Data Visualization

Mid Term Project Report

Radhika Yogesh Kulkarni
Email: rykulkar@iu.edu

FA17-BL-INFO-I590-14120

Project Title: Analysis and Visualization of Mortality With Disease Names Across The Globe

Team : I will be working solely on this project starting from Mortality data downloading from WHO site ,Data cleaning, Analysis ,Visualization and Reporting.

Abstract :
The research takes Mortality data set downloaded from WHO site through a set of steps that could be seen as a manually executed data pipeline. Analysis is performed on a subset of mortality data from year 1980 to 2014 across all the Countries listed along with the exact causes of mortality and countries where rate is reduced now . This analysis will definitely help in understanding the trend and diseases causing mortality all over the world and deciding the next steps in healthcare development and drug discoveries to reduce this death rate.

Introduction:
The project analyzes the mortality rate across the globe along with the disease causing it. This analysis will definitely be helpful towards finding out the major diseases causing mortalities and possible solutions of introducing new drugs to prevent it .
The analysis steps includes the processing of raw data from WHO datasets and further Cleaning, Analysis and Visualization across globe along with disease the information. This project will highlight the most critical set of diseases in account which has the adverse effects on health.

Research questions and working hypotheses:
Project problem is very important because it gives insights of the most critical disease which are causing the mortalities and what we could do to prevent it or reduce it across the globe. This data will definitely help on different levels of research like drug discovery, factors causing these diseases, regions where the rate is observed more.

The project will try to find the answers for few of the below questions based on Analysis with some data explanation.

- Major factors contributing to Mortality
- How would this analysis help in reducing mortality
- Are there any specific regions where Mortality rate found is more across timeline.

Background And Related Work:

The analysis of mortality data provides an opportunity for developing preventive strategies to improve this indicator of a population's health. All deaths in North Carolina during a 5-year period (1980 through 1984) were analyzed using the International Classification of Diseases, 9th revision (ICD-9), and a system for linked birth and death records that allows the analysis of birth certificate information on deaths. Causes of death were aggregated based on common etiology such as prematurity or obstetric-related conditions rather than the more traditional organ system taxonomy of the ICD-9 codes.

The Effective visualization part was missing drastically on this vast data by just giving conclusions based on the data with very little visualization using only Line chart or PieChart graphs.

Willing to work on similar lines ,I would prefer to take ICD-10 dataset (Last updated: 29 March 2017) for further analysis. Subset will focus on only Morality for certain period of time for which complete data is available and for all the countries.

This project Analyses Mortality Rate Across Countries with below important data Plotted in Visualizations.

- Location of Mortalities (States)
- Years in which Mortalities has occurred.
- Diseases Causing Mortality
- States which had the largest reduction in mortality rate.

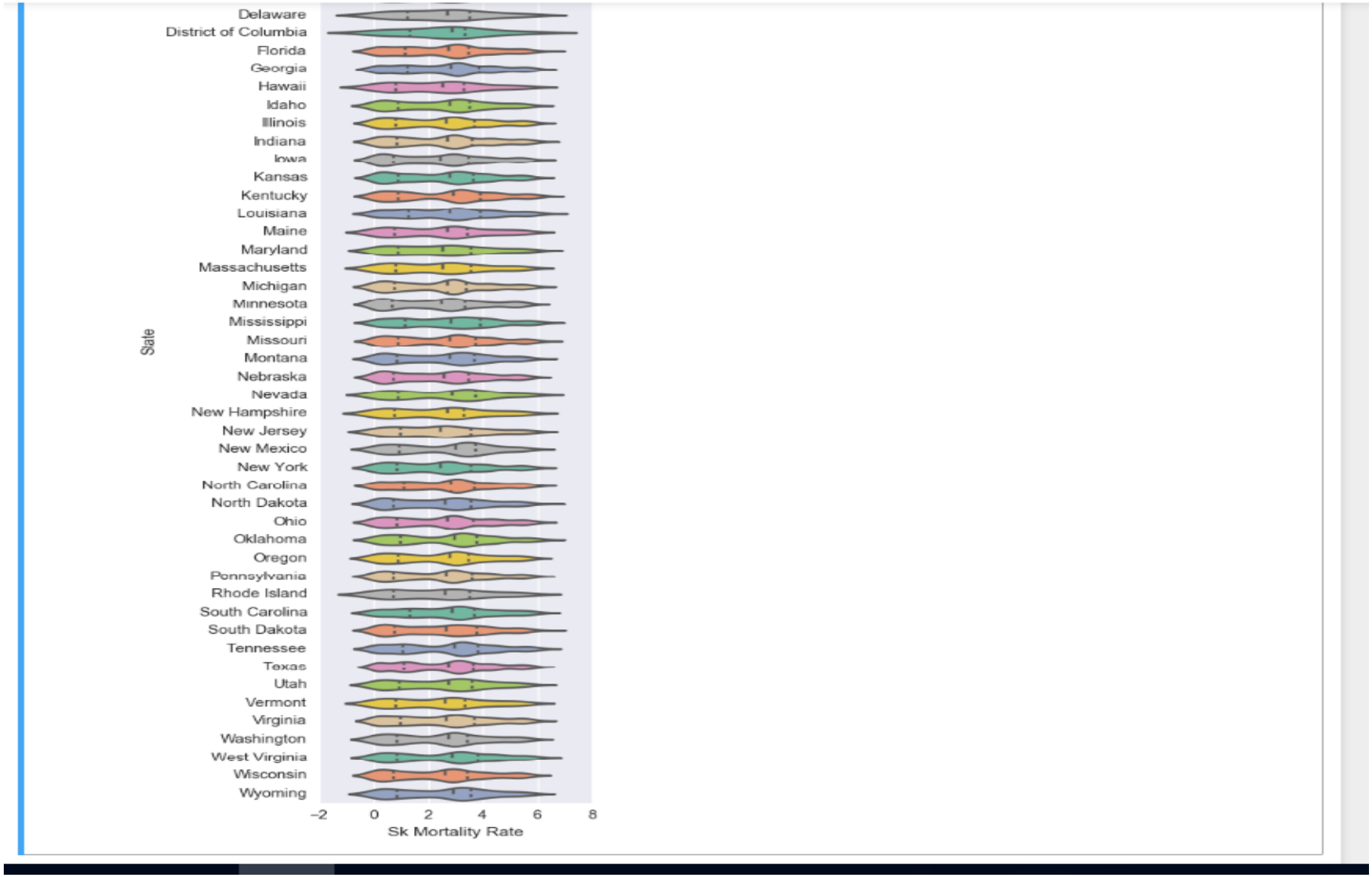
Visualization Techniques used using Seaborn, Matplotlib and Panda Libraries of Python

- Data Grid
- Heatmap
- Violin Plot
- KDE
- Pivot table
- Pair Grid

Visualization and Analysis using Python :

This is the most important step to analyze and visualize the data using different tools like Python Libraries

A) Violin plot using log to visualize States Vs Mortality Rate



B) Finding Correlation matrix -To check at mortality rates correlations .The matrix compares variables- To pivot the data to look at Category as the columns and Cleaning up unwanted columns further.

Out[12]:

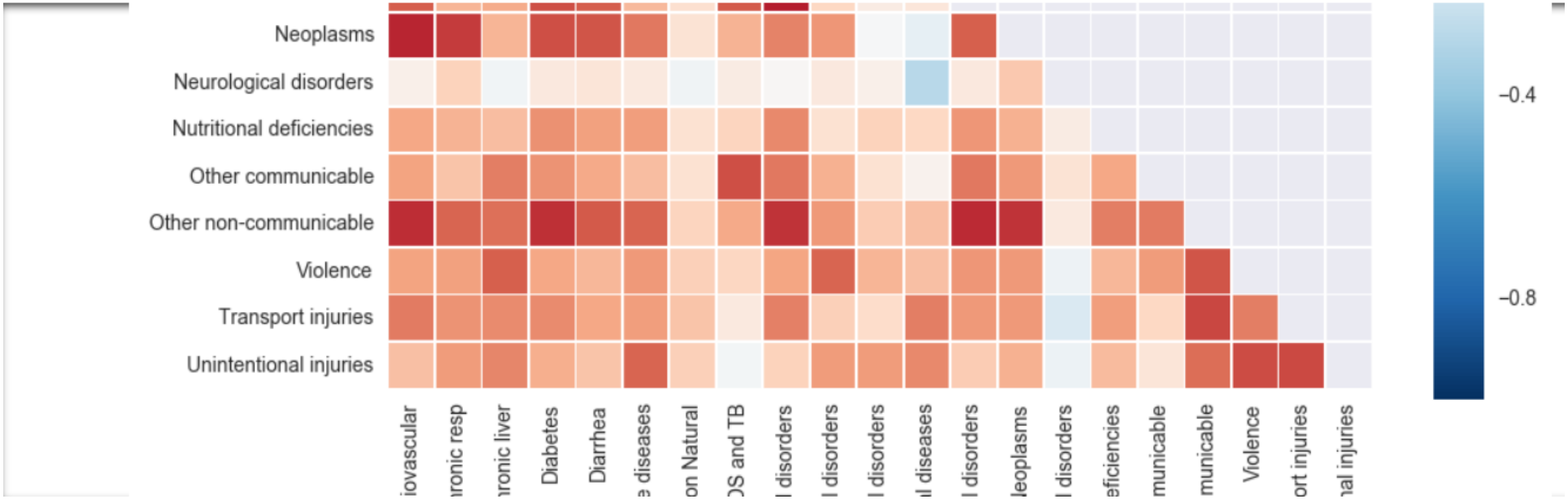
County	Year	State	Cardiovascular	Chronic resp	Chronic liver	Diabetes	Diarrhea	Digestive diseases	Non Natural	HIV/AIDS and TB	Maternal disorders	Mental disorders	...	Tropical diseases	Neonatal disorders
Abbeville County	2014.0	South Carolina	252.42	59.27	20.16	57.49	30.22	15.86	0.06	1.96	0.52	11.14	...	0.05	5.11
Acadia Parish	2014.0	Louisiana	363.23	56.09	19.36	68.58	48.68	14.86	0.09	3.18	0.68	13.62	...	0.09	4.59
Accomack County	2014.0	Virginia	272.88	60.31	17.89	69.63	39.83	14.09	0.08	3.24	0.50	13.53	...	0.05	4.74
Ada County	2014.0	Idaho	211.01	54.67	15.20	42.20	17.79	15.07	0.03	0.56	0.26	10.58	...	0.06	2.15
Adair County	2014.0	Iowa	270.87	47.06	12.03	51.62	46.29	17.04	0.05	0.38	0.27	6.46	...	0.06	2.07

5 rows × 21 columns

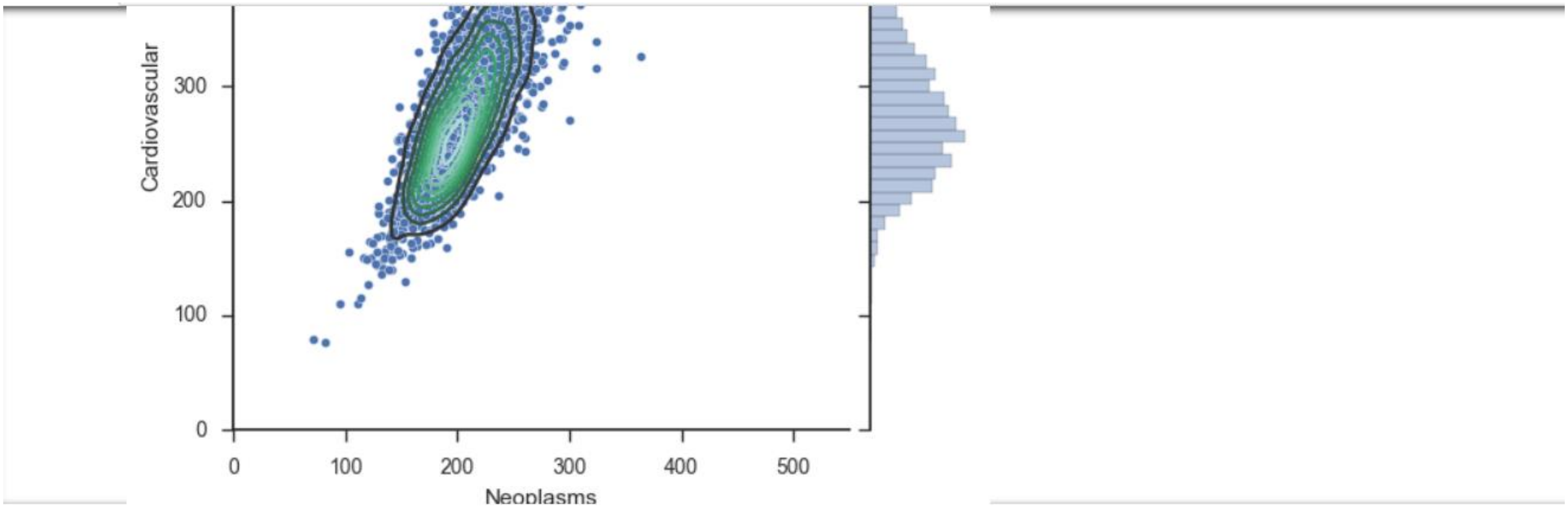
C) Plotting Heatmap of Disease Vs Causes.

Analysis and Observations :

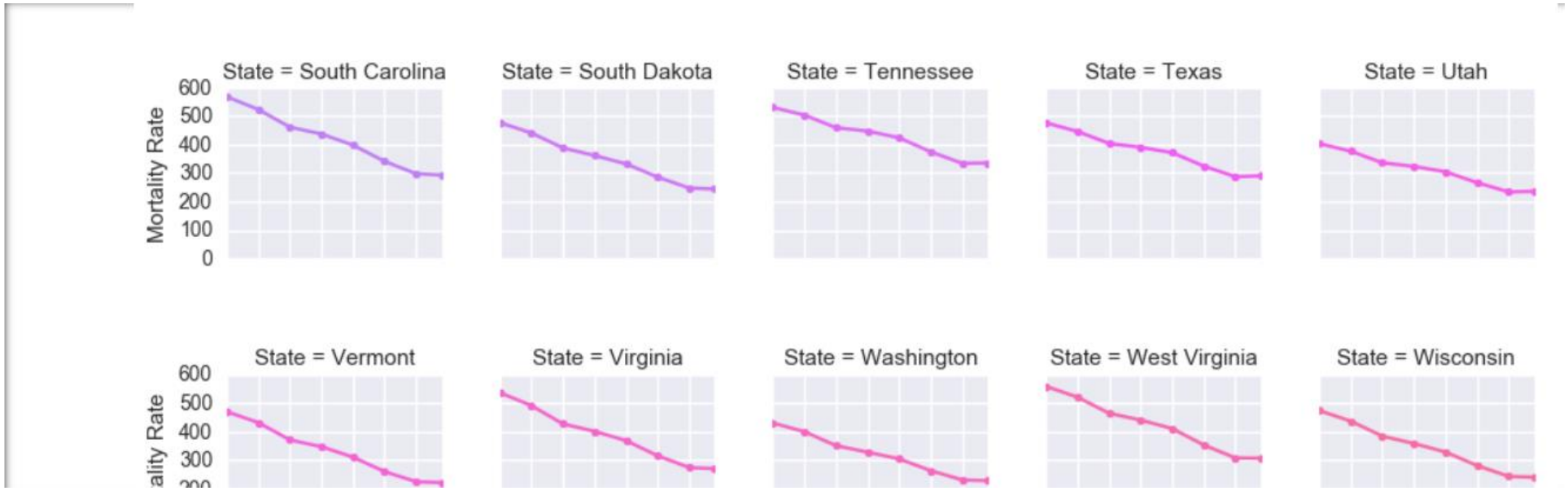
- **Neoplasms and Cardiovascular disease are very highly correlated which are in the highest range of cause.**
- To better the concentration of points in the center ,making use of a KDE and bar graphs on the edges



D) Zooming back out from those two diseases, goal is to find the relationships between other diseases.
Because the top and bottom corners are mirror images, changed the bottom to a kde plot to get a better understanding of the density of values in the concentrations.



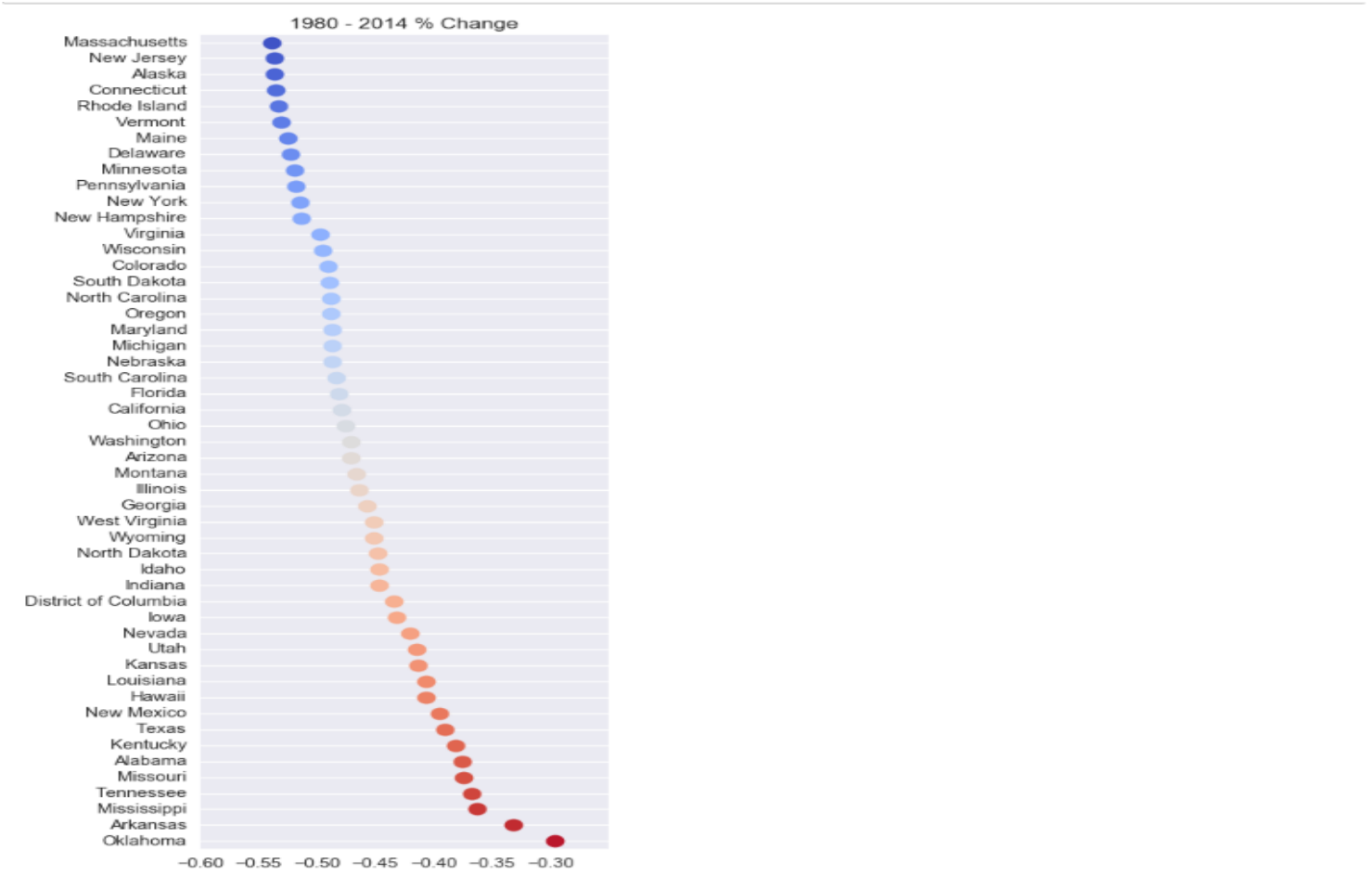
E) Year over year change in mortality rates of cardiovascular disease. We can see that mortality rates are dropping in every state.



F) The last chart showed us that there was a general decline in mortality rate, but it wasn't clear on how each state was doing relative to each other.
For this ,need to find the % change in mortality rates for cardiovascular disease from 1980 to 2014.

This addresses below questions:
1.Which state has had the largest reduction in mortality rate.
2.The lowest number -high drop in mortality.

Massachusetts is in the high drop in mortality.!!



The strengths and weaknesses of the project are internal factors, while opportunities and threats normally are a result of external factors playing their part.
I have listed down SWOT as a part of this project with explanation on each field.

Strengths	**Weaknesses**
• Improved Patient Life Expectancy • Greater Efficiency of data Science • Current Investment in data science technologies	➢ Lack of System Integration ➢ Information sharing Resistance
Opportunities	**Threats**
➢ Drug discovery ➢ Access to work site wellness	➢ Loss of Patient Trust ➢ Costs

Strengths:

- **Improved Patient Life Expectancy:** Improving Patient Life Expectancy is a primary objective of this project. The Mortality analysis can diagnose the problems and disease and better improve the life expectancy by adopting the preventable measures.
- **Greater Efficiency of Data Science**: Data Science has improved operational efficiency and increased productivity by reducing paperwork, automating routine processes, and eliminating waste and duplication and accuracy in analysis.
- **Current Investment in data science technologies** In the past ten years, advances in health information technologies have occurred at an unprecedented rate and healthcare organizations have responded by increasing their IT investments and data science technologies like Bgdata Hadoop,Python,different analytics tools like Python

Weakness:

- **Lack of System Integration:** Healthcare organizations data are not yet at this level of system integration.
- **Information sharing Resistance:** Even WHO dataset is not having full fledge data as hospitals, patients and health organizations are not willing to share the complete data due to concerns of privacy of information.

Opportunities:

- **Drug Discovery:** More drug discovery on diseases causing major mortality need to be done and have lot of opportunities of improvements and research.
- **Access to work site wellness:** To support healthy behaviors, access to work site wellness and health promotion programs can be an opportunity.

Threats:

- **Loss of Patient Trust:** On current treatments on the diseases causing mortalities, patients might lose trusts which can cause patient to ignore the treatment at all.
- **Costs**: Cost involved in drug discoveries for critical illnesses can be tremendous and even the treatment cost is more so there are chances that patients with critical illnesses like Cancer can deny the treatment which in turn leads to more mortalities which can actually would have been prevented with proper treatment

Conclusion, and Future work:

From the analysis of this research ,I have observed the major results as :

- **Neoplasms and Cardiovascular disease are very highly correlated which are in the highest range of cause.**
- **Massachusetts is in the high drop in mortality over the few years when the trend is observed across the years.**
- **Year over year change in mortality rates of cardiovascular disease. We can see that mortality rates are dropping in every state.**

My future work will mainly focus on Infant mortality of certain age range than completely on whole data.
This will help to identify Causes /diseases and the world wise areas where the infant mortality rate is very high and how the technology can be improved to focus on the treatments or new drug discovery.

References:

Dataset: WHO Mortality Data base comprises deaths registered in national vital registration systems
http://www.who.int/healthinfo/statistics/mortality_rawdata/en/

Country Codes:
http://www.who.int/entity/healthinfo/statistics/country_codes.zip?ua=1

Data file containing the detailed mortality data:
http://www.who.int/entity/healthinfo/statistics/morticd9.zip?ua=1

Explanation of all Fields:
http://www.who.int/entity/healthinfo/statistics/documentation.zip?ua=1

Python:
https://docs.Python.com/manual/tutorial/install-Python-enterprise-on-red-hat/

Python libraries for graph plotting:
https://matplotlib.org/users/pyplot_tutorial.html