 Due by 2:30pm, 12^{th} Oct 2018

CMPSC 301
Data Analytics
Fall 2018

Lab 4: Exploratory Data Analysis
28^{th} Sept 2018

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/a4iRigxN

To use this link, please follow the steps below.

- 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`

Lab directory structure: You are to create a labs directory (`mkdir labs` in which you are to add the GitHub Classtoom repositories for each of your weekly labs (use this command `mkdir labs/labsxx`, where xx is the two digit lab number). For example, your first and second lab repository should be located in the paths, `labs/lab01` and `labs/lab02`, respectively.

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

Groupwork: Each person submits own work to own repository

You are to work in a group of 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. Note: an interesting resource for making boxplots and other plots may be found at: in our online textbook at `http://r4ds.had.co.nz/exploratory-data-analysis.html`, at `http://www.r-graph-gallery.com/portfolio/boxplot/` or at the end of this lab in Section .

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, sensationalists media reports and fear mongering from conspiracy theorists, led parts of the public to believe that vaccines were harmful. Some parents stopped vaccinating their children. This dangerous practice can be potentially disastrous given that the Center for Disease Control (CDC) estimates that vaccinations will prevent more than 21 million hospitalizations and 732,000 deaths among children born in the last 20 years. (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:

```
# Run the below only if the library is not already installed.
install.packages(dslabs)
library(dslabs)
data(us_contagious_diseases)
```

1. Use the `us_contagious_disease` and `dplyr` tools to create an object called `dat` that stores only the Measles data, includes a per 100,000 people rate, and removes Alaska and Hawaii since they only became states in the late 50s. Note that there is a `weeks_reporting` column. Take that into account when computing the rate. Hint: one possible rate equation is the following;

$$per100000rate = \frac{count*100000}{population} * \frac{WeeksReporting}{52}$$

Enter your R code in a separate `src/Lab4Program.r` file.

2. Plot the Measles disease rates per year for California. Measles vaccine was introduced in 1965. Add a vertical line to the plot to show this year in relation to the data. Hint: Lookup and use `geom_vline()` for this vertical line addition.

Add and justify your R code to the `src/Lab4Program.r` file.

3. **With and Without the square root transformation of California** We note that it is not always possible to compare the data of the 1950's, 1960's and 1970's directly and so we will use a mathematical transformation to allow us to stabilize the variability between these decades, and to help us make comparisons.

   For California data of the 1950's, 1960's, and 1970's, plot the histogram of the data across states **with** and **without** the square root transformation of the count to determine whether there is any new information shown when comparing the plots.

   For this step, $\sqrt{count}$ is to be placed inside your graphing function where you assign the plotting function's y coordinate to its data. Which seems to have more similar variability across years? Make sure to pick *binwidths* that result in informative plots. Note: Make sure that the numbers $0, 4, 16, 36, \ldots, 100$ appear on the y-axis to be sure that your transformation was successful.

   The following hint is R code to help you get started. :

```
dat_caliFocus <- filter(us_contagious_diseases, state == "California")

dat_caliFocus$yearBlock[dat_caliFocus$year == 1950] <- "1950's"

## TRANSFORMATION, Multi-bar per state,
ggplot(data = dat_caliFocus ) + geom_bar(mapping = aes(x = state,
y = ENTER_TRANSFORMATION_FUNCTION, fill = yearBlock), position = "dodge",
stat = "identity") + theme(axis.text.x = element_text(angle = 90,
hjust = 1, vjust=-0.01))
```

Add and justify your R code to the `src/Lab4Program.r` file.

4. Above we used the hint code to view only California counts. Modify this code to perform the same type of count analysis over all states together. Place the data from all US states into one plot to allow a cross comparison of data over the three decades (i.e., 1950's, 1960's and the 1970's.) Does the pattern you saw above hold for other states, as well?

   Add and justify your R code to the `src/Lab4Program.r` file.

5. One problem with the plot above is that we cannot distinguish states from each other. There are just too many. We have three variables to show: year, state and rate. If we use the two dimensions to show year and state then we need something other than vertical or horizontal position to show the rates. Try using color. Hint: Use the the geometry `geom_tile` to tile the plot with colors representing disease rates.

   Add and justify your R code to the `src/Lab4Program.r` file.

6. The plots above provide strong evidence showing the benefits of vaccines: as vaccines were introduced, disease rates were reduced. But did autism increase? Perform research to find (published) yearly reported autism rates data and provide a plot that shows if it has increased and if the increase coincides with the introduction of vaccines.

   Add and justify your R code to the `src/Lab4Program.r` file.

7. Use data exploration to determine if other diseases (besides Measles) have enough data to explore the effects of vaccines. Be critical of your online research. Are you convinced by the primary or credible articles that you have found? Be sure to cite all articles.

   Report: Prepare a report in the Markdown file, `writing/reportAndReflections.md`, with as many plots as you think are necessary to provide a case for the benefit of vaccines.

8. 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 to answer the following reflection questions to place in your `writing/reportAndReflections.md` file.

   - 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?
   - 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?

- The researcher retracted his paper, which means that he no longer supports its content. Is retracting a paper enough to fix the damage to public medicine? How could the damage be fixed if this you do not think that this is enough?

9. During the talk concerning Global Health, given by Dr. Becky Dawson and Dr. Amelia Finaret on Friday, 28^{th} September, one of the main points is that data is used to fight disease and to track out-breaks of serious ailments in diverse communities all over the world. Find two sets of (publicly available) data from the *World Health Organization* which could be used in a project to determine whether vaccines are beneficial, in any capacity.

10. Write your reactions to the talk and discuss an idea which resonated with you.

    Your completed report should be around two pages in length.

Important Details

All of your R code should be placed into a separate `src/Lab4Program.md` 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.

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

Boxplots

Your slides and notes contain the code to create many different types of plots. You might need to make a boxplot using your current data. The code to perform this operation using the `diamonds` data set is the following.

```
library(tidyverse)
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_boxplot()
```

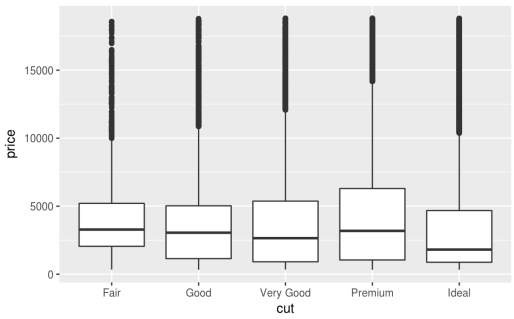

Figure 1: Box plot from the code above.

`writing/reportAndReflections.md`

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. In File, `src/Lab4Program.md`; Your R program source where each line is justified in your responses to items 1-6. You are to use an appropriate header file with your name and an Honor pledge.
2. In File, `writing/reportAndReflections.md`; Your written report, with appropriate graphs and reflection responses for items 7-10.