CMPSC 390
Data Analytics
Fall 2017

Lab 4: Exploratory Data Analysis
Save this lab assignment to: `labs/lab4`

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.

Groupwork

You are to work in a group of not more than three 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.

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:

```
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.

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

2. Plot the Measles disease rates per year for California. Find out when the Measles vaccine was introduced and add a vertical line to the plot to show this year.

   Add your R code to the `Lab4Program.r` file.

3. Note these rates start off as counts. For larger counts we can expect more variability. There are statistical explanations for this which we don't discuss here. But transforming the data might help stabilize the variability such that it is closer across levels. For 1950, 1960, and 1970, plot the histogram of the data across states with and without the square root transformation. Which seems to have more similar variability across years? Make sure to pick binwidths that result in informative plots.

   Add your R code to the `Lab4Program.r` file.

4. Make a plot from step 3 with the square root transformation. Make sure that the numbers $0, 4, 16, 36, \ldots, 100$ appear on the y-axis.

   Add your R code to the `Lab4Program.r` file.

5. Now, this is just California. Does the pattern hold for other states? Use boxplots to get an idea of the distribution of rates for each year, and see if the pattern holds across states. (Note: an interesting resource for making boxplots 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 .

   Add your R code to the `Lab4Program.r` file.

6. One problem with the boxplot is that it does not let us see state-specific trends. Make a plot showing the trends for all states. Add the US average to the plot. Hint: Note there are missing values in the data.

   Add your R code to the `Lab4Program.r` file.

7. One problem with the plot above is that we can't 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 your R code to the `Lab4Program.r` file.

8. The plots above provide strong evidence showing the benefits of vaccines: as vaccines were introduced, disease rates were reduced. But did autism increase? Find 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.

   Start a report document named `Lab4Report` and add your conclusions there.

9. Use data exploration to determine if other diseases (besides Measles) have enough data to explore the effects of vaccines.

   Prepare a report with as many plots as you think are necessary to provide a case for the benefit of vaccines.

10. Read the article titled"Law, Ethics, and Public Health in the Vaccination Debates: Politics of the Measles Outbreak" that you can find in the shared course repository. Based on the facts outlined in the article and your data exploration work for this lab, add at least one paragraph to your report reflecting on the following issues/questions:

    - *ethics*: is it ethical to force vaccinations? is it ethical for parents to refuse vaccines?
    - *privilege* : is the refusal to vaccinate an expression of privilege?
    - *inequality*: does refusal to vaccinate draw on structural inequality?
      (Def: structural inequality is a condition where one category of people are attributed an unequal status in relation to other categories of people.)

    Add to the report your thoughts and reflections to the above questions.

Your completed report should be at least 2 pages long.

Important Details

Lab directory structure: Make sure you have placed your submission materials for this lab into `labs/lab4` directory in your Bitbucket repository (`cs390f2017-billb`).

Submission Information Your 2-3 page report is to be typed up using a word processor such as LibreOffice or using LaTeX. All of your R code should be placed into a separate `Lab4Program.r` file. Your are to submit your document to the instructor using your Bitbucket repository.

Note: Please remember to include your name on everything you submit for the class. Although the instructor collects your work from Bitbucket, each work must be graded outside of the Bitbucket directory and so without adding your name, the instructor will be unable to award you credit for your work.

Required Deliverables

Submit electronic versions of the following deliverable through your Bitbucket repository (`cs390f2017-billb`) by correctly using using appropriate Git commands, such as `git add -A`, `git commit -m ‘‘your message’’` and `git push`. When you have finished, please ensure that the Bitbucket Web site has your pushed work. Please contact the instructor if you have any questions about assignment submission.

1. Your written report, named Lab4Report, with appropriate graphs and reflection answers for points 8-10.
2. Your R program, named Lab4Program, with appropriate header file with your name and an Honor pledge, as well as comments for each point 1-7.

Boxplots

You can make a box plot for the *diamonds* data using the following code:

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

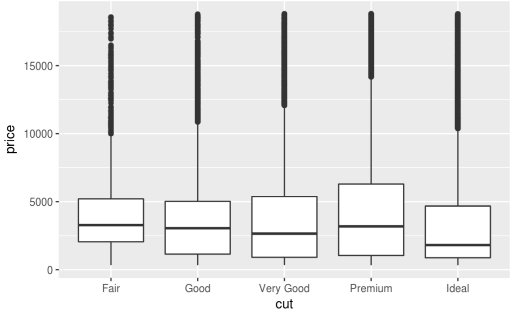

Figure 1: Box plot from the code above.