PHW251 Problem Set 5

Clara Voong

10/14/2024

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 "two_data_sheets" 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.

```r
library(readxl)

rat_dat <- read_xlsx(
  "~/PHW251_2024/problem_sets/problem_set_5/data/two_data_sheets.xlsx",
  sheet= 1)
home_visit_dat <-
  read_xlsx("~/PHW251_2024/problem_sets/problem_set_5/data/two_data_sheets.xlsx",
  sheet= 2)
```

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.

```
rat_dat <- pivot_longer(data=rat_dat,
cols=c("1", "2", "3"),
names_to="cheese",
values_to = "time")
```

2B Please use the `head()` function to print the first few rows of your data frame.

```
head(rat_dat)
```

```
## # A tibble: 6 x 3
##   subject cheese  time
##   <chr>   <chr>  <dbl>
## 1 rat_101 1      14.4
## 2 rat_101 2       9.01
## 3 rat_101 3       8.20
## 4 rat_102 1      11.7
## 5 rat_102 2       8.59
## 6 rat_102 3       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.

```
rat_dat %>%
  group_by(cheese) %>%
  summarize(mean_time = mean(time, na.rm = TRUE),
            sd_time = sd(time, na.rm = TRUE)) %>% ungroup()
```

```
## # A tibble: 3 x 3
##   cheese mean_time sd_time
##   <chr>      <dbl>   <dbl>
## 1 1          12.8    1.43
## 2 2           9.88   0.904
## 3 3           8.51   0.279
```

Question 3

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

2A 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.

```
home_visit_dat <- home_visit_dat %>% pivot_wider(
  names_from = action,
  values_from = count
)
```

2B Please print the whole resulting dataframe.

```
print(home_visit_dat)
```

```
## # A tibble: 9 x 5
##   location       year interview `home visit` questionnaire
##   <chr>         <dbl>     <dbl>        <dbl>         <dbl>
## 1 Washington DC  2015       103           76           200
## 2 Washington DC  2016        71           43           168
## 3 Washington DC  2017        45           60            90
## 4 St Louis       2015        90           86           210
## 5 St Louis       2016        95           82           175
## 6 St Louis       2017        78           71           106
## 7 Tucson         2015       130           98           303
## 8 Tucson         2016       120           88           280
## 9 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]).

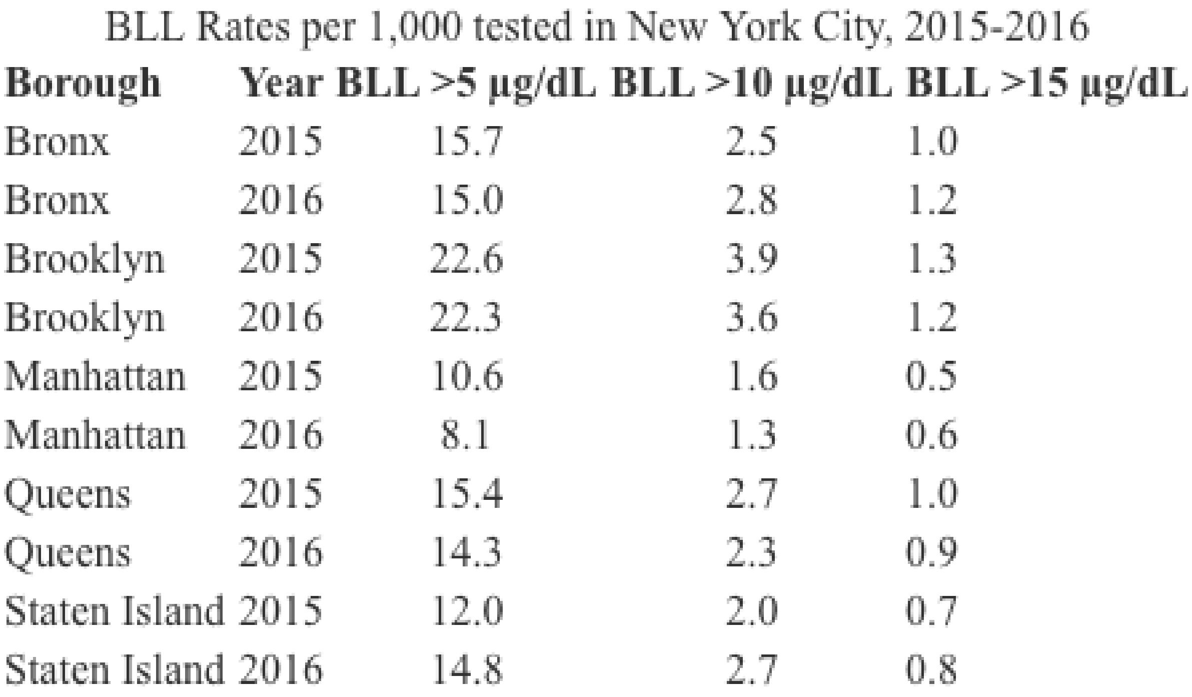
About the data:

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) of 5 mcg/dL or greater. 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 4

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.

```r
knitr::include_graphics('data/question_1_table.png')
```



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

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

4A 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.

```
bll_nyc_2015_16 <- bll_nyc %>% filter(time_period == 2015 | time_period == 2016) %>%
mutate(
  borough_id =
    factor(borough_id, labels = c(
"Bronx",
"Brooklyn",
"Manhattan",
"Queens",
"Staten Island"

)
)
)
```

4B 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 5 mcg/dL, and 15 5 mcg/dL.

```
total_bll <- bll_nyc_2015_16 %>% group_by(borough_id, time_period) %>% summarize(
  sum_tested = sum(total_tested, na.rm=T),
  sum_bll_5 = sum(bll_5, na.rm=T),
  sum_bll_10 = sum(bll_10, na.rm=T),
  sum_bll_15 = sum(bll_15, na.rm=T),
) %>%
  ungroup()

total_bll
```

```
## # A tibble: 10 x 6
##    borough_id    time_period sum_tested sum_bll_5 sum_bll_10 sum_bll_15
##    <fct>               <dbl>      <dbl>     <dbl>      <dbl>      <dbl>
##  1 Bronx                2015     123100      1937        310        122
##  2 Bronx                2016     117800      1763        324        142
##  3 Brooklyn             2015     217400      4911        846        284
##  4 Brooklyn             2016     207500      4627        752        244
##  5 Manhattan            2015      74000       787        118         38
##  6 Manhattan            2016      70400       567         92         44
##  7 Queens               2015     178900      2750        488        174
##  8 Queens               2016     174600      2490        406        150
##  9 Staten Island        2015      27400       328         54         18
## 10 Staten Island        2016      25900       384         70         20
```

4C Third, calculate the rate at which each blood lead level occurred in each year in each borough (BLL per 1,000).

```
rate_bll <- total_bll %>% mutate(
  rate_bll5 = (sum_bll_5/sum_tested)*1000,
  rate_bll10 = (sum_bll_10/sum_tested)*1000,
  rate_bll15 = (sum_bll_15/sum_tested)*1000
)

rate_bll
```

```
## # A tibble: 10 x 9
##    borough_id   time_period sum_tested sum_bll_5 sum_bll_10 sum_bll_15 rate_bll5
##    <fct>              <dbl>      <dbl>     <dbl>      <dbl>      <dbl>     <dbl>
##  1 Bronx               2015     123100      1937        310        122     15.7
##  2 Bronx               2016     117800      1763        324        142     15.0
##  3 Brooklyn            2015     217400      4911        846        284     22.6
##  4 Brooklyn            2016     207500      4627        752        244     22.3
##  5 Manhattan           2015      74000       787        118         38     10.6
##  6 Manhattan           2016      70400       567         92         44      8.05
##  7 Queens              2015     178900      2750        488        174     15.4
##  8 Queens              2016     174600      2490        406        150     14.3
##  9 Staten Isla~        2015      27400       328         54         18     12.0
## 10 Staten Isla~        2016      25900       384         70         20     14.8
## # i 2 more variables: rate_bll10 <dbl>, rate_bll15 <dbl>
```

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

Borough	Year	BLL > 5 ug/dL	BLL > 10 ug/dL	BLL > 15 ug/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

4D 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.

```r
kable(
  rate_bll[c(1,2,7,8,9)],
  digits = 1,
  col.names = c("Borough",
                "Year",
                "BLL > 5 ug/dL",
                "BLL > 10 ug/dL",
                "BLL > 15 ug/dL"),
  caption = "BLL Rates per 1,000 tested in New York City, 2015-2016",
  booktabs = T)
```

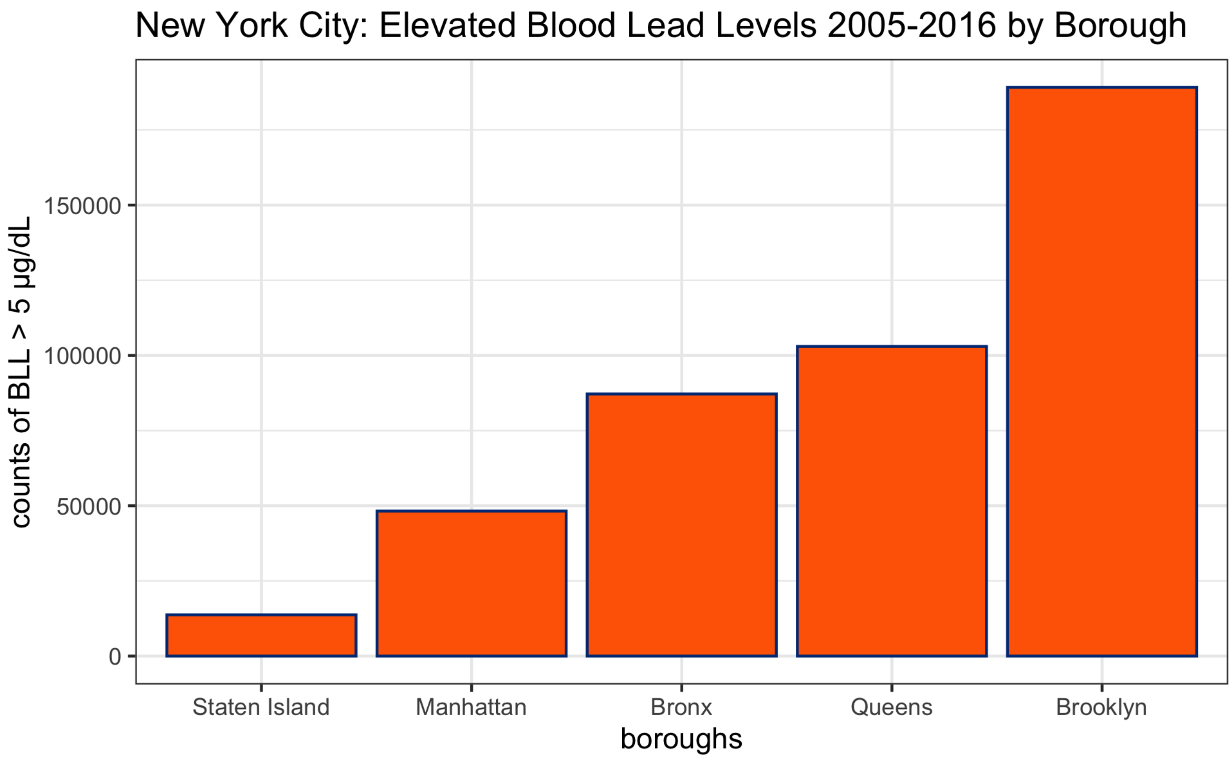
Question 5

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

```
knitr::include_graphics('data/question_2_bar.png')
```



5A First, summarize the original dataset.

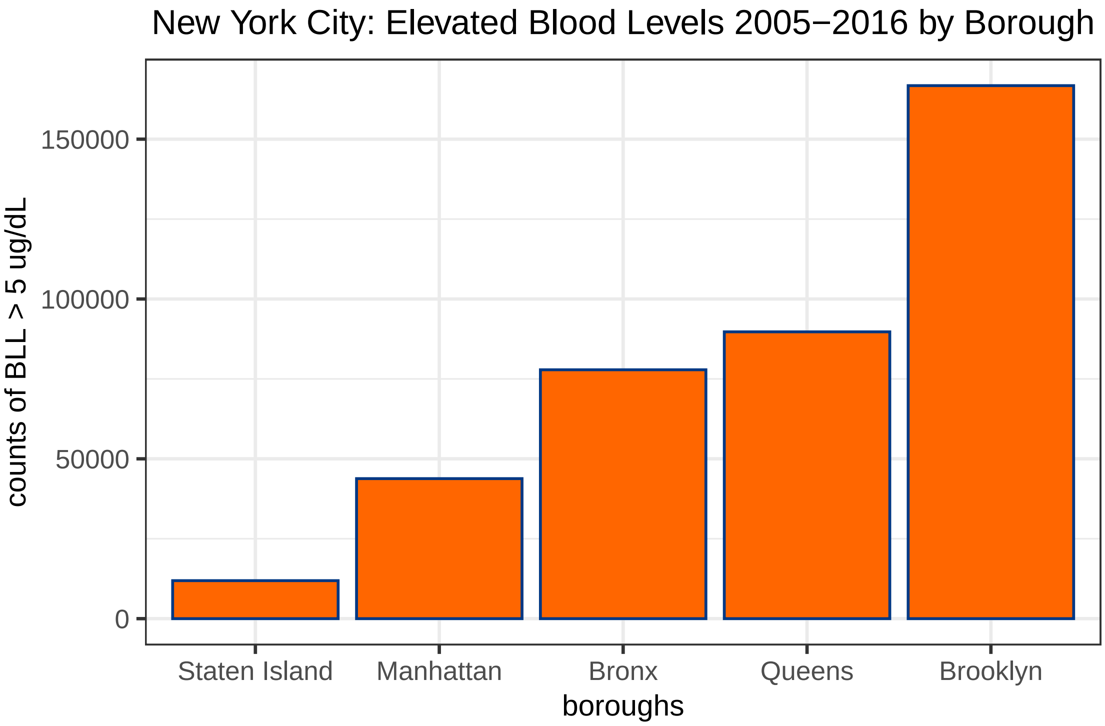
```
bll_nyc$borough_id <- factor(bll_nyc$borough_id, labels = c(
"Bronx",
"Brooklyn",
"Manhattan",
"Queens",
"Staten Island"
))


count_bll5 <- bll_nyc %>%
  group_by(borough_id) %>%
  summarize(
  sum_bll_5 = sum(bll_5, na.rm=T)
) %>%
  ungroup()
```

```
count_bll5
```

```
## # A tibble: 5 x 2
##   borough_id    sum_bll_5
##   <fct>             <dbl>
## 1 Bronx             77860
## 2 Brooklyn         166755
## 3 Manhattan         43804
## 4 Queens            89735
## 5 Staten Island     11886
```

5B Then make the graph!

```
ggplot(data = count_bll5,
       mapping=aes(x=reorder(borough_id, sum_bll_5), y=sum_bll_5)) +
  geom_bar(stat = "identity",
           color= "#003884",
           fill="#ff6600") +
  theme_minimal(base_size=13) +
  theme_bw(base_size=13) +
  theme(
plot.title = element_text(size = 13, hjust = 0.5),
    axis.title.x = element_text(size = 11),
    axis.title.y = element_text(size = 11),
plot.margin = unit(c(1, 1, 1, 1), "cm"),
) +
labs(
title=
"New York City: Elevated Blood Levels 2005-2016 by Borough",
x = "boroughs",
y= "counts of BLL > 5 ug/dL")
```



You're done! Please knit to pdf and upload to gradescope.