PHW251 Problem Set 5

your name here

today

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.

```
# your code here
```

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.

```
# your code here
```

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

```
# your code here
```

Question 3

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

```
# your code here
```

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.

```
# your code here
```

2B Please print the whole resulting dataframe.

```
# your code here
```

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.

```
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.

```
# your code here
```

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.

```
# your code here
```

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

```
# your code here
```

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.

```
# your code here
```

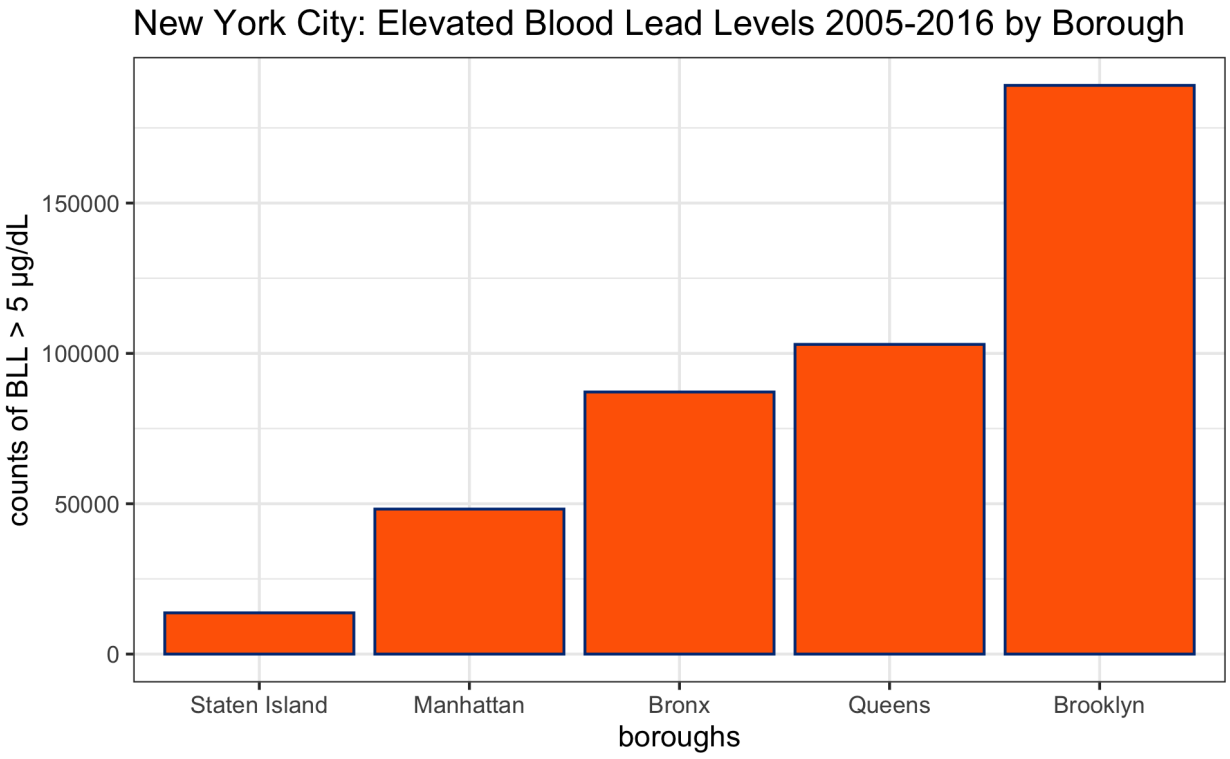
Question 5

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.

```
# your code here
```

5B Then make the graph!

```
# your code here
```

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