Who falls for Misinformation?

Sanjay Chhetri

2021-12-05

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.

```
raw_data <- read_csv("covid_africa_data_untouched.csv")
```

```
## Rows: 563 Columns: 35
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (35): Timestamp, Gender, Age class, Highest level of education, Country ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
names(raw_data)
```

```
##  [1] "Timestamp"
##  [2] "Gender"
##  [3] "Age class"
##  [4] "Highest level of education"
##  [5] "Country of Origin"
##  [6] "Region of your country"
##  [7] "Religion faith/belief"
##  [8] "Do you know anyone/have seen anyone infected with the novel COVID19?"
##  [9] "Source of information about Coronavirus (COVID19). Please tick as many as applicable."
## [10] "If you will love to receive a copy of the published work, please include your email address bel
## [11] "Coronavirus (COVID19) is a man-made virus to reduce the human population"
## [12] "China made the novel Coronavirus (COVID19) in order to become the global power"
## [13] "African Blood compositions resist COVID19."
## [14] "Africa Weather and humid system prevent the spread of COVID19."
## [15] "Africans are naturally resilient and resistant to most diseases."
## [16] "Black skin resists COVID19."
## [17] "Older people are most prone to the danger of COVID19, younger people are less prone."
## [18] "Alcohol consumption can prevent/resist/kill COVID19."
## [19] "Drinking hot water can prevent/resist/kill COVID19."
```

```
## [20] "Smoking weeds (Cannabis) can prevent/resist/kill COVID19."
## [21] "High-temperature cure COVID19"
## [22] "COVID19 is a Bioweapon engineered by the Chinese/US or Bill Gate Foundation."
## [23] "COVID19 is a Digital contagion (the result of 5G, 60GHz network launched in Wuhan, China week
## [24] "Coronavirus (COVID19) has taught us humility and draw humanity closer to GOD."
## [25] "Rubbing of Anointing Oil (or prayer ablution) on the body can prevent/resist COVID19."
## [26] "Coronavirus (COVID19) cannot affect believers."
## [27] "Coronavirus (COVID19) cannot be contracted in the Church/Mosque."
## [28] "Coronavirus (COVID19) is a divine punishment from God to humanity."
## [29] "Social distancing in Africa is in theory, it cannot be practice."
## [30] "Total lockdown in Africa will cause a more dangerous hunger outbreak than the COVID19 itself."
## [31] "The media is causing more panicking about COVID19 than it should."
## [32] "There are many other dangerous diseases (e.g. Malaria, Cancer, etc) that kill more people daily
## [33] "Hunger kills more than COVID19."
## [34] "Thank you for your time! Please leave us with any comments about COVID19 you think we should in
## [35] "Kindly provide your name if you will want us to list your name in the appreciation section of
```

```
renamed_data <- raw_data %>%
  rename_with(.cols = 11:23, .fn = ~ paste0("q", 1:13)) %>%
  rename(familiarity = contains("Do you know anyone"))
names(renamed_data[,c(8,11:23)])
```

```
##  [1] "familiarity" "q1"          "q2"          "q3"          "q4"
##  [6] "q5"          "q6"          "q7"          "q8"          "q9"
## [11] "q10"         "q11"         "q12"         "q13"
```

The dataset was in csv file. After reading it in R, I subset it 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 <- renamed_data %>% mutate(gender = factor(Gender),
         age_group = factor(`Age class`),
         education = factor(`Highest level of education`),
         familiarity = factor(familiarity)) %>%
  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
```

```
table(df1$education)
```

```
##
##                       No formal education
##                                         2
##                       Post Graduate level
##                                       307
##                   Primary/Basic education
##                                         2
## Secondary/high school/form 4/5 education
##                                        33
##                        Tertiary education
##                                       219
```

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

table(df1$education)
```

```
##
##  Post_grad    College No_college
##        307        219         37
```

```
sum(is.na(renamed_data$`Highest level of education`)) #checking na in raw data
```

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

```
sum(is.na(df1$education)) #checking NA's after manipulation
```

```
## [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 three groups (post graduate level, college educated, and no college) facilitated a better 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    Post_grad No          0     1     1     4     4     0
## 2 Male   Adults    Post_grad No          4     1     1     1     4     1
## 3 Male   Adults    Post_grad No          0     0     0     1     0     0
## 4 Female Youths    Post_grad No          1     1     0     0     0     1
## 5 Male   Adults    College   No          1     1     1     1     1     1
## 6 Male   Adults    Post_grad 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    Post_grad No              0     1     1     4     4     0
## 2 Male   Adults    Post_grad No              4     1     1     1     4     1
## 3 Male   Adults    Post_grad No              0     0     0     1     0     0
## 4 Female Youths    Post_grad No              1     1     0     0     0     1
## 5 Male   Adults    College   No              1     1     1     1     1     1
## 6 Male   Adults    Post_grad 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    Post_grad No              0     1     1     4     4     0
## 2 Male   Adults    Post_grad No              4     1     1     1     4     1
## 3 Male   Adults    Post_grad No              0     0     0     1     0     0
## 4 Female Youths    Post_grad No              1     1     0     0     0     1
## 5 Male   Adults    College   No              1     1     1     1     1     1
## 6 Male   Adults    Post_grad 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)
```

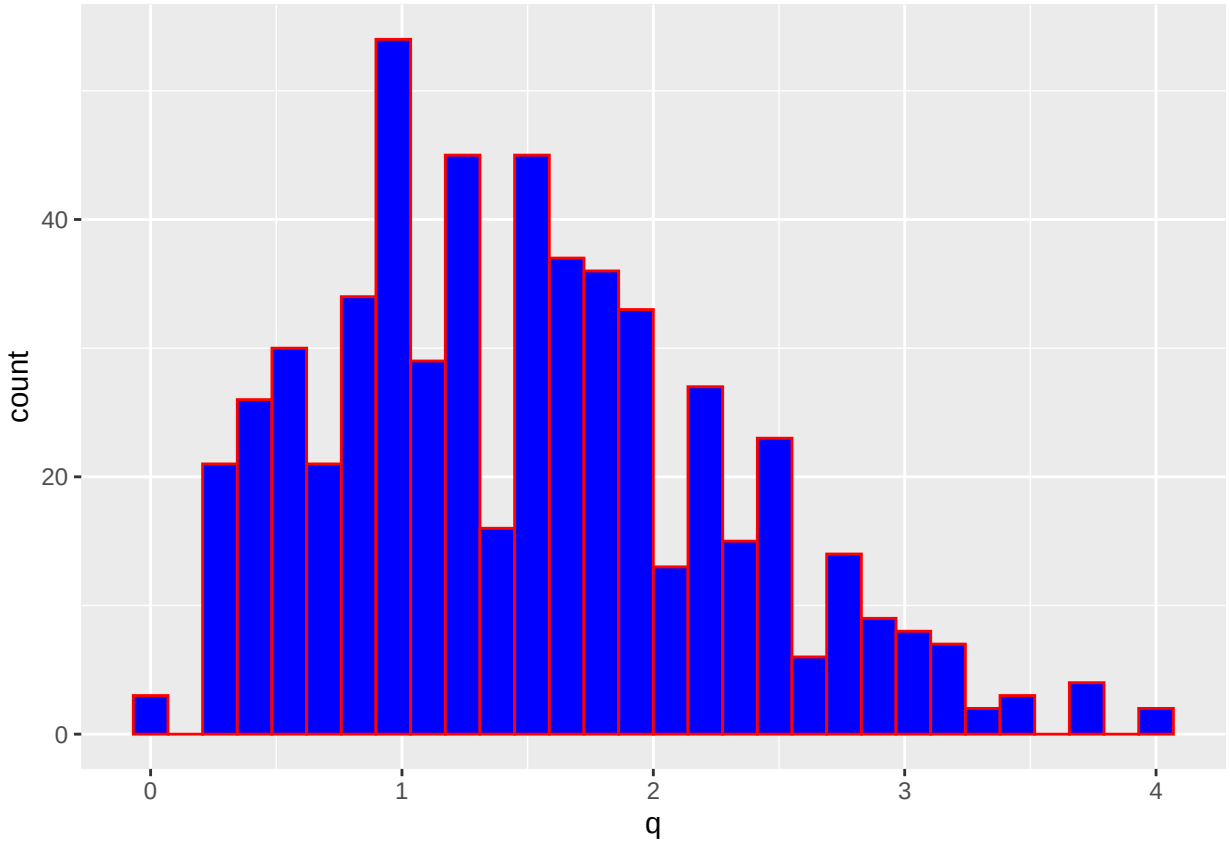
```
## # A tibble: 6 x 5
##   gender age_group education familiarity     q
##   <fct>  <fct>     <fct>     <fct>       <dbl>
## 1 Male   Adults    Post_grad No          1.46
## 2 Male   Adults    Post_grad No          2.08
## 3 Male   Adults    Post_grad No          0.462
## 4 Female Youths    Post_grad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

```
ggplot(df, aes(q))+geom_histogram(color = "red", fill="blue")
```

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



Regression

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

```
##
## Call:
## lm(formula = q ~ gender + age_group + education + familiarity,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45337 -0.60202 -0.07946  0.47185  2.53592
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.449079   0.079250  18.285  < 2e-16 ***
## genderMale          -0.009089   0.069888  -0.130    0.897
## age_groupAdults     -0.051024   0.073807  -0.691    0.490
## age_groupElderly     0.013376   0.120474   0.111    0.912
## educationCollege     0.044380   0.073319   0.605    0.545
## educationNo_college  0.653128   0.138735   4.708 3.17e-06 ***
## familiarityYes       0.075115   0.084150   0.893    0.372
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7687 on 556 degrees of freedom
## Multiple R-squared:  0.0437, Adjusted R-squared:  0.03339
## F-statistic: 4.235 on 6 and 556 DF,  p-value: 0.0003509
```

The regression model with 'q' as the outcome and gender, age_group, education, and familiarity as explanatory variables is significant, $F(6, 556) = 4.235$, $p<0.001$. The model explains 4.37% of the variance in the misinformation index (outcome, denoted by q), with adjusted R squared at 0.034. Education is the only variable that significantly predicts the misinformation index. With no significant difference between college and post-grad educated people. People with no college significantly differed in misinformation index. On average, people with no colleged measured 0.65 points higher in misinformation index compared to to the post graduated educated group.

##Is there Any interaction effect?

```
#check a suspect interaction
model1 <- lm(q~gender+age_group+education+familiarity+familiarity:education, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = q ~ gender + age_group + education + familiarity +
##     familiarity:education, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46139 -0.59211 -0.06798  0.47022  2.54741
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.448188   0.080618  17.963  < 2e-16 ***
## genderMale                        -0.008945   0.070007  -0.128    0.898
## age_groupAdults                   -0.043713   0.074194  -0.589    0.556
## age_groupElderly                   0.022145   0.120414   0.184    0.854
## educationCollege                   0.047345   0.080232   0.590    0.555
## educationNo_college                0.595787   0.143010   4.166 3.59e-05 ***
## familiarityYes                     0.057065   0.100233   0.569    0.569
## educationCollege:familiarityYes   -0.022818   0.187793  -0.122    0.903
## educationNo_college:familiarityYes  1.057279   0.567466   1.863    0.063 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7676 on 554 degrees of freedom
## Multiple R-squared:  0.0498, Adjusted R-squared:  0.03608
## F-statistic:  3.63 on 8 and 554 DF,  p-value: 0.0003951
```

Looks like there is not.

Some visualizations

Let the plots speak!

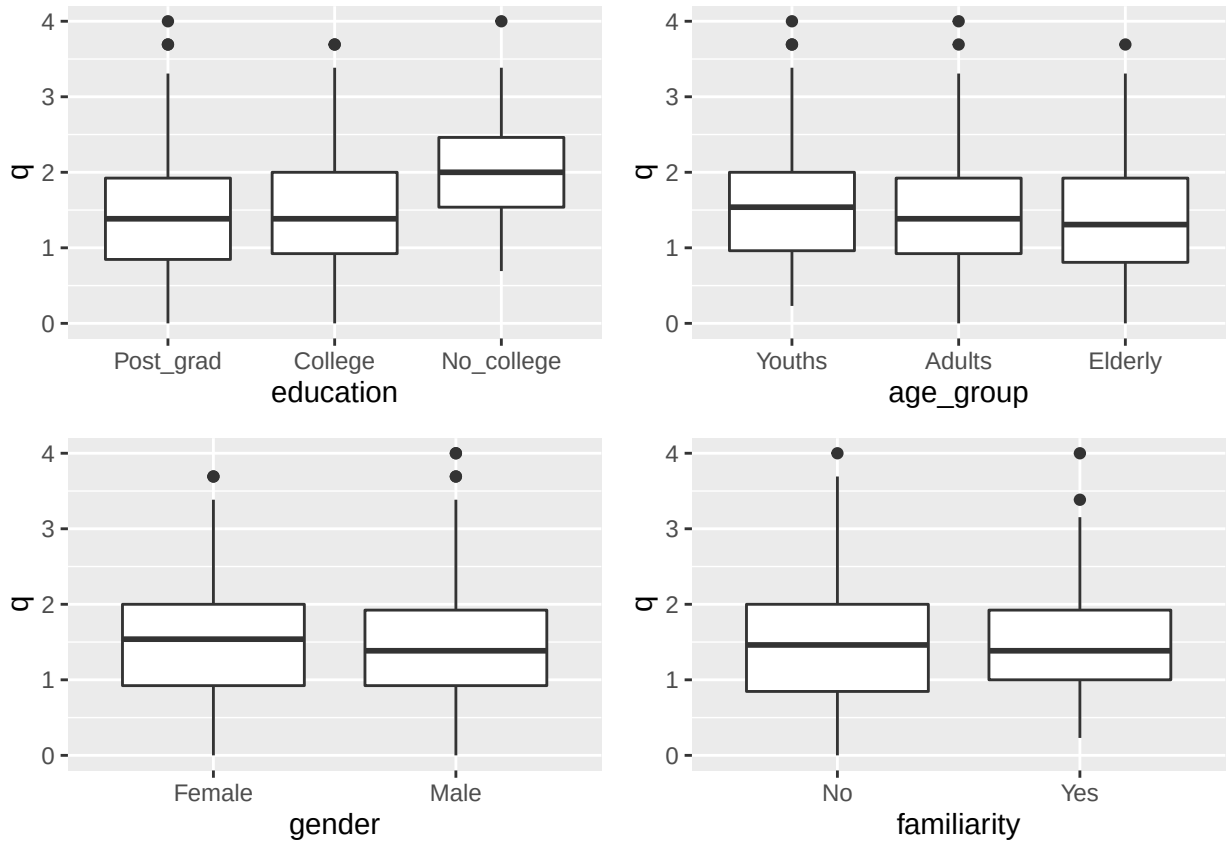
```
require(gridExtra)
```

```
## 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

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

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

```
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    Post_grad No             41 1.46
## 2 Male   Adults    Post_grad No             28 2.08
## 3 Male   Adults    Post_grad No             40 0.462
## 4 Female Youths    Post_grad No             19 1.62
## 5 Male   Adults    College   No             58 1
## 6 Male   Adults    Post_grad No             28 2.15
```

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

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