Summary Document- Data Miners

Preface

This is a summary document prepared by Team Data Miners as part of the final deliverable for the team project in the course “ISQA 8086- From Data to Decisions”.

Table of Contents

1. Introduction to the Data Set

This data represented the percentage of expected death, observed deaths and potential death in different states in the United States. All data is collected based on five leading causes of death in metropolitan and non-metropolitan areas. Heart disease, Cancer, Unintentional injury, Chronic lower respiratory disease and Stroke are the major causes of death that patients are spending so much money for any treatment for them.

With analyzing this data and finding differences between Expected Deaths and Observed Deaths, health care can be improved by public health programs, that support healthier behaviors to better access the health care services for reducing the rate of potentially excess deaths.

2. Target Audience and Analysis targets

The potential users of this data set could be:

- Doctors who are passionate about creating awareness about causes of death and take steps for prevention of the leading causes of death
- Medicine Companies who would like to produce medicines based on the demand region
- Health Researchers' Groups who would like to initiate programs to share the necessary information to the public for their welfare

  The analysis target made by Data Miners focuses more on creating awareness in people and take necessary precautions or steps to reduce the number of deaths in future. Along with that this analysis also targets doctors to do more research on creating tools which can help in detecting diseases well in advance and to cure them once it is found. The analysis can also help Health Researchers' Groups to conduct more programs in the regions where the number of deaths are more in number.

3. Research Questions

1. What is the trend of observed deaths for all the five-leading cause of deaths over time?
2. What is the trend of expected deaths, observed deaths, potentially excess deaths in each age group?
3. Compare the number of deaths for various regions of US. Which region has the maximum and minimum number of deaths recorded for the years 2005 to 2015?
4. Compare the number of deaths for each locality. Which locality has the maximum and minimum number of deaths recorded for the years 2005 to 2015?
5. What is the ratio between the observed deaths and the Population? What is the trend of the ratio over time by region?
6. What is the ratio between the expected deaths and the Population? What is the trend of the ratio over time by region?
7. Are the ratios in questions 5 and 6 correlated?

4. License:

The License details for the Excess Deaths Data Set can be found in License Info

There are no constraints on this data in regard to the license. This is public data set and can be downloaded by any individual within or outside the organization.

5. Metadata

This data set was collected for the years 2005-2015

State FIPS Code were numeric and two-letter alphabetic codes defined in U.S. Federal Information Processing Standard Publication ("FIPS PUB") 5-2 to identify U.S. states and certain other associated areas.

Mortality data for U.S. residents come from the National Vital Statistics System. Estimates based on fewer than 10 observed deaths are not shown and shaded yellow on the map.

Cause of death is based on the International Classification of Diseases, 10th Revision (ICD-10)

Heart disease (I00-I09, I11, I13, and I20–I51)
Cancer (C00–C97)
Unintentional injury (V01–X59 and Y85–Y86)
Chronic lower respiratory disease (J40–J47)
Stroke (I60–I69)

Locality (nonmetropolitan vs. metropolitan) is based on the Office of Management and Budget's 2013 county-based classification scheme.

Benchmarks are based on the three states with the lowest age and cause-specific mortality rates.

Expected deaths are the number of deaths that would be expected if the death rates of the states with the lowest rates occured across all states.

HHS Region is the number of the region allocated by the Office of Intergovernmental and External Affairs. It hosts ten Regional Offices that directly serve state and local organizations. A President-appointed Regional Director leads each office.

Potentially excess deaths for each state are calculated by subtracting deaths at the benchmark rates (expected deaths) from observed deaths.

Users can explore three benchmarks:

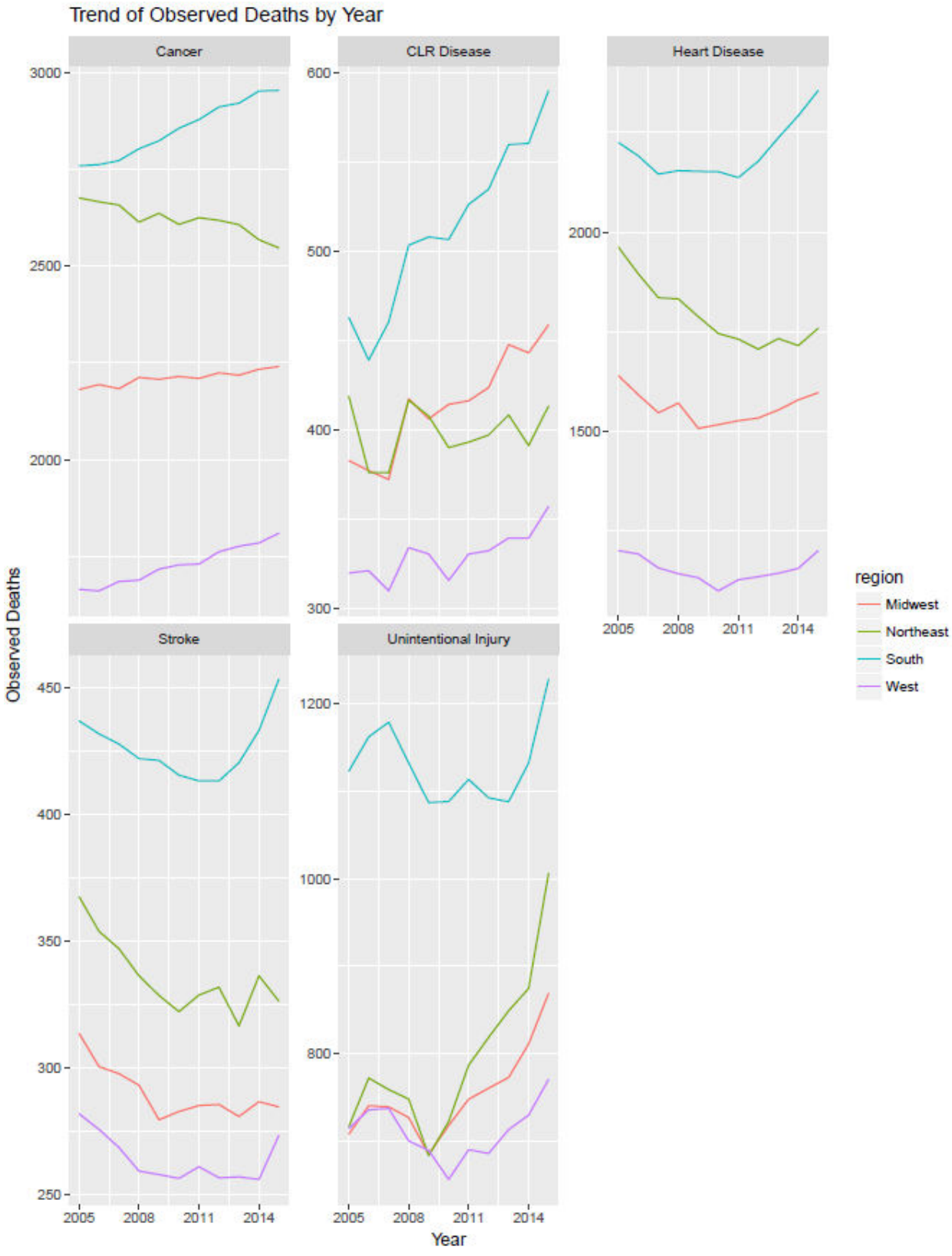
"2010 Fixed" is a fixed benchmark based on the best performing States in 2010.
"2005 Fixed" is a fixed benchmark based on the best performing States in 2005.
"Floating" is based on the best performing States in each year so change from year to year.

Page last reviewed: July 14, 2017
Page last updated: August 28, 2017
Data Last Updated: August 15, 2017
Metadata Last Updated: August 15, 2017
Date Created: January 19, 2017

Content source: CDC/National Center for Health Statistics

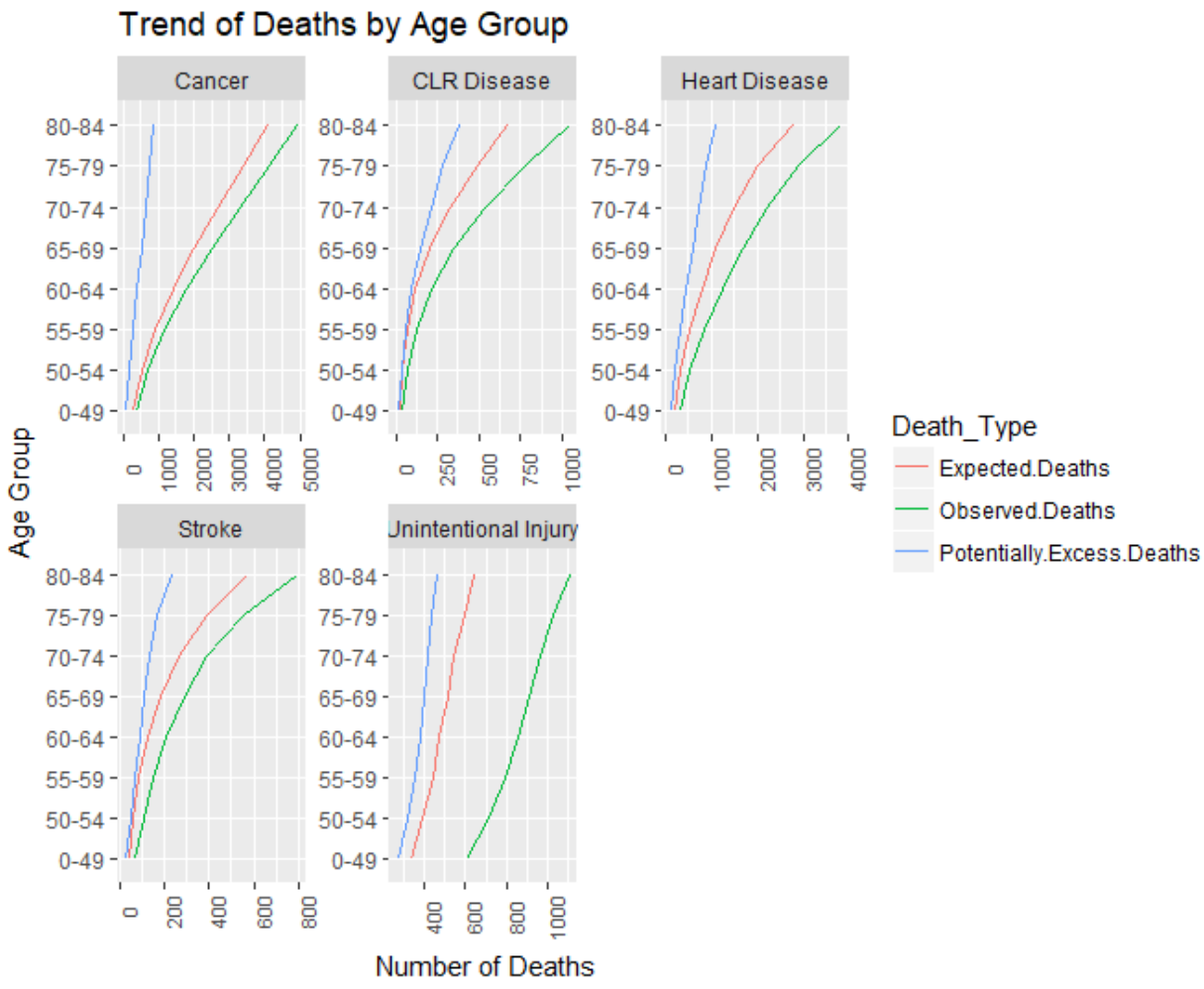
6. R Plots

Plot which represents the trend of all the five leading causes of deaths over time is shown below:


Trend of Observed Deaths by Year
Cancer
CLR Disease
Heart Disease
Stroke
Unintentional Injury
Observed Deaths
Year
region
Midwest
Northeast
South
West
3000
2500
2000
600
500
400
300
2000
1500
450
400
350
300
250
1200
1000
800
2005
2008
2011
2014

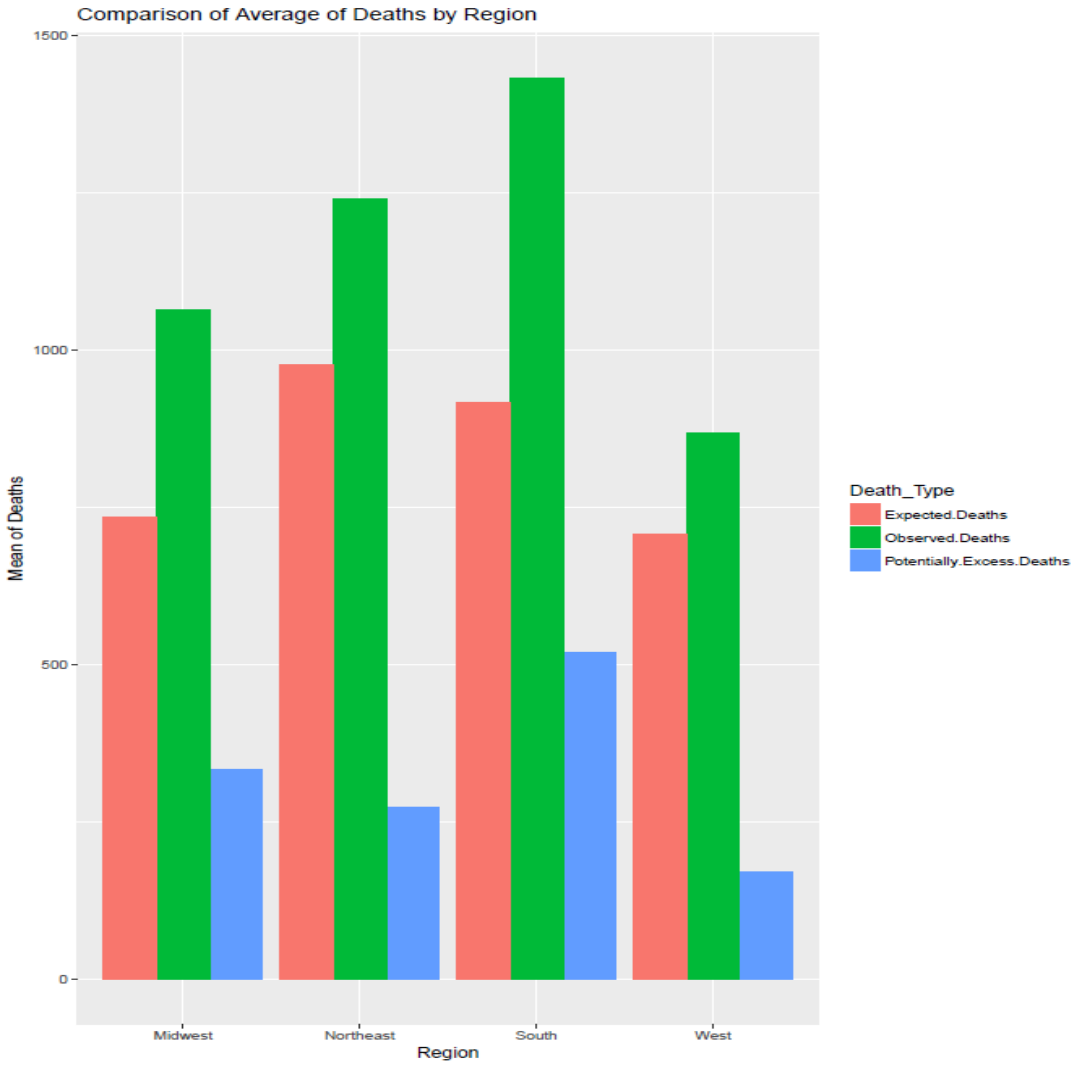
This plot provides the overview of trend for the five-leading cause of deaths (observed deaths) over time. From the plot we could see that "Cancer" has the highest number of deaths and that is in the southern region of United States. The lower number of deaths could be seen for "Stroke" in the western region of United States. The order of the trends in descending order are Cancer, Heart Disease, Unintentional Injury, Chronic Lower Respiratory Disease, Stroke.

The trend of all types of deaths in each age group is plotted as below:



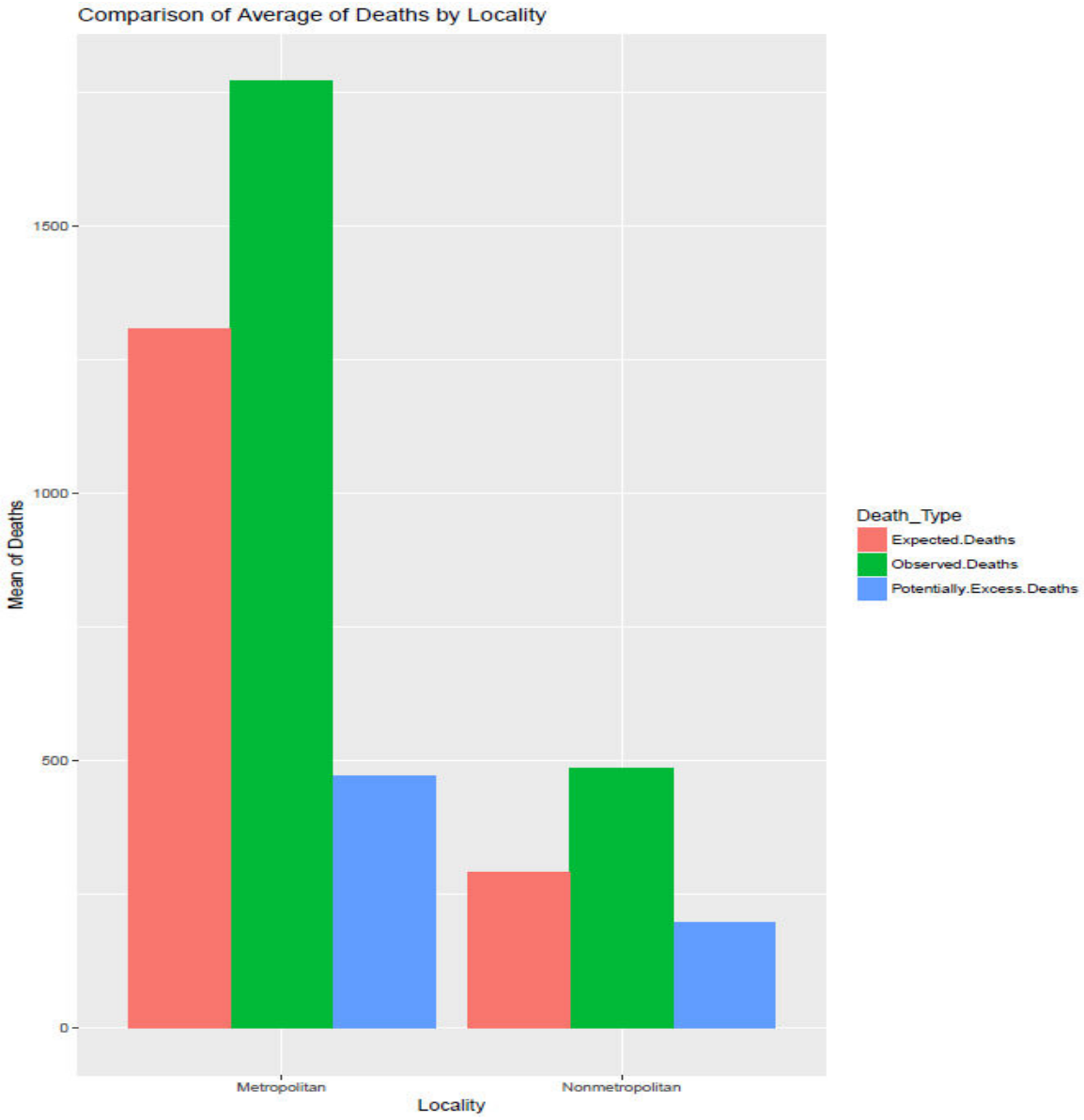
This plot provides the trend of all types of deaths for each age group. In this plot we could see that the observed deaths are always greater than the expected deaths. As the age increases the numbers of deaths kept increasing. This behavior is consistent for all the causes of death.

Plots were also developed for comparing all types of deaths averages across regions and localities as below:



This plot gives the mean of all types of deaths over various regions. The observed deaths are more in the southern part of the United States followed by Northeast, Midwest and West. The Expected Deaths average is almost equal in both South and Northeast. Midwest and West also have equal averages but lower than South and

Northeast. South region also has the maximum average for potentially excess deaths followed by Midwest, Northeast and West.



This plot provides the means of all the deaths by locality.

7. Interpretation of R plots

The southern and north-east regions of United States need to be taken care. The habit of cigarette smoking is found to be more among adults in non-metropolitan. It is also found in a research that cigarette smoking is the leading cause of preventable disease and death in the United States. The use of tobacco need to be reduced. Multiple health programs or demos need to be conducted by health researcher's groups to reduce the excess deaths due to the use of tobacco. This can help reducing the deaths due to heart disease, stroke and Chronic Lower Respiratory Disease. Doctors and Government need to think of arranging health care to areas which recorded high number of deaths due to long-distance traveling for emergency.

No physical activity is also a factor for many chronic related diseases. Regular physical activity through various activity sessions will be of much help. Also in few non-metropolitan areas there might be no health care services for regular checkups, quality care for patients and cancer cervical care. This is where Dr. Quinn can have his expertise get into implementation. He can have a health camp every now and then in those areas or can even think of establishing a new health care building which can be centric in location with respect to all the affected areas.

About 85% of the U.S population (275.3 million) lived in metropolitan localities in 2015. As the number of people increased so did the number of deaths. From the uneven distribution of population between metropolitan and non-metropolitan, a graph which is plotted based on locality will lead to false assumptions that the number of deaths are more in metropolitan than in non-metropolitan. The difference in both the localities with respect to demography, environment, economy and social characteristics also affect the types of health problems. The poverty levels, drinking and smoking habits and obesity would be more in non-metropolitan when compared to metropolitan. But the access to health care would be less for the non-metropolitan. The metropolitan areas would have many health care providers. The deaths in both localities for all the five-leading cause of deaths led to more than 50 percent of the deaths in United States. It might not be possible to prevent excess deaths in few areas which might include long-distance traveling to emergency care. But there are few areas in which few careful precautions and few health programs can make a huge impact.

More deaths for the age groups less than 80 years are recorded for the non-metropolitan areas. When considering the cause of deaths individually, it is observed that more than half of the deaths recorded in non-metropolitan are due to unintentional injury and Chronic Lower Respiratory Disease which is more when compared with metropolitan which is around 30-35%.

The number of deaths keep increasing with the age and year. Highest number of deaths are recorded for the age group 80-84. The observed deaths are always found to be more in number than the expected deaths.

"Cancer" has the highest number of deaths and that is in the southern region of United States. The lower number of deaths could be seen for "Stroke" in the western region of United States. The order of the trends in descending order are Cancer, Heart Disease, Unintentional Injury, Chronic Lower Respiratory Disease, Stroke.

The observed deaths are more in the southern part of the United States followed by Northeast, Midwest and West. The Expected Deaths average is almost equal in both South and Northeast.

Word Count: 1687

References

National Center for Health Statistics. (2017, 08 28). Retrieved September 04, 2017 from https://www.cdc.gov/nchs/data-visualization/potentially-excess-deaths/index.htm

US Department of Health and Human Services. Tobacco use. Healthy people 2020. Washington, DC: U.S. Department of Health and Human Services; 2013. https://www.healthypeople.gov/2020/topics-objectives/topic/tobacco-use?topicid=41