UNIVERSITY of
BRADFORD
School of Electrical
Engineering &
Computer Science

Visualization and Analysis of Health Condition against Public Health Awareness

BIG DATA VISUALIZATION - SMAN

Introduction and Dataset

Advancement of information technology has helped the humanity in various areas by solving the complex computational problems. Like any other field, these computational methodologies have assisted health care practitioners too by providing systems to help them taking wise decisions. From the past few years, scholars have started to work in predictive analysis in which various techniques are used to predict disease risks and for formulating a better health plan for individuals (Harris, May and Vargas, 2016).

A similar effort is made in United States by initiating Behavioural Risk Factor Surveillance System (BRFSS), which conducted telephonic surveys to collect data about health-related habits of people and their health conditions. In this report, BRFSS's data of five years, from 2011 to 2015, is used for visualization and to extract useful information from it.

In the telephonic survey to collect data, apart from asking people about how they feel their health is, information about any chronic disease and factors affecting individual's health were collected. A subsection of survey was also related to health awareness in which people were asked questions related to eating, and drinking habits as well as how frequently they tested their cholesterol and blood pressure level. The purpose was to see how much they are aware of the things, which are essential for maintaining a healthy lifestyle.

Question and Method of Analysis

As mentioned, advancement of machine learning, data mining and data visualization technologies have opened new doors for exciting area of predictive analytics, which combines all these methodologies to make predictions about unknown future events (Amarasingham et al., 2014). But, before predicting it is necessary, and relatively easier to find trends of past event and analyse the correlation between different attributes influencing the result. Therefore, this report will be answering the question that **do the people of United States getting more aware of healthy habits by the passage of time and is this awareness improving the health of people?**

First step in Analysis Methodology – Data Collection

To answer this question, the first step was to collect the data from the official website of BRFSS (Behavioural Risk Factor Surveillance System, 2019). After collecting the data in .csv format, it was observed that the column names in the given data were not self-explanatory so official BRFSS website was consulted to find more documentation. BRFSS website contained code book of every year explaining each and every data variable in detail (CDC, 2016). Therefore, these codebooks were used to understand the data better.

Second Step – Data Cleaning

There were five main data files of the collected data of each year from 2011 to 2015. The data in its initial form was enormous which can be seen from the following table 1 describing the number of columns in every year's data:

Table 1

Year	Number of Columns in the Data
2011	454
2012	359
2013	336
2014	279
2015	330

As the main purpose of the report was to answer the above-mentioned question, so a data cleaning process was done to extract only those attributes which can be helpful in answering this question. Codebook of every year's data was helpful in this phase and following are the key attributes of the data being used:

General Health of People, Hypertension Awareness, Cholesterol Awareness, Chronic Health Conditions, Demographics, Other health-related habits (Smoking, Alcohol consumption, Eating fruits, Exercise, Seatbelt Use), HIV/AIDS

Step Three – Preparing the summary of Data

As the data was constructed in such a way that each entry in the data corresponds to the responses given by one person, so it was hard to visualize data in that form. Therefore, the final step before visualizing the data was to generate summarized table answering 'How many' questions. For instance, following summarized table (table 2) was generated using `count()` function in R language from general health responses of people in 2011 to answer 'How many people rated their health condition as excellent, very good, good, fair, poor, or not sure?'.

Table 2

Health Condition	General Health 2015
Excellent	76032
Very Good	145065
Good	136975
Fair	58962
Poor	23175
Not sure	799
Refused	446
No Data	2

This kind of summarized tables were generated for all the important data attributes before start visualizing the results and analysing them.

Visualizations and Results

As the goal was to see the data trends for awareness in people regarding their health and their health condition, the visualizations were started from bar graphs to plot the response when people were asked about their general health by themselves. The data was of five years and figure 1 provides an overview of their answers.

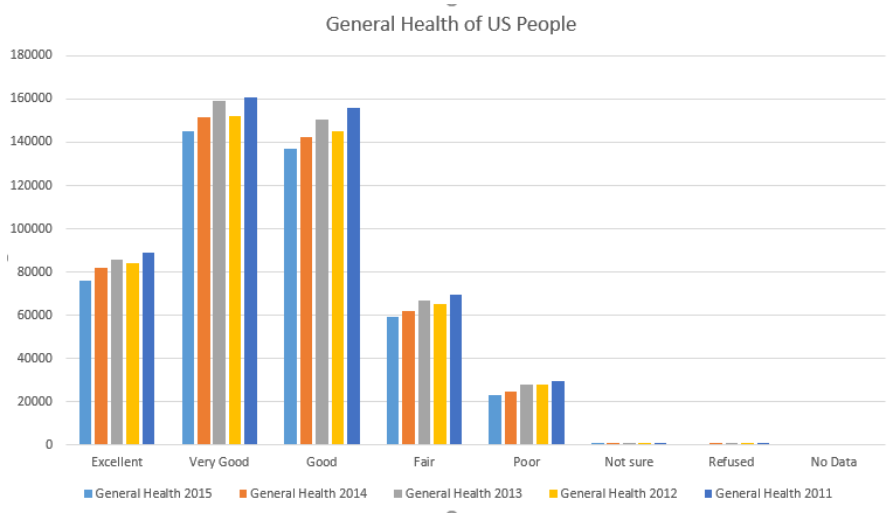

Figure 1

It was observed from the bar plots that there were less number of responses for year 2015 than year 2011 in each of the category from excellent to poor health. This was an interesting observation as the difference in these two years' data was even more clear in categories 'Very Good' and 'Good' health. Therefore, without proceeding any further, it was decided to see the percentages of responses in each category for every year so that we can get a better view of health changes against time. With that in mind, pie-charts for the same data were made to get a holistic view of the data as shown in figure 2.

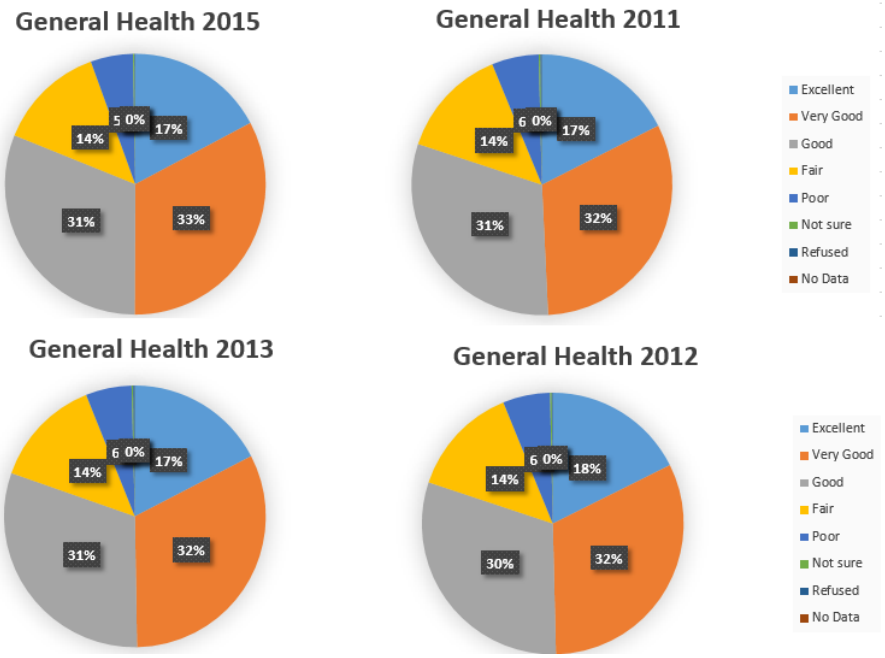

Figure 2

Very interestingly, pie charts gave a changed view of the same data. It was clear from the pie-charts that the number of people with excellent health in both the years, 2011 and 2015, were the same. But, it was seen in the pie-charts that 32% of responses in 2011 were as good health whereas there were 33% such responses for year 2015.

Then moving forward to answer question, we summarized all the responses we had from people about whether they are suffering from asthma, heart diseases or any kind of cancer. On the other side, to visualize the health awareness of people, some attributes were selected for visualization including whether they had undergone medical tests on their own and they do exercise or not. Figure 3 shows the diseases' situation in these years however figure 4 gives the information of public awareness for their health.

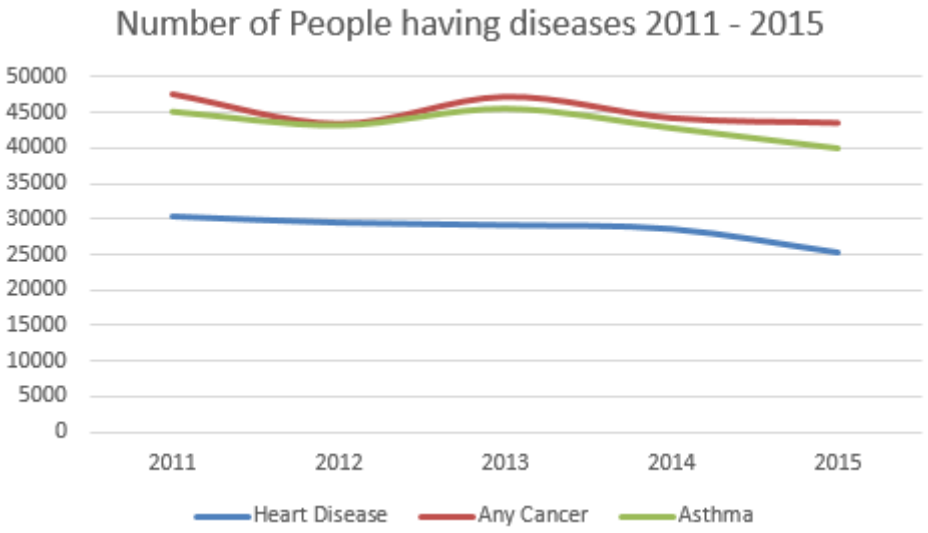

Figure 3

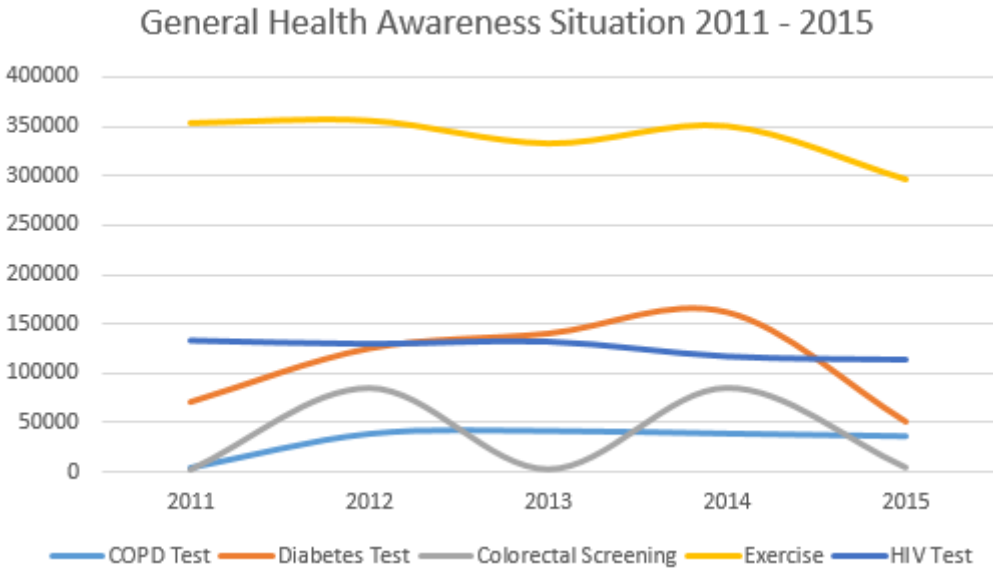

Figure 4

It can be seen from the figures 4 and 5 that there is no persistent trends of either diseases or awareness parameters in these years other than Heart diseases and COPD test screening which have seen a constant decrease and increase respectively. But the remaining attributes were not giving a clear picture of the data so it was decided to use multivariable bar-graphs to see the trends of awareness attributes against the diseases which is shown in figure 5.

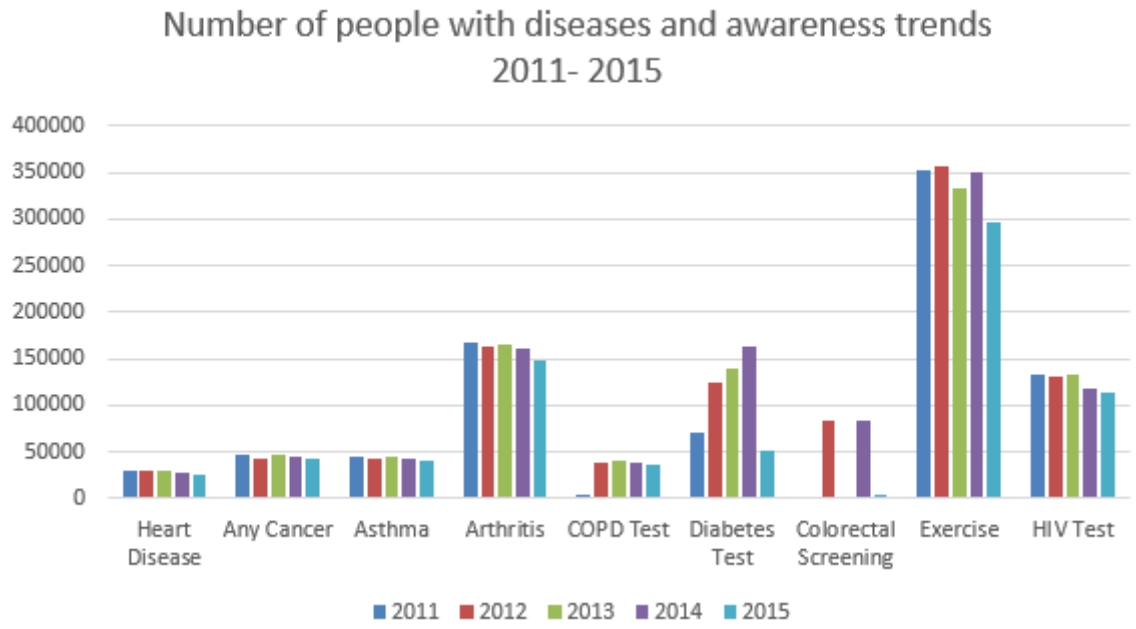

Figure 5

It can be seen from the figure 5 that though the number of people suffering from any of the given diseases has seen overall decrease over the course of five years but it cannot be directly mapped to any of the awareness parameter selected from the data. But, despite the possibility of less number of data objects in the year 2015, it can be concluded that health awareness in people is generally decreasing in the aspect of being proactive for health screening to know about any health disorder. Though, this decrease, as per the data we have, has not yet fully affected the health of people. So, the question asked at the start of this report is successfully answered by the visualization of this data.

Discussions

As correctly identified by West et. al in their study, the impact of electronic data is increasing and there is a need to come up with effective visualization methodologies to make the most of this data (West, Borland and Hammond, 2014). This necessity is being met by specific and focused work on healthcare data to come up with the solutions like interactive visualization (Shneiderman, Plaisant and Hesse, 2013) and Mircomaps (Pickle and Carr, 2010) to name a few. Adding to that, cartograms are also used for visualizing health data clusters with focus on visualizing the geological information (Kronenfeld and Wong, 2017). Apart from these, machine-learning methods are also utilized for building prediction models on the same BRFSS data (Xie et al., 2019). Moreover, Kwon et. al have worked on the same data and have used multivariate regression models for comparison of physical activity levels with cancer survivors and those who have not being diagnosed with cancer (Kwon, Hou and Wang, 2011). In short, efforts are being made to get as much benefit as possible from the healthcare data and this report also provides some insights for choosing the right visualization methods for this kind of data.

In the process of answering the question asked at the start, though bar-plots gave the most meaningful results. However, a very interesting observation was made from figure 1 and 2. Figure 1 was indicating that the number of people with ‘very good health’ has decreased in 2015 as compared to the number in 2011. But, visualizing the same data with different visualization technique i.e. Pie-charts indicated some kind of biasedness because in comparison to the total responses, there were more people in 2015 with ‘very good health’ than there were in 2011. This was a clear indication of how the

methodology of visualization influenced the conclusion taken from data. In these two visualization methods, pie charts gave more information about the data as compared to bar graphs.

Conclusion

Data visualization and analysis techniques are becoming a useful tool for finding the causes of different health disorders and devising the prevention methods. This article has extended the effort of BRFSS system to find out the trends of health awareness in people of US and the number of people suffering from major diseases.

Different visualization techniques were used in the process of finding data pattern in two different aspects, i.e. awareness and disease results. In this effort, though bar-plots were very helpful in visualizing the holistic data and their limitation of ignoring the total number of objects was solved using pie-charts. Moreover, linear interpolation was used to individually analyse the trends in awareness and diseases. It is advised that further study of the data with some more records of different years can be helpful in coming up with the major healthy habits which should be adopted for a better lifestyle. The future study of the data can also be useful in answering questions like 'What is the most important habit to be adopted for protecting oneself from heart disease' and in finding the critical areas in which people are not aware of. These studies can define directions of public awareness programs for healthy world. In this regard, different visualization techniques as well modern technologies like machine learning and data mining can also be very beneficial.

References

Amarasingham, R., Patzer, R., Huesch, M., Nguyen, N. and Xie, B. (2014). Implementing Electronic Health Care Predictive Analytics: Considerations And Challenges. *Health Affairs*, 33(7), pp.1148-1154.

Harris, S., May, J. and Vargas, L. (2016). Predictive analytics model for healthcare planning and scheduling. *European Journal of Operational Research*, 253(1), pp.121-131.

Behavioral Risk Factor Surveillance System. (2018). *CDC - BRFSS Annual Survey Data*. [online] Available at: https://www.cdc.gov/brfss/annual_data/annual_data.htm [Accessed 25 Oct. 2019].

West, V., Borland, D. and Hammond, W. (2014). Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, [online] 22(2), pp.330–339. Available at: https://academic.oup.com/jamia/article/22/2/330/695186.

Shneiderman, B., Plaisant, C. and Hesse, B. (2013). Improving Healthcare with Interactive Visualization. *Computer*, 46(5), pp.58-66.

Pickle, L. and Carr, D. (2010). Visualizing health data with micromaps. *Spatial and Spatio-temporal Epidemiology*, 1(2-3), pp.143-150.

Kronenfeld, B. and Wong, D. (2017). Visualizing statistical significance of disease clusters using cartograms. *International Journal of Health Geographics*, 16(1).

Xie, Z., Nikolayeva, O., Luo, J. and Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, 16.

Kwon, S., Hou, N. and Wang, M. (2011). Comparison of physical activity levels between cancer survivors and non-cancer participants in the 2009 BRFSS. *Journal of Cancer Survivorship*, 6(1), pp.54-62.

CDC. (2016). *Behavioral Risk Factor Surveillance System - 2015 Codebook Report*. [online] Available at: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf [Accessed 29 Oct. 2019].