Potentially Excess Death from the Five Leading Causes - USA

DSC 465 - Data Visualization

Dr. Eli T Brown

Umair Chaanda | Shashank Srikanth | Natalia Bies | Rikesh Patel | Xiaotong He

Introduction

The aim of this deep analysis is to visualize a large "Potentially Excess Deaths from five leading causes of deaths" dataset with several different rich visualization techniques. We have created some of the high-quality explanatory visualizations using Tableau and R that communicates a story from the data using Animation and interactive Dashboards.

The data used in this project is acquired from cdc.gov[1]. The publisher is Centers for Disease Control and Prevention. There are over 200,000 observations in total from the year 2005 to 2015. However, there is overlapping data so we had to apply filters to exclude records. For example under variable Locality All = Metropolitan + Nonmetropolitan so we excluded All. There were 5390 out of 205920 total records selected after applying filters.

The dataset was primarily collected for examining the prevalence of potentially excess diseases, leading to a study into access to healthcare and initiatives geared towards bettering health and wellness.

Variables Involved

Following are the variables involved in our analysis:

Numerical	Categorical / Ordinal	Time & Geographic

• Observed Deaths *(Actual Observed Deaths in each state)*
• Population
• Expected Deaths *(Number of Deaths that would be expected if the death rates of the states with the lowest rates occurred across all states.)*
• Potentially Excess Deaths ([Observed Deaths] - [Expected Deaths])
• Percent Potentially Excess Deaths ([Potentially Excess Deaths] / [Observed Deaths] * 100)
• Percent Death Rate ([Observed Deaths] / [Population] * 100)
	• Cause of Death
	  ○ Cancer
	  ○ Chronic Lower Respiratory Disease
	  ○ Heart Disease
	  ○ Stroke
	  ○ Unintentional Injury
	• Age Range
	  ○ 0-49, 0-54, 0-59, 0-64, 0-69, 0-74, 0-79, 0-84
		• Year
		  ○ 2005 - 2015
		• State
		  ○ 50 States
		• Locality
		  ○ All
		  ○ Metropolitan
		  ○ Nonmetropolitan

[1] "NCHS Data Visualization Gallery - Potentially Excess Deaths ...."
https://www.cdc.gov/nchs/data-visualization/potentially-excess-deaths/index.htm. Accessed 25 Feb. 2020.

Exploratory Analysis

To understand the structure and trends in the data, a few exploratory visualizations were created. There are several variables that proved to be important for drawing conclusions. Since we primarily explored the potentially excess deaths and death rate, those variables were valuable in our data exploration and visualization. The years, states, locality, expected, and observed deaths were also important, depending on the direction taken for this investigation.

Stories

There are three directions that stand out as front runners for investigating:

Potentially Excess Deaths

One of them is examining the Percent potentially excess deaths, where we looked into the relationship between the excess deaths for each cause of death and each locality over time (i.e., how does the number of potentially excess deaths change for each cause of death and each locality (metropolitan vs nonmetropolitan), how has research for each cause of death affect this number, how aspects of healthcare in both localities differ). We have also looked into the excess deaths per state, seeing how it varies per state and how this has changed over time for each state. A similar thread between these ideas is that significant outside research regarding our healthcare system and studies on the five causes of death is required.

Percent Death Rate

The second direction is to examine the Percentage of Actual Deaths observed, where we found the relationship between the Death Rate by each cause of death and each locality over time. A new Death Rate variable was calculated by dividing Observed Deaths by Population and multiplying by 100. This gave us a trend over time to see how the actual number of deaths have increased or decreased for each cause of death and each locality. We went further deep into this story by looking at the Death Rate per state and how it has changed over time for each state.

Comparison

Finally, We also looked into the comparison between Excess Death and Death Rate in combination with the other variables e.g. Cause of Death, Locality, Year, and State.

Comparing Multiple Distributions

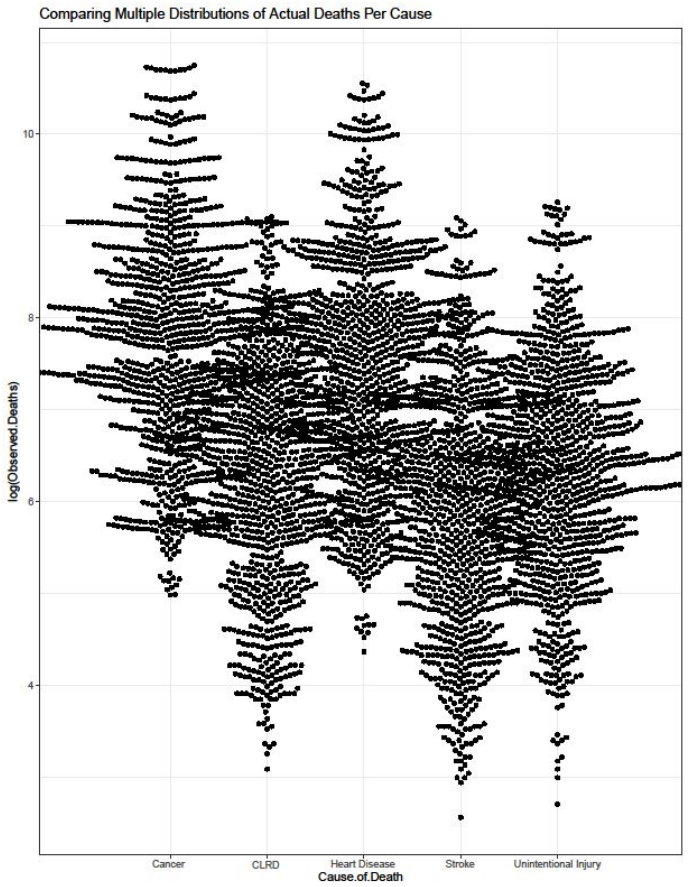

Figure 1.1

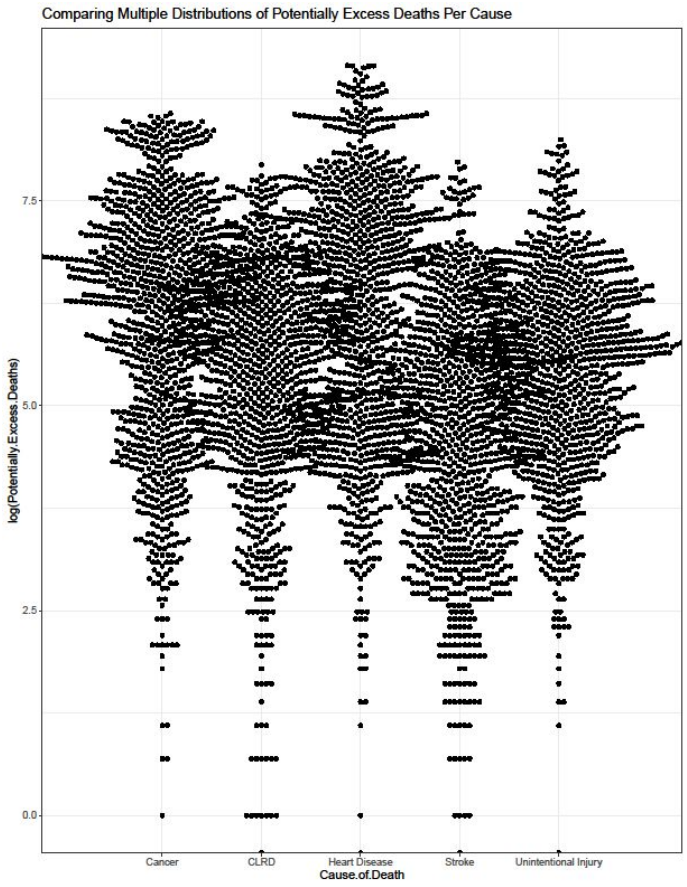

Figure 1.2

- The first exploratory visualization (Figure 1.1) shows the distribution of Actual Deaths per Cause from the years 2005 to 2015.
- Using the beeswarm plot method, we are able to compare the distributions as it relates to each cause of death: Cancer, Chronic Lower Respiratory Disease, Heart Disease, Stroke, and Unintentional Injury.
- We can quickly gather that Cancer is the leading cause of death, while Heart Disease is the second most leading cause of death.
- Overall, the distributions do not show signs of being left or right skewed, concluding that this data is fairly normal.

- Figure 1.2 shows the distribution of Potentially Excess Deaths per Cause from the years 2005 to 2015.
- As opposed to actual deaths, the excess deaths are mostly in Heart Disease followed by Cancer.
- Overall, the distributions do not show signs of being left or right skewed, concluding that this data is fairly normal with few outliers.
- From analyzing the distributions, we wanted to research further as to which years and which locations attribute the most to the cause of deaths.

Average Expected vs Average Actual Deaths

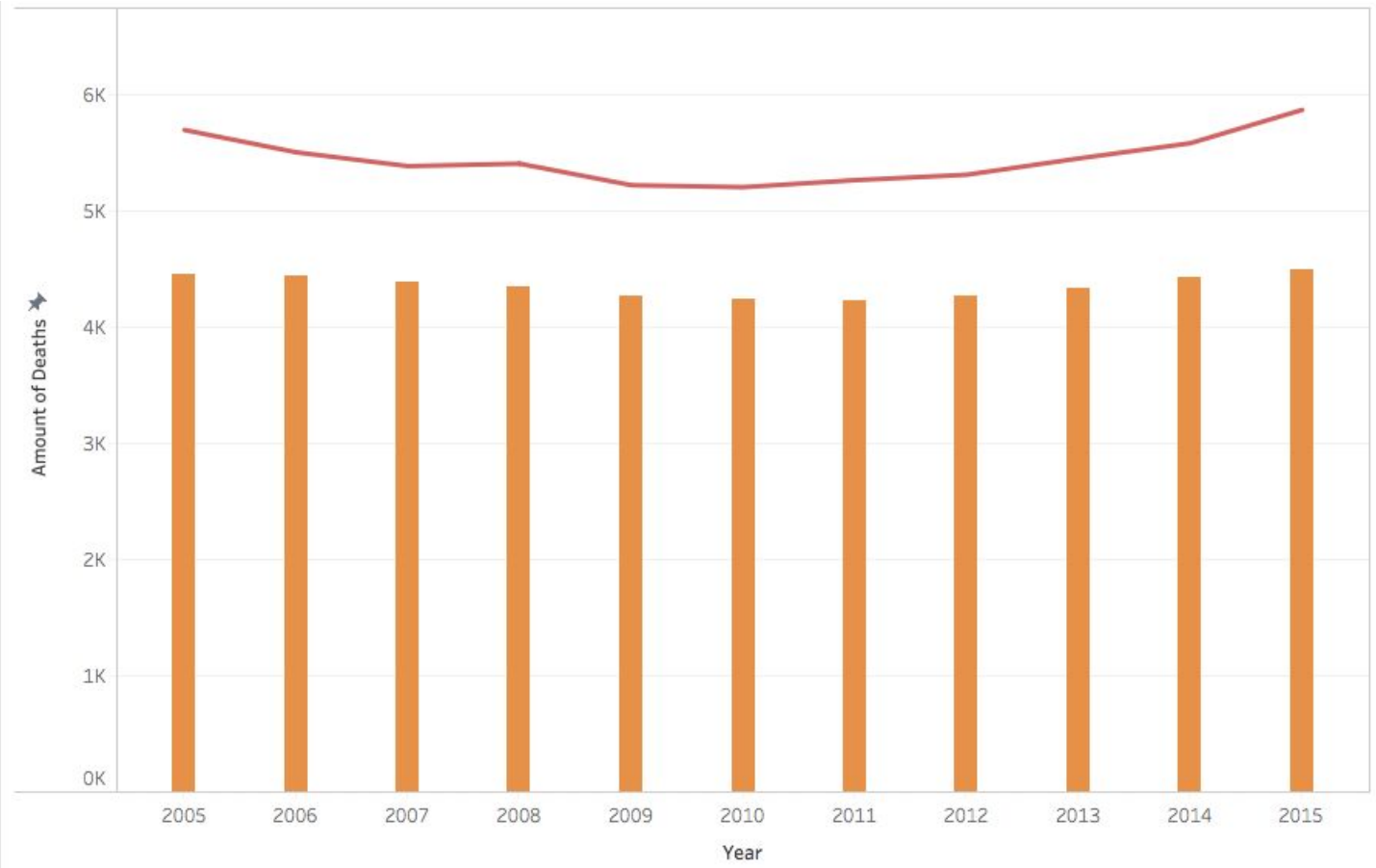

Figure 1.3

- The third exploratory visualization (Figure 1.3) explores the relationship between the average expected deaths and the average actual deaths from the years 2005-2015, from the metropolitan and nonmetropolitan areas, holding ages to 0-84 and the benchmark as floating.
- We can gather the average expected deaths slightly decreased from 2005-2010, but then gradually increased.
- We can also gather that the average actual deaths were above the expected of about 2,000 deaths on average.
- The average actual deaths follow a similar pattern as the average expected deaths of slightly decreasing from 2005-2010, but then gradually increasing.
- However, there is a minor spike in average actual deaths in 2008, perhaps due to the Great Recession of 2008.

Expected Deaths Vs Potentially Excess Deaths

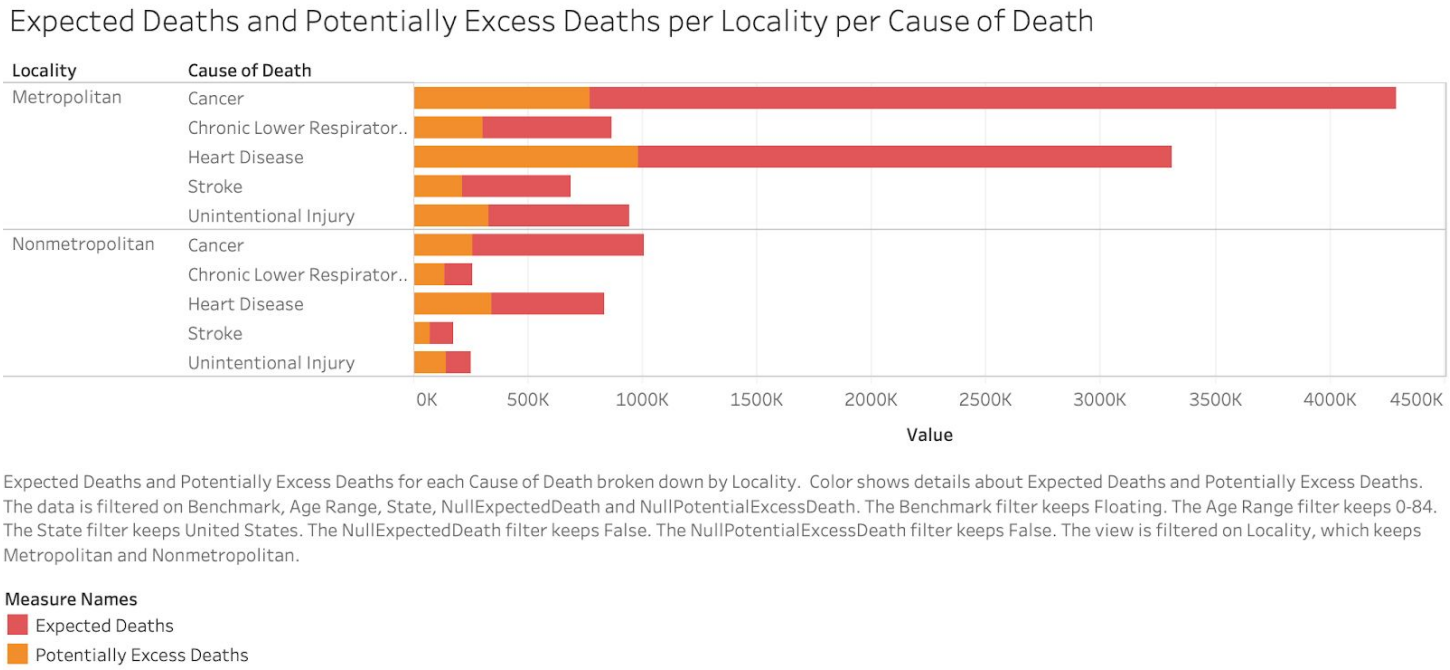

Figure 1.4

- The visualization (Figure 1.4) shows the expected deaths, potentially excess deaths, and the overall deaths for metropolitan and nonmetropolitan areas, divided by the cause of death.
- A bar, as a whole, represents the overall death.
- Potentially excess deaths are deaths that “exceed the numbers that would be expected if the death rates of states with the lowest rates occurred across all states.”
- Potentially excess deaths are calculated by subtracting overall deaths and expected deaths.
- The story here is that the popularity of potentially excess deaths portrays the need for improved health literacy and programs, combined with improved public health initiatives and better access to health care services.

Average Death Rate per Year per Locality

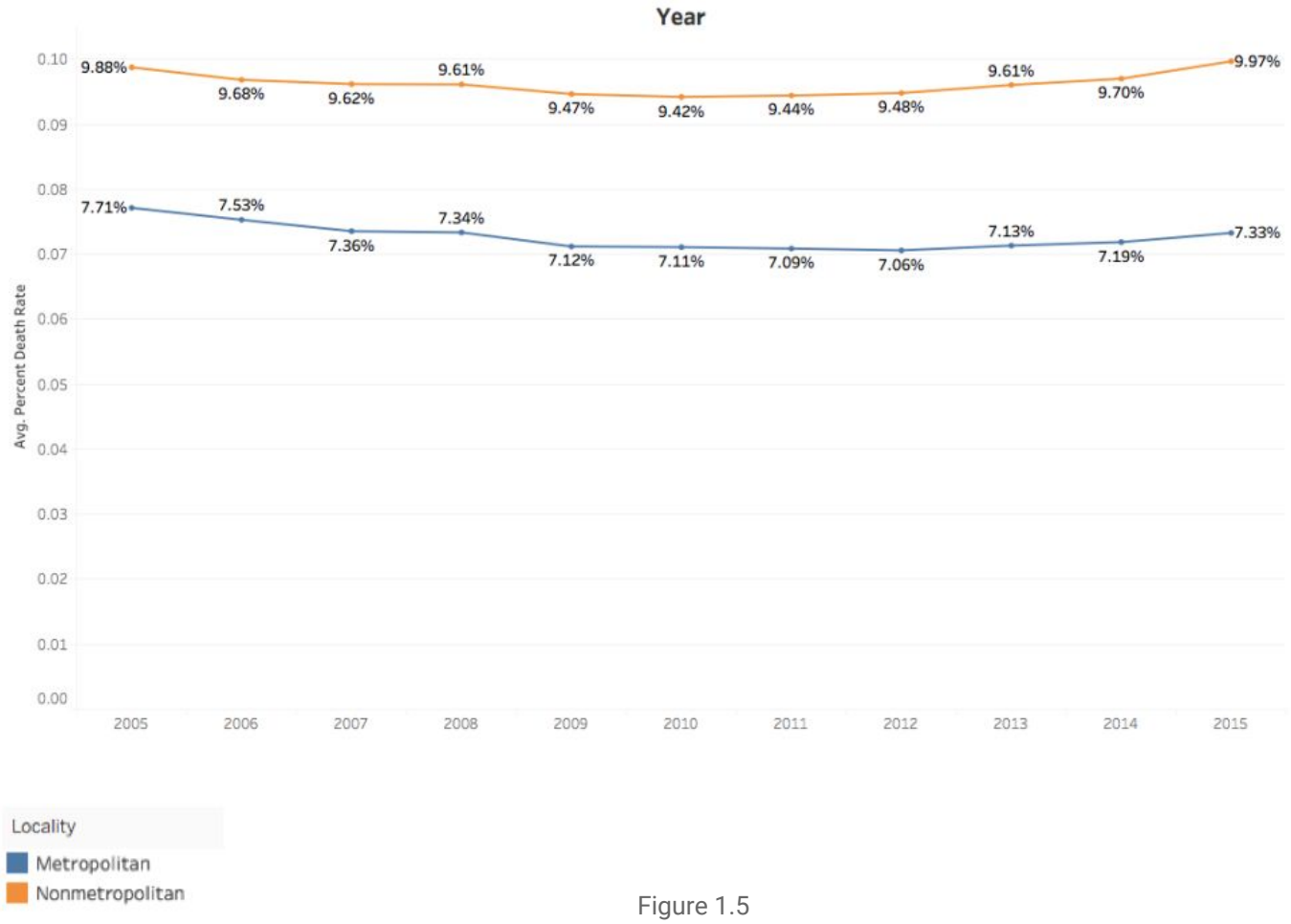

Figure 1.5

- The visualization (Figure 1.5) shows average death rate per year, per locality.
- The story here is that on average, nonmetropolitan areas have a higher death rate of about 2% than metropolitan areas.
- Residents of rural American tend to be older, have higher rates of cigarette smoking, high blood pressure, and obesity as compared to their rural counterparts.
- This shows that there is a gap in health between rural and urban Americans. In order to close this gap, there needs to be a better understanding to address health threats which puts rural Americans at increased risk of early death.

Explanatory Visualizations

After having the idea of our stories, we transition to building explanatory visualizations which are more sophisticated, detailed visualizations. These explanatory visualizations, highly polished, are the core visualizations of our final report. **We would like following five visualizations to be graded.**

Visualization 1: Interactive Dashboard - Potentially Excess Deaths

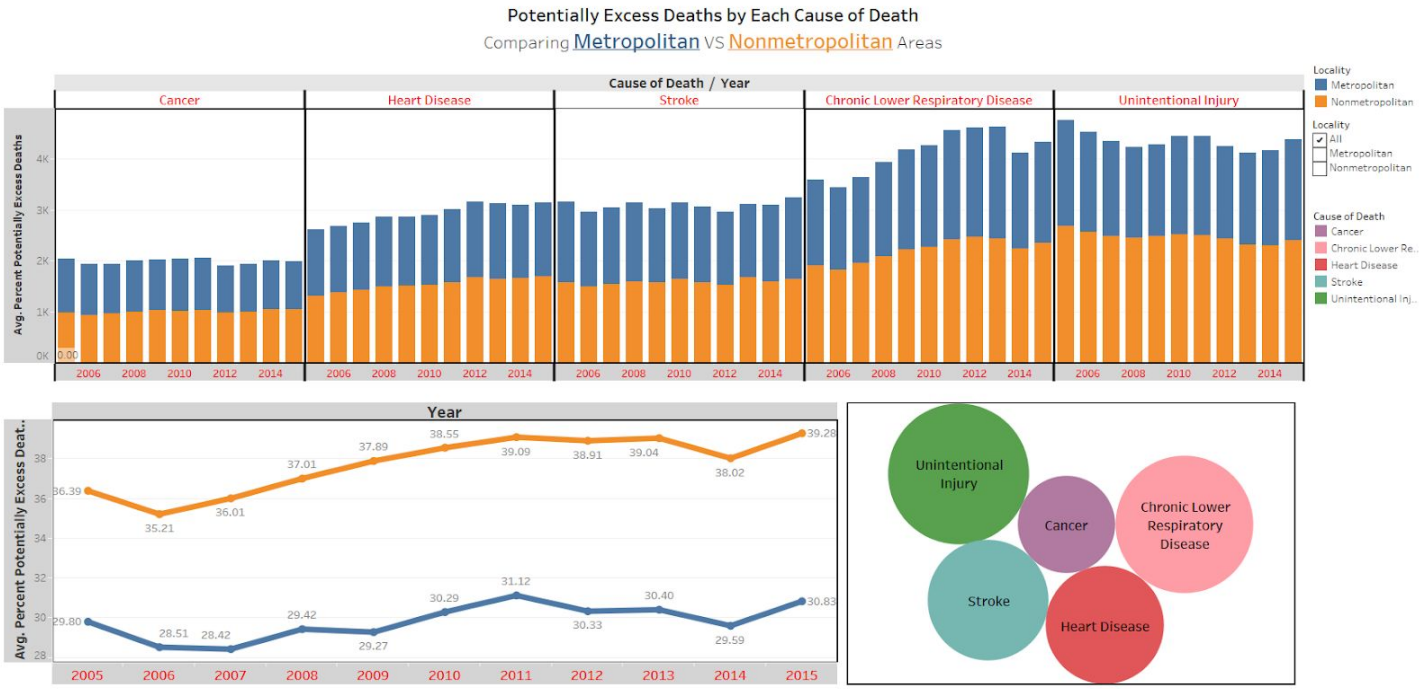

Figure 2.1 - Overall Dashboard

- For our first story, we are starting with the Interactive Dashboard in Tableau to examine the Percent potentially excess deaths, where we looked into the relationship between the excess deaths for each cause of death and each locality over time (i.e., how does the number of potentially excess deaths change for each cause of death and each locality (metropolitan vs nonmetropolitan).
- The first visualization (Figure 2.1) above is the snapshot of the overall dashboard and the (Figure 2.2) on the next page is the snapshot of the interactive element of the dashboard where brushing and linking is enabled for each cause of death.
- There are a total three types of graphs included in the dashboard (Stacked Bar Chart, Line Chart, and Bubble Chart).
- Blue color lines and bars represent Metropolitan areas, whereas Orange color lines and bars represent Nonmetropolitan areas.

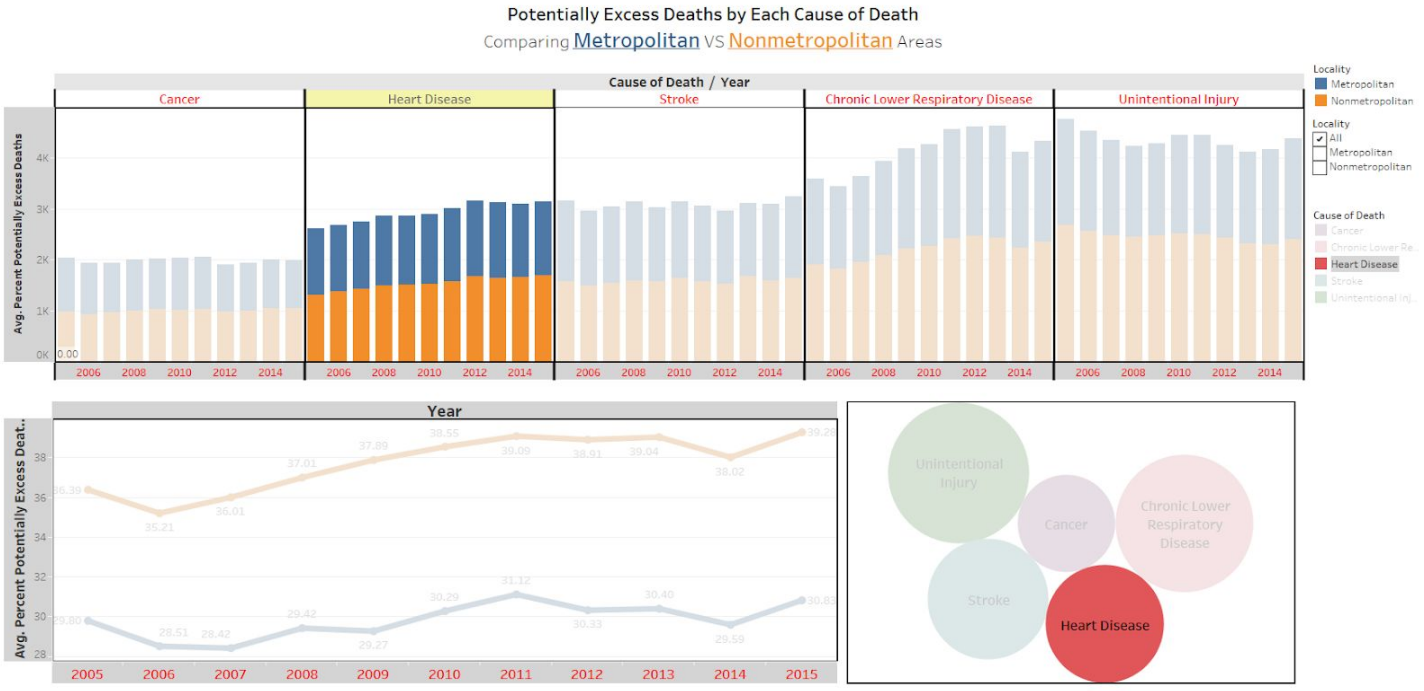

Figure 2.2 - Interactivity in Action - Refer to attached gif file[2]

- The Line Graph shows the trend of Avg. Percent Potentially Excess Deaths by each Locality (Metropolitan Vs Nonmetropolitan) from 2005 to 2015. We can clearly see that the nonmetropolitan areas have a significant number of potentially excess deaths compared to metropolitan areas overall.
- The bubble chart and stacked bar graph are comparing the Avg. Percent Potentially Excess Deaths for each Cause of Death. The highest potentially excess deaths are from "Unintentional Injury" and "Chronic Lower Respiratory Disease" followed by "Stroke", "Heart Disease", and "Cancer" for both metropolitan and nonmetropolitan areas.
- These potentially excess deaths for all causes except "Unintentional Injury" could be prevented through improved public health programs that support healthier behaviors and neighborhoods or better access to health care services.
- The deaths from "Unintentional Injury" can be prevented by implementing different measures such as vehicle safety, safe consumer products, medication management, safeguard home and community environments, and education & training.

[2] Interactive Dashboards and Animation

Visualization 2: Potentially Excess Deaths By Region

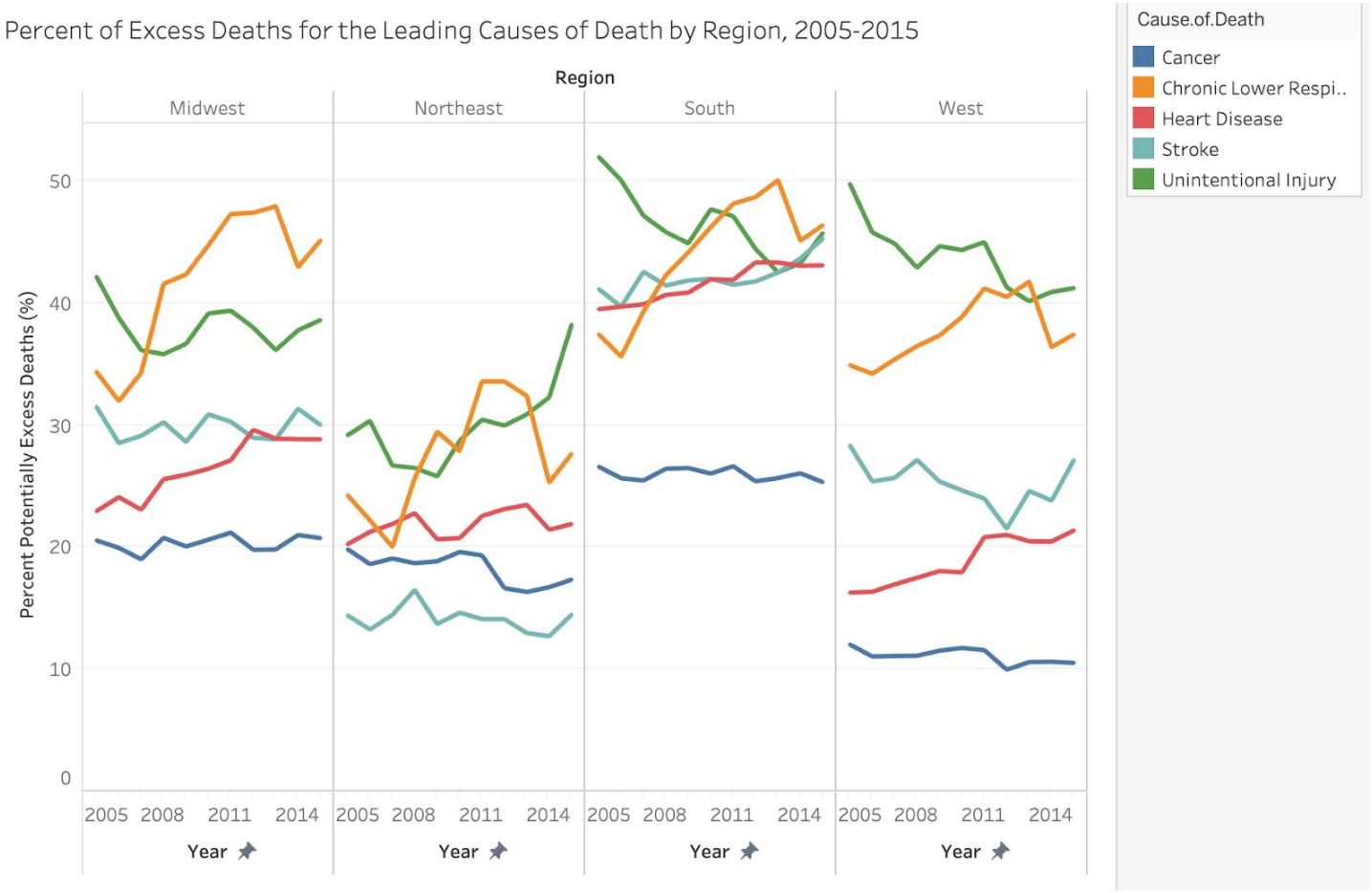

Figure 2.3

- In this visualization, we are looking at the Potentially Excess Deaths by each region in the USA from the year 2005 to 2015.
- The different colored lines represent the five leading causes of deaths.
- Chronic Lower Respiratory Disease and Unintentional Injury have higher excess deaths for all four regions. The lowest excess deaths are for Cancer except the Northeast region where the lowest excess deaths are for Stroke.
- There is a spike in Percent Excess Deaths for Chronic Lower Respiratory Disease for all four regions followed by a downtrend.
- There are not too many spikes noticed for excess deaths from Cancer throughout the period.
- Overall, the South region has the highest Potentially Excess Deaths compared to other regions.

Visualization 3: Interactive Dashboard - Percent Death Rate By Each State

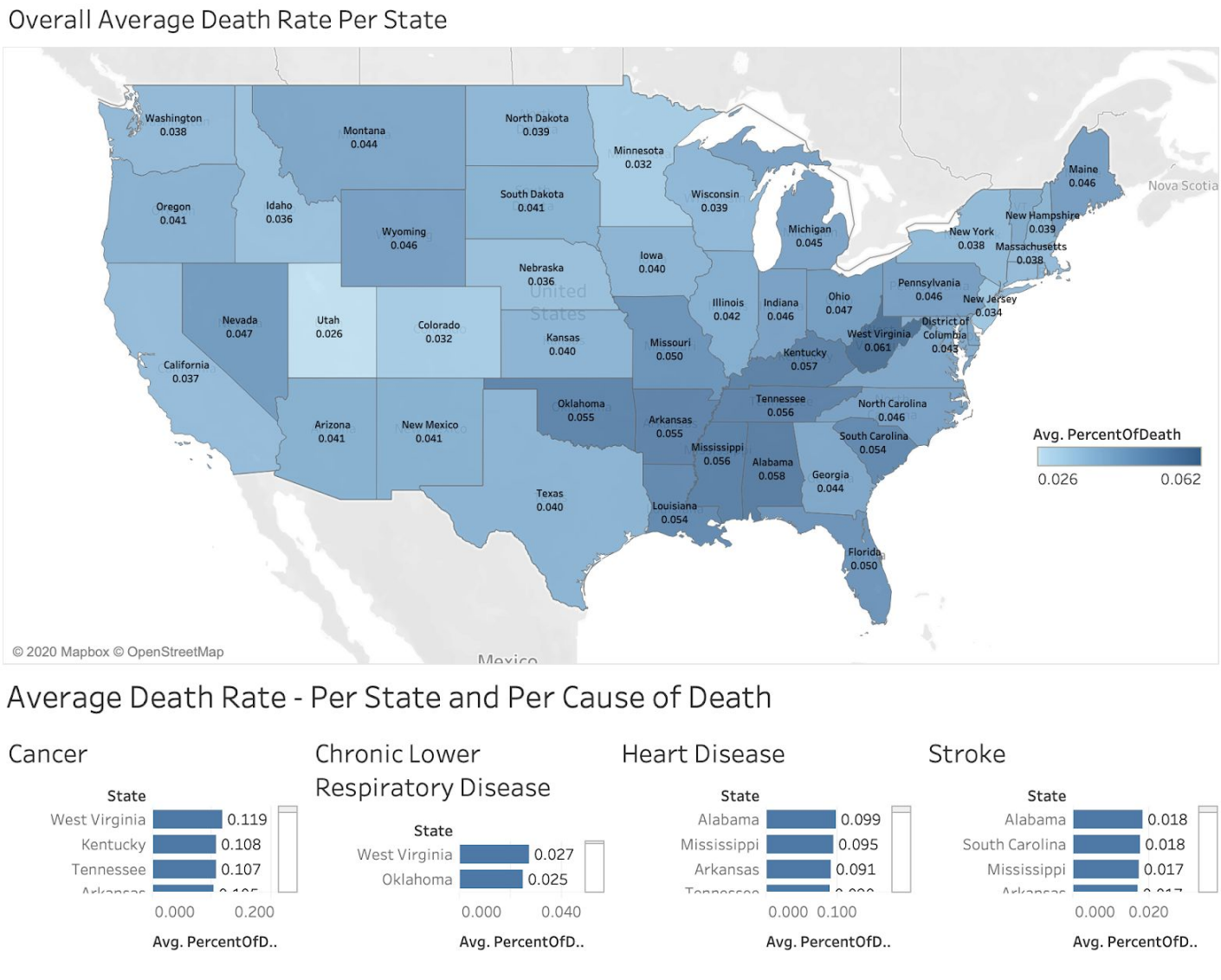

Figure 2.4 - Overall Dashboard

- For our second story, we are starting with the Interactive Dashboard in Tableau to examine the Percent Death Rate, where we looked into the relationship between the Death Rate per state.
- The (Figure 2.4) above shows the overall dashboard and the (Figure 2.5) below shows the interactive element of the dashboard.
- The choropleth shows the average percent rate of death per state. The color scheme was used to correspond with the average percent rate of death (i.e., a lighter color meant the state had a lower average and a darker color meant the state had a higher average).
- The horizontal bar graphs show the average percent rate of death per state per cause of death.

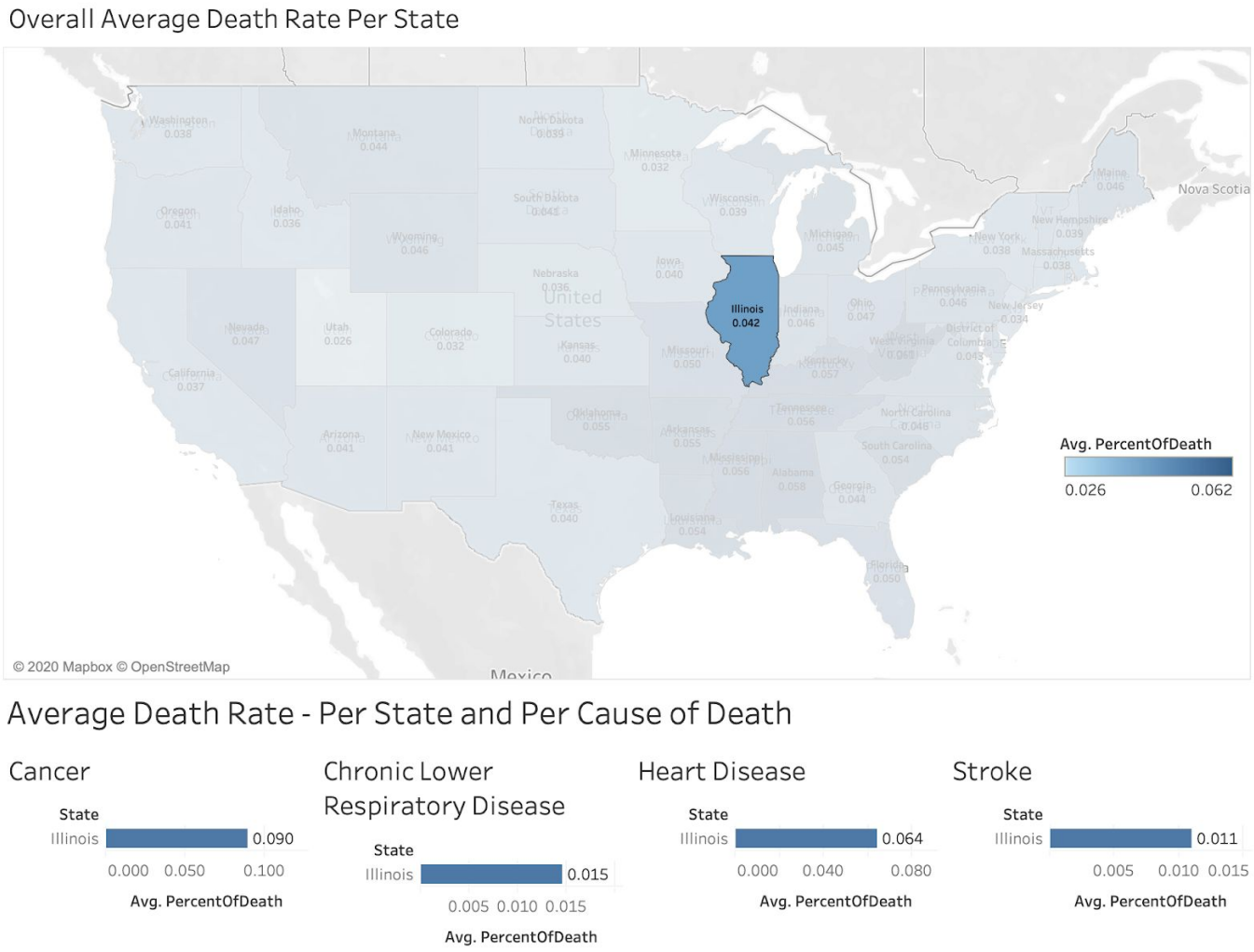

Figure 2.5 - Interactivity in Action - Refer to attached gif file[3]

- Interactivity: when the user scrolls over a state in the map and clicks a state, the average percent rate of death for that state per cause of death will be highlighted in the horizontal bar graphs.
- My audience:
  - A health official or a state official with a scientific background who is trying to evaluate the efficiency of his/her/their own state's health initiatives and act accordingly.
  - A regular citizen who is trying to understand how to make lifestyle changes to decrease their chances of dying due to their state's highest cause of death.
- My message here is that the southeast region of the United States has the highest average percent rate of deaths, and health officials and residents of this region have to be more cautious when it comes to their state's corresponding highest cause of death. Moreover, these could be prevented by the state government through improved public health programs or better access to health care services.

[3] Interactive Dashboards and Animation

Visualization 4: Comparison

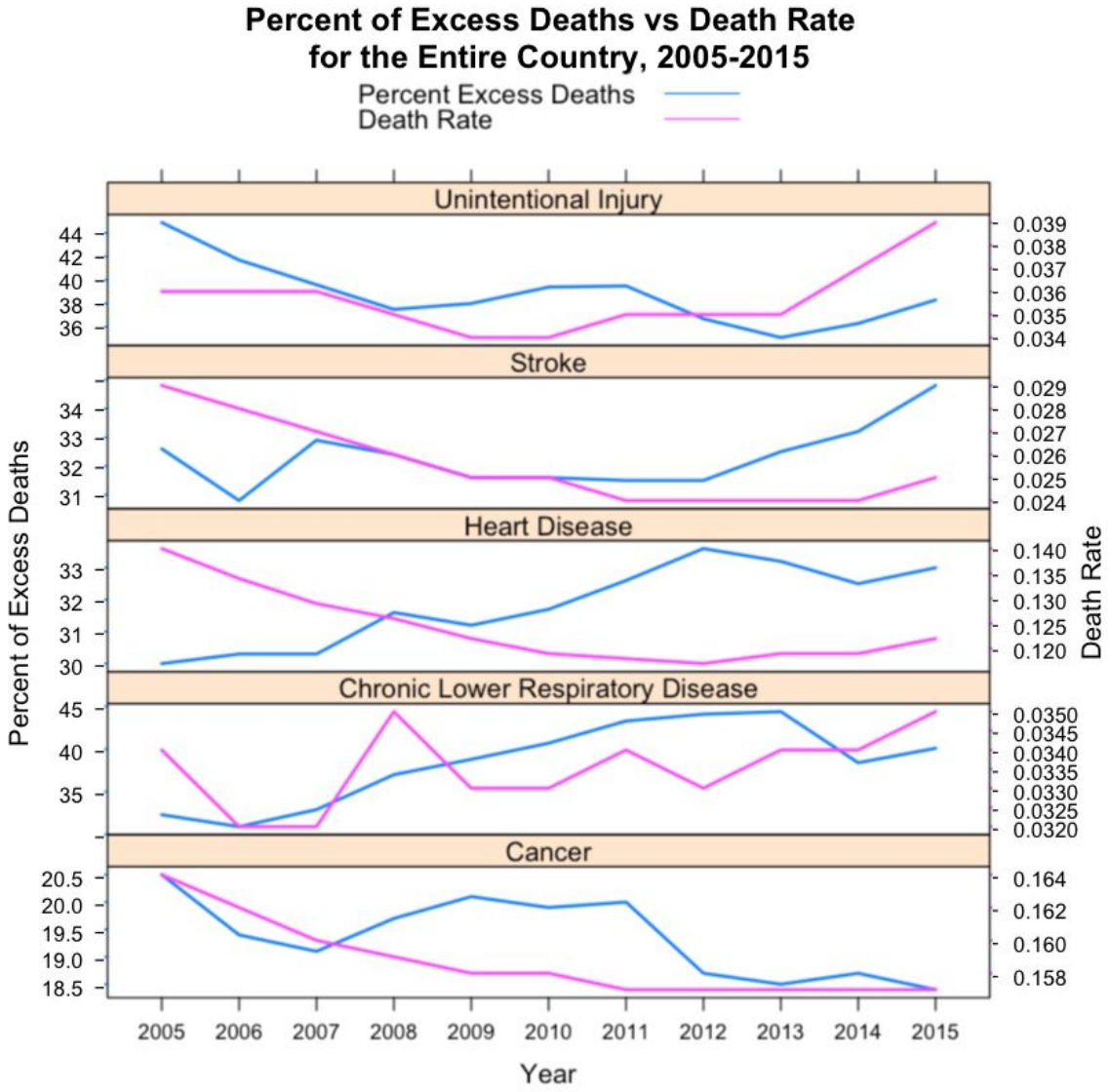

Figure 2.6

- For the double axis panel plot above (Figure 2.6), the data was filtered by Age Range = 0-94, Locality = All, Benchmark = Floating, and State = United States.
- Death Rate is calculated by dividing Observed Deaths by Population and multiplying by 100.
- This is done so we can see the overall trend. of Percent of Excess Deaths and Death Rate for the whole country.
- We can see that the Death Rate and Perc. of Excess Deaths caused by Cancer decrease over time.
- Interestingly, The Percent of Excess Deaths caused by Stroke increases over time, while the Death Rate for Stroke decreases.
- Comparing these two trends is interesting as they tell different stories, so looking at one over the other could be misleading.

Visualization 5: Comparison Using Animation

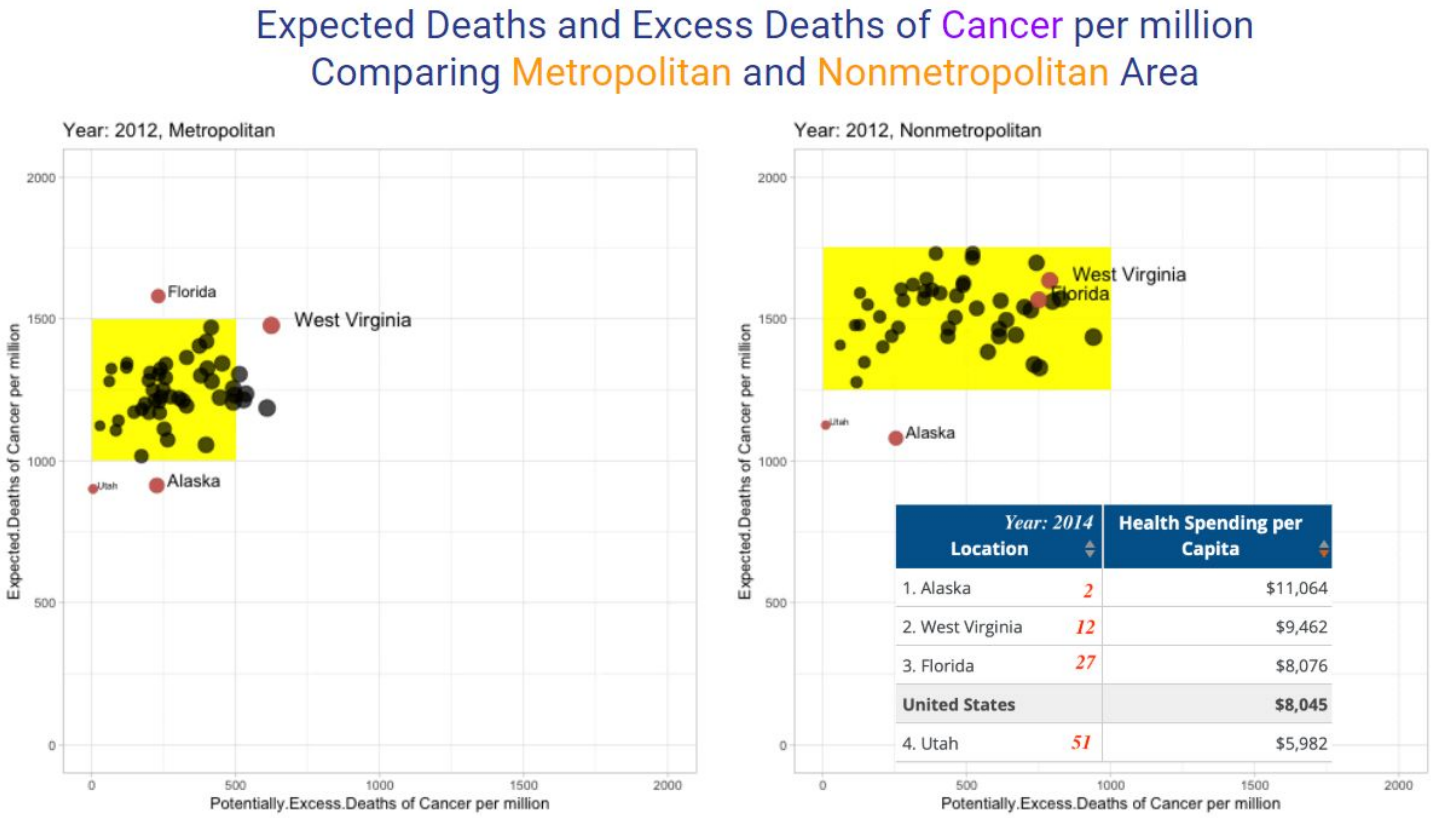

Figure 2.7 - Animation - Refer to attached gif file[4]

- This animation scatter plot shows the changing of excess deaths and expected deaths between metropolitan and nonmetropolitan areas in each state from 2005 till 2015. Both x-axis and y-axis represent the number of deaths in a million population.
- Changes happened mostly in the yellow highlight area. Compared to the highlight area, rural areas tend to have more deaths in both expected deaths and excess deaths. According to demographics, residents of rural areas tend to be older, poorer, sicker and doing less physical activity than their urban counterparts.
- Utah and Alaska have lower expected deaths and excess deaths and remain stable in these years. But Florida and West Virginia have higher deaths. The black colored point represents other states. They all remain relatively stable in expected deaths but num of excess deaths goes up and down by years. Utah and Alaska can keep deaths num stable and low is because both state residents have healthy lifestyle choices, Utah has the lowest percentage of adult smokers in the nation. And both states have a health program to provide free or low cost cancer exams and screening.
- One point for sure, there are risk factors that people can reduce by changing their lifestyle.

[4] Interactive Dashboards and Animation

Analysis & Discussion

Story-1: Potentially Excess Deaths

- ➢ Nonmetropolitan areas have a significant number more compared to metropolitan areas.
- ➢ Leading causes of Death:
  - Unintentional Injury
  - Chronic Lower Respiratory Disease
  - Stroke
  - Heart Disease
  - Cancer
- ➢ The Southeast region of the USA has the greatest percentage of 45%.
  - Use this imbalance distribution to focus on these certain areas and implement more research into causes and solutions.

Story-2: Death Rate

- ➢ Cancer is the leading cause of death.
- ➢ Alabama has the greatest average death rate for Heart Disease and Stroke.
- ➢ West Virginia has the greatest average death rate for Cancer and Chronic Lower Respiratory Disease. This goes along Story 1's conclusion regarding the greatest percentage in the Southeast region.
- ➢ On average, nonmetropolitan areas have a higher death rate of about 2% than metropolitan areas.

Story-3: Comparison

- ➢ We can observe an upward trend in Average Observed Deaths per Year.
- ➢ There is also a downwards and then sharp upward trend in 2014 in the Average of Potentially Excess Deaths per Year.
- ➢ For both, cancer is the leading cause of death, followed by heart disease.

Statistical Techniques Suggested to Data Analyst

- Mean and standard deviation of observed deaths, expected deaths, potentially excess deaths, percent potentially excess deaths, percent death Rate per cause of death, year, state, and locality.
- Regression analysis to find the trend between one dependent variable (death rate in metropolitan area) and one or more independent variables (cause of death, age).
- Correlation analysis to determine co-relationship of two quantitative variables (population v death rate).

If we had more time….

- Age distribution among leading causes of death, observed deaths, death rate.
- Further analysis and visualization among expected deaths vs observed deaths.
- Percentage of population change among the leading causes of death and death rate.
- Combining datasets to get a granular look at distribution of healthcare coverage, cost of healthcare, race, education, etc.

Appendices

R Code for Figure 2.6

```
library(latticeExtra)

# -- import data
deaths <- read.csv("deaths.csv")

# -- filter data so it focuses on entire country, all age groups, all localities, and floating benchmark
filtered <- deaths[ which(deaths$State == 'United States' & deaths$Age.Range == '0-84' &
deaths$Benchmark=='Floating' & deaths$Locality =='All'),]

# -- create Death Rate Percent variable, which is simply (Observed Deaths / Population * 100)
filtered$Death.Rate <- round((filtered$Observed.Deaths / filtered$Population) * 100, digits=3)

# -- create two sets of line graphs, each with 5 individual graphs showing the trends for the 5 different causes of death
excessDeaths <- xyplot(Percent.Potentially.Excess.Deaths ~ Year | Cause.of.Death,filtered,type = "l", scales = list(y="free",x=list(tick.number=10)), lwd=2, ylab="Percent of Excess Deaths",col.axis="blue", main="Percent of
```

```
Excess Deaths vs Death Rate \n for the Entire Country, 2005-2015")
deathRate <- xyplot(Death.Rate ~ Year | Cause.of.Death, filtered, scales =
list(y="free",x=list(tick.number=10)),type = "l", lwd=2, ylab="Death Rate")
 # -- combine sets of graphs with two y axis and a legend, to compare them
chart <- doubleYScale(excessDeaths, deathRate, text = c("Percent Excess Deaths", "Death Rate"),add.ylab2 =
TRUE)

# -- stack the graphs on top of each other to make panel plot
update(chart, layout=c(1,5))
```

R Code for Figure 2.7

```
library(gganimate)
library(gifski)
library(gapminder)
library(dplyr)
options(scipen=200)

# -- import data
data<- read.csv('deaths.csv')

# -- create excess deaths and expected deaths per million population
data$excess_death_rate<-data$Potentially.Excess.Deaths/data$Population*1000000
data$expected_death_rate<-data$Expected.Deaths/data$Population*1000000

# -- filter data
  ## --for Metropolitan
data_cancer<-data%>%filter(Locality=='Metropolitan')
data_cancer<-data_cancer%>%filter(Cause.of.Death=='Cancer')
highlight<-data_cancer%>%filter(State=='Florida'|State=='West Virginia'|State=='Alaska'|State=='Utah')

  ## --for Nonmetropolitan
data_cancer_n<-data%>%filter(Locality=='Nonmetropolitan')
data_cancer_n<-data_cancer_n%>%filter(Cause.of.Death=='Cancer')
highlight_n<-data_cancer_n%>%filter(State=='Florida'|State=='West Virginia'|State=='Alaska'|State=='Utah')

# -- create two sets of animation scatter plots

## --for Metropolitan
plot_cancer <- ggplot(data_cancer,
```

```
  aes(x = excess_death_rate, y=expected_death_rate, size = Percent.Potentially.Excess.Deaths))+
  geom_rect(aes(xmin=0,xmax=500,ymin=1000,ymax=1500),fill='yellow',alpha=0.2,show.legend = FALSE)+
  geom_point(show.legend = FALSE, alpha = 0.7) +
  geom_point(data=highlight,aes(x = excess_death_rate,
y=expected_death_rate,color='red'),alpha=0.7,show.legend = FALSE)+
  geom_text(data=highlight,aes(label=State),hjust=-0.2, vjust=0,show.legend = FALSE)+
  scale_size(range = c(2, 5))+
  xlim(0,2000)+
  ylim(0,2000)+
  theme_light()+
  ggtitle('the Expected Deaths and Excess Deaths of Cancer Changes by years')+
  labs(x = "Potentially.Excess.Deaths of Cancer per million", y = "Expected.Deaths of Cancer per million")
metro<-plot_cancer+transition_time(Year) +
  labs(title = "Year: {frame_time}, Metropolitan")
animate(metro,nframes = 120,renderer = gifski_renderer('cancer.gif'))

## --for Nonmetropolitan
plot_cancer_n <- ggplot(data_cancer_n,
                        aes(x = excess_death_rate, y=expected_death_rate, size = Percent.Potentially.Excess.Deaths))+
  geom_rect(aes(xmin=0,xmax=1000,ymin=1250,ymax=1750),fill='yellow',alpha=0.2,show.legend = FALSE)+
  geom_point(show.legend = FALSE, alpha = 0.7) +
  geom_point(data=highlight_n,aes(x = excess_death_rate,
y=expected_death_rate,color='red'),alpha=0.7,show.legend = FALSE)+
  geom_text(data=highlight,aes(label=State),hjust=-0.2, vjust=0,show.legend = FALSE)+
  scale_size(range = c(2, 5))+
  xlim(0,2000)+
  ylim(0,2000)+
  theme_light()+
  ggtitle('the Expected Deaths and Excess Deaths of Cancer Changes by years')+
  labs(x = "Potentially.Excess.Deaths of Cancer per million", y = "Expected.Deaths of Cancer per million")
nonmetro<-plot_cancer+transition_time(Year) +
  labs(title = "Year: {frame_time}, Nonmetropolitan")
animate(nonmetro,nframes = 120,renderer = gifski_renderer('cancer_n.gif'))
```