Deaths in USA due to major diseases (2005 – 2015)

STAT 515 – applied statistics and visualization of analytics

Dilip Molugu
MS in Data Analytics Engineering
George Mason University
dmolugu@gmu.edu

Abstract

This project deals with various visualization techniques applied to a dataset, to make it easy to draw conclusions about the data by looking at the plot. I have used mortality data, population estimates to estimate heart disease and cancer deaths and other major diseases from 2005 through 2015 to find changes in deaths resulting from population risk, growth, and aging. The raw data set is very difficult to understand and analyze to get to a conclusion by looking at the datasets. Finding the right data for showing the required output is also important for this project. The challenging part in this project is the cleaning process involved with the raw data. Because a cleaned data is key for effective data analysis. According to my analysis cancer has always been the leading cause of death from 2005 through 2015. Deaths due to heart diseases decreased from 2005 to 2010 but has again increased from 2010 to 2015. When I have done more deep analysis by partitioning the data according to the ages I found an interesting trend which is also good news that for the population below the age 50, the cases of heart disease deaths have decreased from 2005 to 2015. Which means that younger population is more more health conscious and may be also due to improvement in health care and medical services in the country.

Introduction

For most of the last century, the main cause of death in the United States, was heart disease, followed by cancer, Stroke, chronic lower respiratory diseases and unintentional injury. Cancer overtook heart disease to become the leading cause of death in 1993, although the trend slowed or stopped in recent years.
Increasing death due to these deaths indicate that the overall population dying from heart disease or cancer has increased.
The objective of this study was to use mortality data, current population estimates to find the general patterns in the deaths in the country. And to provide more age standardized death rate by using age data and population data from the given data set. The results in this analysis which are quite surprising and interesting as they are totally different when we do some deep analysis from the raw data. The results obtained from this visualization can also be used for predictions which will be useful for predicting the potential deaths in future and take measures to reduce the mortality rate.

Dataset Implemented

I have taken the data from the CDC website which has all the statistical data for the number of people dying due to various diseases in the USA. The data has various attributes like the year of deaths, the names of the states, the observed number of people dying due to various diseases, the predicted number of deaths (by NCHS- National

center for health statistics), the number of people dying in metropolitan region and non-metropolitan region, population for every state for every year, age-range for different ages like 0-49, 0-54, 0-79, 0-84 etc.

	Year	Cause_of_Death	State	State_FIPS_Code	HHS_Region	Age_Range	Benchmark	Locality	Observed_Deaths	Population	Expected_Deaths	Potential_Excess_Deaths	Percent_Pot
1	2005	Cancer	Alabama	AL	4	0-49	2005 Fixed	All	756	3148377	451	305	40.30%
2	2005	Cancer	Alabama	AL	4	0-49	2005 Fixed	Metropolitan	556	2379871	341	217	39%
3	2005	Cancer	Alabama	AL	4	0-49	2005 Fixed	Nonmetropolitan	200	768506	111	89	44.50%
4	2005	Cancer	Alabama	AL	4	0-49	2010 Fixed	All	756	3148377	421	335	44.30%
5	2005	Cancer	Alabama	AL	4	0-49	2010 Fixed	Metropolitan	556	2379871	318	238	42.80%
6	2005	Cancer	Alabama	AL	4	0-49	2010 Fixed	Nonmetropolitan	200	768506	103	97	48.50%
7	2005	Cancer	Alabama	AL	4	0-49	Floating	All	756	3148377	451	305	40.30%
8	2005	Cancer	Alabama	AL	4	0-49	Floating	Metropolitan	556	2379871	341	217	39%
9	2005	Cancer	Alabama	AL	4	0-49	Floating	Nonmetropolitan	200	768506	111	89	44.50%
10	2005	Cancer	Alabama	AL	4	0-54	2005 Fixed	All	1346	3463216	784	562	41.80%
11	2005	Cancer	Alabama	AL	4	0-54	2005 Fixed	Metropolitan	968	2615416	590	379	39.20%

Fig1: this is the image of the dataset (imported to R)

As you can see in the image there are a lot of redundant values. We can see a lot of repeated values in the expected deaths and observed deaths column. This redundancy is due to various categories like age, locality etc. All these redundant values must be removed from the dataset before we can proceed to the visualization part of this analysis.

Data cleansing

For good visualization of data, the data must be cleaned before we plot any of the graphs, otherwise the plots don't make sense. The plots must show accurate information as they are used to draw important conclusion about the data. The data has different levels of cleaning so that it can be used for several graphs. The first step is to remove the redundant values by using the filter function R from the data, which will make it easy for us to plot the graphs. I have used different filters depending on the graphs and the attributes required for the plot was trying to show. I have also used the aggregate function to sum up a lot of number depending on their category. Aggregation of the data is very important to show the categorical data. The grouping of data is done at various levels like grouping on year, type of deaths, locality of deaths, ages of deaths. All these are done to show a variety of plots.

Clustering of data

I have used the k means clustering in R to find the top 5 diseases and filtered the data according to their ranking. When I have performed the k means clustering I was able to see that top 5 reasons of deaths were cancer, heart diseases, unintentional injury, chronic lower respiratory diseases and stroke. I have done this to consider only the top 5 reasons for death in the country from 2005 through 2015.

Fig2: clustering of raw dataset

I have performed this clustering just to find the top 5 diseases and their ranking. Because it is going to be difficult to manually go through the vast dataset to find their ranking. From the plot we can see that highest number of deaths are due to cancer. Depending on their concentration in the plot we can find their ranking.

List of top 5 causes of deaths in the USA

- Cancer
- Heart diseases
- Unintentional Injury
- Chronic lower Respiratory Diseases
- stroke

Implementation of graphs

a. total deaths in the USA due to 5 major diseases (2005-2015): To get the Basic idea of the information about the total number of death in the country from 2005 through 2015, I have plotted this graph. From the graph we can observe that the maximum number of deaths are observed in California followed by Texas and Florida.

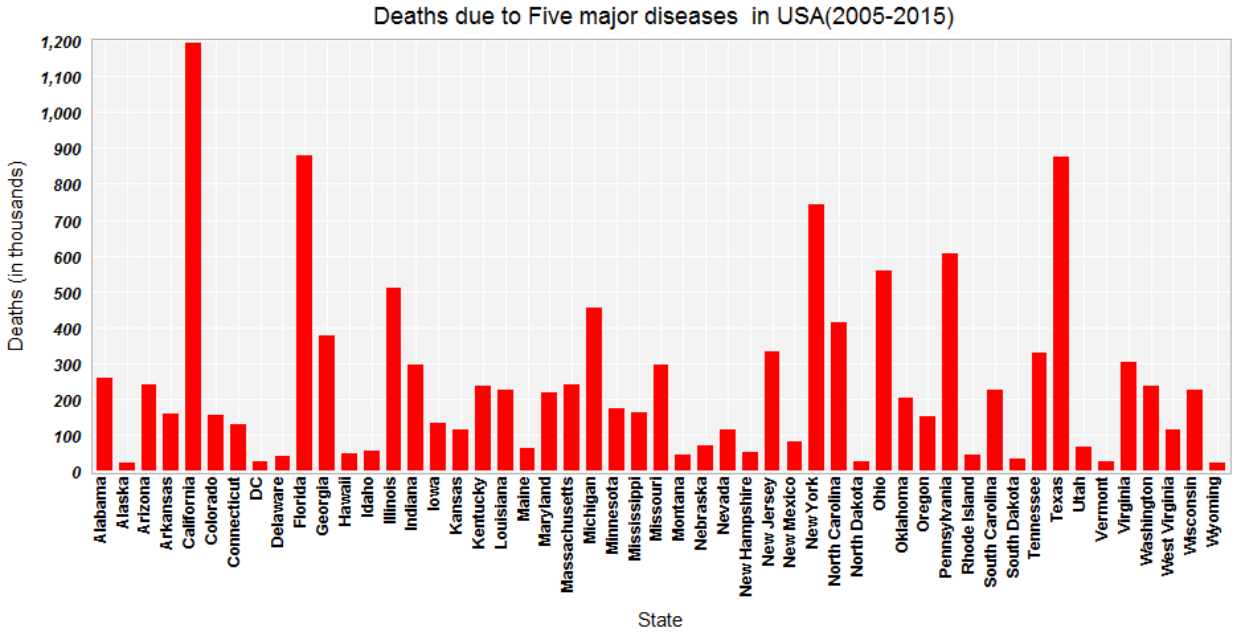


Fig3: total deaths in the USA due to 5 major diseases (2005-2015)

b. comparison- type wise top causes of deaths: From the previous graph(fig:3), I have realized the need of studying the deaths in each year and grouped by cause of death and below is the graph representing the same. Here we are observing how the number of deaths are changing over the years (2005 to 2015). For this I have used the facet wrap to group the data depending on the cause of death and to show this data for every year. I have used a differnt color for each disease type as it will make it easy to differentiate while looking at the graph. **An important observation is that the number of death due to heart diseases has decreases from 2005 to 2010 but has increased from 2010 to 2015**. This observation important to note. And for all the other diseases the number is increasing slightly but this may also be due to the increase in population.

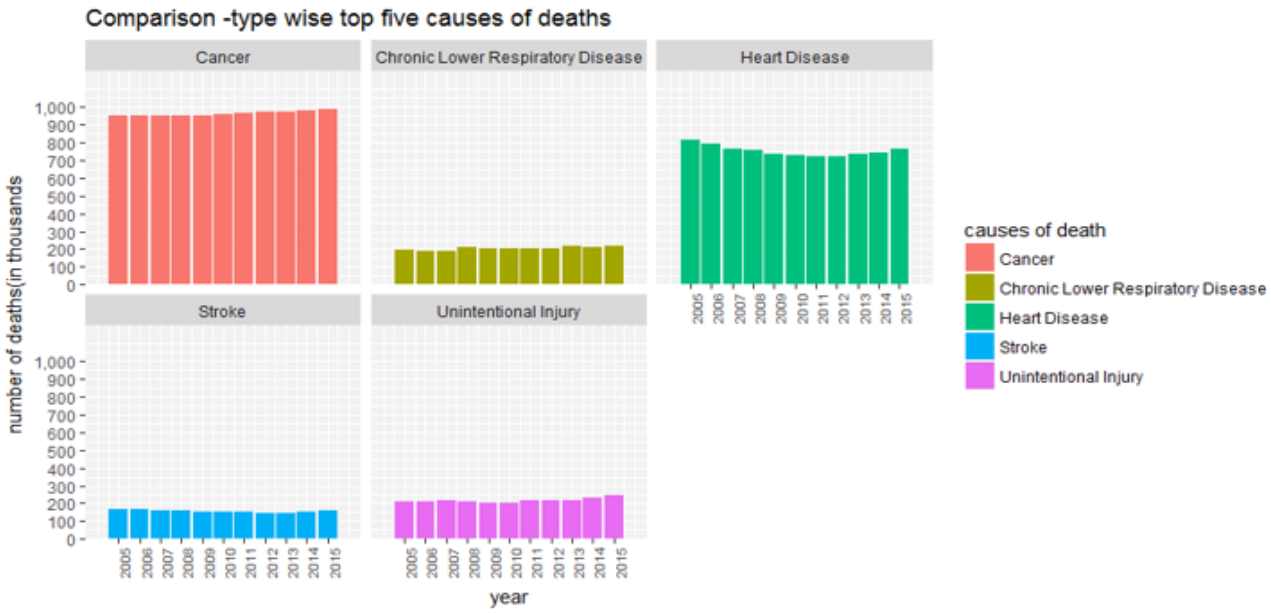


Fig 4: comparison- type wise top causes of deaths

c. comparison- year wise top causes of deaths: this is an interactive graph implemented using plotly and in this graph deaths due to top 5 causes are grouped by years. In this graph we can select different causes of deaths to compare them. In the graph below, I have selected only cancer and heart diseases to compare. Plotly provides this feature to select different legends. Moreover, when we hover over the bar we also get the information about the observed deaths, year and the cause of deaths. This interactivity makes the graphs more interesting for users and makes it possible for using them in web apps. I have also implemented the dashboard to present all the interactivity as I have shown in the presentation.

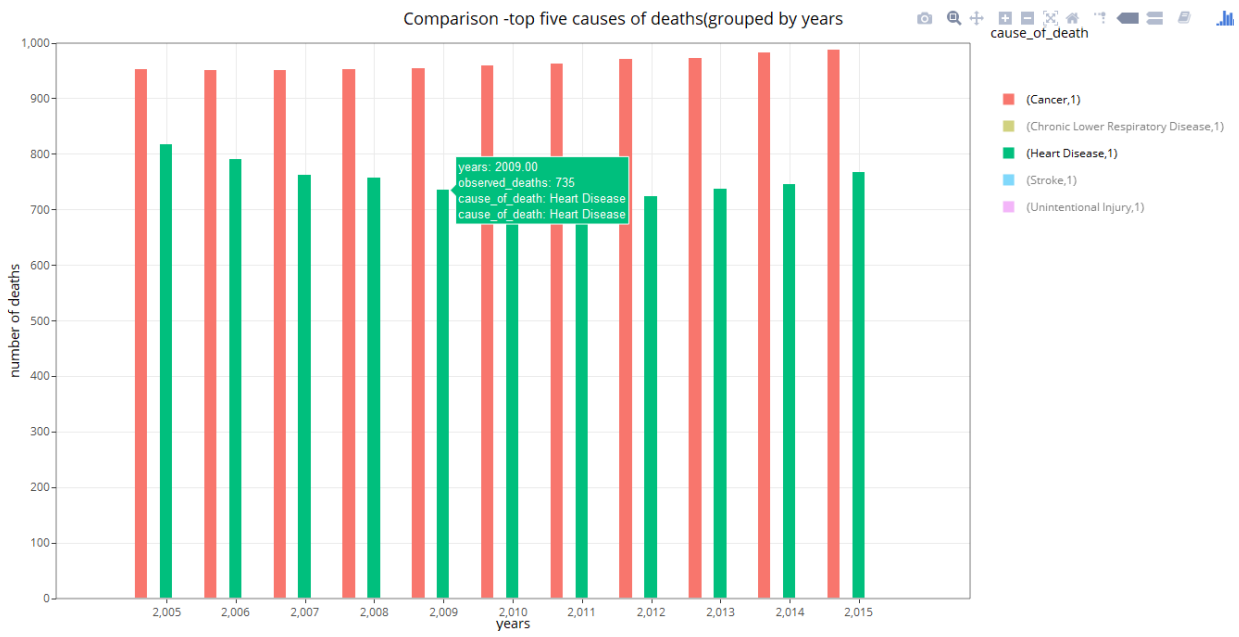
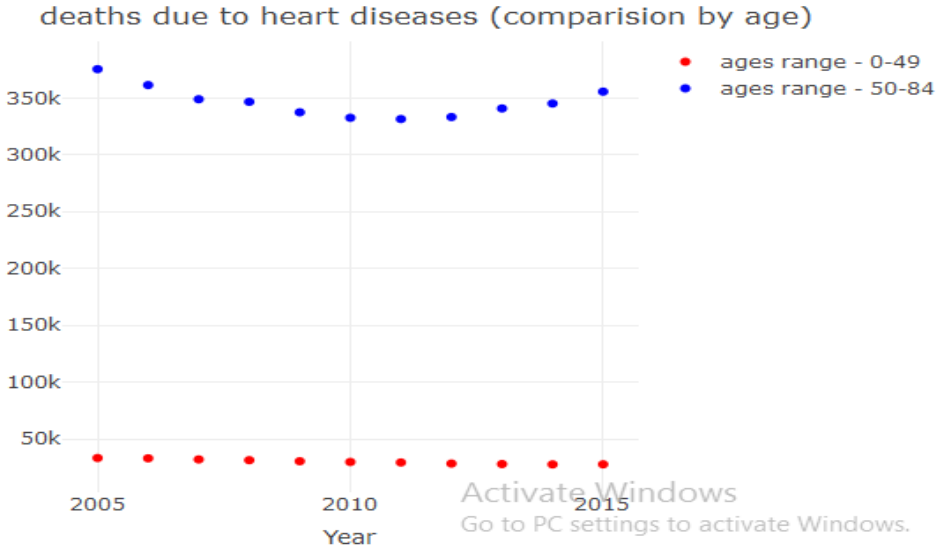


Fig 5: comparison- year wise top causes of deaths

d. Age standardized death analysis: I have made this plot to show the number of deaths by categorizing data according to the ages. The actual data doesn't have the data for the ages 50-84. I had to calculate this required data from the available data in the data set. This plot shows the number of deaths in y-axis and the years in the x axis. From this graph we can see that for the ages 0-49 the number of deaths is decreasing which is good thing. And only the number of deaths of ageing population (50-84) has the same pattern we have observed in fig5. Which is a normal thing for population in old age and we need not worry a lot about it. From this we can say that younger population is taking measures to avoid heart diseases.



d. A Micromap to look at the difference between observed and expected deaths (2015)- NCHS (national center for health statistics) predicts the number of deaths every year to take measures in futures. I have implemented this Micromap to show potential difference in observed and expected deaths for all the states in the country. The first column in the Micromap shows the observed deaths and the second column shows the expected deaths and the third column with the arrows show the difference between observed and expected deaths. Here, by looking at the third column we observe that Texas has the highest difference in expected to observed deaths. And we can also observe that the number of observed deaths is always higher than the values predicted are always lower than the observed values.

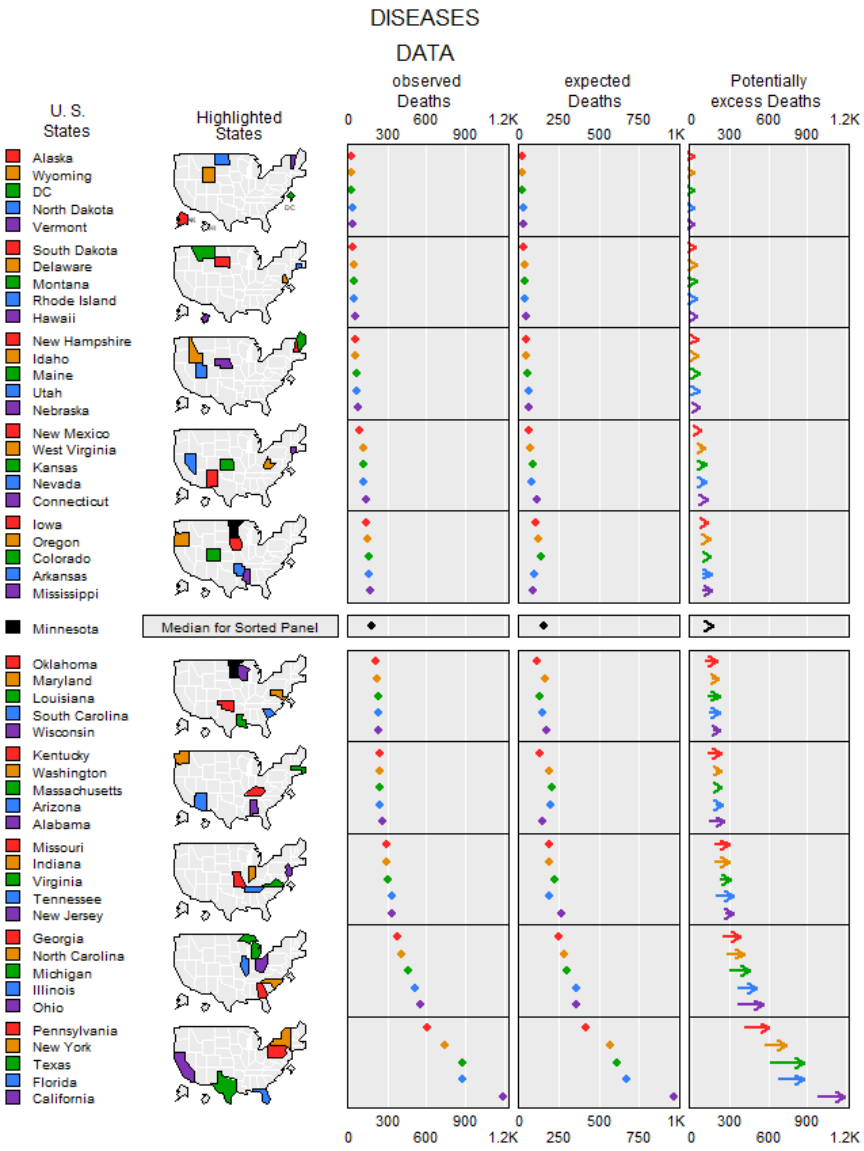


Fig 6: micromap

e. Population vs deaths: this plot is implemented to show there is relationship between the population of each state and the number of deaths. This graph shows the implementation of linear regression. This scatterplot with smooth clearly shows that as the population is increasing the number of deaths is also increasing. By looking at this graph we can clearly say that the number of deaths in a state is dependent on the population. I have highlighted the states with highest deaths. This has motivated me to make a graph with **population vs deaths per hundred thousand.**

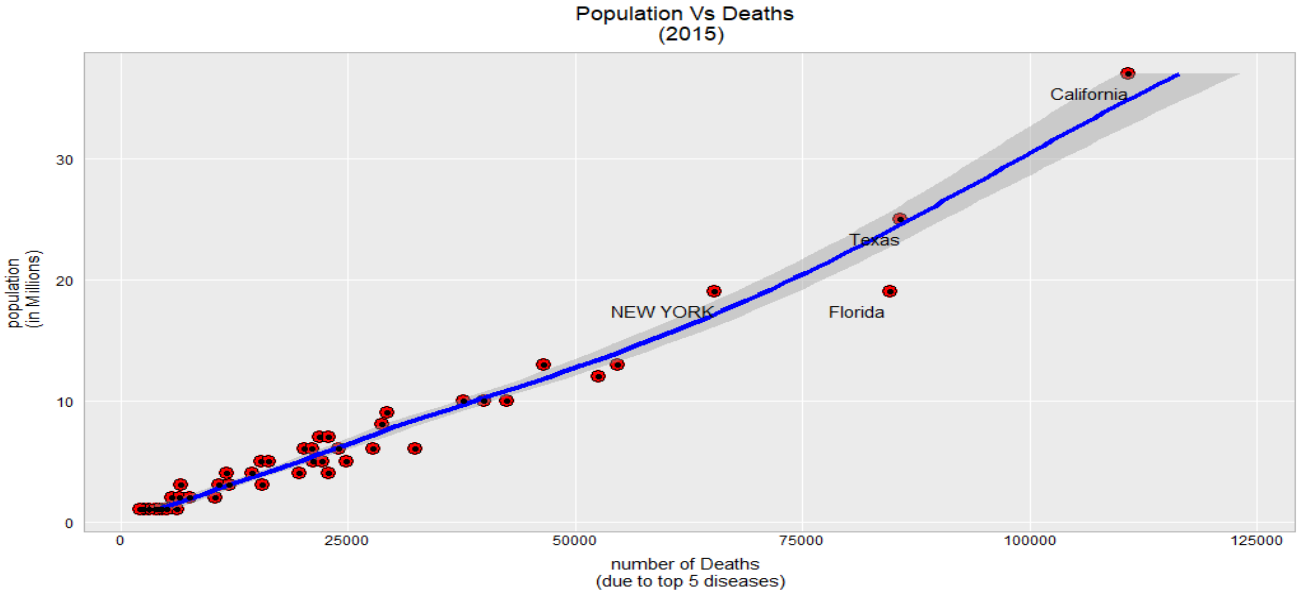


Fig 7: Population vs deaths

f. Population vs Deaths/100,000: I have made this plot to show more detailed analysis with the number of deaths per hundred thousand doesn't increase with the population. Here I have highlighted the states which had the highest number of deaths in **fig7** along with the states with more number of deaths per hundred thousand. Here we can clearly see that death per hundred thousand gives more useful information than the plot in **fig7**. Though States like California, Texas had high deaths from the previous graphs. We need worry about the states like Maine, Tennessee, Kentucky and Mississippi as they have more deaths per hundred thousand as we can see from the graph below.

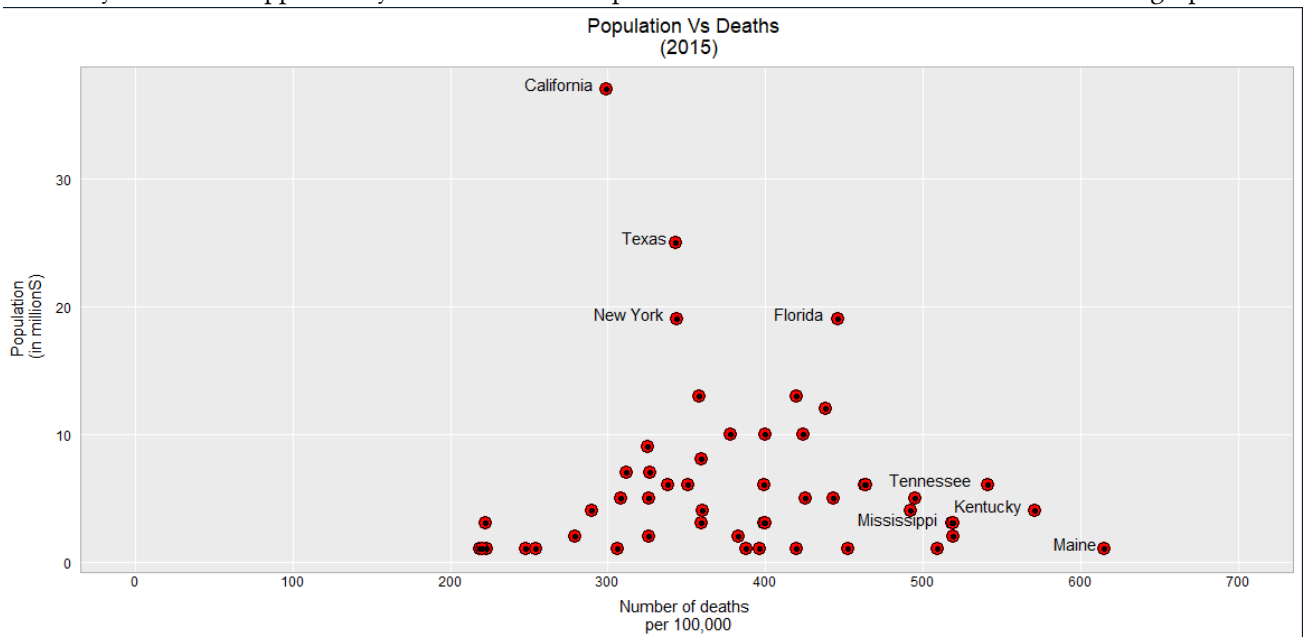
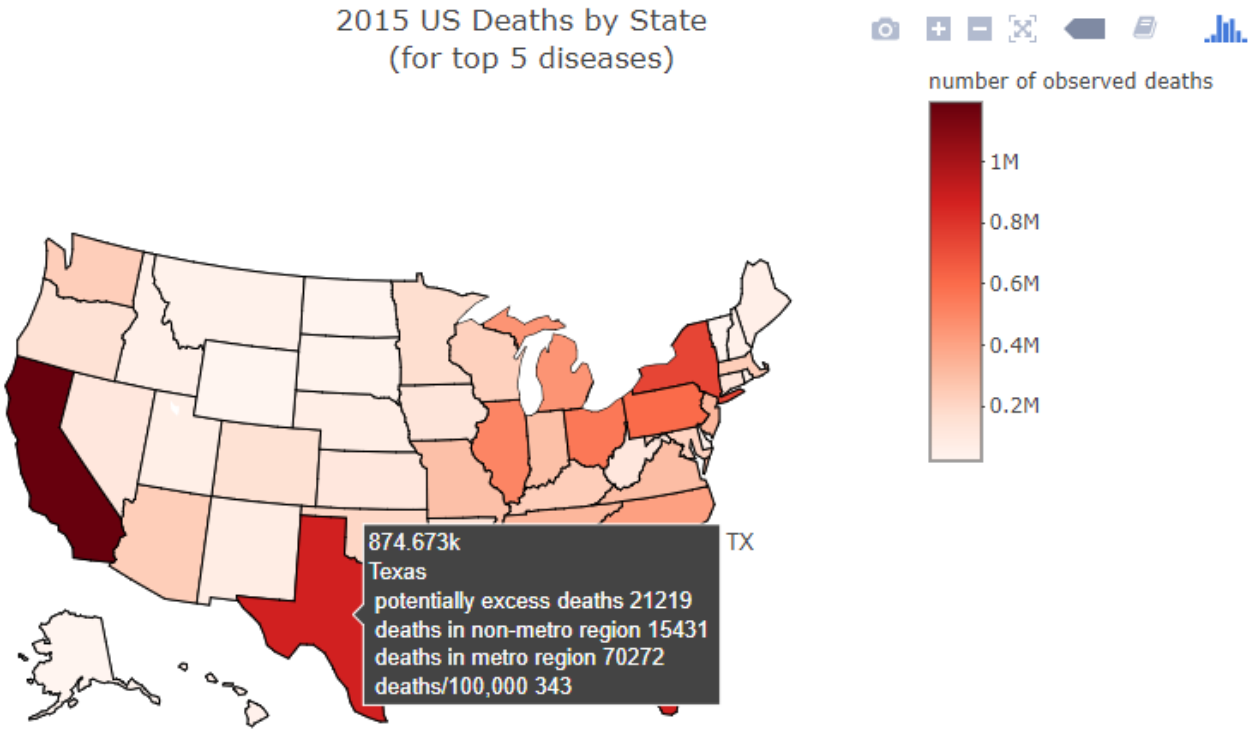


Fig 8: Population vs Deaths/100,000

g. Interactive summary graph using choropleth: I have implemented this summary map to make it interesting and add interactivity to the map to valuable information of deaths due to diseases about every state. This map is colored according to the number of observed deaths in the state.



Conclusions

We have seen that the maximum number of deaths from the year 2005 to 2015 are observed in highly populated states like California, Texas and Florida. Though these states have more number of deaths, states like Maine, Mississippi and Kentucky with low population have higher number of deaths per hundred thousand. This shows that states with more number of deaths per hundred thousand are in more need of attention. I was able to find the top 5 reason for death by using k means clustering. And was able to find the pattern in deaths due to these top diseases

References:

- https://www.cdc.gov/pcd/issues/2016/16_0211.htm
- https://stackoverflow.com/
- https://cran.r-project.org/web/packages/micromap/index.html
- R for everyone (Book by Jared P. Lander)
- R for dummies (Joris mays)
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html