A study on the evaluation of project papers produced in a class

Bhoomika Moorjani
bhoomikamoorjani@cmu.edu
Master of Statistical Practice, Carnegie Mellon University

29 November 2021

ABSTRACT
WIP

INTRODUCTION

Dietrich College of Humanities and Social Sciences at Carnegie Mellon University is in the process of implementing a new "General Education"(GenEd) program for undergraduate students. This program specifies a set of mandatory courses and experiences for undergraduate students and in order to determine whether the new program was successful, the college hopes to rate student work performed in each of the GenEd courses each year. This paper focuses on a recent experiment where project papers produced by students in the Freshmen Statistics class were rated by raters from different departments in the college based on a common set of rubrics and we aim to address the following research questions:

1.
    a. 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?
    b. 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 the 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?

DATA

As part of the experiment, 91 project papers - referred to as "artifacts" - were randomly sampled from a Fall and Spring section of the Freshman Statistics class and three raters from different departments were asked to rate these artifacts on seven rubrics, as shown in Table 1, not knowing which class or student produced the artifact they will be rating. The rating scale for all rubrics is shown in Table 2. Thirteen of the 91 artifacts were rated by all three raters (13 x3 = 39 observations) and each of the remaining 78 artifacts were rated by only one rater (78 x1 = 78 observations). Variables available in the dataset are defined in Table 3 and Table 4. The same data is contained in two files ratings.csv (organized so that each row contains one observation and a different column for ratings in each rubric i.e. wide data format) and tall.csv (organized so that each row contains one rating for each rubric per observation i.e. long data format).

Table 1: Rubrics used for rating project papers produced by students in Freshman Statistics class

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 dataset, the student appropriately describes the data and provides initial Exploratory Data Analysis
SelMeth	Method Selection	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: Rating scale used for all rubrics in Table 1

Rating	Criteria
1	Student does not generate any relevant evidence
2	Student generates evidence with significant flaws
3	Student generates competent evidence with no flaws or only minor ones
4	Student generates outstanding evidence which is comprehensive and sophisticated

Table 3: Variables available in the file ratings.csv

Variable Name	Values	Description
X	1,2,3,.....,117	Row number in the dataset
Rater	1,2, or 3	Which of the three raters gave a rating
Sample	1,2,3,.....,118 (14 doesn't exist)	Sample number

Overlap	1,2,3,.....,13	Unique identifier for each artifact seen by all 3 raters
Semester	Fall or Spring	Which semester the artifact came from
Sex	M or F	Sex of the the student who produced the artifact
RsrchQ	1,2,3 or 4	Rating on Research Question
CritDes	1,2,3 or 4	Rating on Critique Design
InitEDA	1,2,3 or 4	Rating on Initial EDA
SelMeth	1,2,3 or 4	Rating on Method Selection
InterpRes	1,2,3 or 4	Rating on Interpret Results
VisOrg	1,2,3 or 4	Rating on Visual Organization
TxtOrg	1,2,3 or 4	Rating on Text Organization
Artifact	Text labels	Unique identifier for each artifact
Repeated	0 or 1	0 = Artifact was only seen by 1 rater 1 = Artifact was seen by all 3 raters

Table 4: Variables available in the file tall.csv

Variable Name	**Values**	**Description**
X	1,2,3,.....,819	Row number in the dataset
Rater	1,2, or 3	Which of the three raters gave a rating
Artifact	Text labels	Unique identifier for each artifact
Repeated	0 or 1	0 = Artifact was only seen by 1 rater 1 = Artifact was seen by all 3 raters
Semester	F19 or S19	Which semester the artifact came from
Sex	M or F	Sex of the the student who produced the artifact
Rubric	RsrchQ, CritDes, InitEDA, SelMeth, InterpRes, VisOrg or TxtOrg	Rubric the rater is giving rating for
Rating	1,2,3 or 4	Rating for corresponding rubric

Table 5: Ratings Summary

Ratings Summary

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.35	3	4	0.59
InitEDA	1	2	2	2.44	3	4	0.70
SelMeth	1	2	2	2.07	2	3	0.49
InterpRes	1	2	3	2.49	3	4	0.61
TxtOrg	1	2	3	2.60	3	4	0.70
CritDes	1	1	2	1.87	3	4	0.84
VisOrg	1	2	2	2.41	3	4	0.67

METHODS

Our analysis, consisting of four parts, was carried out using the R language and environment for statistical computing.

Research Question 1:

a. 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?
b. 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?

We visually compared barplots (Figures 1-4 in Results) to study the distribution of ratings across rubrics and raters for the full dataset and subset of 13 artifacts which were rated by all three raters.

Research Question 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?

We calculated the intraclass correlation to quantify the degree of association between ratings within each rubric group. Additionally, we computed percent exact agreement for each pair of raters as a measure of inter-rater reliability.

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

To account for the fixed and random effects in the experiment, we fit a multilevel regression model. At the first level, the model studied the relationship between individual ratings and the various factors in the

experiment such as rater, semester, sex, repeated and rubric. At the second level, the model studied the relationship between ratings and predictors for each of the 91 artifacts. We leveraged boxplots and barplots to better visualize the results of the model.

Research Question 4: Is there anything else interesting to say about this data?
WIP

RESULTS

Research Question 1:

a. 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?

The distribution for 13 artifacts that were rated by all three raters in Figure 2 is similar to that of the full dataset in Figure 1. This suggests that the sample of 13 artifacts is representative of the population i.e., all 91 artifacts. We see the distributions for *CritDes* and *SelMeth* in Figure 1 are right skewed, indicating that they tend to get low ratings. On the other hand, *TxtOrg* and *InterpRes* are left skewed, indicating that they tend to get high ratings. This is also evident from Table 5 in the data section - *CritDes* and *SelMeth* have a lower mean rating and *TxtOrg* and *InterpRes* have a higher mean rating.

Figure 1: Full dataset, grouped by Rubrics

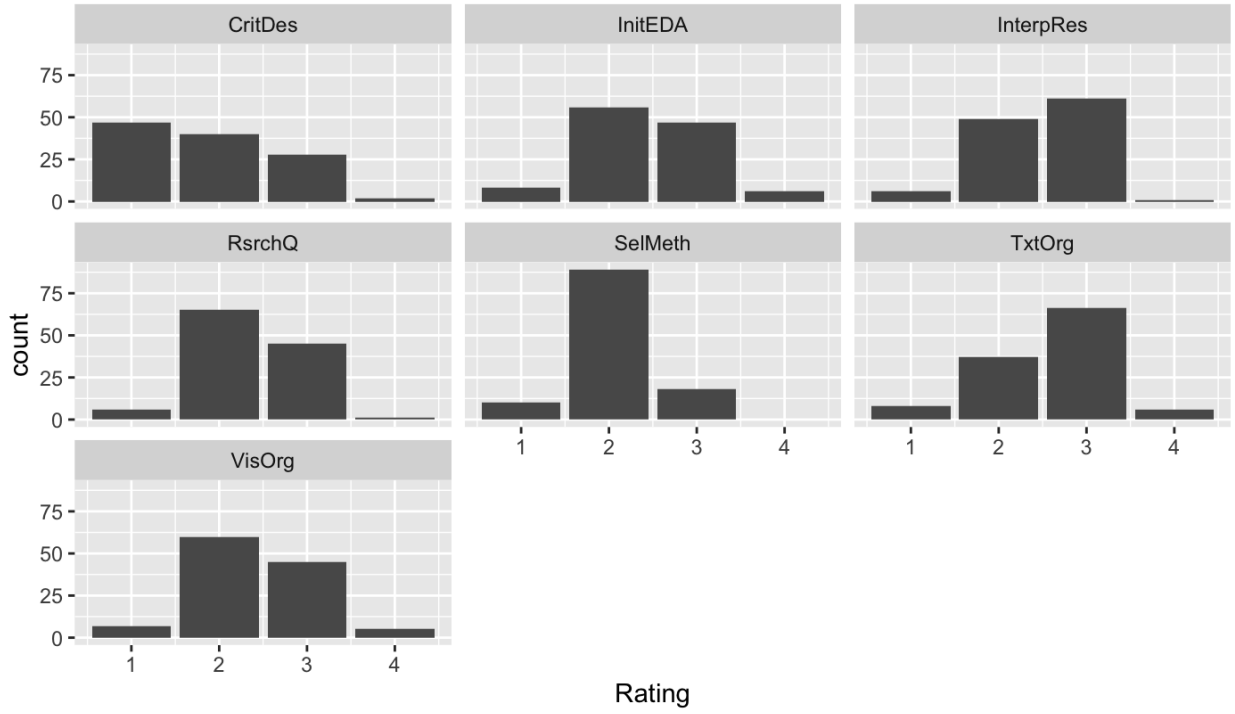
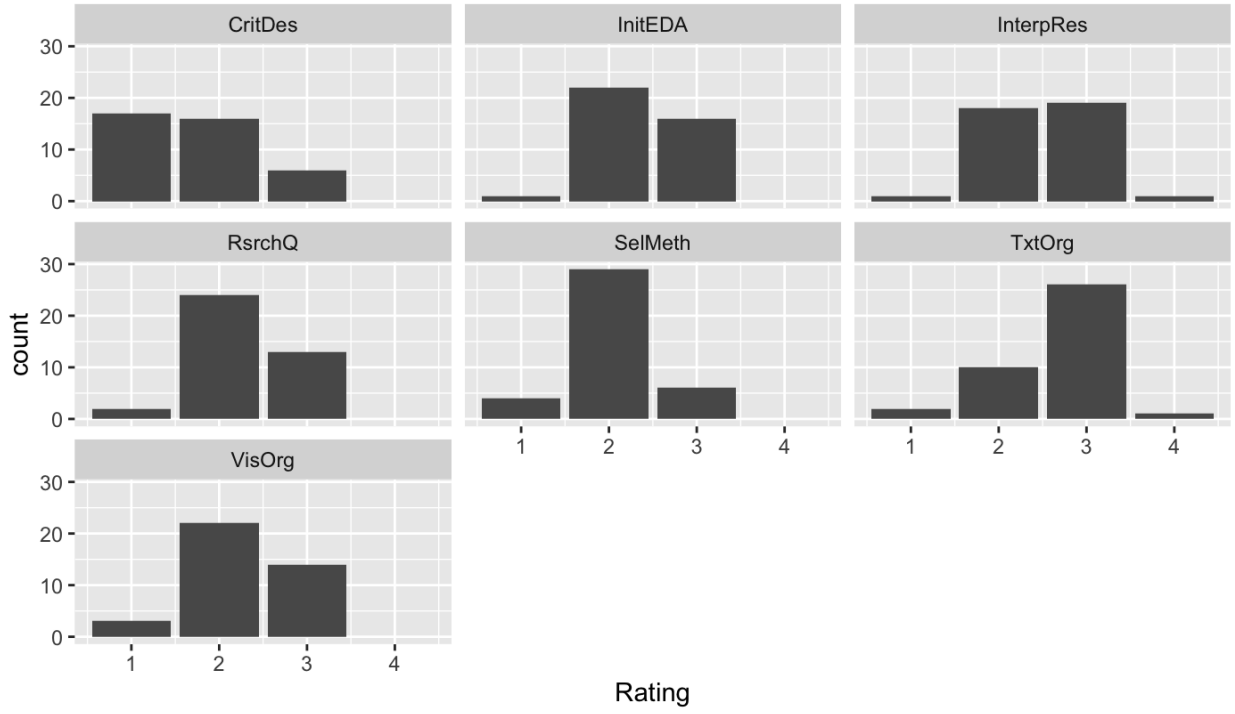


Figure 2: 13 Artifacts seen by all three raters, grouped by Rubrics



Research Question 1:

b. 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?

The distribution for 13 artifacts that were rated by all three raters in Figure 4 is similar to that of the full dataset in Figure 3. This suggests that the sample of 13 artifacts is representative of the population i.e., all 91 artifacts. We see that the distribution of ratings given by Rater 3 is most right skewed and that of Rater 2 is least right skewed. This suggests that Rater 3 tends to rate artifacts lower while Rater 2 tends to rate artifacts higher.

Figure 3: Full dataset, grouped by Raters

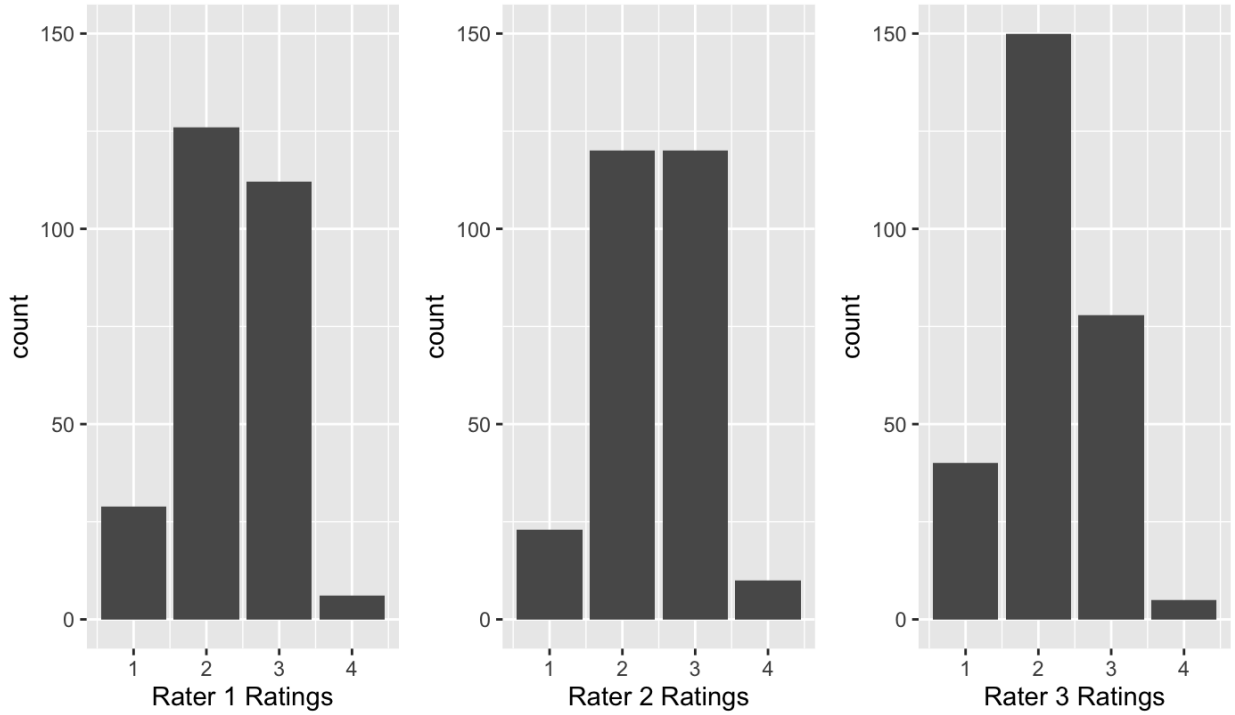
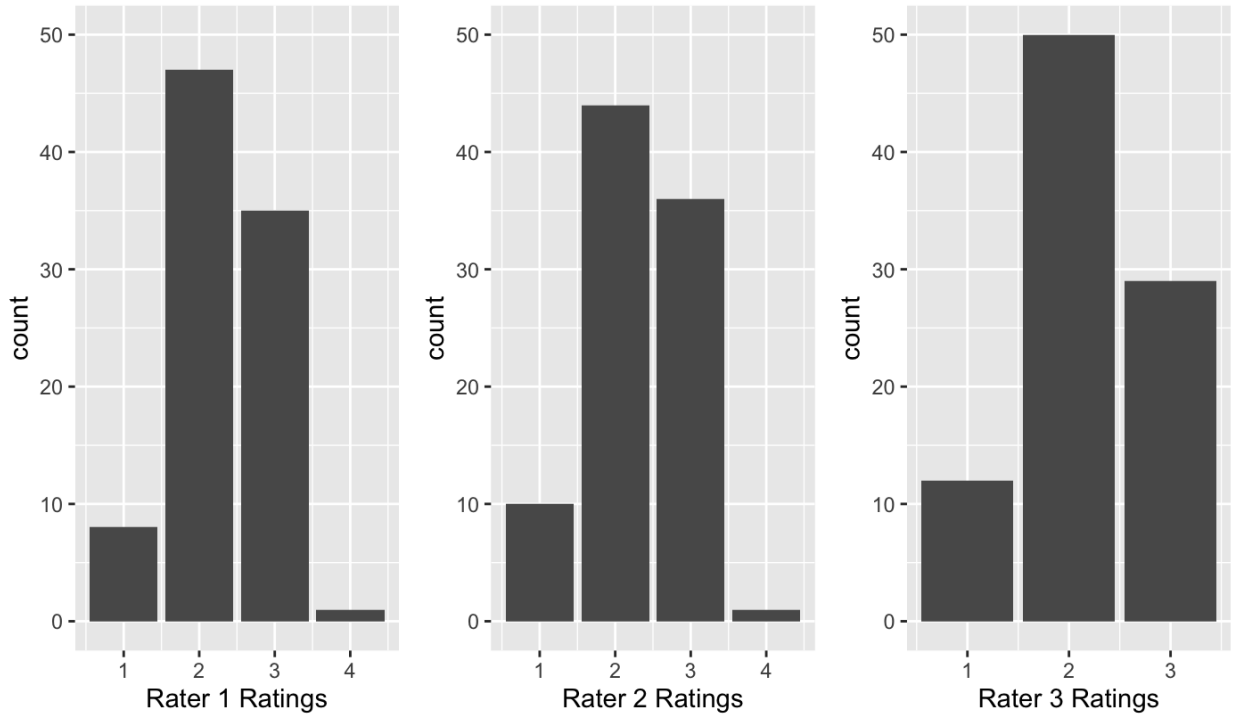


Figure 4: 13 Artifacts seen by all three raters, grouped by Raters



Research Question 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?

Table 6: Intraclass Correlation (ICC) and Inter-rater Reliability

Rubric	**ICC (13 Common Artifacts)**	**ICC (Full data)**	**Percent Exact Agreement for Rater 1 & 2**	**Percent Exact Agreement for Rater 1 & 3**	**Percent Exact Agreement for Rater 2 & 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	0.19	0.69	0.62	0.54

The ICCs for the full dataset seems to be higher than the ICCs for the subset of 13 common artifacts for all but 2 rubrics - InterpRes and SelMeth. The low ICCs for TxtOrg, RsrchQ, InterpRes suggest that raters usually tend to disagree on ratings for these rubrics. On the other hand, high ICCs for CritDes, InitEDA, SelMeth and VisOrg suggest that raters usually tend to agree on ratings for these rubrics. The percent exact agreement indicates that none of the pairs of raters agree or disagree more than the others.

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

Our final model:

Rating ~ Rater + Rubric + Rater:Rubric + (1|Artifact) + (0+ Rater|Artifact) + (0+ Rubric|Artifact)

According to the fixed effects in our model,

- Rater 2 tends to give higher ratings than rater 1 and 3 on average. This is consistent with our results for research question 1(b).
- Average ratings for CritDes < SelMeth < VisOrg < RsrchQ < InitEDA < InterpRes < TxtOrg which is consistent with Table 5 in the Data section which gives us the summary of ratings.
- Significant coefficients for interaction term between raters and rubrics suggests that raters tend to use rubrics differently. This is evident from the facet plots in Figure 5.

According to the random effects in our model,

- At the artifact level, rater 1 tends to have the least variation from the mean rating for that specific artifact.
- At the artifact level, SelMeth rubric tends to have the least variation from the mean rating and CritDesign tends to have the largest variance as shown in Figure 6.

Figure 5

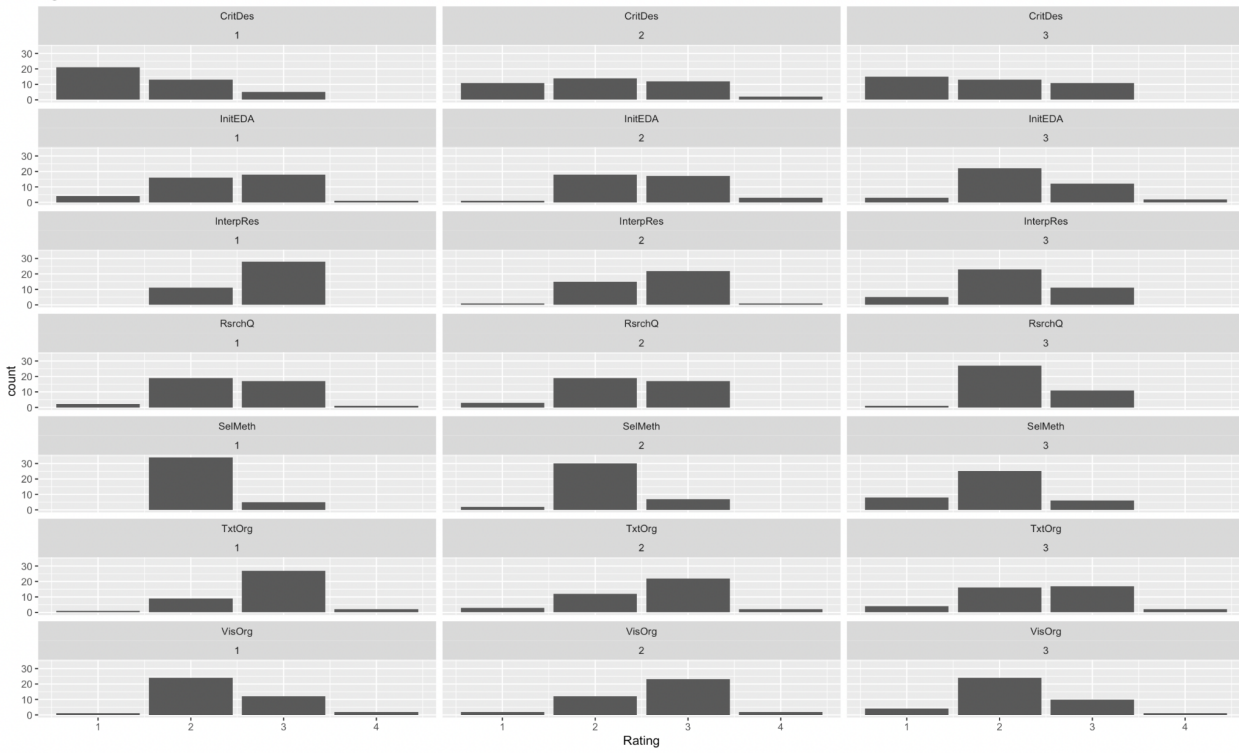
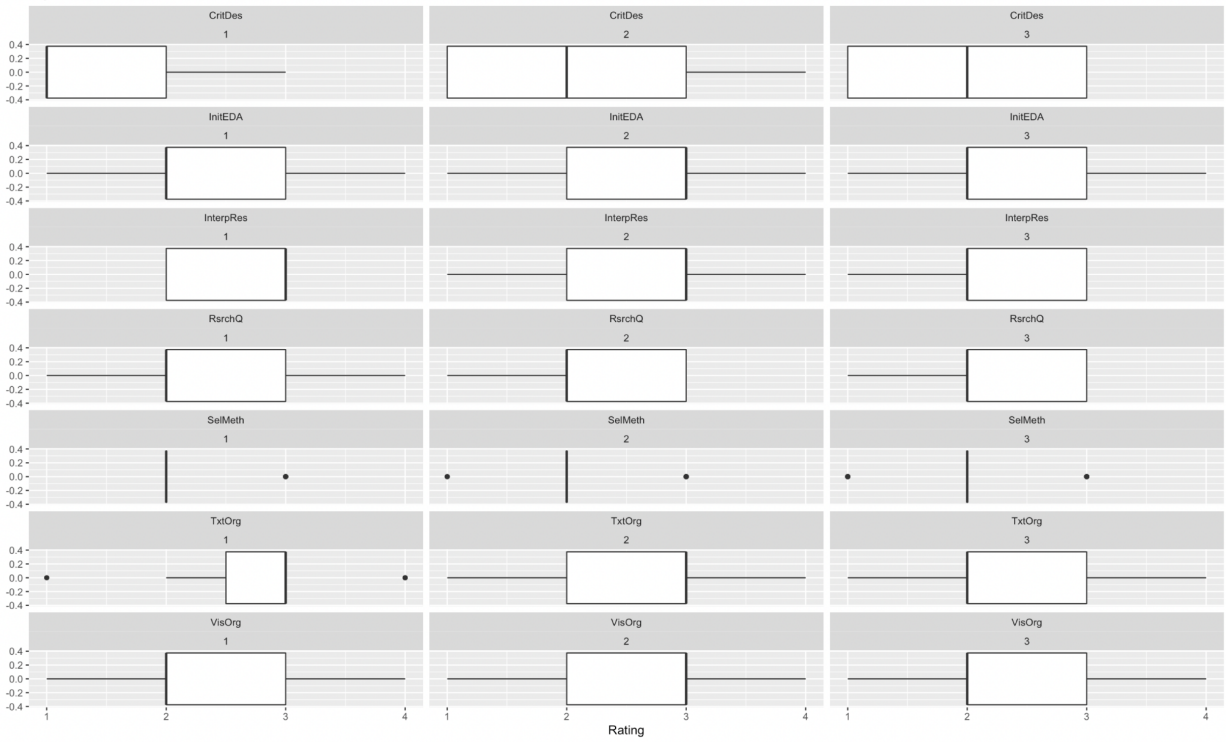


Figure 6



Research Question 4: Is there anything else interesting to say about this data?

WIP

DISCUSSION

WIP

REFERENCES

WIP

Final Project - Technical Appendix

Bhoomika Moorjani

11/29/2021

```
ratings <- read.csv("/Users/bhoomikamoorjani/Downloads/ratings.csv")
tall_data <- read.csv("/Users/bhoomikamoorjani/Downloads/tall.csv")
```

```
# Checking for missing values
tall_data[apply(tall_data, 1, function(x) {
    any(is.na(x))
}), ]
```

```
##       X Rater Artifact Repeated Semester Sex  Rubric Rating
## 161 161     2       45        0      S19   F CritDes     NA
## 684 684     1      100        0      F19   F  VisOrg     NA
```

```
ratings[apply(ratings[, -4], 1, function(x) {
    any(is.na(x))
}), ]
```

```
##     X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth
## 44 44     2     45      NA   Spring   F      2      NA       2       2
## 99 99     1    100      NA     Fall   F      2       3       2       3
##    InterpRes VisOrg TxtOrg Artifact Repeated
## 44         2      2      3       45        0
## 99         3     NA      2      100        0
```

```
# Assigning missing values(NAs) - Rating
getmode <- function(v) {
    uniqv <- unique(v)
    uniqv[which.max(tabulate(match(v, uniqv)))]
}
# Most common rating given by the rater for that rubric
# tall_data
tall_data$Rating[tall_data$X == 684] <- getmode(tall_data$Rating[which((tall_data$Rubric ==
    "VisOrg") & (tall_data$Rater == "1"))])

tall_data$Rating[tall_data$X == 161] <- getmode(tall_data$Rating[which((tall_data$Rubric ==
    "CritDes") & (tall_data$Rater == "2"))])

## ratings
ratings$VisOrg[ratings$X == 99] <- getmode(tall_data$Rating[which((tall_data$Rubric ==
    "VisOrg") & (tall_data$Rater == "1"))])
```

```
ratings$CritDes[ratings$X == 44] <- getmode(tall_data$Rating[which((tall_data$Rubric ==
    "CritDes") & (tall_data$Rater == "2"))])

# Assigning missing values(NAs) - Sex
tall_data$Sex[which(tall_data$Sex == "")] <- "Unknown"
ratings$Sex[which(ratings$Sex == "--")] <- "Unknown"
```

Research Question 1

```
rubric_ratings <- ratings[, 7:13]
# summary(rubric_ratings) %>% kable
temp_summary <- apply(rubric_ratings[, c(1, 3, 4, 5, 7)], 2,
    function(x) c(summary(x), SD = sd(x))) %>%
    as.data.frame %>%
    t() %>%
    round(digits = 2)
temp_summ_na <- apply(na.omit(rubric_ratings[, c(2, 6)]), 2,
    function(x) c(summary(x), SD = sd(x))) %>%
    as.data.frame %>%
    t() %>%
    round(digits = 2)

rbind(temp_summary, temp_summ_na) %>%
    kable(caption = "Ratings Summary") %>%
    kable_styling(latex_options = "HOLD_position")
```

Table 1: Ratings Summary

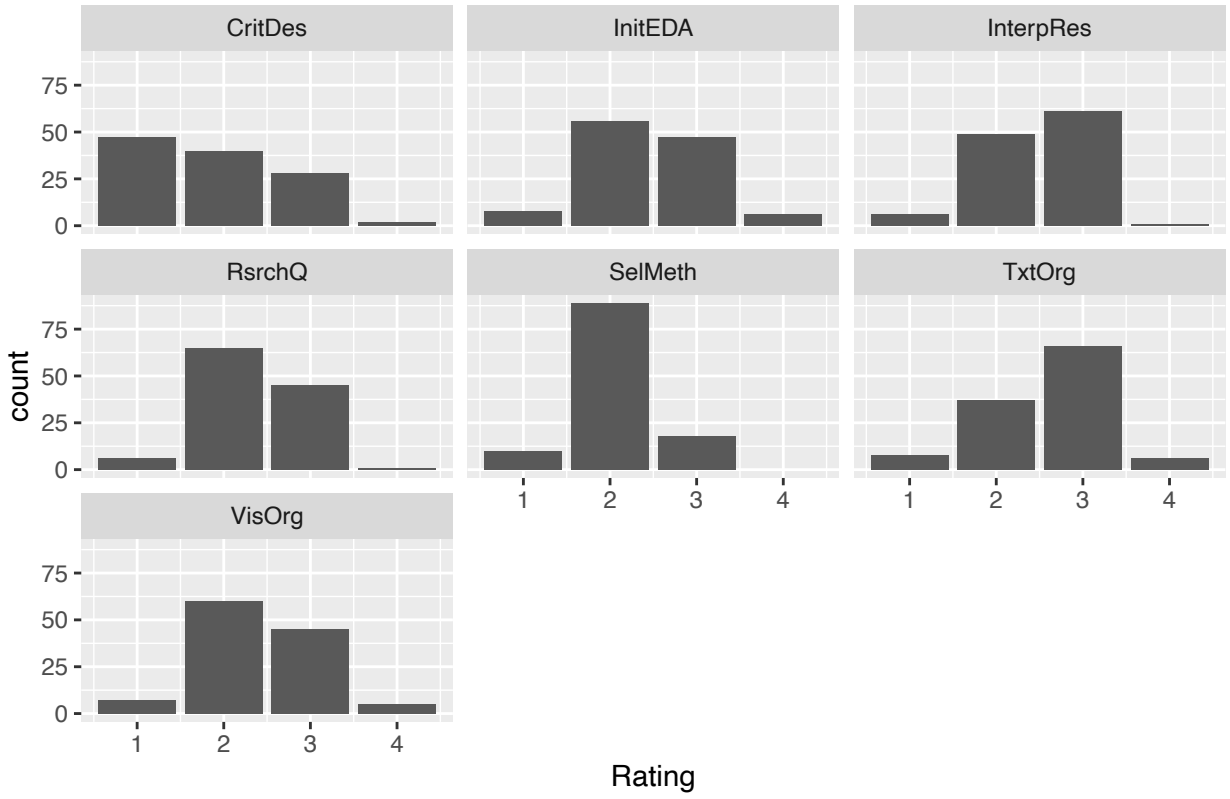
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.35	3	4	0.59
InitEDA	1	2	2	2.44	3	4	0.70
SelMeth	1	2	2	2.07	2	3	0.49
InterpRes	1	2	3	2.49	3	4	0.61
TxtOrg	1	2	3	2.60	3	4	0.70
CritDes	1	1	2	1.87	3	4	0.84
VisOrg	1	2	2	2.41	3	4	0.67

Raters have given lower scores for Method Selection and Critique Design on average.

```
# 13 common artifacts all three raters saw
tall13 <- tall_data[which(tall_data$Repeated == 1), ]
ratings13 <- ratings[which(ratings$Repeated == 1), ]
```

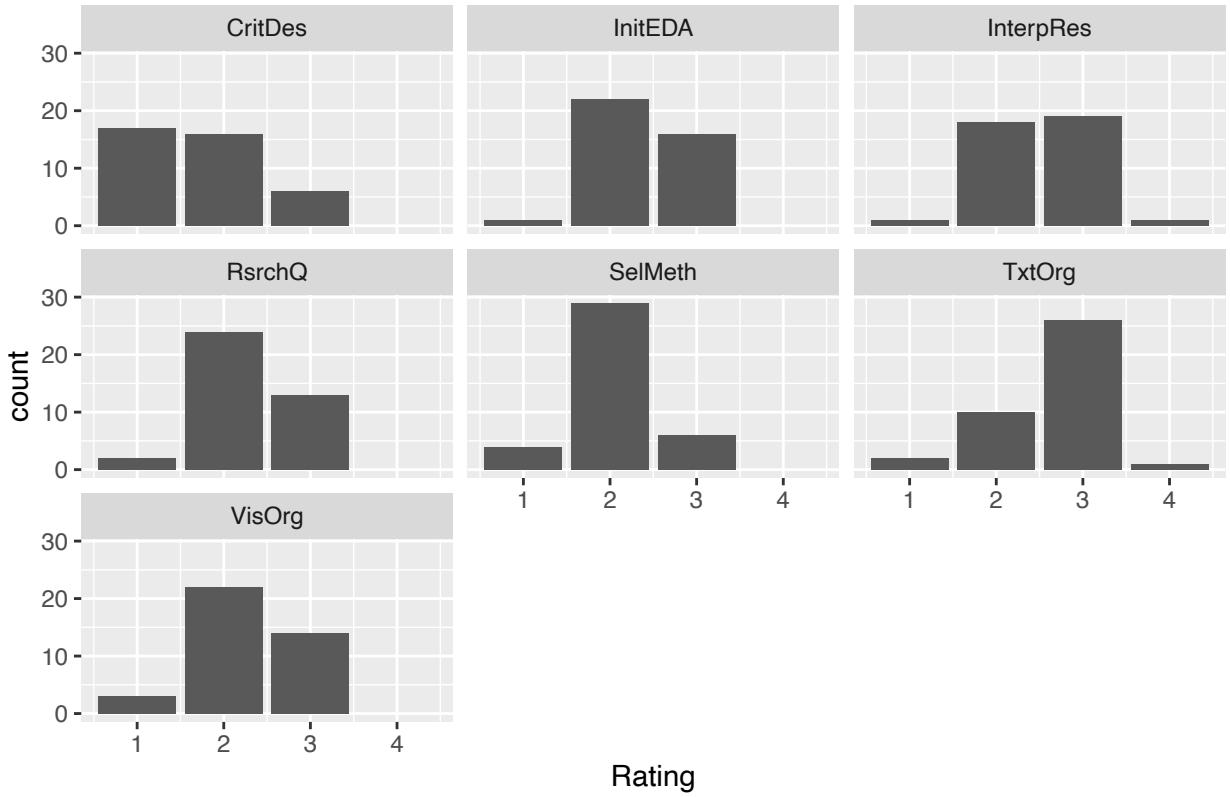
```
# Barplots for full dataset
g <- ggplot(tall_data, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar() + ggtitle("Figure 1: Full dataset, grouped by Rubrics")
g
```

Figure 1: Full dataset, grouped by Rubrics




```
# Barplots for 13 common artifacts
g <- ggplot(tall13, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar() + ggtitle("Figure 2: 13 Artifacts seen by all three raters, grouped by Rubrics")
g
```

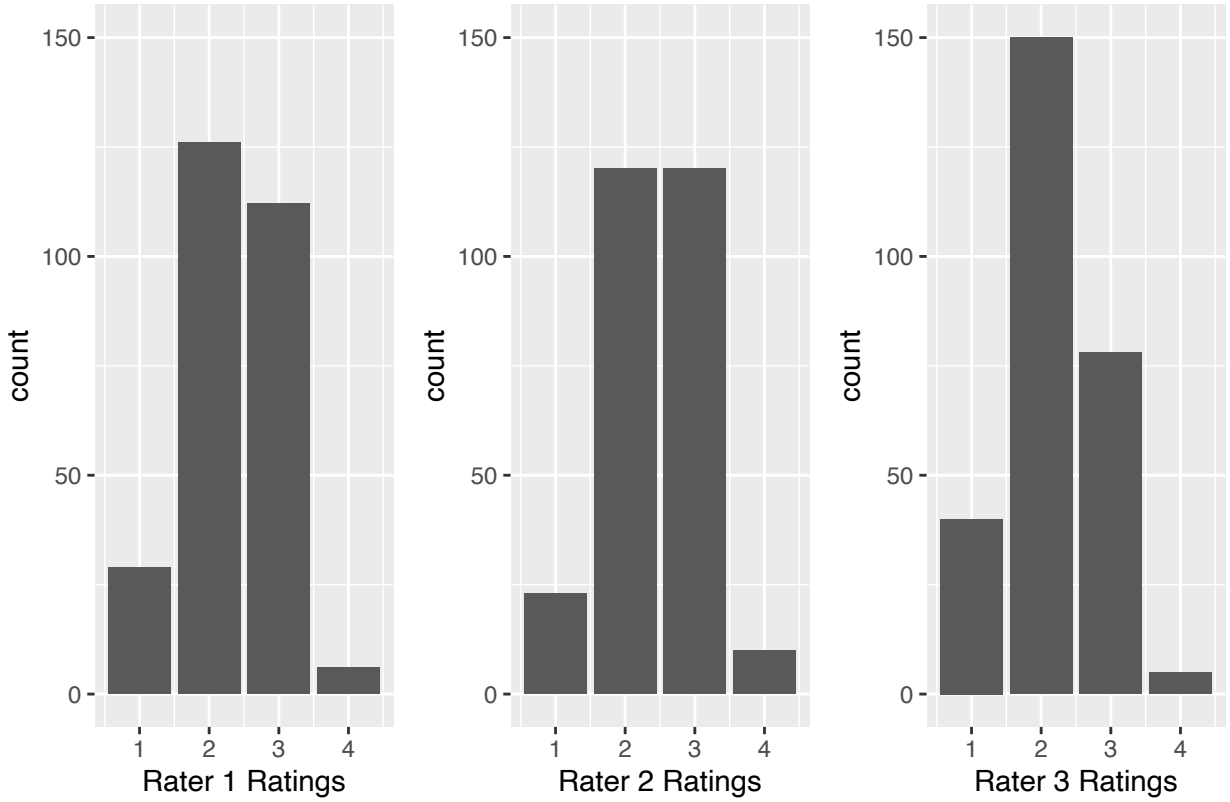
Figure 2: 13 Artifacts seen by all three raters, grouped by Rubrics



The distribution for 13 artifacts that were rated by all three raters is similar to rest of the artifacts that were only rater by one rater for all the rubrics. This suggests that this subset of 13 artifacts is representative of the whole 91 artifacts.

```
# Barplots for full dataset
tall1 <- tall_data$Rating[which(tall_data$Rater == 1)]
tall2 <- tall_data$Rating[which(tall_data$Rater == 2)]
tall3 <- tall_data$Rating[which(tall_data$Rater == 3)]
f <- ggarrange(ggplot(as.data.frame(tall1), aes(tall1)) + geom_bar() +
    labs(x = "Rater 1 Ratings") + ylim(0, 150), ggplot(as.data.frame(tall2),
    aes(tall2)) + geom_bar() + labs(x = "Rater 2 Ratings") +
    ylim(0, 150), ggplot(as.data.frame(tall3), aes(tall3)) +
    geom_bar() + labs(x = "Rater 3 Ratings") + ylim(0, 150),
    ncol = 3, nrow = 1)
annotate_figure(f, top = text_grob("Figure 3: Full dataset, grouped by Raters"))
```

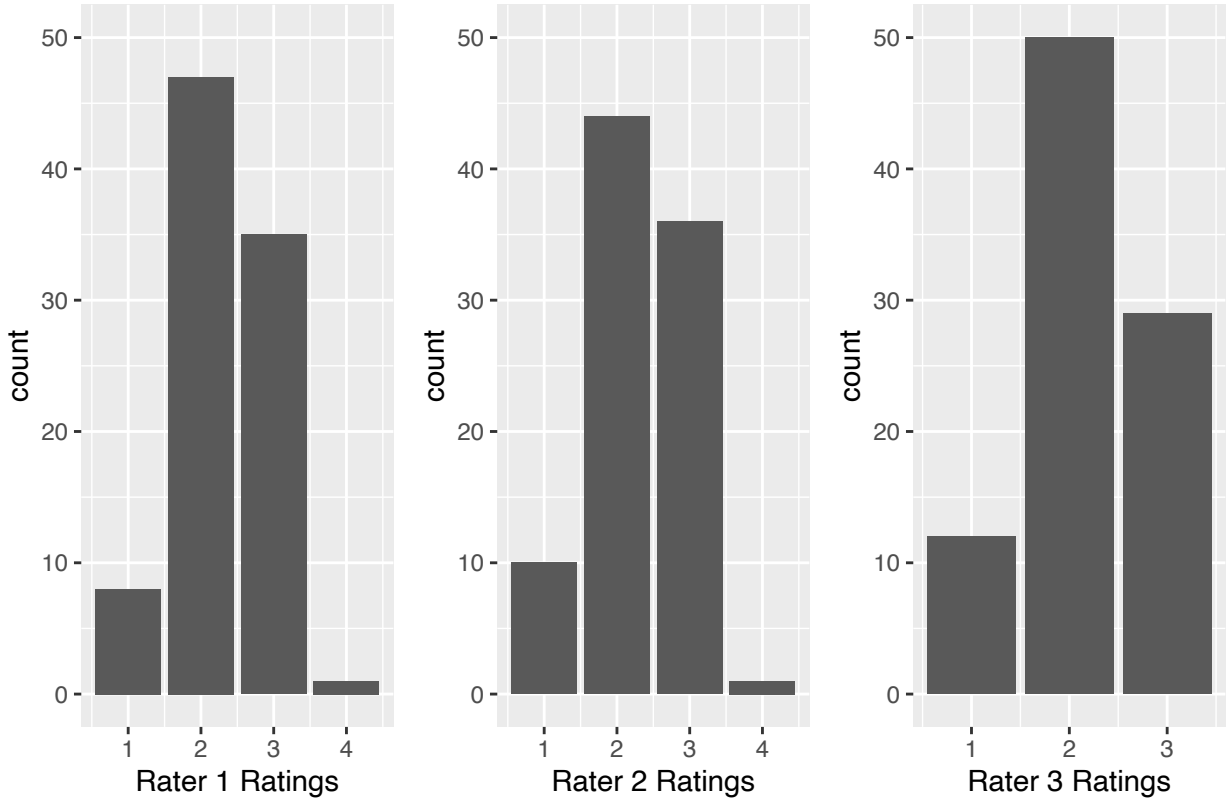
Figure 3: Full dataset, grouped by Raters



The above plot suggests that Rater 3 tends to rate artifacts lower i.e., is stricter while rater 2 is lenient and rates artifacts higher. Distribution of ratings for all three raters is right skewed i.e., they don't give a rating of 4 very often.

```
# Barplots for 13 common artifacts
tall1 <- tall13$Rating[which(tall13$Rater == 1)]
tall2 <- tall13$Rating[which(tall13$Rater == 2)]
tall3 <- tall13$Rating[which(tall13$Rater == 3)]
h <- ggarrange(ggplot(as.data.frame(tall1), aes(tall1)) + geom_bar() +
    labs(x = "Rater 1 Ratings") + ylim(0, 50), ggplot(as.data.frame(tall2),
    aes(tall2)) + geom_bar() + labs(x = "Rater 2 Ratings") +
    ylim(0, 50), ggplot(as.data.frame(tall3), aes(tall3)) + geom_bar() +
    labs(x = "Rater 3 Ratings") + ylim(0, 50), ncol = 3, nrow = 1)
annotate_figure(h, top = text_grob("Figure 4: 13 Artifacts seen by all three raters, grouped by Raters"
```

Figure 4: 13 Artifacts seen by all three raters, grouped by Raters



Research Question 2

```
# Function to calculate ICC
calculate_icc <- function(tau, sigma) {
    icc <- tau^2/(tau^2 + sigma^2)
    return(icc)
}
```

```
# Rater Agreement (ICC) - RsrchQ
RsrchQ.ratings <- tall13[tall13$Rubric == "RsrchQ", ]
lmer(Rating ~ 1 + (1 | Artifact), data = RsrchQ.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
## REML criterion at convergence: 66.1533
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2446
##  Residual             0.5064
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.282
```

```r
RsrchQ.icc <- calculate_icc(0.2446, 0.5064)
```

```r
# Rater Agreement (ICC) - CritDes
CritDes.ratings <- tall13[tall13$Rubric == "CritDes", ]
lmer(Rating ~ 1 + (1 | Artifact), data = CritDes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
## REML criterion at convergence: 75.1397
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.5560
##  Residual             0.4804
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       1.718
```

```r
CritDes.icc <- calculate_icc(0.556, 0.4804)
```

```r
# Rater Agreement (ICC) - InitEDA
InitEDA.ratings <- tall13[tall13$Rubric == "InitEDA", ]
lmer(Rating ~ 1 + (1 | Artifact), data = InitEDA.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
## REML criterion at convergence: 56.7573
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3867
##  Residual             0.3922
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.385
```

```r
InitEDA.icc <- calculate_icc(0.3867, 0.3922)
```

```r
# Rater Agreement (ICC) - SelMeth
SelMeth.ratings <- tall13[tall13$Rubric == "SelMeth", ]
lmer(Rating ~ 1 + (1 | Artifact), data = SelMeth.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
## REML criterion at convergence: 50.8562
## Random effects:
##  Groups   Name        Std.Dev.
```

```
##  Artifact (Intercept) 0.3736
##  Residual             0.3581
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.051
```

```
SelMeth.icc <- calculate_icc(0.3736, 0.3581)
```

```
# Rater Agreement (ICC) - InterpRes
InterpRes.ratings <- tall13[tall13$Rubric == "InterpRes", ]
lmer(Rating ~ 1 + (1 | Artifact), data = InterpRes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
## REML criterion at convergence: 71.0715
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2899
##  Residual             0.5311
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.513
```

```
InterpRes.icc <- calculate_icc(0.2899, 0.5311)
```

```
# Rater Agreement (ICC) - VisOrg
VisOrg.ratings <- tall13[tall13$Rubric == "VisOrg", ]
lmer(Rating ~ 1 + (1 | Artifact), data = VisOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
## REML criterion at convergence: 60.5245
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.4729
##  Residual             0.3922
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.282
```

```
VisOrg.icc <- calculate_icc(0.4729, 0.3922)
```

```
# Rater Agreement (ICC) - TxtOrg
TxtOrg.ratings <- tall13[tall13$Rubric == "TxtOrg", ]
lmer(Rating ~ 1 + (1 | Artifact), data = TxtOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
## REML criterion at convergence: 74.6212
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2357
##  Residual             0.5774
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.667
```

```
TxtOrg.icc <- calculate_icc(0.2357, 0.5774)
```

```
df <- data.frame(Rubric = c("RsrchQ", "CritDes", "InitEDA", "SelMeth",
    "InterpRes", "VisOrg", "TxtOrg"), ICC = c(RsrchQ.icc, CritDes.icc,
    InitEDA.icc, SelMeth.icc, InterpRes.icc, VisOrg.icc, TxtOrg.icc))
df
```

```
##      Rubric       ICC
## 1    RsrchQ 0.1891711
## 2   CritDes 0.5725587
## 3   InitEDA 0.4929391
## 4   SelMeth 0.5211740
## 5 InterpRes 0.2295545
## 6    VisOrg 0.5924793
## 7    TxtOrg 0.1428337
```

For the 13 artifacts which were rated by all three raters, the ratings are highly correlated (ICC greater than 0.5) for three rubrics Visual organization, Method Selection, Initial EDA and Critique Design i.e., they tend to agree on the ratings in these rubrics for more than 50% of the artifacts. They disagree on ratings i.e, have lower ICC and ratings are weakly correlated for Research Question, Interpret Results and Text Organization.

```
Rubrics <- unique(tall13$Rubric)
Artifacts <- unique(tall13$Artifact)
perf_agree = rep(0, length(Rubrics))
for (i in Rubrics) {
    for (j in Artifacts) {
        if (tall13$Rating[which((tall13$Rubric == i) & (tall13$Artifact ==
            j) & (tall13$Rater == 1))] == tall13$Rating[which((tall13$Rubric ==
            i) & (tall13$Artifact == j) & (tall13$Rater == 2))])
            perf_agree[which(Rubrics == i)] = perf_agree[which(Rubrics ==
                i)] + 1
    }
}
rater1_rater2 <- perf_agree

perf_agree = rep(0, length(Rubrics))
for (i in Rubrics) {
    for (j in Artifacts) {
        if (tall13$Rating[which((tall13$Rubric == i) & (tall13$Artifact ==
            j) & (tall13$Rater == 1))] == tall13$Rating[which((tall13$Rubric ==
```

```
            i) & (tall13$Artifact == j) & (tall13$Rater == 3))])
            perf_agree[which(Rubrics == i)] = perf_agree[which(Rubrics ==
                i)] + 1
    }
}
rater1_rater3 <- perf_agree

perf_agree = rep(0, length(Rubrics))
for (i in Rubrics) {
    for (j in Artifacts) {
        if (tall13$Rating[which((tall13$Rubric == i) & (tall13$Artifact ==
            j) & (tall13$Rater == 2))] == tall13$Rating[which((tall13$Rubric ==
            i) & (tall13$Artifact == j) & (tall13$Rater == 3))])
            perf_agree[which(Rubrics == i)] = perf_agree[which(Rubrics ==
                i)] + 1
    }
}
rater2_rater3 <- perf_agree

# Percent Exact Agreement
df2 <- data.frame(Rubric = c("RsrchQ", "CritDes", "InitEDA",
    "SelMeth", "InterpRes", "VisOrg", "TxtOrg"), rater1_rater2 = round(rater1_rater2/13,
    2), rater1_rater3 = round(rater1_rater3/13, 2), rater2_rater3 = round(rater2_rater3/13,
    2))
df2
```

```
##      Rubric rater1_rater2 rater1_rater3 rater2_rater3
## 1    RsrchQ          0.38          0.77          0.54
## 2   CritDes          0.54          0.62          0.69
## 3   InitEDA          0.69          0.54          0.85
## 4   SelMeth          0.92          0.62          0.69
## 5 InterpRes          0.62          0.54          0.62
## 6    VisOrg          0.54          0.77          0.77
## 7    TxtOrg          0.69          0.62          0.54
```

```
mean(df2$rater1_rater2)
```

```
## [1] 0.6257143
```

```
mean(df2$rater1_rater3)
```

```
## [1] 0.64
```

```
mean(df2$rater2_rater3)
```

```
## [1] 0.6714286
```

Rater 2 and 3 agree 67% on average. Rater 1 and 3 agree 64% on average. Rater 1 and 2 agree 63% on average.

```
icc_full = rep(0, length(Rubrics))
for (x in Rubrics) {
    model <- lmer(Rating ~ 1 + (1 | Artifact), data = tall_data[tall_data$Rubric ==
        x, ])
    icc_full[which(Rubrics == x)] = performance::icc(model = model)[1]
}

df3 <- data.frame(Rubric = c("RsrchQ", "CritDes", "InitEDA",
    "SelMeth", "InterpRes", "VisOrg", "TxtOrg"), ICC = c(RsrchQ.icc,
    CritDes.icc, InitEDA.icc, SelMeth.icc, InterpRes.icc, VisOrg.icc,
    TxtOrg.icc), icc_full = unlist(icc_full))
df3
```

```
##      Rubric       ICC  icc_full
## 1    RsrchQ 0.1891711 0.2096214
## 2   CritDes 0.5725587 0.6699202
## 3   InitEDA 0.4929391 0.6867210
## 4   SelMeth 0.5211740 0.4719014
## 5 InterpRes 0.2295545 0.2200285
## 6    VisOrg 0.5924793 0.6586320
## 7    TxtOrg 0.1428337 0.1879927
```

The ICC for the full data set is higher than ICC for the 13 common artifacts in all rubrics except Method Selection and Interpret Results. But similar to the 13 common artifacts, the ratings in Critique Design, Initial EDA, Method Selection, Visual Organization in the full data set are highly correlated.

Research Question 3

```
tall_data$Rater <- as.factor(tall_data$Rater)
# Full model
model1 <- lmer(Rating ~ Rater + Repeated + Semester + Rubric +
    Sex + (1 | Artifact), data = tall_data, REML = FALSE)
# Removing Sex as a fixed effect
model2 <- lmer(Rating ~ Rater + Repeated + Semester + Rubric +
    (1 | Artifact), data = tall_data, REML = FALSE)
anova(model1, model2)  #Likes Model 2
```

```
## Data: tall_data
## Models:
## model2: Rating ~ Rater + Repeated + Semester + Rubric + (1 | Artifact)
## model1: Rating ~ Rater + Repeated + Semester + Rubric + Sex + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2   13 1520.6 1581.8 -747.28   1494.6
## model1   15 1521.2 1591.8 -745.60   1491.2 3.3622  2     0.1862
```

```
# Removing Rubric as a fixed effect
model3 <- lmer(Rating ~ Rater + Repeated + Semester + (1 | Artifact),
    data = tall_data, REML = FALSE)
anova(model2, model3)  #Likes Model 2
```

```
## Data: tall_data
## Models:
```

```
## model3: Rating ~ Rater + Repeated + Semester + (1 | Artifact)
## model2: Rating ~ Rater + Repeated + Semester + Rubric + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3    7 1643.8 1676.8 -814.90   1629.8
## model2   13 1520.6 1581.8 -747.28   1494.6 135.23  6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adding Rubric back and removing semester
model4 <- lmer(Rating ~ Rater + Repeated + Rubric + (1 | Artifact),
    data = tall_data, REML = FALSE)
anova(model2, model4)  #Likes Model 4
```

```
## Data: tall_data
## Models:
## model4: Rating ~ Rater + Repeated + Rubric + (1 | Artifact)
## model2: Rating ~ Rater + Repeated + Semester + Rubric + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model4   12 1520.4 1576.9 -748.22   1496.4
## model2   13 1520.6 1581.8 -747.28   1494.6 1.8743  1      0.171
```

```
# Removing Repeated
model5 <- lmer(Rating ~ Rater + Rubric + (1 | Artifact), data = tall_data,
    REML = FALSE)
anova(model4, model5)  #Likes Model 5
```

```
## Data: tall_data
## Models:
## model5: Rating ~ Rater + Rubric + (1 | Artifact)
## model4: Rating ~ Rater + Repeated + Rubric + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model5   11 1518.8 1570.6 -748.40   1496.8
## model4   12 1520.4 1576.9 -748.22   1496.4 0.3682  1      0.544
```

```
# Removing Rater
model6 <- lmer(Rating ~ Rubric + (1 | Artifact), data = tall_data,
    REML = FALSE)
anova(model5, model6)  #Likes Model 5
```

```
## Data: tall_data
## Models:
## model6: Rating ~ Rubric + (1 | Artifact)
## model5: Rating ~ Rater + Rubric + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model6    9 1523.5 1565.8 -752.74   1505.5
## model5   11 1518.8 1570.6 -748.40   1496.8 8.6701  2     0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Rating ~ Rater + Rubric + (1 | Artifact)

# Checking for interaction
```

```
model7 <- lmer(Rating ~ Rater * Rubric + (1 | Artifact), data = tall_data,
    REML = FALSE)
anova(model5, model7)  #Likes Model 7
```

```
## Data: tall_data
## Models:
## model5: Rating ~ Rater + Rubric + (1 | Artifact)
## model7: Rating ~ Rater * Rubric + (1 | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model5   11 1518.8 1570.6 -748.40   1496.8
## model7   23 1503.2 1611.5 -728.63   1457.2 39.551 12  8.534e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Random effects
model8 <- lmer(Rating ~ Rater * Rubric + (1 | Artifact) + (0 +
    Rater | Artifact), data = tall_data, REML = FALSE)
```

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

```
anova(model7, model8)  #Likes Model 8
```

```
## Data: tall_data
## Models:
## model7: Rating ~ Rater * Rubric + (1 | Artifact)
## model8: Rating ~ Rater * Rubric + (1 | Artifact) + (0 + Rater | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model7   23 1503.2 1611.5 -728.63   1457.2
## model8   29 1492.7 1629.2 -717.34   1434.7 22.579  6  0.0009504 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model9 <- lmer(Rating ~ Rater * Rubric + (1 | Artifact) + (0 +
    Rater | Artifact) + (0 + Rubric | Artifact), data = tall_data,
    REML = FALSE)
```

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

```
anova(model8, model9)  #Likes Model 9
```

```
## Data: tall_data
## Models:
## model8: Rating ~ Rater * Rubric + (1 | Artifact) + (0 + Rater | Artifact)
## model9: Rating ~ Rater * Rubric + (1 | Artifact) + (0 + Rater | Artifact) + (0 + Rubric | Artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model8   29 1492.7 1629.2 -717.34   1434.7
## model9   57 1431.9 1700.2 -658.94   1317.9 116.79 28  8.218e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

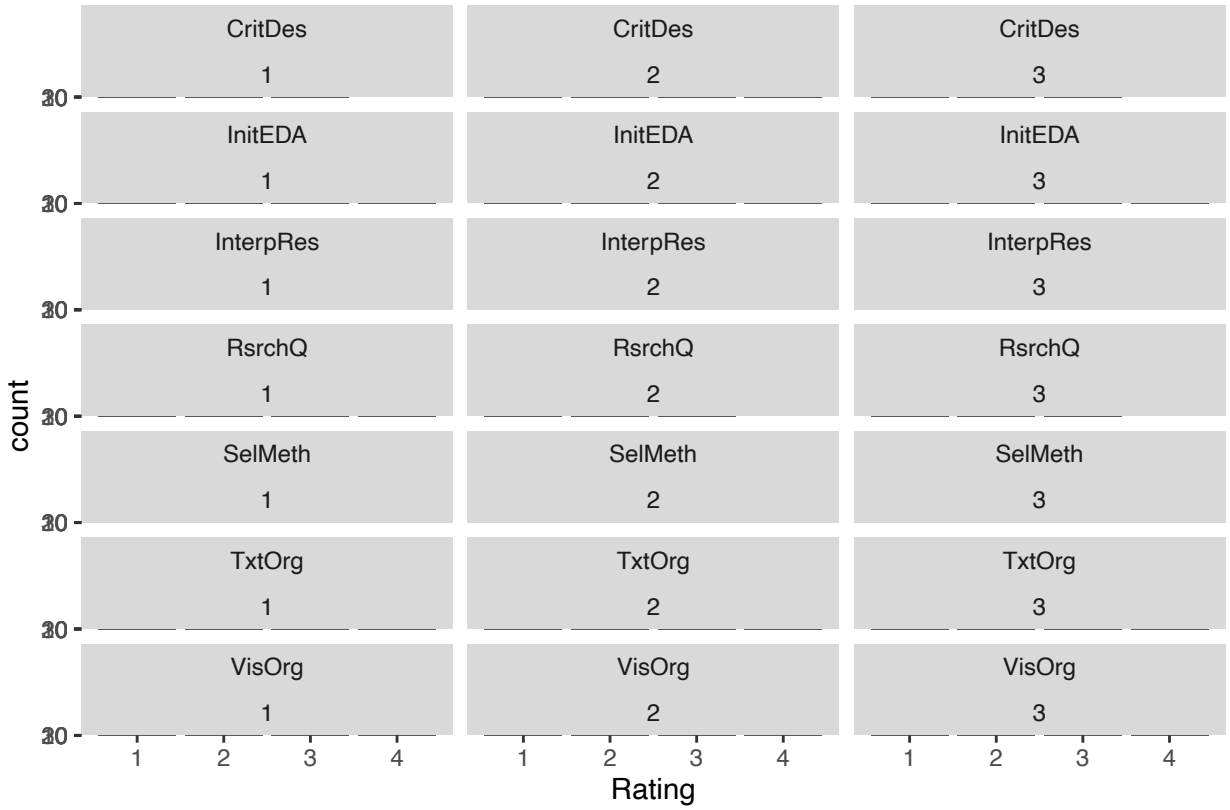
```
display(model9)
```

```
## lmer(formula = Rating ~ Rater * Rubric + (1 | Artifact) + (0 +
##     Rater | Artifact) + (0 + Rubric | Artifact), data = tall_data,
##     REML = FALSE)
##                         coef.est coef.se
## (Intercept)              1.72     0.11
## Rater2                   0.37     0.14
## Rater3                   0.21     0.13
## RubricInitEDA            0.74     0.13
## RubricInterpRes          0.99     0.13
## RubricRsrchQ             0.72     0.12
## RubricSelMeth            0.41     0.12
## RubricTxtOrg             1.01     0.13
## RubricVisOrg             0.65     0.13
## Rater2:RubricInitEDA    -0.30     0.15
## Rater3:RubricInitEDA    -0.30     0.15
## Rater2:RubricInterpRes -0.51     0.15
## Rater3:RubricInterpRes -0.72     0.15
## Rater2:RubricRsrchQ     -0.49     0.14
## Rater3:RubricRsrchQ     -0.33     0.14
## Rater2:RubricSelMeth    -0.39     0.15
## Rater3:RubricSelMeth    -0.37     0.15
## Rater2:RubricTxtOrg     -0.55     0.15
## Rater3:RubricTxtOrg     -0.45     0.15
## Rater2:RubricVisOrg     -0.11     0.16
## Rater3:RubricVisOrg     -0.28     0.16
##
## Error terms:
##  Groups     Name            Std.Dev. Corr
##  Artifact   (Intercept)     0.00
##  Artifact.1 Rater1          0.12
##             Rater2          0.34     -0.31
##             Rater3          0.34      0.46  0.70
##  Artifact.2 RubricCritDes   0.69
##             RubricInitEDA   0.54      0.31
##             RubricInterpRes 0.30      0.13  0.65
##             RubricRsrchQ    0.40      0.50  0.15  0.49
##             RubricSelMeth   0.20      0.18  0.20  0.36 -0.27
##             RubricTxtOrg    0.48      0.26  0.41  0.32  0.27  0.20
##             RubricVisOrg    0.47      0.17  0.49  0.42  0.24 -0.13  0.52
##  Residual                   0.36
## ---
## number of obs: 819, groups: Artifact, 91
## AIC = 1431.9, DIC = 1317.9
## deviance = 1317.9
```

Final Model: **model9: Rating ~ Rater * Rubric + (1 | Artifact) + (0 + Rater | Artifact) + (0 + Rubric | Artifact)**

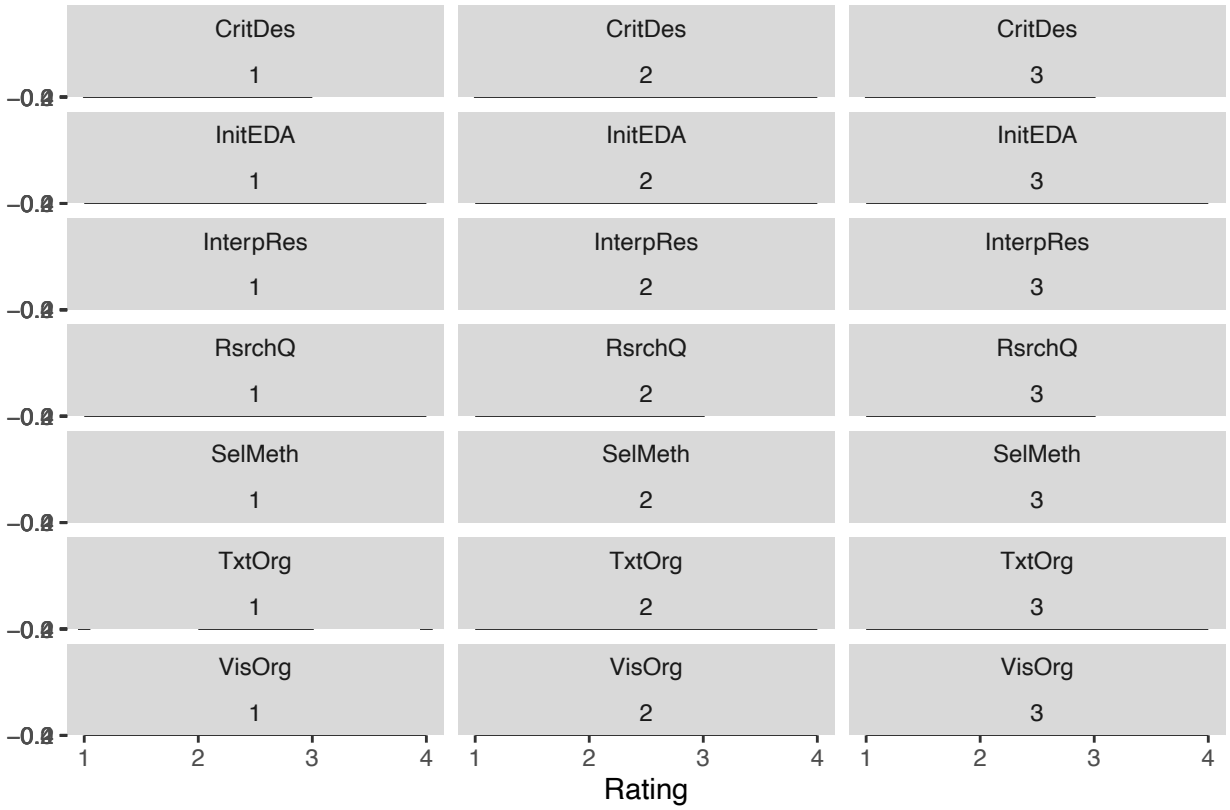
```
g <- ggplot(tall_data, aes(x = Rating)) + geom_bar() + facet_wrap(~Rubric +
    Rater, nrow = 7) + ggtitle("Figure 5")
g
```

Figure 5




```
g <- ggplot(tall_data, aes(x = Rating)) + geom_boxplot() + facet_wrap(~Rubric +
    Rater, nrow = 7) + ggtitle("Figure 6")
g
```

Figure 6



Question 4 We observe that rater 3 tends to be more strict than rater 2 more lenient. We would therefore expect them to disagree on ratings of high number of artifacts. However, they have a high percentage of exact agreement. There are certain rubrics for which ratings are more strongly correlated than others. The ones that raters disagree on could be associated to the department they are from and the differences in the way research is conducted in that specific domain.