p8105_hw2_yf2735

Yujing FU

2024-09-26

```
## -- 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
```

problem 2

```
library(readxl)
library(dplyr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Mr. Trash Wheel dataset

```
mr_trash_df =
  read_excel("data/202309 Trash Wheel Collection Data.xlsx", sheet = "Mr. Trash Wheel", skip = 1) |>
  select(-starts_with("...")) |>
  janitor::clean_names() |>
  filter(!is.na(dumpster)) |>
  mutate(sports_balls = as.integer(round(sports_balls, 0)))
```

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

```
mr_trash_df
```

```
## # A tibble: 584 x 14
##    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 574 more rows
## # i 8 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, sports_balls <int>, homes_powered <dbl>
```

Professor Trash Wheel dataset

```
prof_trash_df =
  read_excel("data/202309 Trash Wheel Collection Data.xlsx", sheet = "Professor Trash Wheel", skip = 1) |>
  select(-starts_with("...")) |>
  janitor::clean_names() |>
  filter(!is.na(dumpster)) |>
  mutate(year = as.character(year))
prof_trash_df
```

```
## # A tibble: 106 x 13
##    dumpster month    year  date                weight_tons volume_cubic_yards
##       <dbl> <chr>    <chr> <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 96 more rows
## # i 7 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, glass_bottles <dbl>, plastic_bags <dbl>,
## #   wrappers <dbl>, homes_powered <dbl>
```

Gwynnda Trash Wheel

```
gwynnda_trash_df =
  read_excel("data/202309 Trash Wheel Collection Data.xlsx", sheet = "Gwynnda Trash Wheel", skip = 1) |>
  select(-starts_with("...")) |>
  janitor::clean_names() |>
  filter(!is.na(dumpster)) |>
  mutate(year = as.character(year))
gwynnda_trash_df
```

```
## # A tibble: 155 x 12
##    dumpster month  year  date                weight_tons volume_cubic_yards
##       <dbl> <chr>  <chr> <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 145 more rows
## # i 6 more variables: plastic_bottles <dbl>, polystyrene <dbl>,
## #   cigarette_butts <dbl>, plastic_bags <dbl>, wrappers <dbl>,
## #   homes_powered <dbl>
```

Combing three datasets

```
all_trash_wheels = bind_rows(
  mr_trash_df |>  mutate(trash_wheel_name = "Mr. Trash Wheel"),
  prof_trash_df |>  mutate(trash_wheel_name = "Professor Trash Wheel"),
  gwynnda_trash_df |>  mutate(trash_wheel_name = "Gwynnda Trash Wheel")
) |>
  select(trash_wheel_name, everything())
all_trash_wheels
```

```
## # A tibble: 845 x 15
##    trash_wheel_name dumpster month year  date                weight_tons
##    <chr>               <dbl> <chr> <chr> <dttm>                    <dbl>
##  1 Mr. Trash Wheel         1 May   2014  2014-05-16 00:00:00        4.31
##  2 Mr. Trash Wheel         2 May   2014  2014-05-16 00:00:00        2.74
##  3 Mr. Trash Wheel         3 May   2014  2014-05-16 00:00:00        3.45
##  4 Mr. Trash Wheel         4 May   2014  2014-05-17 00:00:00        3.1
##  5 Mr. Trash Wheel         5 May   2014  2014-05-17 00:00:00        4.06
##  6 Mr. Trash Wheel         6 May   2014  2014-05-20 00:00:00        2.71
##  7 Mr. Trash Wheel         7 May   2014  2014-05-21 00:00:00        1.91
##  8 Mr. Trash Wheel         8 May   2014  2014-05-28 00:00:00        3.7
##  9 Mr. Trash Wheel         9 June  2014  2014-06-05 00:00:00        2.52
## 10 Mr. Trash Wheel        10 June  2014  2014-06-11 00:00:00        3.76
## # i 835 more rows
## # 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>
```

Trash Wheel Data Summary

This combined dataset `all_trash_wheels` has 845 observations.

Key variables are listed as follows: The name of the trash wheel: `trash_wheel_name`, e.g. Mr. Trash Wheel.

The date:`date`, e.g.2014-05-16.

The weight of trash collected(tons): `weight_tons`,e.g. 4.31.

The volume of trash collected(cubic):`volume_cubic_yards`, e,g, 18.

The amount of plastic bottles it collected: `plastic_bottles`, e.g.1450.

The amount of polystyrene it collected: `polystyrene`, e.g. 1820.

The total weight of trash collected by Professor Trash Wheel is 216.26 tons.

The total number of cigarette butts collected by Gwynnda in June of 2022 is 1.812×10^4

problem 3

```
library(readr)
library(janitor)
library(dplyr)

bakers_df =
  read_csv("data/gbb_datasets/bakers.csv") |>
  janitor::clean_names() |>
  mutate(baker = word(baker_name, 1, sep = " "),
         series = as.numeric(series))
```

```
## Rows: 120 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Baker Name, Baker Occupation, Hometown
## dbl (2): Series, Baker Age
##
## 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.
```

```
bakers_df
```

```
## # A tibble: 120 x 6
##    baker_name       series baker_age baker_occupation             hometown baker
##    <chr>             <dbl>     <dbl> <chr>                        <chr>    <chr>
##  1 Ali Imdad             4        25 Charity worker               Saltley~ Ali
##  2 Alice Fevronia       10        28 Geography teacher            Essex    Alice
##  3 Alvin Magallanes      6        37 Nurse                        Brackne~ Alvin
##  4 Amelia LeBruin       10        24 Fashion designer             Halifax  Amel~
##  5 Andrew Smyth          7        25 Aerospace engineer           Derby /~ Andr~
##  6 Annetha Mills         1        30 Midwife                      Essex    Anne~
##  7 Antony Amourdoux      9        30 Banker                       London   Anto~
##  8 Beca Lyne-Pirkis      4        31 Military Wives' Choir Singer Aldersh~ Beca
##  9 Ben Frazer            2        31 Graphic Designer             Northam~ Ben
## 10 Benjamina Ebuehi      7        23 Teaching assistant           South L~ Benj~
## # i 110 more rows
```

```
bakes_df =
  read_csv("data/gbb_datasets/bakes.csv") |>
  janitor::clean_names() |>
  mutate(series = as.numeric(series),
         episode = as.numeric(episode))
```

```
## Rows: 548 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Baker, Signature Bake, Show Stopper
## dbl (2): Series, Episode
##
## 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.
```

```
bakes_df
```

```
## # A tibble: 548 x 5
##    series episode baker     signature_bake                          show_stopper
##     <dbl>   <dbl> <chr>     <chr>                                   <chr>
##  1      1       1 Annetha   "Light Jamaican Black Cakewith Strawbe~ Red, White ~
##  2      1       1 David     "Chocolate Orange Cake"                 Black Fores~
##  3      1       1 Edd       "Caramel Cinnamon and Banana Cake"      N/A
##  4      1       1 Jasminder "Fresh Mango and Passion Fruit Humming~ N/A
##  5      1       1 Jonathan  "Carrot Cake with Lime and Cream Chees~ Three Tiere~
##  6      1       1 Lea       "Cranberry and Pistachio Cakewith Oran~ Raspberries~
##  7      1       1 Louise    "Carrot and Orange Cake"                Never Fail ~
##  8      1       1 Mark      "Sticky Marmalade Tea Loaf"             Heart-shape~
##  9      1       1 Miranda   "Triple Layered Brownie Meringue Cake\~ Three Tiere~
## 10      1       1 Ruth      "Three Tiered Lemon Drizzle Cakewith F~ Classic Cho~
## # i 538 more rows
```

```
results_df =
  read_csv("data/gbb_datasets/results.csv", skip=2) |>
  janitor::clean_names() |>
  mutate(series = as.numeric(series),
         episode = as.numeric(episode))
```

```
## Rows: 1136 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): baker, result
## dbl (3): series, episode, technical
##
## 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.
```

```
results_df
```

```
## # A tibble: 1,136 x 5
##    series episode baker     technical result
```

```
##     <dbl>   <dbl> <chr>         <dbl> <chr>
##  1      1       1 Annetha           2 IN
##  2      1       1 David             3 IN
##  3      1       1 Edd               1 IN
##  4      1       1 Jasminder        NA IN
##  5      1       1 Jonathan          9 IN
##  6      1       1 Louise           NA IN
##  7      1       1 Miranda           8 IN
##  8      1       1 Ruth             NA IN
##  9      1       1 Lea              10 OUT
## 10      1       1 Mark             NA OUT
## # i 1,126 more rows
```

create a single dataset

```
merged_df = bakes_df |>
  left_join(bakers_df, by = c("baker" = "baker", "series" = "series")) |>
  left_join(results_df, by = c("baker" = "baker", "series" = "series", "episode" = "episode")) |>
  select(-baker) |>
  select(baker_name,everything())

merged_df
```

```
## # A tibble: 548 x 10
##    baker_name         series episode signature_bake       show_stopper baker_age
##    <chr>               <dbl>   <dbl> <chr>                <chr>            <dbl>
##  1 Annetha Mills           1       1 "Light Jamaican Bla~ Red, White ~        30
##  2 David Chambers          1       1 "Chocolate Orange C~ Black Fores~        31
##  3 Edd Kimber              1       1 "Caramel Cinnamon a~ N/A                 24
##  4 Jasminder Randhawa      1       1 "Fresh Mango and Pa~ N/A                 45
##  5 Jonathan Shepherd       1       1 "Carrot Cake with L~ Three Tiere~        25
##  6 Lea Harris              1       1 "Cranberry and Pist~ Raspberries~        51
##  7 Louise Brimelow         1       1 "Carrot and Orange ~ Never Fail ~        44
##  8 Mark Whithers           1       1 "Sticky Marmalade T~ Heart-shape~        48
##  9 Miranda Browne          1       1 "Triple Layered Bro~ Three Tiere~        37
## 10 Ruth Clemens            1       1 "Three Tiered Lemon~ Classic Cho~        31
## # i 538 more rows
## # i 4 more variables: baker_occupation <chr>, hometown <chr>, technical <dbl>,
## #   result <chr>
```

export as csv

```
write_csv(merged_df, "data/gbb_datasets/merged_data.csv")
head(merged_df)
```

```
## # A tibble: 6 x 10
##   baker_name         series episode signature_bake        show_stopper baker_age
##   <chr>               <dbl>   <dbl> <chr>                 <chr>            <dbl>
## 1 Annetha Mills           1       1 Light Jamaican Black~ Red, White ~        30
## 2 David Chambers          1       1 Chocolate Orange Cake Black Fores~        31
## 3 Edd Kimber              1       1 Caramel Cinnamon and~ N/A                 24
## 4 Jasminder Randhawa      1       1 Fresh Mango and Pass~ N/A                 45
```

```
## 5 Jonathan Shepherd       1       1 Carrot Cake with Lim~ Three Tiere~        25
## 6 Lea Harris              1       1 Cranberry and Pistac~ Raspberries~        51
## # i 4 more variables: baker_occupation <chr>, hometown <chr>, technical <dbl>,
## #   result <chr>
```

Data cleaning process: I first import these three datasets and standardize their names. Because the `baker_name` in the `bakers_df` is the full name and the `baker` in other three datasets are the first name, so I generate a new variable also called `baker` in the `bakers_df` for convenience and future merge. And while import the `results_df`, there are two unnecessary lines before the data, so I skipped those two rows. For merging these three data frame, I used the variable `baker` and `series` as the benchmark and delete the `baker` because we already have `baker_name`. After merging, I put the `baker_name` in the first column for a clearer view.

The final dataset includes key information and variables such as their name by `baker_name`, the series and episodes they participated by `series` and `episodes` and the competition results by `result`. Additionally, there are also some main information of the baker such as their age by `baker_age`, their job by`baker_occupation` and their home town by `hometown`.

star baker or winner of each episode in Seasons 5 through 10.

```
star_bakers_and_winners = merged_df |>
  filter(result == "STAR BAKER" | result == "WINNER") |>
  filter(series >= 5 & series <= 10) |>
  select(baker_name, series, episode, result)
view(star_bakers_and_winners)
```

It's surprising that in series 5, although Nancy only became the STAR BAKER once but still won the competition.In series 6redictable overall winners? Any surprises? ## viewership Import, clean, tidy, and organize the viewership data in viewership.csv. Show the first 10 rows of this dataset. What was the average viewership in Season 1? In Season 5?

```
viewership =
  read_csv("data/gbb_datasets/viewers.csv") |>
  janitor::clean_names() |>
  mutate(episode = as.numeric(episode)) |>
  mutate(across(starts_with("series"), ~as.numeric(.)))
```

```
## Rows: 10 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): Episode, Series 1, Series 2, Series 3, Series 4, Series 5, Series ...
##
## 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.
```

```
head(viewership)
```

```
## # A tibble: 6 x 11
##   episode series_1 series_2 series_3 series_4 series_5 series_6 series_7
##     <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1       1     2.24     3.1      3.85     6.6      8.51     11.6     13.6
```

```
## 2       2     3        3.53     4.6      6.65     8.79     11.6     13.4
## 3       3     3        3.82     4.53     7.17     9.28     12.0     13.0
## 4       4     2.6      3.6      4.71     6.82    10.2      12.4     13.3
## 5       5     3.03     3.83     4.61     6.95     9.95     12.4     13.1
## 6       6     2.75     4.25     4.82     7.32    10.1      12       13.1
## # i 3 more variables: series_8 <dbl>, series_9 <dbl>, series_10 <dbl>
```

```
avg_viewership_s1 = mean(pull(viewership, `series_1`), na.rm = TRUE)
avg_viewership_s5 = mean(pull(viewership, `series_5`), na.rm = TRUE)

avg_viewership_s1
```

```
## [1] 2.77
```

```
avg_viewership_s5
```

```
## [1] 10.0393
```

The average viewership in Season 1 is 2.77, and is 10.04 in Season 5.