Homework 2

Lupe Antonio

10/2/2023

Problem 1

```
#loading pols-month data
pols_month <- read_csv("data_hw2/fivethirtyeight_datasets/pols-month.csv")
```

```
#cleaning the data
pols_month <- janitor::clean_names(pols_month) %>%
  #breaking up variable mon into integer variables: year, month, & day
  separate(mon, into = c("year", "month", "day"), convert = TRUE) %>%
  mutate(
    #replacing month num with month name
    month = case_match(
      month,
      1 ~ "january",
      2 ~ "february",
      3 ~ "march",
      4 ~ "april",
      5 ~ "may",
      6 ~ "june",
      7 ~ "july",
      8 ~ "august",
      9 ~ "september",
      10 ~ "october",
      11 ~ "november",
      12 ~ "december"),
    #creating president variable
    president = case_when(
      prez_dem == 1 & prez_gop == 0 ~ "democratic",
      prez_dem == 0 & prez_gop == 1 ~ "republican")) %>%
  #removing prez_dem, prez_gop, and day variables
  select(-prez_dem, -prez_gop, -day)
```

```
#loading snp.csv data
snp <- read_csv("data_hw2/fivethirtyeight_datasets/snp.csv") %>%
  #cleaning data
  janitor::clean_names() %>%
  #fixing the dates
  mutate(date = lubridate::mdy(date)) %>%
  #breaking up date variable into integer variables: year, month, day
  separate(date, into = c("year", "month", "day"), convert = TRUE) %>%
```

```
  mutate(
    #fixing the year problem
    year = ifelse(year > 2023, year - 100, year),
    #replacing month num with month name
    month = case_match(
      month,
      1 ~ "january",
      2 ~ "february",
      3 ~ "march",
      4 ~ "april",
      5 ~ "may",
      6 ~ "june",
      7 ~ "july",
      8 ~ "august",
      9 ~ "september",
      10 ~ "october",
      11 ~ "november",
      12 ~ "december")) %>%
  #removing day variable
  select(-day)
```

```
#loading unemployment.csv data
unemployment <- read_csv("data_hw2/fivethirtyeight_datasets/unemployment.csv") %>%
  #cleaning the data
  janitor::clean_names() %>%
  #going from "wide" to "long" format
  pivot_longer(jan:dec,
               names_to = "month",
               values_to = "unemp_perctg") %>%
  #fixing month names
  mutate(
    month = case_match(
      month,
      "jan" ~ "january",
      "feb" ~ "february",
      "mar" ~ "march",
      "apr" ~ "april",
      "may" ~ "may",
      "jun" ~ "june",
      "jul" ~ "july",
      "aug" ~ "august",
      "sep" ~ "september",
      "oct" ~ "october",
      "nov" ~ "november",
      "dec" ~ "december"
    ))
```

```
#merging snp into pols
final_data <- left_join(pols_month, snp)
```

```
## Joining with `by = join_by(year, month)`
```

```
final_data <- left_join(final_data, unemployment)
```

```
## Joining with `by = join_by(year, month)`
```

The `pols_month` data has a total of 822 observations and 9 variables. This dataset contains information regarding the party affiliations of governors and senators, and the current president (at that time) between 1947 to 2015. The `snp` data has a total of 787 observations and 3 variables. It contains information from the years 1950 to 2015. Additionally, the `unemployment` dataset has a total of 816 observations and 3 variables. It contains information from the years 1948 to 2015.

Problem 2

```
#loading Mr. Trash Wheel data
mr_trash_wheel <- read_excel("data_hw2/Trash_Wheel_Collection_Data.xlsx", sheet = "Mr. Trash Wheel") %>%
  #cleaning data
  janitor::clean_names() %>%
  #removing x15 & x16 columns
  select(-x15, -x16) %>%
  #creating new homes_powered variable
  mutate(
    homes_powered = (weight_tons * 500)/30) %>%
  #adding trash wheel variable
  mutate(trash_wheel = "Mr. Trash Wheel") %>%
  #filter NA's in dumpster
  filter(!is.na(dumpster)) %>%
  #making year numeric
  mutate(year = as.numeric(year))
```

```
#loading Professor Trash Wheel data
professor_trash_wheel <- read_excel("data_hw2/Trash_Wheel_Collection_Data.xlsx", sheet = "Professor Tras
  #cleaning data
  janitor::clean_names() %>%
  #adding trash wheel variable
  mutate(trash_wheel = "Professor Trash Wheel") %>%
  #filter NA's in dumpster
  filter(!is.na(dumpster)) %>%
  #making year numeric
  mutate(year = as.numeric(year))
```

```
#loading Gwynnda data
gwynnda_trash_wheel <- read_excel("data_hw2/Trash_Wheel_Collection_Data.xlsx", sheet = "Gwynnda Trash Wh
  #cleaning data
  janitor::clean_names() %>%
  #adding trash wheel variable
  mutate(trash_wheel = "Gwynnda Trash Wheel") %>%
  #filter NA's in dumpster
  filter(!is.na(dumpster)) %>%
  #making year numeric
  mutate(year = as.numeric(year))
```

```
#combining datasets
trash_wheels_tidy <- bind_rows(mr_trash_wheel, professor_trash_wheel, gwynnda_trash_wheel)
```

The final tidy dataset contains information from `mr_trash_wheel`, `professor_trash_wheel`, and `gwynnda_trash_wheel`. It has a total of 845 observations and 15 variables. This dataset contains information regarding the amount of trash per different trash type between 2014 and 2023 for each type of trash wheel.

From the available data, the total weight (in tons) of trash collected by Professor Trash Wheel was 216.26. The total weight (in tons) of trash collected by Mr. Trash Wheel was 1875.1. Simiarly, the total weight (in tons) of trash collected by Gwynnda Trash Wheel was 451.65. Additionally, the total number of cigarette butts collected by Gwynnda in July of 2021 was 1.63×10^4.

Problem 3

```
#loading the baseline dataset
baseline_demos <- read_csv("data_hw2/data_mci/MCI_baseline.csv", skip = 1) %>%
  #cleaning dataset
  janitor::clean_names() %>%
  #renaming the column names
  rename(
    'baseline_age' = 'current_age',
    'education_years' = 'education',
    'apoe4_carrier' = 'apoe4',
    'onset_age' = 'age_at_onset')
```

```
#continuation of tidying data
baseline_demos <- baseline_demos%>%
  mutate(
    #changing sex values
    sex = case_when(
      sex == 0 ~ "female",
      sex == 1 ~ "male"),
    #changing APOE4 carrier values
    apoe4_carrier = case_when(
      apoe4_carrier == 0 ~ "non-carrier",
      apoe4_carrier == 1 ~ "carrier"),
    #adding NA's in blank spaces
    onset_age = case_when(
      onset_age == "." ~ NA,
      TRUE ~ as.numeric(onset_age)
    )) %>%
  #filtering no MCI
  filter(onset_age > baseline_age | is.na(onset_age))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `onset_age = case_when(onset_age == "." ~ NA, TRUE ~
##   as.numeric(onset_age))`.
## Caused by warning:
## ! NAs introduced by coercion
```

The `baseline_demos` dataset was imported and tidyied up by renaming column names when necessary, mutating the `sex` variable values into female or male, and similarly the `apoe4_carrier` into carrier or non-carrier. Additionally, NA's were needed to `onset_age` variable, and we filtered for participants without MCI at baseline. The resulting dataset contains a total of 479 observations and 6 variables. There was a total of 483 participants recruited and of those 479 developed MCI. The average baseline age was 65.0286013. The proportion of women in the study that were APOE4 carriers were 1.3883×10^4 / 479.

```
#loading other datasets
amyloid_dataset <- read_csv("data_hw2/data_mci/mci_amyloid.csv", skip = 1) %>%
  #cleaning dataset
  janitor::clean_names() %>%
  #renaming colname
  rename('id' = 'study_id') %>%
  #transforming dataset
  pivot_longer(baseline:time_8,
               names_to = 'time_elapsed',
               values_to = 'biomarkers_ratio')
```

```
## Rows: 487 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Baseline, Time 2, Time 4, Time 6, Time 8
## dbl (1): Study ID
##
## 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.
```

The `amyloid_dataset` dataset was imported and tidyied up by renaming column names when necessary, and transforming the dataset from 'wide' format to 'long' format. The resulting dataset contains a total of 2435 observations and 3 variables.

```
#observations in the baseline_demos dataset, but not in amyloid_dataset
anti_join(baseline_demos, amyloid_dataset, by = 'id')
```

```
## # A tibble: 8 x 6
##      id baseline_age sex    education_years apoe4_carrier onset_age
##   <dbl>        <dbl> <chr>            <dbl> <chr>             <dbl>
## 1    14         58.4 female              20 non-carrier        66.2
## 2    49         64.7 male                16 non-carrier        68.4
## 3    92         68.6 female              20 non-carrier        NA
## 4   179         68.1 male                16 non-carrier        NA
## 5   268         61.4 female              18 carrier            67.5
## 6   304         63.8 female              16 non-carrier        NA
## 7   389         59.3 female              16 non-carrier        NA
## 8   412         67   male                16 carrier            NA
```

Of the participants that were in the `baseline_demos` dataset, most were non-carriers of APOE4, and most of their onset age is missing.

```
#observations in amyloid_dataset, but not in baseline_demos dataset
anti_join(amyloid_dataset, baseline_demos, by = 'id')
```

```
## # A tibble: 80 x 3
##       id time_elapsed biomarkers_ratio
##    <dbl> <chr>        <chr>
##  1    72 baseline     0.106965463
##  2    72 time_2       <NA>
##  3    72 time_4       0.107266218
##  4    72 time_6       0.106665207
##  5    72 time_8       <NA>
##  6   234 baseline     0.110521689
##  7   234 time_2       0.110988335
##  8   234 time_4       0.110318671
##  9   234 time_6       0.107334344
## 10   234 time_8       0.108868811
## # i 70 more rows
```

Based on these findings, we can see that the participants that were not in the `baseline_demos`has simialr biomarker ratios.

```
#dataset with common observations
common_amyloid_dataset <- inner_join(baseline_demos, amyloid_dataset, by = "id")

#export csv
write_csv(common_amyloid_dataset,"data_hw2/common_amyloid_dataset.csv")
```

After combining both datasets, the `common_amyloid_dataset` is produced. This dataset contains at total of 2355 observations and 8 variables. This dataset contains information of participants' biomarkers ratio obtained at different times/stages in the study, and shows whether the participant is a APOE4 carrier or not, the participants' sex, ages at baseline and onset, and their years of education.