## **Flu Mortality Impact Analysis**

## **Saurabh Biswas**

## **DSC540 T303**

**Introduction:** Each year lot of people dies because of influenza aka flu in United States of America. Impact of flu related death varies from state to state. Few months of the year, the impact of flu is much more than the other months. I wanted to carry out a detail analysis based on data provided by CDC, data.gov, datausa.io and 50states.com.

**On milestone2,** I gathered dataset from data.gov on flu. This dataset contains data from 1962-2016 for different cities. I selected data between 2005 to 2016 for my purpose. Then I cleaned the dataset and dropped all records having null data on flu death and all death. Further I had to get the data at month & state level from weekly city data. Then I removed all outliers using box plots.

**On milestone3,** I web scarped a website that provides short US state abbreviation and corresponding full name. Outcome of our milestone2 is a dataset that contains details on flu death, other deaths per state year wise. The state in this dataset is a two-position short abbreviation. I have used the dataset from web scraped data and merged it with previous one to get full state name in the combined dataset.

Also, I web-scarped CDC provided flu effectiveness dataset. I used timeseries technique to get the flu vaccine effectiveness at year level. I used these data in next steps to perform a comparison between flu death rate and flu vaccine effectiveness.

**On milestone4,** I used API provided datausa.io to get the total population count for USA between 2013-2106 and population count at each state during the same window.

**On milestone5,** I merged these counts with original dataset and calculated flu death count per million for at state level and flu death count per ten million at country level. Also, I merged vaccine effectiveness dataset with country level data.

I loaded state level data and country level data into sqllite3 databases and then created visualization.

I plotted bar plot for flu death rate year wise. From this I can see year 2013 has highest flu death count per 10 million people.

From the scatter plot of year/month vs flu death rate, we can see 2015 Jan has highest flu death count per 10 million people.

From flu death rate and vaccine effectiveness plot we can see, In 2013 vaccine effectiveness was highest, even though flu caused more death than other years. I didn't expect this behavior. Whereas in 2016, flu death rate dropped with increase in vaccine effectiveness than year 2014 and 2015. This is in line with my expectation. But due to this mixed behavior, we can't conclude what is the effect on vaccine effectiveness on flu death rate.

From state wise scatter plot, we can see DC was most affected state by flu on 2013.

**Things I learned:**

* How to split a data frame column into multiple column and then extract the information of interest.
* Cleaning and dropping null values.
* Removing outliers using box plot.
* Grouping data based on few columns based on my need.
* web scraping and clean that with beautifulsoup.
* Figuring out how we can merge different datasets to get a combined view.
* API call and parsing a JSON and load that into a data frame.
* Loop through data frame rows
* creating new views at different aspects from the data frame.
* Loading data frame into a database.
* plotting and adjust the font size, orientation etc.