Day 4 Worksheet

California Diseases

Many states keep track of multiple types of health-related data. In California the California Health and Human Services Agency has a nice website with 436 open datasets, on all types of databases about health linked to demographics. The goal is to advance health care by making databases open and accessible for analysis by individuals and companies. Here is the website: <https://data.chhs.ca.gov/dataset>. I’ve downloaded a dataset that describes the number of cases of 11 different diseases that are preventable with modern vaccines. For this question I want you to answer the following questions using the dataset “virus\_cases\_ca.csv”, which includes over 8,000 data points for 11 diseases. In this dataset there are 3 categorical variables; the type of disease (11 types are recorded here, you should set this as a factor, ***as.factor ()*** ), the county in California, the year (this dataset goes from 2001-2018 you should set this as a factor), and the last column is a count of the number of cases recorded. Check structure ***str ()*** to make sure that all of your categorical variables are factors, and the disease counts is an int (for integer). Below are some commonly used R functions, but I also suggest you have your script from yesterday open for reference, for any relevant action you can copy from your old script to make a new script. Here is a brief summary of some of the R functions you might want to use;

* To understand the structure/column titles of a dataset: ***str ()***, ***head ()***
* Summarize a column*:* ***mean (), sd (), var (), min (), max (), summary (), table ()***
* Manipulate the dataset: ***subset (), aggregate ()***
* Ggplot2 functions you might need: ***ggplot(), geom\_points(), geom\_smooth(), geom\_histogram (), geom\_density (), geom\_boxplot (), geom\_col (), geom\_line ()***

Part 1: Load and Analyze the Data

Answer the following questions about the disease dataset (“virus\_cases\_ca.csv”):. Make sure you set your working directory ***setwd()*** at the beginning of the script and ***read.csv ()*** to import the datafile.

1. Which disease has the highest total number of infections? Which disease has the least number of occurrences?
2. Which counties had a total number of infections greater than 2,000? Which county had the most infections and which had the least?
3. Which year had the most infections for Tetanus? Which year had the most infections for Pertussis? Which year had the most infections for Diphtheria?
4. For every county and every year, which county-year combination had the highest number of infections? How many county-year combinations had more than 1000 infections? What type of trends in county-year combinations do you see, what was a bad year, and which counties were the worst?
5. For measles, which county has the highest total number of infections? For Tetanus which counties had more than 10 infections? For Rubella which counties had at least one infection?

Part 2: Graphing Patterns of Infections

1. Make a bar graph of the total number of cases for each disease (you should already have this dataframe from your analysis above. Can you make each bar a different color? Take some time to format you ggplot script to label the axes, add a caption and make it look as good as possible.
2. Make a bar graph of the total number of infections for each disease for San Mateo County. Can you make each disease a different color? Take some time to format you ggplot script to label the axes, add a caption and make it look as good as possible.
3. Make a graph of the mean number of measle cases (pool all the counties together) for each year in the database. You could make this as a bar chart or a boxplot (I like boxplot because it shows the variance and the outliers). Can you make the different years different colors? Take some time to format you ggplot script to label the axes, add a caption and make it look as good as possible.
4. Pick another disease you are interested in. Make the same graph as above but for the disease you picked. To make this graph you can just copy the aggregate and the ggplot script from above and insert the values for the dataframe of the disease you picked. Do you see how you could really rapidly copy script to make the same graph for all 9 disease (it is faster if the name of the dataframes are easy to type)? Time yourself, how fast could you make all 9 graphs for the remaining diseases?
5. If you look at just mumps, which county had the highest mean number of cases? What is the total number of cases for Santa Clara County? If we consider Los Angeles, San Diego and Orange county (as south California) and Alameda, San Francisco, San Mateo, and Santa Clara (as Northern California) is there a difference in the number of cases between Southern California and Northern California? Make a Graph that is a comparison of mean number of mumps cases for Northern and Southern California. Can you figure out how to add error bars to a bar graph? You might be able to find this script on the internet, or ask me and I can show you my method of scripting for mean and error bars in ggplot.

Part 3: Bonus Activity: Find a new dataset

Navigate to the website: <https://data.chhs.ca.gov/dataset> This is where I got the disease database that we analyzed above. Can you find another database that has data you are interested in? Try to download this database and import it into R. Sometimes databases are different formats and we need to do some manipulation to get them into R. Try to write a script that imports the data and produces one graph that you can show me to explain what the database is about.

Part 4: Bonus Activity: Make a great graph

Tomorrow we are going to work on making an infographic that explains one of your graphs. To do this you need to pick a graph we have made in R during this class. Find the data set that is your favorite and make a great graph. Use all of the components we have talked about adding in ggplot. Make sure your graph has a caption, this will help to explain what the graph is to your reader.