored significantly more 2s and 3s combined than males did. When looking at the numerical summaries for each rubric by sex, we can see that for 5 out of 7 rubrics, males scored higher on average. The 2 rubrics that females scored higher than males did on average are visorg and txtorg. However, when looking at the grid of bar plots for each rubric, we can see that females are scored lower more often for critdes, initeda, and selmeth.

Looking the bar plots by semester, both fall and spring semester ratings are right skewed. However, the number of artifacts chosen from fall semester is about 2.5 times the number chosen from spring semester. This accounts for the discrepancy in number of ratings between fall and spring semester. It is hard to determine whether there is truly a difference between the rating distributions between fall and spring semester due to the huge difference in artifacts from each semester. when looking at the grid of bar plots for each rubric, we can see that most of the distributions for each rubric looks somewhat similar for fall and spring.