p8105_hw2_csc2233

Cameron Chesbrough

2024-09-27

Problem 2

```r
trash_df = read_excel("202409 Trash Wheel Collection Data.xlsx",
                      sheet = 1, cell_cols("A:N"))

trash_df = janitor::clean_names(trash_df) %>%
  drop_na(dumpster) %>%
  mutate(sports_balls = round(sports_balls)) %>%
  mutate(sports_balls = as.integer(sports_balls)) %>%
  mutate(year = as.integer(year))

name = rep("Mr_Trash_Wheel", nrow(trash_df))
Mr_trash_df = cbind(name, trash_df)

###

trash_df2 = read_excel("202409 Trash Wheel Collection Data.xlsx",
                      sheet = 2)

trash_df2 = janitor::clean_names(trash_df2) %>%
    drop_na(dumpster) %>%
    mutate(year = as.integer(year))

name = rep("Professor_Trash_Wheel", nrow(trash_df2))
Professor_trash_df = cbind(name, trash_df2)

###

trash_df3 = read_excel("202409 Trash Wheel Collection Data.xlsx",
                      sheet = 4)

trash_df3 = janitor::clean_names(trash_df3) %>%
    drop_na(dumpster)

name = rep("Gwynnda", nrow(trash_df3))
Gwynnda_df = cbind(name, trash_df3)

###

all_trash_df = full_join(Mr_trash_df, Professor_trash_df)
```

```
## Joining with ‘by = join_by(name, dumpster, month, year, date, weight_tons,
## volume_cubic_yards, plastic_bottles, polystyrene, cigarette_butts,
## glass_bottles, plastic_bags, wrappers, homes_powered)‘
```

```
final_trash_df = full_join(all_trash_df, Gwynnda_df)
```

```
## Joining with ‘by = join_by(name, dumpster, month, year, date, weight_tons,
## volume_cubic_yards, plastic_bottles, polystyrene, cigarette_butts,
## plastic_bags, wrappers, homes_powered)‘
```

```
rows = nrow(final_trash_df)
cols = ncol(final_trash_df)

prof_trash_total = filter(final_trash_df, name == "Professor_Trash_Wheel")
prof_trash_weightsum = sum(prof_trash_total$weight_tons, na.rm = TRUE)

gwy_total = filter(final_trash_df,
                   name == "Gwynnda", month == "June", year == "2022")
gwy_cigs = sum(gwy_total$cigarette_butts, na.rm = TRUE)
```

Writeup for 2

This dataset describes trash collecting boats and the garbage that they collect. There are 1033 rows and 15 columns. Each garbage dumpster filled by the boat is recorded and information on what garbage fills the dumpster is given. The total weight of trash collected by Professor Trash Wheel is 246.74 tons. The total number of cigarette butts collected by Gwynnda is 1.812×10^4.

Problem 3

```
bakers_df = read_csv(file = "./gbb_datasets/bakers.csv")
```

```
## 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.
```

```
bakes_df = read_csv(file = "./gbb_datasets/bakes.csv")
```

```
## 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.
```

```r
results_df = read_csv(file = "./gbb_datasets/results.csv", skip = 2)
```

```
## 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.
```

```r
bakers_df = janitor::clean_names(bakers_df)
bakes_df = janitor::clean_names(bakes_df)
results_df = janitor::clean_names(results_df) %>%
  drop_na(result)

firsts = separate(bakers_df, col = baker_name, c("first_Name","last_Name"))
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 10 rows [8, 20, 60, 76, 80,
## 90, 96, 102, 108, 110].
```

```r
bakes_df$baker = str_replace_all(bakes_df$baker, '"', "")
results_df$baker = str_replace_all(results_df$baker, 'Joanne', "Jo")

testing = full_join(results_df, bakes_df, by = c("episode" = "episode",
                                                  "series" = "series",
                                                  "baker" = "baker"))

final_merge = right_join(testing, firsts, by = c("baker" = "first_Name",
                                                 "series" = "series"))

final_merge = arrange(final_merge, series, baker, episode)

write.csv(final_merge, "./gbb_datasets/merged_gbb_data.csv")

bake_rows = nrow(final_merge)
bake_cols = ncol(final_merge)
```

Explaining Process

I began the data cleaning process by importing the three datasets so I could look through them. After doing so I identified that the all the datasets included the first name of the baker; I decided that I would center my cleaning to focus on getting that name column as the connecting piece between the three. Besides cleaning the column names, my first step was to separate the baker name in the bakers dataset, as that was a full name and the other datasets only used first names. Next, in the bakes dataset, Jo was entered as “Jo” (with parentheses) so those needed to be removed. Similarly, in the results dataset, Jo was entered as Joanne, I needed to replace that with Jo as well. I finally was able to join the datasets together, starting with results and bakes. I tried to use all columns held in common to join them as there were some potential areas for repeats. For example, there were multiple Toms in separate series. I chose to sort my final dataset to begin with series, then by baker, then by episode. I chose this because I thought it would be a helpful way to look through the dataset, going through each series and seeing the results of each baker in the order of episodes.

This completed dataset describes the show Great British Bake Off (GBB). It gives information on the seasons, episodes, bakers, foods, and results. The dataset has 684 rows and 11 columns.

```r
later_winners = filter(final_merge, result %in% c("WINNER", "STAR BAKER")) %>%
  filter(series > 4) %>%
  select(series, episode, baker, result, baker_age, baker_occupation) %>%
  knitr::kable(col.names = c('Series', 'Episode', 'Baker', 'Result',
                             'Baker Age', 'Baker Occupation'),
               align = "cccccc")

later_winners
```

Series	Episode	Baker	Result	Baker Age	Baker Occupation
5	6	Chetna	STAR BAKER	35	Fashion Designer
5	5	Kate	STAR BAKER	41	Furniture Restorer
5	3	Luis	STAR BAKER	42	Graphic Designer
5	1	Nancy	STAR BAKER	60	Retired Practice Manager
5	10	Nancy	WINNER	60	Retired Practice Manager
5	2	Richard	STAR BAKER	38	Builder
5	4	Richard	STAR BAKER	38	Builder
5	7	Richard	STAR BAKER	38	Builder
5	8	Richard	STAR BAKER	38	Builder
5	9	Richard	STAR BAKER	38	Builder
6	2	Ian	STAR BAKER	41	Travel photographer
6	3	Ian	STAR BAKER	41	Travel photographer
6	4	Ian	STAR BAKER	41	Travel photographer
6	1	Marie	STAR BAKER	66	Retired
6	6	Mat	STAR BAKER	37	Fire fighter
6	5	Nadiya	STAR BAKER	30	Full-time mother
6	8	Nadiya	STAR BAKER	30	Full-time mother

Series	Episode	Baker	Result	Baker Age	Baker Occupation
6	9	Nadiya	STAR BAKER	30	Full-time mother
6	10	Nadiya	WINNER	30	Full-time mother
6	7	Tamal	STAR BAKER	29	Trainee anaesthetist
7	7	Andrew	STAR BAKER	25	Aerospace engineer
7	9	Andrew	STAR BAKER	25	Aerospace engineer
7	4	Benjamina	STAR BAKER	23	Teaching assistant
7	2	Candice	STAR BAKER	31	PE teacher
7	5	Candice	STAR BAKER	31	PE teacher
7	8	Candice	STAR BAKER	31	PE teacher
7	10	Candice	WINNER	31	PE teacher
7	1	Jane	STAR BAKER	61	Garden designer
7	3	Tom	STAR BAKER	26	Project engagement manager
7	6	Tom	STAR BAKER	26	Project engagement manager
8	3	Julia	STAR BAKER	21	Aviation Broker
8	4	Kate	STAR BAKER	29	Health and safety inspector
8	6	Liam	STAR BAKER	19	Student
8	5	Sophie	STAR BAKER	33	Former army officer and trainee stuntwoman
8	9	Sophie	STAR BAKER	33	Former army officer and trainee stuntwoman
8	10	Sophie	WINNER	33	Former army officer and trainee stuntwoman
8	8	Stacey	STAR BAKER	42	Former school teacher
8	1	Steven	STAR BAKER	34	Marketer
8	2	Steven	STAR BAKER	34	Marketer
8	7	Steven	STAR BAKER	34	Marketer
9	6	Briony	STAR BAKER	33	Full-time parent
9	4	Dan	STAR BAKER	36	Full-time parent
9	1	Manon	STAR BAKER	26	Software project manager
9	2	Rahul	STAR BAKER	30	Research scientist

Series	Episode	Baker	Result	Baker Age	Baker Occupation
9	3	Rahul	STAR BAKER	30	Research scientist
9	10	Rahul	WINNER	30	Research scientist
9	8	Ruby	STAR BAKER	29	Project manager
9	9	Ruby	STAR BAKER	29	Project manager
10	2	Alice	STAR BAKER	28	Geography teacher
10	9	Alice	STAR BAKER	28	Geography teacher
10	10	David	WINNER	36	International health adviser
10	7	Henry	STAR BAKER	20	Student
10	3	Michael	STAR BAKER	26	Theatre manager/fitness instructor
10	1	Michelle	STAR BAKER	35	Print shop administrator
10	4	Steph	STAR BAKER	28	Shop assistant
10	5	Steph	STAR BAKER	28	Shop assistant
10	6	Steph	STAR BAKER	28	Shop assistant
10	8	Steph	STAR BAKER	28	Shop assistant

```r
viewers_df = read_csv(file = "./gbb_datasets/viewers.csv")
```

```
## 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.
```

```r
viewers_df = janitor::clean_names(viewers_df)
head(viewers_df, n=10)
```

```
## # A tibble: 10 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
##  7       7    NA        4.42     5.1      7.76    10.3      12.4     13.4
##  8       8    NA        5.06     5.35     7.41     9.02     11.1     13.3
##  9       9    NA       NA        5.7      7.41    10.7      12.6     13.4
## 10      10    NA       NA        6.74     9.45    13.5      15.0     15.9
## # i 3 more variables: series_8 <dbl>, series_9 <dbl>, series_10 <dbl>
```

```
avg_view1 = mean(viewers_df$series_1, na.rm = TRUE)
avg_view5 = mean(viewers_df$series_5, na.rm = TRUE)
```

Table and Viewership

Looking at the table, it appears that most bakers that went on to win the competition won in at least one other round. Candice and Nadiya both won 3 other rounds besides their overall win. David is the surprise here, as he was the big winner, but that was the only round that he won. Richard is the other surprise, as he won 5 separate rounds but did not win the competition.

Looking at the viewership, the average viewership in season 1 was 2.77 and the average viewership in season 5 was 10.0393.