Glasses For Intelligence Impression

BA830 - Fall 2019

team 7

12/13/2019

Team 7 Members: Kunpeng Huang, Yoki Liu, Lyufan Pan, Yunlei Zhou, Jiayuan Zou, Sherry Zuo

Key Elements of Project Writeup When you write up your project paper, there is no predefined format. However, these factors must be included: a. Your causal question of interest.
b. Your experimental design and what data you collected.
c. Why you chose the experiment design.
d. Summary statistics about the data.
e. Treatment effects.
f. Limitations of your study.
Make sure the paper is well-written and that there are figures and tables that are easy to understand.

```
library(data.table)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```
library(lfe)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

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

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year

## The following object is masked from 'package:base':
##
##     date
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(gridExtra)
```

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

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

```r
library(naniar)
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```r
smart <- fread("survey.csv")
```

```r
#skimr::skim(smart)
#View(smart)
dim(smart)
```

```
## [1] 228  43
```

```r
#head(smart, 2)
```

```r
smart=smart%>%select(-Status, -DistributionChannel, -RecipientEmail, -RecipientFirstName, -RecipientLas
```

```r
#only keep finished records
smart=smart%>%filter(Finished=="True")%>%filter(Progress==100)
```

```r
smart=smart%>%select(-Progress, -Finished)
```

```r
length(unique(smart$ResponseId))
```

```
## [1] 226
```

```r
length(unique(smart$IPAddress))
```

```
## [1] 211
```

```r
#some people take more than once, we only want to keep each person once.
smart=smart%>%distinct(IPAddress, .keep_all = TRUE)
```

```r
# Add treatment/control dummy variable:
smart2 = smart %>%
  mutate(Treatment = ifelse(`Q4.1_First Click`=="", FALSE, TRUE))

# Variable selection:
smart_cleaned = smart2 %>%
  select(-c(`Q2.1_First Click`:`Q2.1_Click Count`, `Q3.1_First Click`:`Q3.1_Click Count`,
            `Q4.1_First Click`:`Q4.1_Click Count`, Q48, IPAddress, StartDate, EndDate, ResponseId)) %>%
  select(1:8,
         class_baseline=Q2.3, personality_baseline=Q2.4, grade_baseline=Q2.5,
         gender_p=Q5.1, class_p=Q5.2, personality_p=Q5.3, grade_p=Q5.4,
         everything())
```

```r
control = smart_cleaned %>%
  filter(Treatment == F) %>%
  select(-c(Q4.3, Q4.4, Q4.5), class=Q3.3, personality=Q3.4, grade=Q3.5)

treatment = smart_cleaned %>%
  filter(Treatment == T) %>%
  select(-c(Q3.3, Q3.4, Q3.5), class = Q4.3, personality=Q4.4, grade=Q4.5)

data = rbind(control, treatment)
#write_csv(data, "data.csv")
```

```r
data2<-read.csv("data.csv", na.strings = c("", "NA"))
```

```r
data2 = na.omit(data2)
```

```r
#write_csv(data2, "cleaned_data.csv")
```

```r
dim(data2)
```

```
## [1] 205  15
```

```r
# Convert GPA
data2$grade = ifelse(data2$grade == "A(3.75-4.0)", 9,
                     ifelse(data2$grade == "A-(3.5-3.75)", 8,
                            ifelse(data2$grade == "B+(3.25-3.5)", 7,
                                   ifelse(data2$grade == "B(3.0-3.25)", 6,
                                          ifelse(data2$grade == "B-(2.75-3.0)", 5,
                                                 ifelse(data2$grade == "C+(2.5-2.75)", 4,
                                                        ifelse(data2$grade == "C(2.25-2.5)", 3,
                                                               ifelse(data2$grade == "C-(2.0-2.25)", 2,
                                                                      ifelse(data2$grade == "Prefer not
                                                               ))))))))
table(data2$grade)
```

```
##
##  2  3  4  5  6  7  8  9
##  1  1  3  7 18 40 76 59
```

```r
# Convert GPA_P
data2$grade_p = ifelse(data2$grade_p == "A(3.75-4.0)", 9,
                     ifelse(data2$grade_p == "A-(3.5-3.75)", 8,
                            ifelse(data2$grade_p == "B+(3.25-3.5)", 7,
                                   ifelse(data2$grade_p == "B(3.0-3.25)", 6,
                                          ifelse(data2$grade_p == "B-(2.75-3.0)", 5,
                                                 ifelse(data2$grade_p == "C+(2.5-2.75)", 4,
                                                        ifelse(data2$grade_p == "C(2.25-2.5)", 3,
                                                               ifelse(data2$grade_p == "C-(2.0-2.25)", 2
                                                                      ifelse(data2$grade_p == "Prefer no
                                                                             1)))))))))
table(data2$grade_p)
```

```
##
##  3  4  5  6  7  8  9
##  2  3  5 30 33 79 42
```

```r
# Convert GPA_baseline
data2$grade_baseline = ifelse(data2$grade_baseline == "A(3.75-4.0)", 9,
                     ifelse(data2$grade_baseline == "A-(3.5-3.75)", 8,
                            ifelse(data2$grade_baseline == "B+(3.25-3.5)", 7,
                                   ifelse(data2$grade_baseline == "B(3.0-3.25)", 6,
                                          ifelse(data2$grade_baseline == "B-(2.75-3.0)", 5,
                                                 ifelse(data2$grade_baseline == "C+(2.5-2.75)", 4,
                                                        ifelse(data2$grade_baseline == "C(2.25-2.5)", 3
                                                               ifelse(data2$grade_baseline == "C-(2.0-2
                                                                      ifelse(data2$grade_baseline == "P
                                                                             1)))))))))
table(data2$grade_baseline)
```

```
##
##   1   4   5   6   7   8   9
##   1   1   2  19  55  96  31
```

```
data2$gender_p = ifelse(data2$gender_p == "Male", 1,
                 ifelse(data2$gender_p == "Female", 0, NA))

table(data2$gender_p)
```

```
##
##   0   1
## 130  67
```

```
data2$personality_baseline = ifelse(data2$personality_baseline == "Logical", 1,
                             ifelse(data2$personality_baseline == "Emotional", 0, NA))

table(data2$personality_baseline)
```

```
##
##   0   1
##  79 126
```

```
data2$personality_p = ifelse(data2$personality_p == "Logical", 1,
                             ifelse(data2$personality_p == "Emotional", 0, NA))

table(data2$personality_p)
```

```
##
##  0  1
## 69 84
```

```
data2$personality = ifelse(data2$personality == "Logical", 1,
                             ifelse(data2$personality == "Emotional", 0, NA))

table(data2$personality)
```

```
##
##   0   1
##  53 152
```

```
#save final cleaned version of data
#write_csv(data2, "data2.csv")
```

```
analysis<-fread("data2.csv")
```

```
summary(analysis)
```

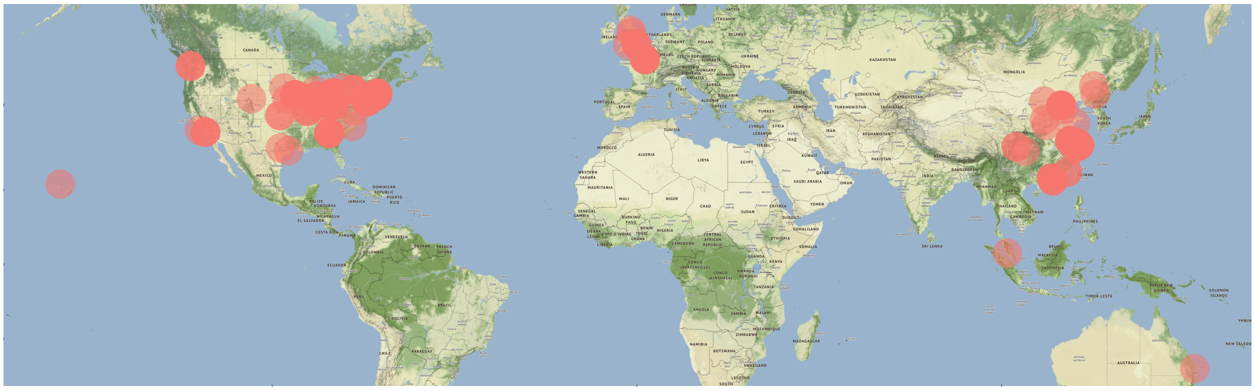
```
##  Duration..in.seconds. RecordedDate       LocationLatitude
##  Min.   :   35.00      Length:205         Min.   :-27.473
```

```
##  1st Qu.:   62.00      Class :character   1st Qu.: 34.065
##  Median :   82.00      Mode  :character   Median : 42.037
##  Mean   :  312.92                         Mean   : 38.942
##  3rd Qu.:  116.00                         3rd Qu.: 42.347
##  Max.   :17896.00                         Max.   : 53.796
##
##  LocationLongitude   class_baseline     personality_baseline
##  Min.   :-158.0419   Length:205         Min.   :0.00000
##  1st Qu.: -83.3713   Class :character   1st Qu.:0.00000
##  Median : -71.0975   Mode  :character   Median :1.00000
##  Mean   : -27.1948                      Mean   :0.61463
##  3rd Qu.:   2.3387                      3rd Qu.:1.00000
##  Max.   : 153.0215                      Max.   :1.00000
##
##  grade_baseline      class              gender_p        class_p
##  Min.   :1.0000   Length:205         Min.   :0.0000   Length:205
##  1st Qu.:7.0000   Class :character   1st Qu.:0.0000   Class :character
##  Median :8.0000   Mode  :character   Median :0.0000   Mode  :character
##  Mean   :7.6146                      Mean   :0.3401
##  3rd Qu.:8.0000                      3rd Qu.:1.0000
##  Max.   :9.0000                      Max.   :1.0000
##                                      NA's   :8
##  personality_p        grade_p        personality          grade
##  Min.   :0.00000   Min.   :3.0000   Min.   :0.00000   Min.   :2.0000
##  1st Qu.:0.00000   1st Qu.:7.0000   1st Qu.:0.00000   1st Qu.:7.0000
##  Median :1.00000   Median :8.0000   Median :1.00000   Median :8.0000
##  Mean   :0.54902   Mean   :7.5464   Mean   :0.74146   Mean   :7.7024
##  3rd Qu.:1.00000   3rd Qu.:8.0000   3rd Qu.:1.00000   3rd Qu.:9.0000
##  Max.   :1.00000   Max.   :9.0000   Max.   :1.00000   Max.   :9.0000
##  NA's   :52        NA's   :11
##  Treatment
##  Mode :logical
##  FALSE:102
##  TRUE :103
##
##
##
##
```

```r
names(analysis)
```

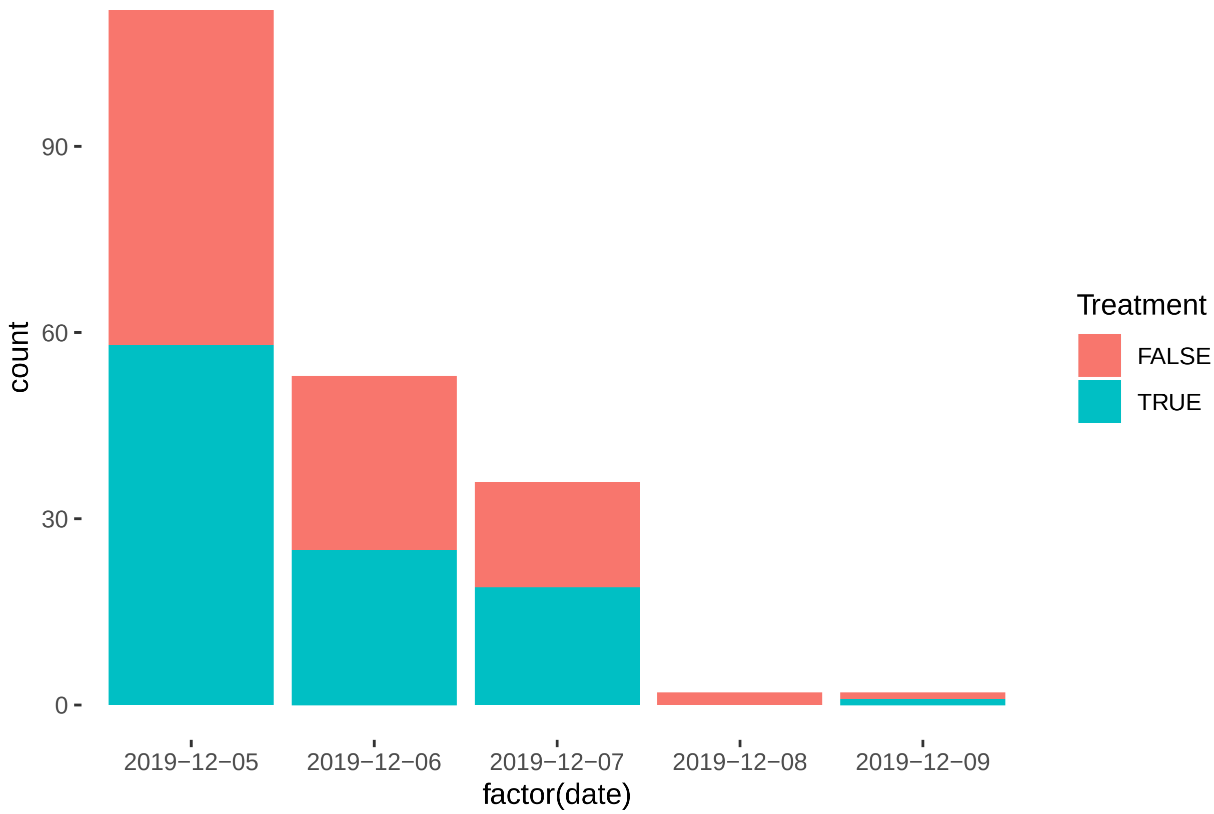
```
##  [1] "Duration..in.seconds." "RecordedDate"
##  [3] "LocationLatitude"      "LocationLongitude"
##  [5] "class_baseline"        "personality_baseline"
##  [7] "grade_baseline"        "class"
##  [9] "gender_p"              "class_p"
## [11] "personality_p"         "grade_p"
## [13] "personality"           "grade"
## [15] "Treatment"
```

```r
qmplot(x = LocationLongitude, y = LocationLatitude, data = analysis, colour = 'blue', alpha=0.4,  size =
  theme(legend.position="")
```

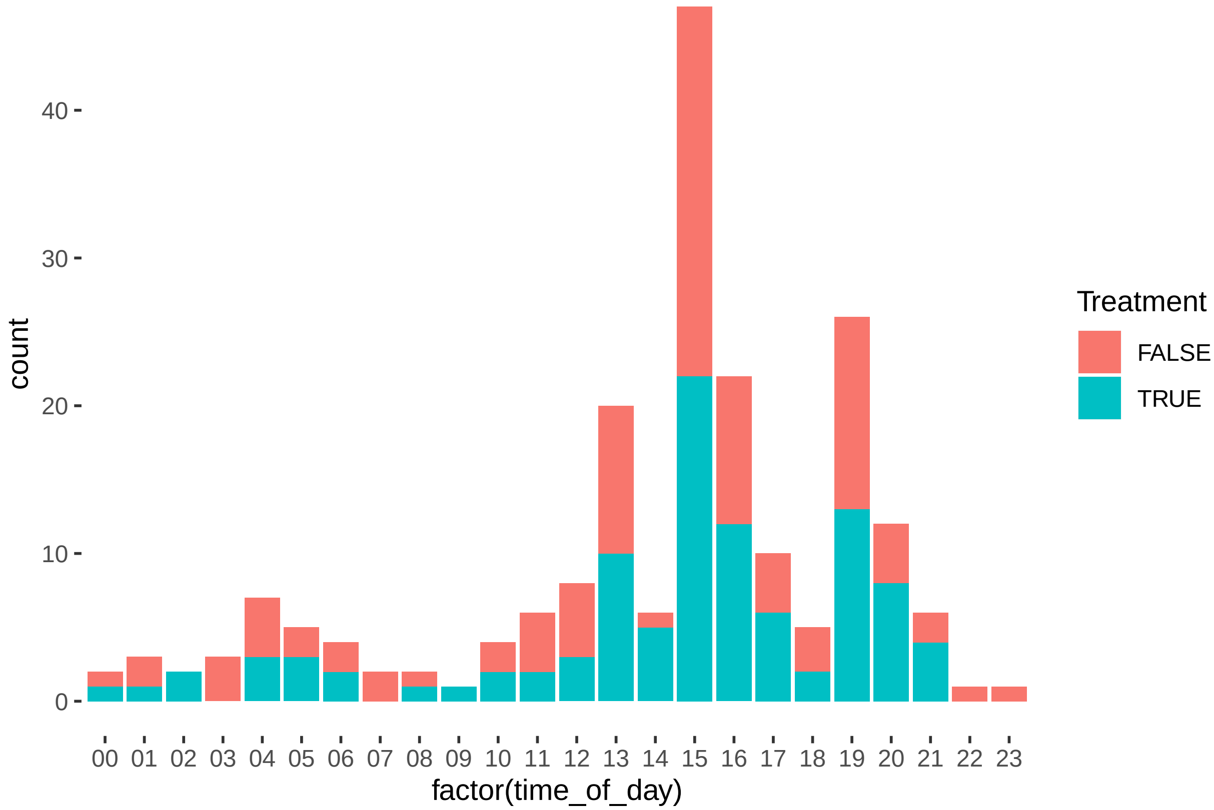


```
analysis[, time_of_day := substr(RecordedDate, 12, 13)]
analysis[, date := as_date(RecordedDate)]
```

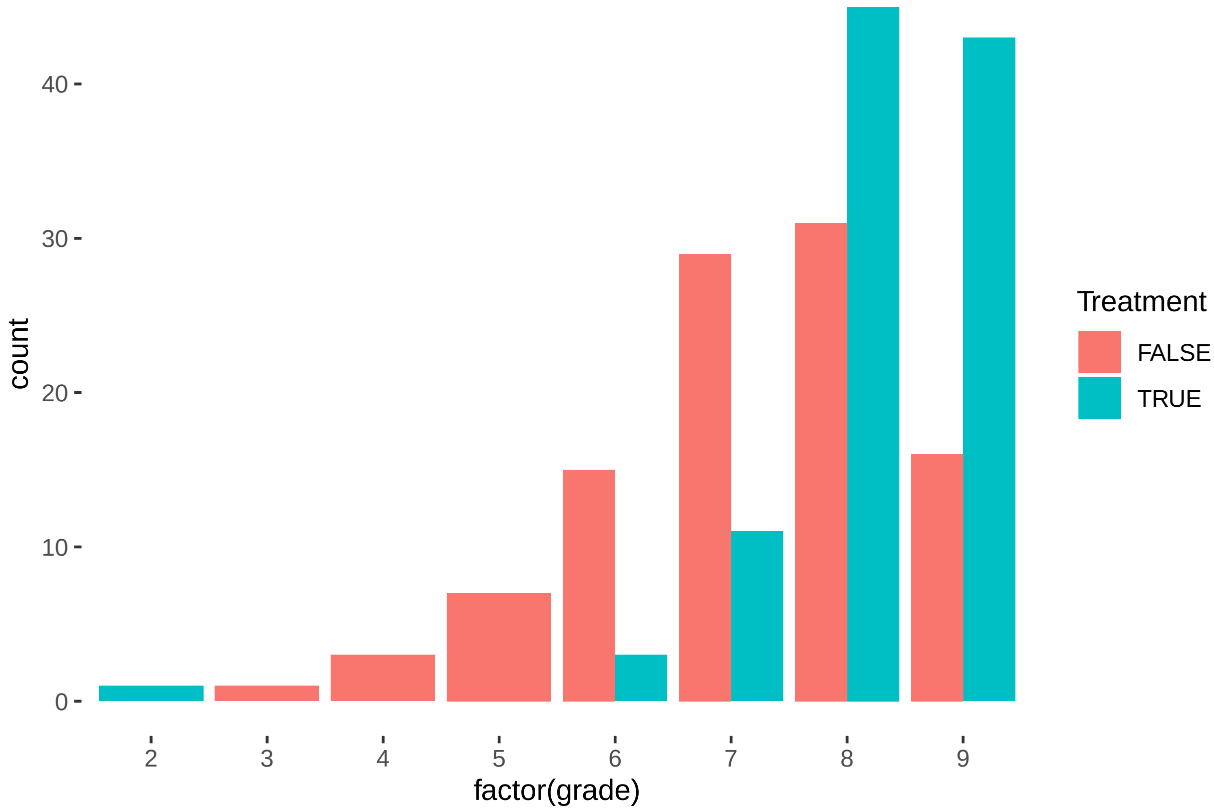
```
ggplot(analysis, aes(x=factor(date)))+
  geom_bar(aes(fill=Treatment))+
  theme(panel.background = element_rect(fill="white"))
```




```
ggplot(analysis, aes(x=factor(time_of_day)))+
   geom_bar(aes(fill=Treatment))+
   theme(panel.background = element_rect(fill="white"))
```




```r
ggplot(analysis, aes(x=factor(grade), fill=Treatment))+
  geom_bar(position = "dodge")+
  theme(panel.background = element_rect(fill="white"))
```



##Randomization Check/Balance Check

```
summary(lm(grade_baseline~Treatment, analysis))
```

```
##
## Call:
## lm(formula = grade_baseline ~ Treatment, data = analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.62136 -0.62136  0.37864  0.39216  1.39216
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.607843   0.101792 74.7392   <2e-16 ***
## TreatmentTRUE 0.013516   0.143606  0.0941   0.9251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.028 on 203 degrees of freedom
## Multiple R-squared:  4.3636e-05, Adjusted R-squared:  -0.0048823
## F-statistic: 0.0088585 on 1 and 203 DF,  p-value: 0.92511
```

Since the p-value is 0.9251 which is larger than 0.05, the R-square is -0.0048823 which is too small, so we fail to reject the null hypothesis, thus we think the ATE is 0, so we have 95% confidence that the glasses treatment on grade taken before the experiment started does not have effect. So this results indicate that the validity of the study's conclusion is high.

##Causal Effect

```
summary(lm(grade~Treatment, data=analysis))
```

```
##
## Call:
## lm(formula = grade ~ Treatment, data = analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.19417 -0.20588 -0.19417  0.80583  1.79412
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    7.20588    0.11520  62.551 < 2.2e-16 ***
## TreatmentTRUE  0.98829    0.16252   6.081 5.838e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1635 on 203 degrees of freedom
## Multiple R-squared:  0.15409,    Adjusted R-squared:  0.14992
## F-statistic: 36.978 on 1 and 203 DF,  p-value: 5.8378e-09
```

```
lowerci<-0.98829-1.96*0.16252
lowerci
```

```
## [1] 0.6697508
```

```
upperci<-0.98829+1.96*0.16252
upperci
```

```
## [1] 1.3068292
```

The ATE of treatment on grade is **0.98829**, the 95% confidence interval is [**0.6697508, 1.3068292**].

##Fixed Effect Check ###Grade_Baseline

```
summary(felm(grade~grade_baseline, data=analysis))
```

```
##
## Call:
##    felm(formula = grade ~ grade_baseline, data = analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.31903 -0.87395  0.12605  0.68097  2.46126
##
## Coefficients:
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    4.313387   0.618596  6.9729 4.278e-11 ***
## grade_baseline 0.445071   0.080514  5.5278 9.889e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1794 on 203 degrees of freedom
## Multiple R-squared(full model): 0.13083   Adjusted R-squared: 0.12655
## Multiple R-squared(proj model): 0.13083   Adjusted R-squared: 0.12655
## F-statistic(full model):30.557 on 1 and 203 DF, p-value: 9.8894e-08
## F-statistic(proj model): 30.557 on 1 and 203 DF, p-value: 9.8894e-08
```

Since the p-value is less than 0.05, so we reject the null hypothesis, so we have 95% confidence that the baseline has effect for the grade, that's what we want to do baseline during our survey. Thus, it's a fixed effect in our result.

###Personal Gender

```
with(analysis, table(gender_p, grade))
```

```
##          grade
## gender_p  2  3  4  5  6  7  8  9
##         0  0  0  2  3  7 23 56 39
##         1  0  0  1  3 11 16 19 17
```

```
male_grade<-analysis[gender_p==1, grade]
table(male_grade)
```

```
## male_grade
##  4  5  6  7  8  9
##  1  3 11 16 19 17
```

```
female_grade<-analysis[gender_p==0, grade]
table(female_grade)
```

```
## female_grade
##  4  5  6  7  8  9
##  2  3  7 23 56 39
```

```
t.test(male_grade, female_grade)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_grade and female_grade
## t = -2.1912, df = 116.782, p-value = 0.030423
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.746452065 -0.037704077
## sample estimates:
## mean of x mean of y
## 7.4925373 7.8846154
```

With the t-test, since the p-value is 0.030423 which is less than 0.05, so we reject the null hypothesis, so we have 95% confidence that the gender of each subject has an effect on the grade. Thus, it's a fixed effect on our results.

###Personal Personality

```r
with(analysis, table(personality_p, grade))
```

```
##              grade
## personality_p  2  3  4  5  6  7  8  9
##             0  0  0  1  2  5 14 21 26
##             1  1  1  1  3  9 20 30 19
```

```r
logical_grade<-analysis[personality_p==1, grade]
table(logical_grade)
```

```
## logical_grade
##  2  3  4  5  6  7  8  9
##  1  1  1  3  9 20 30 19
```

```r
emotional_grade<-analysis[personality_p==0, grade]
table(emotional_grade)
```

```
## emotional_grade
##  4  5  6  7  8  9
##  1  2  5 14 21 26
```

```r
t.test(logical_grade, emotional_grade)
```

```
##
##  Welch Two Sample t-test
##
## data:  logical_grade and emotional_grade
## t = -1.92402, df = 150.802, p-value = 0.056235
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.802585092  0.010659626
## sample estimates:
## mean of x mean of y
## 7.4880952 7.8840580
```

With the t-test, since the p-value is 0.056235 which is larger than 0.05, so we fail to reject the null hypothesis, so we don't have 95% confidence that the personality of each subject has an effect on the grade. Thus, it's not a fixed effect in our result.

###Class

```r
analysis$class_baseline<-factor(analysis$class_baseline, levels=c("Freshman", "Sophomore", "Junior", "Se
p1<-ggplot(analysis, aes(x=class_baseline, fill=factor(Treatment)))+
  geom_bar()+
  theme(panel.background = element_rect(fill="white"))
analysis$class<-factor(analysis$class, levels=c("Freshman", "Sophomore", "Junior", "Senior", "Graduates
p2<-ggplot(analysis, aes(x=class, fill=factor(Treatment)))+
  geom_bar()+
  theme(panel.background = element_rect(fill="white"))
analysis$class_p<-factor(analysis$class_p, levels=c("Freshman", "Sophomore", "Junior", "Senior", "Gradua
```

```r
p3<-ggplot(analysis, aes(x=class_p, fill=factor(Treatment)))+
  geom_bar()+
  theme(panel.background = element_rect(fill="white"))
grid.arrange(p1,p2,p3, nrow=3, ncol=1)
```

While comparing the distributions of the class of photo A (baseline), photo B, the subject itself, we find that although most of the subjects are graduates, their judgment on photo B is still almost evenly distributed. Hence, we don't believe that it's a fixed effect.

##Final Model

```r
summary(felm(grade~Treatment | grade_baseline + gender_p, data=analysis))
```

```
##
## Call:
##    felm(formula = grade ~ Treatment | grade_baseline + gender_p,      data = analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86147 -0.57468  0.13280  0.55207  1.85032
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)
## TreatmentTRUE  1.00573    0.13825   7.275 8.977e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.95269 on 189 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared(full model): 0.33205   Adjusted R-squared: 0.30731
## Multiple R-squared(proj model): 0.21877   Adjusted R-squared: 0.18983
## F-statistic(full model):13.422 on 7 and 189 DF, p-value: 5.028e-14
## F-statistic(proj model): 52.925 on 1 and 189 DF, p-value: 8.9765e-12
```

Based on our fixed effect check, we would like to estimate those fixed effects of grade_baseline and gender_p in our final regression to find the effect of classes for letter grade impression. There is a smaller p-value which is less than 0.05, so the treatment effect on glasses with fixed effects is statistically significant. Here, we could find after considering fixed effects, compared with the original causal effect model, this model has a smaller standard error, a larger adjusted R-squared. It's more precise. Thus, we believe that the causal effect model with the fixed effect would be a better model.

##Limitation

```
summary(felm(grade~Treatment+personality | grade_baseline + gender_p, data=analysis))
```

```
##
## Call:
##    felm(formula = grade ~ Treatment + personality | grade_baseline +      gender_p, data = analysis)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.951298 -0.672204 -0.009534  0.578417  1.757184
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## TreatmentTRUE  0.66267    0.12955  5.1153 7.69e-07 ***
## personality    1.11184    0.14692  7.5676 1.65e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8363 on 188 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared(full model): 0.48801   Adjusted R-squared: 0.46622
## Multiple R-squared(proj model): 0.40118   Adjusted R-squared: 0.3757
## F-statistic(full model):22.399 on 8 and 188 DF, p-value: < 2.22e-16
## F-statistic(proj model): 62.975 on 2 and 188 DF, p-value: < 2.22e-16
```

We have one optional question in the end of our survey to ask whether people can detect our purpose or not, and they can take a guess. We could try this if we have more time.

```
guess <- read_csv("glasses.csv")
```

```
## Parsed with column specification:
## cols(
##   `Can you guess the purpose of our experiment? (optional)` = col_character()
## )
```

```
names(guess)
```

```
## [1] "Can you guess the purpose of our experiment? (optional)"
```

```r
names(guess)<-"purpose_guess"
```