AE 06: Importing and cleaning lottery data

Suggested answers

APPLICATION EXERCISE ANSWERS

MODIFIED
September 24, 2024

Packages

We will use the following four packages in this application exercise.

- **tidyverse**: For data import, wrangling, and visualization.
- **readxl:** For importing data from Excel.
- **janitor:** For cleaning column names.
- **scales:** For formatting **ggplot2** scales.

```
library(tidyverse)
library(readxl)
library(janitor)
library(scales)
```

Powerball

A **lottery** is form of gambling that involves the drawing of numbers at random for a prize.[1] In the United States, Powerball is a popular multi-state lottery played in 45 states, Washington D.C., Puerto Rico, and the US Virgin Islands. [2] The basic rules are (relatively) straightforward :

[1] Source: Wikipedia.

[2] Source: Powerball.com

- Powerball costs $2 per play.
- Players select five numbers between 1 and 69 for the white balls, then select one number between 1 and 26 for the red Powerball.
- Drawings are held every Monday, Wednesday, and Saturday night.
- The Powerball jackpot grows until it is won. Players win a prize by matching one of the 9 ways to win. The jackpot is won by matching all five white balls in any order and the red Powerball.[3]

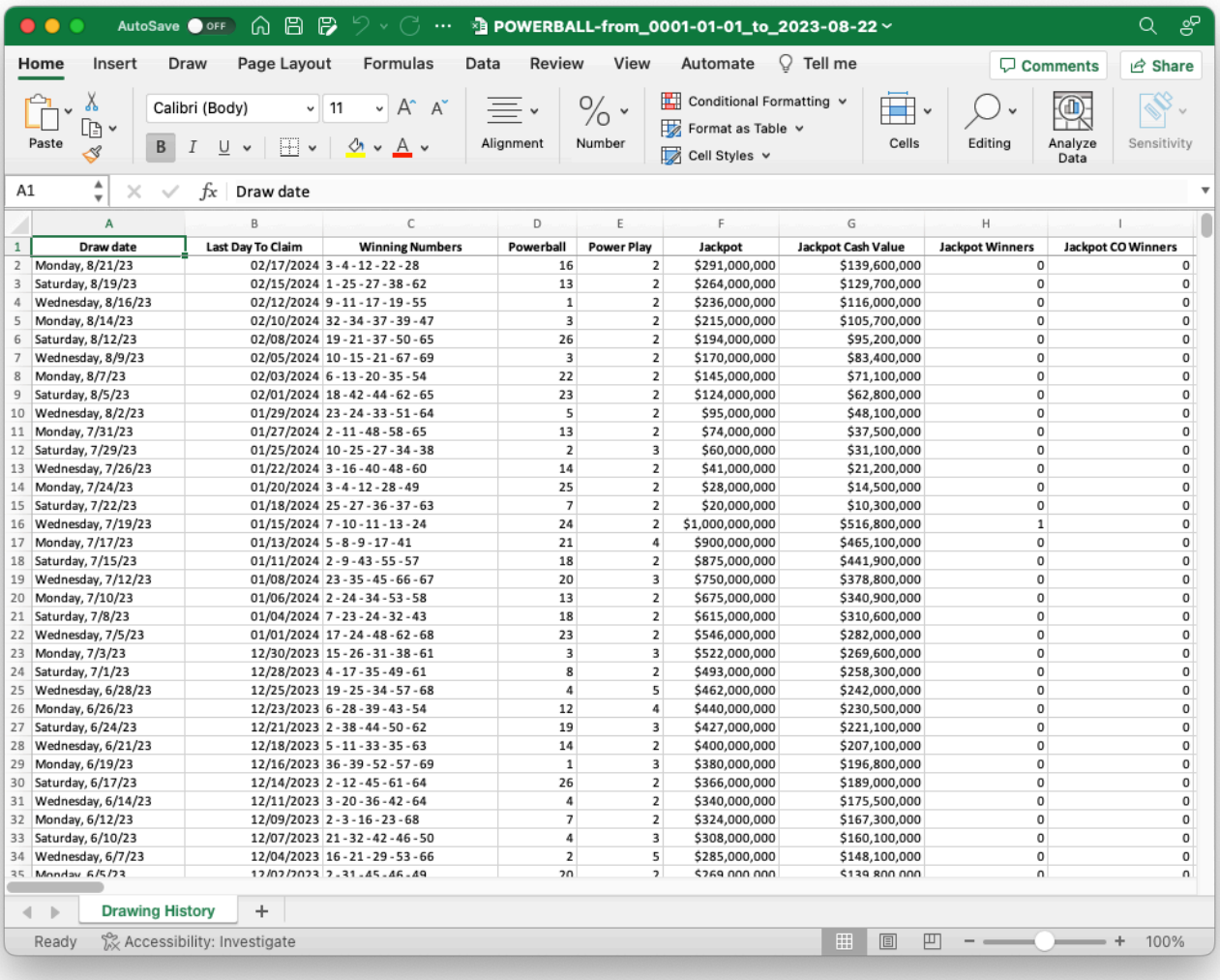
[3] For our purposes here, we will only examine the Powerball jackpot results.

[4] Drawing history can be obtained from their website.

The Colorado Lottery provides detailed information on Powerball drawings dating back to August 2001.[4] For these exercises we will work with a dataset containing every Powerball drawing in the Colorado Lottery's database.

Import and clean the data

The dataset is available for download as an Excel spreadsheet.



POWERBALL-from_0001-01-01_to_2023-08-22

Draw date	Last Day To Claim	Winning Numbers	Powerball	Power Play	Jackpot	Jackpot Cash Value	Jackpot Winners	Jackpot CO Winners
Monday, 8/21/23	02/17/2024	3-4-12-22-28	16	2	$291,000,000	$139,600,000	0	0
Saturday, 8/19/23	02/15/2024	1-25-27-38-62	13	2	$264,000,000	$129,700,000	0	0
Wednesday, 8/16/23	02/12/2024	9-11-17-19-55	1	2	$236,000,000	$116,000,000	0	0
Monday, 8/14/23	02/10/2024	32-34-37-39-47	3	2	$215,000,000	$105,700,000	0	0
Saturday, 8/12/23	02/08/2024	19-21-37-50-65	26	2	$194,000,000	$95,200,000	0	0
Wednesday, 8/9/23	02/05/2024	10-15-21-67-69	3	2	$170,000,000	$83,400,000	0	0
Monday, 8/7/23	02/03/2024	6-13-20-35-54	22	2	$145,000,000	$71,100,000	0	0
Saturday, 8/5/23	02/01/2024	18-42-44-62-65	23	2	$124,000,000	$62,800,000	0	0
Wednesday, 8/2/23	01/29/2024	23-24-33-51-64	5	2	$95,000,000	$48,100,000	0	0
Monday, 7/31/23	01/27/2024	2-11-48-58-65	13	2	$74,000,000	$37,500,000	0	0
Saturday, 7/29/23	01/25/2024	10-25-27-34-38	2	3	$60,000,000	$31,100,000	0	0
Wednesday, 7/26/23	01/22/2024	3-16-40-48-60	14	2	$41,000,000	$21,200,000	0	0
Monday, 7/24/23	01/20/2024	3-4-12-28-49	25	2	$28,000,000	$14,500,000	0	0
Saturday, 7/22/23	01/18/2024	25-27-36-37-63	7	2	$20,000,000	$10,300,000	0	0
Wednesday, 7/19/23	01/15/2024	7-10-11-13-24	24	2	$1,000,000,000	$516,800,000	1	0
Monday, 7/17/23	01/13/2024	5-8-9-17-41	21	4	$900,000,000	$465,100,000	0	0
Saturday, 7/15/23	01/11/2024	2-9-43-55-57	18	2	$875,000,000	$441,900,000	0	0
Wednesday, 7/12/23	01/08/2024	23-35-45-66-67	20	3	$750,000,000	$378,800,000	0	0
Monday, 7/10/23	01/06/2024	2-24-34-53-58	13	2	$675,000,000	$340,900,000	0	0
Saturday, 7/8/23	01/04/2024	7-23-24-32-43	18	2	$615,000,000	$310,600,000	0	0
Wednesday, 7/5/23	01/01/2024	17-24-48-62-68	23	2	$546,000,000	$282,000,000	0	0
Monday, 7/3/23	12/30/2023	15-26-31-38-61	3	3	$522,000,000	$269,600,000	0	0
Saturday, 7/1/23	12/28/2023	4-17-35-49-61	8	2	$493,000,000	$258,300,000	0	0
Wednesday, 6/28/23	12/25/2023	19-25-34-57-68	4	5	$462,000,000	$242,000,000	0	0
Monday, 6/26/23	12/23/2023	6-28-39-43-54	12	4	$440,000,000	$230,500,000	0	0
Saturday, 6/24/23	12/21/2023	2-38-44-50-62	19	3	$427,000,000	$221,100,000	0	0
Wednesday, 6/21/23	12/18/2023	5-11-33-35-63	14	2	$400,000,000	$207,100,000	0	0
Monday, 6/19/23	12/16/2023	36-39-52-57-69	1	3	$380,000,000	$196,800,000	0	0
Saturday, 6/17/23	12/14/2023	2-12-45-61-64	26	2	$366,000,000	$189,000,000	0	0
Wednesday, 6/14/23	12/11/2023	3-20-36-42-64	4	2	$340,000,000	$175,500,000	0	0
Monday, 6/12/23	12/09/2023	2-3-16-23-68	7	2	$324,000,000	$167,300,000	0	0
Saturday, 6/10/23	12/07/2023	21-32-42-46-50	4	3	$308,000,000	$160,100,000	0	0
Wednesday, 6/7/23	12/04/2023	16-21-29-53-66	2	5	$285,000,000	$148,100,000	0	0

Demo: Import the data file so it looks like below. Store it as `powerball_raw`.

```
# A tibble: 2,577 × 61
  `Draw date`     `Last Day To Claim` `Winning Numbers` Powerball `Power Play`
  <chr>           <dttm>              <chr>                 <dbl>        <dbl>
1 Monday, 9/23/… 2025-03-22 00:00:00 15 - 21 - 25 - 3…        19            3
2 Saturday, 9/2… 2025-03-20 00:00:00 17 - 19 - 21 - 3…        14            2
3 Wednesday, 9/… 2025-03-17 00:00:00 1 - 11 - 22 - 47…         7            4
4 Monday, 9/16/… 2025-03-15 00:00:00 8 - 9 - 11 - 27 …        17            5
5 Saturday, 9/1… 2025-03-13 00:00:00 29 - 34 - 38 - 4…        16            2
```

```
 6 Wednesday, 9/… 2025-03-10 00:00:00 10 - 12 - 55 - 6…         3             3
 7 Monday, 9/9/24 2025-03-08 00:00:00 1 - 16 - 21 - 47…         5             3
 8 Saturday, 9/7… 2025-03-06 00:00:00 14 - 34 - 37 - 5…        20             2
 9 Wednesday, 9/… 2025-03-03 00:00:00 7 - 10 - 21 - 33…        20             3
10 Monday, 9/2/24 2025-03-01 00:00:00 8 - 42 - 46 - 48…        22             3
# ℹ 2,567 more rows
# ℹ 56 more variables: Jackpot <dbl>, `Jackpot Cash Value` <dbl>,
#   `Jackpot Winners` <dbl>, `Jackpot CO Winners` <dbl>,
#   `Match 5 Prize` <dbl>, `Match 5 CO Winners` <dbl>,
#   `Match 5 Prize (with Power Play)` <dbl>,
#   `Match 5 CO Winners (with Power Play)` <dbl>,
#   `Match 4 + Powerball Prize` <dbl>, …
# ℹ Use `print(n = ...)` to see more rows
```

```
powerball_raw <- read_excel("data/POWERBALL-from_0001-01-01_to_2024-09-24.xlsx",
    col_types = c("text", "date", "text",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "text", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric"))
powerball_raw
```

```
# A tibble: 2,577 × 61
   `Draw date`      `Last Day To Claim` `Winning Numbers` Powerball `Power Play`
   <chr>            <dttm>              <chr>                 <dbl>        <dbl>
 1 Monday, 9/23/24  2025-03-22 00:00:00 15 - 21 - 25 - 3…        19            3
 2 Saturday, 9/21/… 2025-03-20 00:00:00 17 - 19 - 21 - 3…        14            2
 3 Wednesday, 9/18… 2025-03-17 00:00:00 1 - 11 - 22 - 47…         7            4
 4 Monday, 9/16/24  2025-03-15 00:00:00 8 - 9 - 11 - 27 …        17            5
 5 Saturday, 9/14/… 2025-03-13 00:00:00 29 - 34 - 38 - 4…        16            2
 6 Wednesday, 9/11… 2025-03-10 00:00:00 10 - 12 - 55 - 6…         3            3
 7 Monday, 9/9/24   2025-03-08 00:00:00 1 - 16 - 21 - 47…         5            3
 8 Saturday, 9/7/24 2025-03-06 00:00:00 14 - 34 - 37 - 5…        20            2
 9 Wednesday, 9/4/… 2025-03-03 00:00:00 7 - 10 - 21 - 33…        20            3
10 Monday, 9/2/24   2025-03-01 00:00:00 8 - 42 - 46 - 48…        22            3
```

```
# ℹ 2,567 more rows
# ℹ 56 more variables: Jackpot <dbl>, `Jackpot Cash Value` <dbl>,
#   `Jackpot Winners` <dbl>, `Jackpot CO Winners` <dbl>, `Match 5 Prize` <dbl>,
#   `Match 5 CO Winners` <dbl>, `Match 5 Prize (with Power Play)` <dbl>,
#   `Match 5 CO Winners (with Power Play)` <dbl>,
#   `Match 4 + Powerball Prize` <dbl>, `Match 4 + Powerball CO Winners` <dbl>,
#   `Match 4 + Powerball Prize (with Power Play)` <dbl>, …
```

Your turn: Clean the raw data to fix the following issues:

- Standardize the column names using `snake_case` format.
- Create columns with appropriate data types for the date of the drawing as well as the weekday. Append these columns to the beginning of the data frame.
- Our analysis focuses specifically on jackpot outcomes. Drop columns related to other prizes offered through the Powerball lottery (e.g. Match N, Double Play)

Store the cleaned data frame as `powerball`.

```
# standardize column names
powerball <- powerball_raw |>
  clean_names() |>
  # separate draw_date into two variables, clean both
  separate_wider_delim(
    cols = draw_date,
    delim = ",",
    names = c(NA, "draw_date")
  ) |>
  mutate(
    draw_date = mdy(draw_date),
    draw_weekday = wday(x = draw_date, label = TRUE),
    .before = last_day_to_claim
  ) |>
  # keep only a smaller subset to work with
  select(draw_date:jackpot_co_winners)
powerball
```

```
# A tibble: 2,577 × 10
   draw_date  draw_weekday last_day_to_claim   winning_numbers        powerball
   <date>     <ord>        <dttm>              <chr>                      <dbl>
 1 2024-09-23 Mon          2025-03-22 00:00:00 15 - 21 - 25 - 37 - 45        19
 2 2024-09-21 Sat          2025-03-20 00:00:00 17 - 19 - 21 - 37 - 45        14
 3 2024-09-18 Wed          2025-03-17 00:00:00 1 - 11 - 22 - 47 - 68          7
 4 2024-09-16 Mon          2025-03-15 00:00:00 8 - 9 - 11 - 27 - 31          17
 5 2024-09-14 Sat          2025-03-13 00:00:00 29 - 34 - 38 - 48 - 56        16
 6 2024-09-11 Wed          2025-03-10 00:00:00 10 - 12 - 55 - 65 - 67         3
 7 2024-09-09 Mon          2025-03-08 00:00:00 1 - 16 - 21 - 47 - 60          5
 8 2024-09-07 Sat          2025-03-06 00:00:00 14 - 34 - 37 - 55 - 63        20
 9 2024-09-04 Wed          2025-03-03 00:00:00 7 - 10 - 21 - 33 - 59         20
10 2024-09-02 Mon          2025-03-01 00:00:00 8 - 42 - 46 - 48 - 53         22
# ℹ 2,567 more rows
```

```
# ℹ 5 more variables: power_play <dbl>, jackpot <dbl>,
#   jackpot_cash_value <dbl>, jackpot_winners <dbl>, jackpot_co_winners <dbl>
```

Why does it seem like everyone is winning big?

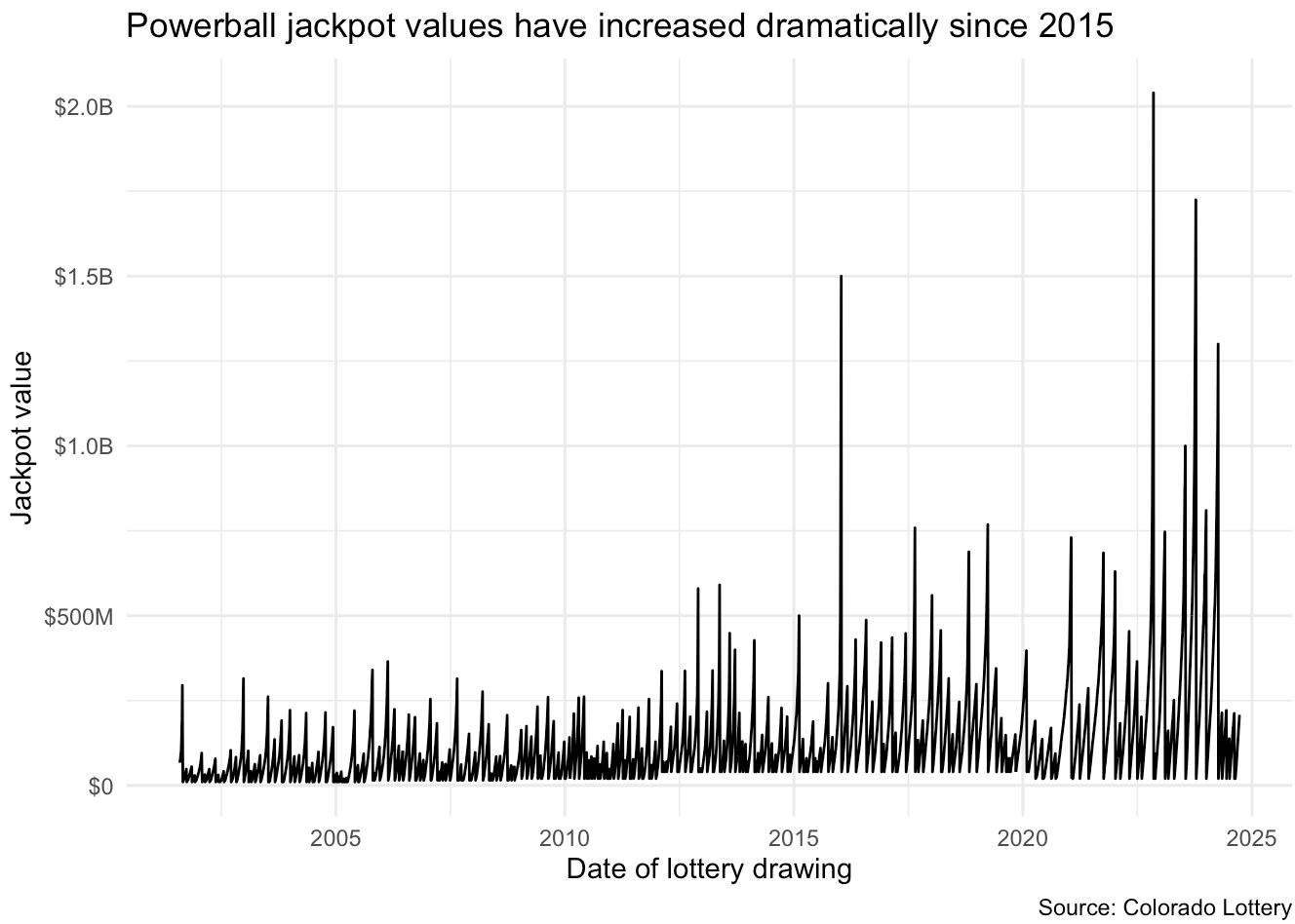
Anyone living in the United States in the past few years is likely to have seen news reports whenever the jackpot grows dramatically. The 10 biggest lottery jackpots in the United States have all occurred since 2015. What is driving this trend?

How the jackpot value has changed over time

In order to address this question, let's start first with a simpler question: **how has the jackpot value changed over time?** The jackpot amount varies for each drawing depending on the number of tickets sold as well as if the jackpot is rolling over from the previous drawing.

Demo: Create a line graph visualizing the jackpot value for every Powerball drawing over time.

```
ggplot(data = powerball, mapping = aes(x = draw_date, y = jackpot)) +
  geom_line() +
  scale_y_continuous(labels = label_currency(scale_cut = cut_short_scale())) +
  theme_minimal() +
  labs(
    x = "Date of lottery drawing",
    y = "Jackpot value",
    title = "Powerball jackpot values have increased dramatically since 2015",
    caption = "Source: Colorado Lottery"
  )
```



Your turn: What do you observe from the graph? *Add response here.*

There definitely seems to be an increase in the typical Powerball jackpot values since 2015.

Distribution of winning numbers

To investigate this further, let's look at the distribution of the white balls + the red Powerball. Presumably since the numbers are drawn at random, then they should be uniformly distributed.

Your turn: Convert `winning_numbers` into numeric values with one row for each drawing for each number. Keep just two columns: the drawing date and the winning numbers. Store this as `powerball_white`.

```
powerball_white <- powerball |>
  separate_longer_delim(
    cols = winning_numbers,
    delim = " - "
  ) |>
  mutate(winning_numbers = parse_number(x = winning_numbers)) |>
  select(draw_date, number = winning_numbers)
powerball_white
```

```
# A tibble: 12,885 × 2
   draw_date  number
   <date>      <dbl>
 1 2024-09-23     15
 2 2024-09-23     21
 3 2024-09-23     25
 4 2024-09-23     37
 5 2024-09-23     45
 6 2024-09-21     17
 7 2024-09-21     19
 8 2024-09-21     21
 9 2024-09-21     37
10 2024-09-21     45
# ℹ 12,875 more rows
```

Your turn: Create a similarly structured data frame for the red Powerball called `powerball_red`. Ensure it has the same column names as `powerball_white`.

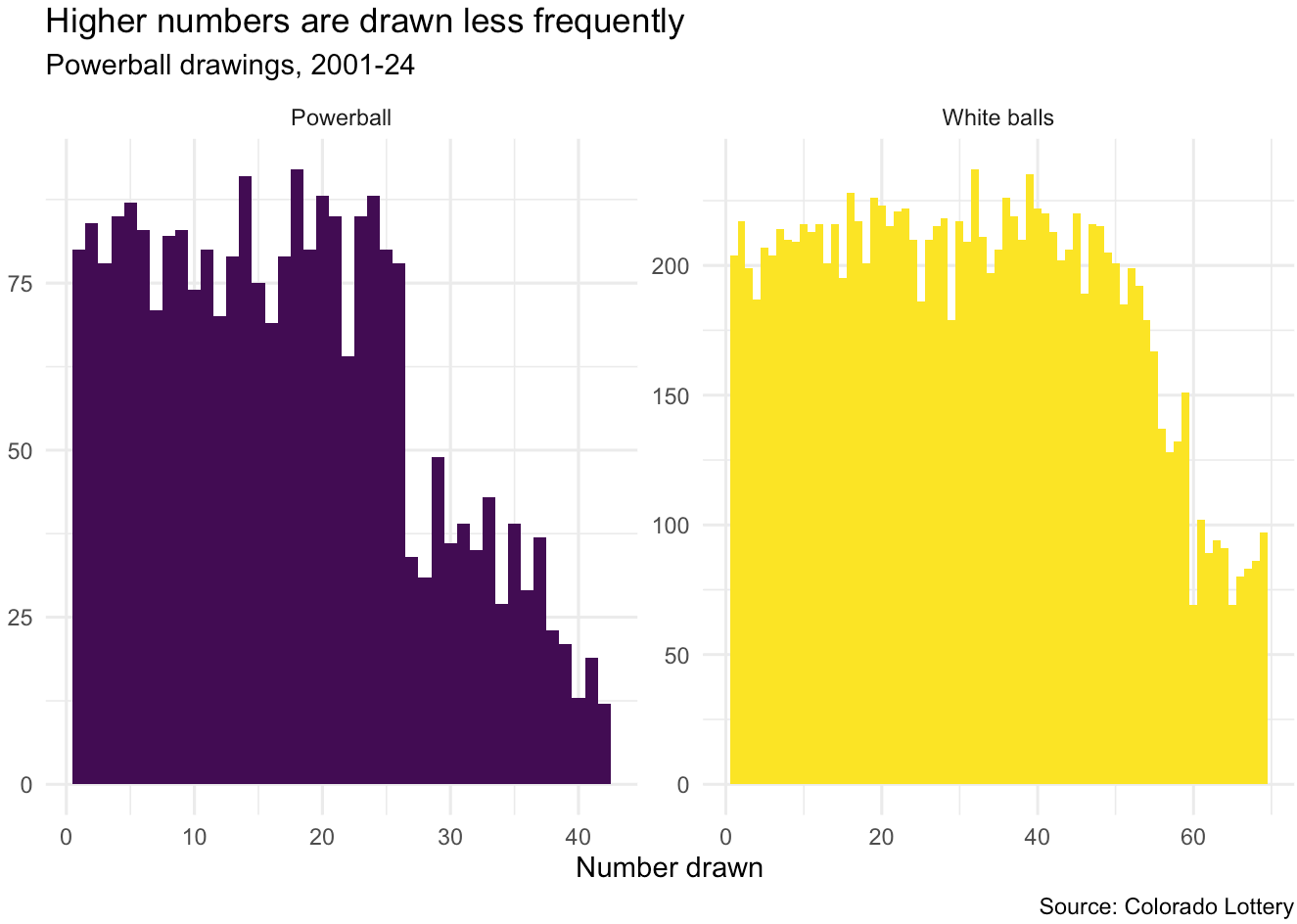
```
powerball_red <- powerball |>
  select(draw_date, number = powerball)
powerball_red
```

```
# A tibble: 2,577 × 2
   draw_date  number
   <date>      <dbl>
 1 2024-09-23     19
 2 2024-09-21     14
 3 2024-09-18      7
 4 2024-09-16     17
 5 2024-09-14     16
 6 2024-09-11      3
 7 2024-09-09      5
 8 2024-09-07     20
 9 2024-09-04     20
10 2024-09-02     22
# ℹ 2,567 more rows
```

Your turn: Combine the two data frames and create a histogram visualizing the distribution of the winning numbers, faceted between the white balls and the Red Powerballs.

```
bind_rows(
  `White balls` = powerball_white,
  Powerball = powerball_red,
  .id = "num_type"
) |>
  ggplot(mapping = aes(x = number, fill = num_type)) +
  geom_histogram(binwidth = 1) +
  scale_fill_viridis_d(guide = "none") +
  facet_wrap(facets = vars(num_type), scales = "free") +
  theme_minimal() +
```

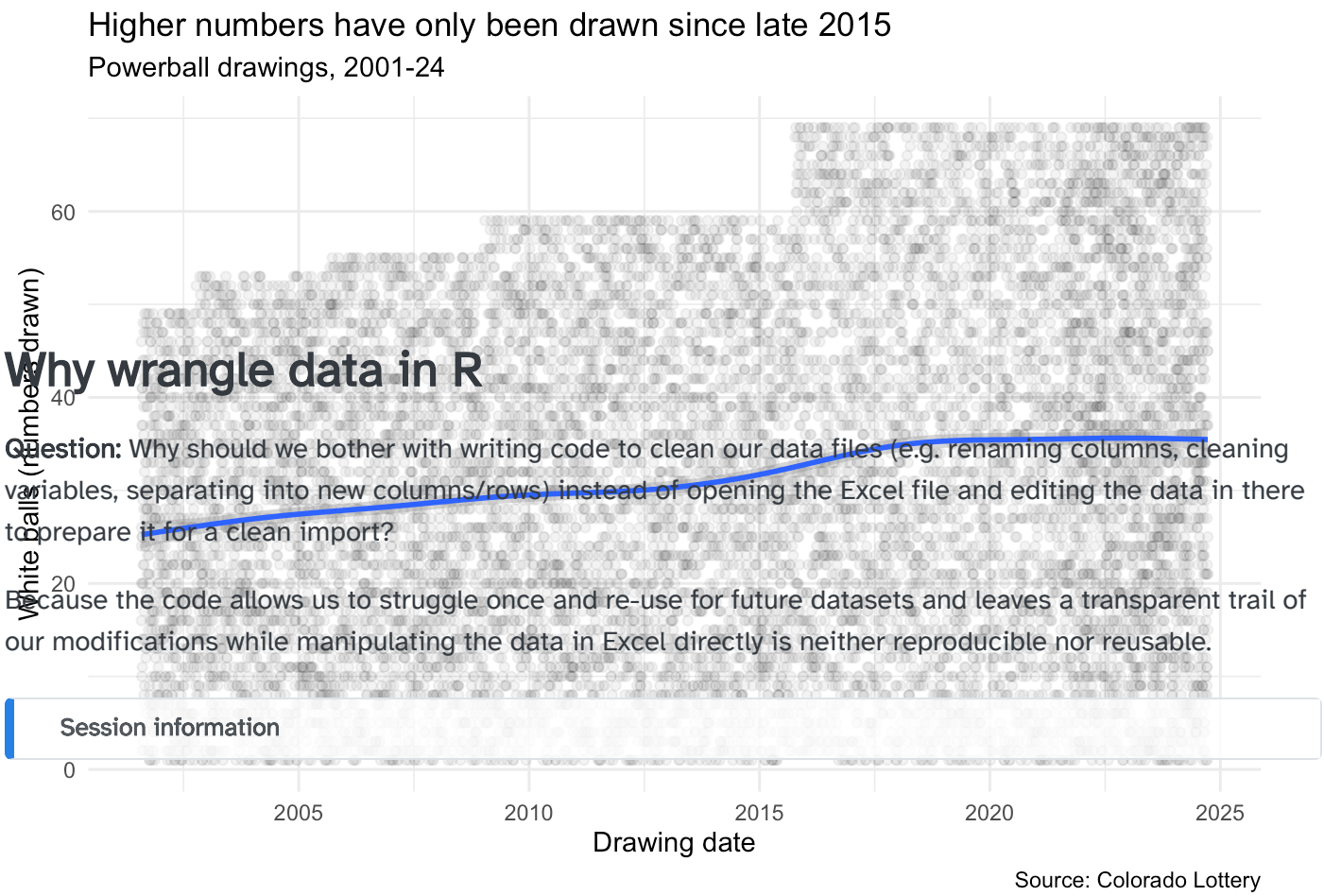
```
  labs(
    x = "Number drawn",
    y = NULL,
    title = "Higher numbers are drawn less frequently",
    subtitle = "Powerball drawings, 2001-24",
    caption = "Source: Colorado Lottery"
  )
```



Your turn: Visualize the distribution of white balls (numbers drawn) over time using a scatterplot + a smoothing line.

```
ggplot(data = powerball_white, mapping = aes(x = draw_date, y = number)) +
  geom_point(alpha = 0.06) +
  geom_smooth() +
  theme_minimal() +
  labs(
    x = "Drawing date",
    y = "White balls (numbers drawn)",
    title = "Higher numbers have only been drawn since late 2015",
    subtitle = "Powerball drawings, 2001-24",
    caption = "Source: Colorado Lottery"
  )
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Why wrangle data in R

Question: Why should we bother with writing code to clean our data files (e.g. renaming columns, cleaning variables, separating into new columns/rows) instead of opening the Excel file and editing the data in there to prepare it for a clean import?

Because the code allows us to struggle once and re-use for future datasets and leaves a transparent trail of our modifications while manipulating the data in Excel directly is neither reproducible nor reusable.

Session information

Turns out lottery officials have changed the rules of Powerball over time to change the odds of winning the jackpot. By increasing the pool of choices for the white balls, they double the available combinations of white balls and make it harder to win the jackpot. Higher jackpots spur more casual lottery players to purchase tickets, and overall generate more revenue.