CSE 587:DIC Lab-1 Part-2

Project Documentation

Yasha Ballal
yashaash@buffalo.edu

Abhinav Neelakantan
aneelaka@buffalo.edu

Overview:
We use the R programming language working with Jupyter notebook and Rstudio to recreate plots from the flu.gov website. The challenge is to find the right data (clean and in proper format) and work with it using the appropriate functions to create the plots.

Steps of execution:
1. Gather data
- Data is gathered in the form of csv files from the flu.gov site to recreate the plots.
- The data is available as per requirements and can be parameterized as required
- Distribution of data is according to seasons, years and weeks

2. Choose appropriate plot functions
- **R provides several plot functions such as bar plots, histograms and line plots**
- **R also provides a lot of customization options to display the chosen data**

Important Concepts:
- **Data Cleaning**

1. Data cleaning is important before plotting it on a graph.
2. Here the step is unnecessary as flu.gov provides clean data

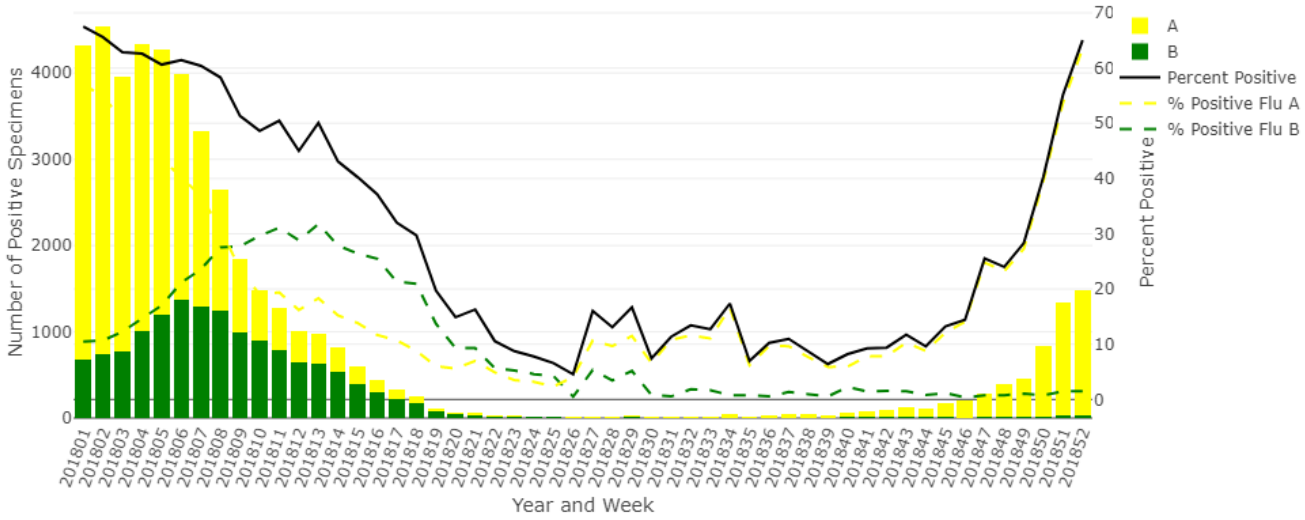
- **Data Formatting**

1. Data in the csv file is sometimes not directly loadable into R data frames and some minor tweaks are essential.
2. Something as simple as skipping reading the top line in the csv file because it has no purpose in the graph and hence the data frame can be considered as a part of this process

- **R libraries**

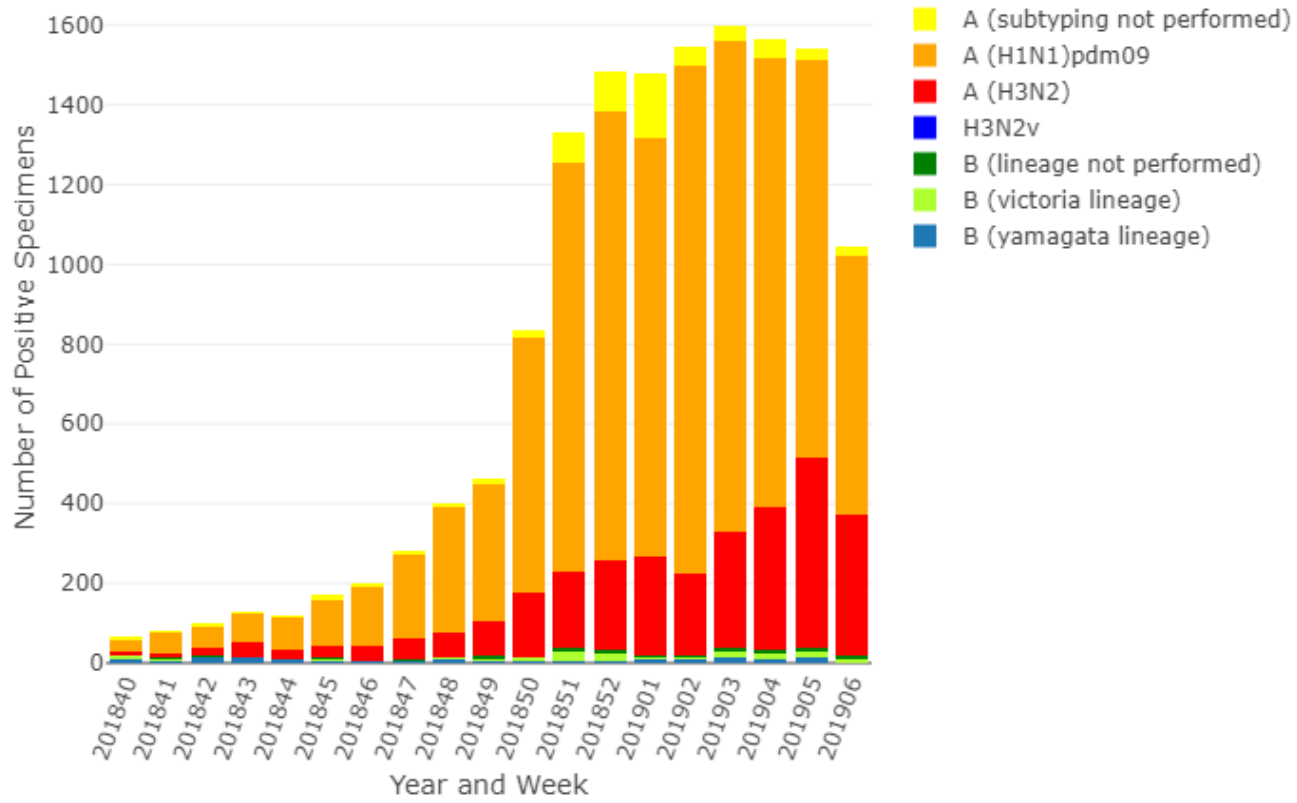
1. stringr- used to perform string manipulation in data frames
2. ggpolt2- the basic plotting library has a variety of plot options using aes(aesthetics) parameter among many others.
3. plotly- provides higher customization to plots than ggplot2 provides
4. dplyr – used to get the latitude and longitude for plotting the states in the heat map.

Plots:
1) Positive tested
- We use the plotly library to plot his graph
- The graph has two Y axes one for the bar graph for the number of cases and the other for the line graph showing the percentages
- It categorizes the flu roughly into types A and B



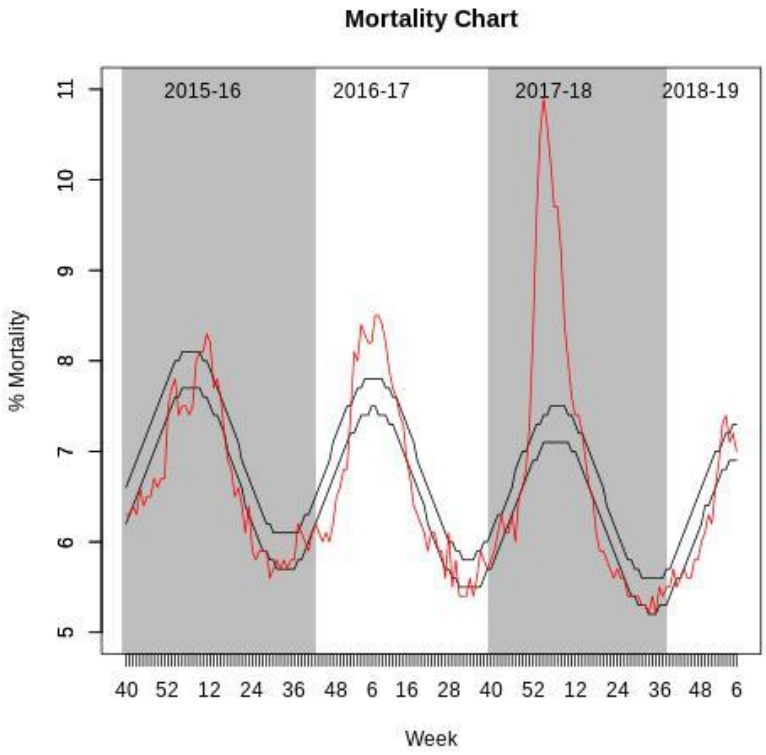
2) Positive tested with sub types

- We use plotly to plot this graph
- It is a bar graph that plots the number of flu cases categorizing them into the sub types
- It customizes individual bars to plot subtypes according to color



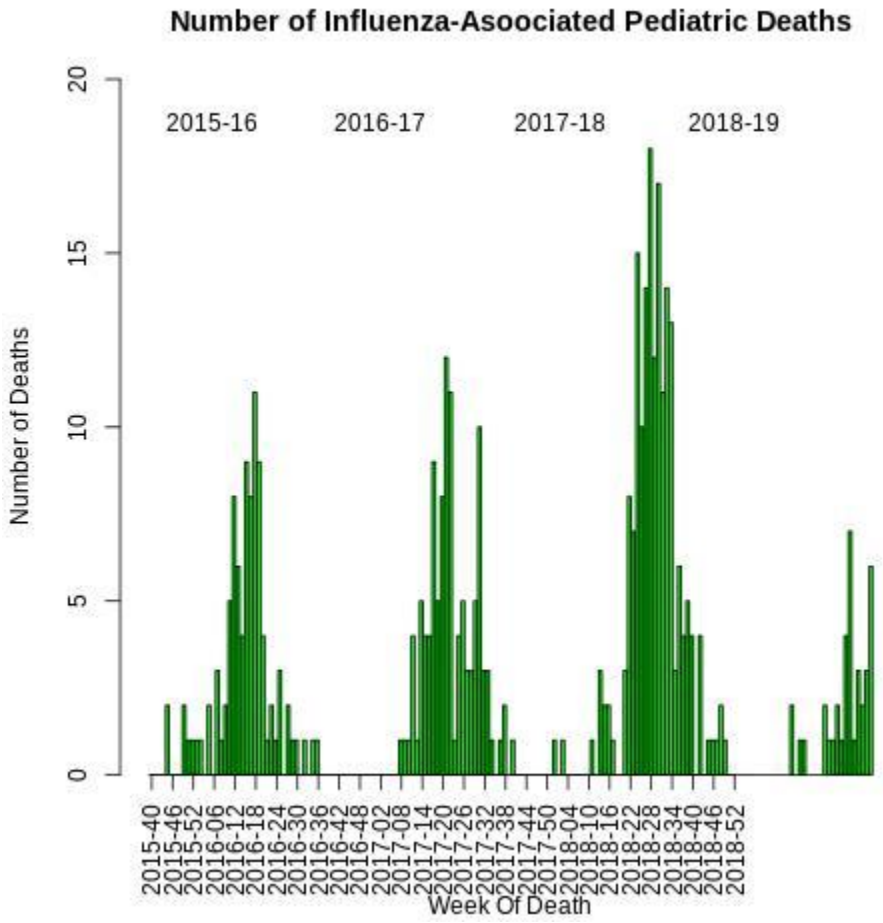
3) Mortality

- We use the R plot function for this graph
- It is a line graph that plots the percentage of deaths due to the flu considering three parameters- actual percentage, threshold allowed and the base percentage



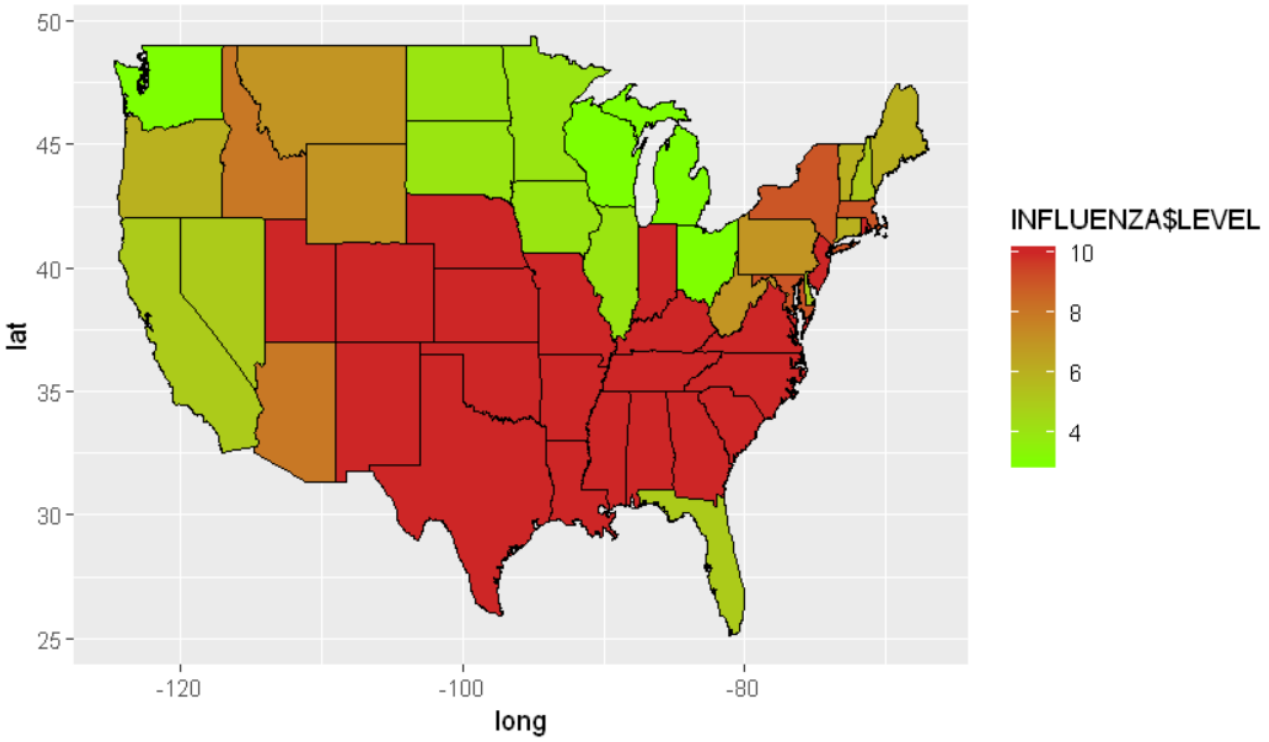
4) Infant Mortality

- we use the R plot function for this graph
- It is a bar graph that plots the number of infants that died due to influenza according to week of the year.



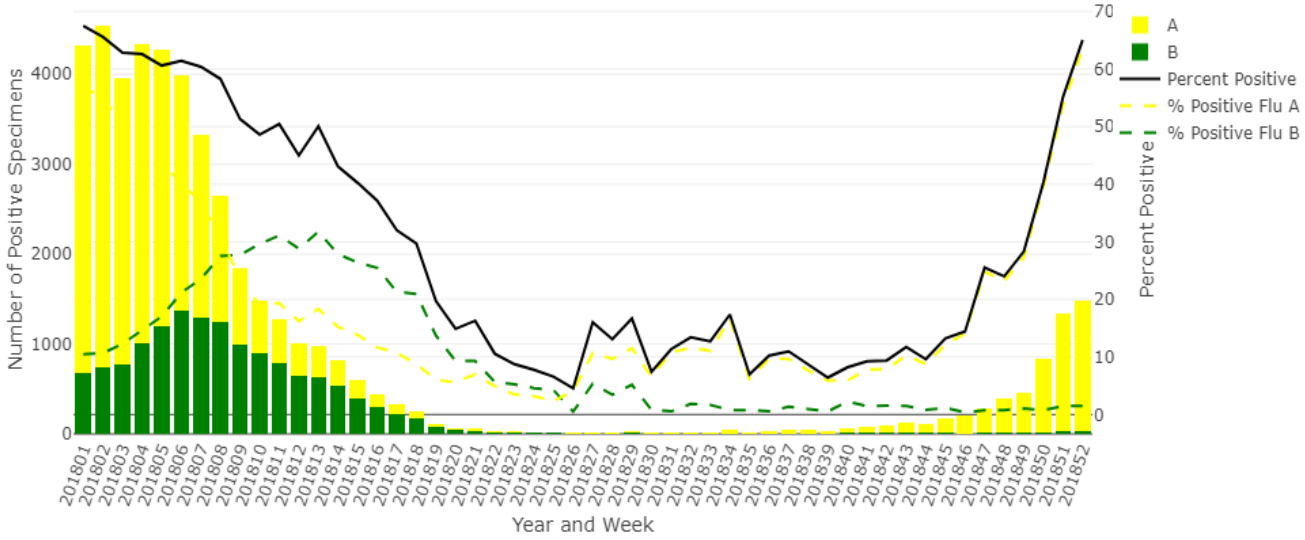
5) Heat Map of the US

- Libraries required – ggplot, dplyr
- The following steps were involved in the plotting of the heat map:
  - From the data read from csv, the STATENAME column needs to be converted to lower case, so that this can be merged with the state dataframe that already exists in R dplyr library. This needs to be done to plot the latitude and longitude-wise map of the United States using ggplot.
  - There are a number of options available for customization with ggplot. To fill our heat map with color gradients according to the risk each state has of influenza, we need to convert the levels into integers and fill continuous values from green to red for low to high.



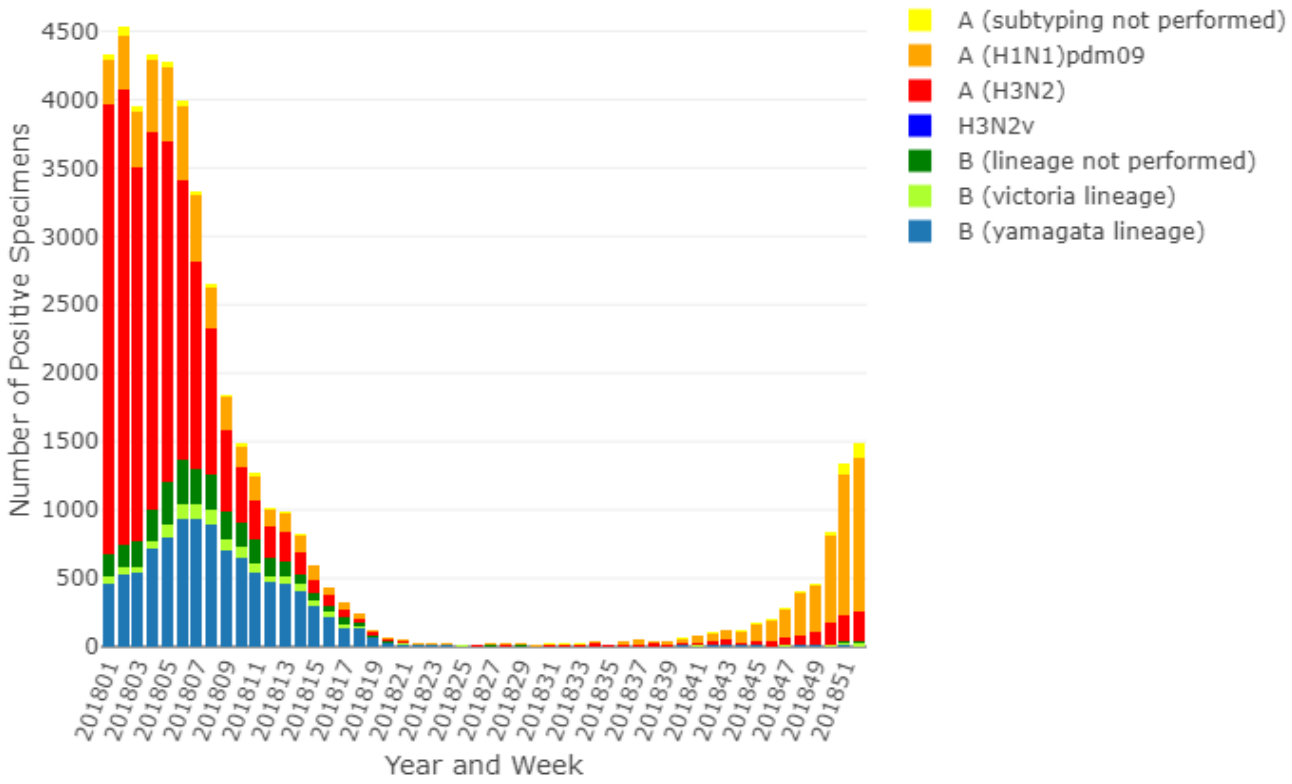
6) Positive tested for 52 weeks of 2018

- In order to get this data, two csv files(2017-2018, 2018-2019) needed to be merged and the data needed to be cleaned, so that we got only the columns relevant for 52 weeks of 2018.
- The rest of the procedure followed is the same as done for plot number 1.



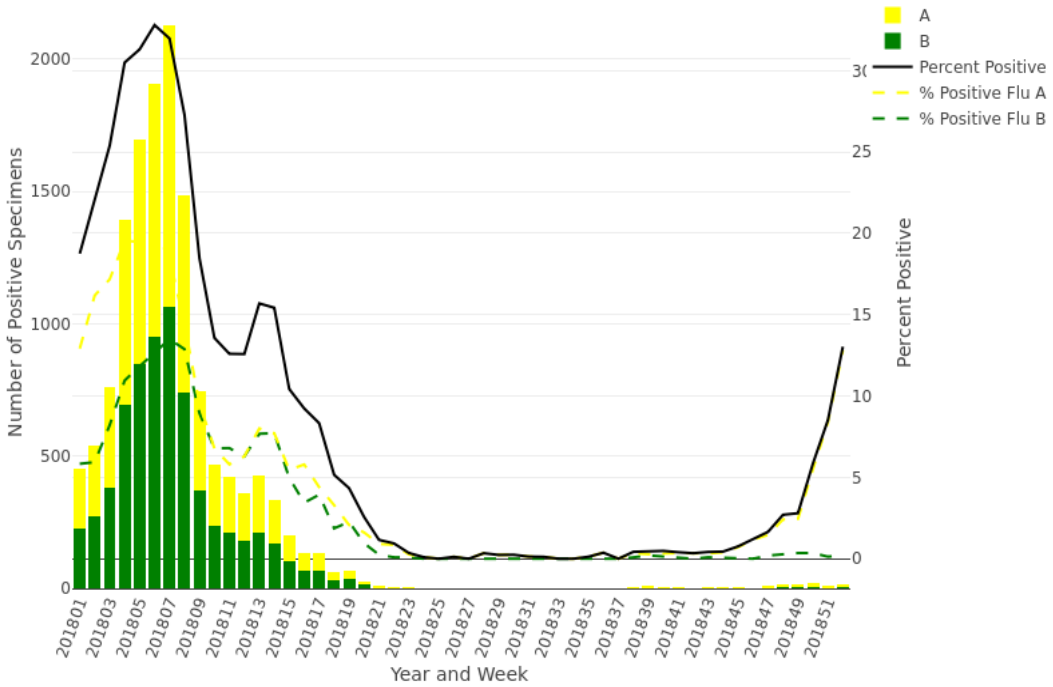
7) Positive tested for 52 weeks of 2018 with sub types

- In order to get this data, two csv files(2017-2018, 2018-2019) needed to be merged and the data needed to be cleaned, so that we got only the columns relevant for 52 weeks of 2018.
- The rest of the procedure followed is the same as done for plot number 2.



8) Positive tested for 52 weeks for NY state

- we use plotly for this graph
- It plots the number of influenza cases in NY state



References:

- https://www.r-bloggers.com/palettes-in-r/
- http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually
- https://ggplot2.tidyverse.org/reference/aes_colour_fill_alpha.html
- https://ggplot2.tidyverse.org/reference/scale_colour_continuous.html
- https://stackoverflow.com/questions/25449093/ggplot2-geom-polygon-with-no-fill
- https://stackoverflow.com/questions/24496984/how-to-add-legend-to-ggplot-manually-r/24497113
- https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf
- https://swcarpentry.github.io/r-novice-inflammation/02-func-R/
- https://gis.stackexchange.com/questions/234942/keep-customized-colors-with-geom-polygon
- http://r-statistics.co/Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html#Legend%20Labels%20and%20Point%20Color
- https://ggplot2.tidyverse.org/reference/scale_colour_continuous.html
- https://stackoverflow.com/questions/29278153/plotting-with-ggplot2-error-discrete-value-supplied-to-continuous-scale-on-c
- http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually#change-colors-manually
- http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf