Data Analysis

Team 7

December 10, 2021

Data Cleaning

```
# removing the header
df = df[c(-1,-2),]
# Renaming columns
df <- df%>%rename(Age=`Â `,Gender=Q2,Education=Q3,subgroup = FL_49_DO)
# Keeping only the finished surveys
df_finished = df%>%filter(Finished == 'True')
# making the blank cells to be the control group
df_finished$subgroup[df_finished$subgroup == ""] <- 'Control'
df_finished$group<- ifelse(grepl('^Treatment', df_finished$subgroup), 'Treatment', 'Control')
# extract out the dollar values
df_finished$willing_to_pay <- str_extract(df_finished$Q16, '\\$(\\d+)')
# Remove the dollar sign
df_finished$willing_to_pay <- (gsub("\\$", "", df_finished$willing_to_pay))
# Add 0 to the NA  values
df_finished$willing_to_pay[is.na(df_finished$willing_to_pay)] <- 0
#Create binary outcome
df_finished$is_willing_to_pay<- ifelse(df_finished$willing_to_pay>0,1,0)
#Create binary indicators for age
df_finished$equal_over35<- ifelse(df_finished$Age=="35 or more than 35",1,0)
df_finished$equal_over30<- ifelse(df_finished$Age=="30-34" | df_finished$Age=="35 or more than 35" ,1,0
# Select just the columns needed
df_finished <- df_finished %>%
              select(ResponseId,Age,Gender,Education,subgroup,group,equal_over35,equal_over30,willing_to
              mutate(willing_to_pay = as.numeric(willing_to_pay))
```

```
control_group = df_finished%>%filter(group == "Control")
treatment_group = df_finished%>%filter(group == "Treatment")
```

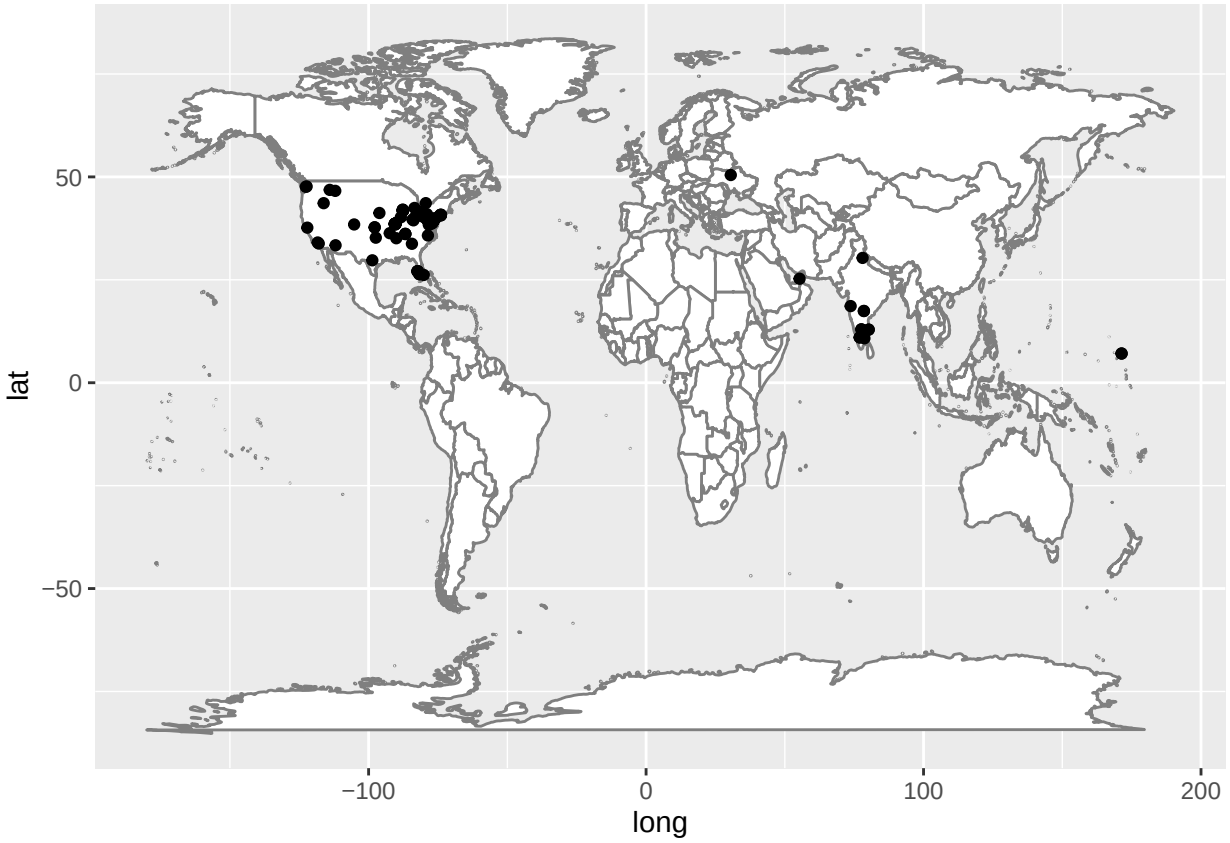
Survey Metadata

- We have collected so far 194 surveys.
- These include 167 finished surveys.
- We have 32 subjects in the control group and 135 subjects in the treatment groups

Where Survey Participants Are Coming From

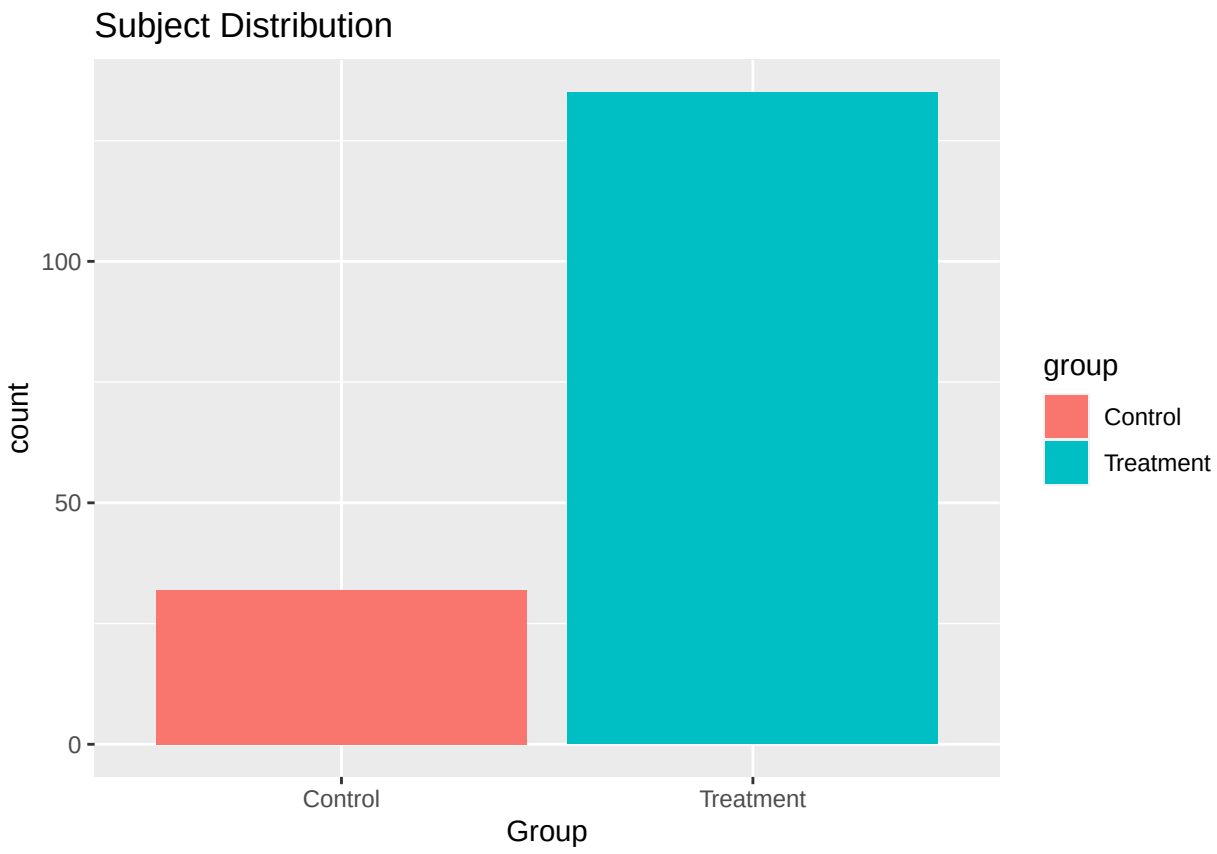
```
map_df <- df%>%select(LocationLatitude,LocationLongitude)
mapWorld <- borders("world", colour="gray50", fill="white")
mp <- ggplot() + mapWorld
mp + geom_point(data = map_df, aes(x =as.numeric(LocationLongitude), y=as.numeric(LocationLatitude)))
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```



Distribution among Treatment and Control Groups

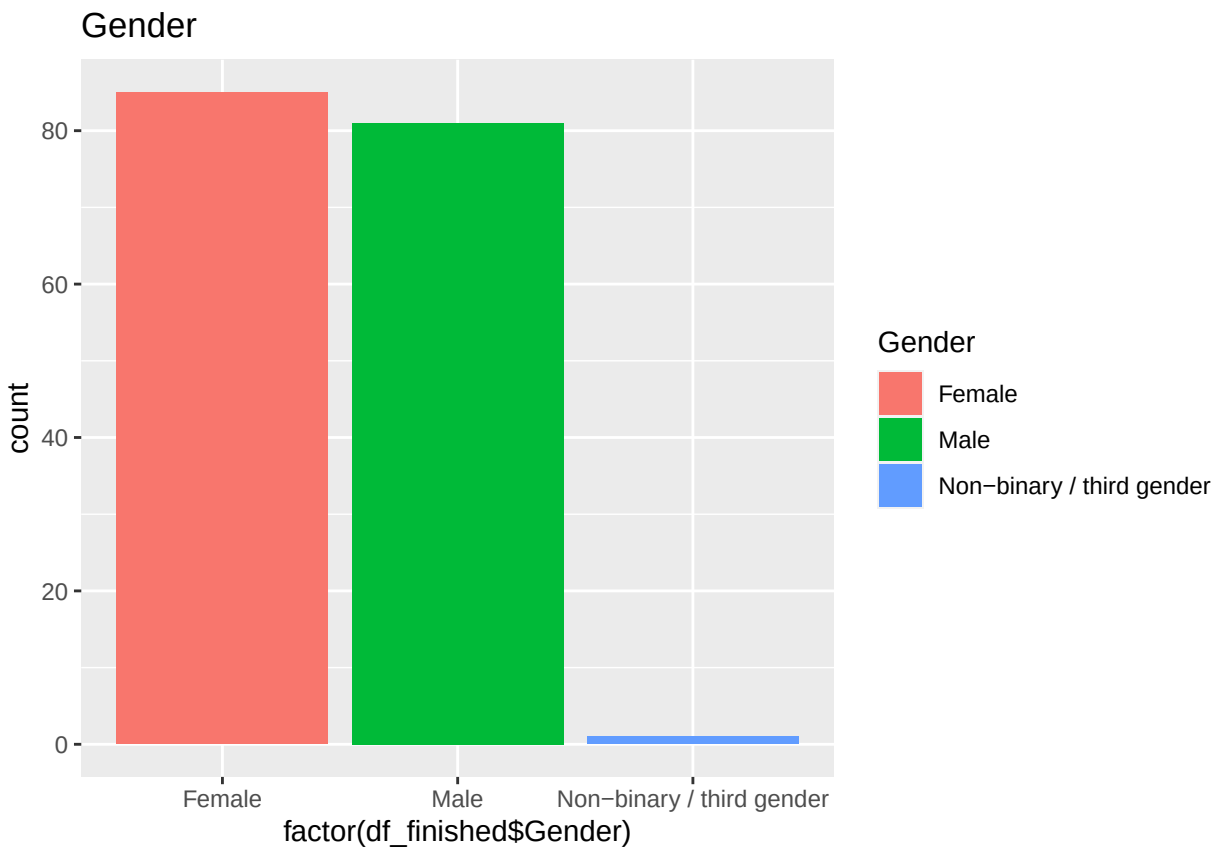
```
ggplot(data = df_finished)+
  aes(x = factor(group),fill = group)+
  geom_bar(stat = "count")+ggtitle("Subject Distribution")+xlab("Group")
```



Demographic Distribution

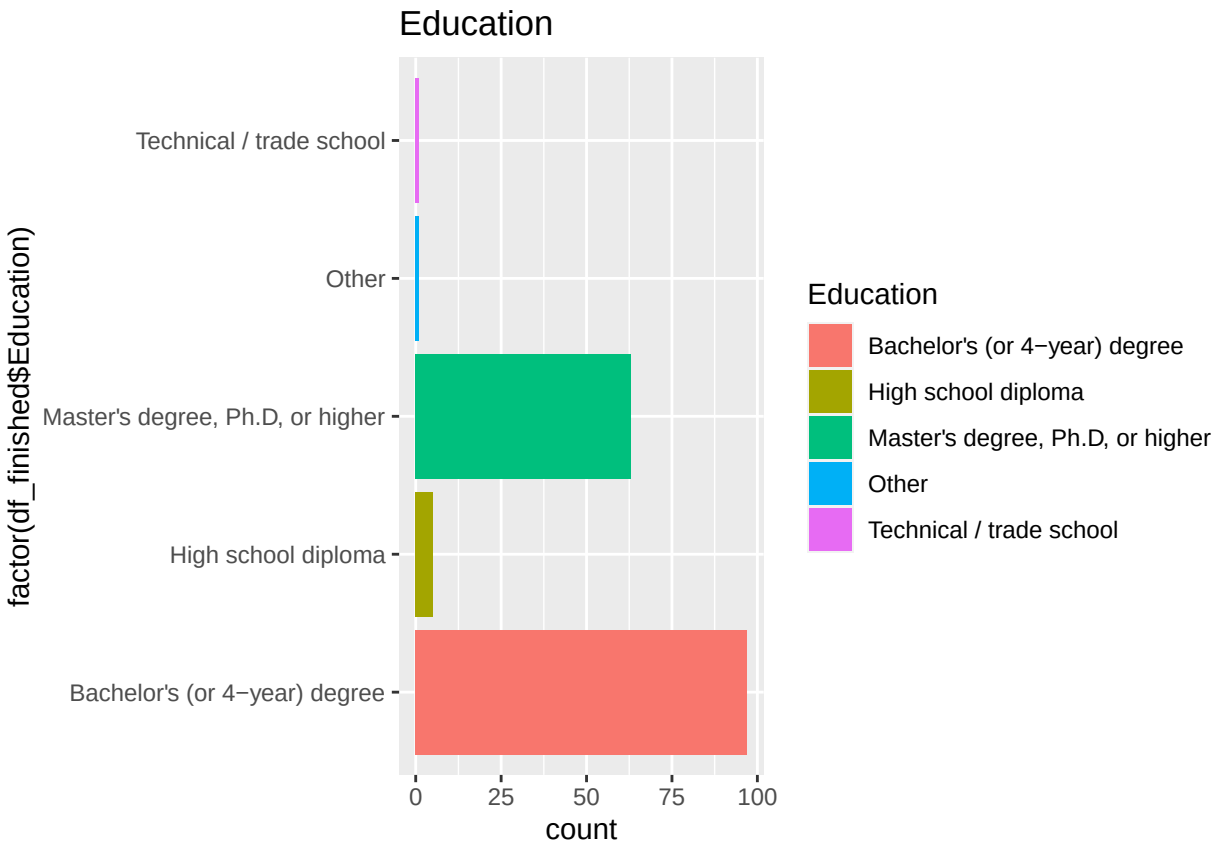
Gender

```
ggplot(data = df_finished)+
aes(x = factor(df_finished$Gender),fill = Gender)+
geom_bar(stat = "count")+ggtitle("Gender")
```



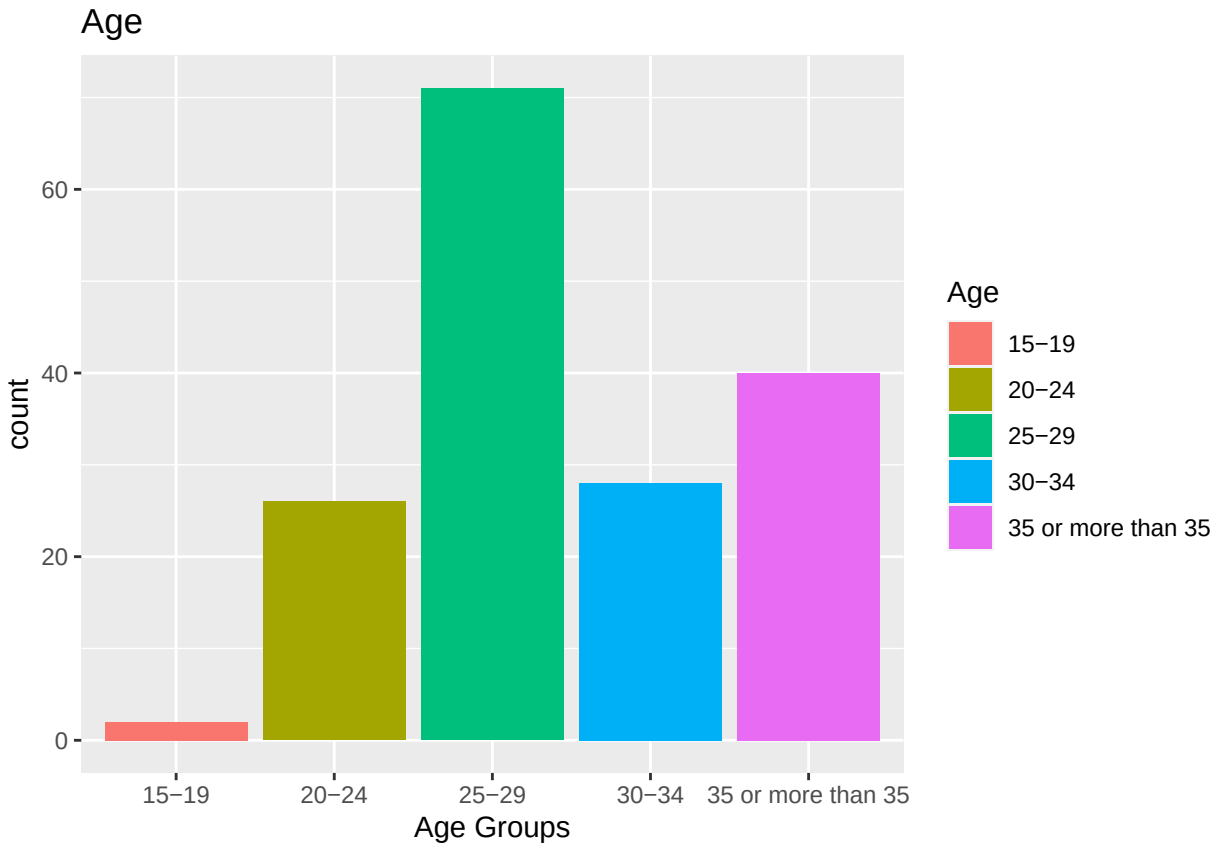
Education

```
ggplot(data = df_finished)+
  aes(x = factor(df_finished$Education),fill = Education)+
  geom_bar(stat = "count")+ggtitle("Education")+coord_flip()
```



Age Distribution

```
ggplot(data = df_finished)+
  aes(x = factor(df_finished$Age),fill= Age)+
  geom_bar(stat = "count")+ggtitle("Age")+xlab("Age Groups")
```



Testing to see if there is a Association between Gender and being in the control group

```
# Is the Gender distribution in the treatment and control group the same or different
# Ran a Fisher Test since the  the smallest expected frequency is lower than 5
df <- table(df_finished$Gender, df_finished$group)
fisher.test(df)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df
## p-value = 0.02789
## alternative hypothesis: two.sided
```

```
# seems like there is an association between gender and being the control or treatment group
```

Comparing the Control and Treatment group willingness to pay

Between The Control Group and All Treatment Groups

```
(group_summary <- df_finished%>%
                  group_by(group)%>%summarize(`avg willingness in dollars`=mean(willing_to_pay),
                                              `sd willing to pay in dollars` = sd(willing_to_pay),
                                              lower = t.test(willing_to_pay)$conf.int[1],
                                              upper = t.test(willing_to_pay)$conf.int[2]))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 5
##   group     `avg willingness in dollar~ `sd willing to pay in dolla~ lower upper
##   <chr>                           <dbl>                        <dbl> <dbl> <dbl>
## 1 Control                         0.906                         1.20 0.473  1.34
## 2 Treatment                       0.919                         1.22 0.711  1.13
```

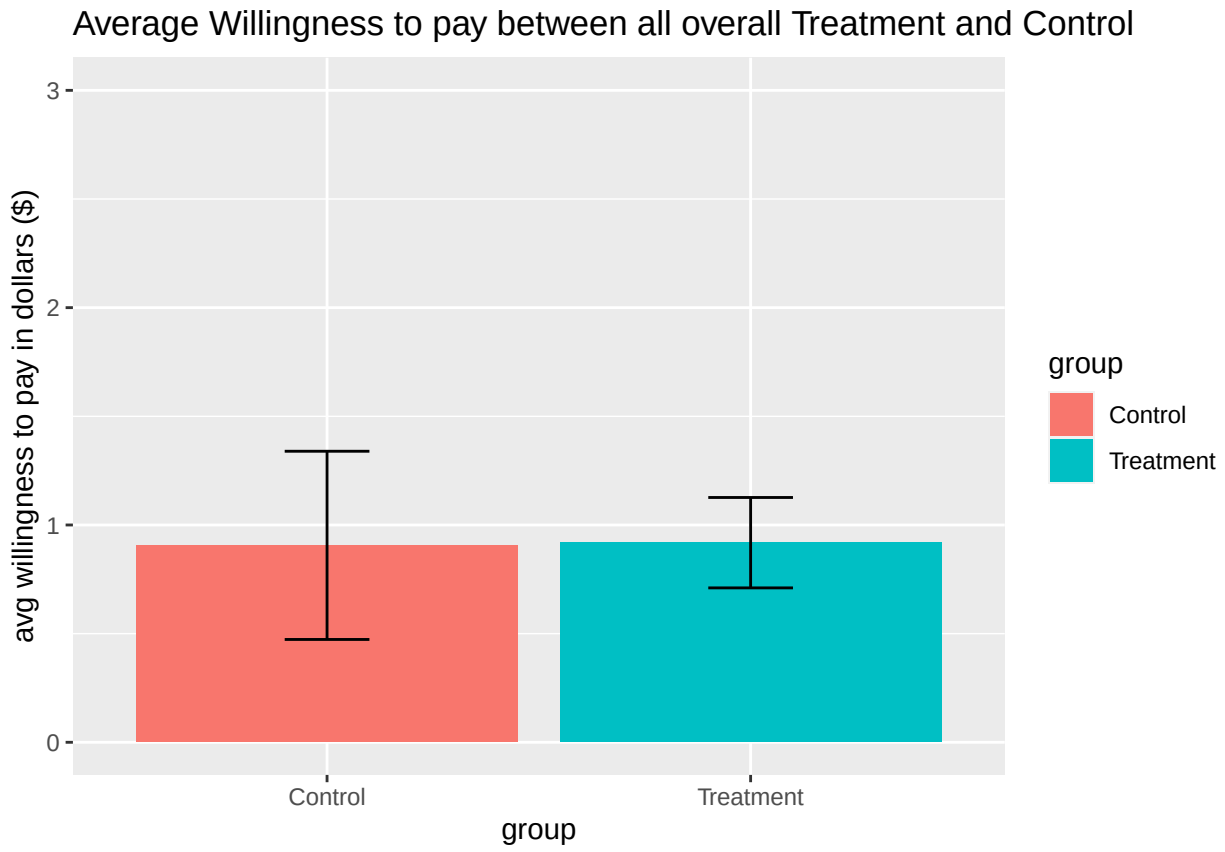
Betweem The Control and Treatment Sub Groups

```
(sub_group_summary <-  df_finished%>%
                       group_by(subgroup)%>%
                       summarize(`avg willingness in dollars`=mean(willing_to_pay),
                                 `sd willing to pay in dollars` = sd(willing_to_pay),
                                 lower = t.test(willing_to_pay)$conf.int[1],
                                 upper = t.test(willing_to_pay)$conf.int[2]))
```

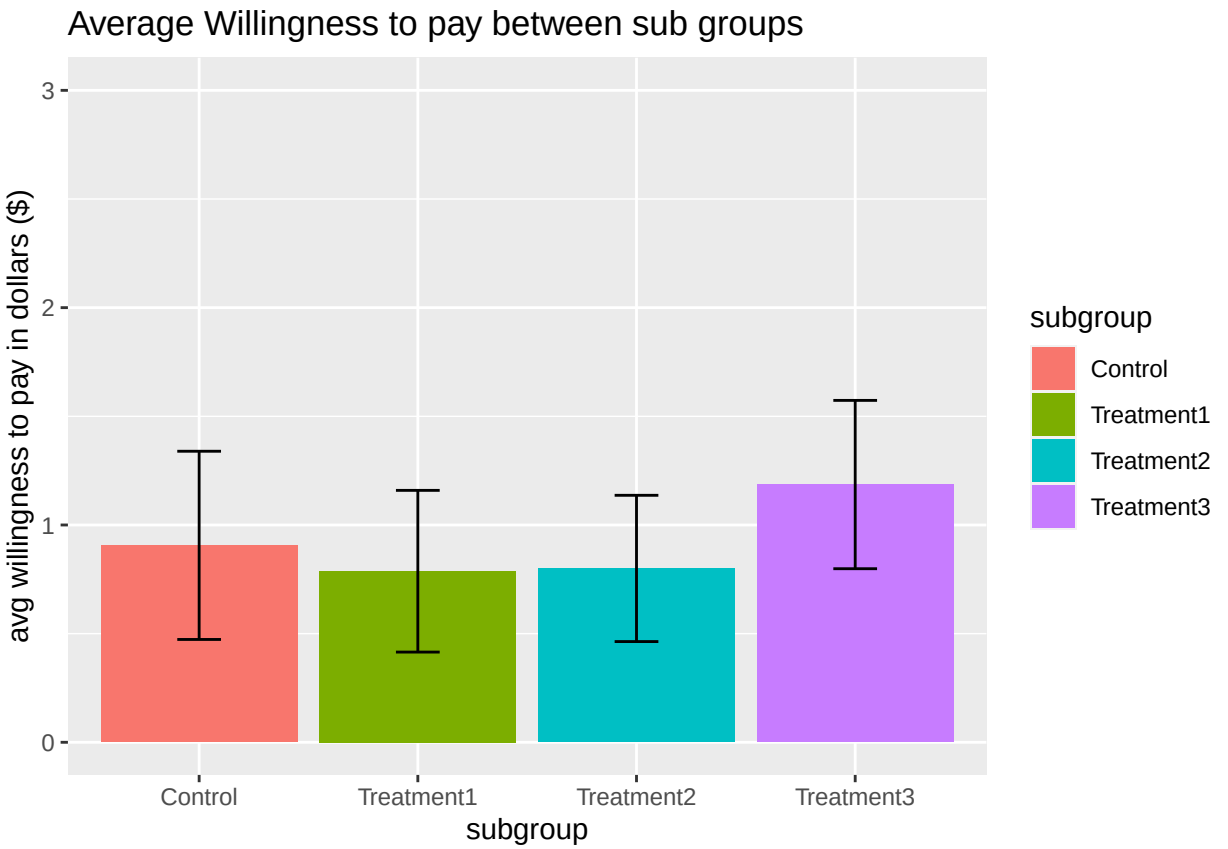
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 5
##   subgroup   `avg willingness in dolla~ `sd willing to pay in dolla~ lower upper
##   <chr>                           <dbl>                        <dbl> <dbl> <dbl>
## 1 Control                         0.906                         1.20 0.473  1.34
## 2 Treatment1                      0.787                         1.27 0.415  1.16
## 3 Treatment2                      0.8                           1.12 0.463  1.14
## 4 Treatment3                      1.19                          1.26 0.799  1.57
```

```
# plotting the values
ggplot(data = group_summary,
       aes(x=group,y=`avg willingness in dollars`,fill = group))+geom_bar(stat = 'identity')+
       ylim(0,3)+ylab("avg willingness to pay in dollars ($)")+
       geom_errorbar(aes(ymin=lower, ymax=upper),
                 width=.2,                    # Width of the error bars
                 position=position_dodge(0.9))+
       ggtitle ("Average Willingness to pay between all overall Treatment and Control")
```




```
ggplot(data = sub_group_summary,
       aes(x=subgroup,y=`avg willingness in dollars`,fill = subgroup))+geom_bar(stat = 'identity')+
       ylim(0,3)+ylab("avg willingness to pay in dollars ($)")+
       geom_errorbar(aes(ymin=lower, ymax=upper),
                 width=.2,                    # Width of the error bars
                 position=position_dodge(0.9))+
       ggtitle ("Average Willingness to pay between sub groups")
```



Regression Analysis

```
# Run basic linear models for both continuous and binary outcome
lm.willing_to_pay <- lm(willing_to_pay ~ subgroup, data=df_finished)
lm.is_willing_to_pay <- lm(is_willing_to_pay ~ subgroup, data = df_finished)

stargazer(lm.willing_to_pay,lm.is_willing_to_pay, title = "Effect of Information Warning on Privacy",
          type = "text",
          column.labels = c("(Continuous)","(Binary)"))
```

```
##
## Effect of Information Warning on Privacy
## ===============================================================
##                                   Dependent variable:
##                               ------------------------------
##                               willing_to_pay is_willing_to_pay
##                                (Continuous)      (Binary)
##                                    (1)              (2)
## ---------------------------------------------------------------
## subgroupTreatment1                -0.119          -0.118
##                                  (0.278)          (0.113)
##
## subgroupTreatment2                -0.106          -0.015
##                                  (0.281)          (0.114)
##
## subgroupTreatment3                0.280            0.121
##                                  (0.283)          (0.115)
```

```
##
## Constant                            0.906***        0.437***
##                                     (0.215)          (0.087)
##
## ------------------------------------------------------------
## Observations                          167              167
## R2                                   0.018            0.031
## Adjusted R2                          0.0004           0.014
## Residual Std. Error (df = 163)       1.214            0.493
## F Statistic (df = 3; 163)            1.023            1.764
## ============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Heterogeneous Treatment Effects

```
# Run linear model for heterogeneous treatment effects for Age >= 35
lm.willing_to_pay.hetero <- lm(willing_to_pay ~ subgroup + equal_over35 + subgroup:equal_over35, data=di
lm.is_willing_to_pay.hetero <- lm(is_willing_to_pay ~ subgroup + equal_over35 + subgroup:equal_over35, d

stargazer(lm.willing_to_pay, lm.willing_to_pay.hetero, lm.is_willing_to_pay, lm.is_willing_to_pay.hetero
          type = "text",
          column.labels = c("(Continuous)","(Continuous)","(Binary)","(Binary)"))
```

```
##
## Effect of Information Warning on Privacy
## ================================================================================================
##                                                  Dependent variable:
##                                 ---------------------------------------------------------------
##                                          willing_to_pay                      is_willing_to_pay
##                                 (Continuous)        (Continuous)          (Binary)            (Bi
##                                      (1)                 (2)                 (3)
## ------------------------------------------------------------------------------------------------
## subgroupTreatment1                  -0.119              -0.302              -0.118              -0
##                                     (0.278)             (0.328)             (0.113)             (0
##
## subgroupTreatment2                  -0.106              -0.164              -0.015              -0
##                                     (0.281)             (0.318)             (0.114)             (0
##
## subgroupTreatment3                   0.280               0.093               0.121               0
##                                     (0.283)             (0.324)             (0.115)             (0
##
## equal_over35                                            -0.708                                  -0
##                                                         (0.496)                                 (0
##
## subgroupTreatment1:equal_over35                          0.727                                  0
##                                                         (0.624)                                 (0
##
## subgroupTreatment2:equal_over35                          0.039                                  0
##                                                         (0.685)                                 (0
##
## subgroupTreatment3:equal_over35                          0.754                                  0
```

```
##                                                (0.673)                                (0
##
## Constant                                    0.906***           1.083***           0.437***           0.5
##                                              (0.215)            (0.248)            (0.087)            (0
##
## ------------------------------------------------------------------------------------------------------------
## Observations                                   167                167                167
## R2                                            0.018              0.043              0.031               0
## Adjusted R2                                  0.0004              0.001              0.014               0
## Residual Std. Error                1.214 (df = 163)   1.214 (df = 159)   0.493 (df = 163)   0.494 (d
## F Statistic                       1.023 (df = 3; 163) 1.017 (df = 7; 159) 1.764 (df = 3; 163) 1.256 (df
## ============================================================================================================
## Note:                                                                        *p<0.1; **p<0.05
```

Interaction Coefficients are not statistically significant, suggesting there are no Heterogenous Treatment Effects

```
# Run linear model for heterogeneous treatment effects for Age >= 30
lm.willing_to_pay.hetero <- lm(willing_to_pay ~ subgroup + equal_over30 + subgroup:equal_over30, data=d
lm.is_willing_to_pay.hetero <- lm(is_willing_to_pay ~ subgroup + equal_over30 + subgroup:equal_over30, d

stargazer(lm.willing_to_pay, lm.willing_to_pay.hetero, lm.is_willing_to_pay, lm.is_willing_to_pay.hetero
          type = "text",
          column.labels = c("(Continuous)","(Continuous)","(Binary)","(Binary)"))
```

```
##
## Effect of Information Warning on Privacy
## ============================================================================================================
##                                                      Dependent variable:
##                                    ---------------------------------------------------------------------
##                                               willing_to_pay                          is_willing_to_pay
##                                    (Continuous)       (Continuous)          (Binary)            (Bi
##                                         (1)                (2)                (3)
## ------------------------------------------------------------------------------------------------------------
## subgroupTreatment1                     -0.119             -0.173             -0.118              -0
##                                         (0.278)            (0.396)            (0.113)            (0
##
## subgroupTreatment2                     -0.106              0.098             -0.015               0
##                                         (0.281)            (0.380)            (0.114)            (0
##
## subgroupTreatment3                      0.280              0.326              0.121               0
##                                         (0.283)            (0.391)            (0.115)            (0
##
## equal_over30                                               -0.051                                -0
##                                                            (0.430)                               (0
##
## subgroupTreatment1:equal_over30                             0.109                                -0
##                                                            (0.557)                               (0
##
## subgroupTreatment2:equal_over30                            -0.750                                -0
##                                                            (0.586)                               (0
##
## subgroupTreatment3:equal_over30                            -0.146                                0
```

```
##                                                (0.575)                                    (0
##
## Constant                           0.906***       0.933***       0.437***       0.4
##                                     (0.215)        (0.313)        (0.087)        (0
##
## ---------------------------------------------------------------------------------------------
## Observations                          167            167            167            
## R2                                   0.018          0.045          0.031          0
## Adjusted R2                         0.0004          0.002          0.014          0
## Residual Std. Error              1.214 (df = 163)    1.213 (df = 159)    0.493 (df = 163)    0.494 (
## F Statistic                    1.023 (df = 3; 163) 1.058 (df = 7; 159) 1.764 (df = 3; 163) 1.280 (d
## =============================================================================================
## Note:                                                                    *p<0.1; **p<0.05
```

Interaction Coefficients are not statistically significant, suggesting there are no Heterogenous Treatment Effects