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

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 variables related to Rating. 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 the implication of Sex and Semester on 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 7 rubric items for the sample of 91 project papers for the freshman statistics course. 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 3 raters, while the remaining 78 were rated only by one rater. We were provided 2 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 one rating, shown in the column Rating and the rubric for that rating in the column Rubric. *Table 2* (page 3) shows the 7 rubric items that the 3 raters rated the artifacts on, while *Table 3* (page 3) shows the rating scale for the rubric items.

Numeric summaries for each rubric is shown in Table 4 *(need to insert numerical summaries tables).*

After initial EDA, we can see the distributions of the ratings in *Figure 1* (page 3). Looking at *Figure 1* (page 3), `CritDes`, `InitEDA`, and `VisOrg` seem to be right skewed, while `TxtOrg` and `InterpRes` are left skewed. The other variables' distributions are harder to tell, simply

because some rubrics do not take on every rating from 1 to 4. Lastly, we see the distributions of ratings by each rater in *Figure 2* (page 4). Upon initial investigation, it seems like rater 3 on the far right in *Figure 2* (page 4) 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.

Variable Name	**Description**
X*	Row number in the data set
Rater	Which of the 3 raters gave a rating
Sample*	Sample number
Overlap*	Unique identifier for artifact seen by all 3 raters
Semester	Fall or Spring, 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	0 or 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: Rubric items for 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 3: Rating scale for each rubric item.

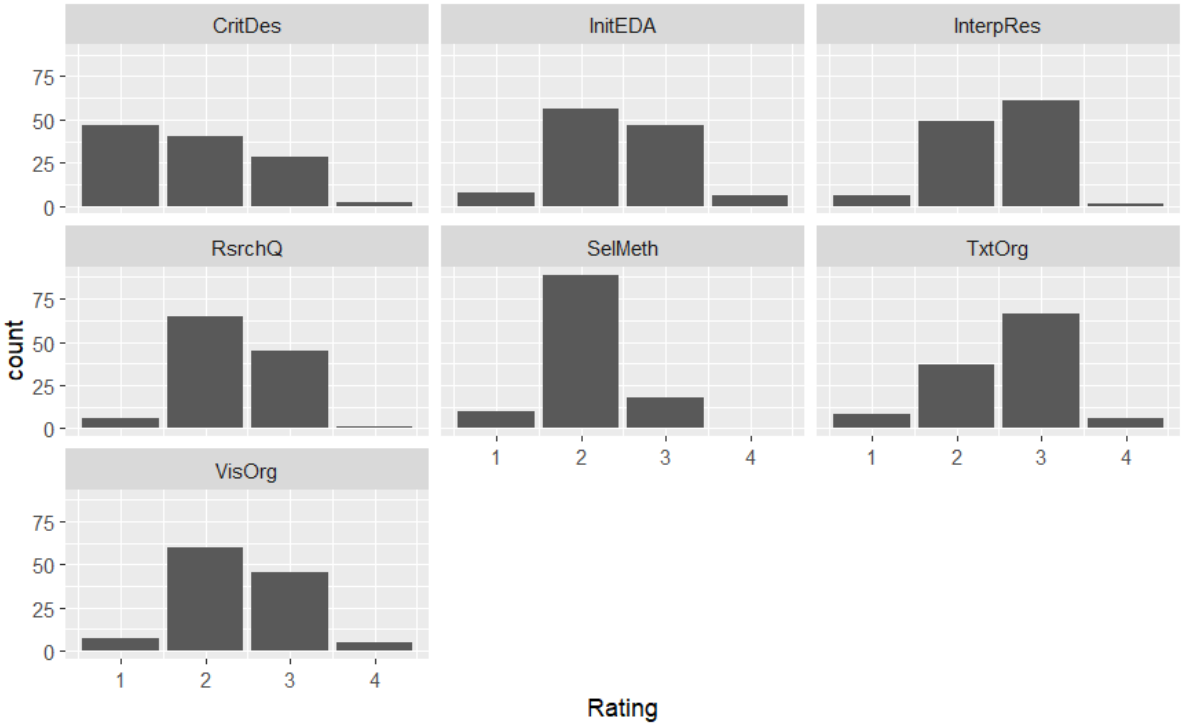


Figure 1: Bar plot of each rubric item for the entire dataset `tall.csv`.

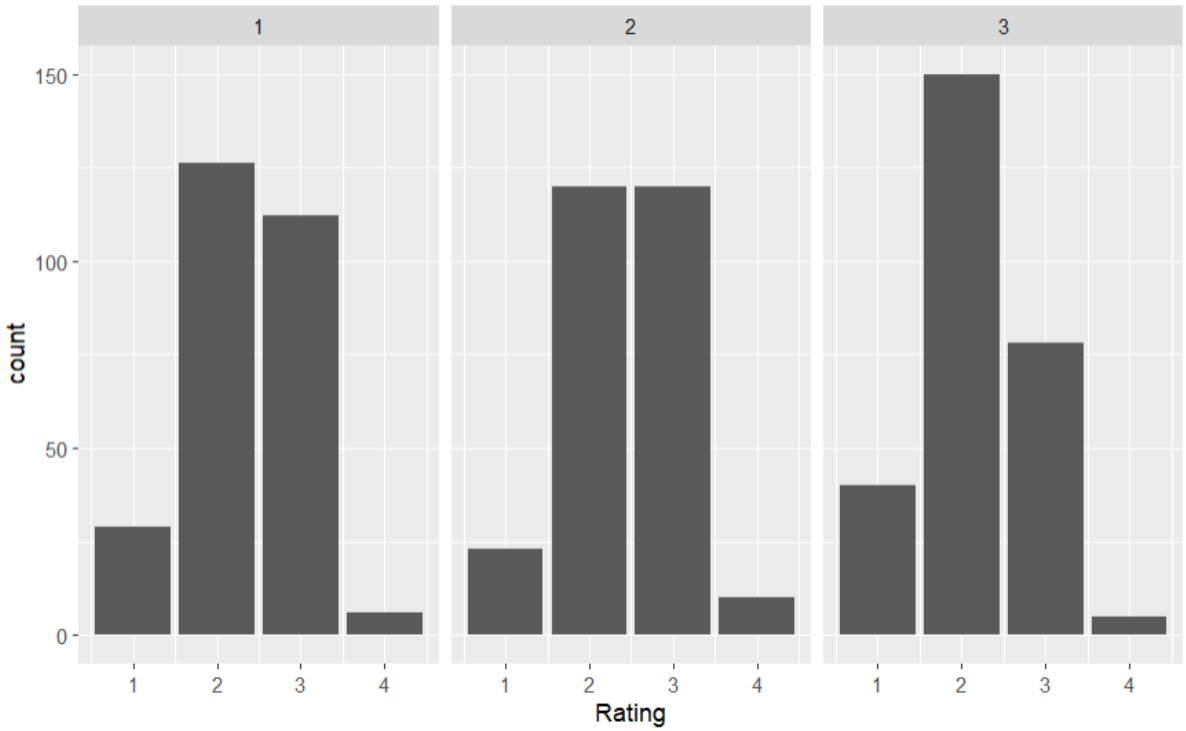


Figure 2: Bar plot of each rater's ratings.

4. METHODS

For the first research question, we look at numerical summaries, histograms, and bar plots for to determine the distribution of ratings for each rubric and to determine the distribution of ratings given by each rater. Additionally, we looked at the distribution of ratings mentioned above for the 13 artifacts that were rated by all 3 raters compared to the distribution of ratings mentioned above for the remaining 91 artifacts that were only rated by 1 rater.

For the second research question, we initially focus only on the 13 artifacts that were rated by all 3 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 7 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 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 5 variables. Lastly, we explore interactions

between the 5 variables and add the meaningful interactions to the model. To determine whether the model with interactions performs better than the model without interactions, we do model selection using an ANOVA test. The final mixed effects model is created using the `fitLMER.fnc()` function in R, which automatically does backward selection on fixed effects, forward selection in random effects, and then backward selection again on fixed effects. We take the final model that fitLMER.fnc() produces to add in interaction terms between the variables.

For the fourth research question, we do further exploratory data analysis to see what insights are surprising and may need further investigation. *(need to add more on methods for this question)*

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 9-11 in Technical Appendix). Looking at the distributions of the scores for each of the 7 rubrics in *Figure 3* (page 6), it seems like `CritDes` scored lowest (right skewed), while `RsrchQ`, `InitEDA`, and `VisOrg` scored lower (right skewed). `SelMeth` seemed to be scored very fairly (nearly uniform distribution). Lastly, `TxtOrg` scored slightly better than all 7 rubrics, 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.

Secondly, to determine whether there is a difference between each rater's ratings we look at numerical summaries, histograms, and bar plots for each rater's ratings (pages 11-18 in Technical Appendix). When we look at the distributions of each rater's ratings for each rubric (*Figures 4*, *5*, and *6* on pages 6-7), it looks like rater 3 is a bit 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, as shown earlier in initial EDA (page 4). 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. Therefore, 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 low 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 4* (page 8), we see the ICC values for each rubric for the 13 artifacts seen by all 3 raters (page 23 in Technical Appendix) . Comparing these values, `CritDes`, `InitEDA`, `SelMeth`, and `VisOrg` have the highest ICCS, meaning that the 3 raters agreed the most on these 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.

In order to combat this issue of broad insight, we look at contingency tables between pairs of raters to determine the percentage of agreement for each rubric (pages 24-34 in Technical Appendix) . In *Table 5* (page 8), we see the agreement rates between each pair of raters for each rubric item. *(need to insert table that has all agreement rates)* 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 6 (page 8) shows a comparison of the ICC values for the full dataset and the 13 artifacts seen by all 3 raters. For the ICCs for the full dataset, `CritDes`, `InitEDA`, `VisOrg`, and `TxtOrg` have the highest ICCs. This means the raters agree the most for these 4 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 5 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:

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

An interpretation of the final model (Model 1.1) is as follows: *(need to put interpretation for final model)*

- In the US, for every 1 unit of per capita income increase, there is a ∼1% increase in crime. This increase is statistically significant.

We look at the model diagnostics plots for Model 1.1. *(need to add residuals)*

Our fourth question asks whether there are any other interesting insights that should be mentioned to the Dean. After looking at the initial EDA and Model 1.1, it shows that `Sex` and `Semester` have no significant effect as a fixed or random effect on `Rating`. It's interesting that `Sex` doesn't seem to affect `Rating`, since usually gender is usually an apparent factor

that leads to differences. It may also be interesting to conduct further analysis on whether the `Semester` that this Freshman Statistics class was taken makes a difference in the way that the grades are distributed. Different professors have different guidelines and grading scales that could lead to differences in the rating distributions.

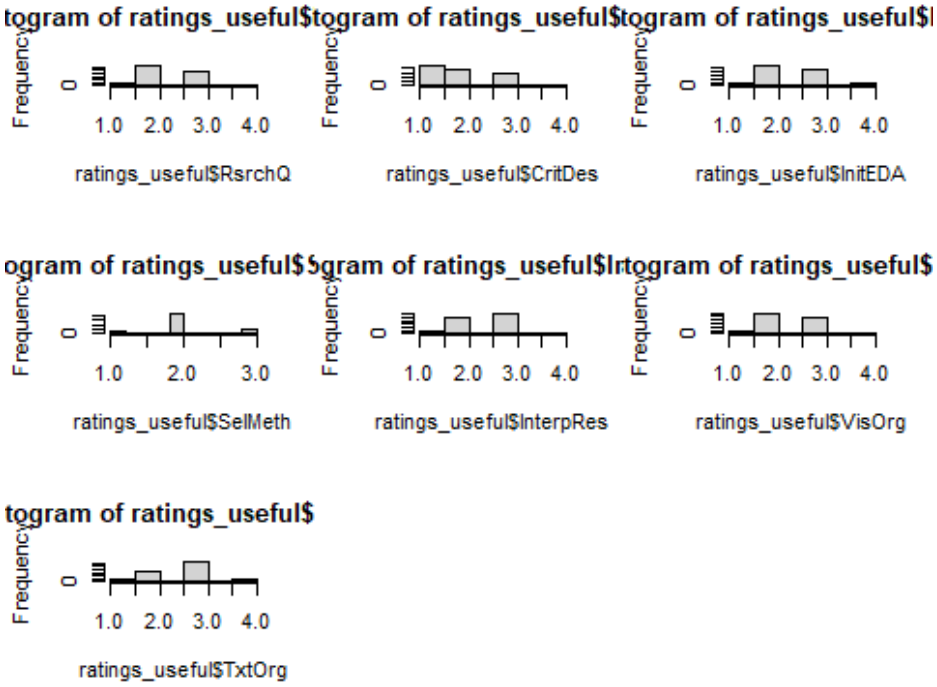


Figure 3: Histograms for each rubric's ratings.

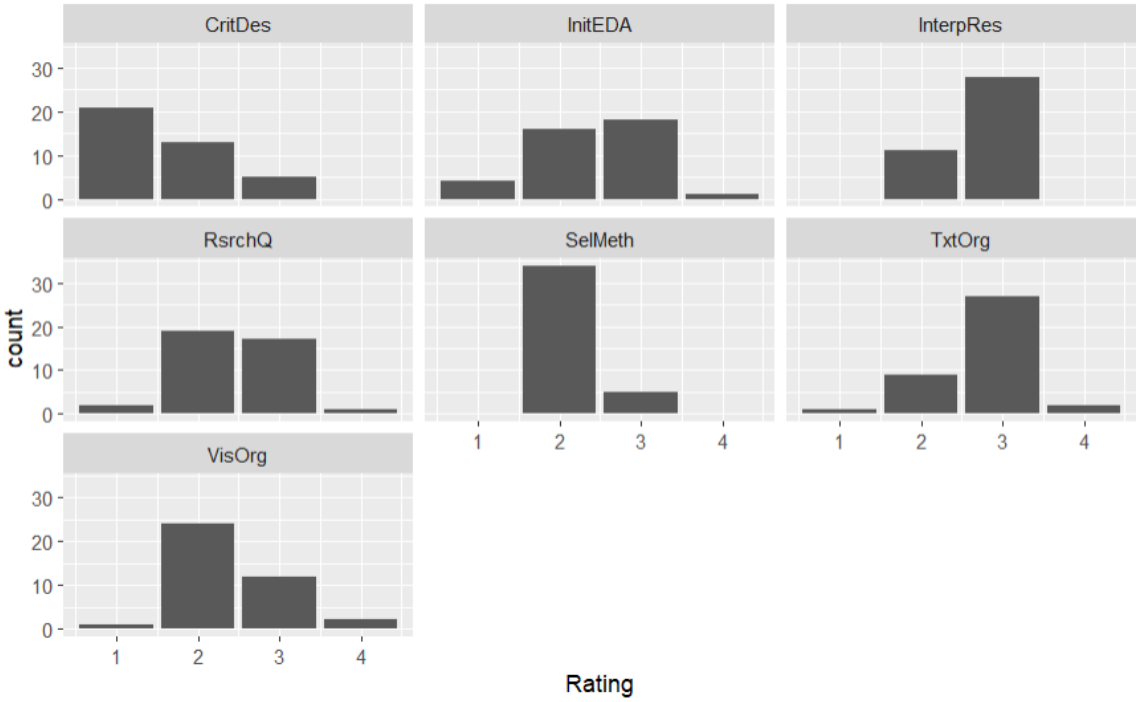


Figure 4: Bar plots of rater 1's ratings for each rubric.

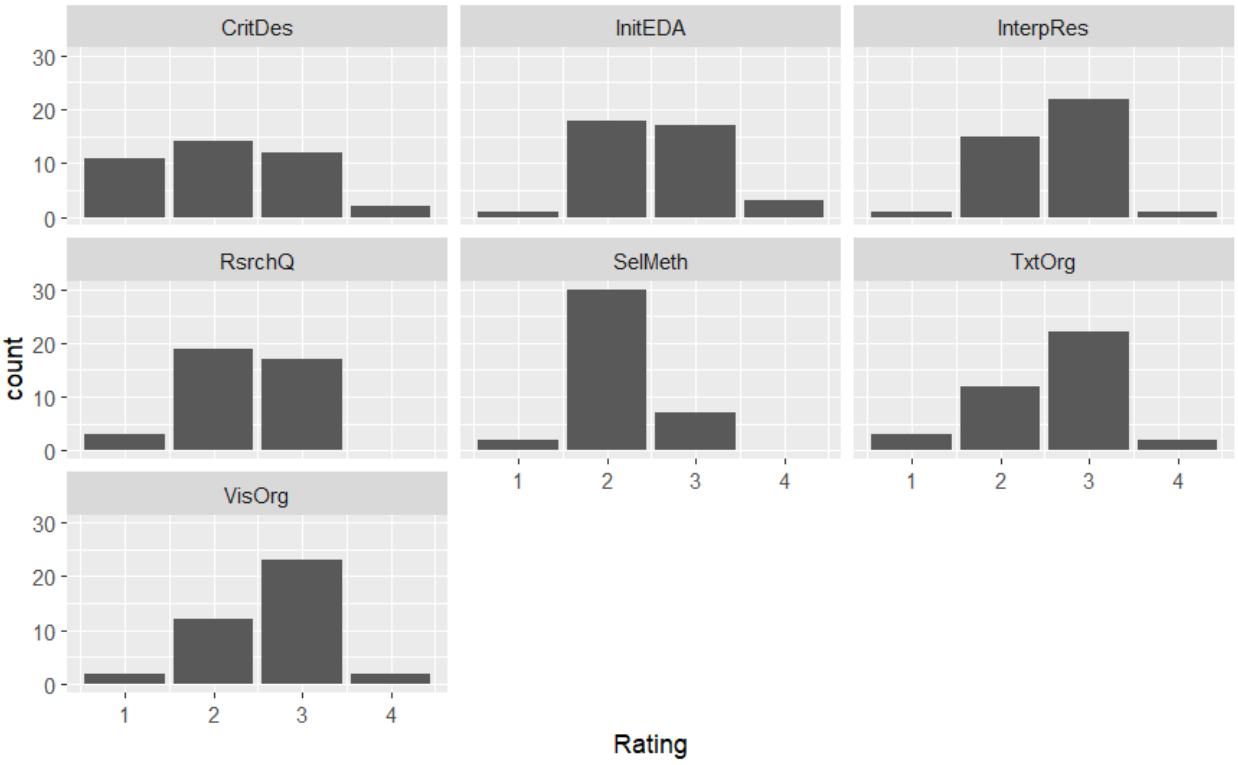


Figure 5: Bar plots of rater 2's ratings for each rubric.

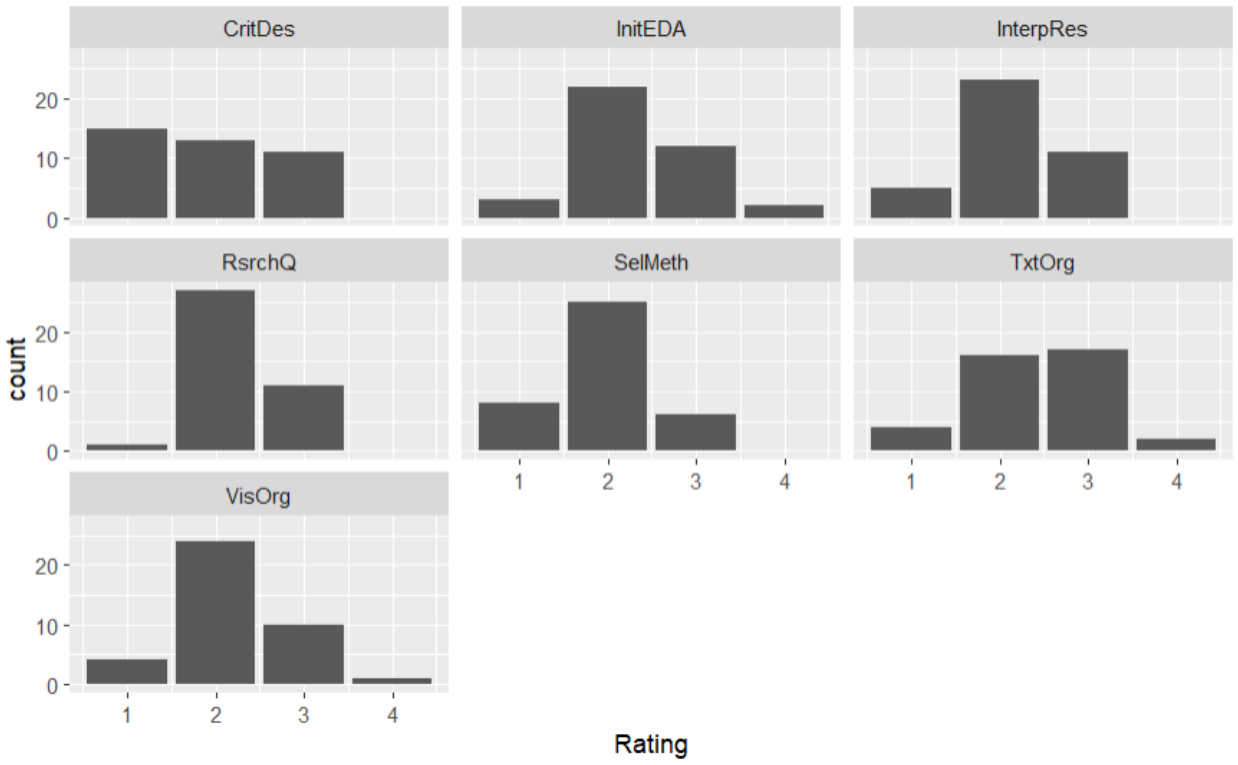


Figure 6: Bar plots of rater 3's ratings for each rubric.

Rubric	ICC
RsrchQ	0.1891892
CritDes	0.5725594
InitEDA	0.4929577
SelMeth	0.5212766
InterpRes	0.2295720
VisOrg	0.5924529
TxtOrg	0.1428571

Table 4: Intraclass correlations for each rubric for 13 artifacts seen by all 3 raters.

Rubric	ICC for Full Dataset	ICC for Subset
RsrchQ	0.2096214	0.1891892
CritDes	0.6699202	0.5725594
InitEDA	0.6867210	0.4929577
SelMeth	0.4719014	0.5212766
InterpRes	0.2200285	0.2295720
VisOrg	0.6586320	0.5924529
TxtOrg	0.6699202	0.1428571

Table 6: Intraclass correlations for each rubric for 13 artifacts vs full dataset.

6. DISCUSSION

As a reminder, our analysis and modeling all aimed to determine the success and fairness of Dietrich College's new "General Education" undergraduate program. 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`.

Lastly, for the fourth question, we looked at the results from question 3 and from initial 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.

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 3, 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. We chose to fill in the missing values with mode of `Rating` and `Sex`, which were 2 and Female respectively. It is possible that the way in which we handled missing data could 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. It might be interesting to look additionally at different courses in the "General Education" program, 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

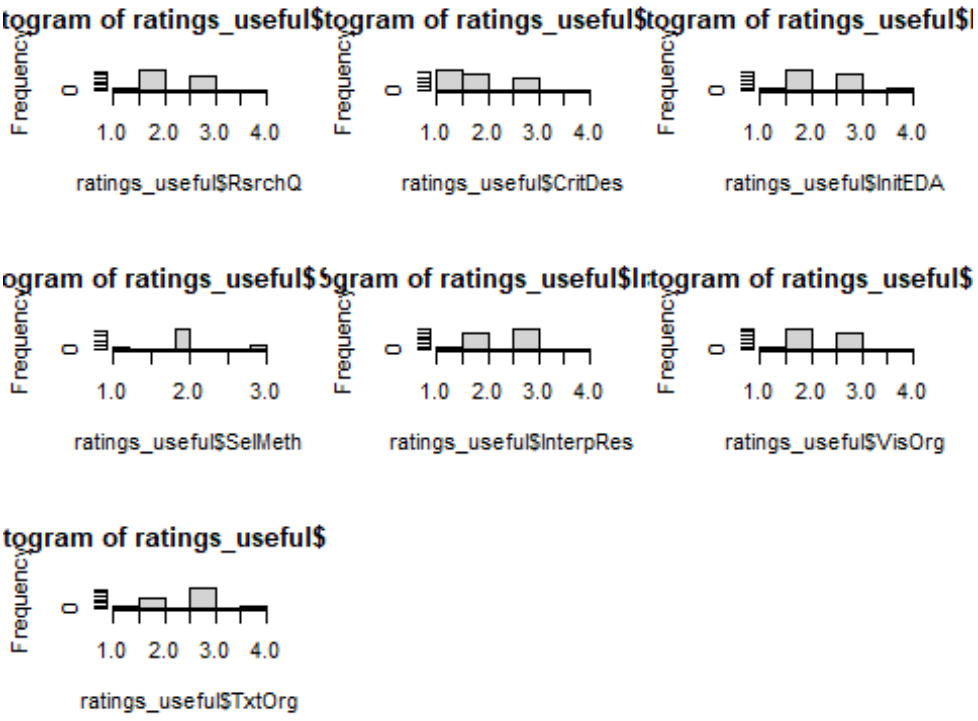
Sheather, S. J. (2009), "A Modern Approach to Regression with R," *Springer eBooks.*

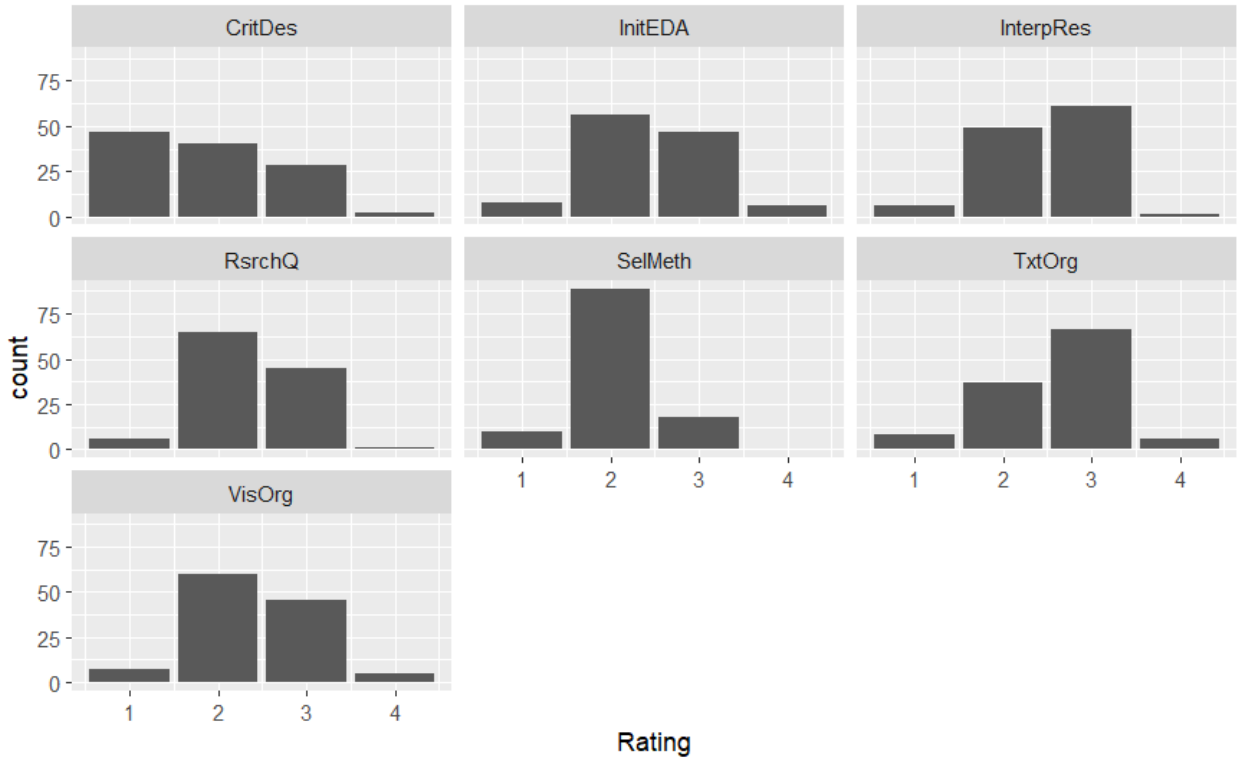
8. TECHNICAL APPENDIX

question 1

```
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)

## 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()
```

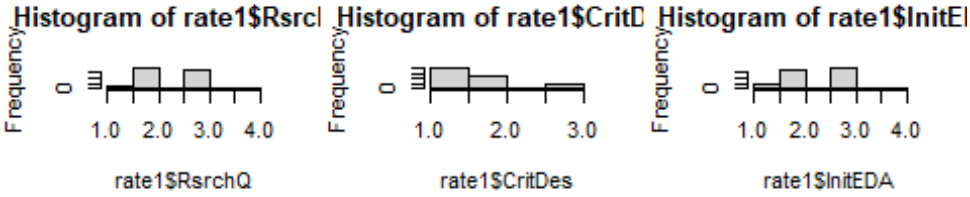


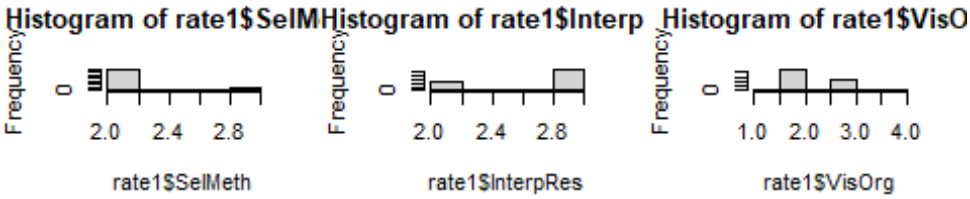


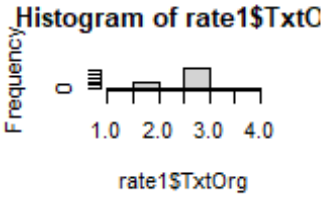
looking at the distributions of the scores for each of the 7 rubrics, it looks like critique design scored lowest (extremely right skewed), while research question, initial eda, and visual organization scored lower (right skewed). selection method seemed to be scored very fairly (almost uniform distribution).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

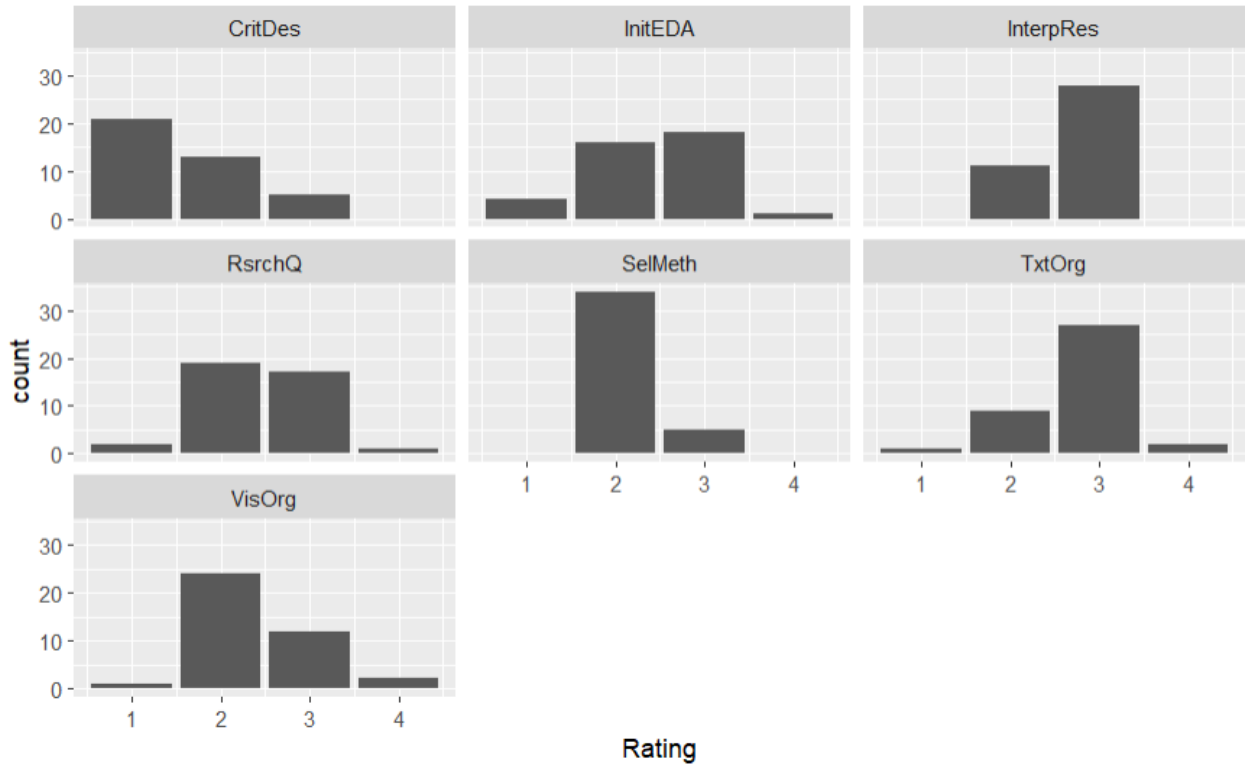
```
## subset data for each rater
rate1 <- ratings_useful %>%
  filter(Rater == 1)
rate1.tall <- tall %>% filter(Rater == 1)

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()
```



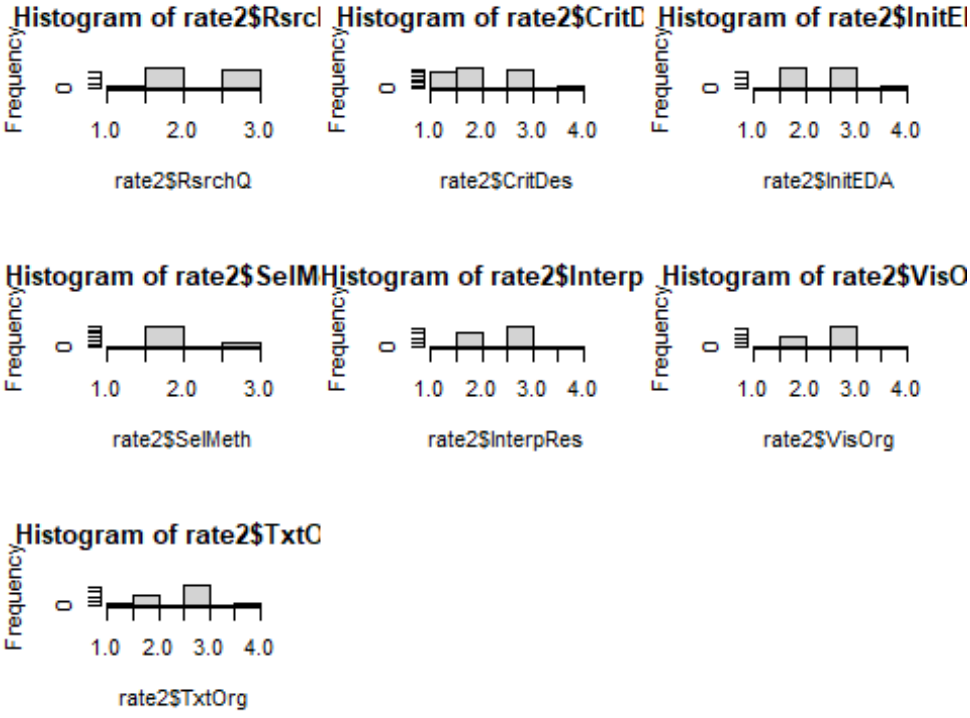


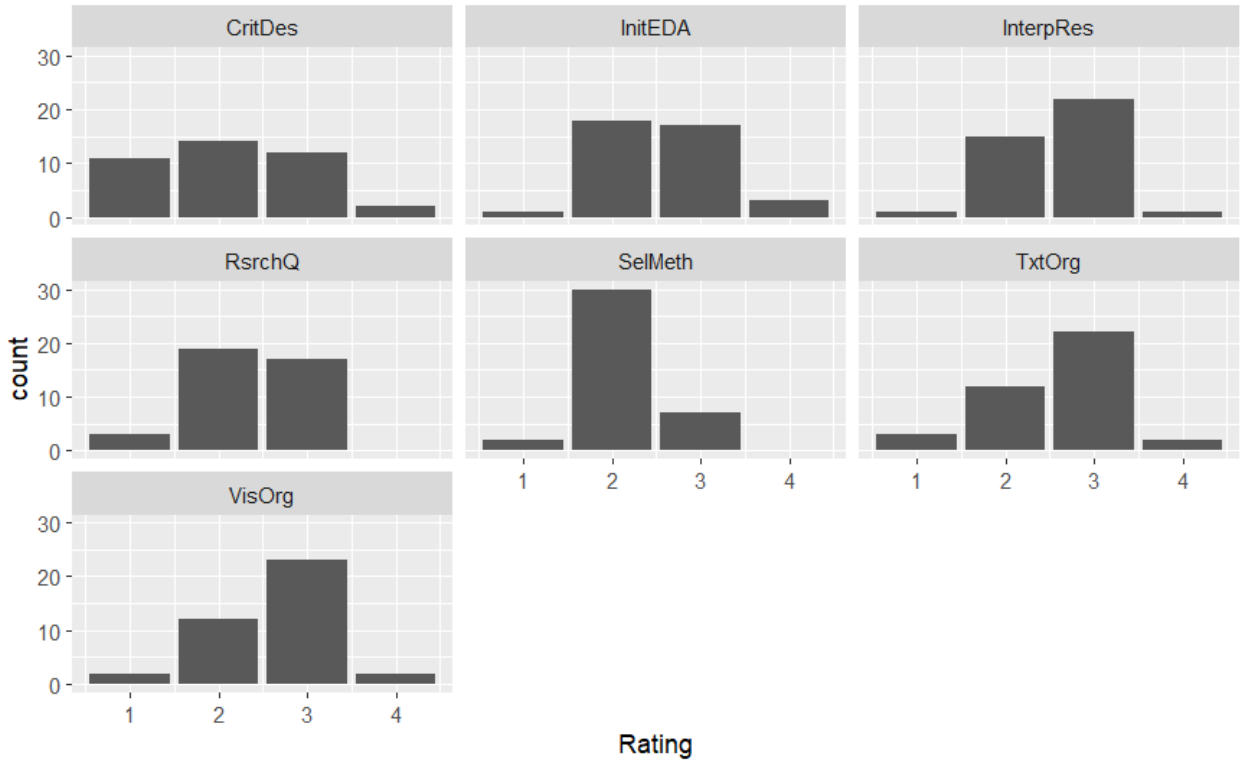





```
## rater 2 distributions
par(mfrow=c(3,3))
```

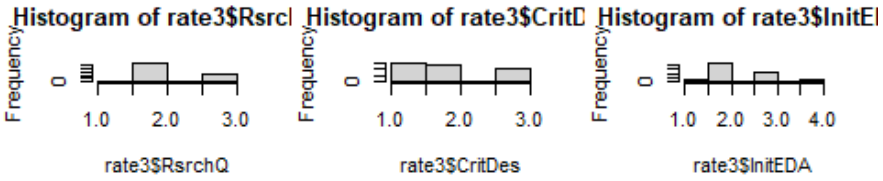
```
hist(rate2$RsrchQ)
summary(rate2$RsrchQ)

##    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()
```

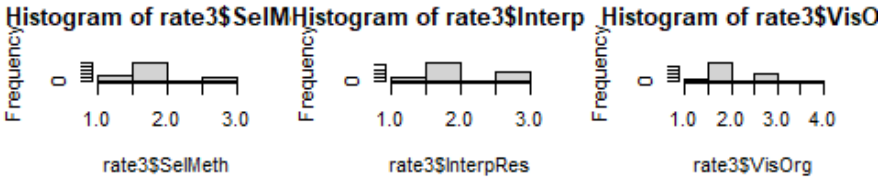


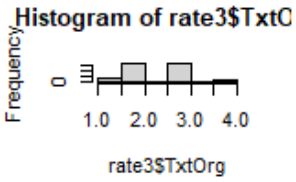


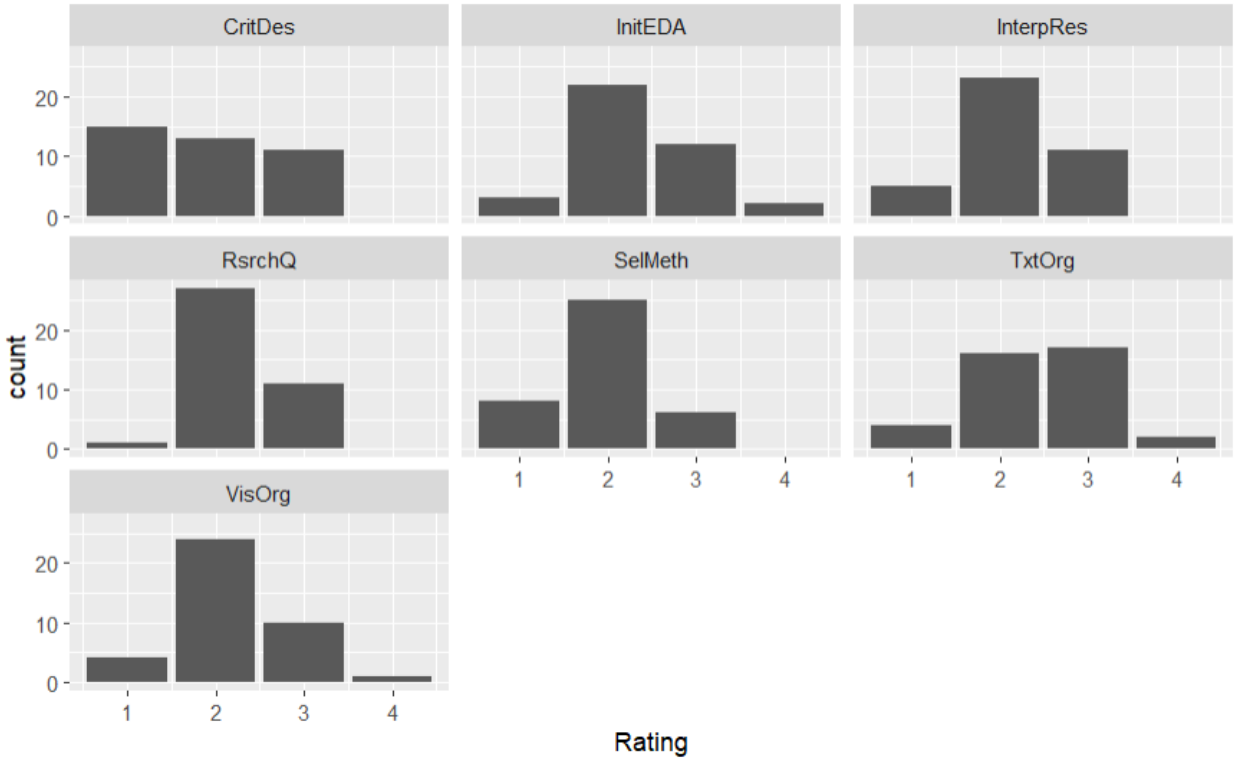

```
## rater 3 distributions
par(mfrow=c(3,3))
```

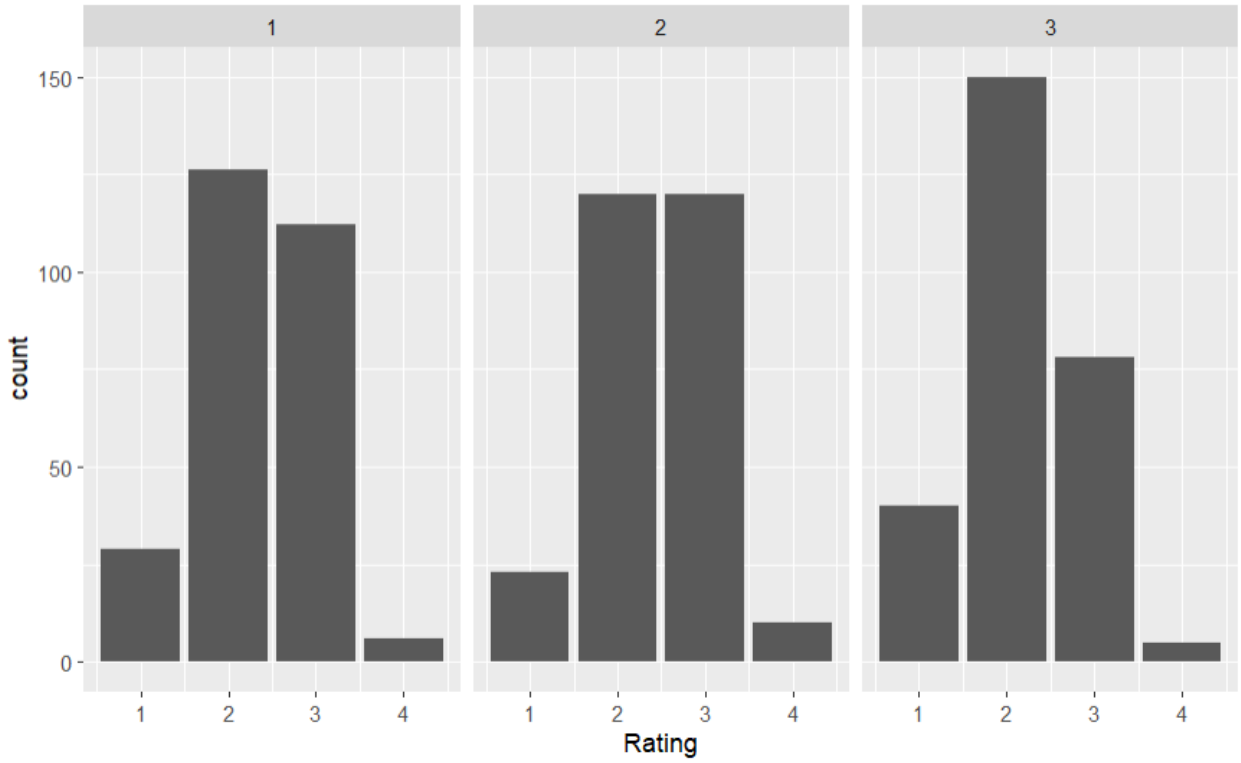
```
hist(rate3$RsrchQ)
summary(rate3$RsrchQ)

##    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$Rsrc
Frequency
0
1.0
2.0
3.0
rate3$RsrchQ
Histogram of rate3$CritD
Frequency
0
1.0
2.0
3.0
rate3$CritDes
Histogram of rate3$InitEl
Frequency
0
1.0
2.0
3.0
4.0
rate3$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

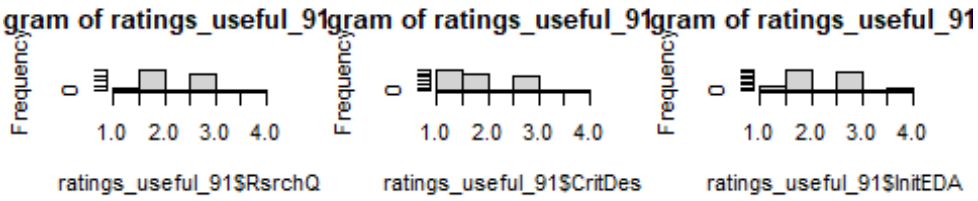

CritDes
InitEDA
InterpRes
RsrchQ
SelMeth
TxtOrg
VisOrg
count
0
10
20
1
2
3
4
Rating

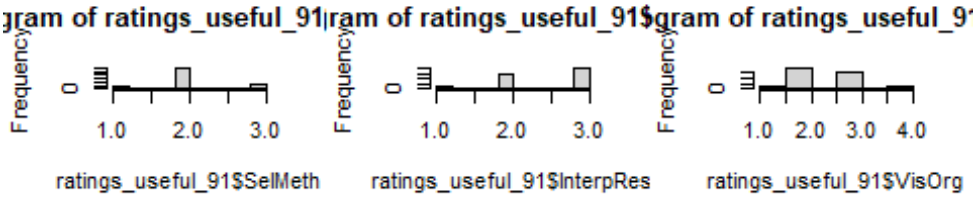


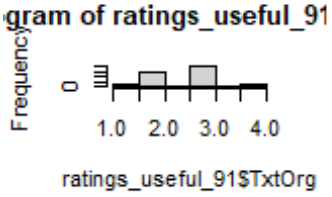
it looks like rater 3 is a bit harsher than the other 2 raters. most of the distributions for the rubrics for rater 3 are closer to right skewed. rater 1 is the only rater that sometimes gives binary ratings, meaning only rating 2 values, as opposed to 3 or 4 ratings. this is confirmed by the bar plot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

```
## 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
```

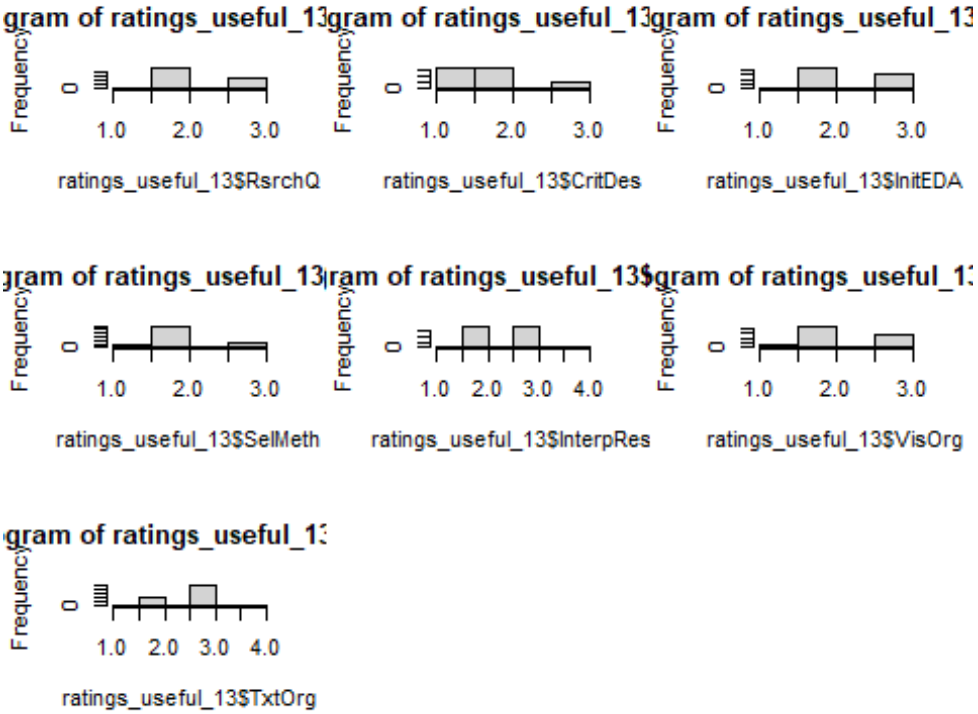







```
## distributions for subset of 13 artifacts rated by all 3 raters
par(mfrow=c(3,3))
hist(ratings_useful_13$RsrchQ)
summary(ratings_useful_13$RsrchQ)

##    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.

question 2

```
## 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 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

##    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

##    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

##    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

##    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

##    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

##    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

##    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

##    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

##    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

##    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
```

```
##    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
```

```
##    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
```

```
##    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
```

```
##    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
```

```
##    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

##    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

##    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
```

```
##    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
```

```
##    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

##    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

##    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

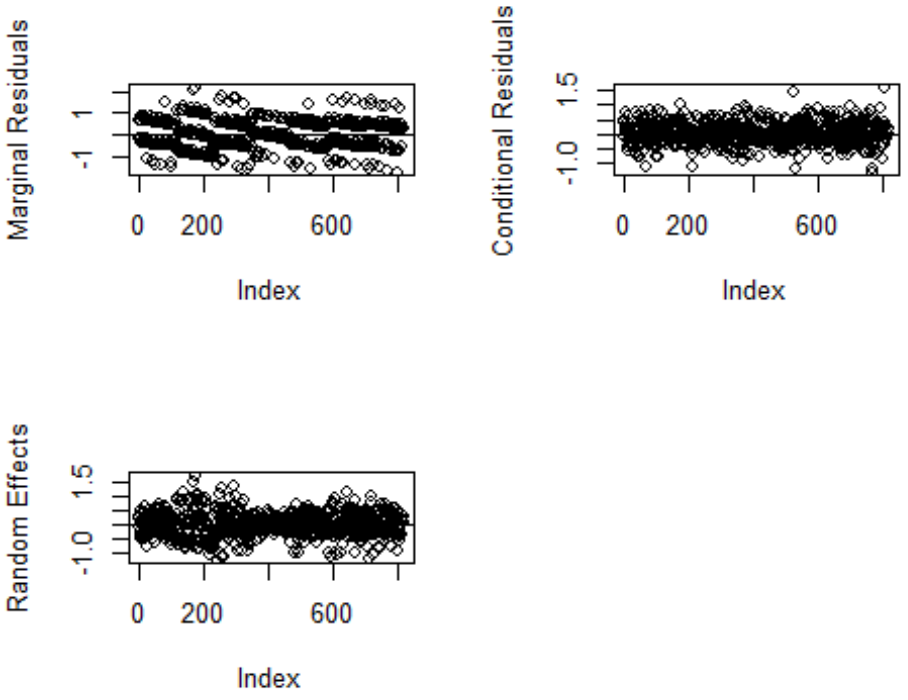
critdes, initeda, visorg, and txtorg have the highest icc's. this means the raters agree the most for these 4 rubrics. when comparing to the subset of 13 artifcats, the icc's are not the same, especially for txtorg - icc is much higher for full dataset. otherwise, the icc's are similar enough.

question 3

```
# 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=list(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).

```
## 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
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkCon
v, :
## Model failed to converge with max|grad| = 0.00214365 (tol = 0.002, compone
nt 1)

##     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)

## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.

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

## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.

##  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 and compare.

```
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).

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.

question 4

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.