Who falls for Misinformation?

Sanjay Chhetri

11/28/2021

Data Source

I obtained the dataset from a repository publicly shared by the authors of a published article. Link to the article For this article, the authors used the dataset to run a Confirmatory Factor Analysis. The sample is entirely from Africa, by design. The dataset represents a total of 21 African countries. I used the dataset to run a multiple regression to find out what factors predict the subscription of COVID-related misinformation.

```
library(readxl)
cad <- read_excel("covid_africa_data.xlsx")
dim(cad)
```

```
## [1] 563  35
```

```
head(cad, 10)
```

```
## # A tibble: 10 x 35
##    Timestamp         Gender `Age class`  `Highest level of ed~ `Country of Orig~
##    <chr>             <chr>  <chr>        <chr>                 <chr>
##  1 2020/03/31 1:18:~ Male   35 - 39 yea~ Post Graduate level   Nigeria
##  2 2020/03/31 1:26:~ Male   40 - 49 yea~ Post Graduate level   Nigeria
##  3 2020/03/31 1:34:~ Male   35 - 39 yea~ Post Graduate level   Nigeria
##  4 2020/03/31 1:34:~ Female 25 - 29 yea~ Post Graduate level   Nigeria
##  5 2020/03/31 1:35:~ Male   30 - 34 yea~ Tertiary education    Nigeria
##  6 2020/03/31 1:35:~ Male   35 - 39 yea~ Post Graduate level   Nigeria
##  7 2020/03/31 1:35:~ Male   30 - 34 yea~ Tertiary education    Nigeria
##  8 2020/03/31 1:35:~ Male   18 - 24 yea~ Post Graduate level   Nigeria
##  9 2020/03/31 1:42:~ Male   18 - 24 yea~ Tertiary education    Nigeria
## 10 2020/03/31 1:42:~ Female 40 - 49 yea~ Post Graduate level   Nigeria
## # ... with 30 more variables: Region of your country <chr>,
## #   Religion faith/belief <chr>, Know_a_covid_patient <chr>,
## #   Source of information about Coronavirus (COVID19). Please tick as many as applicable. <chr>,
## #   If you will love to receive a copy of the published work, please include your email address belo
## #   q1 <chr>, q2 <chr>, q3 <chr>, q4 <chr>, q5 <chr>, q6 <chr>, q7 <chr>,
## #   q8 <chr>, q9 <chr>, q10 <chr>, q11 <chr>, q12 <chr>, q13 <chr>,
## #   Coronavirus (COVID19) has taught us humility and draw humanity closer to GOD. <chr>, ...
```

```
glimpse(cad)
```

```
## Rows: 563
```

```
## Columns: 35
## $ Timestamp
## $ Gender
## $ ‘Age class‘
## $ ‘Highest level of education‘
## $ ‘Country of Origin‘
## $ ‘Region of your country‘
## $ ‘Religion faith/belief‘
## $ Know_a_covid_patient
## $ ‘Source of information about Coronavirus (COVID19). Please tick as many as applicable.‘
## $ ‘If you will love to receive a copy of the published work, please include your email address below
## $ q1
## $ q2
## $ q3
## $ q4
## $ q5
## $ q6
## $ q7
## $ q8
## $ q9
## $ q10
## $ q11
## $ q12
## $ q13
## $ ‘Coronavirus (COVID19) has taught us humility and draw humanity closer to GOD.‘
## $ ‘Rubbing of Anointing Oil (or prayer ablution) on the body can prevent/resist COVID19.‘
## $ ‘Coronavirus (COVID19) cannot affect believers.‘
## $ ‘Coronavirus (COVID19) cannot be contracted in the Church/Mosque.‘
## $ ‘Coronavirus (COVID19) is a divine punishment from God to humanity.‘
## $ ‘Social distancing in Africa is in theory, it cannot be practice.‘
## $ ‘Total lockdown in Africa will cause a more dangerous hunger outbreak than the COVID19 itself.‘
## $ ‘The media is causing more panicking about COVID19 than it should.‘
## $ ‘There are many other dangerous diseases (e.g. Malaria, Cancer, etc) that kill more people daily th
## $ ‘Hunger kills more than COVID19.‘
## $ ‘Thank you for your time! Please leave us with any comments about COVID19 you think we should inclu
## $ ‘Kindly provide your name if you will want us to list your name in the appreciation section of this
```

The dataset was in csv file, which I downloaded and saved as an Excel worksheet. I then subsetted the dataset to include only the variables needed for my analysis. Two variables pertaining to the source of covid information and religious orientation could be included in the model but I opted not to use them. FOr the former, the questionnaire gave multiple choices on COVID information source and most participants chose various options. For religion, vast majority chose Christianity and Islam, with only a handful with ‘don’t want to tell’/‘other’/‘neutral’. It would basically be comparing one religion vs another.

Reshaping the dataset

Here I select only the variables that are of use for my analysis.

```
df1 <- cad %>% mutate(gender = factor(Gender),
         age_group = factor(`Age class`),
         education = factor(`Highest level of education`),
         familiarity = factor(Know_a_covid_patient)) %>%
  select(gender, age_group, education,  familiarity, q1:q13)
```

Data Wrangling

Here I recategorieze some variables to make them more readable.

```
table(df1$age_group)
```

```
##
##      18 - 24 years      25 - 29 years      30 - 34 years      35 - 39 years
##                122                118                100                 75
##      40 - 49 years      50 - 59 years      60 - 69 years 70 years and above
##                 86                 46                  9                  4
## Less than 18 years
##                  3
```

```
# collapsing age class varible to three categories
levels(df1$age_group) <- list("Youths" = c("Less than 18 years", "18 - 24 years", "25 - 29 years"),
                          "Adults" = c("30 - 34 years", "35 - 39 years", "40 - 49 years"),
                          "Elderly" = c("50 - 59 years", "60 - 69 years", "70 years and above"))
sum(is.na(df1$age_group)) #checking missing values
```

```
## [1] 0
```

```
levels(df1$education) <- list("college_educated" = c("Tertiary education", "Post Graduate level"),
                              "No_college" = c("No formal education", "Primary/Basic education", "Secon

sum(is.na(cad$`Highest level of education`))
```

```
## [1] 0
```

```
sum(is.na(df1$education))
```

```
## [1] 0
```

Collapsing various age groups into three general groups made sense to see the age effect with better clarity. Collapsing various education levels into two groups (college educated or not) facilitated a plain comparison.

More Data Manipulation

Here, I first convert the likert scale texts into their numerical equivalents as suggested by the data source. Then I aggregate the misinformation related columns taking average.

```
df1[df1 == "Strongly Disagree"] <- "0"
df1[df1 == "Disagree"] <- "0"
df1[df1 == "Neutral"] <- "1"
df1[df1=="Agree"] <- "4"
df1[df1=="Strongly Agree"] <- "4"
head(df1)
```

```
## # A tibble: 6 x 17
##   gender age_group education     familiarity q1    q2    q3    q4    q5    q6
##   <fct>  <fct>     <fct>         <fct>       <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Male   Adults    college_educ~ No          0     1     1     4     4     0
## 2 Male   Adults    college_educ~ No          4     1     1     1     4     1
## 3 Male   Adults    college_educ~ No          0     0     0     1     0     0
## 4 Female Youths    college_educ~ No          1     1     0     0     0     1
## 5 Male   Adults    college_educ~ No          1     1     1     1     1     1
## 6 Male   Adults    college_educ~ No          1     1     4     4     4     1
## # ... with 7 more variables: q7 <chr>, q8 <chr>, q9 <chr>, q10 <chr>,
## #   q11 <chr>, q12 <chr>, q13 <chr>
```

```
class(df1$q1)
```

```
## [1] "character"
```

```
df1[, 5:17] <- lapply(df1[, 5:17], as.numeric)
class(df1$q1)
```

```
## [1] "numeric"
```

```
head(df1)
```

```
## # A tibble: 6 x 17
##   gender age_group education     familiarity    q1    q2    q3    q4    q5    q6
##   <fct>  <fct>     <fct>         <fct>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Male   Adults    college_educ~ No              0     1     1     4     4     0
## 2 Male   Adults    college_educ~ No              4     1     1     1     4     1
## 3 Male   Adults    college_educ~ No              0     0     0     1     0     0
## 4 Female Youths    college_educ~ No              1     1     0     0     0     1
## 5 Male   Adults    college_educ~ No              1     1     1     1     1     1
## 6 Male   Adults    college_educ~ No              1     1     4     4     4     1
## # ... with 7 more variables: q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>,
## #   q11 <dbl>, q12 <dbl>, q13 <dbl>
```

```
df1$q <- rowMeans(subset(df1, select = q1:q13), na.rm = T)
head(df1)
```

```
## # A tibble: 6 x 18
##   gender age_group education     familiarity    q1    q2    q3    q4    q5    q6
##   <fct>  <fct>     <fct>         <fct>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Male   Adults    college_educ~ No              0     1     1     4     4     0
## 2 Male   Adults    college_educ~ No              4     1     1     1     4     1
## 3 Male   Adults    college_educ~ No              0     0     0     1     0     0
## 4 Female Youths    college_educ~ No              1     1     0     0     0     1
## 5 Male   Adults    college_educ~ No              1     1     1     1     1     1
## 6 Male   Adults    college_educ~ No              1     1     4     4     4     1
## # ... with 8 more variables: q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>,
## #   q11 <dbl>, q12 <dbl>, q13 <dbl>, q <dbl>
```

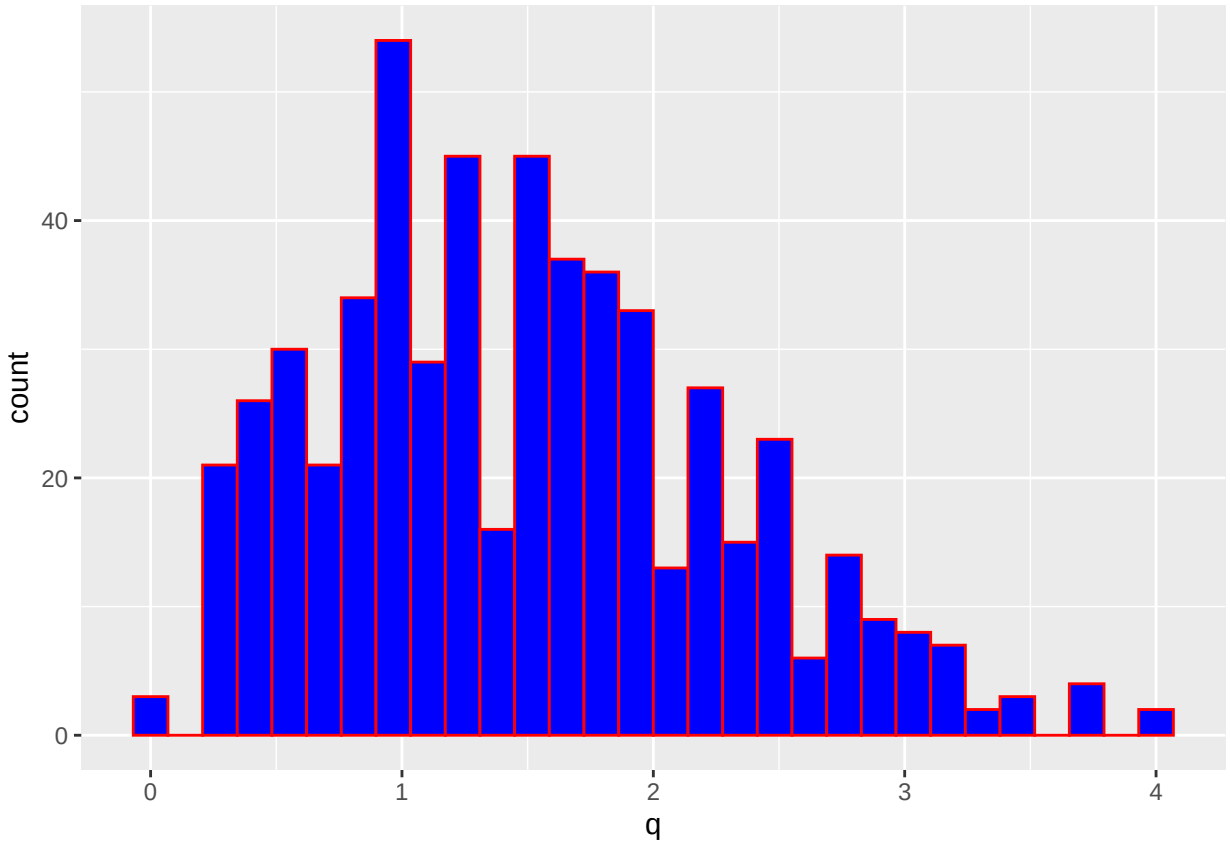
```
df <- df1 %>% select(gender, age_group, education, familiarity, q )
head(df)
```

```
## # A tibble: 6 x 5
##   gender age_group education        familiarity     q
##   <fct>  <fct>     <fct>            <fct>       <dbl>
## 1 Male   Adults    college_educated No          1.46
## 2 Male   Adults    college_educated No          2.08
## 3 Male   Adults    college_educated No          0.462
## 4 Female Youths    college_educated No          1.62
## 5 Male   Adults    college_educated No          1
## 6 Male   Adults    college_educated No          2.15
```

Here, q is the outcome variable and represents the extent to which one is misinformed about COVID (subscribes to COVID related misinformation)

##What's the distribution of misinformation score (q)?

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Regression

```
model <- lm(q~gender+age_group+education+familiarity, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = q ~ gender + age_group + education + familiarity,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46375 -0.61969 -0.08445  0.45401  2.52661
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.47594    0.06563  22.490  < 2e-16 ***
## genderMale          -0.00687    0.06975  -0.098    0.922
## age_groupAdults     -0.06459    0.07028  -0.919    0.358
## age_groupElderly    -0.00532    0.11638  -0.046    0.964
## educationNo_college  0.62901    0.13282   4.736 2.77e-06 ***
## familiarityYes       0.06892    0.08348   0.826    0.409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7683 on 557 degrees of freedom
## Multiple R-squared:  0.04307,    Adjusted R-squared:  0.03448
## F-statistic: 5.015 on 5 and 557 DF,  p-value: 0.0001671
```

The regression model with 'q' as the outcome and gender, age_group, education, and familiarity as explanatory variables is significant, $F(5, 557) = 5.015$, $p<0.001$. The model explains less than 4% of the variance in the misinformation index (outcome), with adjusted R squared at 0.034. Education is the only variable that significantly predicts the misinformation index.

##Is there Any interaction effect?

Some visualizations

Let the plots speak!

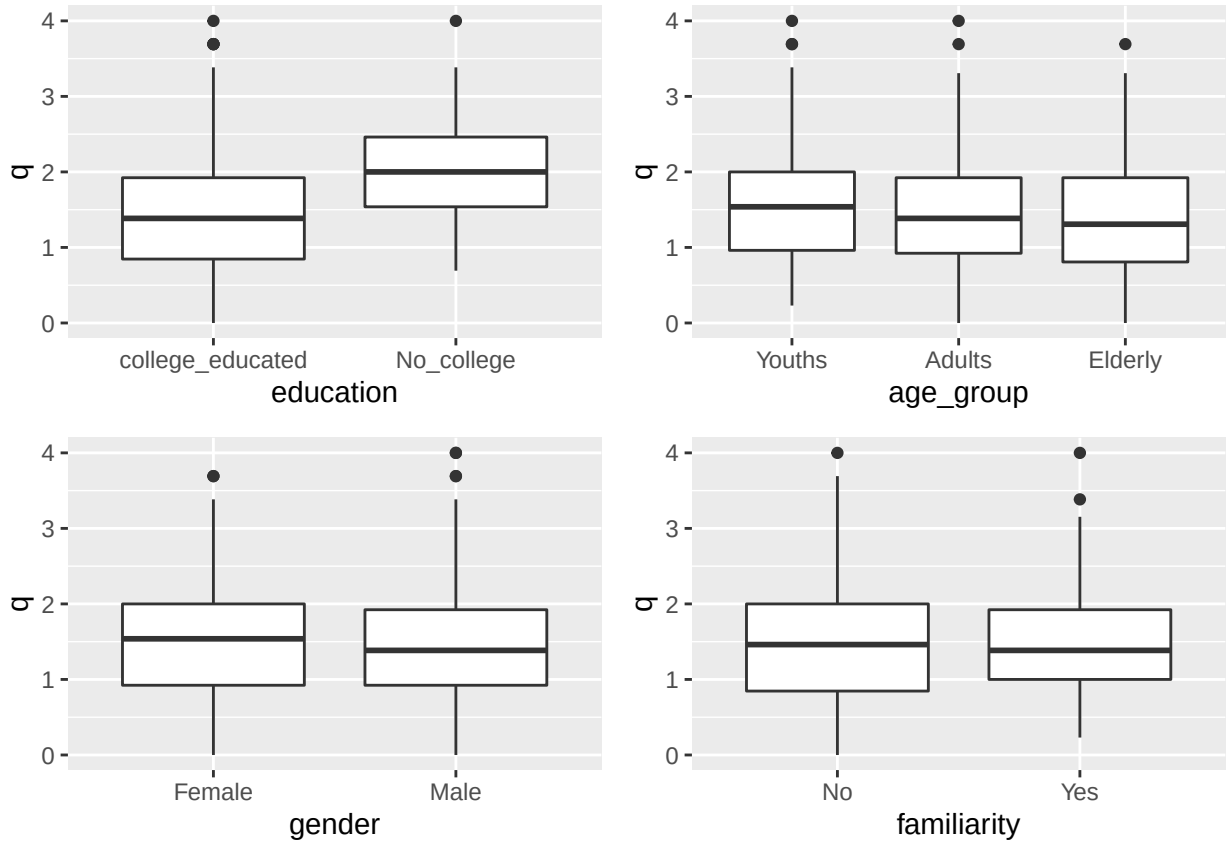
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

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

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

```
p1 <- df %>% ggplot(aes(x=education, y=q))+
 geom_boxplot()
p2 <- df %>% ggplot(aes(age_group, y=q))+geom_boxplot()
p3 <- ggplot(df, aes(gender, q,))+geom_boxplot()
p4 <- ggplot(df, aes(familiarity, q))+geom_boxplot()
grid.arrange(p1, p2, p3, p4, ncol =2)
```



##Figment of Imagination

```
df2 <- df
age_img <- sample(18:70, size = 563, replace = TRUE)
age_img
```

```
##   [1] 63 38 48 50 43 29 68 58 49 64 69 57 67 28 60 68 68 35 70 50 36 64 62 52 18
##  [26] 18 52 29 19 66 32 30 70 70 59 34 34 33 41 51 61 70 46 48 25 19 66 32 33 33
##  [51] 66 31 60 22 18 42 66 45 21 35 61 54 18 29 70 28 67 40 24 43 49 51 62 23 63
##  [76] 38 56 70 38 46 32 21 43 50 51 45 23 33 24 67 35 46 34 66 45 70 45 28 61 35
## [101] 34 28 33 25 22 21 47 55 26 24 33 51 58 57 26 18 37 47 35 54 39 25 39 36 41
## [126] 37 60 70 65 49 21 68 24 30 25 64 69 26 59 30 66 34 65 59 63 52 38 52 58 47
## [151] 63 29 69 20 53 47 64 32 63 47 62 21 24 31 38 19 64 21 41 49 57 21 40 20 64
## [176] 35 21 55 31 23 46 51 19 52 55 42 48 27 22 43 57 68 47 70 68 61 31 26 38 69
## [201] 37 66 21 33 52 62 45 42 52 51 59 43 28 26 60 20 37 47 42 56 65 39 48 58 45
## [226] 48 26 42 25 63 65 61 22 42 59 42 70 59 60 66 62 18 39 18 52 67 31 20 50 58
## [251] 29 62 40 43 66 48 52 61 61 33 39 57 43 64 36 58 39 44 48 23 23 69 53 41 33
## [276] 27 37 49 27 42 29 63 60 50 33 24 18 32 20 37 37 22 32 46 61 61 48 33 19 39
```

```
## [301] 58 61 70 22 54 48 67 53 47 33 27 37 27 48 57 18 41 45 42 61 63 59 46 29 55
## [326] 29 70 31 30 61 44 29 41 53 24 41 47 25 61 25 18 66 60 36 24 29 25 69 49 45
## [351] 54 48 59 67 54 59 38 50 38 55 64 63 61 42 52 65 52 31 43 55 21 41 36 19 62
## [376] 51 49 63 49 31 50 26 27 68 37 64 61 62 68 18 25 22 55 18 64 70 68 31 60 54
## [401] 57 37 33 69 30 42 67 40 27 58 25 70 47 49 66 27 57 50 42 33 21 68 59 33 67
## [426] 38 20 30 51 22 47 54 28 33 43 65 27 22 24 33 29 65 60 41 25 49 47 58 39 65
## [451] 19 23 37 69 59 60 20 19 68 61 24 63 67 60 36 52 45 24 34 34 20 22 39 57 41
## [476] 56 53 41 62 47 39 47 40 43 64 30 36 66 63 36 40 70 34 49 18 67 49 50 21 67
## [501] 58 56 41 63 34 52 21 41 22 23 21 65 30 34 66 66 60 57 42 24 57 41 55 30 48
## [526] 46 31 24 36 37 22 21 40 43 29 34 27 35 24 55 41 63 63 33 70 49 31 42 54 55
## [551] 19 43 57 21 66 26 48 35 24 63 67 48 55
```

```
summary(age_img)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   45.00   44.45   59.00   70.00
```

```
df2$age <- age_img
head(df2)
```

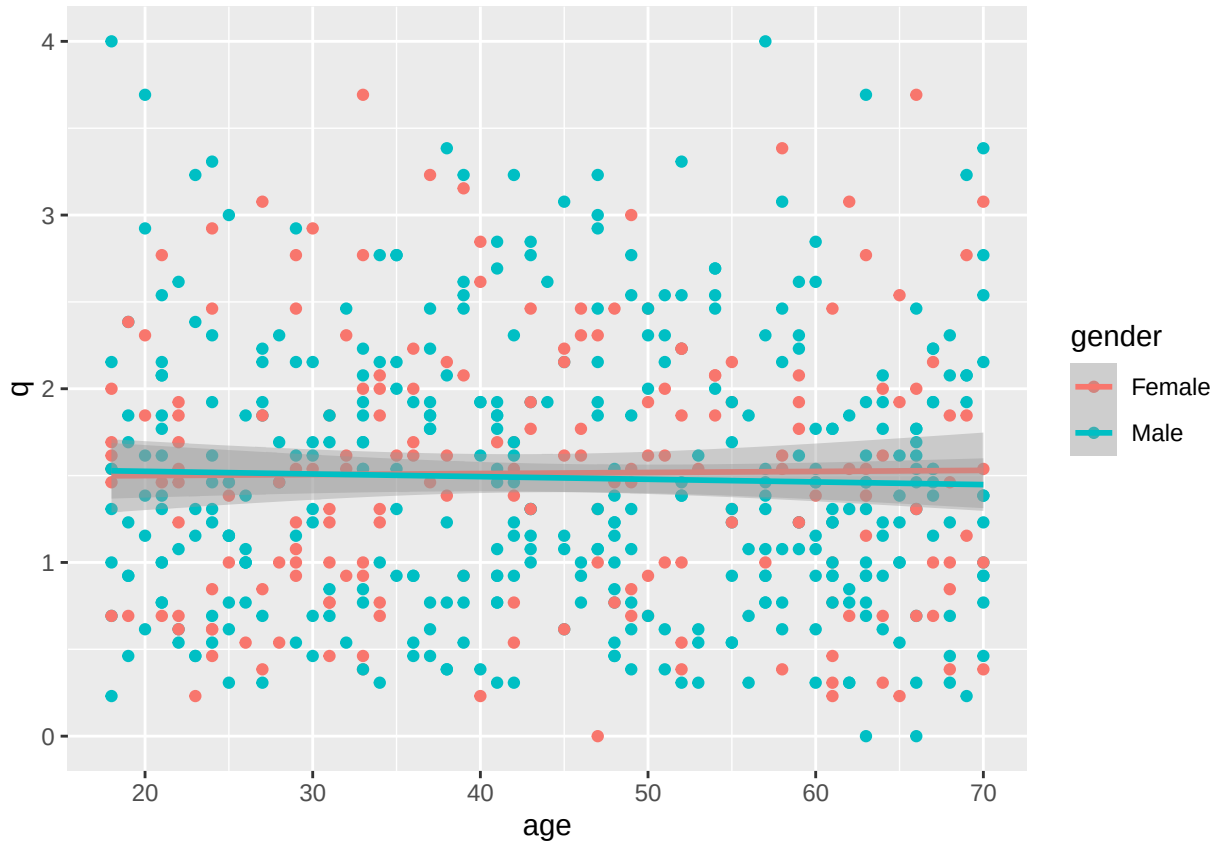
```
## # A tibble: 6 x 6
##   gender age_group education        familiarity     q   age
##   <fct>  <fct>     <fct>            <fct>       <dbl> <int>
## 1 Male   Adults    college_educated No          1.46     63
## 2 Male   Adults    college_educated No          2.08     38
## 3 Male   Adults    college_educated No          0.462    48
## 4 Female Youths    college_educated No          1.62     50
## 5 Male   Adults    college_educated No          1        43
## 6 Male   Adults    college_educated No          2.15     29
```

```
df2 <- df2 %>% relocate(age, .before = q)
head(df2)
```

```
## # A tibble: 6 x 6
##   gender age_group education        familiarity   age     q
##   <fct>  <fct>     <fct>            <fct>       <int> <dbl>
## 1 Male   Adults    college_educated No             63 1.46
## 2 Male   Adults    college_educated No             38 2.08
## 3 Male   Adults    college_educated No             48 0.462
## 4 Female Youths    college_educated No             50 1.62
## 5 Male   Adults    college_educated No             43 1
## 6 Male   Adults    college_educated No             29 2.15
```

```
df2 %>% ggplot(aes(x=age, y = q, color = gender))+
 geom_point()+geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



If we had a continuous/descrete predictor such as age, visualization would be more fun