Name: Sharik Shaik

MID: M15396597

HI-7072 – Final Project Proposal

1) 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 is 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.

2) 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.

3) 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.