Analysis Report of Health Data:

Death due to Disease

PRAGNYA SRINIVASAN
psrinivasan@umassd.edu

ABSTRACT: There has been a lot of deaths in the past years because of various diseases- that probably didn't have any cure for initially- or maybe there was a cure, but it wasn't available for the common man. The datasets[1] which are analyzed, talk through the death count per disease per year in different states, or focuss a particular state. This paper walks through the comparison of these attributes for the states of US, and then focusses on a few states, which observed to have the highest number of deaths. It also gives statistical or visualized results to a few questions related to the data.

INTRODUCTION

Health related issues have been in the limelight for a long time, and one of the best ways to find a solution to any problem is by trying to find an answer through its past occurrences. This is called analysis of data. With the help of huge information present about the diseases, and which state it prevailed in, one can find out about the birth and characteristics of the disease- whether it is generic or not, and how the external environment plays role to ameliorate the disease or make it worse, whether the medicines were available, or if the medicines available eradicated the disease or just pacified it. A lot of information can be gathered by just rows and columns of data. This project would be one such experiment, to find out some information about deaths that occurred , particularly in New York, because of a few diseases. The dataset about deaths in New York City, have compared the deaths in New York per disease per ethnicity per year. Another dataset, that has information about the death count per state per year per disease in the USA.

DATA AND DESCRIPTION

This project involves analysis of the available health data sets for various states in the US, and finding out more detail about the major causes of death. The data set contains information about the list of disease that are a leading cause of death in New York City[1]. The set is in the form of comma separated values, which have the death count with respect to the disease, gender, Ethnicity , Year, and Percentage.

For an overall review, there is another dataset that is available, that contains the leading causes of death in the US[4], by State, year, cause of death, and the count. To find out in depth, about the disease, and whether they are in situ or genetic, a data set collected from Global Burden of Disease Study 2013[3] is used which contains information about deaths due to various diseases in different countries of the world. So the countries could be categorized according to the ethnicity i.e., each region has its list of countries associated.

STAGES OF ANALYSIS

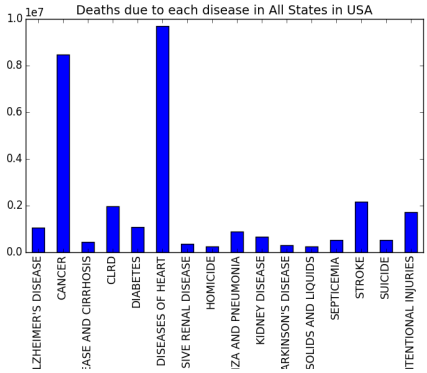
There are stages in which the analysis is carried out. Initially, the All States in USA dataset is analysed, which contains information about death due to disease in each state. This is a good dataset for comparison between states.

In the first stage, the data fetched from different source is cleaned - where the same disease names are stored in different way- there could be spelling differences, or the names of the disease could be different, or they could be stored in their abbreviated format. Regular expression search could be an option instead of cleaning, but these names are analysed quite often, so replacing the names to a common type was better. Also, the disease names can be easily confused even after using a regular expression. To replace the diseases with the common terms, the Python function replace is used. Cleaning also involved using the same column names for every different DataFrames. More about DataFrames will be explained in the later sections. After cleaning the data, to know more about its states death count, the data is grouped together with respect to their attributes- for example, to get the result of deaths in all states in all year, but individually for every disease, the count of death of every disease is added, so this automatically adds all the values irrespective of the year or the state. The result was visualized to see clearly which diseases affected USA the most. To take it to the next stage, to check if any diseases were common in all states, something similar can be done.

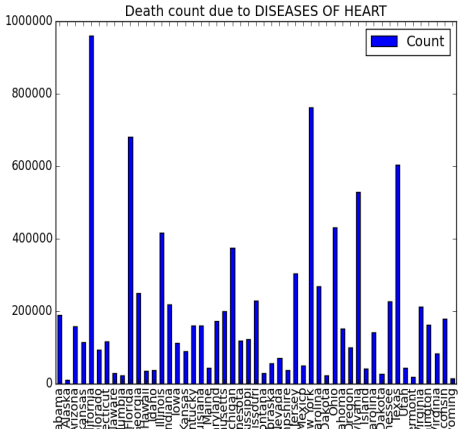
The first data that is visualized describes about the death count because of each disease in all the states in the US.

From the below image, one can conclude that most of the deaths in the US occurred because of Diseases of the heart which, according to this[4] data has 19383466 number of death. To check if there was just few states that had this rate for Disease of Heart, the data was visualized to find that California topped with 960725 deaths and next state was New York 763438 deaths. The figure describes that majority of states showed quite a lot of deaths due to this disease. This brought out the

question if Diseases of heart had such an impact in the total deaths in the US? Calculations showed that deaths due to diseases of heart accounted to 31.94% of total deaths in the US out of a total of 15 types of death causes(diseases).



Death count each disease



Death count cuz of diseases of heart

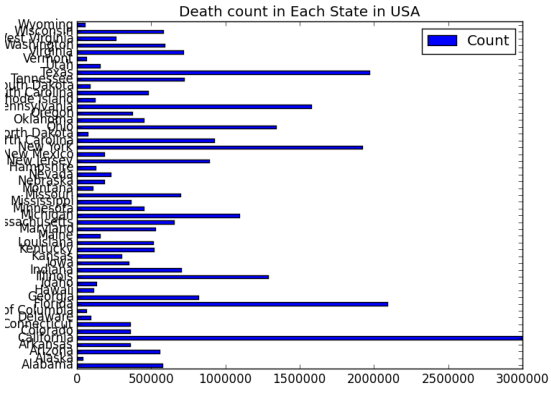
To now check for the total death counts in each state in the US, because of all the diseases, another plot is visualized. This plot shows that a few states have prominently large values of death rate when compared to others. California again, has the highest number of total deaths with a count of 2998036, followed by Florida with a count of 2090865 whicc is followed by Texas, with a count of 1970694.New York has a death count of 1923260 , even though it stand next to California in deaths due to Diseases of heart. The states with lower death counts include Vermont, District of Columbia, Wyoming and Alaska, ofcourse the population is comparatively lesser in these states.

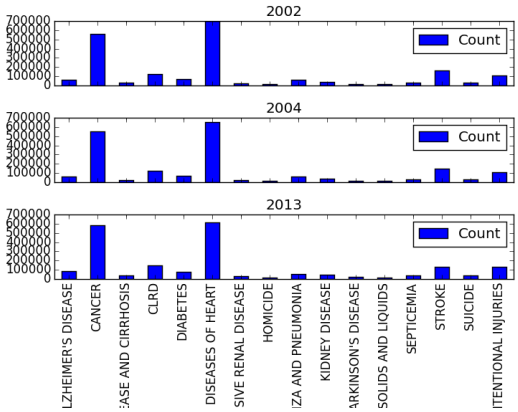
The next task was to compare the death count in the US in the growing years. A comparative analysis showed that the total deaths in the US kept fluctuating in years, and had a steep rise twice- once in 2002,and another in 2013. It also showed a drastic decrease in 2004. This resulted in comparison of deaths in 2002, 2013 and 2004 in all states per disease.

The below comparison shows that there isn't a drastic difference between the three years, but every disease cause have comparatively lesser death count. This could be a reason for the fall in death count.[3] According to US studies, there was a 3% decline in deaths due to accidents by 2004. Deaths due to alcohol also decreased. As of 2004, few states had met the Healthy People 2010 objective to reduce deaths due to accidents. Another study says that smoking had reduced a great deal between 2000 and 2004.[6]

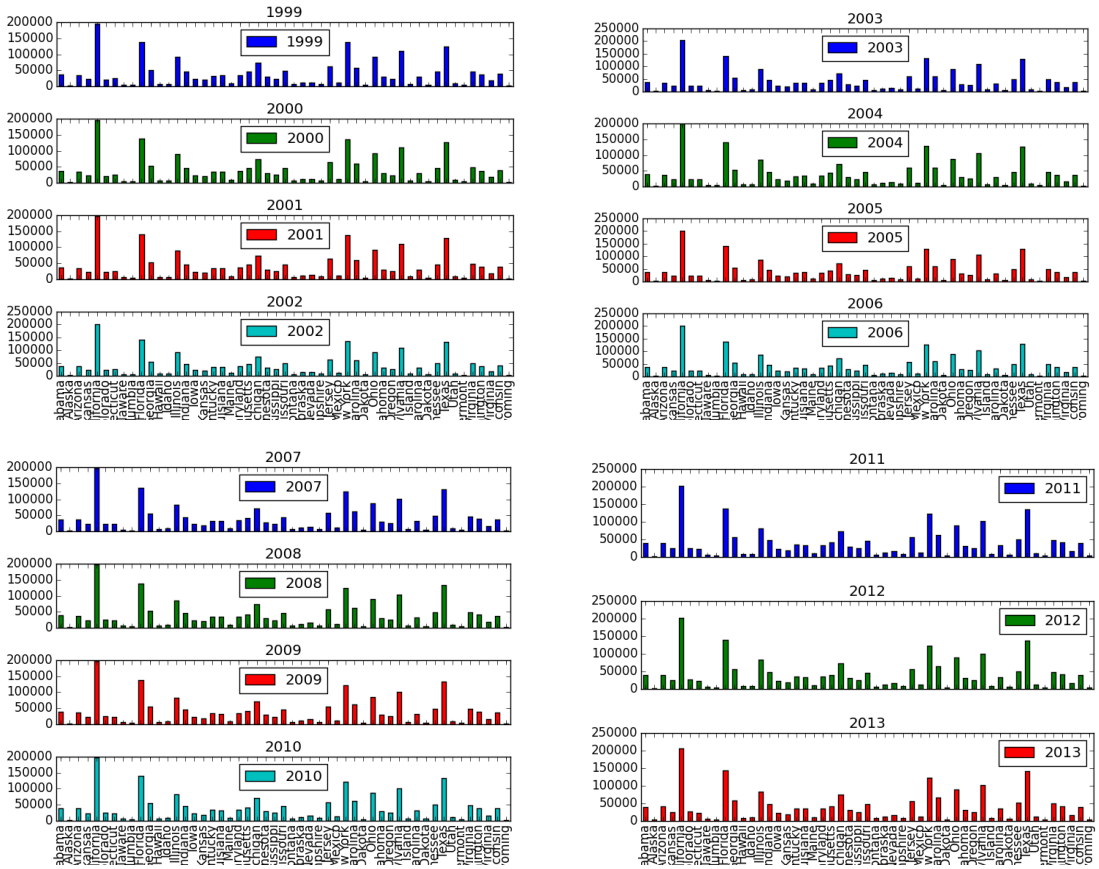
To explain in detail, the death count in each state per year is visualized. The graph shows a similar trend for states in all the years.

Again California tops the death count in all years. Looking at only California, it seemed to increase from 1999 to 2002 and go over 200000 in 2002, while it stayed between 210000 and 190000 between 2003 and 2013. In case of New York, it showed stable changed between 130000 and 120000, and gradually decreased to near 110000 by 2013.





a) All states per year. b)Comparison of deaths in 2002,2004,2013



Death count per State per year

These analyses resulted in another study to find the death rate of disease per year in particular states - especially for the highest death count states and the lowest. Which states with California. The diseases that caused major deaths were Diseases of Heart and Cancer both of which seemed to be pretty high. Though death due to heart disease showed a decreased over the years, Cancer seemed to have taken a lot of lives in California.

For New York according to the USA data[4] , the major diseases that caused deaths were again, Diseases of Heart, which gradually decreased over the year, and Cancer which showed very slow fluctuations.

Alaska, which showed the lowest death rate, on the other hand, displayed a higher death count due to cancer, than due to heart diseases. The rate of deaths due to cancer also looks like it is increasing over the year. Diseases due to heart, though a lower count than Cancer, still showed an increasing rate.

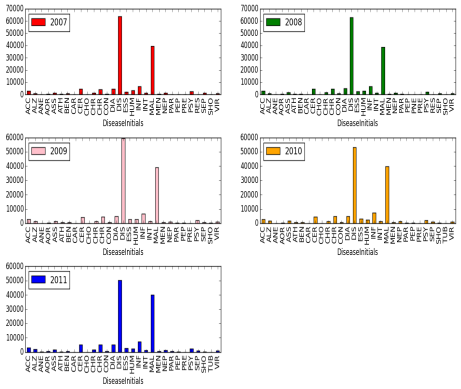
A detailed analysis of health aspects in New York City is studied with the help of the NYC dataset[1]. More explanations about it are given in the next section.

NEW YORK CITY

New York City, that showed the second highest death rate due to heart disease, is analysed to find out what is the reason behind it? Another dataset is used, which has details about death cause, Gender , Year and the Ethnicity. This can answer many questions like how prominent is the death and in which gender or origin of people etc.,

Year

The data is first analyzed to show the number of death due to disease in each year. This data set compared only years between 2007 and 2011. It plot shows that major deaths are due to diseases of the heart, which matches to the data plot that was plotted earlier with another dataset. This could be one of the criteria that tests if the analysis is right. This plot shows that the death rate due to Malignant Neoplasm - which is another term for Cancer, which again matches the previous dataset plot. Deaths due to both the disease seem to be decreasing over time.

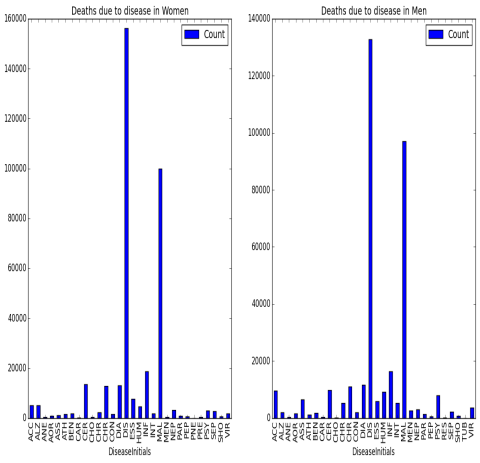


NYC Disease per Year

Men vs Women

With the dataset one can also compare to see the death rate between men and women for various diseases in NYC. The analysis shows that the ratio of deaths due to diseases of heart was more in women than in men.

On the other hand, cancerous deaths were more in men than in women. Why was this so?



NYC Disease Deaths Men vs Women

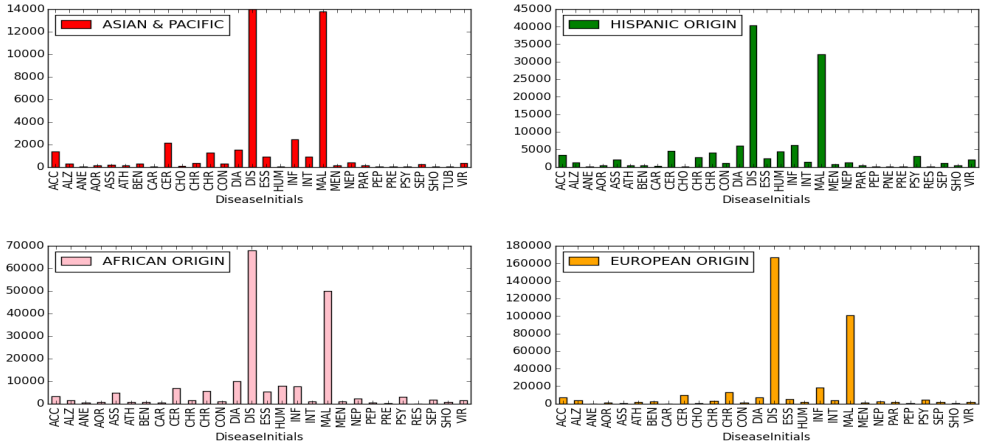
Was it because of Origin?

Another analysis was made to see the comparison of deaths in people according to their ethnicity. People of European Origin had the highest death rate, and those of Asian Origin had the least. The analysis explains that people of Asian Origin have high death rate of both Cancer and heart Diseases, while those of Hispanic origin have higher heart disease rate, but lower Cancer rate. African Origin have higher diseases of heart rate than Cancer. European origin have very high heart death rate, and comparatively lesser Cancer death rate

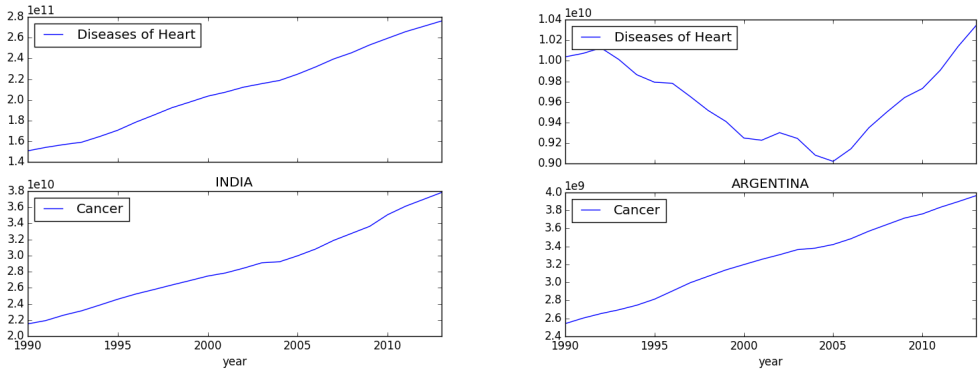
Does it really have anything to do with Ethnicity?

To test this, another dataset was analyzed, which was taken from [3] which had the deaths due to disease in every country of the world. A group of countries were studied for every ethnicity/origin and the death rate for heart diseases and cancer were analyzed to see if the deaths in New York had anything to do with the Ethnicity of the people.

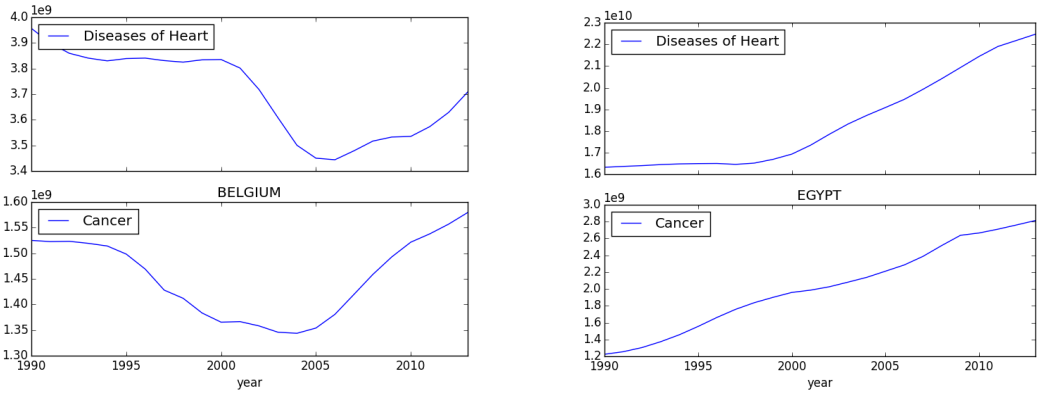
For example, to check Zimbabwe, Egypt Ethiopia, Sudan and Ghana data were analyzed for African origins; India, China, Singapore, Pakistan and Malaysia for Asian origin, Norway , Belgium, Sweden, Ireland, Netherland for European origin, Cuba , Argentina, Chile, Costa Rica etc for Hispanic Origin. The analysis show that all these countries with Asian origin have both Diseases of Heart and Cancer rate increasing over year. For Hispanic origin ,Diseases of Heart seemed to decrease and then increase over the year for most of the countries, and an increasing Cancer death rate. For European origin, diseases of heart seemed to decrease over the year, and then increase , but for Ireland alone cancer deaths decreased and then increased. For African Origin both Heart disease deaths and Cancer deaths seemed to increase over the year, except for Ghana and Zimbabwe where cancer rate seemed to decrease and then increase.



NYC Disease Death per Ethnicity



Death Rate due to Disease of Heart and Cancer in a)Asian Origin b)Hispanic Origin



Death Rate due to Disease of Heart and Cancer of a)European Origin b)African Origin

LANGUAGE AND LIBRARY

To do this analysis, Python library called Pandas[5] was used. Pandas is an easy and effective tool for analysis, which supports a lot of libraries for statistics along with NumPy and SciPy. To accelerate the multiple call of methods, multiprocessing thread library was used, which reduced the execution time by a great deal. Another functionality called pivot was used to transform the columns to required type. The figures are plotted using matplotlib. Matplot also has a functionality to plot multiple plots (subplots) with same x or y axis, which was an efficient way than having three or four different figures of the same plot. Regular expression was using to find out all the cancer deaths in each country in the world, because there were many causes due to different types of Cancer .

QUESTIONS AND RESULTS

The main purpose of analysis of the data is to be able to answer any questions imposed in the data, and try to get a solution to it through the visualization. Here are a few questions, but the analysis is not limited to this.

Does gender play any role

One major question regarding deaths due to disease in New York was if Gender played any role, or if, were there any diseases that showed more deaths in men than women or vice versa? From **Figure** it is seen that death due to Disease of Heart is more in women than in men, while cancer was more prominent in men than in women- again it is just mentioned as cancer, we do not know what type of cancer it is. Psych. Substance use and accidental drug poisoning showed more deaths in men, probably because men were exposed to drugs more than women. Assault (Homicide), accidents other than drug poisoning were also more in men than in women. This could also be because of the craze for high speed driving among men, or probably deaths due to accident accidents. Population also plays a role in the death ratio, the ratio of men is greater than that of women.

Ethnicity

The images show that Ethnicity might not be the cause of deaths due to diseases of the heart, but it could be the case for cancerous deaths. As a matter of fact few studies say that cancer could be hereditary. So there could be a chance that Cancerous deaths in New York could be because of the ethnicity of people. On the other hand, if we look at diseases of the heart, it does look like most of the deaths in other countries of the world has a **higher rate because of diseases of the heart.** But looking at other states in the US, it does look like Cardiovascular disease is the leading cause of death even

for the people of the US (i.e., not of other origin). Death due to heart disease was highest in the year 2007, according to the New York data[1] which decreased over the year. This decrease could be because of better availability of medical aid. On the other hand, cancerous deaths in NYC showed a slow increase over the year. This could be because of increase in pollution too, as an external factor, other than the genetic factor.

Is it common in more than one States for that particular Year

From the previous plot it is understood that 2007 is the year with maximum deaths due to Diseases of the heart in NYC. To check if in the Year 207, are there other states in the US that are affected because of Heart Disease. The analysis shows that next to California, New York shows the maximum deaths, followed by Florida and Texas. The pattern is the same of all diseases, and for Disease of Heart, hence death due to cardiovascular problems account to a large portion of total deaths.

What were the problems and how were they solved?

The initial problems faced were to match the disease names. For example Cancer was also named as Malignant Neoplasm in one dataset, and Stomach Cancer Breast Cancer etc in another dataset. To clean them all to be able to be recognized as a same disease was the main task. Data from Global Burden of Disease Study 2013 had different types of cancer, so had to use regex and sum up all cancer deaths. This has to be done to get th desired result. Another major problem was during the pivoting of the data-frame. It first created different columns for each year which contained the death rate, but the graph was too congested. The columns were then grouped together according to years, for example the death count between 1999 and 2002 could be plot together, to show the comparison. Another problem was with the world wide data, where the mean value for a few rows were zero, for the metric 'rate per hundred thousands'. Therefore the metric 'number' had to be used, which plots the mean value, hence the exact rate is unknown. But these data sets are used just to see if origin of people plays any role in the death. Also initially the data sets of diseases in California and deaths due to diseases in California were proposed to be used, but these data after careful analysis wasn't of any help. Hence these data sets were ignored.

FUTURE WORK AND CONCLUSION

Thus with the help of these available data sets, some information about deaths in New York could be obtained based on the particular year, gender and origin. Future work includes doing detailed analysis of all diseases in New York and if the diseases have any effect due to country of origin. The Visualization could be made user-interactive by adding bokeh widgets and its functionalities. More information can be fetched about the services provided for the diseases by collecting health data about the hospitals in New York that provide cure for the diseases. The initial proposal included a choropleth map to plot all the states of USA, that could be implemented.

REFERENCE

[1]https://catalog.data.gov/dataset/new-york-city-leading-causes-of-death-ce97f

[2]Global Burden of Disease Study 2013

[3]]http://www.nhtsa.gov/About+NHTSA/Press+Releases/2014/traffic-deaths-decline-in-2013

[4]http://blogs.cdc.gov/nchs-data-visualization/2015/06/01/leading-causes-of-death/

[5]http://pandas.pydata.org/

[6]http://www.cdc.gov/nchs/data/nvsr/nvsr55/nvsr55_19.pdf