# title: "p8105_hw2_hs3478"

## author: "Huachen Shan"

## date: "2024-10-02"

## output: github_document

In [ ]:
```
Problem 1
1.1 Import the dataset and clean names

transit_df =
  read_csv("hw2data/NYC_Transit_Subway_Entrance_And_Exit_Data.csv", na = c("NA", "",
".")) |>
  janitor::clean_names()
```

**Rows: 1868 Columns: 32**

**── Column specification ──────────────────────────────────────────────────────**

**Delimiter: ","**

**chr (22): Division, Line, Station Name, Route1, Route2, Route3, Route4, Rout...**

**dbl (8): Station Latitude, Station Longitude, Route8, Route9, Route10, Rout...**

**lgl (2): ADA, Free Crossover**

**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.**

In [ ]:
```
tail(transit_df, 10)
```

In [ ]:
```
# # A tibble: 10 × 32
####    division line   station_name station_latitude station_longitude route1 route
2
####    <chr>    <chr>  <chr>                  <dbl>             <dbl> <chr>  <chr>
####  1 IRT      White… Simpson St              40.8              -73.9 2      5
####  2 IRT      White… Simpson St              40.8              -73.9 2      5
####  3 IRT      White… Simpson St              40.8              -73.9 2      5
####  4 IRT      White… Simpson St              40.8              -73.9 2      5
####  5 IRT      White… Simpson St              40.8              -73.9 2      5
####  6 IRT      White… Wakefield-2…            40.9              -73.9 2      5
####  7 IRT      White… Wakefield-2…            40.9              -73.9 2      5
####  8 IRT      White… Wakefield-2…            40.9              -73.9 2      5
####  9 IRT      Flush… 34 St Hudso…            40.8              -74.0 7      <NA
>
#### 10 IRT      Flush… 34 St Hudso…            40.8              -74.0 7      <NA
>
### # ℹ 25 more variables: route3 <chr>, route4 <chr>, route5 <chr>, route6 <chr>,
### #   route7 <chr>, route8 <dbl>, route9 <dbl>, route10 <dbl>, route11 <dbl>,
### #   entrance_type <chr>, entry <chr>, exit_only <chr>, vending <chr>,
### #   staffing <chr>, staff_hours <chr>, ada <lgl>, ada_notes <chr>,
### #   free_crossover <lgl>, north_south_street <chr>, east_west_street <chr>,
### #   corner <chr>, entrance_latitude <dbl>, entrance_longitude <dbl>,
### #   station_location <chr>, entrance_location <chr>
```

**Select variables**

**Select columns and clean the data; convert entry to logical.**

In [ ]:
```
Select columns and clean the data; convert `entry` to logical.

transit_select <- transit_df |>
  select(
    line,
    station_name,
    station_latitude,
    station_longitude,
    starts_with("route"),
    entry,
    vending,
    entrance_type,
    ada
  ) |>
  mutate(
    entry = entry == "YES"
  )
```

### 1.3 Analysis the data

#### 1.3.1 Calculate the distinct station.

In [ ]:
```
distinct_stations <- transit_select |>
  distinct(line, station_name) |>
  nrow()

distinct_stations
```

**[1] 465**

**There are 465 distinct stations are there.**

### 1.3.2 Calculate the number of ADA compliant.

In [ ]:
```
ADA_compliant_stations <-
  transit_select |>
  filter(ada == "TRUE") |>
  distinct(line, station_name) |>
  nrow()

ADA_compliant_stations
```

**[1] 84**

**84 stations are ADA compliant.**

### 1.3.3 Propotion of station

**Calculate the total number of stations without vending.**

In [ ]:
```
no_vending <- transit_select |>
  filter(vending == "NO") |>
  distinct(line, station_name) |>
  nrow()
```

**Calculate how many of them allow entry.**

In [ ]:
```
allow_entry_no_vending <- transit_select |>
  filter(vending == "NO", entry == TRUE) |>
  distinct(line, station_name) |>
  nrow()
```

**[1] 0.4343434**

### 1.4 Reform the data.

### Problem 2

### 2.1 Import the Mr. Trash Wheel sheet and clean names

In [ ]:
```
mr_df =
  read_excel("hw2data/202409 Trash Wheel Collection Data.xlsx", sheet = "Mr. Trash Wheel", skip = 1) |>
  janitor::clean_names()
```

**New names:**

- `` -> ...15`
- `` -> ...16`

In [ ]:
```
## # A tibble: 10 × 16
##    dumpster month year  date                weight_tons volume_cubic_yards
##       <dbl> <chr> <chr> <dttm>                    <dbl>              <dbl>
##  1        1 May   2014  2014-05-16 00:00:00        4.31                 18
##  2        2 May   2014  2014-05-16 00:00:00        2.74                 13
##  3        3 May   2014  2014-05-16 00:00:00        3.45                 15
##  4        4 May   2014  2014-05-17 00:00:00        3.1                  15
##  5        5 May   2014  2014-05-17 00:00:00        4.06                 18
##  6        6 May   2014  2014-05-20 00:00:00        2.71                 13
##  7        7 May   2014  2014-05-21 00:00:00        1.91                  8
##  8        8 May   2014  2014-05-28 00:00:00        3.7                  16
##  9        9 June  2014  2014-06-05 00:00:00        2.52                 14
## 10       10 June  2014  2014-06-11 00:00:00        3.76                 18
## # ℹ 10 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, sports_balls <dbl>, homes_powered <dbl>, x15 <lgl>,
## #   x16 <lgl>
```

In [ ]:
```
### 2.1 Import the `Mr. Trash Wheel` sheet and clean names
```{r}
mr_df <-
  read_excel("hw2data/202409 Trash Wheel Collection Data.xlsx", sheet = "Mr. Trash Wheel", skip = 1) |>
  janitor::clean_names()

head(mr_df, 10)

mr_clean <-
  mr_df |>
  select(dumpster:homes_powered) |>
  drop_na(dumpster) |>  # Remove rows without dumpster-related data
  mutate(
    sports_balls = as.integer(round(sports_balls)),  # Round sports balls count to the nearest integer and convert to integer type
    year = as.numeric(year),  # Ensure year is numeric
    trash_wheel = "Mr"  # Add a new column identifying the trash wheel as "Mr"
  )

tail(mr_clean, 10)
```
```

In [ ]:
```
## # A tibble: 10 × 15
##    dumpster month  year date                weight_tons volume_cubic_yards
##       <dbl> <chr> <dbl> <dttm>                    <dbl>              <dbl>
##  1      642 April  2024 2024-04-04 00:00:00        4.3                15
##  2      643 April  2024 2024-04-09 00:00:00        3.49               15
##  3      644 April  2024 2024-04-09 00:00:00        1.25               15
##  4      645 May    2024 2024-05-02 00:00:00        4.9                15
##  5      646 May    2024 2024-05-10 00:00:00        3.68               15
##  6      647 May    2024 2024-05-10 00:00:00        4.7                15
##  7      648 May    2024 2024-05-30 00:00:00        4.13               15
##  8      649 May    2024 2024-05-30 00:00:00        3.34               15
##  9      650 June   2024 2024-06-11 00:00:00        3.02               15
## 10      651 June   2024 2024-06-11 00:00:00        4                  15
## # ℹ 9 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, sports_balls <int>, homes_powered <dbl>, trash_wheel <chr>
```

### 2.2 Import the Professor Trash Wheel sheet and clean names

In [ ]:
```
professor_df <-
  read_excel("hw2data/202409 Trash Wheel Collection Data.xlsx", sheet = "Professor Trash Wheel", skip = 1) |>
  janitor::clean_names()

head(professor_df, 10)
```

In [ ]:
```
## # A tibble: 10 × 13
##    dumpster month     year date                weight_tons volume_cubic_yards
##       <dbl> <chr>    <dbl> <dttm>                    <dbl>              <dbl>
##  1        1 January   2017 2017-01-02 00:00:00        1.79                 15
##  2        2 January   2017 2017-01-30 00:00:00        1.58                 15
##  3        3 February  2017 2017-02-26 00:00:00        2.32                 18
##  4        4 February  2017 2017-02-26 00:00:00        3.72                 15
##  5        5 February  2017 2017-02-28 00:00:00        1.45                 15
##  6        6 March     2017 2017-03-30 00:00:00        1.71                 15
##  7        7 April     2017 2017-04-01 00:00:00        1.82                 15
##  8        8 April     2017 2017-04-20 00:00:00        2.37                 15
##  9        9 May       2017 2017-05-10 00:00:00        2.64                 15
## 10       10 May       2017 2017-05-26 00:00:00        2.78                 15
## # ℹ 7 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, homes_powered <dbl>
```

In [ ]:
```
professor_clean <-
  professor_df |>
  select(dumpster:homes_powered) |>
  drop_na(dumpster, month) |>  # Remove rows missing dumpster or month data
  mutate(trash_wheel = "Professor")  # Add a column indicating the trash wheel as "P
rofessor"

tail(professor_clean, 10)
```

In [ ]:
```
## # A tibble: 10 × 14
##    dumpster month      year date                weight_tons volume_cubic_yards
##       <dbl> <chr>     <dbl> <dttm>                    <dbl>              <dbl>
##  1      109 August     2023 2023-08-09 00:00:00        2.64                 15
##  2      110 August     2023 2023-08-23 00:00:00        2.82                 15
##  3      111 September  2023 2023-09-06 00:00:00        1.71                 10
##  4      112 September  2023 2023-09-26 00:00:00        2.43                 15
##  5      113 October    2023 2023-10-27 00:00:00        1.28                 10
##  6      114 December   2023 2023-12-05 00:00:00        2.13                 15
##  7      115 January    2024 2024-01-08 00:00:00        2.2                  15
##  8      116 March      2024 2024-03-14 00:00:00        3.75                 15
##  9      117 April      2024 2024-04-16 00:00:00        3                    15
## 10      118 May        2024 2024-05-30 00:00:00        2.48                 15
## # ℹ 8 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, homes_powered <dbl>, trash_wheel <chr>
```

### 2.3 Import the Gwynnda Trash Wheel sheet and clean names

In [ ]:
```
gwynnda_df =
  read_excel("hw2data/202409 Trash Wheel Collection Data.xlsx", sheet = "Gwynnda Tra
sh Wheel", skip = 1) |>
  janitor::clean_names()

head(gwynnda_df, 10)
```

In [ ]:
```
## # A tibble: 10 × 12
##    dumpster month   year date                weight_tons volume_cubic_yards
##       <dbl> <chr>  <dbl> <dttm>                    <dbl>              <dbl>
##  1        1 July    2021 2021-07-03 00:00:00        0.93                 15
##  2        2 July    2021 2021-07-07 00:00:00        2.26                 15
##  3        3 July    2021 2021-07-07 00:00:00        1.62                 15
##  4        4 July    2021 2021-07-16 00:00:00        1.76                 15
##  5        5 July    2021 2021-07-30 00:00:00        1.53                 15
##  6        6 August  2021 2021-08-11 00:00:00        2.06                 15
##  7        7 August  2021 2021-08-14 00:00:00        1.9                  15
##  8        8 August  2021 2021-08-16 00:00:00        2.16                 15
##  9        9 August  2021 2021-08-16 00:00:00        2.6                  15
## 10       10 August  2021 2021-08-17 00:00:00        3.21                 15
## # ℹ 6 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, plastic_bags <dbl>, wrappers <dbl>,
## #   homes_powered <dbl>
```

In [ ]:
```
gwynnda_clean =
  gwynnda_df |>
  select(dumpster:homes_powered) |>
  drop_na(dumpster) |>
  mutate(trash_wheel = "Gwynnda")

tail(gwynnda_clean, 10)
```

In [ ]:
```
## # A tibble: 10 × 13
##    dumpster month  year date                weight_tons volume_cubic_yards
##       <dbl> <chr> <dbl> <dttm>                    <dbl>              <dbl>
##  1      253 April  2024 2024-04-05 00:00:00        3.43                 15
##  2      254 April  2024 2024-04-26 00:00:00        2.87                 15
##  3      255 May    2024 2024-05-14 00:00:00        3.27                 15
##  4      256 May    2024 2024-05-29 00:00:00        2.72                 15
##  5      257 May    2024 2024-05-29 00:00:00        3                    15
##  6      258 May    2024 2024-05-29 00:00:00        3.78                 15
##  7      259 May    2024 2024-05-30 00:00:00        3.35                 15
##  8      260 May    2024 2024-05-31 00:00:00        3.55                 15
##  9      261 June   2024 2024-06-01 00:00:00        2.88                 15
## 10      262 June   2024 2024-06-07 00:00:00        3.43                 15
## # ℹ 7 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, plastic_bags <dbl>, wrappers <dbl>,
## #   homes_powered <dbl>, trash_wheel <chr>
```

### 2.4 Bind the datasets

In [ ]:
```
bind_data =
  bind_rows(mr_clean, professor_clean, gwynnda_clean) |>
  janitor::clean_names() |>
  select(trash_wheel, everything())

head(bind_data, 10)
```

In [ ]:
```
## # A tibble: 10 × 15
##    trash_wheel dumpster month  year date                weight_tons
##    <chr>          <dbl> <chr> <dbl> <dttm>                    <dbl>
##  1 Mr                 1 May    2014 2014-05-16 00:00:00        4.31
##  2 Mr                 2 May    2014 2014-05-16 00:00:00        2.74
##  3 Mr                 3 May    2014 2014-05-16 00:00:00        3.45
##  4 Mr                 4 May    2014 2014-05-17 00:00:00        3.1
##  5 Mr                 5 May    2014 2014-05-17 00:00:00        4.06
##  6 Mr                 6 May    2014 2014-05-20 00:00:00        2.71
##  7 Mr                 7 May    2014 2014-05-21 00:00:00        1.91
##  8 Mr                 8 May    2014 2014-05-28 00:00:00        3.7
##  9 Mr                 9 June   2014 2014-06-05 00:00:00        2.52
## 10 Mr                10 June   2014 2014-06-11 00:00:00        3.76
## # ℹ 9 more variables: volume_cubic_yards <dbl>, plastic_bottles <dbl>,
## #   polystyrene <dbl>, cigarette_butts <dbl>, glass_bottles <dbl>,
## #   plastic_bags <dbl>, wrappers <dbl>, sports_balls <int>, homes_powered <dbl>
```

**Total description**

In [ ]:
```
total_weight_professor =
  professor_clean |>
  summarise(total_weight = sum(weight_tons, na.rm = TRUE)) |>
  pull(total_weight)

total_weight_professor
```

**[1] 246.74**

In [ ]:
```
total_cigarette_butts_gwynnda =
  gwynnda_clean |>
  filter(year == 2022 & month == "June") |>
  summarise(
    total_cigarette_butts = sum(cigarette_butts, na.rm = TRUE)
  ) |>
  pull(total_cigarette_butts)

total_cigarette_butts_gwynnda
```

In [ ]:
```
## [1] 18120
```