1. Introduction

- The objective is to develop a data mining project to investigate the leading causes of death in the United States and the state of California,with a focus on identifying the trends in the number of deaths.
- I will be using features like year, cause of death, state and deaths to give an overview of the various causes of deaths throughout the year and will be presented in a visual map.

2. Literature Review

- The research investigates the actual cause of death in the United States for the year 2000. It tries to identify and have a statistical analysis on the real causes of death. The research was conducted by referencing various articles from 1920 till 2000. Additionally data from Centers for Disease Control and Prevention was used. Conclusions were drawn that smoking remained the leading cause of death but a pattern was observed and it was predicted that due to poor diet and physical inactivity may take over the actual cause of death. [1]
- The research discusses the overall decrease in the total number of deaths in the United States from the year 1970 to 2002. But further research is carried out to unmask mortality due to 6 leading causes. It was found that the age standardized death rate decreased for 4 out of 6 diseases but the absolute number of deaths continued to increase in the United State as the diseases occurred at a higher age. [2]
- The research talks about Stroke declining from third to fourth cause of death in America. A discussion about the history of Stroke in America is analyzed where it has transitioned from second to fourth leading cause over a century. It states the recent decline of Stroke among leading causes of death in America is a testament of how superior progress has been made in terms of cerebrovascular disease and acute care. [3]
- The research investigates analysis of educational differentials in mortality by cause of death. It was observed that there was strong inversion among white between level of education and mortality with consistent declining in mortality with the higher the education level. [4]
- The research talks about Alzheimer's diseases as a cause of death in the United States. The data from National Vital Statistics was observed to draw the conclusion that Alzheimer's disease has increasingly been reported as a cause of death due to a variety of factors such as better diagnosis, awareness and changes in the perception of Alzheimer's disease. [5]
- The research discusses Infant and Youth Mortality trends in the United States. It is observed that the United States has higher infant mortality compared to Canada and Europe/United Kingdom. It is observed that infant mortality rates declined

throughout the years 1999 to 2002 and 2012 to 2015 but other causes have taken over youth mortality rates such as drug poisoning and unintentional injuries. [6]

3. Problem Formulation

- The data mining project tries to investigate leading causes of death in the United States and California.
- Key features such as year, cause of death, state, and deaths are used to understand and identify trends in the number of deaths .
- The project offers an insight on each cause and total number of deaths due to each cause over the years and provides valuable insights into the evolving health scenario at both national and state levels, aiding in informed decision-making.

4. Data Description

- The dataset I used for the project is from Kaggle (Link to the Dataset) which to be precise has 10869 entries.
- This dataset presents the age-adjusted death rates for the 10 leading causes of death in the United States from 1999 to 2017.
- Data Preprocessing: After getting info about the dataset I noticed that it had no null entries and duplicate entries as well. Although the column '113 Cause Name' was not needed as it just had a more descriptive Cause of Death. The data was clean after this point and is ready to be used for the project..

5. Methodology

- Conducted a data analysis from 1999 to 2017, revealing trends in leading causes of death in the United States through line graphs. Comparing 1999 and 2017 causes of death and highlighting shifting mortality patterns for the state of California. Visualized nationwide mortality on a map for 1999 to 2017.
- I selected a clustering algorithm as it can help identify patterns and groupings for this project.
- Employed k-means clustering to make appropriate clusters of states falling into high, medium and low numbers of deaths by respective diseases.

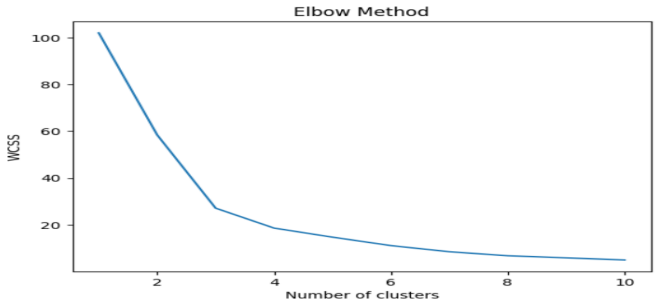
6. Solution Implementation

- Tools:
  - Jupyter Notebook: I used a jupyter notebook as a platform to develop this project.
  - Python: I used python programming language.
  - Pandas: I used pandas libraries to load my .csv file
  - Scikit-Learn: I used this tool to carry out clustering algorithms and verify it with the help of silhouette score.
  - Seaborn/Matplotlib: I used these libraries to implement visually appealing graphs.

    - Plotly Express/PyCountry: I used these libraries to implement the map of the United States.
- Breakdown and Code Snippets:
    - Data Cleaning/Data Preprocessing:
        - df.info() to get an overall insight of the dataset.
        - df.duplicated().sum() to check if there are any duplicates in the dataset.
    - Feature Selection:
        - df1=df.drop('113 Cause Name', axis=1) to drop column '113 Cause Name' from dataset.
    - Elbow Method:
        - for i in range(1, 11): kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0) to use elbow method and determine appropriate number of clusters.

7. Experimental Setup

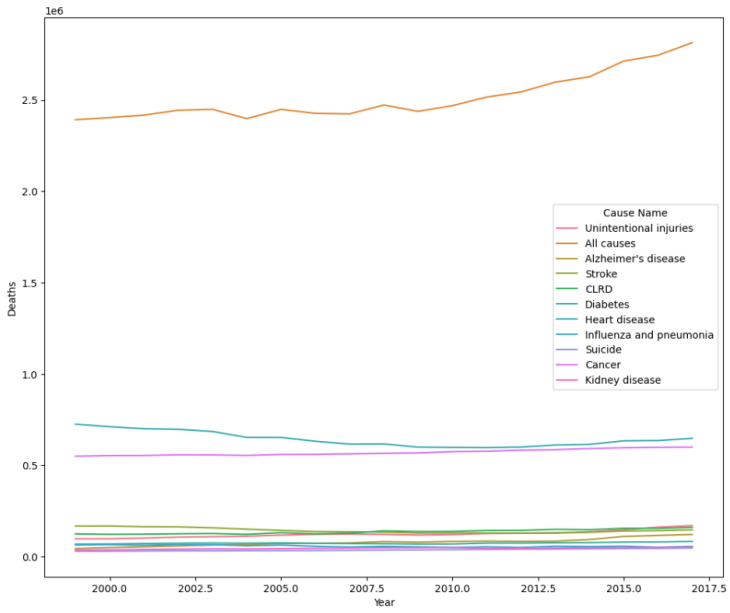
- To identify the appropriate number of clusters, the elbow method was used.



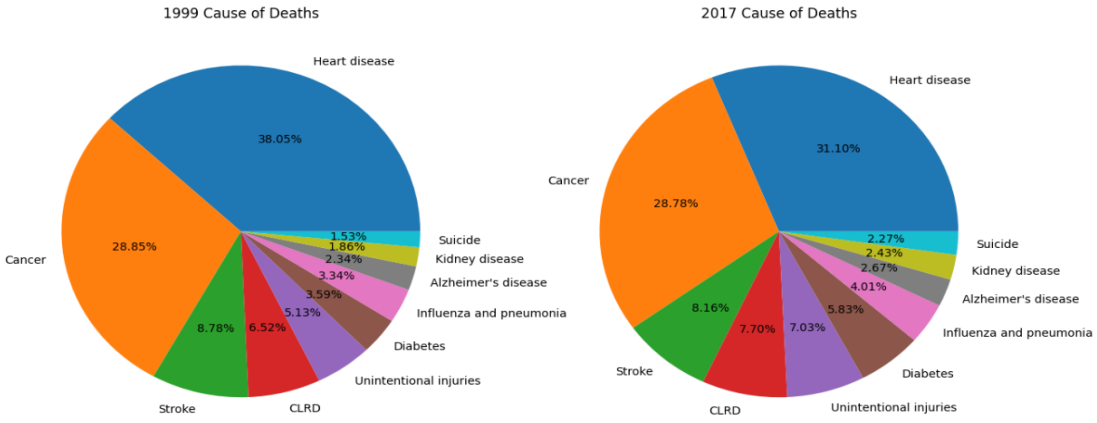
- The optimal results were obtained when the number of clusters were 4.
- To validate the consistency of the clusters formed I employed Silhouette Score as my metric as it is well-suited for providing a comprehensive measure of both internal cohesion and external separation of clusters which is crucial when analyzing death patterns, it works well with clusters of irregular shapes and sizes and does not assume a particular shape of clusters making it flexible to use.
- I tried manipulating the number of clusters and verified that the optimal Silhouette score was obtained when the number of clusters were 4.

8. Results

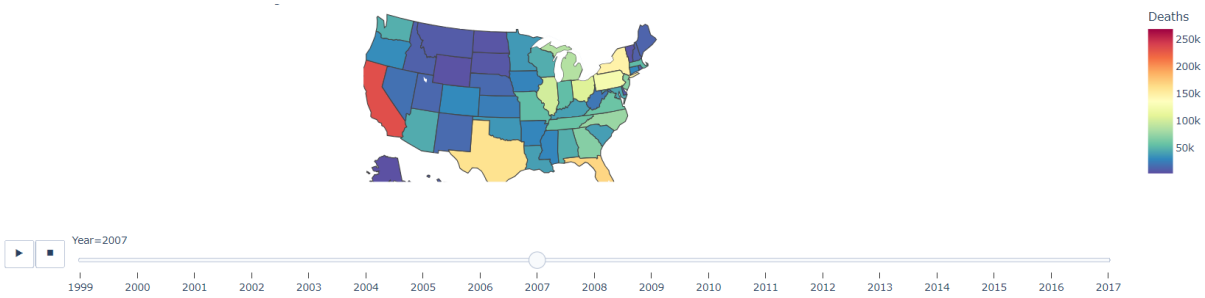
- This graph shows a graphical representation of trends of Causes of Deaths in the United States from year 1999 to 2017.



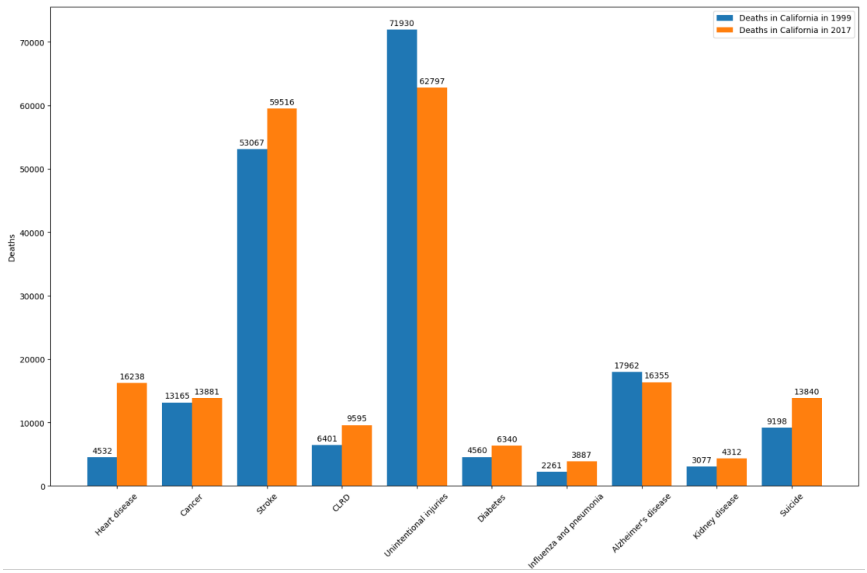
- This graph compares Causes of Deaths in the United States for the year 1999 and 2017.
- Key takeaways:
  - Cancer and Stroke are almost the same throughout the years.
  - Heart Disease has dropped by 7%
  - There is a significant increase in Suicide, Kidney Disease, Influenza and Pneumonia, Unintentional Injuries, CLRD.



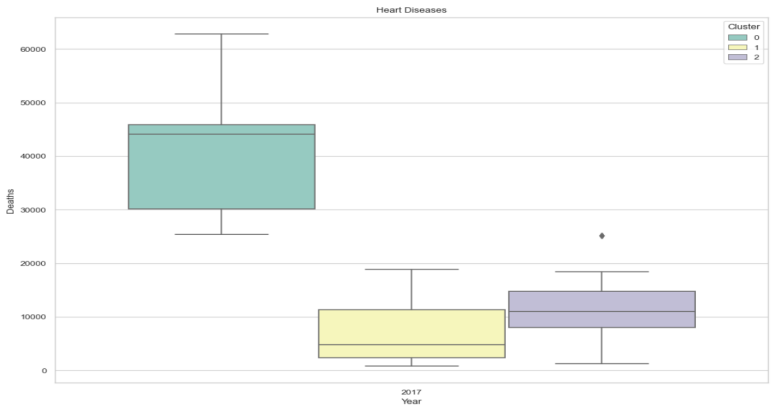
- A visual map of the United States in the form of a heat map plotting the total number of deaths in each state over the years represented by varying color gradients.



- This graph compares Causes of Deaths in the state of California for the year 1999 and 2017.
- Key takeaways:
    - Here, as we can see heart diseases and stroke have decreased overtime but all other diseases have increased over 18 years!



- Cluster based on Heart Disease in the United States for the year 2017. Used kMeans clustering along with silhouette score matrix to understand the result.
- Results:
    - Silhouette Score: 0.543030073897613
    - Cluster 0 States: Illinois, Pennsylvania, Ohio, New York, Florida, California, Texas
    - Cluster 1 States: Hawaii, Idaho, Iowa, Kansas, Maine, Maryland, Massachusetts, Minnesota, Montana, Nebraska, South Dakota, South Carolina, Rhode Island, Oregon, North Carolina, New Mexico, New Jersey, New Hampshire, North Dakota, Delaware, Connecticut, Colorado, Arizona, Alaska, Utah, Vermont, Wyoming, Wisconsin, Washington, Virginia
    - Cluster 2 States: Indiana, Kentucky, Louisiana, Michigan, Mississippi, Missouri, Oklahoma, Georgia, Nevada, Tennessee, Arkansas, District of Columbia, Alabama, West Virginia



9. Discussion

- The project starts with data cleaning and preprocessing steps.
- It was observed that there were no null values in the dataset and neither a duplicate value.
- There is a need to remove a column ‘113 Cause Name’ which gives additional information about the cause of death.
- After removing the column there are only four columns left: ‘Cause Name’, ‘Age Adjusted Death Rate’, ’State’, ‘Deaths’, and ‘Year’.
- Python libraries like pandas, sklearn, matplotlib, plotly express, pycountry and warnings.
- An overall analysis of causes of death is carried out and a line graph is plotted on a time vs total number of deaths.
- It is observed that Heart Disease, Cancer and Unintentional Injuries are the leading causes of death for the latest 2017 year.
- Further a comparison in the form of pie chart is conducted which showcases that Cancer has remained as the leading cause of death for the entirety of our analysis.
- To plot deaths on the United States map an additional column country code is added. This column helped mapping respective states with their position and number of deaths.
- A case study on the state of California is carried out in the form of a bar graph to compare causes of death for the year 1999 and 2017.
- Clustering of the dataset is conducted using kmeans clustering algorithms to determine the of states falling into high, medium and low numbers of deaths by respective diseases.
- To determine the number of clusters, the elbow method is used.

- Comparison with findings from Literature review:
    - The findings aligned in sync with the results from [1], where as predicted death due to inactivity and poor diet like heart diseases and stroke have risen.

- Another key finding from [2] is supported where overall there is an increase in the number of deaths due to listed 6 key diseases.
- As seen in the line graph, stroke has declined to fourth cause of death and thus it bolsters the finding from [3] research. This signifies as years go by there is significant improvement in the medical sector.
- It is also reflected from the line graph that the deaths due to Alzheimer's diseases has increased over time supporting the findings from [5] .

10. Limitations

- Limitation: The project does analysis based on the total number of deaths rather than a more normalized standard
- Reflection:Use of a more normalized standard such as deaths per 1 sq. mile or deaths per 1000 people is required.
- Limitations: The dataset does not take into consideration the age factor and any other such features.
- Reflection: A more unbiased or standardized age adjusted feature is required to normalize the dataset and take into account disease with age into consideration.
- Limitation: The project simply visualizes the dataset.
- Reflection: Need to employ any machine learning algorithms or predictive techniques.

11. Future Work

- There is a need to use and take this project further and investigate more about why certain diseases are leading.
- Data from medical sectors can be extracted and a comparison of health sectors in each state and their respective total number of deaths can be evaluated.
- The data can be predicted and a future analysis can be made which can help the medical sector analyze the trend and help them manage resources and technology effectively.

12. Work Cited

1. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual Causes of Death in the United States, 2000. JAMA. 2004;291(10):1238–1245. doi:10.1001/jama.291.10.1238
2. Jemal A, Ward E, Hao Y, Thun M. Trends in the Leading Causes of Death in the United States, 1970-2002. JAMA. 2005;294(10):1255–1259. doi:10.1001/jama.294.10.1255
3. Towfighi A, Saver JL. Stroke declines from third to fourth leading cause of death in the United States: historical perspective and challenges ahead. Stroke. 2011 Aug;42(8):2351-5. doi: 10.1161/STROKEAHA.111.621904.

4. Evelyn M. Kitagawa, Philip M. Hauser; Education differentials in mortality by cause of death: United States, 1960. Demography 1 March 1968; 5 (1): 318–353. doi: https://doi.org/10.1007/BF03208579
5. Hoyert, D L, and H M Rosenberg. “Alzheimer's disease as a cause of death in the United States.” Public health reports (Washington, D.C. : 1974) vol. 112,6 (1997): 497-505.
6. Khan SQ, Berrington de Gonzalez A, Best AF, et al. Infant and Youth Mortality Trends by Race/Ethnicity and Cause of Death in the United States. JAMA Pediatr. 2018;172(12):e183317. doi:10.1001/jamapediatrics.2018.3317