Code Challenge

Liujing Xie

2024-12-06

Loading Packages and Data

```
# Packages
library(tidyverse)
library(dplyr)
library(ggplot2)
# RData
load("CodeChallenge2024.RData")
# df, participants of interest
ids_of_interest = read.delim("IDs.txt", col.names = "ids", colClasses = "character")
# as.factor
id_map$new_id = as.factor(id_map$new_id)
HAM_sleep$ID = as.factor(HAM_sleep$ID)
```

Question 1 Data Cleaning

Map ID

```
ids_of_interest = ids_of_interest %>%
  left_join(id_map, by = join_by(ids == old_id)) # ids of interest and the corresponding new ids

final_df = HAM_protect %>%
  left_join(id_map, join_by(ID == old_id)) %>% # match old ids to new ids
  select(-ID) %>% # remove old ids column
  select(new_id, timepoint:ham_17_weight) %>%
  rename("ID" = "new_id") %>% # "new_id" renamed to "ID"
  rbind(HAM_sleep) %>%
  inner_join(ids_of_interest, by = join_by(ID == new_id)) %>% # keep only participants of interest
  select(-ids)
```

Calculate HAM scores

```
for (i in c(5:26)) {
  final_df[, i] = as.numeric(final_df[, i]) # make each HAM item variable numeric
}
```

```
## Warning: NAs introduced by coercion
```

```r
HAM_scores = final_df %>% # adding up items except 3a to 3e
  mutate(HAM_Score = rowSums(across(ham_1_dm:ham_17_weight), na.rm = T))
HAM_scores %>% select(ID, HAM_Score) %>% head() # first 6 rows of HAM scores
```

```
##     ID HAM_Score
## 1 2027        26
## 2 2027        27
## 3 2027        24
## 4 2027        14
## 5 2027        27
## 6 2027        13
```

Calculate mean HAM of each participant

```r
HAM_scores %>%
  group_by(ID) %>%
  summarise(mean = round(mean(HAM_Score), 2)) # mean of each participant
```

```
## # A tibble: 89 x 2
##    ID     mean
##    <fct> <dbl>
##  1 2027  17.1
##  2 2056   3.88
##  3 2068  22.7
##  4 2069  21.2
##  5 2072   7.15
##  6 2074  24.0
##  7 2086  10.8
##  8 2090  14.1
##  9 2104   9.13
## 10 2110   9.71
## # i 79 more rows
```

Latest HAM

```r
HAM_scores %>%
  group_by(ID) %>%
  # the date of each participant's latest HAM
  slice_max(fug_date, n = 1, with_ties = FALSE) %>%
  select(ID, fug_date, HAM_Score)
```

```
## # A tibble: 89 x 3
## # Groups:   ID [89]
##    ID    fug_date   HAM_Score
##    <fct> <date>         <dbl>
##  1 2027  2011-09-09         9
##  2 2056  2019-06-10         4
##  3 2068  2024-11-03        61
```

```
##  4 2069  2024-09-19        56
##  5 2072  2023-04-16         6
##  6 2074  2024-08-29        70
##  7 2086  2011-05-02         9
##  8 2090  2024-10-22        58
##  9 2104  2020-01-14         0
## 10 2110  2021-08-01        59
## # i 79 more rows
```

HAM of 1-year timepoint

```
HAM_scores %>% # filter 1-year timepoints of each participant
  filter(grepl("1_year_", timepoint) | grepl("year_1_", timepoint)) %>%
  group_by(ID) %>%
  slice_min(fug_date, with_ties = FALSE) # the dates 1 year after the initial consent
```

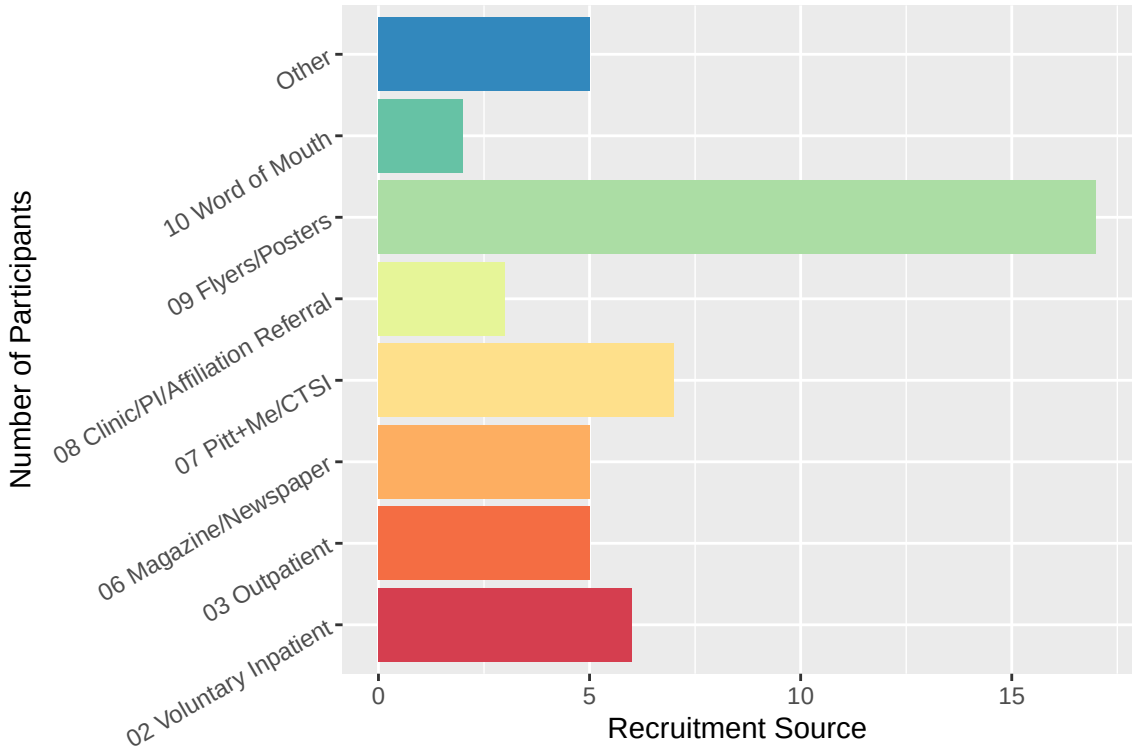
```
## # A tibble: 89 x 27
## # Groups:   ID [89]
##    ID    timepoint    bq_date fug_date   ham_1_dm ham_2_gf ham_3_su ham_3a_wl
##    <fct> <chr>        <date>  <date>        <dbl>    <dbl>    <dbl>     <dbl>
##  1 2027  1_year_arm_1 NA      2004-04-12        3        3        2        NA
##  2 2056  1_year_arm_1 NA      2006-05-01        0        0        0        NA
##  3 2068  year_1_arm_2 NA      2019-01-20        3        3        0         0
##  4 2069  year_1_arm_2 NA      2019-10-24        2        0        0         0
##  5 2072  1_year_arm_1 NA      2015-04-14       NA       NA       NA        NA
##  6 2074  year_1_arm_2 NA      2018-09-09        2        3        2         1
##  7 2086  1_year_arm_1 NA      2007-02-19        0        0        0        NA
##  8 2090  year_1_arm_2 NA      2017-11-13        2        1        0         0
##  9 2104  1_year_arm_1 NA      2007-11-22        2        0        2        NA
## 10 2110  1_year_arm_1 NA      2008-03-10        2        0        2        NA
## # i 79 more rows
## # i 19 more variables: ham_3b_wd <dbl>, ham_3c_rld <dbl>, ham_3d_asa <dbl>,
## #   ham_3e_pdw <dbl>, ham_4_ii <dbl>, ham_5_im <dbl>, ham_6_di <dbl>,
## #   ham_7_wi <dbl>, ham_8_re <dbl>, ham_9_ag <dbl>, ham_10_psya <dbl>,
## #   ham_11_soma <dbl>, ham_12_gi <dbl>, ham_13_gs <dbl>, ham_14_sex <dbl>,
## #   ham_15_hd <dbl>, ham_16_li <dbl>, ham_17_weight <dbl>, HAM_Score <dbl>
```

Question 2

Number of participants by recruitment sources

```
recruitment_data %>%
  ggplot(aes(RecruitSource, fill = RecruitSource)) +
  geom_bar() +
  coord_flip() + # flip coordinate for aesthetics
  scale_fill_brewer(palette = "Spectral") +
  labs(x = "Number of Participants",
       y = "Recruitment Source") +
  theme(axis.text.y = element_text(angle = 30, hjust = 1),
```

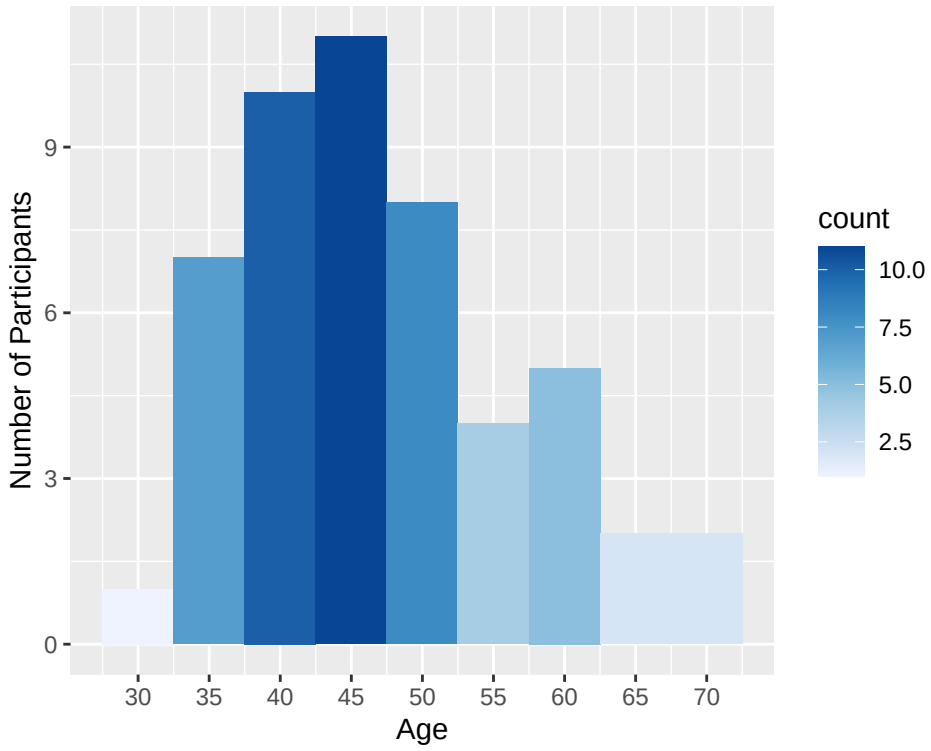
```
        # rotate the label due to limited space
        legend.position = "none")
```



Number of participants by age

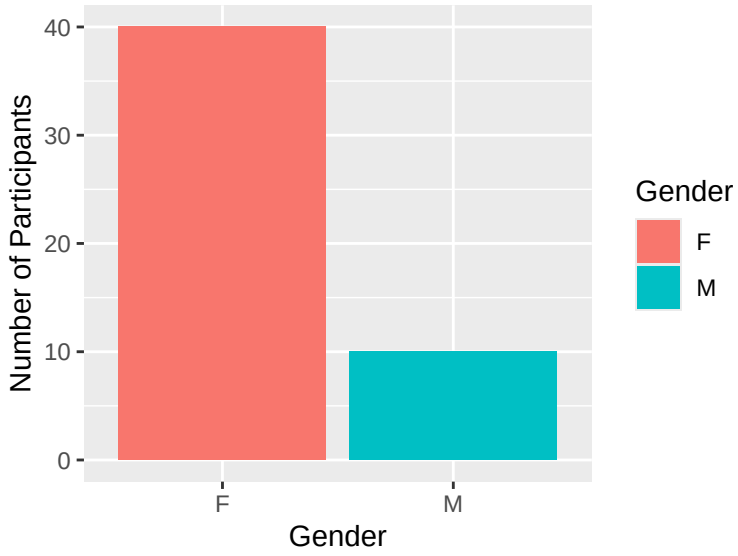
```
recruitment_data %>%
  ggplot(aes(Age, fill = ..count..)) +
  geom_histogram(binwidth = 5) +
  scale_x_continuous(breaks = seq(0, 100, by = 5)) +
  scale_fill_distiller(palette = "Blues", direction = 1) +
  labs(x = "Age",
       y = "Number of Participants")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Number of participants by gender

```
recruitment_data %>%
  ggplot(aes(Gender, fill = Gender)) +
  geom_bar() +
  labs(x = "Gender",
       y = "Number of Participants")
```



Number of participants by group

```
recruitment_data %>%
  ggplot(aes(Group, fill = Group)) +
  geom_bar() +
  scale_fill_brewer(palette = "Spectral") +
  labs(x = "Group",
       y = "Number of Participants") +
  theme(legend.position = "none")
```