Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management
Detroit, Michigan, USA, September 23-25, 2016

Data Analytics and Visualization in Analyzing Mortality Records

Mehul R Patel and Md. Noor-E-Alam
Department of Mechanical and Industrial Engineering
Northeastern University
Boston, MA 02115, USA
Patel.mehu@husky.neu.edu, mnalam@neu.edu

Abstract

Advent of sophisticated technologies has drastically reduced the cost of data collection and storage but unless we exploit the data in order to gain meaningful insights, it is just the *data*. Data Analytics can be considered as an approach to carve out information from a very large data set. Analytics is becoming prevalent in the field of healthcare, manufacturing industries, and service industries. In this paper, an attempt has been made to analyze the mortality data set of the USA-2014 using data analytics and data visualization tools. Findings of the study suggest that number of suicides and level of education are inversely related, and in the USA, men tend to have more suicidal tendencies compared to that of women. The study also includes analysis of deaths in 2014 with respect to marital status, age-group, injury at work, leading causes of death, and manner of death. This analysis can also be used to build predictive models which may help to take preventive steps in order to reduce untimely deaths in the future. It could also be helpful for government agencies and policy makers to set strategies to avoid unwanted incidence in the society.

Keywords
Data Analytics, Data Visualization, Mortality Analysis

1. Introduction

The primary objective of performing this study is to analyze major causes of deaths, and simultaneously also to find what was the main reasons of untimely deaths, such as suicides, accidents etc. in the USA. We also attempted to present our findings in a visually appealing manner. The analysis that we performed can be helpful for the government in order to conduct awareness programs. Further, it can be used by policy makers and healthcare providers to make sustainable policies and disease control systems for well-being of people as it includes detailed information about the deceased people's demographic background as well.

Statistical analysis of death cases is carried out for evaluating the treatment effect, and developing disease control and prevention strategies [1], where the study was performed for only 2115 death cases. Illicit and non-illicit drugs are the cause of Drug Poisoning Deaths (DPD) whether intentional or unintentional [2]. The number of DPD nationally has increased dramatically in the last 15–20 years. And this can be considered as an untimely and unnatural incidence, which can be reduced if pro-active strategies are implemented. Educational materials need to be marketed to the demographic groups at greatest risk and take into account differences in population characteristics between and within States [2]. Other economic facts are that wealthier abusers are more likely to abuse Prescription Opioid (PO) obtained from a doctor while poorer PO abusers tend to obtain them from illicit sources. Previous research has shown a greater tendency for young males to be involved with illicit drugs while women are more likely to use POs. This could explain one of the findings of our study that reason behind the most accidental deaths is exposure to specified or unspecified drugs or narcotics. And the accessibility of these drugs to the abuser is the result of their diversion from legal sources to the illicit market place [3].

For 15 to 24 year olds, the gun homicide rate in the USA was 49.0 times higher. Thus, the USA has a prevalent firearm problem compared with other developed countries, with higher rates of homicide and firearm-related suicide between year 2003 and 2010 [4]. Previous research indicates that homicide rates are generally higher in the regions where divorce rates, suicide rates, unemployment rate, the population, and per capita income are higher [5]. Untimely death such as death by suicide is the third leading cause of death for very young people (13 to 18 years old) and a key issue to prevent suicide is to encourage victim to seek professional care or help [6]. The reason behind committing a suicide could be loss of family member or abusive surrounding or lower self-esteem or exposure to illicit drugs or may be eating behavior disorder [7].

Our objective of this research is to analyze a large-scale death related data set using analytics and visualization techniques to unlock valuable insights which can help government agencies and policy makers to set public awareness strategies and preventive measures for social welfare.

2. Mortality Analysis

2.1 The Data Set

Deaths in the United States-2014 [8] is a publicly available data set, provided by Center of Disease Control and Prevention, which covers the information about the deceased people's educational background, age when died, sex, cause of death, manner of death, marital status and other 32 relevant criteria. This data set consists of 38 variables and 2,631,171 records (~ 2.63 million records) which indicates the total number of deaths in the USA in 2014. The entire database consists of 27 other different tables, or child tables, meaning that there are different tables for different variables. These tables contain specific numbering and description method which can be used to perform an analysis on the data (or parent table) at an aggregate level. The data set is majorly created by using numbers in order to reduce the size of the final data set and these numbers are further linked with other tables (or child tables) so that the specific numbering format can be exploited.

Variable named *MannerOfDeath* means a very general reason for a person's death, which could be either of the following: Natural, Accident, Homicide, Suicide, and Self-inflicted. On the other hand, variable named *Icd10Code,* that is International Classification of Diseases (ICD-10), indicates a specific reason of a person's death, few examples of which are mentioned in [9]-as follows:

- **Natural**: heart disease, dementia, and any other failure of internal body organ
- **Accident**: accidental poisoning by or exposure to drug or narcotics, unspecified fall, motor-vehicle accident
- **Homicide**: by firearm discharge, by sharp object, by handgun discharge
- **Suicide**: by firearm discharge, by strangling or suffocation, self-poisoning by drugs
- **Self-inflicted**: cutting or piercing oneself

2.2 Methodology of Analysis

We first tried to view and analyze the available data using MS-Excel 2013 (64-bit version) but handling of such a large data set of millions of records did not seem convenient in Excel. After this, we came up with a systematic experimental and analysis set-up as shown in the following Figure 1. In this set-up, first, we imported the entire data set in MS SQL Server Management Studio [10]. The objective was to extract required information out of the data set by writing specific set of queries. The size of output data varied from 2 to 5 - in terms of variables, and 9 to few thousand - in terms of records. This size of tables can be manipulated effectively by tools such as Excel. In the second step, we exported the output data, obtained in a tabular format, into MS-Excel to do some data cleaning. Finally, these excel files were exported into Tableau [11], a powerful tool for data visualization, to effectively communicate the findings. All the visuals of the analysis are in the following section.

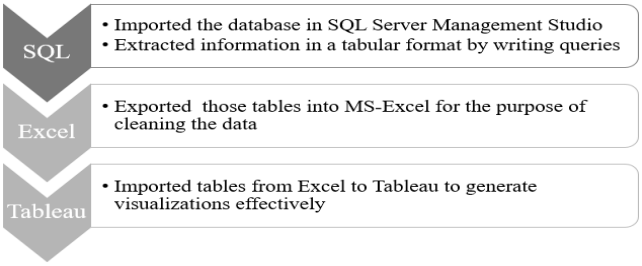

Figure 1: Systematic procedure of the analysis

2.3 Analysis

In this section, we will discuss all the analyses that we carried out with the data set. For the analyses, bar charts, stacked bar charts, tree map, and circle views are generated with the software Tableau [11]. In our experiment, we have used a Laptop with the following configuration:

- RAM: 4 GB
- Processor: Intel Core-i5 52000 CPU @ 2.20Ghz
- System type: 64-bit OS
- Operating system: Windows 10 Education

2.3.1 Suicides, Sex, and Education

The following Figure 2 is a stacked bar chart, which indicates the number of suicides committed by males and females, depending upon their education level. Pink and blue bars represent number of suicides committed by females and males, respectively. We can infer that higher the education level, lesser the number of people committed suicide. These numbers also depict that less number of females committed suicides in 2014 compared to the suicides committed by men in the same year. One may think that this could be true because of high percentage of males in overall population but this is not true. Overall sex ratio in the USA in 2014 was 0.97 male per female [12]. That means even though there are less number of males in total population, proportion of men committing suicides is very high. This finding can be utilized to take necessary steps, which may help in reducing number of suicides in males as well as in females. In this study, the unknown data points for education level of deceased people have been neglected. *Null* indicates no education at all.

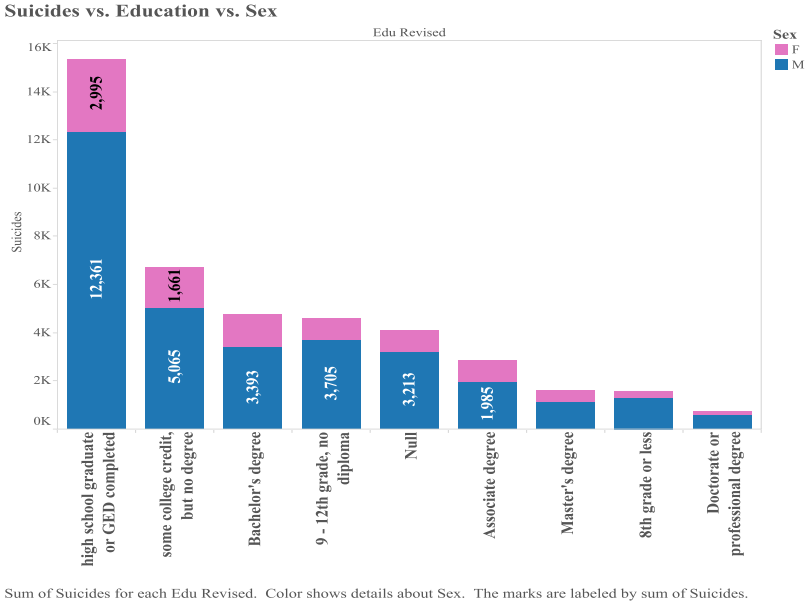

Figure 2: Suicides vs. Sex vs. Education level

2.3.2 Major Causes of Deaths

In the following Figure 3, we analyzed the total number of deaths, segregated on the basis of causes of death. According to the data, maximum number of people died because of Atherosclerotic Heart Disease (i.e. 161,961 deaths), followed by Malignant neoplasm: Bronchus or lung unspecified (i.e. 154,862 deaths). Results obtained in this figure are filtered out for more than 50,000 deaths.

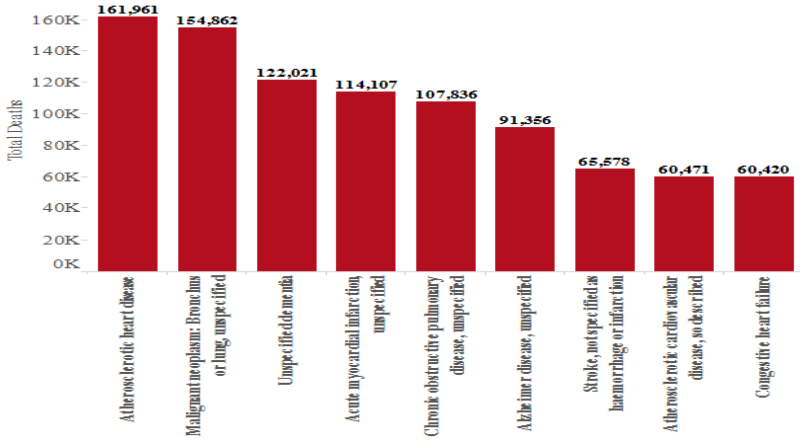

Figure 3: Total deaths vs. Causes of deaths

When we further analyzed the results on the basis of sex, we came to know that most females died because of *Unspecified Dementia* and most males died because of *Atherosclerotic Heart Disease*. The following Figure 4 is a tree-map, divided into females and males, which has blocks of different sizes. Shaded pink color indicates the total number of deaths of females and shaded purple color indicates the total number of deaths of males, because of the specified cause in that block. That means the area of each block is dependent on the total number of deaths because of that cause. The larger the block is, higher the number of deaths for that cause. We see that 82,864 females died of *Unspecified Dementia* but comparatively less males died of the same disease. It means that dementia is more prevalent in females than it is in the males. This fact can be used to bring awareness among females about the disease and preventive steps can be taken as well. On the other hand, 88,792 males and 73,169 females died of *Atherosclerotic Heart Disease*, as we see in the tree-map. Results for this figure are filtered out for more than 50,000 deaths.

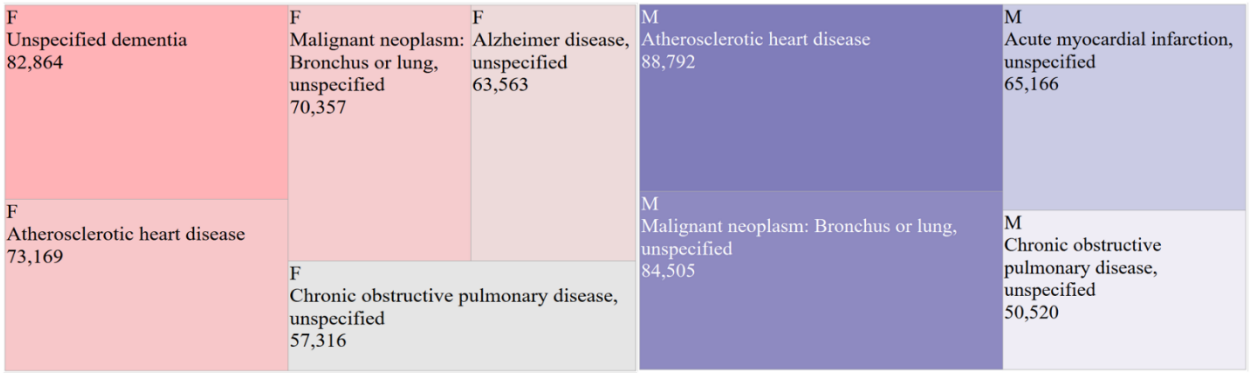

Figure 4: Tree map for major causes of deaths based on sex

2.3.3 Deaths on Days of the Week

This section reveals interesting insights about the time frame of unwanted incidences. The following Figure 5 shows which day of the week was the *deadliest* in the USA in 2014, considering whether a person was injured at work. We see that 747 people died on Tuesday, the highest, who were injured at work and 31,447 people died on Saturday, the highest, who were not injured at work. For this particular analysis, there are some records in the data set for which it is not known whether a person was injured at work. Such status of the variable *InjuryAtWork* is referred as *Unknown* in the data set. We have neglected those data points for this analysis.

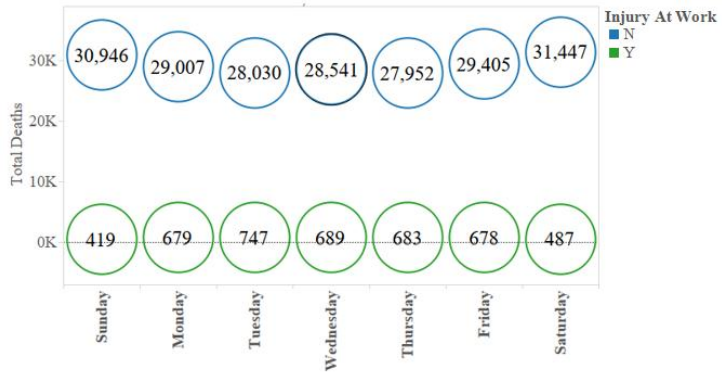

Figure 5: Total deaths based on days of the week and injury at work

On an aggregate measure, most people died on Wednesday (total = 380,526), and the least number of people died on Sunday (total = 372,388) which is shown in Table-1.There are few records in the data set for which it is not known that on which day of the week a person died. Therefore, we have not taken into consideration such data points for the specific variable *DayOfWeekOfDeath*.

Here, a sample SQL query is shown to better understand the procedure of our analysis (**Query run time was 5 to 6 Seconds)**:

```
select dr.DayOfWeekOfDeath, dwd.Description,  COUNT(*) as TotalDeaths
from DeathRecords dr
join DayOfWeekOfDeath dwd
on dr.DayOfWeekOfDeath = dwd.Code
group by dr.DayOfWeekOfDeath, dwd.Description
order by TotalDeaths Desc
```

A sample output is also shown in Table 1:

WeekDay	Description	TotalDeaths
4	Wednesday	380526
6	Friday	378949
7	Saturday	378070
2	Monday	375854
3	Tuesday	372792
5	Thursday	372464
1	Sunday	372388
9	Unknown	128

Table 1: Deaths on days of week

2.3.4 Deaths Based on Age Groups and Sex

Figure 6 depicts the total number of deaths, segregated in different age groups. Bar charts, on the left and right sides, indicate total deaths of females and males, respectively, based on age groups. It can be inferred from this figure that most of females died when their age was 85 or more, which is clearly not the case with males. Most men died when they were in the age range of 75 to 84 years. It means that in 2014, females lived longer than males did, in the USA. This finding supports the fact that life expectancy of a female in the USA is more than that of male [13].

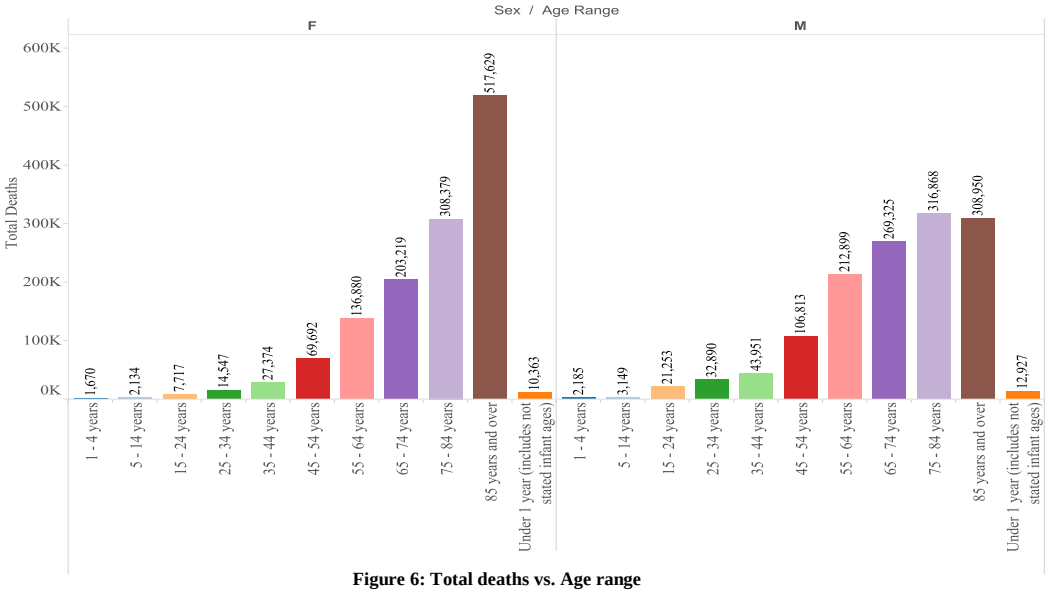

Figure 6: Total deaths vs. Age range

2.3.5 Manner of Death Based on Age Groups

Figure 7 represents total number of deaths occurred because of accidents, homicides, and suicides. It is clear from the figure that most people died of accidents and suicides when their age was between 45 and 54. 21,016 people died of accidents indicated by blue color and 8,855 people died of suicides indicated by pink color, when they were in that particular age range. Our findings indicate that major causes of accidental deaths are: Accidental poisoning by and exposure to unspecified drug or narcotics, unspecified fall, unspecified motor-vehicle accident etc. It is also clear from the figure that most people died of homicides when their age was between 25 and 34. Indicated by green color, 4397 people died because of homicides when they were in that particular age range. Our findings suggest that most homicides were committed by using firearms, and sharp objects.

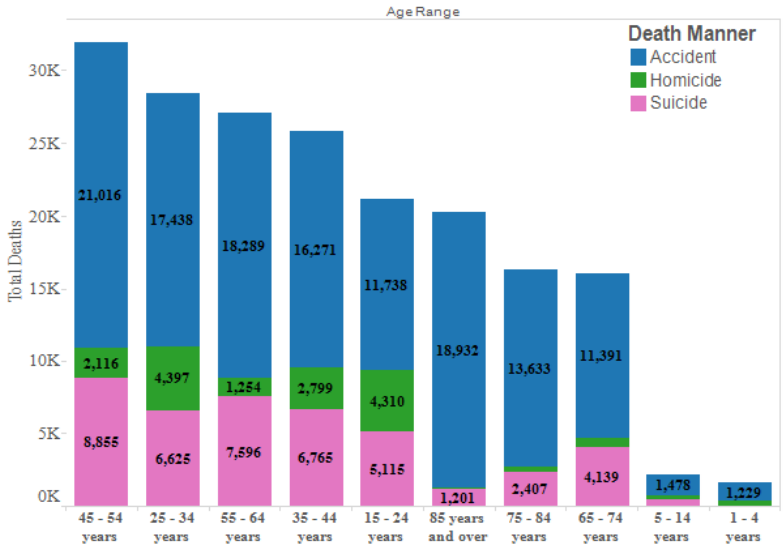

Figure 7: Manner of death vs. Age range vs. Total deaths

2.3.6 Marital Status, Education Level, and deaths

Stacked bar chart and simple bar chart of this section represent total number of deaths, segregated based on people's marital status and education level. Figure 8(a) is for those people whose education level was other than *High School Graduate* or *General Educational Development (GED)*. Figure 8(b) is only for those deceased people whose education level was *High School Graduate* or *GED*. Reason behind separating the GED data from rest of the data is their comparatively high numbers of deaths for any marital status, which is clear from the following figures.

For Figure 8(a), blue, orange, green and purple colors indicate marital status of divorced, married, single and widowed, respectively. Deeper analysis of the stacked bar chart indicates that 124,530 married people died who had no college degree, followed by 110,451 deaths of married people who had only Bachelor's degree. We can also see that 54,049 people died who were *single* and had education level of 8^{th} *Grade or less.* On the other hand, based on the marital status, the highest number of people died who were *Widowed* and had education level of *GED*, shown in Figure 8(b).

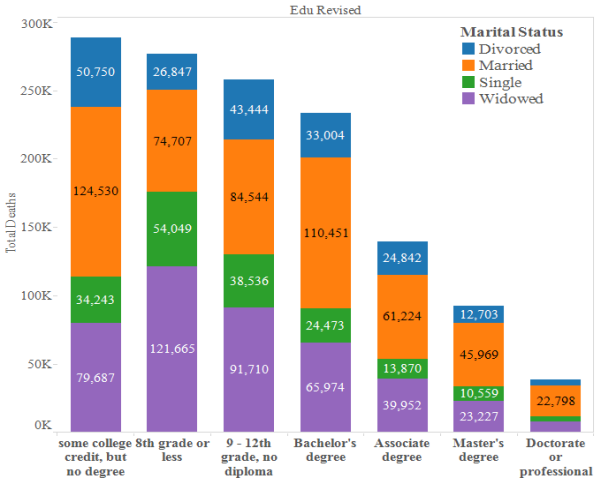

Figure 8 (a): Total deaths vs. Marital status vs. Education (w/o GED)

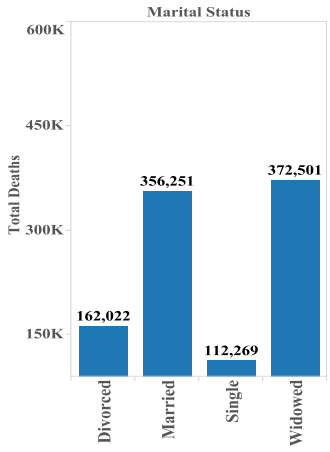

Figure 8 (b): Total Deaths vs. Marital Status vs. Education (GED only)

3. Conclusion

This research carried out a systematic analysis with visualization and analytics tools for a large-scale death data set. It has revealed many interesting insights as a wakeup call, which we claim helpful for policy makers and government agencies to develop pro-active strategies for the social welfare. For example, our study revealed that maximum number of suicides are committed by the people having very basic educational background. As such, awareness programs can be conducted by the government or other welfare agencies to show the importance of the education, and training program can be introduced to avoid such incidences. We also found that accidental deaths are prevalent when person's age is more than 45 years. Therefore, policymakers can bring forth necessary policies to have suitable insurance for appropriate age groups. On the other hand, homicides are dominant in the age group when most people are in their late twenties or early thirties. Most people died of homicides when their age was between 25 and 34, followed by second highest homicides when their age was between 15 and 24. This represents a shift of age range when people died of homicide, an unnatural violent death. This finding suggests that the policy makers or government agencies should come up with a set of strict regulations or strategies for firearm ownership and usage to avoid homicides.

References

[1] Hu Shengli, Luo Yi, Statistical Analysis of 2115 Hospitalization Death Cases, *7th International Conference on Information Technology in Medicine and Education,* 2015.

[2] Ruth Kerry, Pierre Goovaerts, Maureen Vowles, Ben Ingram Spatial analysis of drug poisoning deaths in the American West, particularly Utah. *International Journal of Drug Policy*, 2016.

[3] T.J. Cicero, S.P. Kurtz, H.L. Surratt, G.E. Ibanez, M.S. Ellis, M.A. Levi-Minzi, et al. Multiple determinants of specific modes of prescription opioid diversion, Journal of Drug Issues, 41 (2012), pp. 283–304

[4] Erin Grinshteyn, David Hemenway, Violent Death Rates: The US Compared with Other High-income OECD Countries, *The American Journal of Medicine*, 2010.

[5] David Lester (2001). Regional studies of homicide:A meta analysis, *Death Studies, pp.705 – 708,* 2001.

[6] Verena Venek, Stefan Scherer, Louis-Philippe Morency, Albert Rizzo, and John Pestian (2015). Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction, *IEEE (2015).*

[7] Celina Stafie, Maria Manuela Apostol (2009). The need for multidisciplinary approach in the treatment of eating behaviour disorders of the young population, *Advanced Technologies for Enhanced Quality of Life-IEEE* (2009).

[8] *National Vital Statistics Reports*. (2014). Retrieved on 25th June, 2016 from Center for Disease Control and Prevention: http://www.cdc.gov/nchs/products/nvsr.htm

[9] 2016 ICD-10-CM Codes. (2016). Retrieved on 25th June, 2016 from ICCD10 data: http://www.icd10data.com/ICD10CM/Codes

[10] *Edition and Components of SQL Server 2014*. (2016) Retrieved on 24th June, 2016 from Microsoft Developer Network: https://msdn.microsoft.com/en-us/library/ms144275(v=sql.120).aspx

[11] Tableau Support. (2003-2016). Retrieved on 24th June, 2016 from Tableau: http://www.tableau.com/support.

[12] The World Factbook. (2015). Retrieved on 25th June, 2016 from Central Intelligence Agency: https://www.cia.gov/library/publications/the-world-factbook/fields/2018.html

[13] Felicitie C. Bell, M. L. (2005, August). Retrieved on 25th June, 2016 from Social Security Administration: https://www.ssa.gov/oact/NOTES/pdf_studies/study120.pdf

Biography

Mehul Patel is currently a graduate student of Industrial Engineering at Northeastern University, Boston, MA. He earned his B.Tech. in Industrial Engineering from Pandit Deendayal Petroleum University (PDPU), Gujarat, India. He has previously worked as an Analyst at Indus Momentus Business Solutions (IMBS). His research interest includes data analytics, data visualization, data mining, design of experiments, operations research, and supply chain management.

Md Noor-E-Alam is an Assistant Professor in the Department of Mechanical & Industrial Engineering at the Northeastern University. Prior to his current role, he was working as a Postdoctoral Research Fellow at Massachusetts Institute of Technology. His current research interests lie in the intersection of operations research and data analytics, particularly as applied to healthcare, manufacturing systems and supply chain. He has completed his PhD in Engineering Management in the Department of Mechanical Engineering at the University of Alberta (UofA) in 2013. Before coming to the UofA, he served as a faculty member (first as a Lecturer and then as an Assistant Professor) in the Department of Industrial and Production Engineering at Bangladesh University of Engineering & Technology (BUET). He also previously received a B.Sc. and M.Sc. in Industrial and Production Engineering from BUET.