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)
head(pols_month)
```

```
##   year    month gov_gop sen_gop rep_gop gov_dem sen_dem rep_dem president
## 1 1947  January      23      51     253      23      45     198       gop
## 2 1947 February      23      51     253      23      45     198       gop
## 3 1947    March      23      51     253      23      45     198       gop
## 4 1947    April      23      51     253      23      45     198       gop
## 5 1947      May      23      51     253      23      45     198       gop
## 6 1947     June      23      51     253      23      45     198       gop
```

```
# 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())
head(snp)
```

```
##   year    month day   close
## 1 2015     July   1 2079.65
## 2 2015     June   1 2063.11
## 3 2015      May   1 2107.39
## 4 2015    April   1 2085.51
## 5 2015    March   2 2067.89
## 6 2015 February   2 2104.50
```

```r
# 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)
head(unemployment)
```

```
## # A tibble: 6 x 3
##   year  month     unemployment
##   <chr> <chr>            <dbl>
## 1 1948  May                3.4
## 2 1948  April              3.8
## 3 1948  August             4
## 4 1948  January            3.9
## 5 1948  September          3.5
## 6 1948  July               3.6
```

```r
# 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)
head(merged_data)
```

```
##   year    month gov_gop sen_gop rep_gop gov_dem sen_dem rep_dem president  day
## 1 1947    April      23      51     253      23      45     198       gop <NA>
## 2 1947   August      23      51     253      23      45     198       gop <NA>
## 3 1947 December      24      51     253      23      45     198       gop <NA>
## 4 1947 February      23      51     253      23      45     198       gop <NA>
## 5 1947  January      23      51     253      23      45     198       gop <NA>
## 6 1947     July      23      51     253      23      45     198       gop <NA>
##   close unemployment
## 1    NA           NA
## 2    NA           NA
## 3    NA           NA
## 4    NA           NA
## 5    NA           NA
## 6    NA           NA
```

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))

head(mrTrash)
```

```
## # A tibble: 6 x 14
##   dumpster year  month day   weight_tons volume_cubic_yards plastic_bottles
##      <dbl> <chr> <chr> <chr>       <dbl>              <dbl>           <dbl>
## 1        1 2014  05    16           4.31                 18            1450
## 2        2 2014  05    16           2.74                 13            1120
## 3        3 2014  05    16           3.45                 15            2450
## 4        4 2014  05    17           3.1                  15            2380
## 5        5 2014  05    17           4.06                 18             980
## 6        6 2014  05    20           2.71                 13            1430
## # i 7 more variables: polystyrene <dbl>, cigarette_butts <dbl>,
## #   glass_bottles <dbl>, grocery_bags <dbl>, chip_bags <dbl>,
## #   sports_balls <dbl>, homes_powered <dbl>
```

```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))

head(profTrash)
```

```
## # A tibble: 6 x 13
##   dumpster year  month day   weight_tons volume_cubic_yards plastic_bottles
##      <dbl> <chr> <chr> <chr>       <dbl>              <dbl>           <dbl>
## 1        1 2017  01    02           1.79                 15            1950
## 2        2 2017  01    30           1.58                 15            9540
## 3        3 2017  02    26           2.32                 18            8350
## 4        4 2017  02    26           3.72                 15            8590
## 5        5 2017  02    28           1.45                 15            7830
## 6        6 2017  03    30           1.71                 15            8210
## # i 6 more variables: polystyrene <dbl>, cigarette_butts <dbl>,
## #   glass_bottles <dbl>, grocery_bags <dbl>, chip_bags <dbl>,
## #   homes_powered <dbl>
```

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

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))

head(gwyTrash)
```

```
## # A tibble: 6 x 11
##   dumpster year  month day   weight_tons volume_cubic_yards plastic_bottles
##      <dbl> <chr> <chr> <chr>       <dbl>              <dbl>           <dbl>
## 1        1 2021  07    03           0.93                 15            1200
## 2        2 2021  07    07           2.26                 15            2000
## 3        3 2021  07    07           1.62                 15            1800
## 4        4 2021  07    16           1.76                 15            1000
## 5        5 2021  07    30           1.53                 15            2100
## 6        6 2021  08    11           2.06                 15            2400
## # i 4 more variables: polystyrene <dbl>, cigarette_butts <dbl>,
## #   plastic_bags <dbl>, homes_powered <dbl>
```

```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.