Exploratory Data Analysis

Zachary Himmelberger

2/15/2022

Import Packages

```
library(knitr)
library(psych)
```

Import Data

```
source("Data-Processing.R")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Exploratory Data Analysis

We will first try to get a basic understanding of the DataFrame.

```
glimpse(analysis.df)
```

```
## Rows: 221
## Columns: 30
## $ age                       <dbl> 24.99178, 19.68767, 18.56438, NA, 19.64110, ~
## $ gender                    <chr> "male", "male", "female", "female", "non-bin~
## $ ethnicity                 <chr> "white", "white", "black", "white", "white",~
## $ attitude                  <dbl> 6.793103, 6.172414, 7.655172, 6.448276, 7.79~
## $ integration               <dbl> 5.714286, 4.428571, 6.857143, 6.571429, 7.42~
## $ social.distance           <dbl> 7.125, 7.250, 8.375, 7.500, 8.375, 8.250, 8.~
## $ private.rights            <dbl> 6.571429, 5.857143, 7.714286, 5.428571, 8.28~
## $ subtle.derogatory.beliefs <dbl> 7.714286, 7.000000, 7.571429, 6.142857, 7.00~
## $ male                      <dbl> 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ non.white                 <dbl> 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0,~
## $ gender.male               <dbl> 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
```

```
## $ gender.non.binary         <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ social.desirability       <int> 18, 15, 18, 20, 12, 15, 20, 22, 25, 22, 20, ~
## $ openness                  <dbl> 4.3, 3.9, 4.6, 3.6, 3.8, 2.5, 4.5, 3.1, 3.9,~
## $ conscientiousness         <dbl> 4.333333, 3.777778, 3.444444, 3.555556, 3.55~
## $ extroversion              <dbl> 3.625, 3.500, 3.375, 2.875, 1.875, 2.125, 2.~
## $ agreeableness             <dbl> 4.444444, 3.555556, 3.888889, 4.000000, 3.55~
## $ neuroticism               <dbl> 3.125, 1.750, 3.500, 3.750, 4.375, 3.750, 3.~
## $ quantity                  <dbl> 1.600000, 2.000000, 7.714286, 2.300000, 5.20~
## $ quality                   <dbl> 3.714286, 6.428571, 8.000000, 6.571429, 6.14~
## $ knowledge                 <dbl> 2.500000, 2.875000, 6.312500, 2.200000, 5.62~
## $ social.desirability.c     <dbl> 1.530172, -1.469828, 1.530172, 3.530172, -4.~
## $ openness.c                <dbl> 0.83405172, 0.43405172, 1.13405172, 0.134051~
## $ conscientiousness.c       <dbl> 0.61590038, 0.06034483, -0.27298851, -0.1618~
## $ extroversion.c            <dbl> 0.244612069, 0.119612069, -0.005387931, -0.5~
## $ agreeableness.c           <dbl> 0.42289272, -0.46599617, -0.13266284, -0.021~
## $ neuroticism.c             <dbl> 0.17510776, -1.19989224, 0.55010776, 0.80010~
## $ quantity.c                <dbl> -3.31038717, -2.91038717, 2.80389854, -2.610~
## $ quality.c                 <dbl> -3.48208617, -0.76780045, 0.80362812, -0.624~
## $ knowledge.c               <dbl> -2.1078120, -1.7328120, 1.7046880, -2.407812~
```

We can now begin to examine the variables. Let's start with some basic summaries.

```
summary(df)
```

```
##     gender           ethnicity              age        social.desirability
##  Length:232         Length:232         Min.   :18.01   Min.   : 3.00
##  Class :character   Class :character   1st Qu.:18.50   1st Qu.:12.00
##  Mode  :character   Mode  :character   Median :18.81   Median :17.00
##                                        Mean   :19.10   Mean   :16.47
##                                        3rd Qu.:19.47   3rd Qu.:20.00
##                                        Max.   :25.41   Max.   :30.00
##                                        NA's   :14
##     openness     conscientiousness  extroversion   agreeableness
##  Min.   :1.800   Min.   :2.222     Min.   :1.500   Min.   :2.444
##  1st Qu.:3.100   1st Qu.:3.333     1st Qu.:2.875   1st Qu.:3.639
##  Median :3.500   Median :3.778     Median :3.375   Median :4.111
##  Mean   :3.466   Mean   :3.717     Mean   :3.380   Mean   :4.022
##  3rd Qu.:3.800   3rd Qu.:4.111     3rd Qu.:4.000   3rd Qu.:4.444
##  Max.   :4.700   Max.   :5.000     Max.   :5.000   Max.   :5.000
##
##   neuroticism       attitude      integration    social.distance
##  Min.   :1.000   Min.   :4.172   Min.   :2.143   Min.   :4.375
##  1st Qu.:2.375   1st Qu.:6.310   1st Qu.:5.143   1st Qu.:7.125
##  Median :3.000   Median :7.034   Median :6.143   Median :8.000
##  Mean   :2.950   Mean   :6.961   Mean   :6.099   Mean   :7.789
##  3rd Qu.:3.500   3rd Qu.:7.759   3rd Qu.:7.143   3rd Qu.:8.656
##  Max.   :4.750   Max.   :8.793   Max.   :9.000   Max.   :9.000
##
##  private.rights  subtle.derogatory.beliefs    quality         quantity
##  Min.   :4.000   Min.   :3.714             Min.   :1.000   Min.   :1.600
##  1st Qu.:6.000   1st Qu.:6.000             1st Qu.:6.429   1st Qu.:3.900
##  Median :7.000   Median :6.857             Median :7.571   Median :5.000
##  Mean   :7.012   Mean   :6.828             Mean   :7.196   Mean   :4.910
##  3rd Qu.:7.893   3rd Qu.:7.571             3rd Qu.:8.429   3rd Qu.:6.075
##  Max.   :9.000   Max.   :9.000             Max.   :9.000   Max.   :8.400
```

```
##                                          NA's   :1       NA's   :10
##    knowledge
##  Min.   :1.000
##  1st Qu.:3.359
##  Median :4.562
##  Mean   :4.608
##  3rd Qu.:5.803
##  Max.   :8.800
##
```

Age and quantity have some missing values. Otherwise, nothing much stands out.

Let's look at a correlation matrix for the quantitative variables. Note that anything about .13 is statistically significant.

```
cor.matrix <- df %>%
  select(where(is.numeric)) %>%
  cor(use="pairwise.complete.obs") %>%
  round(2)

kable(cor.matrix, format="html", align="c")
```

age

social.desirability

openness

conscientiousness

extroversion

agreeableness

neuroticism

attitude

integration

social.distance

private.rights

subtle.derogatory.beliefs

quality

quantity

knowledge

age

1.00

0.02

0.12

0.05

0.08

0.01

0.01

0.01

0.02

-0.03

-0.01

0.04

-0.16

-0.05

-0.08

social.desirability

0.02

1.00

0.15

0.42

0.27

0.47

-0.45

0.21

0.14

0.20

0.22

0.14

0.32

0.21

0.25

openness

0.12

0.15

1.00

0.00

0.12

0.15

-0.07

0.27

0.26

0.20

0.23

0.20

0.21

0.25

0.22

conscientiousness

0.05

0.42

0.00

1.00

0.23

0.32

-0.27

0.12

0.03

0.18

0.10

0.11

0.19

0.07

0.06

extroversion

0.08

0.27

0.12

0.23

1.00

0.23

-0.25

0.15

0.14

0.11

0.12

0.10

0.23

0.25

0.24

agreeableness

0.01

0.47

0.15

0.32

0.23

1.00

-0.38

0.35

0.28

0.37

0.29

0.21

0.39

0.17

0.18

neuroticism

0.01

-0.45

-0.07

-0.27

-0.25

-0.38

1.00

-0.12

-0.05

-0.09

-0.14

-0.14

-0.10

0.00

-0.07

attitude

0.01

0.21

0.27

0.12

0.15

0.35

-0.12

1.00

0.83

0.82

0.81

0.80

0.56

0.42

0.43

integration

0.02

0.14

0.26

0.03

0.14

0.28

-0.05

0.83

1.00

0.55

0.54

0.56

0.45

0.38

0.38

social.distance

-0.03

0.20

0.20

0.18

0.11

0.37

-0.09

0.82

0.55

1.00

0.59

0.55

0.57

0.37

0.38

private.rights

-0.01

0.22

0.23

0.10

0.12

0.29

-0.14

0.81

0.54

0.59

1.00

0.55

0.41

0.35

0.36

subtle.derogatory.beliefs

0.04

0.14

0.20

0.11

0.10

0.21

-0.14

0.80

0.56

0.55

0.55

1.00

0.40

0.25

0.27

quality

-0.16

0.32

0.21

0.19

0.23

0.39

-0.10

0.56

0.45

0.57

0.41

0.40

1.00

0.59

0.60

quantity

-0.05

0.21

0.25

0.07

0.25

0.17

0.00

0.42

0.38

0.37

0.35

0.25

0.59

1.00

0.69

knowledge

-0.08

0.25

0.22

0.06

0.24

0.18

-0.07

0.43

0.38

0.38

0.36

0.27

0.60

0.69

1.00

Obviously there is a lot to go through with all of those correlations. Let's examine the correlation table just among the personality variables and attitude.

```
cor.matrix <- df %>%
  select(attitude, openness, conscientiousness, extroversion,
         agreeableness, neuroticism) %>%
  cor(use="pairwise.complete.obs") %>%
  round(2)

kable(cor.matrix, format="html", align="c")
```

attitude

openness

conscientiousness

extroversion

agreeableness

neuroticism

attitude

1.00

0.27

0.12

0.15

0.35

-0.12

openness

0.27

1.00

0.00

0.12

0.15

-0.07

conscientiousness

0.12

0.00

1.00

0.23

0.32

-0.27

extroversion

0.15

0.12

0.23

1.00

0.23

-0.25

agreeableness

0.35

0.15

0.32

0.23

1.00

-0.38

neuroticism

-0.12

-0.07

-0.27

-0.25

-0.38

1.00

Attitude is moderately correlated with openness and agreeableness. It is also weakly correlated with extroversion. Let's look at how the attitude subscales correlate with the total score.

```
cor.matrix <- df %>%
  select(attitude, integration, social.distance, private.rights,
         subtle.derogatory.beliefs) %>%
  cor(use="pairwise.complete.obs") %>%
  round(2)

kable(cor.matrix, format="html", align="c")
```

attitude

integration

social.distance

private.rights

subtle.derogatory.beliefs

attitude

1.00

0.83

0.82

0.81

0.80

integration

0.83

1.00

0.55

0.54

0.56

social.distance

0.82

0.55

1.00

0.59

0.55

private.rights

0.81

0.54

0.59

1.00

0.55

subtle.derogatory.beliefs

0.80

0.56

0.55

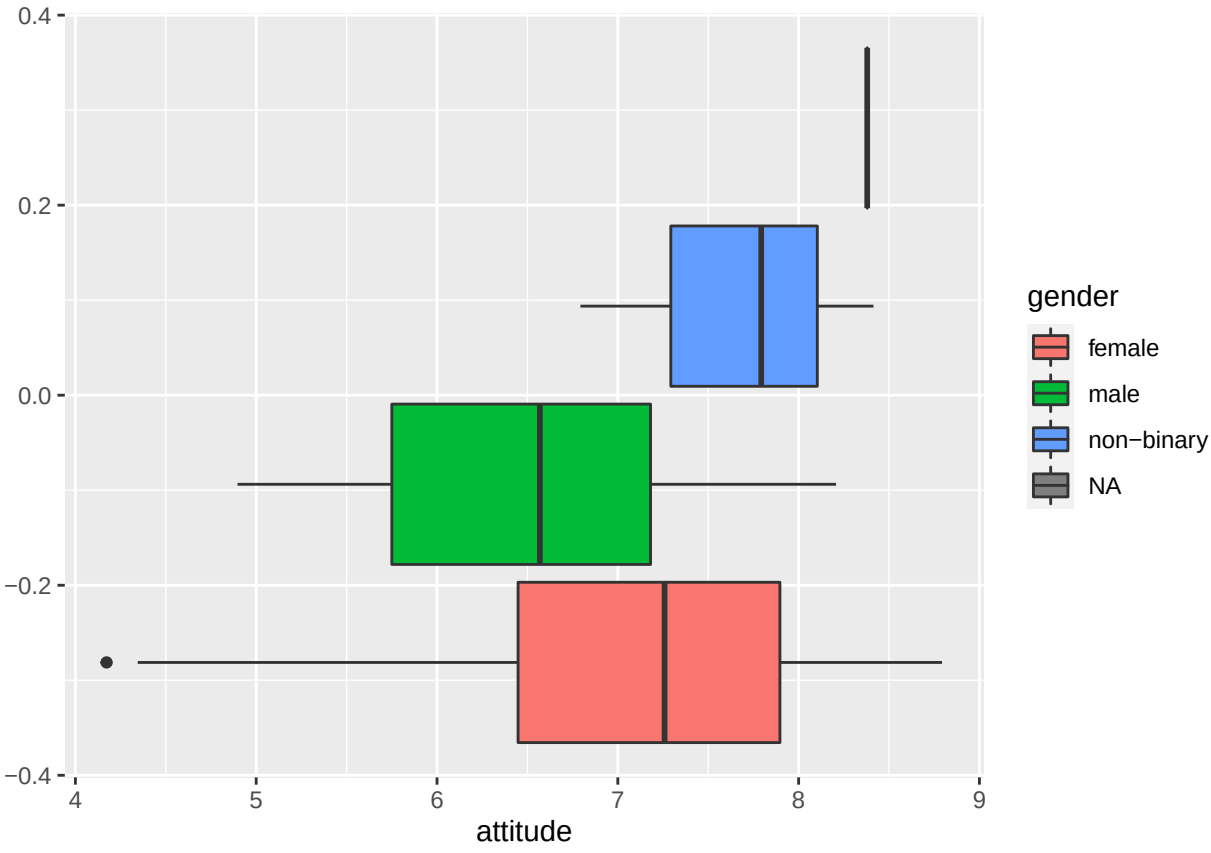
0.55

1.00

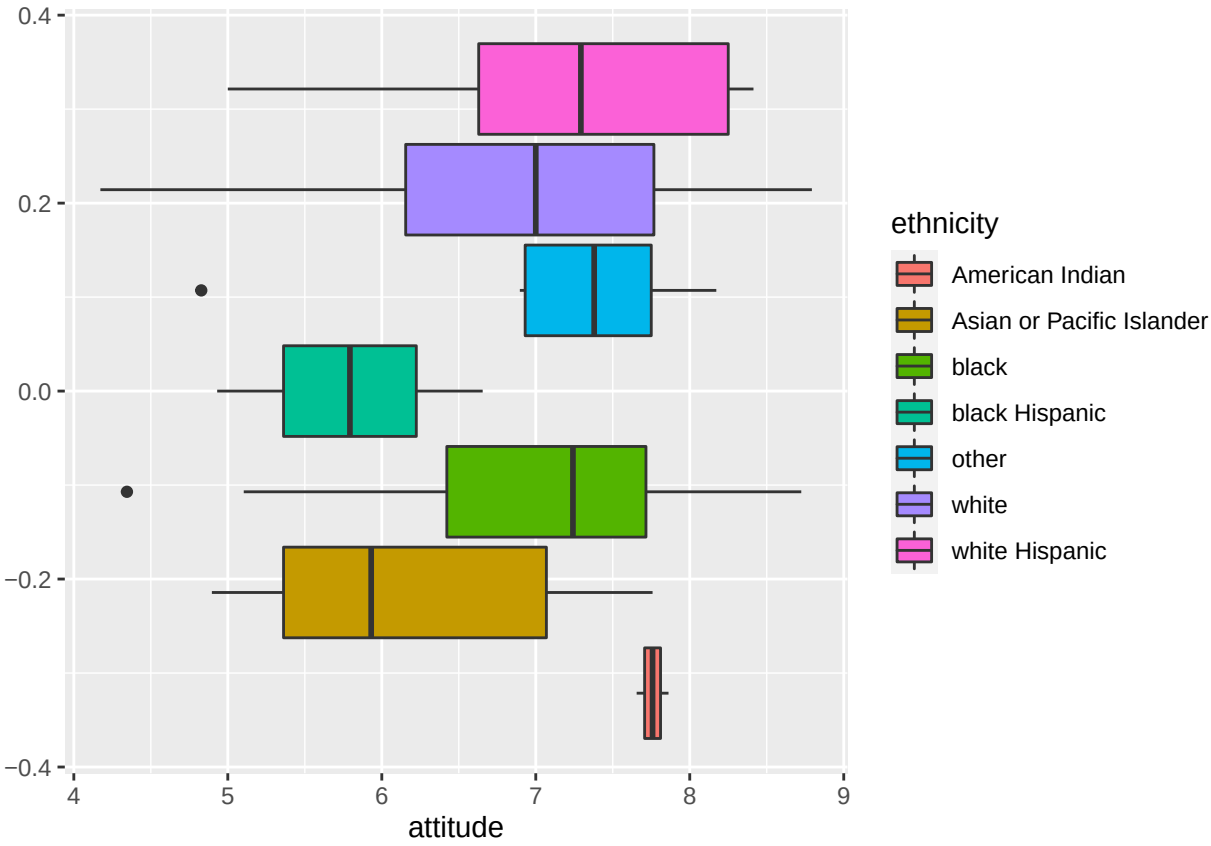
All four subscales correlate very strongly (greater than .8) with the total score and strongly (greater than .5) with each other.

We should also look at the potential covariates. The two that come to mind based on previous research are gender (McManus et al.) and race (Keith et al.).

```
ggplot(data=df, aes(x=attitude, fill=gender)) +
  geom_boxplot()
```




```
ggplot(data=df, aes(x=attitude, fill=ethnicity)) +
  geom_boxplot()
```



It seems like gender may be a useful covariate. It seems to be related to attitudes. It may also be worthwhile to include race/ethnicity as a covariate. However, two groups make up most of our data. Combining the other ethnicities as their own group makes less sense, given how spread out they are. We may want to collapse across some groups (or white v. non-white as Keith et al. did), if only for practical reasons.