Ghana Analysis

Colin Heeneman

2025-03-04

Table of contents

1 General Instructions

You have 24 hours to complete this test. All necessary files for completing the test are included in the attached zip file. When you submit the test, we want to see a complete record of your work. Please attach do-files and any other files you think are necessary to evaluate your work.

Remember that you will be graded not only on your answers, but also on your process (such as the efficiency of your code and comments). This test is designed to assess logical reasoning and problem-solving in addition to coding skills, so if you're unable to answer a question using Stata, please explain what steps you would have taken had you known the appropriate commands. In general, remember to explain your reasoning and give as complete an answer as you can.

Follow coding best practices to ensure your code is efficient, easily readable, and the results are reproducible even if the data set changes over time.

Using resources such as Stata's internal help files and online resources is encouraged (and often necessary!), but please cite the resources you use, and do not consult with other people.

We have provided you with survey data collected from households in Ghana in two waves. These households were part of a Randomized Controlled Trial (RCT), where participants in treated households received Group Therapy to alleviate symptoms of depression.

To explore whether Group Therapy (GT) alleviates depression or not, an RCT was conducted in partnership with a local organization that provides GT sessions designed to improve mental health through guided discussions led by a trained facilitator. After Wave 1 of data collection was complete, half of the households in the sample were randomly selected to receive the opportunity to attend weekly GT sessions for a period of three months at no cost to participants. In each treatment household, the household head and their spouse were invited to GT sessions. For the purposes of this analysis, assume that there was perfect attendance at these sessions. Once all GT sessions were completed, Wave 2 data were collected (6 months after Wave 1).

This is a hypothetical treatment created for this exercise. The Ghana Panel Survey data has been slightly modified for this purpose.

Through the exercises below, you will prepare the data, conduct exploratory analysis, and present findings for the RCT.

There are three data sets, which include data collected at Wave 1 (before the intervention), and Wave 2 (after the intervention). The data sets are as follows:

- *Demographics: This data set includes treatment assignment at the household level and demographic information for each member in the sampled households.*
- *Assets: This data set includes the quantity and monetary value of assets owned by the household, under three categories of assets: Animals, Tools, and Durable Goods.*

- *Depression Information: This data set includes information for the Kessler Psychological Distress Scale collected for household heads and their spouses.*

Note that these are imperfect data sets. Not all the variables are cleaned, so watch out for things like missing values, ordering of categories, and strings.

2 Part 1

```
# setting up for analysis by loading necessary libraries and data

# loading the libraries I will need for this analysis
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.1
```

```
Warning: package 'dplyr' was built under R version 4.3.1
```

```
Warning: package 'stringr' was built under R version 4.3.1
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(haven)
library(dplyr)
library(ggplot2)
library(summarytools)
```

```
Attaching package: 'summarytools'
```

```
The following object is masked from 'package:tibble':

    view
```

```r
library(readr)
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```r
 # loading raw data sets naming them _raw
assets_raw <- read_dta("data/assets.dta")
demographics_raw <- read_dta("data/demographics.dta")
depression_raw <- read_dta("data/depression.dta")


 # check the structure of the data sets
str(assets_raw)
str(demographics_raw)
str(depression_raw)

 # Inspect the first few rows of each data set
head(assets_raw)
head(demographics_raw)
head(depression_raw)
```

2.1 Q1:

At what level is each data set uniquely identified (i.e., what does each row represent, and which variables are unique identifiers)?

```r
# In order to answer this question, we need to inspect the structure of each datase and ID u
```

Q1 ANSWER:

- Asset: Rows are individual **assets** owned by households. Unique identifiers are survey wave (*wave*) and household ID (*hhid*).

- Demographics: Rows are demographic characteristics of **individuals**. Unique identifier is a combination of survey wave (*wave*) household ID (*hhid*) and household member ID (*hhmid*).
- Depression: Rows are **individuals** once again, this time their responses on mental health survey in wave X. Unique identifier is a combination of survey wave (*wave*) household ID (*hhid*) and household member ID (*hhmid*).

Demographics

2.2 Q2:

Import the demographics data set and calculate a proxy variable for household size based on the number of members surveyed in each household in Wave 1. Assume the household size for Wave 2 remains the same.

```
 # calculate household size proxy
demographicsv1 <- demographics_raw %>%
  group_by(hhid, wave) %>%
  mutate(hh_size_proxy = ifelse(wave == 1, n(), NA)) %>%
  group_by(hhid) %>%
  mutate(hh_size_proxy = max(hh_size_proxy, na.rm = TRUE)) %>%
  ungroup()

 # display a summary of this variable
summary(demographicsv1$hh_size_proxy)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   5.000   5.244   7.000  20.000
```

Q2 ANSWER:

Household size was determined here using count from Wave 1 data, and assuming it holds constant for Wave 2. Unsurprisingly, the minimum value was 1 with the median being 5 and maximum of 20!

Assets

Asset data is useful as a proxy for wealth. This data set contains information on the quantity and monetary value of each individual asset owned by surveyed households.

Through the questions below, you will aggregate this information to estimate a monetary value for three asset categories.

2.3 Q3:

To calculate the monetary value of all assets, you should use the 'currentvalue' variable, which reports the monetary value of a single unit of the asset. However, you will notice that this variable is often missing. Please use the median of "current- value" for each type of asset (by type we mean, for example, "chickens", "Cutlass", "Room Furniture", "Radio", "Cell (mobile) Phone handset", etc.) to impute the missing values.

```
# impute missing 'currentvalue' with the median for each asset type
assetsv1 <- assets_raw %>%
  group_by(Asset_Type) %>%
  mutate(currentvalue = ifelse(is.na(currentvalue), median(currentvalue, na.rm = TRUE), curre
  ungroup()

# display summary of the imputed 'currentvalue'
summary(assetsv1$currentvalue)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
     0.0      15.0      50.0     175.3      50.0  100000.0 
```

Q3 ANSWER:

Missing values for *currentvalue* of assets were imputed by using the median for each asset type. Overall, the median asset value was 50 units and the maximum value was 100,000.

2.4 Q4:

Create a variable that contains the total monetary value for each observation, by multiplying quantity and the imputed current value.

```
# create total monetary value variable
assetsv2 <- assetsv1 %>%
  mutate(total_value = quantity * currentvalue)

# display summary of the new variable
summary(assetsv2$total_value)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
       0        30        50      1383       150  10000000 
```

Q4 ANSWER:

After creating a column for total asset value for each asset observation, the median remained at 50 units and maximum increased to 10 million.

2.5 Q5:

Produce a data set at the household-wave level (for each household, there should be at most two observations, one for each wave) which contains the following variables:

- *Household ID (hhid)*
- *Wave ID (wave)*
- *Total value of animals*
- *Total value of tools*
- *Total value of durable goods*
- *Total asset value*

```r
# categorize assets into animals, tools, and durable goods
assetsv3 <- assetsv2 %>%
  mutate(category = case_when(
    !is.na(animaltype) ~ "animals",
    !is.na(toolcode) ~ "tools",
    !is.na(durablegood_code) ~ "durable_goods",
    TRUE ~ "other"
  ))

# aggregate data at household-wave level
assetsv4 <- assetsv3 %>%
  group_by(hhid, wave) %>%
  summarise(
    total_value_animals = sum(total_value[category == "animals"], na.rm = TRUE),
    total_value_tools = sum(total_value[category == "tools"], na.rm = TRUE),
    total_value_durable_goods = sum(total_value[category == "durable_goods"], na.rm = TRUE),
    total_asset_value = sum(total_value, na.rm = TRUE),
    .groups = "drop"
  )

# display the processed data set
print(assetsv4)
```

```
# A tibble: 9,357 x 6
   hhid        wave total_value_animals total_value_tools total_value_durable_~1
   <chr>      <dbl>               <dbl>             <dbl>                  <dbl>
 1 0101001002     1                 250               8                     1500
 2 0101001002     2                 180              11.5                   9325
 3 0101001003     1                   0             269                     1614.
 4 0101001003     2                   0             774                    18382
 5 0101001004     1                1000              48                     2550
 6 0101001004     2                   0              92                     2413
 7 0101001009     1                   0               4                      950
 8 0101001009     2                   0              94                   103311
 9 0101001010     1                   0               6                     2900
10 0101001010     2                   0              20                    44945
# i 9,347 more rows
# i abbreviated name: 1: total_value_durable_goods
# i 1 more variable: total_asset_value <dbl>
```

Q5 ANSWER:

Data set has been restructured such that each household's asset values were aggregated in the three categories: animals, tools, and durable goods.

Mental Health / Depression Data

2.6 Q6:

A Kessler-10 scale is a measure of mental health that uses 10 questions that identify how often people experience symptoms associated with depression. Using this reference, construct the kessler score (name it kessler_score) and a categorical variable named kessler categories with 4 categories: no significant depression, milddepression, moderate depression, and severe depression.Be careful: sometimes variables in the Kessler section are missing, which can complicate construction of a score. Please justify, using comments in your code, how you handle missing values.

```
 # define the Kessler-10 variables
kessler_vars <- c("tired", "nervous", "sonervous", "hopeless", "restless",
                  "sorestless", "depressed", "everythingeffort", "nothingcheerup", "worthles

 # handle missing values:
 # if at least 8 out of 10 responses are available, replace missing with the mean of
depressionv1 <- depression_raw %>%
  rowwise() %>%
  mutate(
    available_values = sum(!is.na(c_across(all_of(kessler_vars)))),
    kessler_score = ifelse(available_values >= 8,
                           sum(c_across(all_of(kessler_vars)), na.rm = TRUE) * 10 / availabl
                           NA_real_)
  ) %>%
  ungroup() %>%
  select(-available_values) # Remove intermediate calculation column

 # categorize Kessler scores into depression levels
mentalhealth <- depressionv1 %>%
  mutate(
    kessler_categories = case_when(
      kessler_score < 20 ~ "No significant depression",
      kessler_score >= 20 & kessler_score < 25 ~ "Mild depression",
      kessler_score >= 25 & kessler_score < 30 ~ "Moderate depression",
      kessler_score >= 30 ~ "Severe depression",
      TRUE ~ NA_character_  # Assign NA if score couldn't be computed
    )
  )

 # display summary of results
summary(mentalhealth$kessler_score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  10.00   14.00   18.00   18.92   23.00   50.00      28
```

```
table(mentalhealth$kessler_categories, useNA = "ifany")
```

```

          Mild depression       Moderate depression No significant depression
                     3190                      1651                      8119
        Severe depression                      <NA>
                      854                        28
```

Q6 ANSWER:

Kessler-10 score was calculated using the depression data set and categories created for ‘No significant depression’, ‘Mild depression’, ‘Moderate depression’ and ‘Severe depression’. Kessler 10 scores ranged from the minimum scale value of 10 to maximum scale value of 50, with the median being 18. A plurality of individuals fall into the 'No significant depression' category with the second most common being 'Mild depression'.

Constructing a single data set

2.7 Q7:

At this point you have created three data sets: demographics, assets, and mentalhealth. Please combine all three of these data sets to create a single data set that you will use for data exploration and analysis. The unit of observation in this data set should be an individual in a given survey round. (There should be at most two observations per individual, one for Wave 1 and another for Wave 2).

```
# first I will convert hhid to character in all data sets to ensure compatibility

  ## doing this because initially got this error

   #Error in `left_join()`:
   #! Can't join `x$hhid` with `y$hhid` due to incompatible types.
   #  `x$hhid` is a <double>.
   #  `y$hhid` is a <character>.

 # ensuring all key columns (`hhid`, `hhmid`, `wave`) have the same type so that they are

demographicsv2 <- demographicsv1 %>% mutate(across(c(hhid, hhmid, wave), as.character))
mentalhealthv2 <- mentalhealth %>% mutate(across(c(hhid, hhmid, wave), as.character))
assetsv5 <- assetsv4 %>% mutate(across(c(hhid, wave), as.character))

 # standardize hhid formatting, ensure leading zeros are consistent as the assets data

demographics <- demographicsv2 %>% mutate(hhid = str_pad(hhid, width = max(nchar(assetsv5$hhi
mentalhealthv3 <- mentalhealthv2 %>% mutate(hhid = str_pad(hhid, width = max(nchar(assetsv5$h
assets <- assetsv5 %>% mutate(hhid = str_pad(hhid, width = max(nchar(demographicsv2$hhid)), p

 # make use of inner_join() to merge datasets based on relevant keys

mergedv1<- inner_join(demographics, mentalhealthv3, by = c("hhid", "hhmid", "wave"))
mergedv2 <- inner_join(mergedv1, assets, by = c("hhid", "wave"))

 # print the first few rows to check if all values are merging correctly
head(mergedv2)
```

```
# A tibble: 6 x 36
  wave  hhid   hhmid villageid treat_hh gender    age relationship maritalstatus
  <chr> <chr>  <chr>     <dbl> <dbl+lb> <dbl+l> <dbl> <dbl+lbl>    <dbl+lbl>
1 1     01010~ 1             1 0 [Cont~ 1 [Mal~    35 1 [Househol~ 1 [Married]
2 1     01010~ 2             1 0 [Cont~ 5 [Fem~    31 2 [Spouse]   1 [Married]
3 2     01010~ 1             1 0 [Cont~ 1 [Mal~    35 1 [Househol~ 1 [Married]
4 2     01010~ 2             1 0 [Cont~ 5 [Fem~    31 2 [Spouse]   1 [Married]
5 1     01010~ 1             1 1 [Trea~ 1 [Mal~    28 1 [Househol~ 1 [Married]
6 2     01010~ 1             1 1 [Trea~ 1 [Mal~    29 1 [Househol~ 1 [Married]
# i 27 more variables: spouseinhouse <dbl+lbl>, agemarried <dbl>,
#   religion <dbl+lbl>, religionother <chr>, fatherinhouse <dbl+lbl>,
#   fathereduc <dbl+lbl>, fathereducother <chr>, motherinhouse <dbl+lbl>,
#   mothereduc <dbl+lbl>, mothereducother <chr>, hh_size_proxy <int>,
#   tired <dbl+lbl>, nervous <dbl+lbl>, sonervous <dbl+lbl>,
#   hopeless <dbl+lbl>, restless <dbl+lbl>, sorestless <dbl+lbl>,
#   depressed <dbl+lbl>, everythingeffort <dbl+lbl>, ...
```

Q7 ANSWER:

Demographics, assets, and mentalhealth have been merged into a single data set 'mergedv2'. Our ultimate data set for analysis contains the individual as the unit of observation, with a maximum of two observations for each (wave 1 and wave 2).

3 Part 2: Exploratory Analysis

Using Wave 1 data, conduct exploratory analysis to understand the relationship between depression and household and demographic characteristics among individuals in Ghana. Specifically, do the following:

3.1 Q1:

Explore the relationship between depression and:

1. *Household wealth, proxied by total asset value.*

```
 # just wave 1 data

# Filter for Wave 1 data
wave1 <- mergedv2 %>%
  filter(wave == 1)
```

```r
# Check for missing values in key variables
summary(wave1$kessler_score)
```

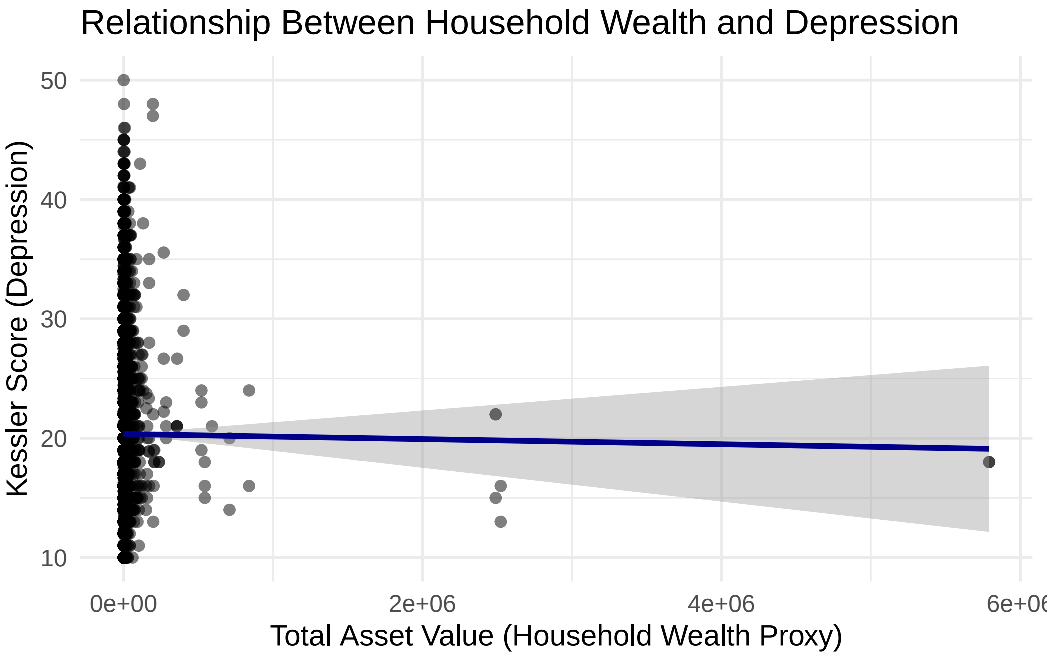
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  10.00   15.00   20.00   20.35   24.00   50.00      28 
```

```r
summary(wave1$total_asset_value)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      2    1400    2550   11152    4961 5791358 
```

```r
# Scatter plot: Kessler Score vs. Household Wealth
ggplot(wave1, aes(x = total_asset_value, y = kessler_score)) +
  geom_point(alpha = 0.5) +  # Transparency to avoid overplotting
  geom_smooth(method = "lm", color = "darkblue", se = TRUE) +  # Linear trend line
  labs(
    title = "Relationship Between Household Wealth and Depression",
    x = "Total Asset Value (Household Wealth Proxy)",
    y = "Kessler Score (Depression)"
  ) +
  theme_minimal()
```

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



```r
# Correlation test between wealth and depression
cor.test(wave1$total_asset_value, wave1$kessler_score, use = "complete.obs")
```

```

	Pearson's product-moment correlation

data:  wave1$total_asset_value and wave1$kessler_score
t = -0.34813, df = 7410, p-value = 0.7278
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02680802  0.01872392
sample estimates:
         cor
-0.004044145
```

```r
# Linear regression: Predicting depression based on household wealth
model.wealth <- lm(kessler_score ~ total_asset_value, data = wave1)
summary(model.wealth)
```

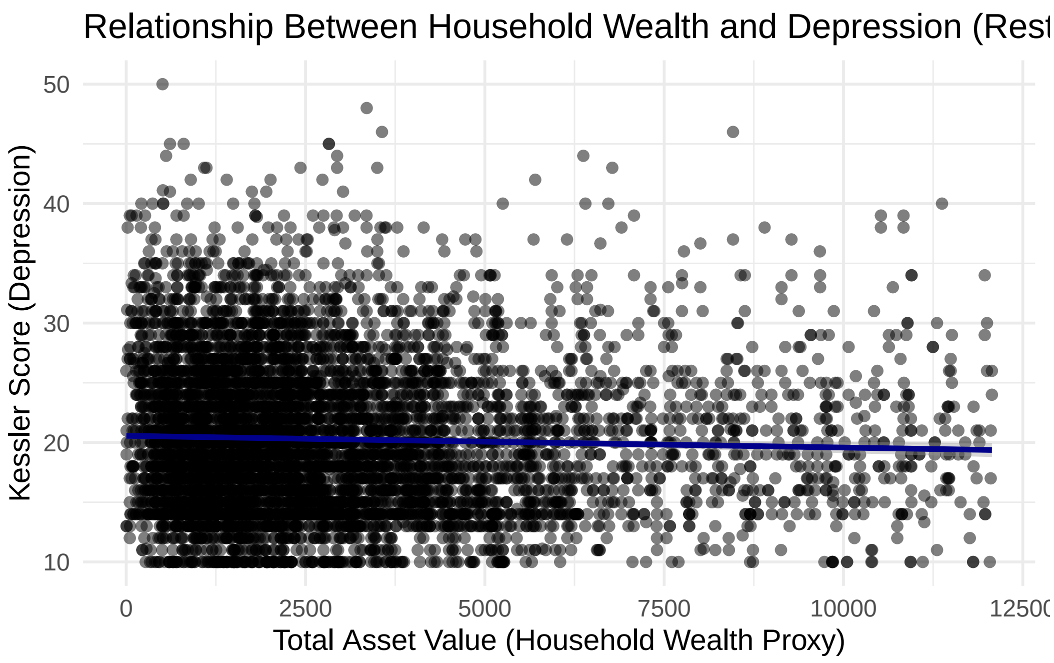
```

Call:
lm(formula = kessler_score ~ total_asset_value, data = wave1)

Residuals:
    Min      1Q  Median      3Q     Max
-10.351  -5.350  -0.351   3.650  29.649

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.035e+01  7.360e-02 276.509   <2e-16 ***
total_asset_value -2.139e-07  6.144e-07  -0.348    0.728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.309 on 7410 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  1.636e-05,	Adjusted R-squared:  -0.0001186
F-statistic: 0.1212 on 1 and 7410 DF,  p-value: 0.7278
```

```r
 #---------------------------------------------------------------------#

 # now restricting by removing outliers in total asset value (essentially getting rid of
 # Define percentile thresholds
lower_threshold <- 0  # Set lower threshold to 0
upper_threshold <- quantile(wave1$total_asset_value, 0.90, na.rm = TRUE)  # 90th percent

# Restrict data: Remove values above the 90th percentile and ensure values are non-negat
wave1_restricted <- wave1 %>%
  filter(total_asset_value >= lower_threshold & total_asset_value <= upper_threshold)

# Check summary after restriction
summary(wave1_restricted$total_asset_value)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.5  1300.0  2287.0  2996.2  3917.0 12069.0
```

```
# Scatter plot: Kessler Score vs. Household Wealth (restricted data)
ggplot(wave1_restricted, aes(x = total_asset_value, y = kessler_score)) +
  geom_point(alpha = 0.5) +  # Transparency to avoid overplotting
  geom_smooth(method = "lm", color = "darkblue", se = TRUE) +  # Linear trend line
  labs(
    title = "Relationship Between Household Wealth and Depression (Restricted Data)",
    x = "Total Asset Value (Household Wealth Proxy)",
    y = "Kessler Score (Depression)"
  ) +
  theme_minimal()
```

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



```r
# Correlation test (restricted data)
cor.test(wave1_restricted$total_asset_value, wave1_restricted$kessler_score, use = "comp
```

```
    Pearson's product-moment correlation

data:  wave1_restricted$total_asset_value and wave1_restricted$kessler_score
t = -3.0653, df = 6667, p-value = 0.002183
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06146011 -0.01352533
sample estimates:
       cor
-0.0375143
```

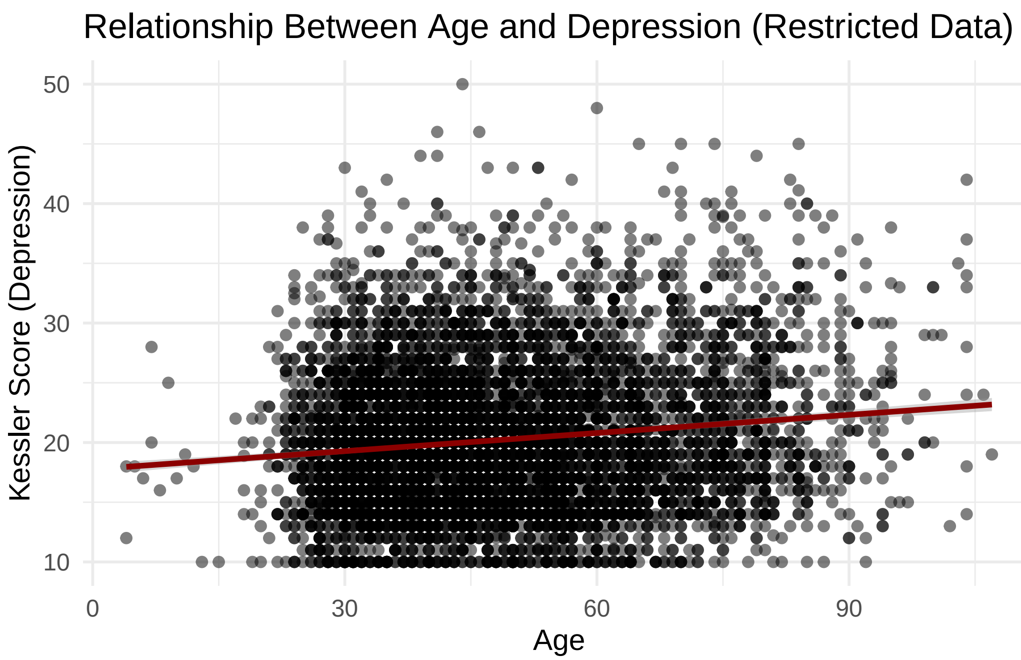
```r
# Linear regression (restricted data)
model.wealth2 <- lm(kessler_score ~ total_asset_value, data = wave1_restricted)
summary(model.wealth2)
```

```
Call:
lm(formula = kessler_score ~ total_asset_value, data = wave1_restricted)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5326  -5.1812  -0.8122   3.8426  29.4905 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2.056e+01  1.230e-01 167.101  < 2e-16 ***
total_asset_value -9.800e-05  3.197e-05  -3.065  0.00218 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.305 on 6667 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.001407,  Adjusted R-squared:  0.001258 
F-statistic: 9.396 on 1 and 6667 DF,  p-value: 0.002183
```

2. *A household or demographic characteristic that seems interesting to you. Present the results from your exploration through tables, plots, a write-up, or anything else you think would be useful.*

```r
# Age

# ensure only positive ages included
wave1_restricted <- wave1_restricted %>%
  filter(age > 0)

 # scatter plot: Kessler Score vs. Age
ggplot(wave1_restricted, aes(x = age, y = kessler_score)) +
  geom_point(alpha = 0.5) +  # Transparency to avoid overplotting
  geom_smooth(method = "lm", color = "darkred", se = TRUE) +  # Linear trend line
  labs(
    title = "Relationship Between Age and Depression (Restricted Data)",
    x = "Age",
    y = "Kessler Score (Depression)"
  ) +
  theme_minimal()
```

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



```r
# correlation test: Age vs. Depression
cor.test(wave1_restricted$age, wave1_restricted$kessler_score, use = "complete.obs")
```

```

    Pearson's product-moment correlation

data:  wave1_restricted$age and wave1_restricted$kessler_score
t = 10.849, df = 6641, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1082623 0.1555214
sample estimates:
      cor
0.1319668
```

```r
# apply linear regression: Predicting depression based on age
model.age <- lm(kessler_score ~ age, data = wave1_restricted)
summary(model.age)
```

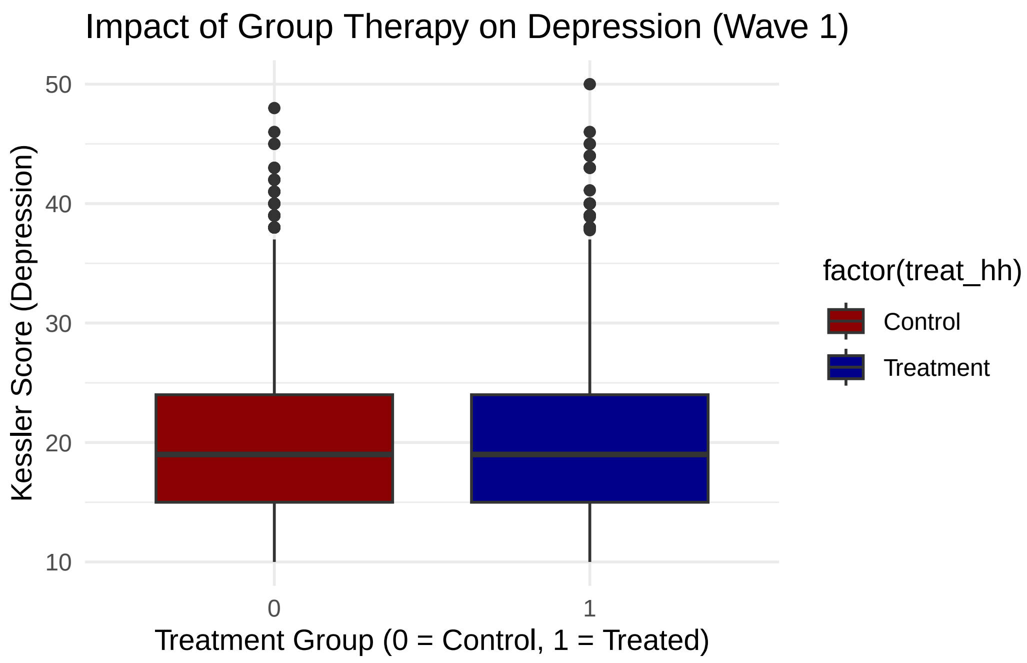
```

Call:
lm(formula = kessler_score ~ age, data = wave1_restricted)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.4193  -4.9883  -0.7857   3.9923  30.0117 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.759895   0.242060   73.37   <2e-16 ***
age          0.050646   0.004668   10.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.245 on 6641 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.01742,   Adjusted R-squared:  0.01727 
F-statistic: 117.7 on 1 and 6641 DF,  p-value: < 2.2e-16
```

3. *Evaluating the RCT*

```r
# summary statistics for Kessler Score by treatment group
wave1 %>%
  group_by(treat_hh) %>%
  summarise(
    mean_kessler = mean(kessler_score, na.rm = TRUE),
    median_kessler = median(kessler_score, na.rm = TRUE),
    sd_kessler = sd(kessler_score, na.rm = TRUE),
    count = n()
  )
```

```
# A tibble: 2 x 5
  treat_hh                 mean_kessler median_kessler sd_kessler count
  <dbl+lbl>                       <dbl>          <dbl>      <dbl> <int>
1 0 [Control household]            20.3             19       6.27  3713
2 1 [Treatment household]          20.4             20       6.35  3727
```

```r
# boxplot: Depression Score by Treatment Group
ggplot(wave1_restricted, aes(x = factor(treat_hh), y = kessler_score, fill = factor(trea
  geom_boxplot() +
  labs(
    title = "Impact of Group Therapy on Depression (Wave 1)",
    x = "Treatment Group (0 = Control, 1 = Treated)",
    y = "Kessler Score (Depression)"
  ) +
  scale_fill_manual(values = c("darkred", "darkblue"), labels = c("Control", "Treatment'
  theme_minimal()
```



```
 # conduct T-test: Difference in Depression Scores Between Control and Treatment
t.test(kessler_score ~ treat_hh, data = wave1_restricted)
```

```
    Welch Two Sample t-test

data:  kessler_score by treat_hh
t = -0.51425, df = 6641, p-value = 0.6071
alternative hypothesis: true difference in means between group 0 and group 1 is not equa
95 percent confidence interval:
 -0.3825464  0.2235507
sample estimates:
mean in group 0 mean in group 1
       20.21098        20.29047
```

```
# apply ;inear regression: Effect of treatment on depression
model.treat1 <- lm(kessler_score ~ treat_hh, data = wave1_restricted)
summary(model.treat1)
```

```
Call:
lm(formula = kessler_score ~ treat_hh, data = wave1_restricted)

Residuals:
    Min      1Q  Median      3Q     Max
-10.290  -5.211  -1.211   3.789  29.709

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.2110     0.1097 184.278   <2e-16 ***
treat_hh      0.0795     0.1546   0.514    0.607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.3 on 6641 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  3.981e-05, Adjusted R-squared:  -0.0001108
F-statistic: 0.2644 on 1 and 6641 DF,  p-value: 0.6071
```

Using Wave 2 data to measure outcomes, answer the following questions, explaining any decisions and assumptions you make, and interpret your results. There is no need for you to address the validity of the random assignment of the intervention.

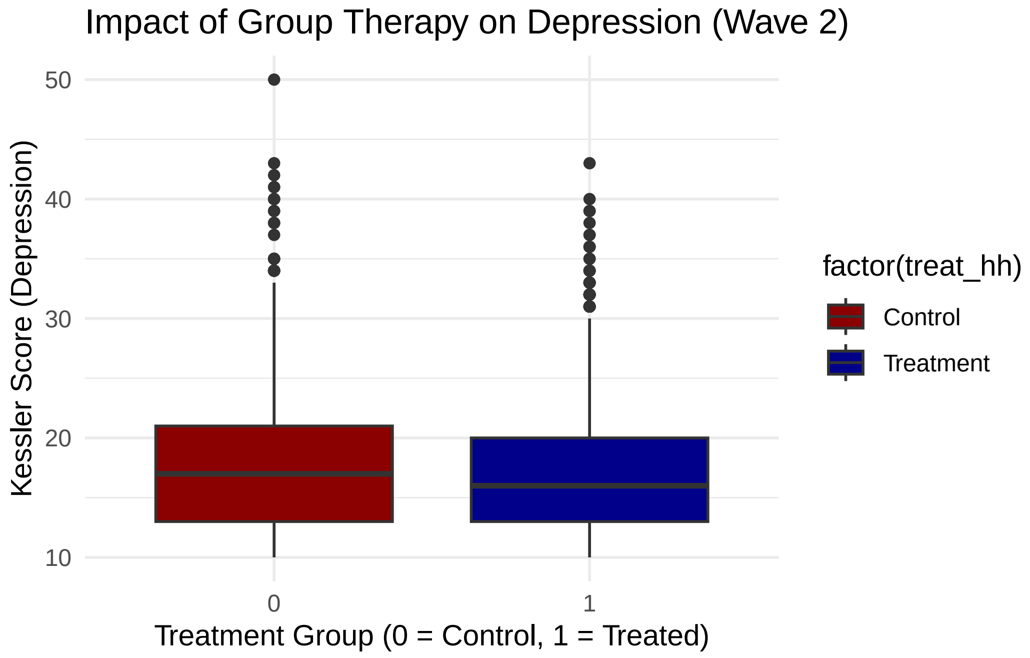
3.2 Q2:

Were the GT sessions effective at reducing depression?

```r
 # first, filter for Wave 2 data
wave2<- mergedv2 %>%
  filter(wave == 2)

 # check summary statistics of Kessler Score by treatment group in Wave 2
wave2 %>%
  group_by(treat_hh) %>%
  summarise(
    mean_kessler = mean(kessler_score, na.rm = TRUE),
    median_kessler = median(kessler_score, na.rm = TRUE),
    sd_kessler = sd(kessler_score, na.rm = TRUE),
    count = n()
  )
```

```
# A tibble: 2 x 5
  treat_hh                mean_kessler median_kessler sd_kessler count
  <dbl+lbl>                      <dbl>          <dbl>      <dbl> <int>
1 0 [Control household]           17.5             17       5.59  3190
2 1 [Treatment household]         17.0             16       5.35  3192
```

```r
# Boxplot: Depression Score by Treatment Group (Wave 2)
ggplot(wave2, aes(x = factor(treat_hh), y = kessler_score, fill = factor(treat_hh))) +
  geom_boxplot() +
  labs(
    title = "Impact of Group Therapy on Depression (Wave 2)",
    x = "Treatment Group (0 = Control, 1 = Treated)",
    y = "Kessler Score (Depression)"
  ) +
  scale_fill_manual(values = c("darkred", "darkblue"), labels = c("Control", "Treatment")) +
  theme_minimal()
```



```r
 # next conduct T-test: Difference in Depression Scores Between Control and Treatment (Wave
t.test(kessler_score ~ treat_hh, data = wave2)
```

```
    Welch Two Sample t-test

data:  kessler_score by treat_hh
t = 3.8796, df = 6367.5, p-value = 0.0001057
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 0.2626988 0.7993474
```

```
sample estimates:
mean in group 0 mean in group 1 
       17.51285        16.98183 
```

```
 # finally aplly linear regression: Effect of treatment on depression (Wave 2)
model.GT <- lm(kessler_score ~ treat_hh, data = wave2)
summary(model.GT)
```

```

Call:
lm(formula = kessler_score ~ treat_hh, data = wave2)

Residuals:
   Min     1Q Median     3Q    Max 
-7.513 -3.982 -0.982  3.487 32.487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.5129     0.0968  180.92  < 2e-16 ***
treat_hh     -0.5310     0.1369   -3.88 0.000106 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.467 on 6380 degrees of freedom
Multiple R-squared:  0.002354,	Adjusted R-squared:  0.002197 
F-statistic: 15.05 on 1 and 6380 DF,  p-value: 0.0001057
```

3.3 Q3:

Did the effect of GT sessions on depression differ by gender? Perform a linear regression of the Kessler Score on:

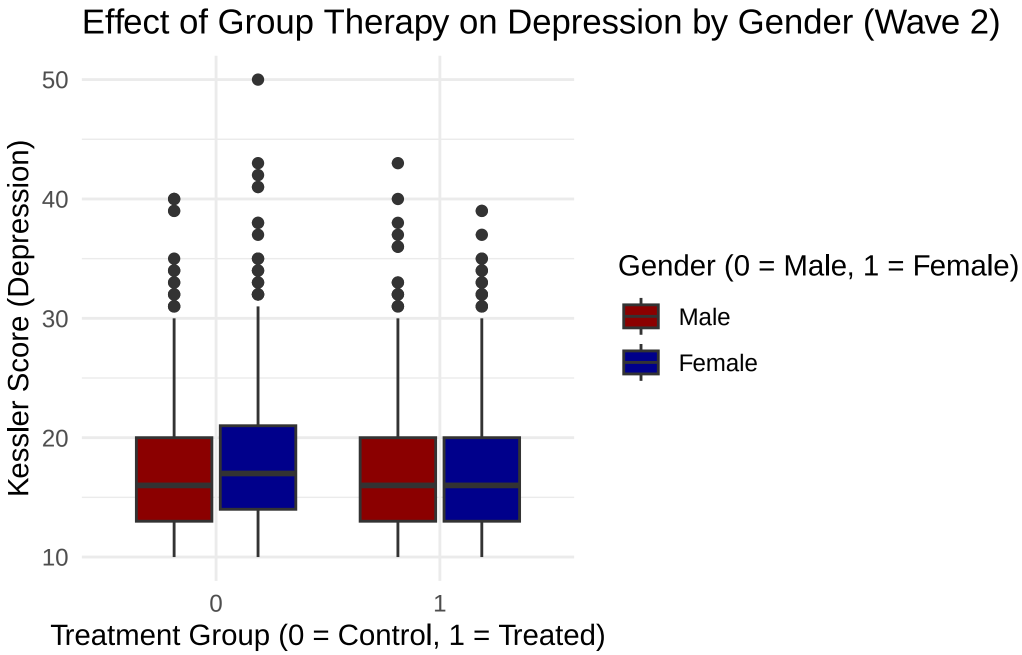
- *Woman (binary variable)*
- *Treated Household (binary variable)*
- *Interaction term: Treated Household * Woman, using only Wave 2 observations.*

Note: Note: In your write-up for this question, please make sure to explain and interpret all coefficients in your specification, keeping in mind units and reference groups.

```r
 # first create binary variable for Woman (assuming 1 = male, 5 = female)
wave2_gender<- wave2 %>%
  mutate(woman = ifelse(gender == 5, 1, 0))

 # next apply linear regression: testing interaction between treatment and gender
model.genderINT <- lm(kessler_score ~ treat_hh * woman, data = wave2_gender)
summary(model.genderINT)
```

```

Call:
lm(formula = kessler_score ~ treat_hh * woman, data = wave2_gender)

Residuals:
   Min     1Q Median     3Q    Max 
-7.859 -4.021 -0.932  3.141 32.141 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     17.0634     0.1466 116.411  < 2e-16 ***
treat_hh        -0.1319     0.2068  -0.638   0.5237    
woman            0.7956     0.1950   4.080 4.56e-05 ***
treat_hh:woman  -0.7059     0.2756  -2.561   0.0105 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.461 on 6378 degrees of freedom
Multiple R-squared:  0.004983,  Adjusted R-squared:  0.004515 
F-statistic: 10.65 on 3 and 6378 DF,  p-value: 5.586e-07
```

```r
 # visualization: Effect of treatment by gender
ggplot(wave2_gender, aes(x = factor(treat_hh), y = kessler_score, fill = factor(woman))) +
  geom_boxplot() +
  labs(
    title = "Effect of Group Therapy on Depression by Gender (Wave 2)",
    x = "Treatment Group (0 = Control, 1 = Treated)",
    y = "Kessler Score (Depression)",
    fill = "Gender (0 = Male, 1 = Female)"
  ) +
  scale_fill_manual(values = c("darkred", "darkblue"), labels = c("Male", "Female")) +
  theme_minimal()
```



4 Summary of Results

Present your results as you see fit and add a write-up with your interpretation and conclusions.

4.1 Results Q1 - Asset Value

Key Findings (Full Data, Before Restriction)

- **Correlation Analysis:**
  - The correlation between total asset value (household wealth proxy) and depression (Kessler-10 score)was very weak and statistically insignificant (cor=−0.004, p=0.728).
  - This suggests that wealth does not meaningfully predict depression in the full sample.
- **Linear Regression Results:**
  - The total asset value coefficient was not significant (effect =−2.14e-7, p=.728).

  - The r-squared value was nearly zero, meaning that household wealth explains almost none of the variation in depression.

Interpretation:

- There is no meaningful relationship between household wealth and depression when analyzing the full data set.

I thought an interesting adjustment may be to check the relationship by restricting the data, removing observations above 90th percentile in total asset value.

Key Findings (Restricted Data: Removing Top 10% Wealthiest Households)

- **Summary Statistics After Restriction:**
  - The maximum **total asset value** was **12,069** currency units, compared to **5.79 million** in the full dataset, removes extreme wealth values that may introduce bias.
- **Correlation Analysis:**
  - Interestingly, the correlation became negative and statistically significant (cor=−0.038, p=0.002).
  - This suggests that higher household wealth is now slightly associated with lower depression scores.
- **Linear Regression Results:**
  - The total asset value coefficient was negative and significant (effect=−9.8e-5, p=0.002).
  - Effect size is very small, but it implies that an increase in wealth corresponds to a slight decrease in depression.

Interpretation:

- After removing the wealthiest 10% of households, a small but statistically significant negative relationship emerges between wealth and depression. This suggests that for most households, increased wealth is weakly associated with better mental health. However, the effect remains minimal, meaning other social and psychological factors likely play a larger role in depression in Ghana than economic status.

4.2 Results Q1 - Age

Findings:

- **Correlation Analysis:** There was a weak but statistically significant positive correlation between age and depression (cor= 0.128 p<0.001). This suggests that older individuals tend to report higher depression scores.
- **Linear Regression:** Age was a significant predictor of depression (effect = 0.049, p<0.001), meaning that each additional year of age is associated with a 0.05 point increase in depression score.
- **Scatter Plot Analysis:** The **positive trend** in the scatterplot supports this conclusion.

Interpretation:

- Older individuals in Ghana tend to report higher depression levels, possibly due to reduced social engagement, health deterioration, or financial insecurity in later life.
- The effect, though small, is statistically significant, suggesting that age should be considered in mental health interventions.

4.3 Results Q1 - RCT

Findings (Pre-Treatment Comparison in Wave 1)

- **Descriptive Statistics**
  - The mean Kessler-10 score in control households: 20.31
  - The mean Kessler-10 score in treatment households: 20.36
  - The difference in means is only 0.05 points, which is very small.
- **T-Test Results:**
  - The p-value = 0.7433, meaning there is **no** statistically significant difference in depression scores before treatment.
- **Linear Regression:**
  - The **t**reatment group indicator is **not** statistically significant p =0.743), confirming that depression levels were similar in both groups before therapy.

Interpretation:

Randomization worked: Before treatment, there was no meaningful difference in depression scores between control and treatment households.
This ensures that any differences observed in Wave 2 are due to Group Therapy and not pre-existing differences between groups.

4.4 Results Q2 - Group Therapy Effect

- **Descriptive Statistics:**
  - The mean Kessler-10 score in control households (no therapy) was 17.51.
  - The mean Kessler-10 score in treatment households (received therapy) was 16.98.
- **T-Test Results:** The difference in depression scores between the control and treatment groups was **statistically significant** $p<0.001$), indicating lower depression in treatment households.
- **Linear Regression:**
  - Belonging to a **treatment household** was associated with a **0.53-point decrease** in depression scores with an effect size of 0.531, $p<0.001$).
  - The effect size is small but meaningful.

4.5 Results Q3 - Gender Effect

To explore whether therapy affected men and women differently, we introduced an interaction term in the regression model.

Findings:

- **Main effects:**
  - Being female **was** associated with higher depression scores $=0.796$, $p < 0.001$), meaning women experienced more depressive symptoms than men on average.
  - Belonging to a treatment household alone was not statistically significant ($p=0.524$), suggesting that GT alone **did not** impact depression when gender isn't accounted for.
- **Interaction Effect:**
  - The interaction term effect $=-0.706$, $p=0.011$) **was** statistically significant, meaning the effect of GT on depression was stronger for women.

- **Visualization:** The box plot showed a clearer reduction in depression among treated women compared to treated men.

Interpretation:

- Women in Ghana experience more depression than men. Group Therapy was more effective for women, reducing their depression scores significantly. This suggests gender-sensitive mental health interventions may be needed, as women benefit more from these socially structured interventions according to our data.

5 Bonus Visualizations & Analysis

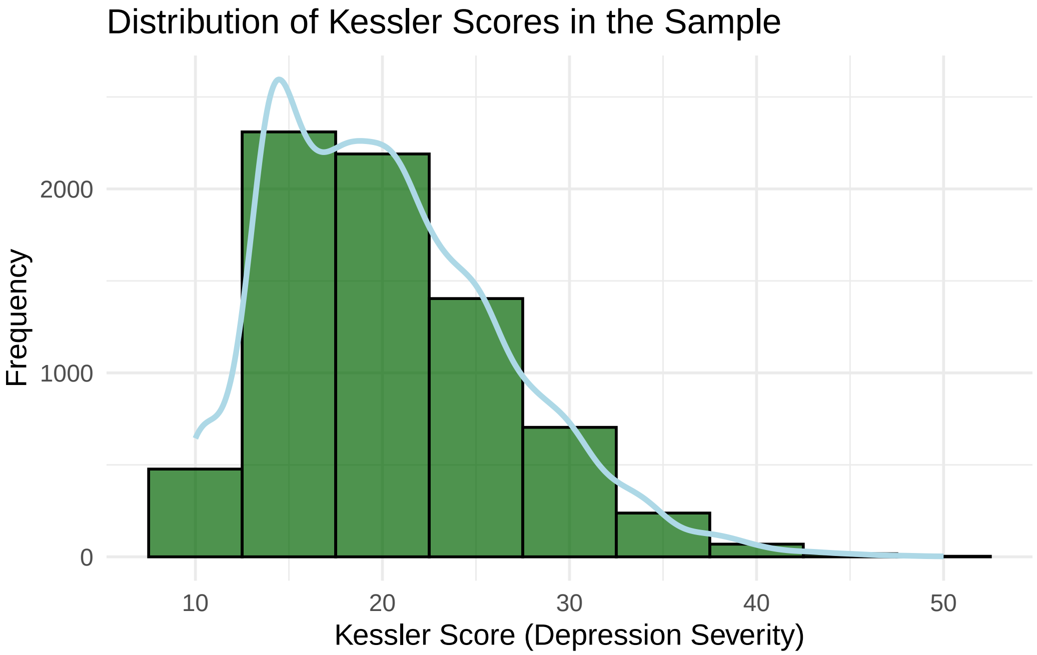
In this bonus section, I do some digging into Kessler scores and demographic variables to add some more types of visualizations for the final render!

```r
 # create histogram of Kessler Scores for each wave

 # wave 1
hist.wave1<- ggplot(data = wave1, aes(x = kessler_score)) +
  geom_histogram(binwidth = 5, fill = "darkgreen", color = "black", alpha = 0.7) +
  geom_density(aes(y = ..count.. * 5), color = "lightblue", lwd = 1) +  # Density overlay
  labs(title = "Distribution of Kessler Scores in the Sample",
       x = "Kessler Score (Depression Severity)",
       y = "Frequency") +
  theme_minimal()

 # wave 2
hist.wave2<- ggplot(data = wave2, aes(x = kessler_score)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black", alpha = 0.7) +
  geom_density(aes(y = ..count.. * 5), color = "darkblue", lwd = 1) +  # Density overlay
  labs(title = "Distribution of Kessler Scores in the Sample",
       x = "Kessler Score (Depression Severity)",
       y = "Frequency") +
  theme_minimal()

# displaying
hist.wave1
```

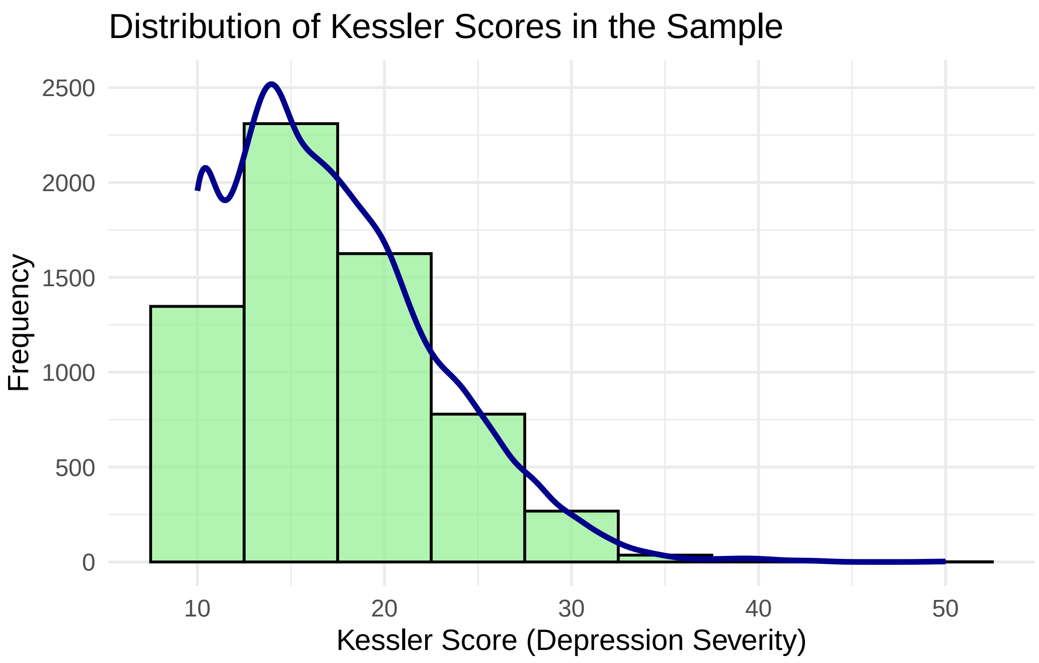
```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```

```
Warning: Removed 28 rows containing non-finite outside the scale range
(`stat_bin()`).
```

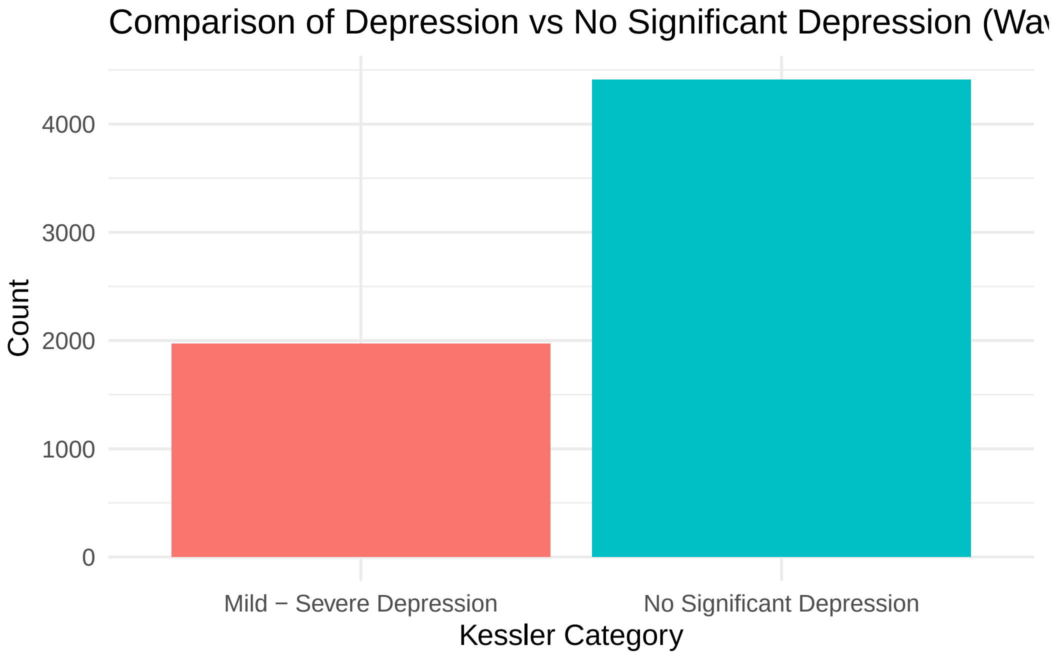
```
Warning: Removed 28 rows containing non-finite outside the scale range
(`stat_density()`).
```



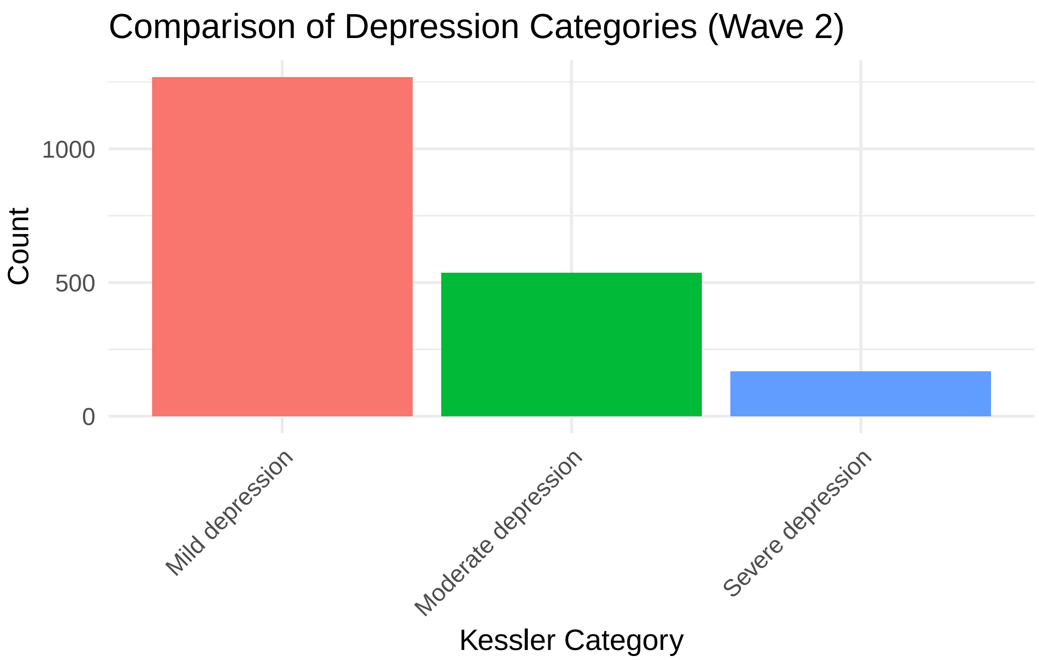
```
hist.wave2
```



```
 # Let's create a visual for Kessler categories, bar chart seems most fitting

 # first, re-code Kessler categories into two groups: "Depression" and "No Significant Depre
wave2 <- wave2 %>%
  mutate(kessler_group = ifelse(kessler_categories == "No significant depression",
                                "No Significant Depression", "Mild - Severe Depression"))

 # bar chart comparing aggregated depression vs no significant depression
wave2 %>%
  count(kessler_group) %>%
  ggplot(aes(x = kessler_group, y = n, fill = kessler_group)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(title = "Comparison of Depression vs No Significant Depression (Wave 2)",
       x = "Kessler Category",
       y = "Count") +
  theme_minimal()
```



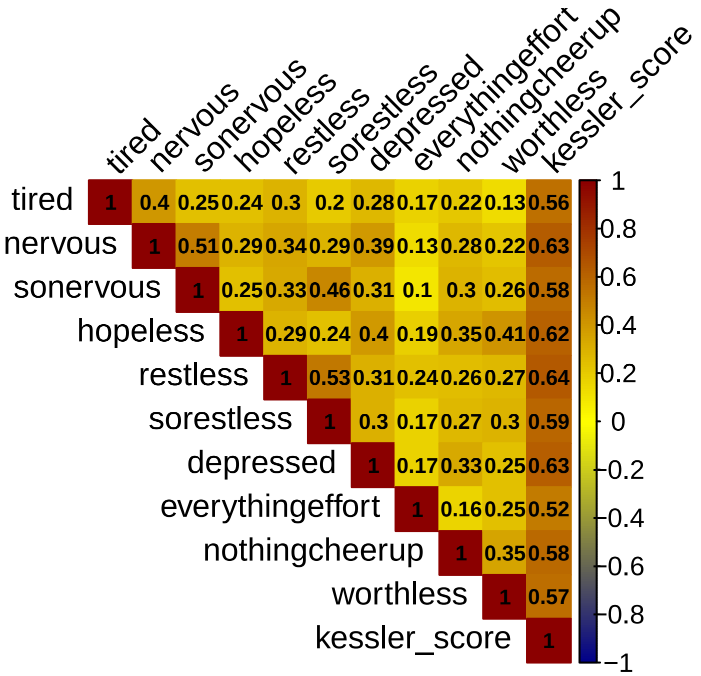
```r
# bar chart comparing the three depression categories only
wave2 %>%
  filter(kessler_categories != "No significant depression") %>%
  count(kessler_categories) %>%
  ggplot(aes(x = reorder(kessler_categories, -n), y = n, fill = kessler_categories)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(title = "Comparison of Depression Categories (Wave 2)",
       x = "Kessler Category",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
 # Let's create a corrplot to analyze how the kessler measures are associated with each othe

library(corrplot)
```

```
corrplot 0.92 loaded
```

```
 # select only the Kessler-related variables
corr_vars <- wave2 %>%
  select(tired, nervous, sonervous, hopeless, restless, sorestless, depressed, everythingeffo

 # compute correlation matrix
corr_matrix <- cor(corr_vars, use = "pairwise.complete.obs")

 # plot the correlation matrix
corrplot(corr_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7, col = colorRampPalette(c("darkblue", "yello
```



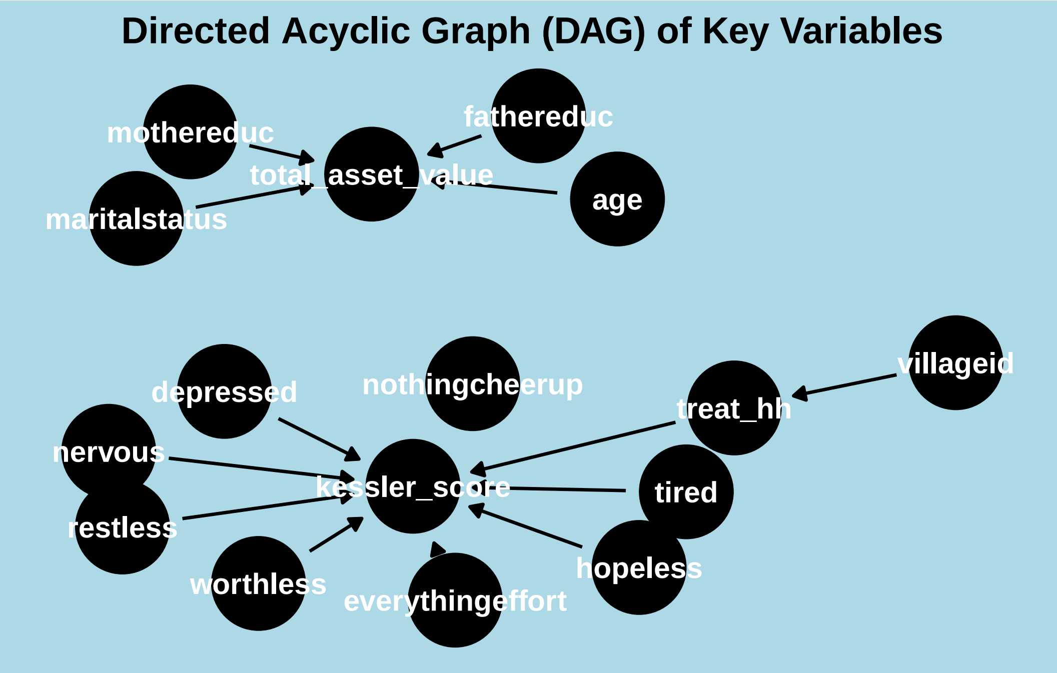
```r
library(ggdag)
```

```
Warning: package 'ggdag' was built under R version 4.3.3


Attaching package: 'ggdag'

The following objects are masked from 'package:summarytools':

    label, label<-

The following object is masked from 'package:stats':

    filter
```

```r
library(DiagrammeR)
```

```
Warning: package 'DiagrammeR' was built under R version 4.3.1
```

```
# Note : This is more hypothetical and theoretical, not necesarilly reflective of the result

# If we wanted to develop causal theories about how these variables are interacting it may b

 # creating a hypothetical dag with some of our variables
dag <- dagify(
  kessler_score ~ treat_hh + tired + nervous + hopeless + depressed + restless + everythinge
  total_asset_value ~ age + maritalstatus + fathereduc + mothereduc,
  treat_hh ~ villageid,
  exposure = "treat_hh",
  outcome = "kessler_score"
)

 # plot
ggdag(dag, text = TRUE) +
  theme_dag_blank() +
  theme(
    plot.background = element_rect(fill = "lightblue", color = NA),  # Light grey background
    panel.background = element_rect(fill = "lightblue", color = NA),
    legend.position = "none",
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5)
  ) +
  ggtitle("Directed Acyclic Graph (DAG) of Key Variables")
```



GPRL team, thank you so much for this opportunity!

Colin