An In-depth Analysis of Serious Diseases in the United States: Identifying and Understanding Contributing Factors

Yuqi Cheng (919630150)

Tianyu Wang (917440012)

Zhiyue Guan (919286782)

Research Questions:

What are the evolving temporal trends in mortality rates for the most severe diseases in the US, and are there discernible patterns, anomalies, or demographic factors linked to heightened mortality rates that warrant focused attention?

Introduction:

In our pursuit of advancing public health understanding and mitigating risks, our research focuses on identifying diseases that pose a substantial threat to individuals in the US on a state level. By evaluating population mortality trends across multiple years for all states, we aim to discern diseases that consistently exhibit a serious impact on human health. Simultaneously, we plan to conduct a comprehensive analysis of potential risk factors intricately linked with the prevalence and severity of these diseases. Given that diseases can significantly influence human life, we would access health data efficiently by utilizing web APIs and web scraping from authoritative databases. This approach allows us to gather extensive information from health websites such as WHO or CDC, including data on death rates and influencing factors. The advantage of using web APIs lies in their structured and organized data retrieval, ensuring accuracy and reliability when dealing with a large volume of data. Meanwhile, web scraping

provides the flexibility to extract valuable data from various sources that may not offer direct API access. To steer our exploration, we have constructed hypotheses as follows:

- Null Hypothesis (H0): There is no correlation between demographic or behavioral factors and the prevalence of high-mortality diseases.
- Alternative Hypothesis (HA): Certain demographic or behavioral factors are associated with an increased risk of specific diseases leading to higher mortality.

In order to test these hypotheses and address our research questions, we will leverage the collected data to create informative plots. The visual representation of the data will not only enhance our understanding of trends and patterns but also serve as a validation tool for our hypotheses. By combining efficient data retrieval methods with visually compelling plots, our research endeavors to provide valuable insights that can inform targeted interventions, ultimately contributing to the reduction of disease-related mortality and the enhancement of public well-being.

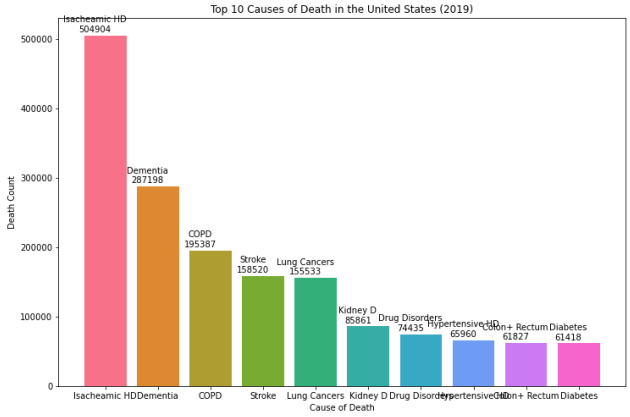
Reports based on code and visualizations:

Part I: Identifying the Most Serious Diseases in the US from 2017 to 2019

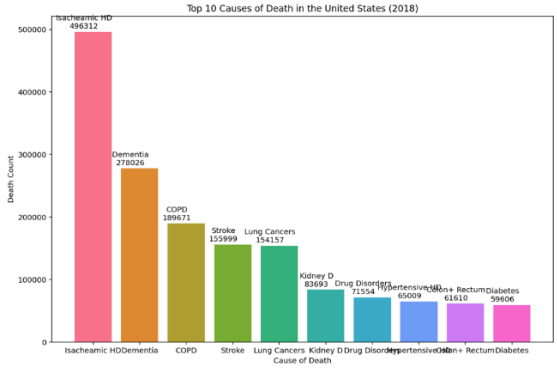
To present information on serious diseases, our focus lies in accessing data related to death populations and death rates based on different diseases. We are working with health-related observations obtained from the World Health Organization (WHO) through an API. The data is retrieved in a JSON format from the web API. We import the JSON function to convert the data into a Python dictionary for further processing, such as filtering and plotting. Each parameter from this API represents Country, Year, Age Group, Sex, Cause code, Population Numeric, Cause Title, DALYRate100K, DALY Count, DeathsRate100K, and Deaths Count. These parameters are combined into a data frame. During the process of scraping data, I discovered that the WHO only

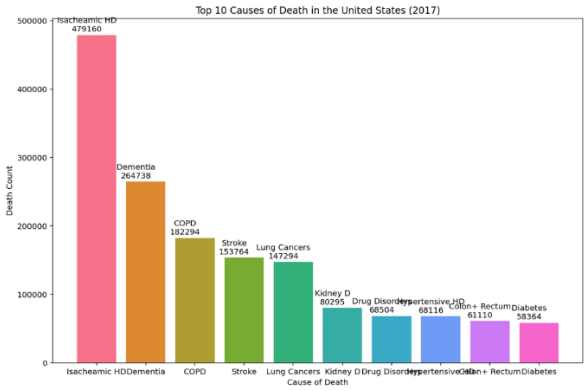
provides an API based on a single year. This means you cannot access data spanning multiple years or within a range of, for example, three years. Therefore, the only option is to collect the data for one year at a time, convert it into a data frame, and then filter the data for comparison later.

Initially, we filter the data for the year 2019 and obtain the top 10 diseases based on death counts. The results and plots reveal a clear trend, indicating that ischemic heart disease is the most prevalent, with a significant difference from the second-ranked disease.



Recognizing that a single year's data may lack precision, we use additional APIs from 2018 and 2017. The goal is to apply the same methods to identify and rank the top 10 diseases. To ensure consistency, we control for other parameters, focusing on the same age group and both sexes.





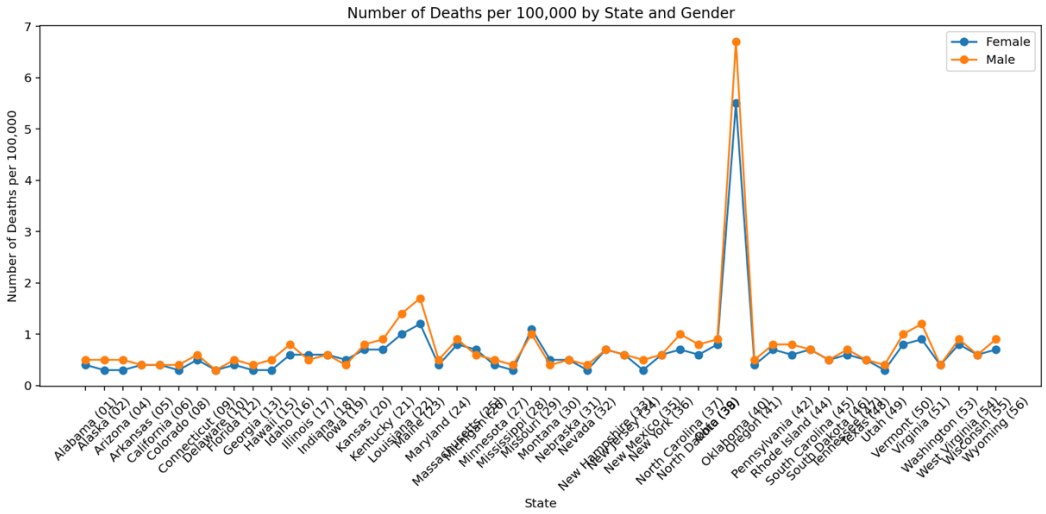
Through the resulting plots, both sets of data re-affirm the seriousness of ischemic heart disease, causing over 400,000 deaths. The linear plot illustrates an increase in death populations across the years. This prompts us to delve deeper into a specific analysis of the geographical influence on disease prevalence.

After we found the prevalence of Ischemic Heart Disease, we are intersted in if there are specific patterns, like demographic or gender-based, of the ischemic HD on a state level. Therefore, we consulted the data of death statistics in the past 20 years of all states in the US from CDC Wonder database. Web scraping is applied here to acquire the data from the CDC Wonder database. However, the Wonder database provide data for research purpose only, and all users need to make an pledge before they are able to access the data. As a result, web scraping could not directly fetch the data, even through the url of the resulting page. Therefore, we downloaded the source code of the webpage after I manually made the agreement for our program to read the content of the page. However, the webpage used a special encoding format, and the routine encoding format like uft-8 and ISO 8859-1 format do not work here. As a result, we used pandas package in python to read the entire source code line by line as a dataframe, and then we analyzed the entire contents from the source code to access the data table in it. The data

includes the death count of the ischemic HD of each age-groups (five-year) and both genders from all states in the US form 1999 to 2020.

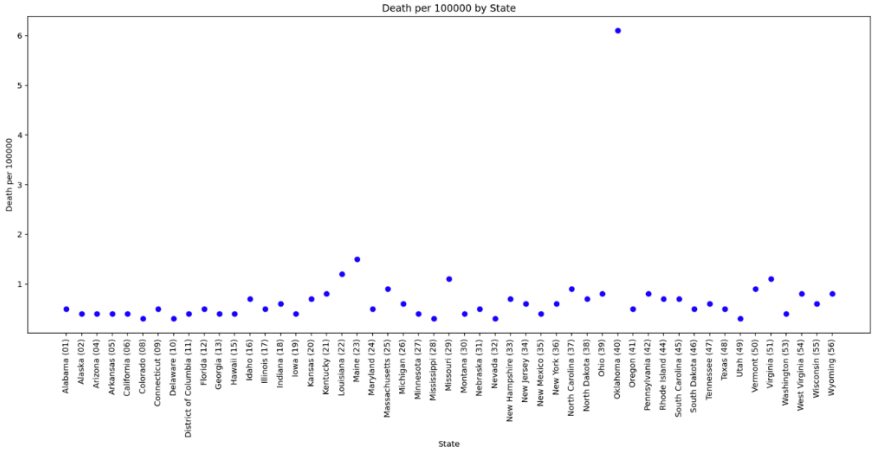
Part II: Ischemic HD statistics analysis of US States based on age-group and gender data from CDC Wonder DB

After acquiring the data, we are interested in looking for which gender is more like to die of the ischemic heart disease. We sorted through the original dataframe for the death rate (number of death per 100000) for each gender (i.e., we count the number of death for all age groups) for all states. The data of district of Columbia is marked as unreliable, so we discarded it from the sorted dataset. Then, we draw a line chart to visually illustrate the results. There are five states having female death rate higher than that of male. Five out of 50 could hardly be explained by chance effect by the pair comparison test, so we can conclude that in the US, males are more likely to suffer the death from ischemic heart disease.



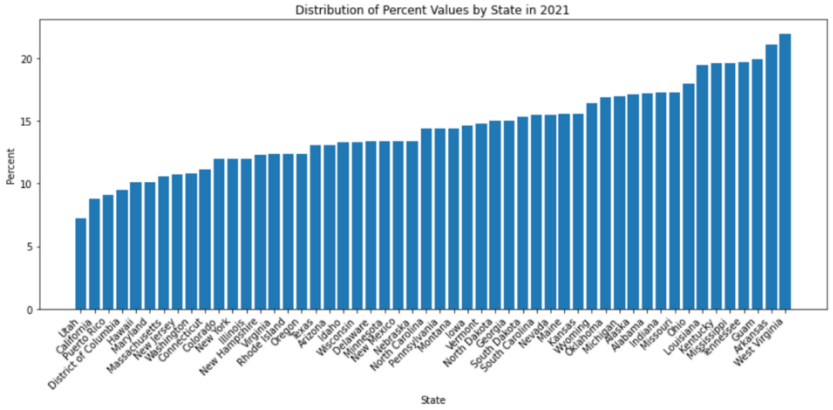
On the other hand, we are also interested to figure out which state has highest patients of ischemic heart disease. So, we then sort through the original dataset for death number per

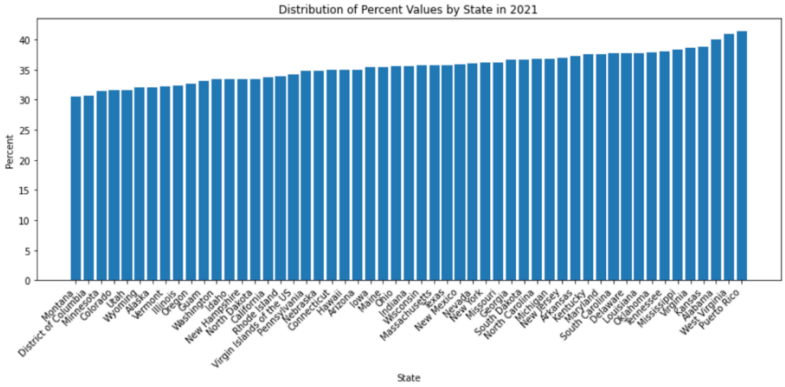
100000 (combine all age groups and both gender) of each state. Then, a scatter plot is drawn to illustrate the result. We found that Oklahoma has an extinct high number of death than any other states. I then sort the entire data of Oklahoma state from the original dataframe. I found that Oklahoma provide the number of death of patients of very high age (over 85), which are not usually provided by other states. And we speculated that since death of ischemic HD is more likely to happen in high age-groups, the average death rate of the state is influenced by the extreme value of high age groups. However, more research, including the census data, habits like smoking, alcohol usage, and epidemiological research of Oklahoma are necessary to make a more sound conclusion.

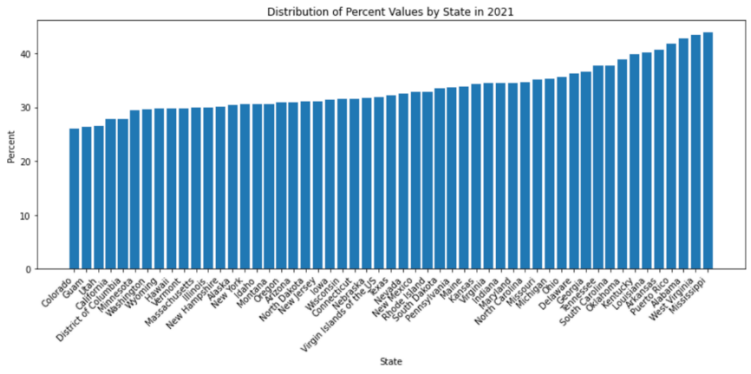


After we get a brief knowledge of statistics of ischemic HD in the US, we are interested to learn the influences of three main factors of heart diseases, including smoking, high blood pressure, and high cholesterol, we extract data from crude prevalence of smoking among adults in 2021 for each state. After drawing histograms based on states, we can decide which state has the highest risk of heart diseases caused by smoking, high blood pressure, and high cholesterol. Moreover, to learn which factor makes more influence to the heart diseases, we draw a line plot to compare based on states.

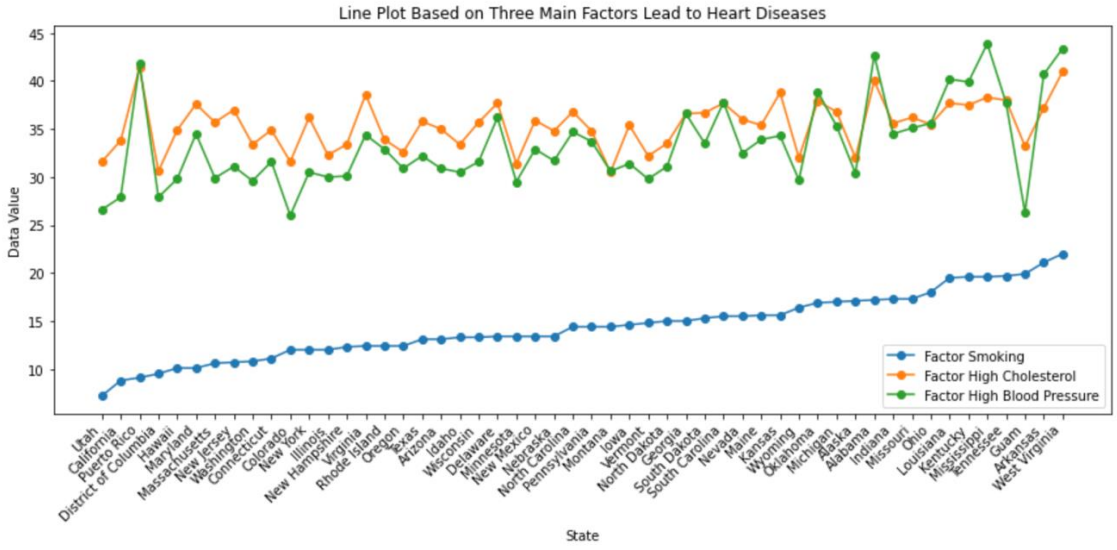
In 2021, West Virginia has highest risk of heart diseases caused by smoking while Puerto Rico has highest risk of heart diseases caused by high cholesterol. Moreover, Mississippi has the highest risk of heart diseases caused by high blood pressure.







After comparing all three main factors of heart diseases, high blood pressure and high cholesterol have higher influence to heart diseases. Puerto Rico has the highest risk of heart diseases caused by high cholesterol and the third highest risk of heart diseases caused by high blood pressure while its smoking risk factors is the third lowest one. Also, West Virgininia has highest risk of heart diseases caused by smoking while its high blood pressure and high cholesterol lead to heart diseases are both in the second highest places. Furthermore, Mississippi has the highest risk of heart diseases caused by high blood pressure with relative lower percentage of smoking and high cholesterol lead to heart diseases comparing to Puerto Rico and West Virgininia.



When we attempt to extract data from website, we face several challenges. The API provider update website may influence us to access the dataset, forcing us to take more time to find new available API links. Moreover, we also meet API limitations: we have to make agreement to not use API to extract data. To solve the issue, we start to combine downloading and web scrabing together to extract useful data for our project.

Conclusion

Our comprehensive research on the evolving trends in mortality rates due to serious diseases in the United States has yielded significant insights. By adeptly utilizing web APIs and web scraping, we were able to analyze a vast array of data from the WHO and CDC, which helped us identify ischemic heart disease as the most prevalent cause of death across different demographics. The gender-specific analysis revealed a higher susceptibility among males to ischemic heart disease, with interesting geographical disparities, most notably in Oklahoma.

Our investigation extended beyond mere identification of prevalent diseases to understanding the underlying risk factors. We found that behavioral and demographic factors, such as smoking, high blood pressure, and high cholesterol, significantly contribute to the risk of heart diseases. The state-specific analysis highlighted West Virginia, Puerto Rico, and Mississippi as having the highest risks due to smoking, high cholesterol, and high blood pressure, respectively. This study not only validates our alternative hypothesis, affirming the correlation between demographic/behavioral factors and the prevalence of high-mortality diseases, but it also provides a crucial foundation for public health interventions.

Declaration of Work:

We discussed the research outline together, and each person then focused on assigned research sub-questions, including data acquisition from online databases, data sorting and analysis, and data explanation. We then write the draft for our part separately and then compiled together.

Zhiyue Guan: Part I & Introduction

Tianyu Wang: Part II & Conclusion

Yuqi Cheng: Part III

All code for this project could be found on the github page: https://github.com/xhs-wang/STA141B-Final-Project