Name: Sharik Shaik

MID: M15396597

HI-7072 – Final Project

Exploring Potential Excess Deaths in the United States

Structured data and its source:

I selected the dataset "NCHS__Potentially_Excess_Deaths_from_the_Five_Leading_Causes_of_Death.csv" which is from the website https://healthdata.gov/dataset/NCHS-Potentially-Excess-Deaths-from-the-Five-Leadi/yt9r-6btk.

Why this structured data:

This dataset was selected because it has a lot of information on potential excess deaths and also percentage of this potential excess deaths associated with the five main causes of death over a different number of years in various states and regions. The data's temporal and geographical characteristics make it appropriate for analysing patterns and trends in death rates in different states or regions.

So, with this dataset I can get to know the trends and patterns of the expected and potential excess deaths over the years in the United States which included the five different causes like Cancer, Heart disease, stroke, chronic lower respiratory disease, unintentional injury.

Unstructured data and its source:

I have selected this unstructured data “cdcgov.csv” from the same website from which I have got the above structured data. The source of the data is “https://healthdata.gov/” website as well. I thought to use the same dataset as above but it doesn’t have unstructured data. So, I have to find the other csv file from the same website for unstructured data.

Why this unstructured data:

I have selected this data because it has the text data related to cancer deaths in it. So, with this text data in one of the columns of the csv file I can do the preprocessing methods and prepare it for the visualizations like bar plot, word cloud etc.

Variables that are captured in the data and their datatypes:

Year (Integer): This represents the year for which the data is recorded.

Cause of Death (Character/String): This represents the particular cause of death that are Cancer, Heart disease, stroke, chronic lower respiratory disease, unintentional injury.

State (Character/String): This refers to the geographic location (state) where the data is recorded.

State FIPS Code (Character/String): This refers to Federal Information Processing Standards code for the state which is 2 letter field, a standardized numeric code to uniquely identify states.

HHS Region (Integer): This represents the region number assigned by the U.S. Department of Health and Human Services.

Age Range (Character/String): This tells the range of ages for which the data is aggregated.

Benchmark (Character/String): This is possibly a benchmark used in the analysis.

Locality (Character/String): This represents specific area like metropolitan, non-metropolitan of that particular state.

Observed Deaths (Integer): This is the actual number of deaths observed for a particular cause, state.

Population (Integer): This is a population count for a specific state.

Expected Deaths (Integer): This is expected number of deaths based on certain benchmarks or predictive models.

Potentially Excess Deaths (Integer): This represents the difference between observed deaths and expected deaths, representing deaths that may be considered "excess."

Percent Potentially Excess Deaths (Numeric/Float): This is the percentage of potentially excess deaths according to the expected deaths.

Questions I am going to answer with my visualizations in R and tableau:

- What are the trends in potentially excess deaths over the years?
    - Data visualization Planned: **Line chart** in Tableau.

- Are there specific age ranges where potentially excess deaths being more common in each death cause?
    - Data visualization Planned: **Bar chart** in Tableau.

- Is there a correlation between observed deaths and expected heart disease deaths?
    - Data visualization Planned: **Scatter plot** both in R and Tableau.

- What are the number of excess deaths in each state for each cause?
    - Data visualization Planned: **Text Table** in Tableau.

- How do expected deaths and observed deaths compare across different States in US?
    - Data visualization Planned: **Map Plot** in Tableau

- What are the top 5 states that are having most potential excess deaths due to different causes?
    - Data visualization Planned: **Bar Chart** in R

- What are the number of deaths in non-metropolitan locality of each state in the year 2013?
    - Data visualization Planned: **Map plot** in R

- What are the prevalent themes or topics mentioned in the descriptions?
    - Data visualization Planned: **Bar plot** and **word cloud** of most frequent words in R.

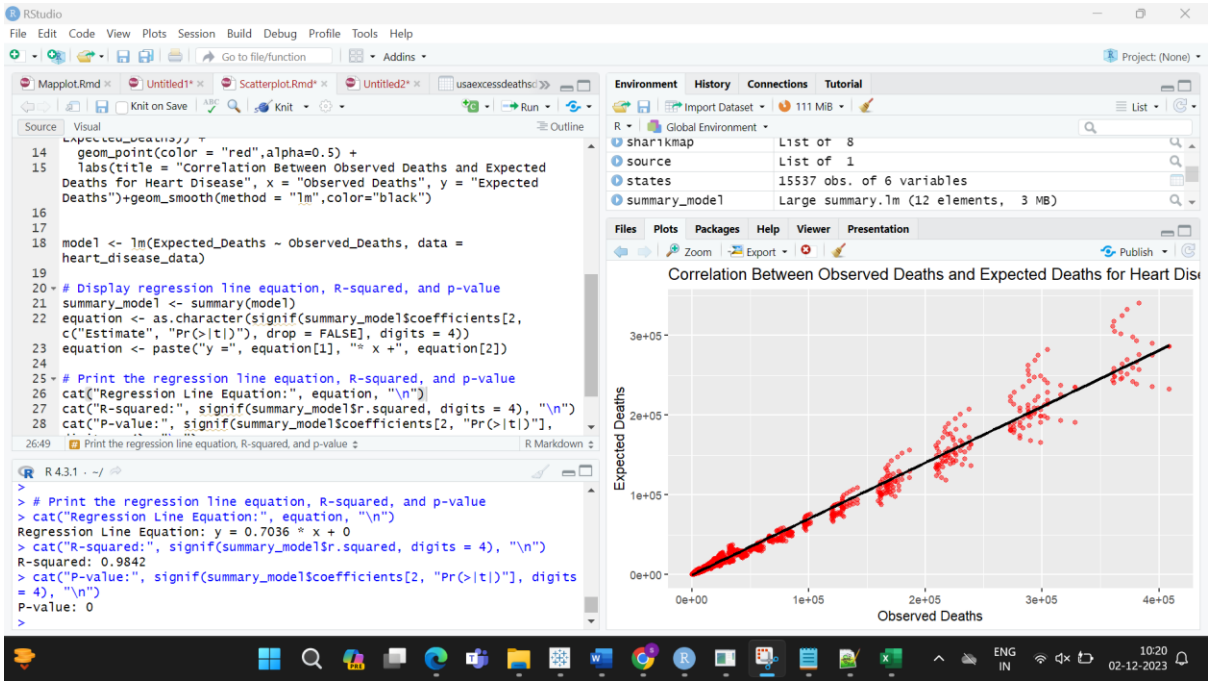
Data cleaning for analysis:

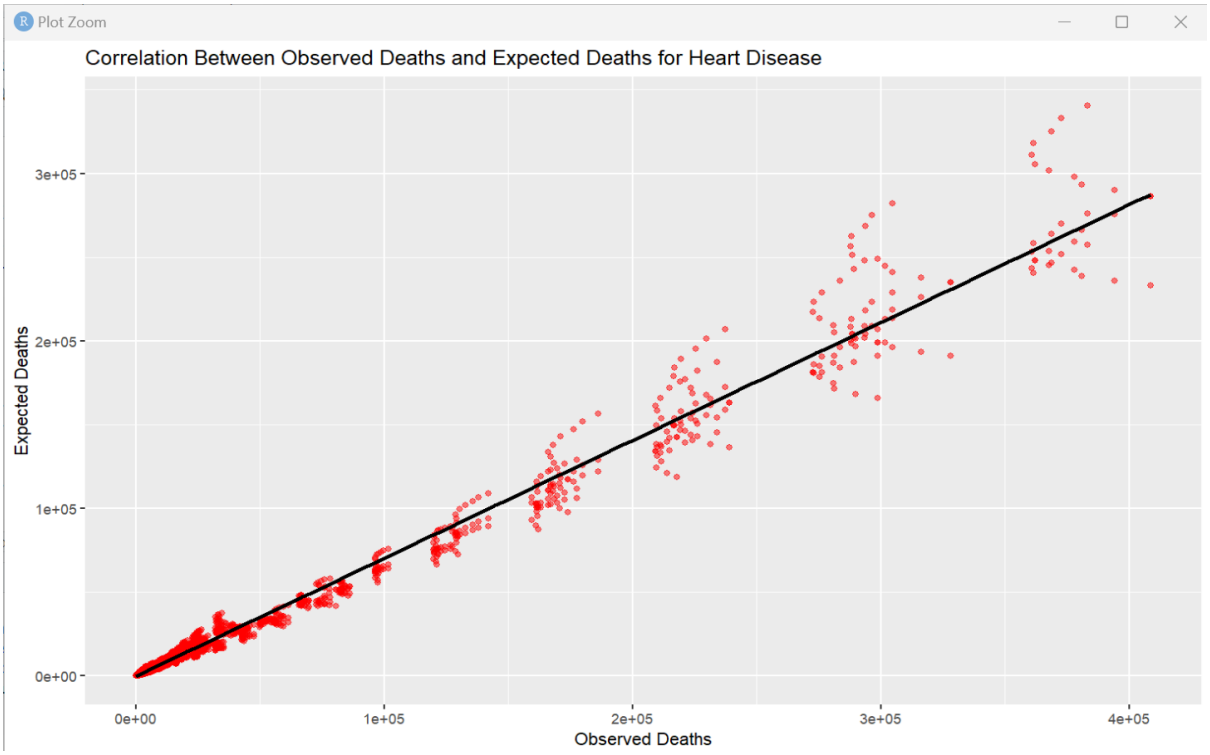
For unstructured data I have performed few data cleaning tasks which includes removing stop words, punctuations, whitespaces, sorting etc. As in the unstructured data csv file I have a column with required unstructured data, where I performed the above-mentioned operations to perform analysis and plot the bar plot of most frequent words and word cloud which has the most used words with these words, we can get know the results like which diseases are frequently occurring and also, we can know the words that are closely related to that disease.

For structured data, to plot a map, as I have a 200k records which is a very large dataset. It was taking a lot of time to read and visualize the data. Due to this reason, I have applied some filters on the required columns to reduce the data and perform the data analysis and visualization tasks. Though this is not a data cleaning process, but I need to filter the records for the faster processing and get the accurate results.

R Visualizations:

1:

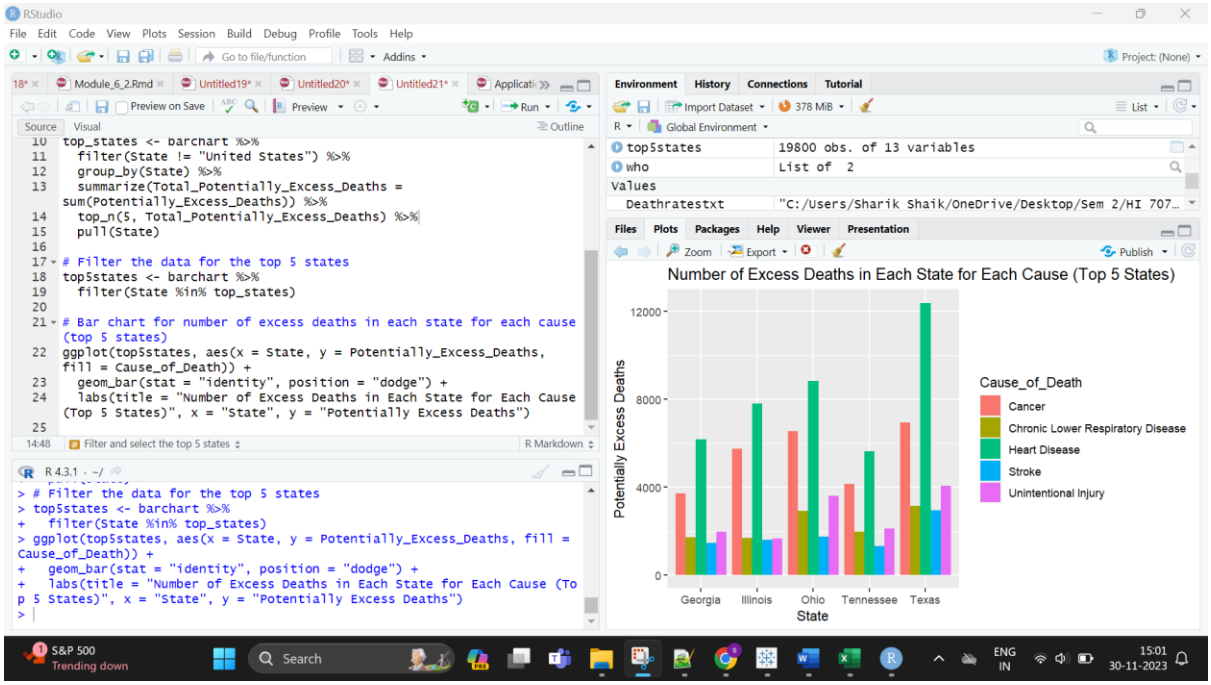


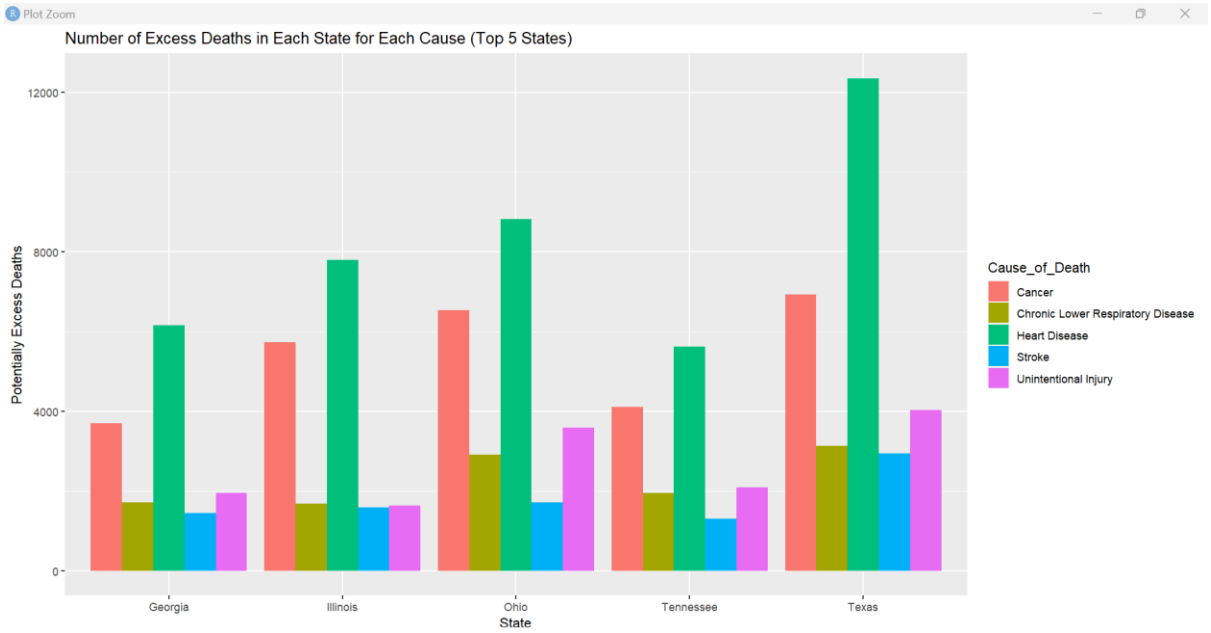


The above scatterplot in R answers the question “Is there a correlation between observed deaths and expected heart disease deaths?”

The regression analysis for the scatterplot comparing expected and observed deaths in cases of heart disease reveals a robust and statistically significant linear relationship. The regression line equation, y = 0.7036 * x + 0, signifies that, on average, each incremental unit increase in observed deaths corresponds to a 0.7036 unit increase in expected deaths. The intercept of 0 implies that when observed deaths are zero, the expected deaths are also zero. The high R-squared value of 0.9842 indicates that approximately 98.42% of the variability in expected deaths can be explained by the linear relationship with observed deaths. The statistical significance of this link is further highlighted by the remarkably low p-value of 0, which indicates that it is extremely unlikely to be the result of chance. Basically, this study highlights a significant and pertinent correlation between observed and projected deaths from heart disease, giving helpful details about the observed deaths variable's predictive ability in this particular context.

2:

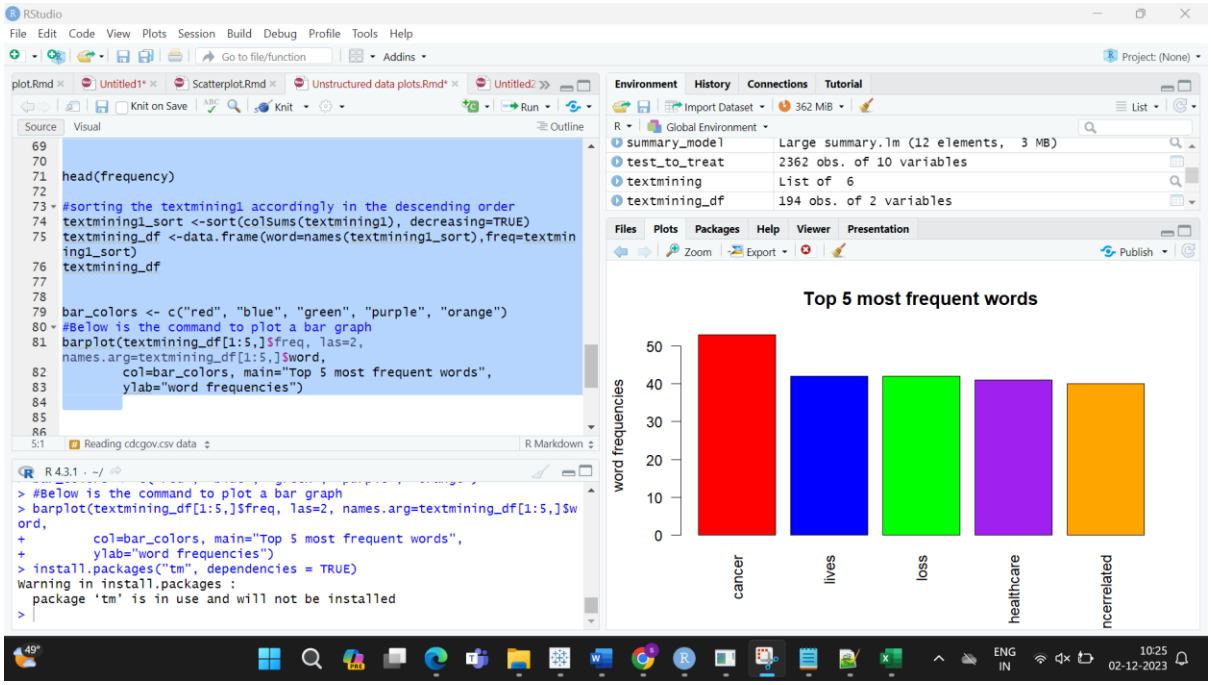




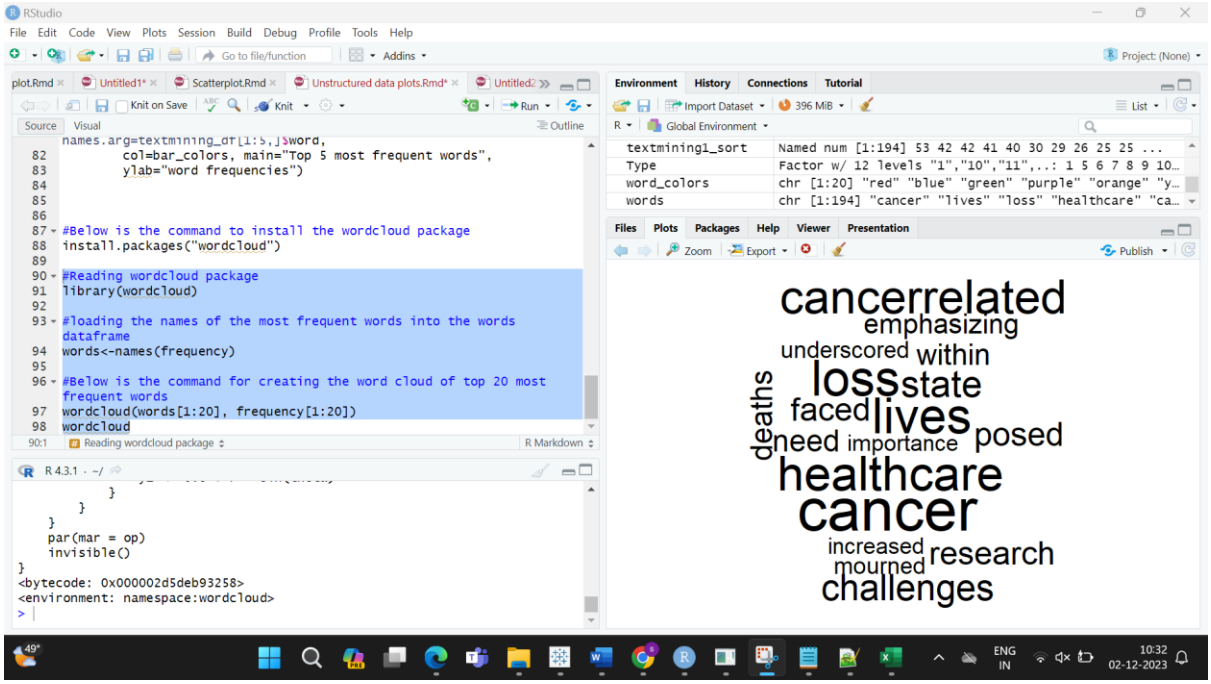
The above bar plot in R answers the question “What are the top 5 states that are having most potential excess deaths due to different causes and what decisions could be taken with the help of this plot?”

This plot represents the top 5 states with most potential excess deaths. In each of the above mentioned 5 states there is a high bar for the heart disease which indicates that in every state there are high excess deaths due to the heart disease from which we can say that there needs to be more increment in hospital equipment and also more availability of doctor appointments to treat patients and also provide them the prescription accordingly so that for some extent we can reduce/prevent these heart diseases in these states. In the same way we can follow for the other diseases as well accordingly.

3:



4:

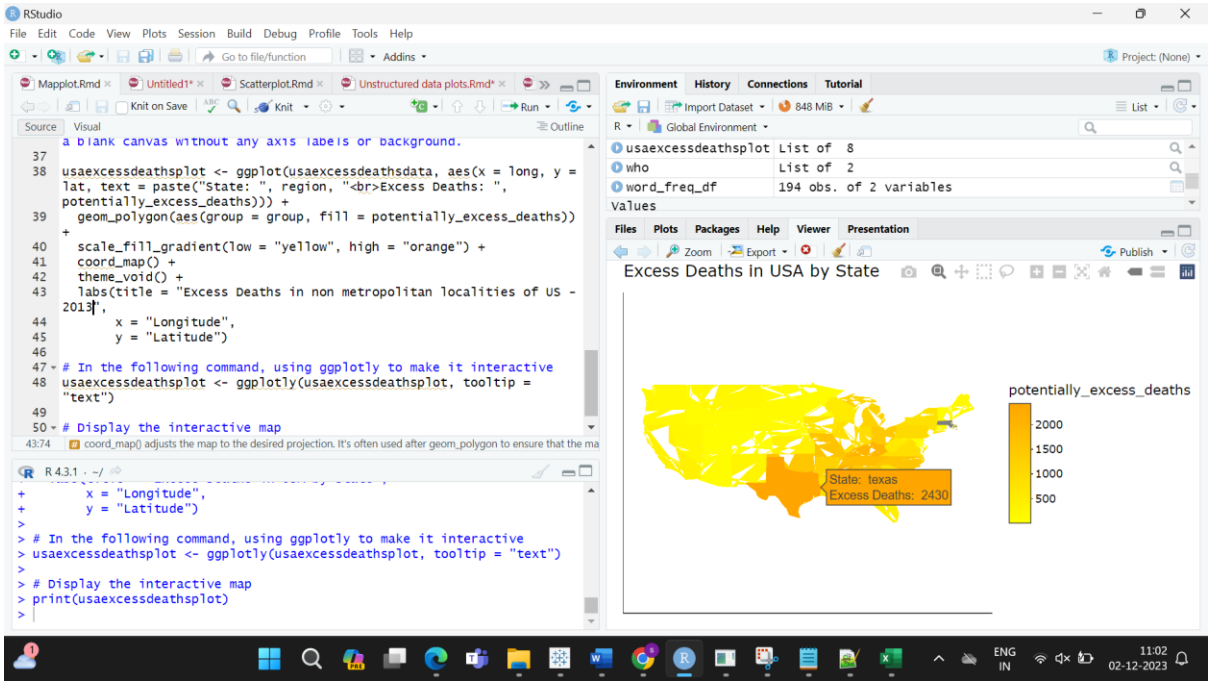


The above 2 visualizations are related to the unstructured data and they answer the question “What are the prevalent themes/topics mentioned in the descriptions, how the plots are helpful?”

So, the above visualizations are bar plot, word cloud for more frequent words. With this we can know that this unstructured data is more relatively about the disease called cancer. So, with this we can check the other words and relate them to what are the requirements that are needed. If we want a

more advanced analysis, we can consider the topic modelling techniques such as Latent Dirichlet Allocation (LDA) to discover latent topics within the text data. We can analyse the word cloud and bar chart to identify clusters of related terms or common themes. Look for recurring words that provide insights into the most discussed aspects of cancer-related descriptions.

5:



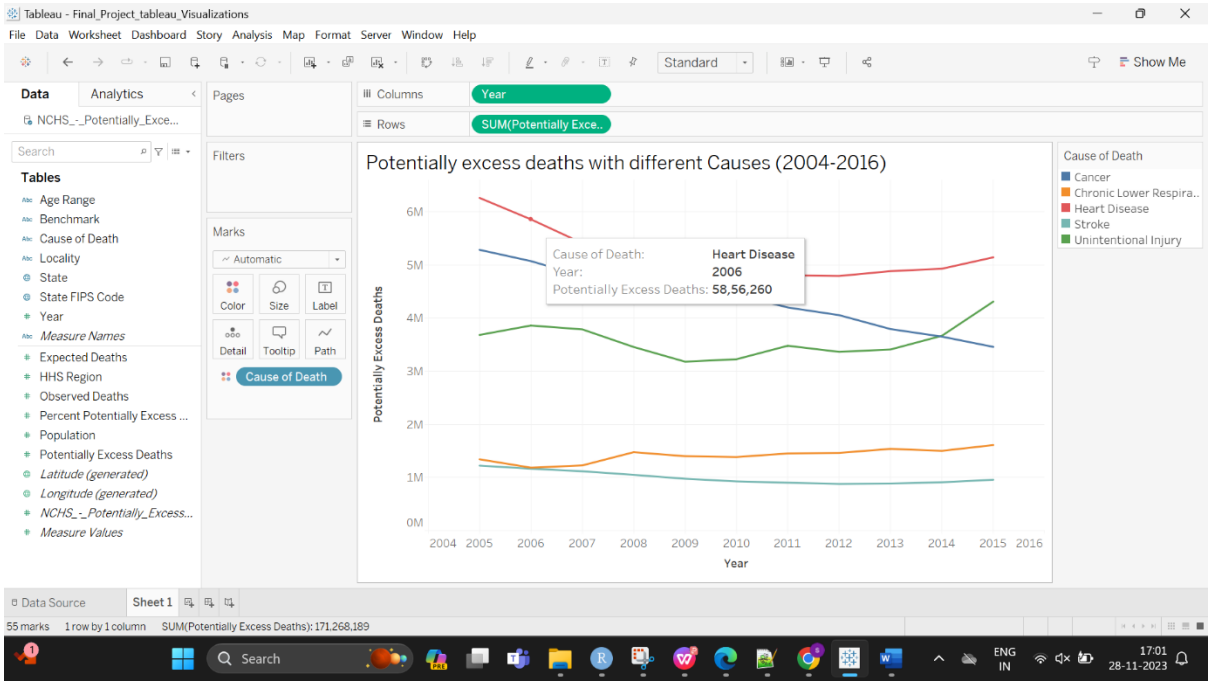
The above map plot answers the question “What are the number of deaths in non-metropolitan locality of each state in the year 2013? “

Using the above map plot, I am showing the number of deaths in non-metropolitan locality of each state in 2013. So, when we place the pointer i.e., hover on the map on any of the states, we can see the state name and the excess deaths count on it. So, to know the number of excess deaths for a particular state in non-metropolitan localities in the year 2013 we can just place the pointer on that state. The white spots that are visible on the map are the places where there are no excess deaths.

To plot this first I have filtered my dataset in excel according to the requirements which includes the data of year 2013, non-metropolitan locality of each state.

Tableau visualizations:

1:

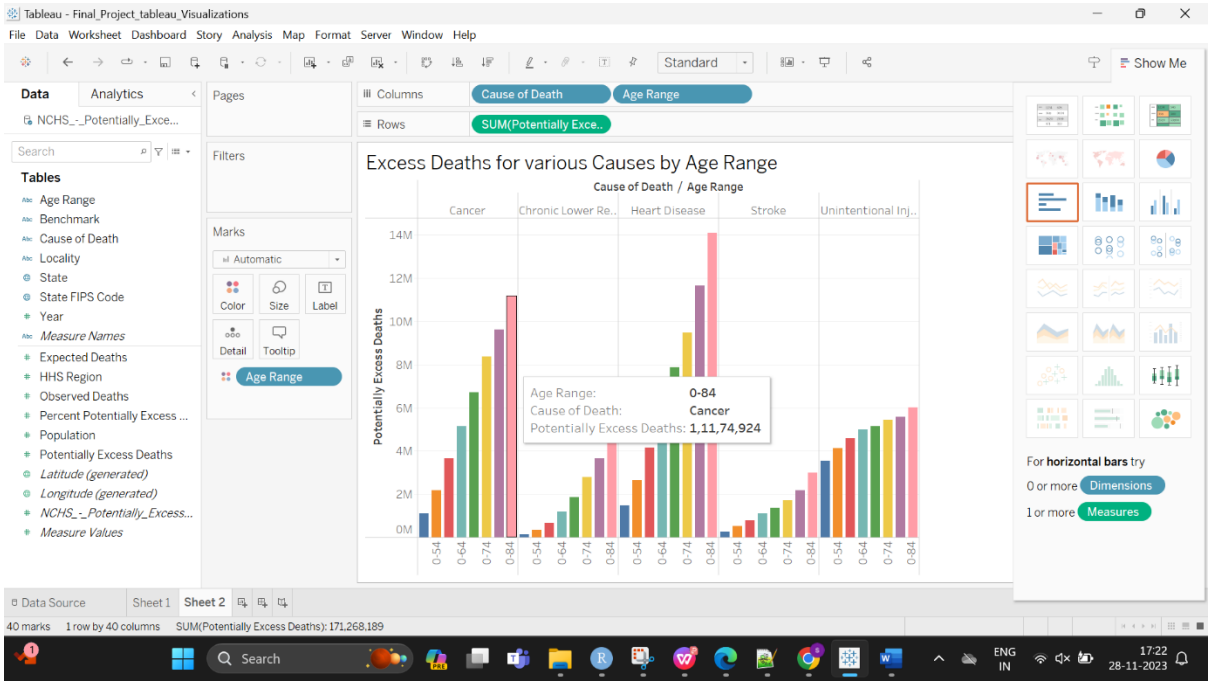


This visualization is helpful to answer the first question. “What are the trends in potentially excess deaths over the years?”

This line chart represents the potentially excess deaths over the years (2005-2015). Here we can see that there are more excess deaths due to heart disease all these years, where it started decreasing after 2005 but again started increasing in 2012. And the lowest death cause is stroke, where it has been almost constant. Regarding Cancer it seems to be decreasing all these years which seems to be good. The other 2 causes mostly have fluctuations over the years but seems like they started increasing from 2013.

So, with this visualization I have shown that, when we point on any colour line at any particular year it shows how many potential deaths happened in that particular year. Each colour represents the different cause of death.

2:

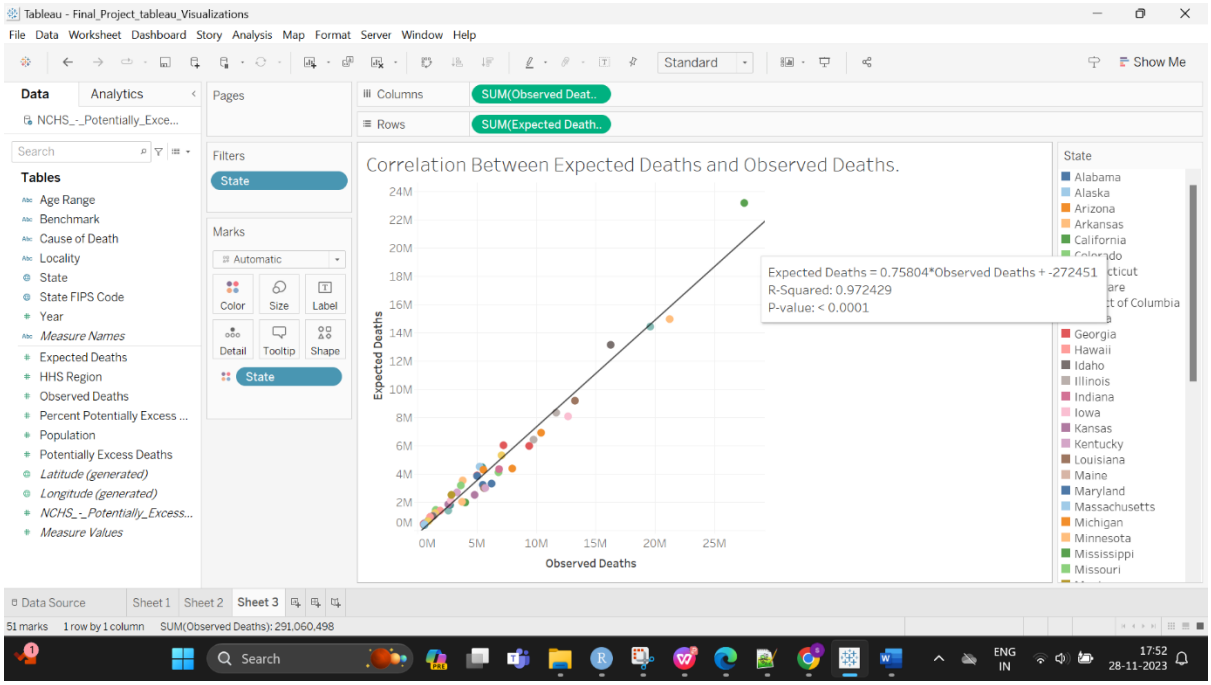


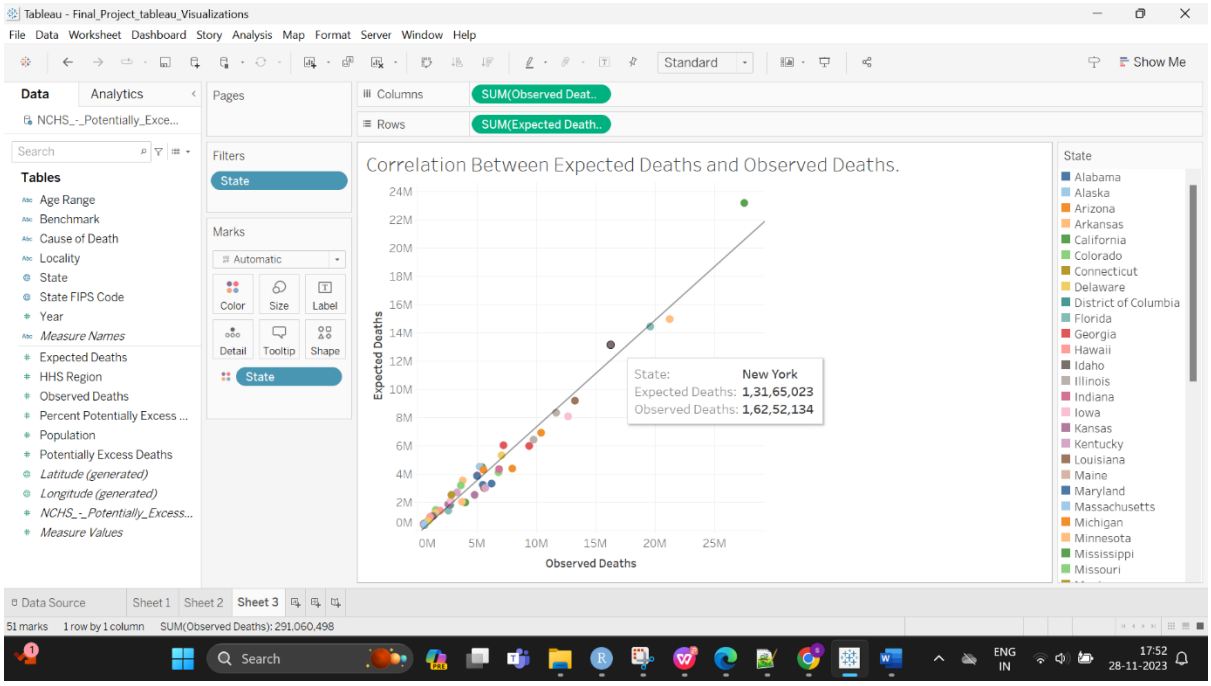
This visualization is helpful to answer the question. “Are there specific age ranges where potentially excess deaths being more common in each death cause?”

This bar plot shows the excess deaths due to different causes for different age groups. As shown in the plot we can see that in every cause of death when there is increase in age there is a chance of increase in deaths. Un-surprisingly when we are growing old there are more chances of getting more diseases and deaths. So, the age range between 60-80 there are more potential excess deaths occurring.

Age range deaths are included for each death cause and there is a definite increase in chance of deaths as the age increases.

3:





The above scatterplot answers the question “Is there a correlation between observed deaths and expected heart disease deaths?”

The strong positive correlation between observed and expected deaths suggests that as the number of observed deaths increases, the expected deaths also tend to increase.

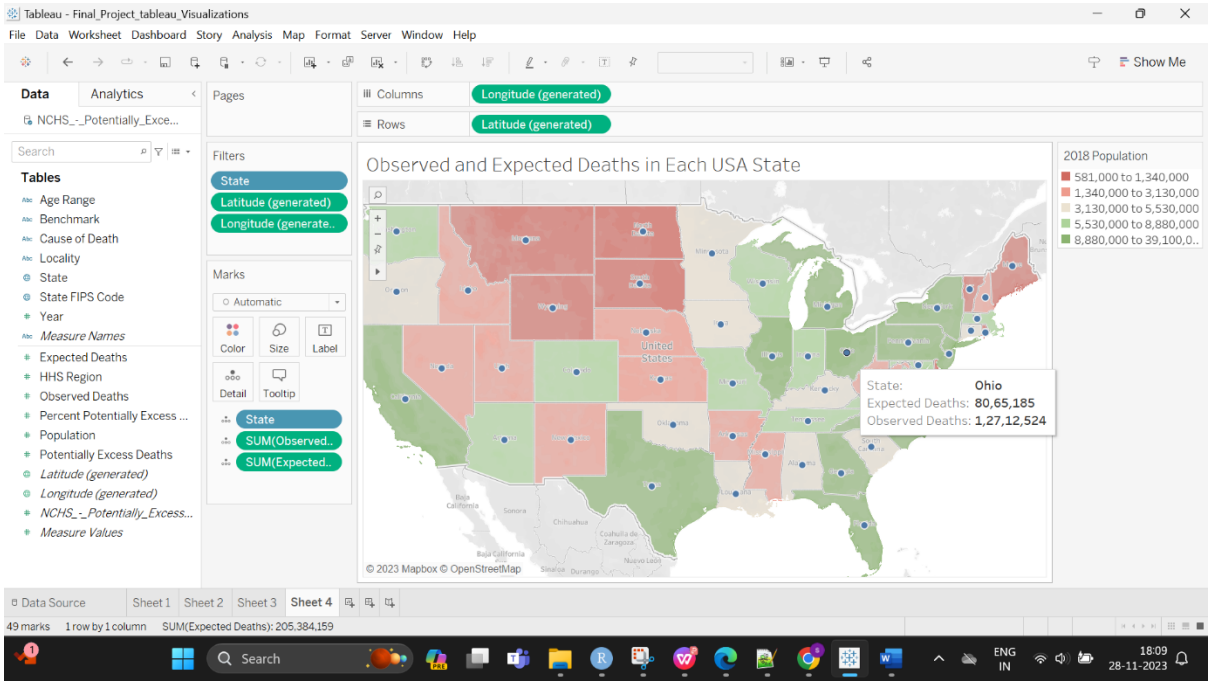
The R-squared value being close to 1 from this we can say that the linear regression model fits the data very well, explaining a high percentage of the variability in expected deaths based on observed deaths.

With a very low p-value, we can see that we have statistical confidence in the significance of the observed correlation.

The scatterplot, along with the provided statistics, shows a robust and statistically significant linear relationship between observed and expected heart disease deaths. The model fits the data well, and the observed correlation is not likely due to random chance.

So, when we place the pointer on the points, it will show the state name and its expected, observed deaths. Along with that if we want to know the p value, r-squared value and equation we have to place the pointer on the trend line, tableau automatically shows these values.

Visualization 4:

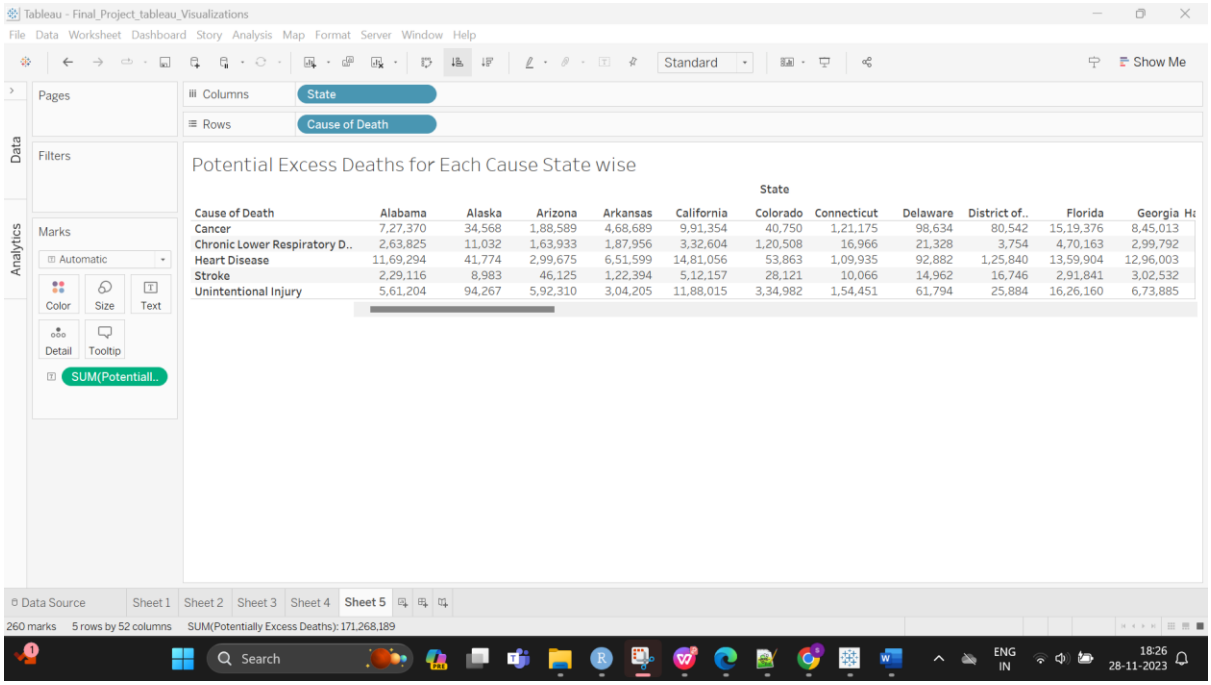


This map plot answers the question “How do expected deaths and observed deaths compare across different States in US?”

As seen in the map plot, we can check the expected and observed deaths of each state. We have to just hover over the pointer on the state which we want to look it shows the required results. There is a colour of each state according to the population. So, the states which are having high population are also having more excess deaths i.e., high population is directly proportional to more excess deaths.

There are points on each state, when we place the mouse pointer on it, it shows the state name, expected, observed deaths so we can have a detailed overview for each state.

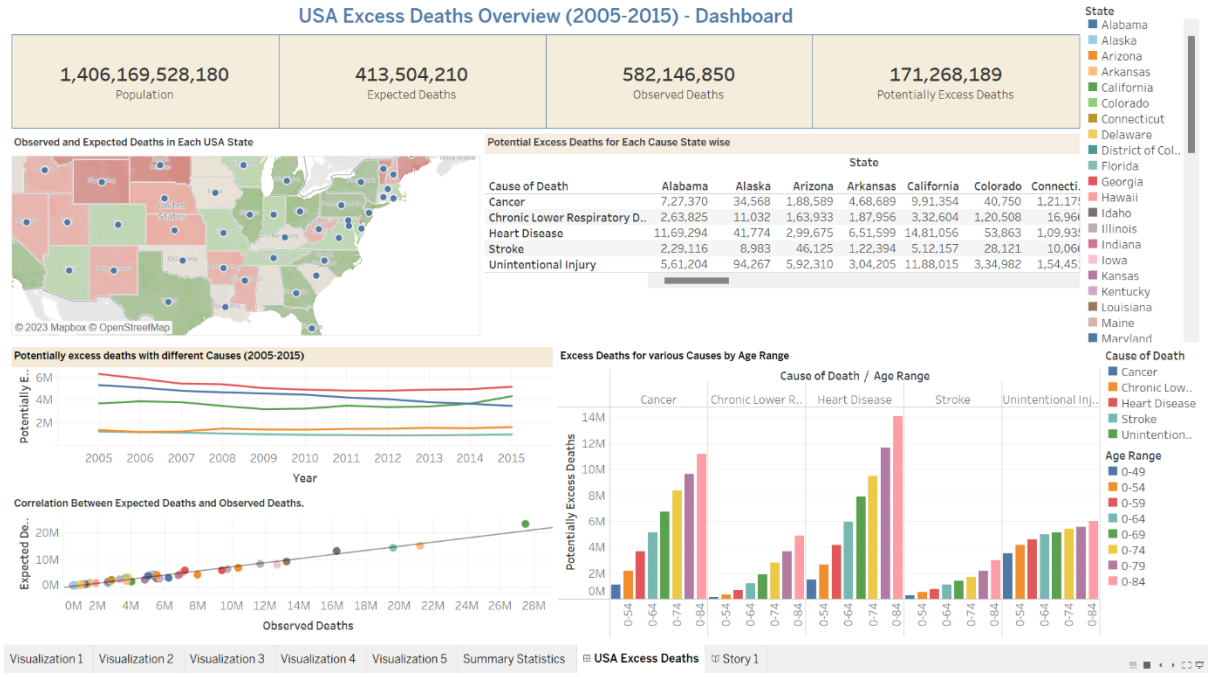
Visualization 5:



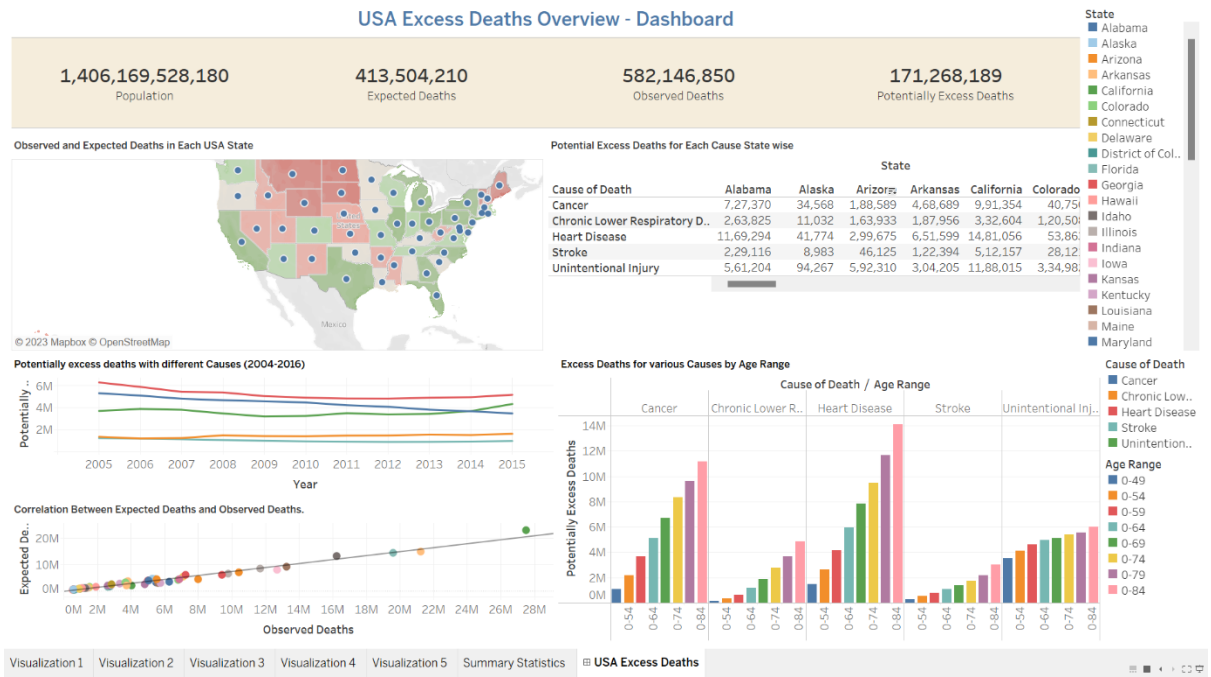
This visualization answers the question “What are the number of excess deaths in each state for each cause?”

Here we can see in the above text chart shows all the excess deaths for each state and also it includes the filter of cause of death. So, as it is interactive where we can move the bar to check the deaths stats for each state. There are five causes of death (Cancer, Chronic lower respiratory disease, Heart disease, Stroke, Unintentional injury) so for each state we have death stats of a particular death cause in each row.

Tableau Interactive Dashboard:



In presenting mode:

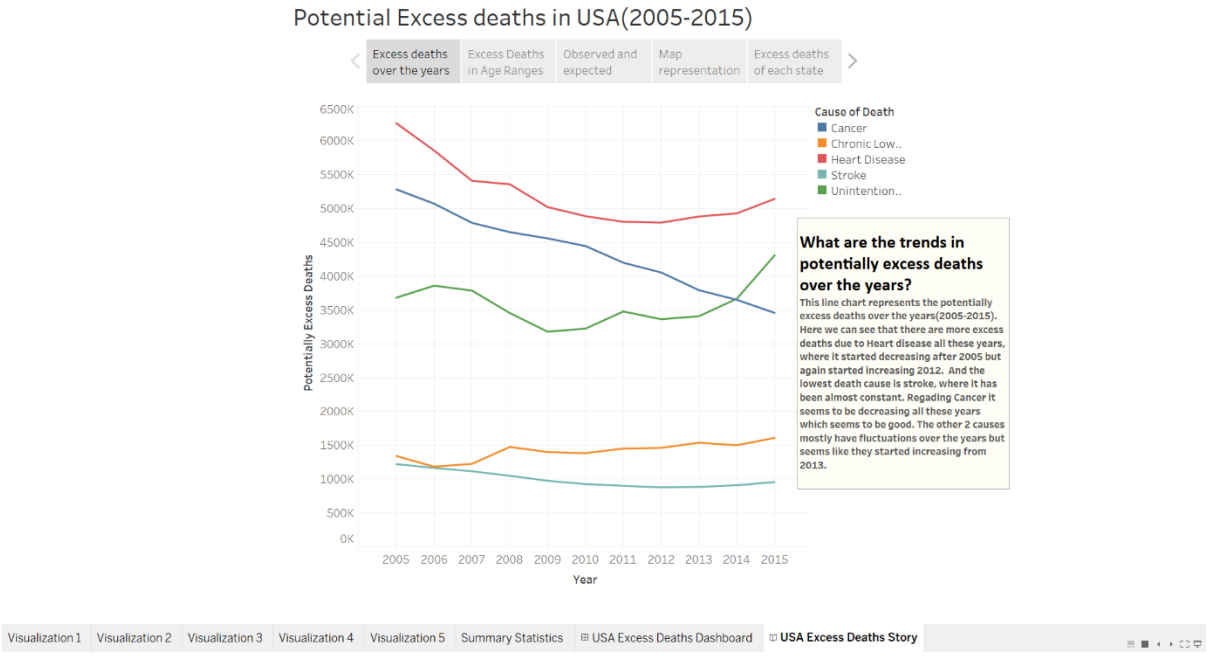


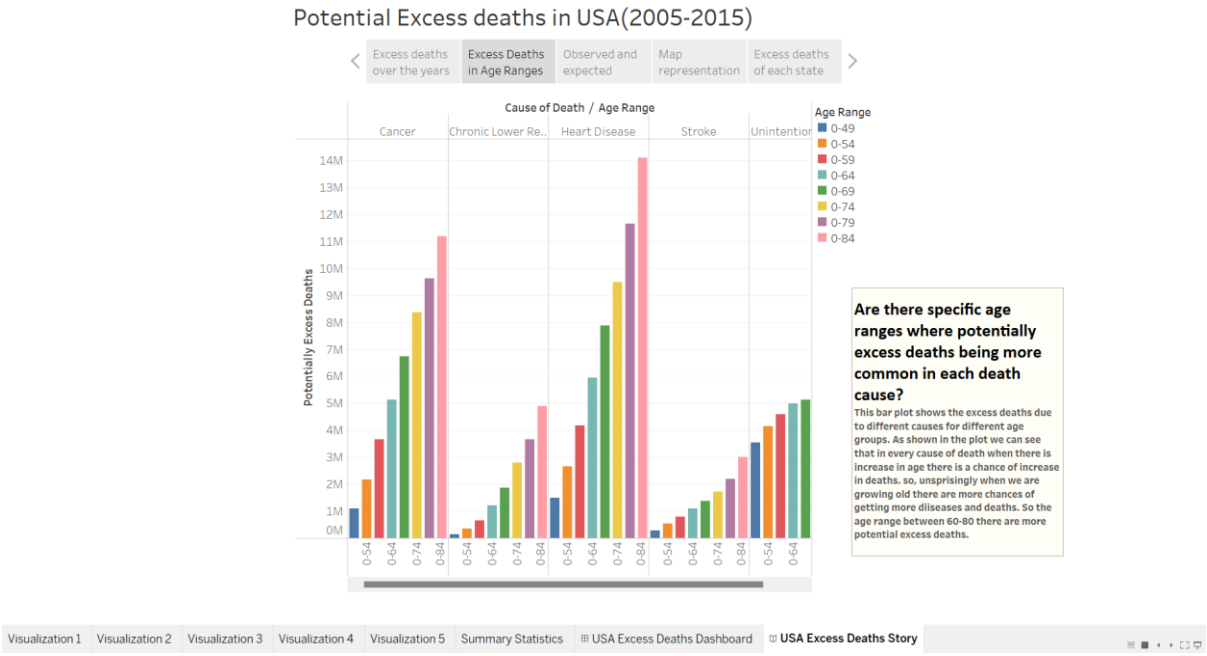
For creating this dashboard, I have opened new dashboard first and then dragged all the required sheets in to it. Later, I have arranged them to represent in visually attractive way.

Then, I got a thought to add summary statistics as well at the top to get the overview of total excess deaths. Then I have created a new sheet in that I started adding statistics that are required. And then did the format and font changes.

And then added this statistics sheet as well to top of the dashboard. Later, I have made the dashboard even more attractive by adding the subtle colours to it.

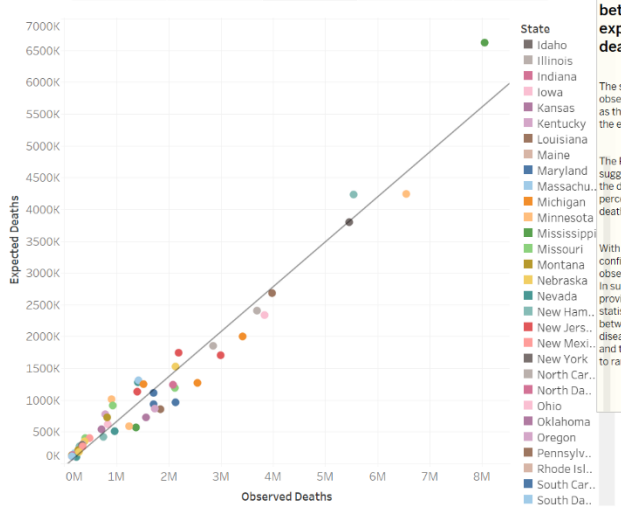
Tableau Story:





Potential Excess deaths in USA(2005-2015)

Excess deaths over the years | Excess Deaths in Age Ranges | Observed and expected | Map representation | Excess deaths of each state



Is there a correlation between observed and expected heart disease deaths?

The strong positive correlation between observed and expected deaths suggests that as the number of observed deaths increases, the expected deaths also tend to increase.

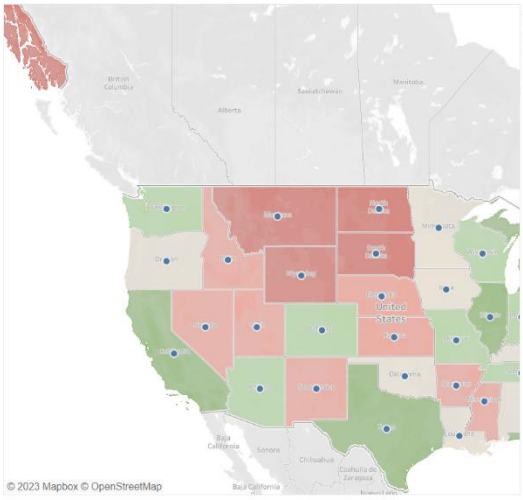
The R-squared value being close to 1 suggests that the linear regression model fits the data very well, explaining a high percentage of the variability in expected deaths based on observed deaths.

With a very low p-value, you have statistical confidence in the significance of the observed correlation.
In summary, the scatterplot, along with the provided statistics, shows a robust and statistically significant linear relationship between observed and expected heart disease deaths. The model fits the data well, and the observed correlation is not likely due to random chance.

Visualization 1 | Visualization 2 | Visualization 3 | Visualization 4 | Visualization 5 | Summary Statistics | USA Excess Deaths Dashboard | **USA Excess Deaths Story**

Potential Excess deaths in USA(2005-2015)

Excess deaths over the years | Excess Deaths in Age Ranges | Observed and expected | Map representation | Excess deaths of each state



How do expected deaths and observed deaths compare across different States in US?

As seen in the map plot we can check the expected and observed deaths of each state. We have to just hover over the pointer on the state which we want to look it shows the required results. There is a color of each state according to the population. So the states which are having high population are also having more excess deaths i.e high population is directly proportional to more excess deaths.

Visualization 1 | Visualization 2 | Visualization 3 | Visualization 4 | Visualization 5 | Summary Statistics | USA Excess Deaths Dashboard | **USA Excess Deaths Story**

To create this tableau story, first I have added new story sheet in tableau, then I have added total of 5 story points according to my visualizations. Where I have added the story points for each visualization. There are total of story points. For each visualization I have added the description in which it has the question that is answered by that particular visualization.

I will be adding the whole tableau workbook which consists of the 5 visualizations, dashboard, story in the canvas while submitting this project.