# Pilot Data Cleaner

Daniel K Baissa, Melani Cammett, and Aytug Sasmaz

December 12, 2020

## Loading in data

I will start by creating a function to automatically set up the data.

All of this is a work in progress

```
# Creating a function to load datasets

data_cleaner <- function(d){

  df <- read_csv(d) %>% #reading in the data
  slice(c(-1, -2)) #qualtrics adds 2 rows of unnecessary headers. This removes them.

# Cleaning up the conjoint data.
# Note, this is not very efficient memory wise
# and may need to be tuned up for larger datasets

names <- colnames(df)

new_names <- paste0(1:243, "_conjoint_friend1")
new_names2 <- paste0(1:243, "_conjoint_friend2")

df2 <- df %>%
  setnames(old = names[246:488], new = new_names) %>%
  setnames(old = names[489:731], new = new_names2)


test2 <- df2 %>%
    pivot_longer(
          cols = ends_with("friend1"),
          names_to = "Conjoint_first_permutation",
          values_to = "Conjoint_first_permutation_answer") %>%
  filter(!is.na(Conjoint_first_permutation_answer)) %>%
    pivot_longer(
          cols = ends_with("_friend2"),
          names_to = "Conjoint_second_permutation",
          values_to = "Conjoint_second_permutation_answer") %>%
    filter(!is.na(Conjoint_second_permutation_answer)) %>%
   pivot_longer(
          cols = starts_with("ptt"),
          names_to = "Petition_Experiment_Treatment",
          values_to = "Petition_Experiment_Treatment_Answer") %>%
  filter(!is.na(Petition_Experiment_Treatment_Answer))
```

```
}
```

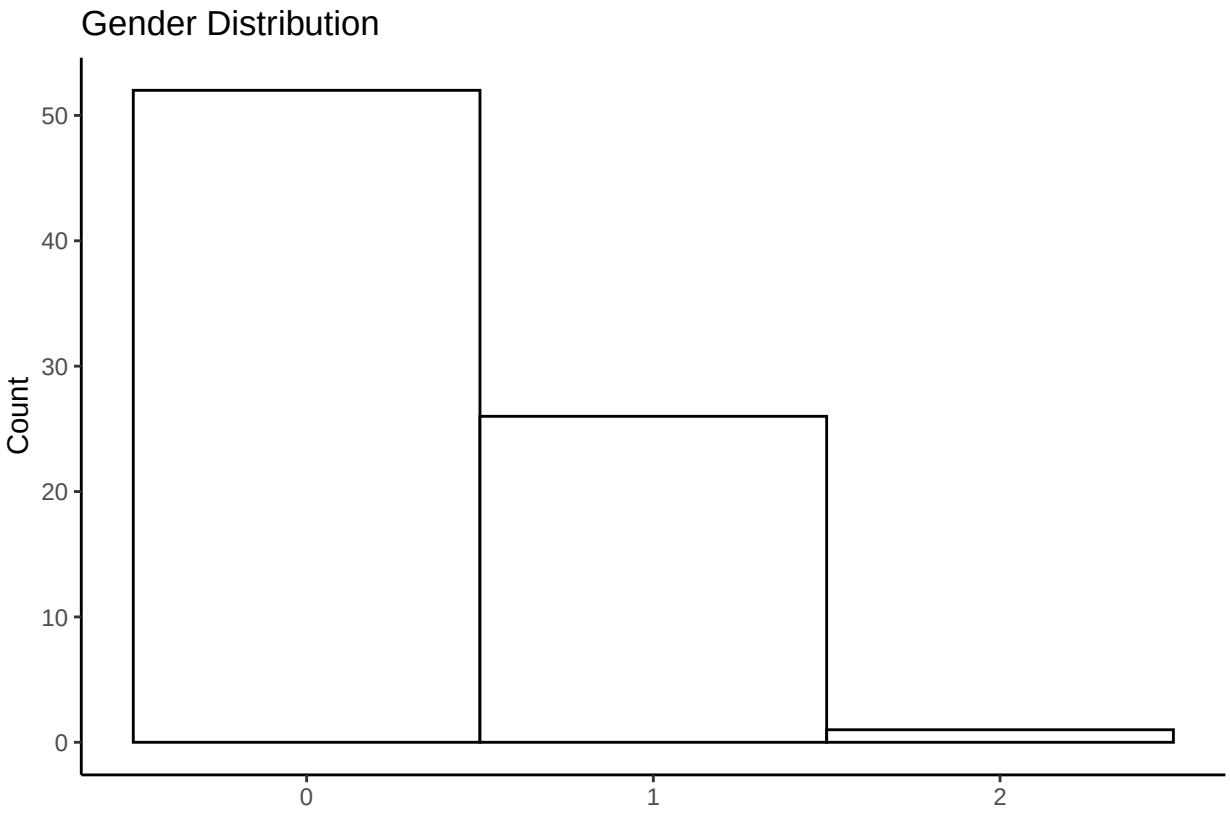
Now that the function is created, we can use it to make our data. There will be a warning, but it is totally ok for now.

```
csv <- c("D:/Lebanon_data/LEB Youth Civic Engagement ENG - postcut_December 11, 2020_21.29.csv",
         "D:/Lebanon_data/LEB Youth Civic Engagement ARA - postcut_December 11, 2020_22.09.csv")

ENG <- data_cleaner(csv[1])
ARA <- data_cleaner(csv[2])

# For now I will remove the Q_RecaptchaScore

ARA$Q_RecaptchaScore <- NULL
ENG$Q_RecaptchaScore <- NULL

# There is an extra att_women in English "att_women_4"

ENG$att_women_4 <- NULL


# For now I am just going to force the two datasets to have the same variable names

ARA <- ARA %>%
  setnames(old = colnames(ARA), new = colnames(ENG))

df <- rbind(ENG, ARA)
```

## Demographic distribution

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3544  1.0000  2.0000
```

```
ggplot(df, aes(x = gender))+
  geom_histogram(bins = 3, fill="white", color="black")+
  labs(title="Gender Distribution",x="", y = "Count")+
  theme_classic()
```

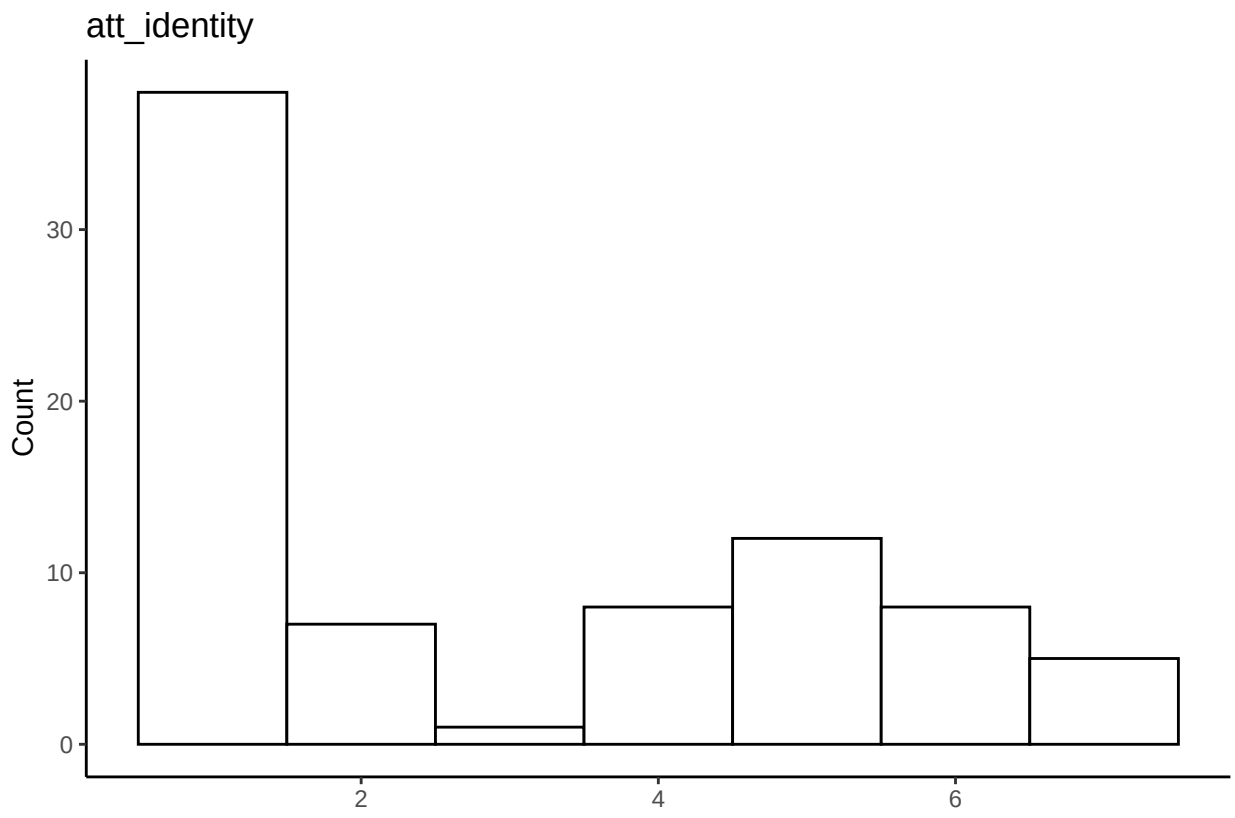



**Religious demographics**

```
summary(as.double(df$att_identity))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.911   5.000   7.000
```

```
ggplot(df, aes(x = as.double(att_identity)))+
  geom_histogram(bins = 7, fill="white", color="black")+
  labs(title="att_identity",x="", y = "Count")+
  theme_classic()
```



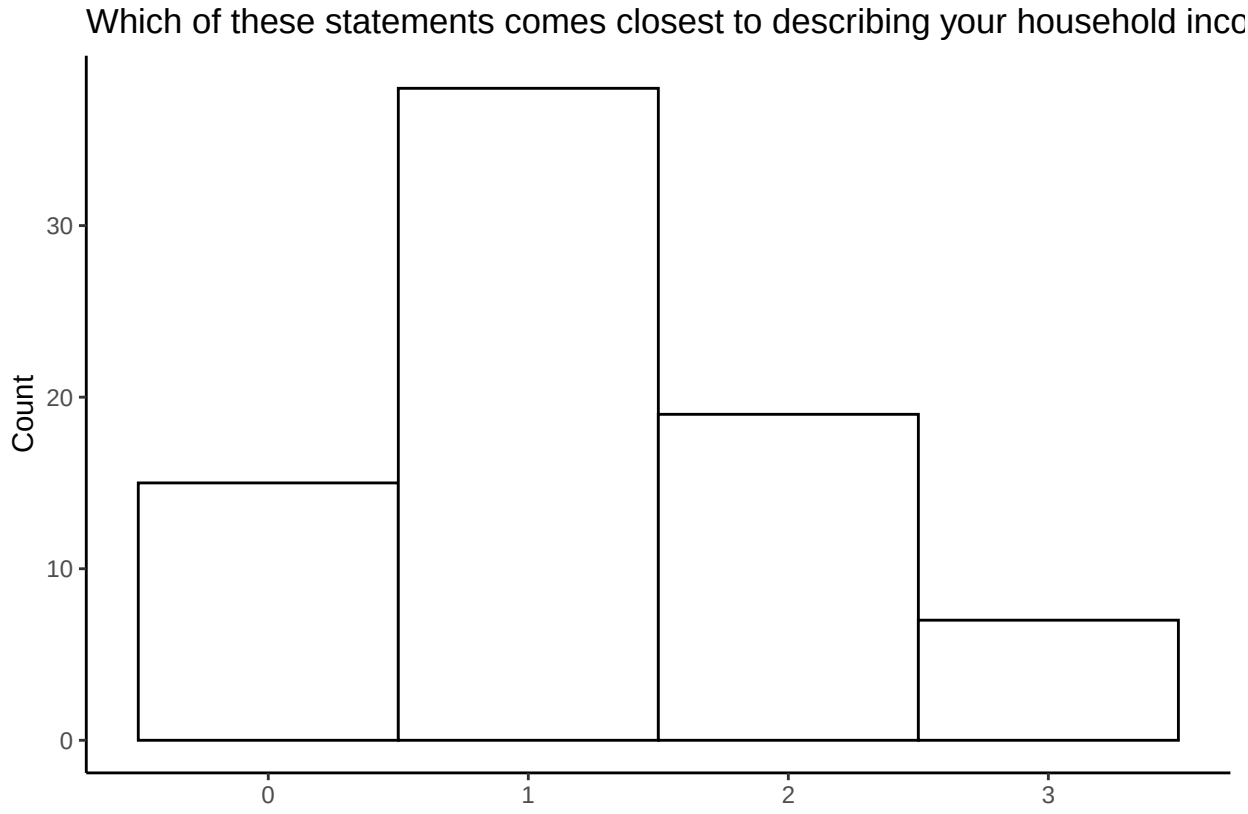

**Income distribution**

Below are some statements related to your household income. Which of these statements comes closest to describing your household income?

```
summary(as.double(df$dem_income1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.228   2.000   3.000
```

```
ggplot(df, aes(x = as.double(dem_income1)))+
  geom_histogram(bins = 4, fill="white", color="black")+
  labs(title="Which of these statements comes closest to describing your household income?",x="", y = "
  theme_classic()
```



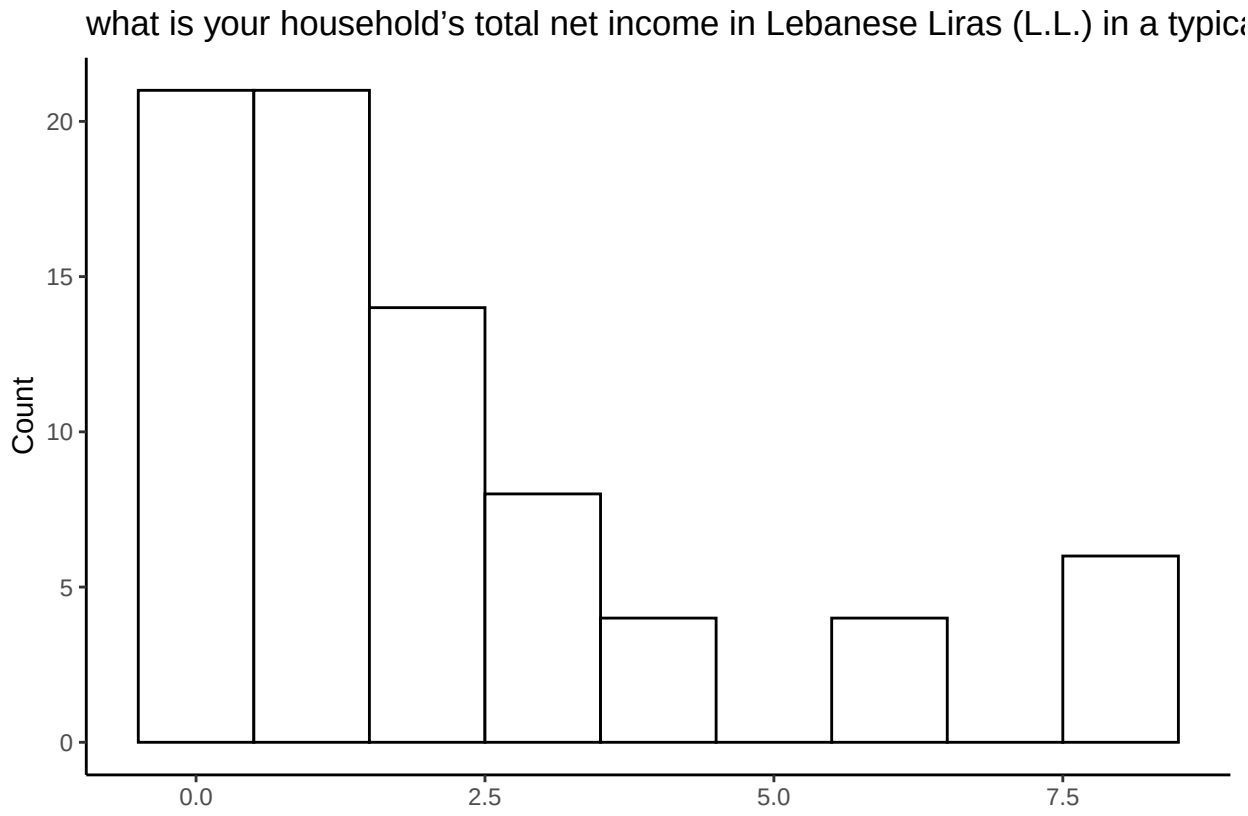

To the best of your knowledge, what is your household's total net income in Lebanese Liras (L.L.) in a typical month?

```
summary(as.double(df$dem_income2))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   1.000   2.064   3.000   8.000       1
```

```
ggplot(df, aes(x = as.double(dem_income2)))+
  geom_histogram(bins = 9, fill="white", color="black")+
  labs(title="what is your household's total net income in Lebanese Liras (L.L.) in a typical month?",x=
  theme_classic()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

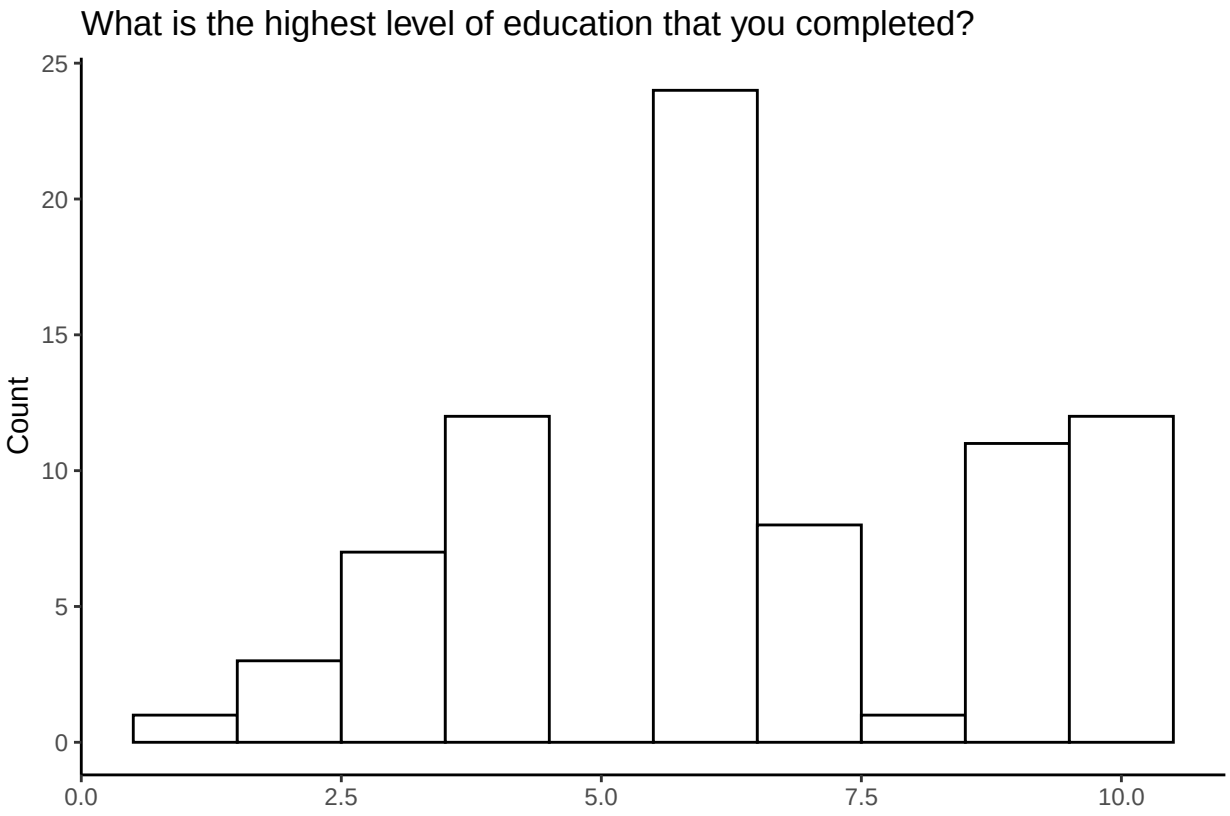



What is the highest level of education that you completed?

```
summary(as.double(df$dem_edu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   6.367   9.000  10.000
```

```
ggplot(df, aes(x = as.double(dem_edu)))+
  geom_histogram(bins = 10, fill="white", color="black")+
  labs(title="What is the highest level of education that you completed?",x="", y = "Count")+
  theme_classic()
```



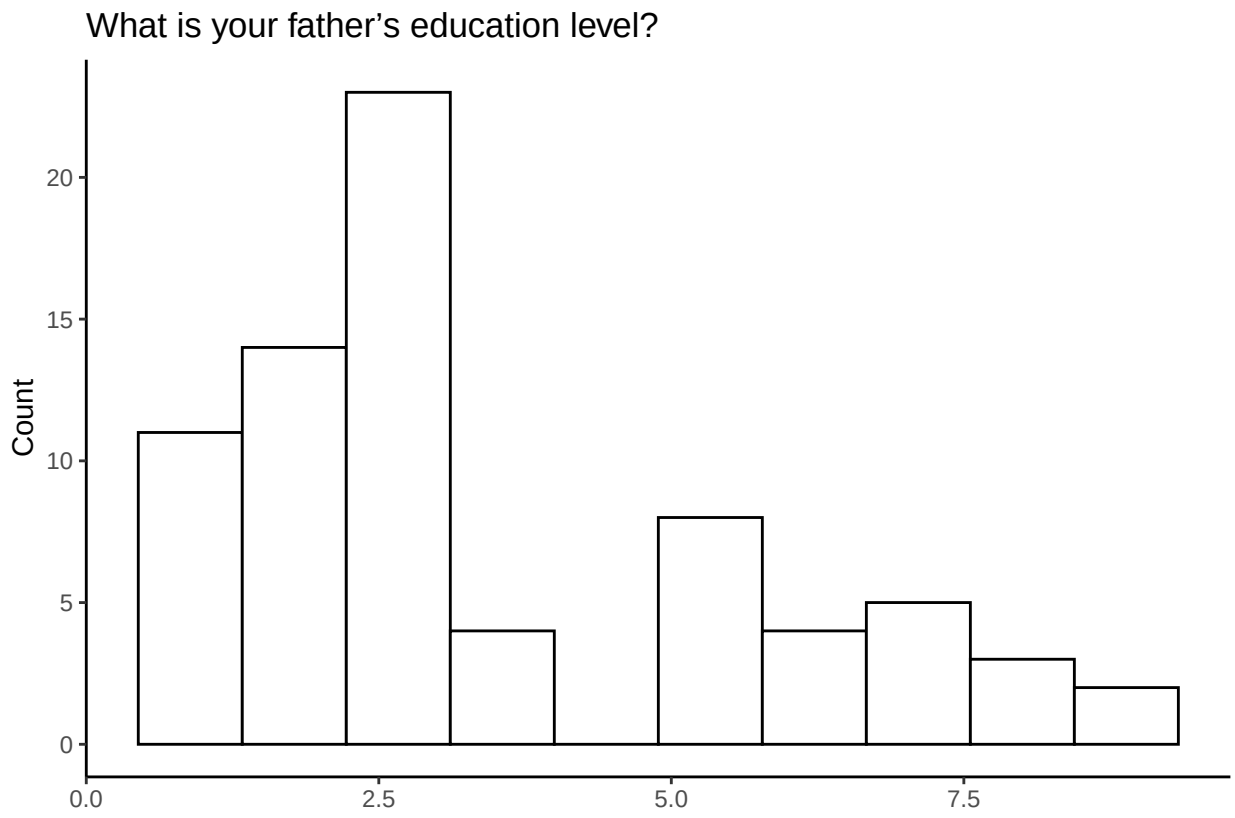

What is your father's education level?

```
summary(as.double(df$dem_fatheredu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   3.000   3.581   5.000   9.000       5
```

```
ggplot(df, aes(x = as.double(dem_fatheredu)))+
  geom_histogram(bins = 10, fill="white", color="black")+
  labs(title="What is your father's education level? ",x="", y = "Count")+
  theme_classic()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

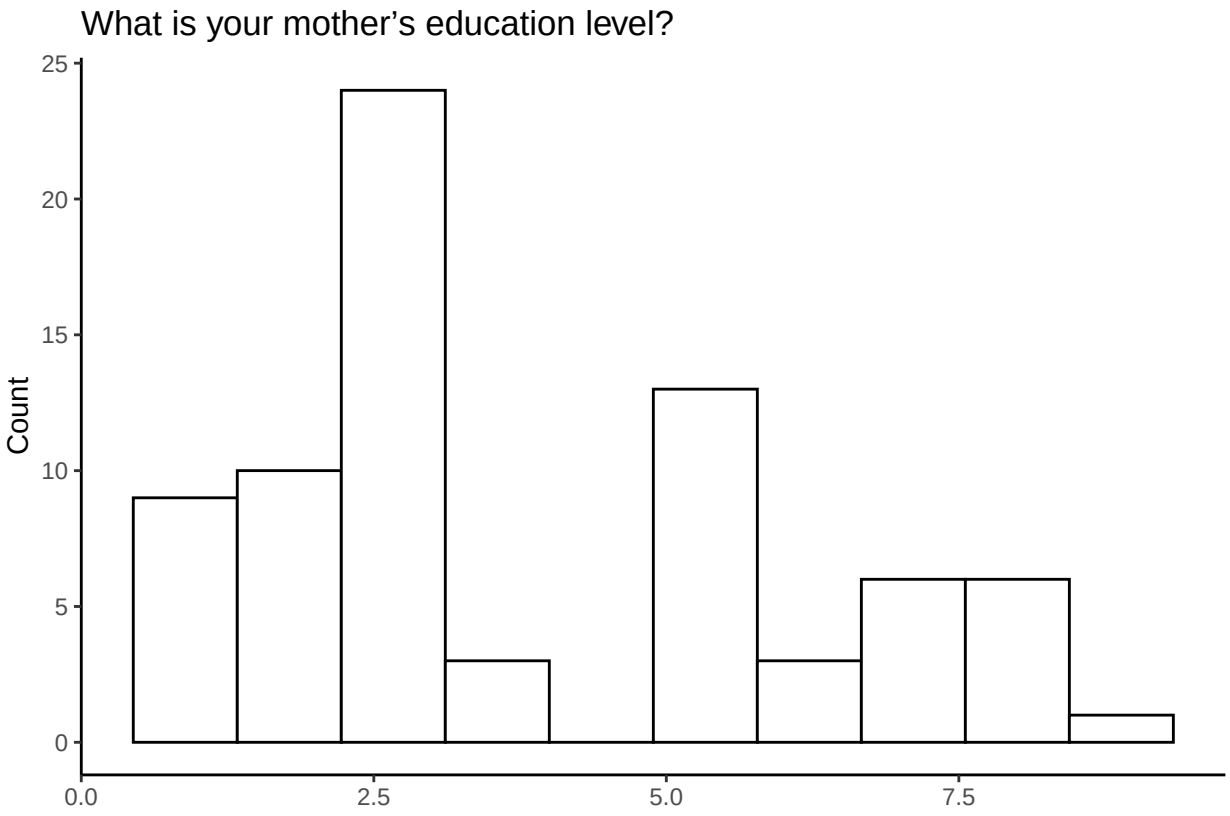



```
summary(as.double(df$dem_motheredu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.500   3.000   3.933   5.000   9.000       4
```

```
ggplot(df, aes(x = as.double(dem_motheredu)))+
  geom_histogram(bins = 10, fill="white", color="black")+
  labs(title="What is your mother's education level?",x="", y = "Count")+
  theme_classic()
```

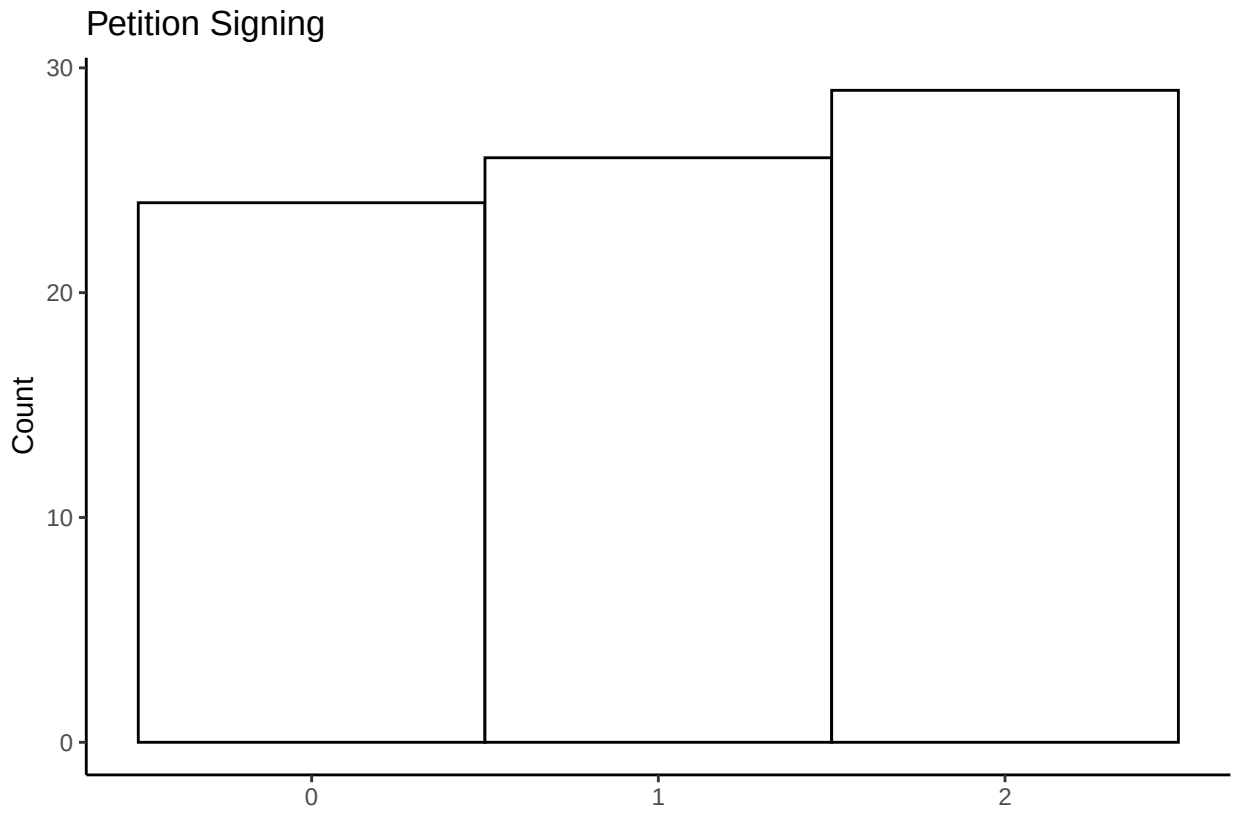
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```




Mother's are better educated than fathers in this sample

### Petition Signing distribution

How many people actually signed the petition? Lets find out.




Lets do some basic stats on the petition experiment. Here is a simple logistic regression to see if the control group is different from any of the treatments.

```
# colnames(df)
# unique(df$Petition_Experiment_Treatment)

df2 <- df %>%
  filter(Petition_Experiment_Treatment == "ptt_treat_control_fo" | Petition_Experiment_Treatment == "pt

df2$pt_econ_treatment <- 0
df2$pt_econ_treatment[which(df2$Petition_Experiment_Treatment == "ptt_treat_eco_foll")] <- 1
df2$signed <- 0
df2$signed[which(df2$Q144 > 0)] <- 1

test <- glm(signed ~ pt_econ_treatment, data = df2, family = binomial())

summary(test)
```

```
##
## Call:
## glm(formula = signed ~ pt_econ_treatment, family = binomial(),
##     data = df2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84648   0.00008   0.00008   0.63352   0.63352
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)          19.57    3584.67   0.005    0.996
## pt_econ_treatment   -18.06    3584.67  -0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.003  on 19  degrees of freedom
## Residual deviance: 10.431  on 18  degrees of freedom
## AIC: 14.431
##
## Number of Fisher Scoring iterations: 18
```

```
## combined table
(ctable <- cbind(ctable, "p value" = p))
```

```
##                                                    Value Std. Error
## Petition_Experiment_Treatmentptt_treat_bft_foll   -0.1667512  0.7582067
## Petition_Experiment_Treatmentptt_treat_control_fo  2.0974644  0.9855129
## Petition_Experiment_Treatmentptt_treat_cor_foll   -0.1171191  0.7088077
## Petition_Experiment_Treatmentptt_treat_eco_foll    0.9024628  0.8248948
## Petition_Experiment_Treatmentptt_treat_sec_foll   -0.6321240  0.7543644
## 0|1                                               -0.7742921  0.5908313
## 1|2                                                0.7744112  0.5908326
##                                                      t value    p value
## Petition_Experiment_Treatmentptt_treat_bft_foll   -0.2199285 0.82592686
## Petition_Experiment_Treatmentptt_treat_control_fo  2.1282973 0.03331245
## Petition_Experiment_Treatmentptt_treat_cor_foll   -0.1652339 0.86875986
## Petition_Experiment_Treatmentptt_treat_eco_foll    1.0940338 0.27394015
## Petition_Experiment_Treatmentptt_treat_sec_foll   -0.8379557 0.40205559
## 0|1                                               -1.3105130 0.19002234
## 1|2                                                1.3107116 0.18995522
```

The Covid Treatment was the treatment left out for comparison by the model. Lets take a look at that.

```
signif((ctable <- cbind(ctable, "p value" = p)),3)
```

```
##                                                   Value Std. Error t value
## Petition_Experiment_Treatmentptt_treat_control_fo  2.040      1.010    2.03
## 0|1                                               -0.713      0.639   -1.12
## 1|2                                                0.714      0.639    1.12
##                                                   p value
## Petition_Experiment_Treatmentptt_treat_control_fo   0.042
## 0|1                                                 0.264
## 1|2                                                 0.264
```

Something appears to be going on here, but the N is so small that it could still be chance