PHW251 Problem Set 5

AUTHOR
Val Stacey

PUBLISHED
October 12, 2025

```
library(tibble)
library(dplyr)
library(tidyr)
library(here)
library(janitor)
library(kableExtra)
library(knitr)
library(readxl)
library(here)
library(readr)
library(ggplot2)
```

At this point in the course we have introduced a fair amount of code, which can be a lot to hold in our memory at once! Thankfully we we have search engines and these helpful cheatsheets. You may find the Base R and Data Transformation Cheatsheet helpful.

Part 1

Question 1

Use the readxl library and load two data sets from the “data/two_data_sheets.xlsx” file. There’s a parameter that you can specify which sheet to load. In this case, we have data about rat reaction time in sheet 1 and home visits in sheet 2.

```
rats <- read_xlsx(path = "data/two_data_sheets.xlsx", sheet = "Sheet1") %>%
         clean_names()
visits <- read_xlsx(path = "data/two_data_sheets.xlsx", sheet = "Sheet2") %>%
         clean_names()
```

Question 2

2A

For the rats data, pivot the data frame from wide to long format. We want the 1, 2, 3 columns, which represent the amount of cheese placed in a maze, to transform into a column called “cheese”. The values in the cheese column will be the time, which represents the amount of time the rat took to complete the maze.

```
rats_lng <- rats %>%
  pivot_longer(
         cols = c("x1", "x2", "x3"),
     names_to = "cheese",
    values_to = "time"
  )
```

2B

Please use the `head()` function to print the first few rows of your data frame.

```
head(rats_lng, 6)
```

```
# A tibble: 6 × 3
  subject cheese  time
  <chr>   <chr>  <dbl>
1 rat_101 x1      14.4
2 rat_101 x2       9.01
3 rat_101 x3       8.20
4 rat_102 x1      11.7
5 rat_102 x2       8.59
6 rat_102 x3       8.49
```

Question 3

Use `summarize()` to compute the mean and standard deviation of the maze time depending on the amount of cheese in the maze.

```
rats_sum <- rats_lng %>%
  group_by(cheese) %>%
  summarise(
    mean_time = mean(time, na.rm = TRUE),
      st_time = sd(time, na.rm = TRUE),
      .groups = "drop") %>%
  arrange(cheese)

rats_sum
```

```
# A tibble: 3 × 3
  cheese mean_time st_time
  <chr>      <dbl>   <dbl>
1 x1         12.8    1.43
2 x2          9.88   0.904
3 x3          8.51   0.279
```

Question 4

The home visits data is a record of how and where some interviews were conducted.

4A

Pivot the home visits data frame from long to wide. We want the names from the action column to become unique columns and the values to represent the counts.

```
visits_wide <- visits %>%
  pivot_wider(
        id_cols = c("location", "year"),
     names_from = "action",
    values_from = "count",
    values_fill = NA_integer_
  )
```

4B

Please print the whole resulting dataframe.

Pivoted Home Visits Dataframe

Location	Year	Interviews	Home Visits	Questionnaires
Washington DC	2015	103	76	200
Washington DC	2016	71	43	168
Washington DC	2017	45	60	90
St Louis	2015	90	86	210
St Louis	2016	95	82	175
St Louis	2017	78	71	106
Tucson	2015	130	98	303
Tucson	2016	120	88	280
Tucson	2017	78	65	230

Part 2

For this part we will use data from New York City that tested children under 6 years old for elevated blood lead levels (BLL). You can read more about the data on their website.

About the data:

All NYC children are required to be tested for lead poisoning at around age 1 and age 2, and to be screened for risk of lead poisoning, and tested if at risk, up until age 6. These data are an indicator of children younger that 6 years of age tested in NYC in a given year with blood lead levels (BLL) greater than 5 mcg/dL. In 2012, CDC established that a blood lead level of 5 mcg/dL is the reference level for exposure to lead in children. This level is used to identify children who have blood lead levels higher than most children's levels. The reference level is determined by measuring the NHANES blood lead distribution in US children ages 1 to 5 years, and is reviewed every 4 years.

Question 5

In this question you will recreate the below table with the "kable" pacakge. Please make sure you follow all of the steps outlined in parts A though D.

BLL Rates per 1,000 tested in New York City, 2015-2016

Borough	Year	BLL >5 µg/dL	BLL >10 µg/dL	BLL >15 µg/dL
Bronx	2015	15.7	2.5	1.0
Bronx	2016	15.0	2.8	1.2
Brooklyn	2015	22.6	3.9	1.3
Brooklyn	2016	22.3	3.6	1.2
Manhattan	2015	10.6	1.6	0.5
Manhattan	2016	8.1	1.3	0.6
Queens	2015	15.4	2.7	1.0
Queens	2016	14.3	2.3	0.9
Staten Island	2015	12.0	2.0	0.7
Staten Island	2016	14.8	2.7	0.8

q1_tbl

You will need to calculate the BLL per 1,000, filter for years 2015-2016, and rename the boroughs based on the following coding scheme:

- 1: Bronx
- 2: Brooklyn
- 3: Manhattan
- 4: Queens
- 5: Staten Island

5A

First, filter your dataframe for the years 2015-2016 and rename the boroughs. If you make your borough names a factor, it will make your life easier when we create tables and graphs.

```
nyc_A <- bll_nyc %>%
  filter(between(time_period, 2015, 2016)) %>%
  mutate(
    borough = factor(borough_id, levels = 1:5,
       labels = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"),
      ordered = TRUE)
  ) %>%
  select(-borough_id) %>%
  relocate(borough, .before = "time_period")
```

5B

Second, group and summarize the data to calculate the total *number* of children in each borough in each year that were tested and the number with blood lead levels that were greater than 5 mcg/dL, 10 mcg/dL, or 15 mcg/dL.

```
nyc_B <- nyc_A %>%
  group_by(borough, time_period) %>%
  summarise(
    total_tested = first(total_tested),
    gr_5 = sum(bll_5),
    gr_10 = sum(bll_10),
    gr_15 = sum(bll_15),
    .groups = "drop"
  )
```

5C

Third, calculate the rate at which each blood lead level occurred in each year in each borough (BLL per 1,000).

```
nyc_C <- nyc_B %>%
  group_by(borough, time_period) %>%
  mutate(
    gr_5_rt = round((1000 * gr_5/total_tested),1),
    gr_10_rt = round((1000 * gr_10/total_tested),1),
    gr_15_rt = round((1000 * gr_15/total_tested),1)
  )
```

5D

Now we have calculated all the numbers we need to recreate the table shown at the beginning of this question. Use `kable()` to produce your table.

```
nyc_D <- nyc_C %>%
  select(-c(total_tested, gr_5, gr_10, gr_15)) %>%
  arrange(borough, time_period)

bll_tble <- nyc_D %>%
  kbl(align = "lcccc",
      col.names = c("Borough", "Year", "BLL ≥5 µg/dL", "BLL ≥10 µg/dL", "BLL
         ≥15 µg/dL"),
      caption = "<h4 style='text-align:center; font-weight:bold;color:#000'>
                  BLL Rates per 1,000 Tested in New York City, 2015–2016</h4>",
    escape = FALSE) %>%
  column_spec(1, bold = TRUE) %>%
  kable_styling(full_width = TRUE,
    bootstrap_options = c("striped", "hover", "condensed"))

bll_tble
```

BLL Rates per 1,000 Tested in New York City, 2015–2016

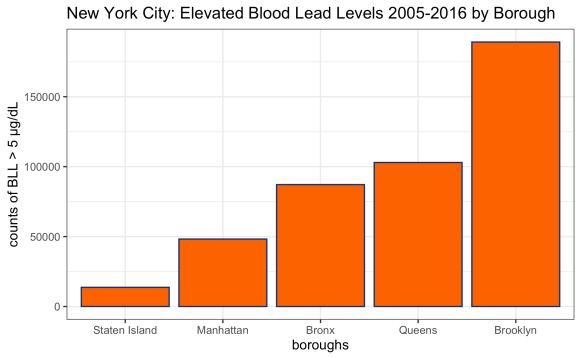
Borough	Year	BLL ≥5 µg/dL	BLL ≥10 µg/dL	BLL ≥15 µg/dL
Bronx	2015	31.4	5.0	2.0
Bronx	2016	29.9	5.5	2.4
Brooklyn	2015	45.1	7.8	2.6
Brooklyn	2016	44.6	7.2	2.4
Manhattan	2015	21.0	3.1	1.0
Manhattan	2016	15.8	2.6	1.2
Queens	2015	30.7	5.5	1.9
Queens	2016	28.5	4.6	1.7
Staten Island	2015	23.9	3.9	1.3
Staten Island	2016	29.8	5.4	1.6

Question 6

In this question you will replicate the following bar chart. Since we want the graph to have an ascending order, we will need to factor borough_id with the levels in a different order than the default. Note that this graph covers the whole time period from the original dataset!

Here are the HEX codes used for the colors:

- #ff6600: orange
- #003884: blue



q2_bar

6A

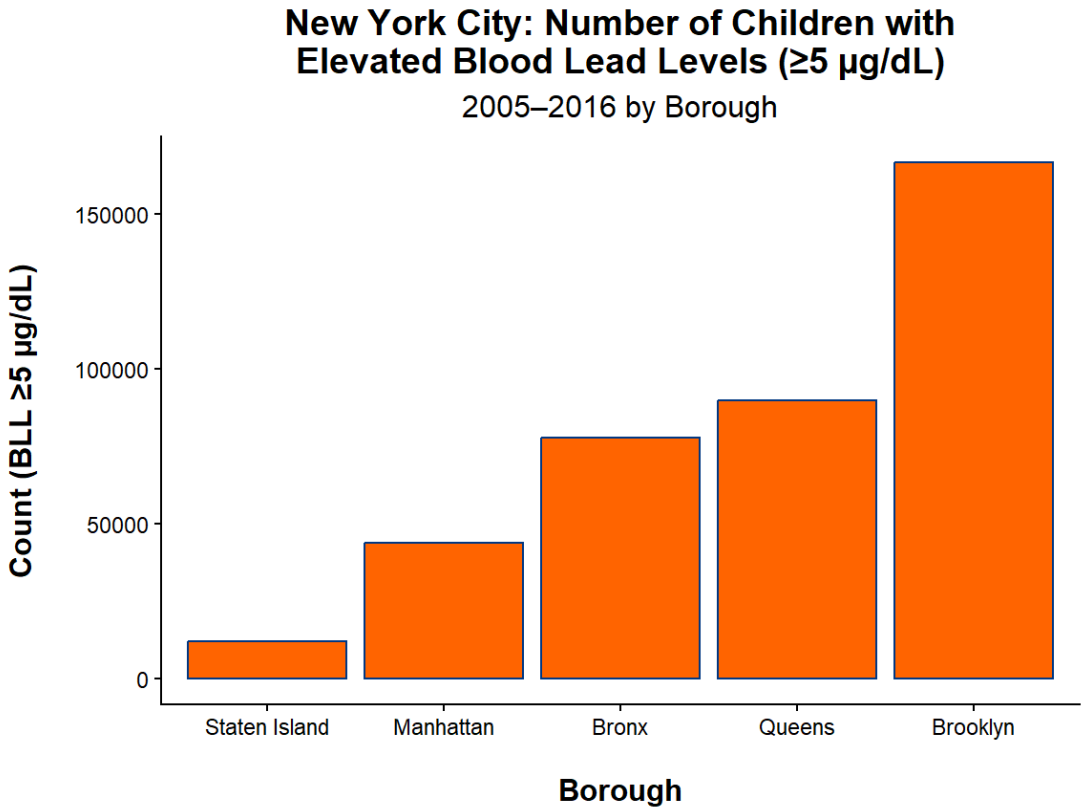
First, summarize the original dataset, counting only the number of kids with BLL > 5mcg/dL. Don't forget to change the order of factor borough_id in accordance with the graph above!

```
bar_df <- bll_nyc %>%
  group_by(borough_id) %>%
  summarise(
       gr_5 = sum(bll_5),
    .groups = "drop") %>%
  mutate(
   borough = factor(borough_id, levels = 1:5,
       labels = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"),
      ordered = TRUE)
  ) %>%
  select(-borough_id) %>%
  relocate(borough, .before = "gr_5") %>%
  arrange(gr_5) %>%
  mutate(borough = factor(borough, levels = borough))
```

6B

Then make the graph!

```
p_6 <- ggplot(bar_df) +
  geom_col(
    aes(x = borough, y = gr_5),
    fill = "#ff6600",
    color = "#003884"
  ) +
  theme_classic() +
  labs(
  title = "New York City: Number of Children with\nElevated Blood Lead Levels
         (≥5 µg/dL)",
    subtitle = "2005–2016 by Borough",
    x = "Borough",
    y = "Count (BLL ≥5 µg/dL)"
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    axis.title.x = element_text(size = 12, face = "bold", margin =
         margin(t=15)),
    axis.title.y = element_text(size = 12, face = "bold", margin =
         margin(r=15))
  )
```


New York City: Number of Children with Elevated Blood Lead Levels (≥5 µg/dL)
2005–2016 by Borough
Count (BLL ≥5 µg/dL)
150000
100000
50000
0
Staten Island
Manhattan
Bronx
Queens
Brooklyn
Borough