Final 241 Project Consolidation

author: Elizabeth Khan, Estrella Ndrianasy, Chandni Shah, Michelle Shen, Catherine Tsai date: "November 8, 2021" output: html_document —

```
# Import data and remove invalid age

d <- fread("survey_cleaned_with_states.csv")
head(d)
```

```
##           ResponseId Gender   Age              Education LocationLatitude
## 1: R_2WC5vlB2DCvHHrH   Male 30-40       Bachelors degree         52.36690
## 2: R_x2RQnFD8LZIZuRH Female 40-50      Associates degree         36.99040
## 3: R_09f8Af9asil81DH Female 30-40 Some College No degree         27.99290
## 4: R_3oTqFe6v8l4jV0p Female 30-40 Some College No degree         36.83070
## 5: R_0ePUcdKMDE2jNe1 Female 30-40       Bachelors degree         39.30370
## 6: R_2owAeGSUxdN9fQR Female 40-50      Associates degree         44.14149
##    LocationLongitude instr_enthusiasm instr_professional instr_subject
## 1:          4.951706                3                  4             4
## 2:        -85.923897                4                  4             4
## 3:        -82.403000                1                  4             3
## 4:        -76.314598                2                  2             2
## 5:        -94.932198                3                  1             1
## 6:       -103.205200                4                  4             5
##    instr_material instr_rating                        subject_1
## 1:              4            4                 All of the above
## 2:              4            4                 Web Applications
## 3:              2            4                 Web Applications
## 4:              2            2                 Web Applications
## 5:              3            2 Artificial intelligence projects
## 6:              4            4                 Web Applications
##                                          subject_2
## 1:                                      Both A & B
## 2:                                      Both A & B
## 3: C. A low level implementation programming language
## 4:                 B. A flexible programming language
## 5:                 B. A flexible programming language
## 6:                 B. A flexible programming language
##
## 1:                                                                                               Python is an objec
## 2: Python is a compiled programming language with a faster and more efficient execution time than in
## 3:                                                                                               Python is an objec
## 4:                                                                                                  Python was crea
## 5:                                                                                                  Python was crea
## 6:                                                                                               Python is an objec
##              subject_4       video_watched assigned assignment ques_1
## 1:           Easy to Learn Video1-1(Treatment)        1  Treatment      1
```

```
## 2: Great Advanced Language   Video3-2(Control)        0    Control      0
## 3:           Easy to Learn   Video2-2(Control)        0    Control      0
## 4:     Wonderful Community Video2-1(Treatment)        1  Treatment      0
## 5:     Wonderful Community   Video1-2(Control)        0    Control      0
## 6:     Wonderful Community   Video2-2(Control)        0    Control      0
##    ques_2 ques_3 ques_4 quiz_avg avg_rating        state state_code dem_percent
## 1:      1      0      0     0.50       3.75                                  NA
## 2:      1      1      1     0.75       4.00     kentucky         KY       0.362
## 3:      0      0      0     0.00       2.75      florida         FL       0.479
## 4:      0      0      0     0.00       2.00     virginia         VA       0.541
## 5:      0      0      0     0.00       2.25       kansas         KS       0.416
## 6:      0      0      0     0.00       4.00 south dakota         SD       0.356
##    rep_percent
## 1:          NA
## 2:       0.621
## 3:       0.512
## 4:       0.440
## 5:       0.562
## 6:       0.618
```

```
d <- d[Age!='Other']


degrees = c("Bachelors degree","Associates degree","Masters degree")

# create column to indicate whether or not subject has a college degree

d<- d[,college_degree:= ifelse(Education %in% degrees, 1,0)]
length(d$college_degree)
```

```
## [1] 222
```

Figures and Tables

```
summary_response_table <- d %>%
  count(Gender, Age, assignment)

summary_response_table
```

```
##     Gender   Age assignment  n
##  1: Female 20-30    Control 14
##  2: Female 20-30  Treatment 11
##  3: Female 30-40    Control 27
##  4: Female 30-40  Treatment 29
##  5: Female 40-50    Control 15
##  6: Female 40-50  Treatment 17
##  7: Female   50+    Control 12
##  8: Female   50+  Treatment 17
##  9:   Male 20-30    Control  9
## 10:   Male 20-30  Treatment 10
## 11:   Male 30-40    Control 12
```

```
## 12:   Male 30-40  Treatment 15
## 13:   Male 40-50    Control 11
## 14:   Male 40-50  Treatment  5
## 15:   Male   50+    Control  8
## 16:   Male   50+  Treatment 10
```

```
## Table printed with 'knitr::kable()', not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include 'message = FALSE' in code chunk header.
```

Characteristic	**Overall**, N = 222	**Female**, N = 142	**Male**, N = 80
assignment			
Control	108 (49%)	68 (48%)	40 (50%)
Treatment	114 (51%)	74 (52%)	40 (50%)
Age			
20-30	44 (20%)	25 (18%)	19 (24%)
30-40	83 (37%)	56 (39%)	27 (34%)
40-50	48 (22%)	32 (23%)	16 (20%)
50+	47 (21%)	29 (20%)	18 (22%)

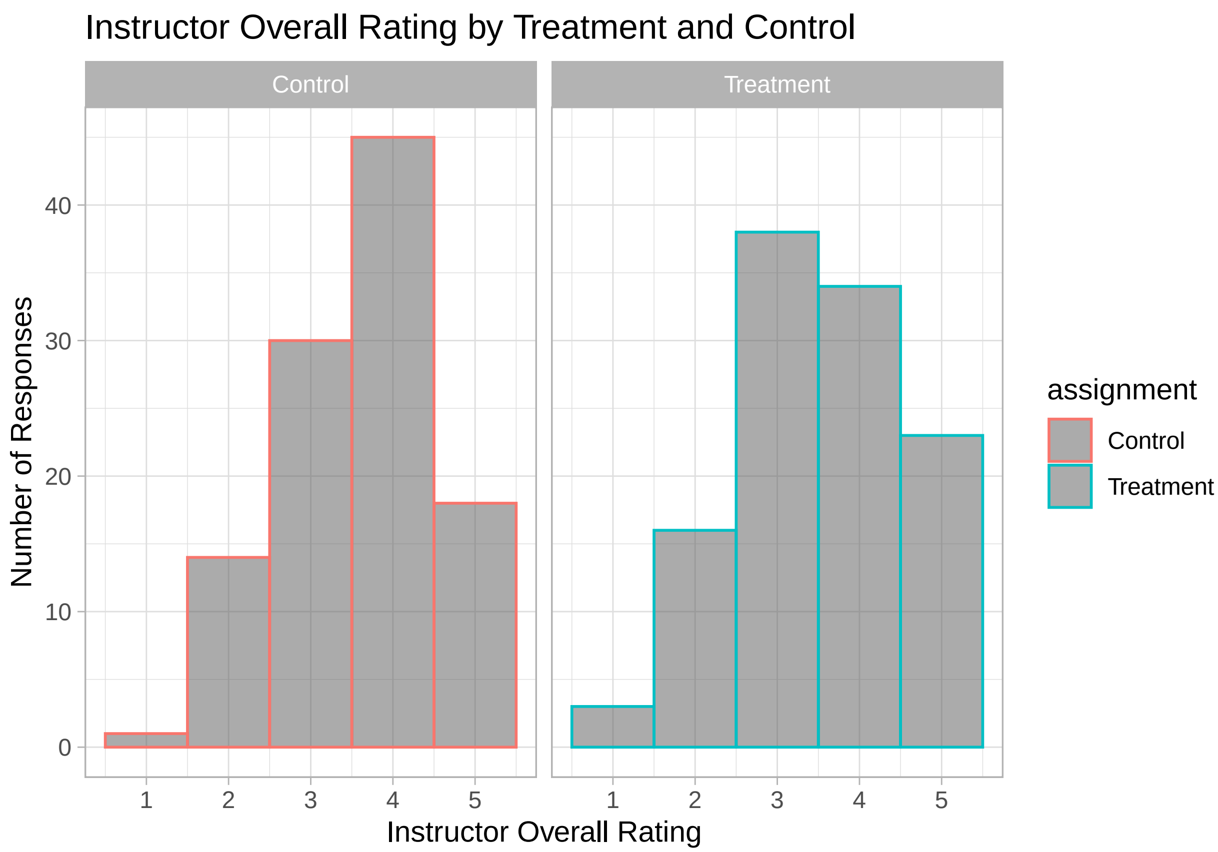
```
d %>%
  select(Gender, Education, assignment) %>%
  tbl_summary(by = assignment) %>%
  add_overall()
```

```
## Table printed with 'knitr::kable()', not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include 'message = FALSE' in code chunk header.
```

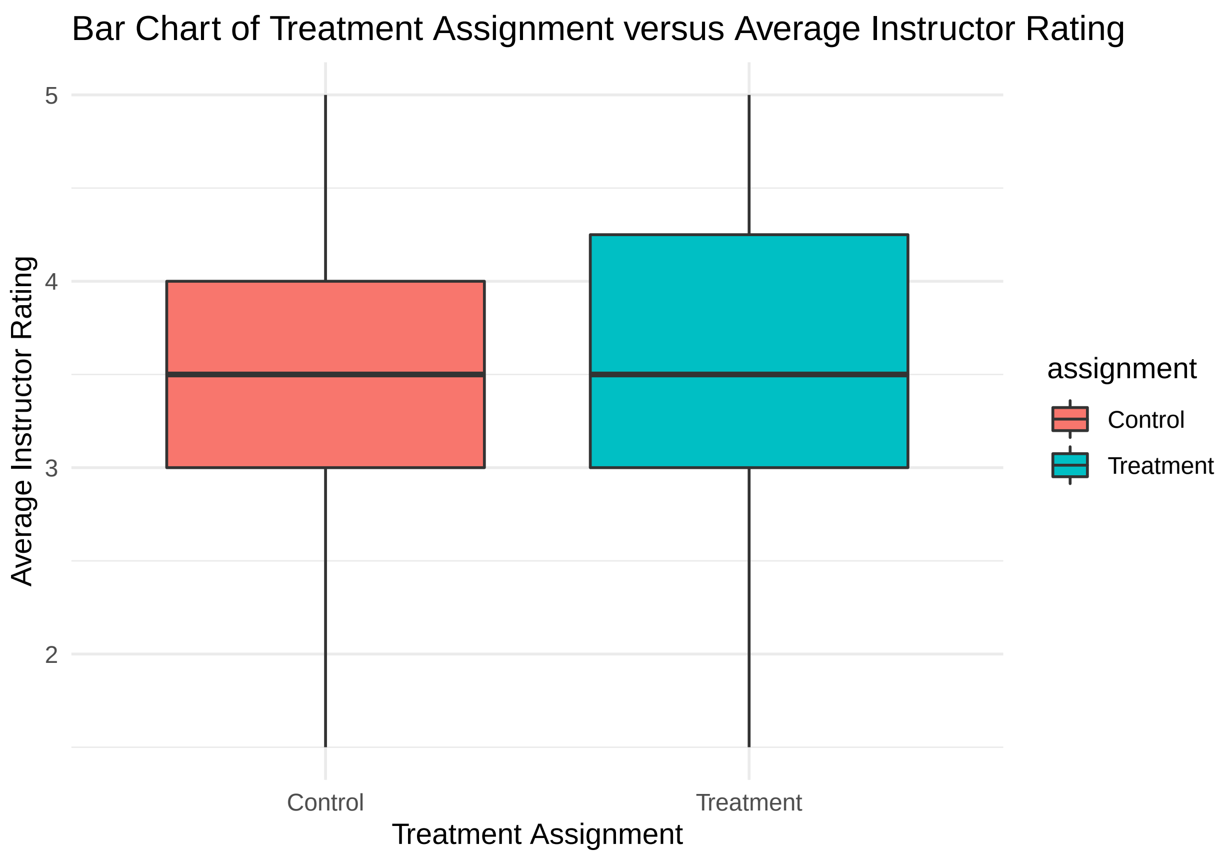
Characteristic	**Overall**, N = 222	**Control**, N = 108	**Treatment**, N = 114
Gender			
Female	142 (64%)	68 (63%)	74 (65%)
Male	80 (36%)	40 (37%)	40 (35%)
Education			
Associates degree	27 (12%)	16 (15%)	11 (9.6%)
Bachelors degree	44 (20%)	21 (19%)	23 (20%)
High school diploma	66 (30%)	33 (31%)	33 (29%)
Less than High school	7 (3.2%)	1 (0.9%)	6 (5.3%)
Masters degree	11 (5.0%)	5 (4.6%)	6 (5.3%)
Some College No degree	67 (30%)	32 (30%)	35 (31%)

```
response_hist <- d %>%
  ggplot() +
  aes(x = instr_rating, color=assignment) +
  geom_histogram(alpha = 0.5, bins=5) +
  labs(
    title  = 'Instructor Overall Rating by Treatment and Control',
    x      = 'Instructor Overall Rating',
    y      = 'Number of Responses'
  ) +
  theme_light() +
  facet_wrap(~ assignment)
```

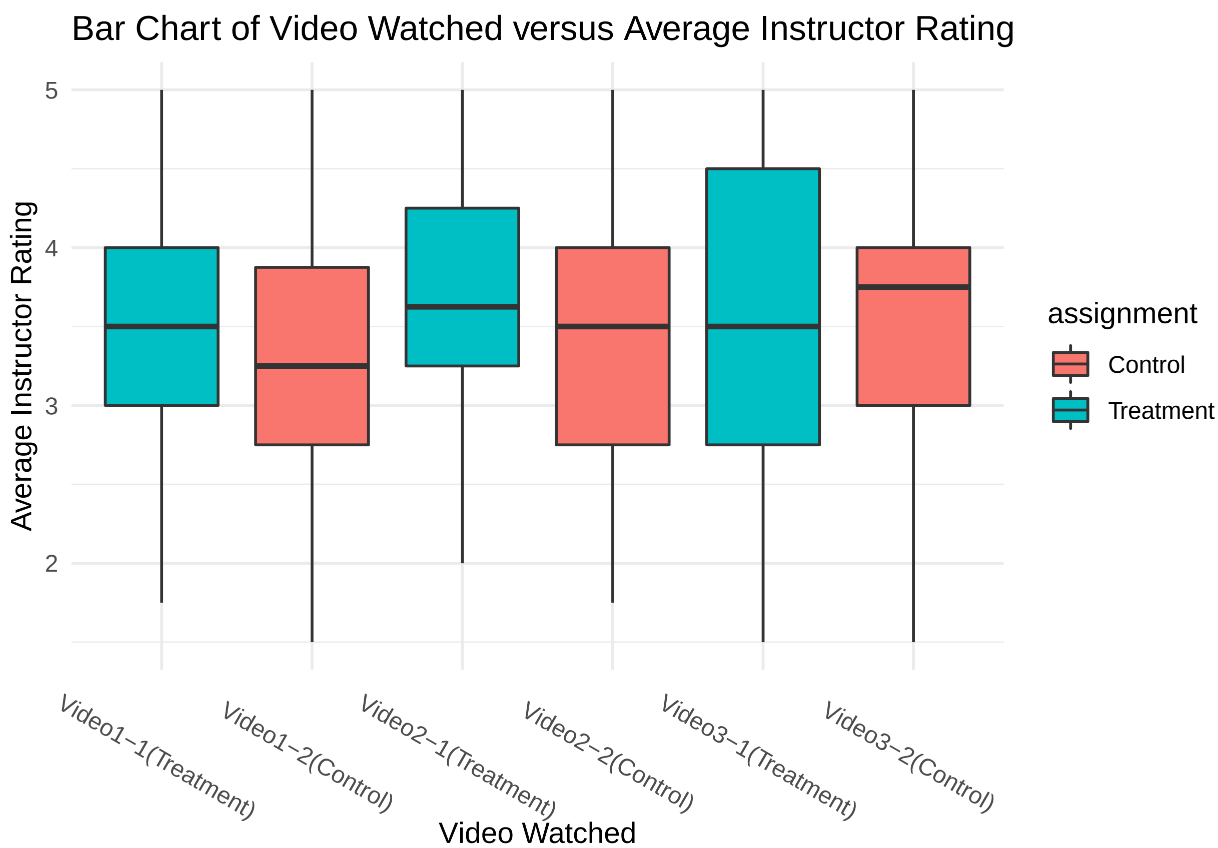
```
response_hist
```



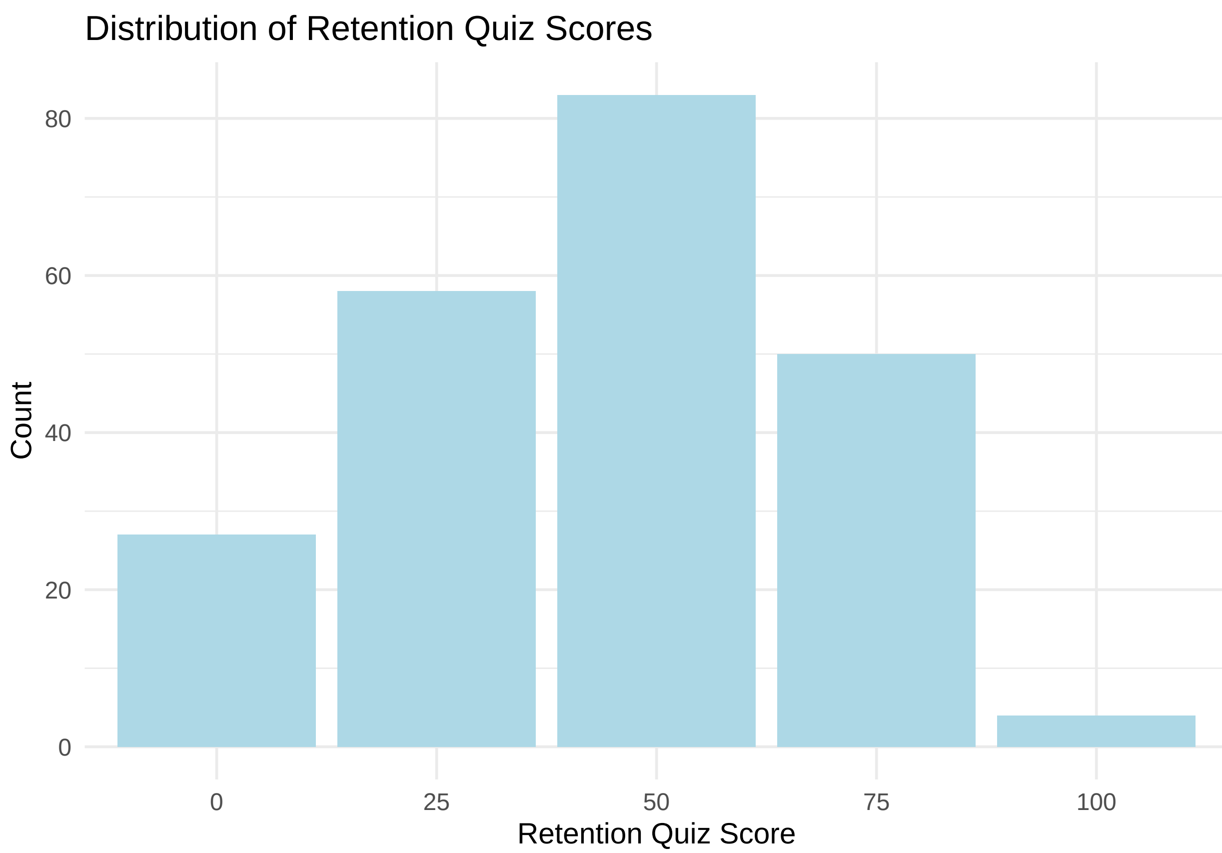

```
ggplot(d, aes(x=assignment, y=avg_rating, fill=assignment))+
  geom_boxplot()+
  labs(title="Bar Chart of Treatment Assignment versus Average Instructor Rating",
        x ="Treatment Assignment", y = "Average Instructor Rating")
```




```
ggplot(d, aes(x=video_watched, y=avg_rating, fill=assignment))+
  geom_boxplot()+
  labs(title="Bar Chart of Video Watched versus Average Instructor Rating",
        x ="Video Watched", y = "Average Instructor Rating")+theme(axis.text.x = element_text(angle = -3
```




```
d$quiz_grade <- as.factor(d$quiz_avg*100)
ggplot(d, aes(x=quiz_grade))+
geom_bar(aes(y = (..count..)),fill="lightblue")+
labs(title="Distribution of Retention Quiz Scores",
        x ="Retention Quiz Score", y = "Count")
```



Randomness Check

In this section, we verify if the block randomization was successful; there should be no increase in predicting whether or not someone is in the treatment or control group based on their blocks.

```
null_model <- lm(assigned~1,data=d)
full_model <- lm(assigned ~ 1 + as.factor(Age)+as.factor(Gender), data=d)
f_test <- anova(full_model, null_model, test ='F')
f_test
```

```
## Analysis of Variance Table
##
## Model 1: assigned ~ 1 + as.factor(Age) + as.factor(Gender)
## Model 2: assigned ~ 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    217 55.036
## 2    221 55.459 -4  -0.42338 0.4173 0.7961
```

```
# Separate data objects for treatment and control groups
d_treat <-
  d %>%
  filter(assignment == 'Treatment')

d_control <-
```

```r
  d %>%
  filter(assignment == 'Control')


# 2 sided t-test for each one of the instructor questions
instructor_enthusiams_ttest <- t.test(d_treat$instr_enthusiasm, d_control$instr_enthusiasm, conf.level =
instructor_enthusiams_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  d_treat$instr_enthusiasm and d_control$instr_enthusiasm
## t = 2.385, df = 214.8, p-value = 0.01795
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0629284 0.6622178
## sample estimates:
## mean of x mean of y
##  3.140351  2.777778
```

```r
instructor_professional_ttest <- t.test(d_treat$instr_professional, d_control$instr_professional, conf.l
instructor_professional_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  d_treat$instr_professional and d_control$instr_professional
## t = 0.38053, df = 213.09, p-value = 0.7039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1955552  0.2891224
## sample estimates:
## mean of x mean of y
##  3.824561  3.777778
```

```r
instructor_subject_ttest <- t.test(d_treat$instr_subject, d_control$instr_subject, conf.level = 0.95)
instructor_subject_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  d_treat$instr_subject and d_control$instr_subject
## t = 0.56564, df = 219.88, p-value = 0.5722
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1706972  0.3081241
## sample estimates:
## mean of x mean of y
##  3.929825  3.861111
```

```r
instructor_material_ttest <- t.test(d_treat$instr_material, d_control$instr_material, conf.level = 0.95)
instructor_material_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  d_treat$instr_material and d_control$instr_material
## t = 1.3029, df = 219.98, p-value = 0.194
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0909308  0.4457066
## sample estimates:
## mean of x mean of y
##  3.807018  3.629630
```

```r
instructor_overall_ttest <- t.test(d_treat$instr_rating, d_control$instr_rating, conf.level = 0.95)
instructor_overall_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  d_treat$instr_rating and d_control$instr_rating
## t = -0.69453, df = 219.48, p-value = 0.4881
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3572076  0.1710477
## sample estimates:
## mean of x mean of y
##  3.508772  3.601852
```

```r
model_0 <- lm(avg_rating ~ assignment, data=d)
stargazer(model_0, type='text')
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             avg_rating
## -----------------------------------------------
## assignmentTreatment           0.123
##                              (0.114)
##
## Constant                     3.447***
##                              (0.082)
##
## -----------------------------------------------
## Observations                   222
## R2                            0.005
## Adjusted R2                   0.001
## Residual Std. Error      0.848 (df = 220)
## F Statistic            1.176 (df = 1; 220)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
model_1 <- lm(avg_rating ~ as.factor(Age) + as.factor(Gender)+assignment, data=d)
model_2 <- lm(instr_rating ~ as.factor(Age) + as.factor(Gender)+ assignment, data=d)
model_3 <- lm(instr_subject ~ as.factor(Age) + as.factor(Gender)+assignment, data=d)
model_4 <- lm(instr_material ~ as.factor(Age) + as.factor(Gender)+assignment, data=d)
model_5 <- lm(instr_enthusiasm ~ as.factor(Age) + as.factor(Gender)+assignment , data=d)
model_6 <- lm(instr_enthusiasm ~ as.factor(Age) + as.factor(Gender)+assignment +as.factor(Gender)*assig
model_7 <- lm(avg_rating ~ as.factor(Age) + as.factor(Gender)+assignment +as.factor(Gender)*assignment,
model_8 <- lm(instr_rating ~ as.factor(Age) + as.factor(Gender)+assignment +as.factor(Gender)*assignmen
```

```r
#models with blocking across instructor ratings
stargazer(model_1, model_2, model_3,model_4, model_5, type = 'text')
```

```
##
## ==========================================================================================================
##                                                           Dependent variable:
##                                   ------------------------------------------------------------------------
##                                   avg_rating instr_rating instr_subject instr_material instr_enthusiasm
##                                      (1)          (2)          (3)           (4)              (5)
## ----------------------------------------------------------------------------------------------------------
## as.factor(Age)30-40                -0.117       -0.103       -0.210        -0.096            0.035
##                                    (0.158)      (0.187)      (0.169)       (0.191)          (0.207)
##
## as.factor(Age)40-50                -0.094       0.098        -0.048        0.003            -0.307
##                                    (0.177)      (0.209)      (0.189)       (0.213)          (0.232)
##
## as.factor(Age)50+                  -0.131       0.003        0.064         -0.035           -0.144
##                                    (0.178)      (0.210)      (0.190)       (0.214)          (0.233)
##
## as.factor(Gender)Male              0.230*       0.243*       0.110         0.221            0.472***
##                                    (0.119)      (0.140)      (0.127)       (0.143)          (0.155)
##
## assignmentTreatment                0.133        -0.081       0.070         0.186            0.364**
##                                    (0.114)      (0.135)      (0.122)       (0.137)          (0.149)
##
## Constant                          3.451***     3.525***     3.896***      3.589***         2.691***
##                                    (0.148)      (0.175)      (0.159)       (0.178)          (0.194)
##
## ----------------------------------------------------------------------------------------------------------
## Observations                         222          222          222           222              222
## R2                                  0.026        0.022        0.021         0.021            0.079
## Adjusted R2                         0.004        -0.001       -0.002        -0.002           0.058
## Residual Std. Error (df = 216)      0.846        1.000        0.905         1.018            1.107
## F Statistic (df = 5; 216)           1.170        0.970        0.915         0.919            3.726***
## ==========================================================================================================
## Note:                                                                   *p<0.1; **p<0.05; ***p<0.01
```

```r
#interaction term models
stargazer( model_6, model_7, model_8, type = 'text')
```

```
##
## ====================================================================================
##                                                           Dependent variable:
```

```
##                                              ----------------------------------------
##                                              instr_enthusiasm avg_rating instr_rating
##                                                    (1)            (2)          (3)
## ----------------------------------------------------------------------------------
## as.factor(Age)30-40                                0.034         -0.116       -0.102
##                                                   (0.208)        (0.159)      (0.188)
##
## as.factor(Age)40-50                               -0.315         -0.087        0.101
##                                                   (0.233)        (0.178)      (0.210)
##
## as.factor(Age)50+                                 -0.147         -0.129        0.004
##                                                   (0.233)        (0.178)      (0.211)
##
## as.factor(Gender)Male                              0.536**        0.180        0.220
##                                                   (0.222)        (0.169)      (0.200)
##
## assignmentTreatment                                0.409**        0.098       -0.097
##                                                   (0.187)        (0.143)      (0.169)
##
## as.factor(Gender)Male:assignmentTreatment         -0.127          0.098        0.045
##                                                   (0.312)        (0.238)      (0.282)
##
## Constant                                           2.670***       3.467***     3.532***
##                                                   (0.201)        (0.154)      (0.181)
##
## ----------------------------------------------------------------------------------
## Observations                                        222            222          222
## R2                                                 0.080          0.027        0.022
## Adjusted R2                                        0.054         -0.00003     -0.005
## Residual Std. Error (df = 215)                     1.110          0.848        1.002
## F Statistic (df = 6; 215)                          3.120***       0.999        0.809
## ==================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

```
# base model
model_0_2 <- lm(quiz_avg~ assignment, data=d)
model_1_2 <- lm(quiz_avg ~ as.factor(Age) + as.factor(Gender)+assignment, data=d)
model_2_2 <- lm(quiz_avg ~ as.factor(Age) + as.factor(Gender)+assignment +as.factor(Gender)*assignment,


#models with blocking across quiz averages
stargazer(model_1_2, model_2_2, type = 'text')
```

```
##
## ==================================================================================
##                                                        Dependent variable:
##                                              ----------------------------------------
##                                                             quiz_avg
##                                                      (1)                  (2)
## ----------------------------------------------------------------------------------
## as.factor(Age)30-40                                 -0.030               -0.030
##                                                     (0.046)              (0.046)
##
## as.factor(Age)40-50                                  0.033                0.032
```

```
##                                                 (0.051)           (0.052)
##
## as.factor(Age)50+                                0.095*            0.095*
##                                                 (0.052)           (0.052)
##
## as.factor(Gender)Male                            -0.050            -0.045
##                                                 (0.034)           (0.049)
##
## assignmentTreatment                              -0.023            -0.019
##                                                 (0.033)           (0.041)
##
## as.factor(Gender)Male:assignmentTreatment                          -0.010
##                                                                   (0.069)
##
## Constant                                        0.453***          0.451***
##                                                 (0.043)           (0.045)
##
## ----------------------------------------------------------------------------------
## Observations                                      222               222
## R2                                               0.046             0.046
## Adjusted R2                                      0.024             0.019
## Residual Std. Error                         0.246 (df = 216)   0.246 (df = 215)
## F Statistic                               2.080* (df = 5; 216) 1.728 (df = 6; 215)
## ==================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```