CMPSC 301
Data Analytics
Fall 2020

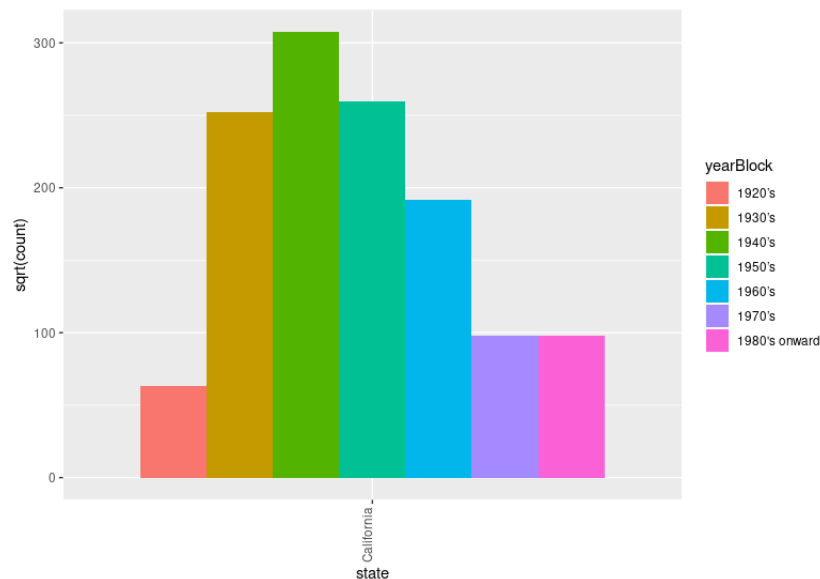**Lab 4: Vaccines Exploratory Analysis**
28th Sept 2020

Figure 1: A grouping the counts of infections by year. The code for this plot is provided in this lab.

Objectives

To enhance the understanding of the exploratory data analysis while practicing skills of data transformation. To investigate the issues of ethics, privilege and inequality surrounding vaccine refusal.

Reading Assignment

Please read Chapters 3 and 5 in the course book, corresponding to Chapters 5 and 7 in the website (online) version of the book. You may be required to look up the syntax of coding to prepare types of plots as you go through this lab.

GitHub Starter Link

https://classroom.github.com/a/8cvnY8QX

To use this link, please follow the steps below.

HANDED OUT: 2nd OCT 2020

- Click on the link and accept the assignment.

- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.

- Clone this repository (bearing your name) and work on the lab locally.

- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):

  - `git commit <nameOfFile> -m ''Your notes about commit here''`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:

  - `git add -A`
  - `git commit -m ''Your notes about commit here''`
  - `git push`

Groupwork Option: Each person submits own work to own repository

You are allowed to work in groups for this assignment. However, if you would rather work individually, then you may choose to work by yourself. Please keep group sizes to not more than four people for this lab. Be sure to discuss each of the questions and proceed after the group has come to a complete agreement. Each person is to turn in his or her own report and code, however all lab partners should be listed in the submission. **If you are working in a group, please be sure to include the names of your group members in you submissions documents.**

Exploratory Data Analysis On Vaccines

Vaccines have helped save millions of lives. In the 19th century, before herd immunization was achieved through vaccination programs, deaths from infectious diseases, like smallpox and polio, were common. However, today, despite all the scientific evidence for their importance, vaccination programs have become somewhat controversial.

The controversy started with a paper published in 1988 and lead by Andrew Wakefield claiming there was a link between the administration of the measles, mumps and rubella (MMR) vaccine, and the appearance of autism and bowel disease. Despite much science contradicting this finding, sensationalistic media reporting and fear mongering from conspiracy theorists, led parts of the public to believe that vaccines were harmful. Some parents stopped vaccinating their children as a result of wide-felt fear. However, the Center for Disease Control (CDC) estimated that vaccinations prevented more than 21 million hospitalizations and 732,000 deaths among children born in the last 20 years. For more information on this, please see *Benefits from Immunization during the Vaccines for Children Program Era United States, 1994-2013*, MMWR `https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6316a4.htm`.

Effective communication of data is a strong antidote to misinformation and fear mongering. In this lab you are going to prepare a report to have ready in case you need to help a family member, friend or acquaintance that is not aware of the positive impact vaccines have had for public health.

The data used for these plots were collected, organized and distributed by the Tycho Project (www.tycho.pitt.edu). They include weekly reported counts data for seven diseases from 1928 to 2011, from all fifty states. We include the yearly totals in the dslabs package:

Part 1. Steps to Follow

Use the below code to get started. Be sure to keep your code in a script source file called, src/vaccines.r. Record only working code. Please label your code for each step in your deliverable.

```
install.packages("dslabs")
library(dslabs)
library(tidyverse)
data(us_contagious_diseases)
```

1. Begin by installing the libraries dplyr and tidyverse in your R code if they are not already installed. Once it is installed, load the libraries into using library(dplyr) and library(tidyverse).

2. Use the command, data(us contagious diseases) to load the dataset.

   - 1. Describe this dataset; What types of data does it contain?
   - 2. Name two types of questions can you expect to answer using this dataset?

3. Create a dataset called dat by assigning it to the us contagious diseases dataset. You should now see this variable, dat, pop-up in your Global Environment section of rStudio.

4. Create another dataset from dat called dat measles rate in which;

   - The disease concerns only *Measles*. Hint: use filter()
   - A new column has been added called rate, using the equation below. Hint: use mutate().

$$rate = \frac{count*100000}{population} * \frac{WeeksReporting}{52}$$

   - 1. What is rate?
   - 2. In terms of its informational content, how could this column be useful in analysis?

5. Remove the two states (Alaska and Hawaii) from your dataset. For this, create a new variable called dat measles rate lessTwoStates that has been assigned to dat measles rate without the two states. During this assignment, use the filter() function to remove *Alaska* and *Hawaii*. Hint: you can filter out these states with similar code to the following;

```
    filter(dat_measles_rate, state != ''myState", state != ''myOtherState")
```

6. Prepare a plot of the `dat_measles_rate_lessTwoStates` dataset that relates the data of 48 states by editing the below code. Set your x and y variables to `year` and `rate`, respectively. Be sure to add a relevant title to the code where appropriate.

```
ggplot(data = dat_measles_rate_lessTwoStates,
  mapping = aes(x = ADD_VARIABLE, y = ADD_VARIABLE, color = year)) +
geom_point() +
geom_vline(xintercept = 1963, color = "red") +
labs(y = "ADD A TITLE")
```

   - 1. Describe this plot; What information does it contain?
   - 2. What is significant about the red vertical line? (You may have to go online to search for this answer.)

7. Create a new dataset from `dat_measles_rate_lessTwoStates`, called `dat_caliFocus` in which *California* is the only state present in the data.

8. Prepare a plot of this dataset where x is the `year` and y is the `rate`. Hint, modify the given plotting code from above to make a plot for `dat_caliFocus`.

   - 1. In clear and meaningful language, interrupt your results from the plot.
   - 2. Compare this plot to the one that you made earlier. Could California be used to represent the rest of the country in terms of general and similar patterns? Why or why not? What are these patterns?
   - 3. Describe what both of these plots are showing. Why is the analysis that you have just completed so revolutionary in medical science?

Part 2. Ethics

Please write your reflections in markdown. Pleaes save your work in the file, `writing/reflections.md`.

- In the New York Times article, entitled, "Journal Retracts 1998 Paper Linking Autism to Vaccines" by Gardiner Harris (`https://www.nytimes.com/2010/02/03/health/research/03lancet.html`) a research article written by Dr. Andrew Wakefield has been retracted by the authors because it suggests that autism followed from the use of vaccines. Read the article (also found in `reading/` to answer the following reflection questions to place in your `writing/reflections.md` file.

   1. What is the damage to the public medicine and public opinion from such an article which states (incorrectly) that autism is a result of vaccines?
   2. What should the role of academic research groups and organizations be to ensure that published information is absolutely correct (i.e., has been properly analyzed) before public exposure?

3. Researchers associated with this paper retracted their names from the article. This means that the researchers no longer support its content and science. What was the main damage to public medicine from this paper? Can retracting names from a paper (or retracting the paper itself) be enough to fix this damage? How could the damage be fixed, in your opinion?

Important Details

All of your R code should be placed into a separate file, `src/vaccines.r` where each of your statements is justified or is explained. You instructor will run your code and so if it does not appear to serve any immediate function, your justification will help in comprehension. In addition, the answers to the questions-in-blue above will be placed in the file, `writing/reflections.md`, along with your response to the ethical questions.

Note: Please remember to include your name on everything you submit for the class.

Required Deliverables

This portion of the assignment invites you to submit an electronic version of the following deliverable through your GitHub Classroom lab repository. Note: this repository is the one which you clone from the above link.

1. File, `src/vaccines.r`; Your R program source where you respond to the code and questions in blue in Part 1, above.

2. File, `writing/reflections.md`; Your written reflection responses for Part 2, above.