

p8105_hw2_mc5698.Rmd

2024-09-27

#Question 1

```
#loading necessary packages
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v readr      2.1.5
```

```
## v forcats    1.0.0      v stringr   1.5.1
```

```
## v ggplot2    3.5.1      v tibble    3.2.1
```

```
## v lubridate  1.9.3      v tidyr     1.3.1
```

```
## 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 become errors
```

```
library(readxl)
```

```
#clean the dataset
```

```
nyc_t =
```

```
  read_csv(
```

```
    "/Users/nicolechen/Downloads/p8105_hw2_mc5698/dataset/NYC_Transit_Subway_Entrance_And_Exit_Data.csv"
```

```
    col_types = cols(Route8 = "c", Route9 = "c", Route10 = "c", Route11 = "c")) |>
```

```
  janitor::clean_names() |>
```

```
  select(
```

```
    line, station_name, station_latitude, station_longitude, route1, route2, route3, route4, route5, route6, route7, route8, route9, route10, route11, entry, vending, entrance_type, ada
```

```
  mutate(
```

```
    entry = ifelse(entry == "YES", TRUE, FALSE))
```

The dataset contains line, station_name, station_latitude, station_longitude, route1, route2, route3, route4, route5, route6, route7, route8, route9, route10, route11, entry, vending, entrance_type, ada. For the data cleaning, I removed unnecessary columns and convert the entry variable from character to a logical variable by using `case_match` function. The dimension of the resulting dataset is 1868, 19. These data are mostly tidy but we could pivot different route columns into one variable.

```
distinct_stations=
```

```
  nyc_t|>
```

```
  distinct(station_name,line)
```

```
nrow(distinct_stations)
```

```
## [1] 465
```

```

ada_stations=
  nyc_t |>
  filter(ada==TRUE) |>
  distinct(station_name, line)

nrow(ada_stations)

```

```
## [1] 84
```

```

no_vending=
  nyc_t |>
  filter(vending == "NO") |>
  pull(entry)

proportion_entry= mean(no_vending)
proportion_entry

```

```
## [1] 0.3770492
```

There are 465 distinct stations. 84 stations are ADA compliant. The proportion of station entrances/exits without vending allow entrance is `proportion_entry`.

```

transfrom_ent=
  nyc_t |>
  pivot_longer(
    route1:route11,
    names_to = "route_num",
    values_to = "route")

A_stations=
  transfrom_ent |>
  filter(route == "A") |>
  select(station_name, line) |>
  distinct()

ada_stations =
  transfrom_ent |>
  filter(route == "A", ada == TRUE) |>
  select(station_name, line) |>
  distinct()

```

There are 56 distinct stations serve the A train and 16 stations serve the A train and ADA compliant.

#Question 2

```

#clean the datasets
mr_trash_wheel =
  readxl::read_excel("/Users/nicolechen/Downloads/p8105_hw2_mc5698/dataset/202409TrashWheelCollectionData.xlsx") |>
  filter(!is.na(Dumpster)) |>
  mutate(Sports_Balls = as.integer(round(`Sports Balls`)),
         Year = as.character(Year),
         Trash_Wheel = "Mr. Trash Wheel")

```

```
## New names:
## * ' -> '...15'
## * ' -> '...16'
```

```
professor_trash_wheel =
  readxl::read_excel("/Users/nicolechen/Downloads/p8105_hw2_mc5698/dataset/202409TrashWheelCollectionDa
  filter(!is.na(Dumpster)) |>
  mutate(Year = as.character(Year),
         Trash_Wheel = "Professor Trash Wheel")

gwynnda_trash_wheel =
  readxl::read_excel("/Users/nicolechen/Downloads/p8105_hw2_mc5698/dataset/202409TrashWheelCollectionDa
  filter(!is.na(Dumpster)) |>
  mutate(Year = as.character(Year),
         Trash_Wheel = "Gwynnda Trash Wheel")

combined_data =
  bind_rows(mr_trash_wheel, professor_trash_wheel, gwynnda_trash_wheel)
combined_data
```

```
## # A tibble: 1,033 x 18
##   Dumpster Month Year Date           'Weight (tons)'
##   <dbl> <chr> <chr> <dtm>           <dbl>
## 1      1 May 2014 2014-05-16 00:00:00 4.31
## 2      2 May 2014 2014-05-16 00:00:00 2.74
## 3      3 May 2014 2014-05-16 00:00:00 3.45
## 4      4 May 2014 2014-05-17 00:00:00 3.1
## 5      5 May 2014 2014-05-17 00:00:00 4.06
## 6      6 May 2014 2014-05-20 00:00:00 2.71
## 7      7 May 2014 2014-05-21 00:00:00 1.91
## 8      8 May 2014 2014-05-28 00:00:00 3.7
## 9      9 June 2014 2014-06-05 00:00:00 2.52
## 10    10 June 2014 2014-06-11 00:00:00 3.76
## # i 1,023 more rows
## # i 13 more variables: 'Volume (cubic yards)' <dbl>, 'Plastic Bottles' <dbl>,
## #   Polystyrene <dbl>, 'Cigarette Butts' <dbl>, 'Glass Bottles' <dbl>,
## #   'Plastic Bags' <dbl>, Wrappers <dbl>, 'Sports Balls' <dbl>,
## #   'Homes Powered*' <dbl>, ...15 <lgl>, ...16 <lgl>, Sports_Balls <int>,
## #   Trash_Wheel <chr>
```

By reading and cleaning the datasets, I combined the three datasets from Mr. Trash Wheel, Professor Trash Wheel and Gwynnda Trash Wheel. There are 1033 observations in the combined dataset. This dataset includes key variables such as `Dumpster`, which shows the the number of dumpster filled by trash, and `Cigarette Butts` which means the number of cigarette they collected. It also includes the specific time of the trash such as `Year`, `Date`, `Month` and `Trash_Wheel` indicates different trash types correspond to the different trash wheel. Moreover, it provides the detailed volumn and types for each trash wheel.

```
tw_professor =
  combined_data |>
  filter(Trash_Wheel == "Professor Trash Wheel") |>
  summarise(total_weight = sum(`Weight (tons)`, na.rm = TRUE))
```

```
cb_gwynnda_june2022 =  
  combined_data |>  
  filter(Trash_Wheel == "Gwynnda Trash Wheel", Year == "2022", Month == "June") |>  
  summarise(total_cig_butts = sum(`Cigarette Butts`, na.rm = TRUE))
```

The total weight of trash collected by Professor Trash Wheel was 246.74. The total number of cigarette butts collected by Gwynnda in June of 2022 was 1.812×10^4 .

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist  
## Min.   : 4.0    Min.   :  2.00  
## 1st Qu.:12.0    1st Qu.: 26.00  
## Median :15.0    Median : 36.00  
## Mean   :15.4    Mean   : 42.98  
## 3rd Qu.:19.0    3rd Qu.: 56.00  
## Max.   :25.0    Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.