p8105_hw2_hc3451

Huanyu Chen

2023-09-27

Problem 1

```
pols_month <- read.csv("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/pols-month.csv")
snp <- read.csv("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/snp.csv")
unemployment <- read.csv("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/unemployment.csv")

# First Step of Data Cleaning: pols_month
pols_month <- pols_month |>
  separate(mon, into = c("year", "month", "day"), sep = "-") |>
  mutate(month = month.name[as.numeric(month)]) |>
  mutate(president = ifelse("prez_dem" == 1, "dem", "gop")) |>
  select(-prez_dem, -prez_gop, -day)
```

```
# Second Step of Data Cleaning: snp
snp <- snp |>
  separate(date, into = c("month", "day", "year"), sep = "/") |>
  mutate(month = month.name[as.numeric(month)]) |>
  mutate(year = ifelse(as.numeric(year) <= 20, paste0("20", year), paste0("19", year))) |>
  select(year, month, everything())
```

```
# Third Step of Data Cleaning: unemployment
unemployment = pivot_longer(unemployment, Jan:Dec, names_to = "month", values_to = "unemployment")
unemployment <- unemployment |>
  mutate(month = month.name[factor(month)]) |>
  mutate(year = tolower(Year)) |>
  select(-Year) |>
  select(year, month, unemployment)
```

```
# Join the datasets
merged_data_1 <- merge(pols_month, snp, by = c("year", "month"), all.x = TRUE)
merged_data <- merge(merged_data_1, unemployment, by = c("year", "month"), all.x = TRUE)
```

Conclusion

The final merged dataset involves three datasets: "pols" containing political data, "snp" with stock market information, and "unemployment" providing economic indicators. It comprises 822 observations and 12 variables, spanning from year 1947 to 2015. Key variables include `year`, `month`, and `unemployment_rate`, alongside some political and stock market indicators.

Problem 2

```r
# mrTrash
mrTrash <- read_excel("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/202207 Trash Wheel C

mrTrash <- janitor::clean_names(mrTrash)
mrTrash <- separate(mrTrash, date, into = c("year", "month", "day"), sep = "-")

mrTrash <- mrTrash |>
  select(dumpster, year, month, everything()) |>
  mutate(homes_powered = round(weight_tons * 500 / 30))
```

```r
# profTrash
profTrash <- read_excel("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/202207 Trash Wheel

profTrash <- janitor::clean_names(profTrash)
profTrash <- separate(profTrash, date, into = c("year", "month", "day"), sep = "-")

profTrash <- profTrash |>
  select(dumpster, year, month, everything()) |>
  mutate(homes_powered = round(weight_tons * 500 / 30))
```

```r
# gwyTrash
gwyTrash <- read_excel("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/202207 Trash Wheel 

gwyTrash <- janitor::clean_names(gwyTrash)
gwyTrash <- separate(gwyTrash, date, into = c("year", "month", "day"), sep = "-")

gwyTrash <- gwyTrash |>
  select(dumpster, year, month, everything()) |>
  mutate(homes_powered = round(weight_tons * 500 / 30))
```

```r
# Combine dataset
mrTrash <- mrTrash |>
  mutate(source = "Mr Trash") |>
  mutate(year = as.character(year))

profTrash <- profTrash |>
  mutate(source = "Prof Trash") |>
  mutate(year = as.character(year))

gwyTrash <- gwyTrash |>
  mutate(source = "Gwy Trash") |>
  mutate(year = as.character(year))

combined_trash <- bind_rows(mrTrash, profTrash, gwyTrash)
combined_trash <- combined_trash |>
  dplyr:::select(dumpster, year, month, day, source, everything()) |>
  arrange(year, match(month, month.name), day)

filtered_data <- combined_trash |>
  filter(source == "Gwy Trash" & year == 2021 & month == "07")
```

```
filtered_data
```

```
## # A tibble: 5 x 16
##   dumpster year  month day   source    weight_tons volume_cubic_yards
##      <dbl> <chr> <chr> <chr> <chr>           <dbl>              <dbl>
## 1        1 2021  07    03    Gwy Trash        0.93                 15
## 2        2 2021  07    07    Gwy Trash        2.26                 15
## 3        3 2021  07    07    Gwy Trash        1.62                 15
## 4        4 2021  07    16    Gwy Trash        1.76                 15
## 5        5 2021  07    30    Gwy Trash        1.53                 15
## # i 9 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, grocery_bags <dbl>,
## #   chip_bags <dbl>, sports_balls <dbl>, homes_powered <dbl>,
## #   plastic_bags <dbl>
```

```
total_cigarette_butts_july_2021 <- sum(pull(filtered_data, cigarette_butts))
```

Conclusion

This combined dataset has 747 observations. The variable `source` represents the origin of the data and `homes_powered` represents the number of homes powered based on electricity from trash. The total weight of trash collected by Professor Trash Wheel is 190.12 and the total number of cigarette butts collected by Gwynnda in July of 2021 is 1.63×10^4.

Problem 3

```
base <- read.csv("/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/MCI_baseline.csv", skip =
base <- janitor::clean_names(base)
base <- base |>
  mutate(sex = if_else(sex == 0, 'female', 'male')) |>
  mutate(apoe4 = if_else(apoe4 == 0, 'non-carrier', 'carrier'))

base_filtered <- base |>
  filter(age_at_onset != ".")
```

Important steps in the import process and relevant features of the dataset include deleting the first row, cleaning variable names to a tidy format, and mutating the 'sex' and 'apoe4' variables from numeric values to their respective real categories. 483 participants were recruited, and of these 97 develop MCI.

```
base_mean = mean(pull(base_filtered, current_age))

data_female <- base_filtered |>
  filter(sex == "female")
proportion <- sum(data_female$apoe4 == "carrier") / nrow(data_female)
```

The average baseline age is 65.6113402. Moreover, 65.22% of women in the study are APOE4 carriers.

```r
amyloid <- read.csv('/Users/huanyu/Documents/CUIMC/Data Science/p8105_hw2_hc3451/mci_amyloid.csv', skip
amyloid <- janitor::clean_names(amyloid)
amyloid <- amyloid |>
  rename(id = study_id,
         ratio_2_year = time_2,
         ratio_4_year = time_4,
         ratio_6_year = time_6,
         ratio_8_year = time_8)
```

Important steps in the import process and relevant features of the dataset include deleting the first row, cleaning variable names to a tidy format, and renaming variables to their respective real categories.

```r
base_participants <- base_filtered$participant_id
amyloid_participants <- amyloid$participant_id

only_in_base <- setdiff(base_filtered$id, amyloid$id)
only_in_amyloid <- setdiff(amyloid$id, base_filtered$id)

combined_dataset <- inner_join(base_filtered, amyloid, by = "id")
write.csv(combined_dataset, "combined_dataset.csv", row.names = FALSE)
```

Conclusion

Some participants (with id: 14, 49, 268) appear in only the baseline; while some participants appear in only the amyloid datasets. `combined_dataset` (has 94 observations) combined participants who appear in both datasets.