CptS 475

Team 40

12/8/2023

# Geospatial analysis of cancer Final Report

**GitHub link:**

JonathanZhen/CptS475-Geospatial-analysis-of-cancer-Project (github.com)


## Abstract

Our project analyzes the issue of cancer prevalence in the United States, focusing on demographic, lifestyle, environmental, and socioeconomic factors. The program aims to develop a web-based application or dashboard for investigating the correlation between cancer incidence and various influencing factors, specifically lung and liver cancer. This project involves geospatial analysis that analyzes the risk factors for human disease by gender, age, race, lifestyle choices (smoking, obesity, alcohol consumption), environmental conditions (air quality, water pollutants), health/socioeconomic factors (health insurance, poverty, specific health conditions). We primarily use Python and R to implement exploratory data analysis and visualize results through interactive maps, plots, and charts.


## Introduction

This project aims to investigate the factors that influence lung cancer incidence and understand their relationships. Its importance lies in identifying potential risk factors to guide public health interventions, improve medical policy, and contribute to the prevention and early detection of lung cancer. Basic methods include data loading and cleaning, loading lung cancer and related data sets, removing duplicate entries to ensure data accuracy, and then selecting different plots according to the situation. The conclusion was that white men were more likely to develop lung cancer, and that air quality itself did not appear to directly affect lung cancer rates. There was a statistically significant correlation between arsenic levels and lung cancer incidence. Alcohol consumption is positively associated with lung cancer incidence, but causality has not been proven. Heart disease, asthma, and diabetes were moderately positively associated with lung cancer, suggesting a potential relationship. Areas with higher proportions of the uninsured population tend to have higher lung cancer rates. Smoking prevalence is positively correlated with lung cancer incidence, reinforcing the established link between smoking and lung cancer. In summary, this project combines statistical analysis, visualization, and correlation assessment to gain insights into the complex relationship between numerous factors and lung cancer incidence. For liver cancer, we did the same to the data: taking the dataset related to liver cancer and the factors that might cause it, cleaning the data and converting some non-numeric value to numeric value. Some of the findings were predicted like people who live in poverty or have obesity are

more likely to get liver cancer, but some are not predicted such as drinking alcohol does not seem to have a direct impact on getting liver cancer.

## Problem Definition

- **Lung cancer**
    - Does gender affect lung cancer prevalence?

Understanding if there is a gender-based difference in lung cancer rates can provide insights into potential biological or lifestyle factors contributing to the disease.

    - Does race affect lung cancer prevalence?

Examining the impact of race on lung cancer prevalence helps identify potential disparities and may guide targeted interventions and healthcare strategies for specific racial groups.

    - Could worse air quality lead to higher lung cancer rates?

Investigating the link between air quality and lung cancer rates is crucial for public health, as it explores the environmental aspect and informs policies aimed at reducing air pollution.

    - Does the amount of arsenic affect the prevalence of lung cancer?

Exploring the link between arsenic exposure and lung cancer prevalence contributes to understanding the role of specific environmental contaminants in the disease's development.

    - Does alcohol affect lung cancer rates?

Analyzing the relationship between alcohol consumption and lung cancer rates helps clarify the impact of lifestyle choices on the disease, potentially influencing health recommendations.

    - Are people with heart disease more likely to be diagnosed with lung cancer?

Examining the likelihood of lung cancer diagnosis in individuals with heart disease provides insights into potential co-morbidities and shared risk factors.

    - Are people with asthma more likely to be diagnosed with lung cancer?

Investigating the connection between asthma and lung cancer helps assess whether respiratory conditions may contribute to an increased risk of developing lung cancer.

    - Does not having health insurance increase lung cancer rates?

Understanding the impact of health insurance coverage on lung cancer rates is vital for identifying healthcare disparities and areas that may require intervention.

    - Are people with diabetes more likely to be diagnosed with lung cancer?

Exploring the relationship between diabetes and lung cancer helps elucidate potential connections between metabolic conditions and cancer development.

    - Where in the United States has the highest lung cancer rate?

Identifying regions with the highest lung cancer rates in the United States is crucial for resource allocation, intervention planning, and understanding potential regional risk factors.

- Does smoking increase the risk of lung cancer?

Examining the well-established link between smoking and lung cancer is essential for reinforcing the understanding of a major preventable cause of the disease.

- **Liver cancer**
  - Are people with diabetes more likely to get Liver cancer?

Finding the pattern of numbers of liver cancer per 100k population and the number of people with diabetes per 100k population helps understanding the relationship between liver cancer and diabetes.

- Does gender affect liver cancer prevalence?

Analyzing liver cancer rate based on gender can help us understand how lifestyle might affect the chance of getting liver cancer and the genetic factors.

- Does obesity affect liver cancer prevalence?

Obesity is known to add more pressure on human's internal organs, so the relationship between obesity and liver cancer is a valid topic to do research on.

- Will drinking alcohol affect the chance of getting liver cancer?

Alcohol can cause damage to the liver, so it is necessary to learn the relationship between the amount of alcohol consumed and the chance of getting liver cancer.

## Models/Algorithms/Measures

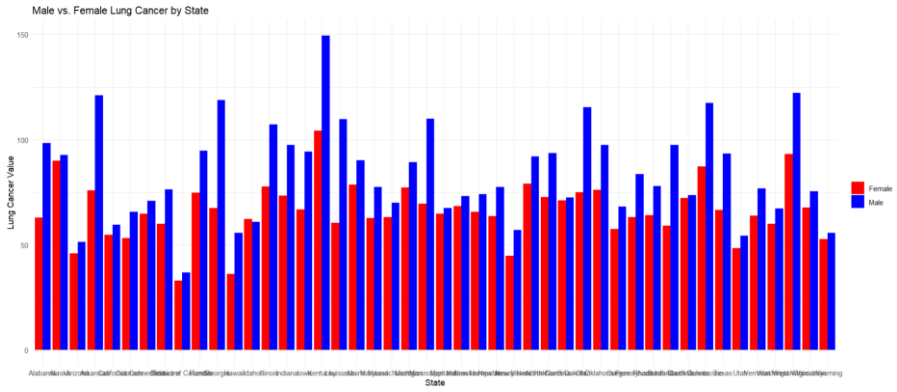
- **Lung Cancer**
  - Air Quality vs. Population:
    - Model/Measure: Scatter plot
    - Description: Examining the relationship between air quality and lung cancer incidence using a scatter plot. The conclusion is drawn based on visual inspection.
  - Arsenic Levels vs. Lung Cancer Rates:
    - Model/Measure: Boxplot, Chi-squared test
    - Description: Investigating the relationship between arsenic levels and lung cancer rates using a boxplot. A chi-squared test for independence is performed to assess the statistical significance of the association.
  - Alcohol Consumption vs. Lung Cancer Rates:
    - Model/Measure: Scatter plot, Correlation coefficient
    - Description: Analyzing the relationship between alcohol consumption and lung cancer rates using a scatter plot. The correlation coefficient is calculated to quantify the strength and direction of the linear relationship.

    - Heart Disease, Asthma, Diabetes vs. Lung Cancer Prevalence:
        - Model/Measure: Linear regression, Correlation coefficient
        - Description: Examining the relationships between heart disease, asthma, and diabetes with lung cancer prevalence using linear regression models. Correlation coefficients are calculated to quantify the strength and direction of these relationships.
    - Health Insurance Coverage vs. Lung Cancer Rates:
        - Model/Measure: Bar plot, Correlation coefficient
        - Description: Investigating the relationship between health insurance coverage and lung cancer incidence rates using a bar plot. The correlation coefficient is calculated to assess the strength and direction of the relationship.
    - Lung Cancer Patients by State:
        - Model/Measure: Choropleth map
        - Description: Visualizing the number of lung cancer patients in each state using a choropleth map. The maximum and minimum values, along with the corresponding states, are identified.
    - Smoking Rate vs. Lung Cancer Incidence:
        - Model/Measure: Scatter plot, Linear regression, Correlation coefficient
        - Description: Analyzing the relationship between smoking rate and lung cancer incidence using a scatter plot. Linear regression and correlation coefficient calculations are performed to quantify the strength and direction of the linear relationship
- **Liver Cancer**
    - Gender vs Liver cancer rate
        - Model/Measure: Bar plot
        - Description: finding the relationship between the likelihood of getting liver cancer for male and female population
    - Race vs Liver cancer rate
        - Model/Measure: Bar plot
        - Description: finding the relationship between the likelihood of getting liver cancer for population of different races.
    - Alcohol
        - Model/Measure: scatterplot
        - Description: finding the relationship between the likelihood of getting liver cancer by drinking alcohol
    - Poverty
        - Model/Measure: scatterplot
        - Description: finding the relationship between the likelihood of getting liver cancer for population that lives in poverty.
    - Diabetes
        - Model/Measure: scatterplot
        - Description: finding the relationship between the likelihood of getting liver cancer because of diabetes.

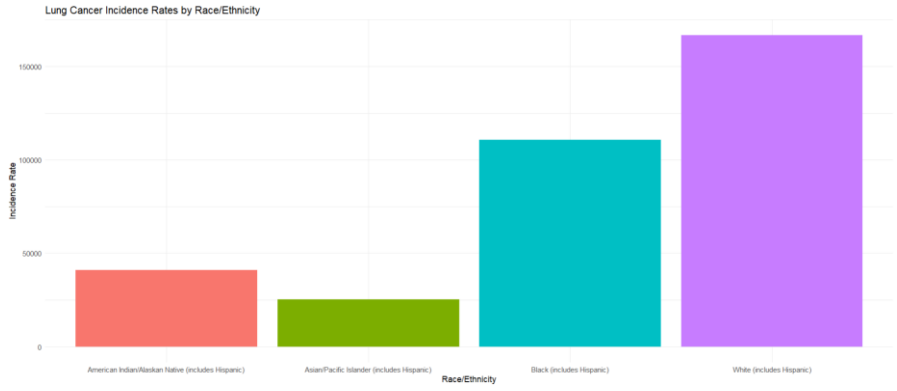
- Obesity
    - Model/Measure: scatterplot
    - Description: finding the relationship between the likelihood of getting liver cancer for obese people.

## Implementation/Analysis

- **Lung Cancer**

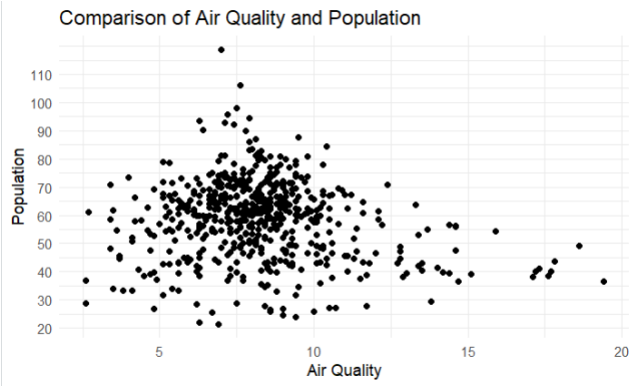



The dataset is named "lungcancer_inc_per100k_pop_2015_2019_gender.csv." It contains information on lung cancer incidence rates per 100,000 population for different states, counties, and genders. The hypothesis I made is about gender, which is there is no significant difference in lung cancer rates between males and females. I use R code to filter out missing values for the "Gender" column. It selects relevant columns (State, County, Gender, Value) and converts the "Value" column to numeric. So, the numeric_gender_data dataframe is created for further analysis. A bar plot is generated using ggplot2, comparing lung cancer rates between males and females for each state. Based on the observed plot, I conclude that males have higher lung cancer rates than females in every state, which disproved my hypothesis.

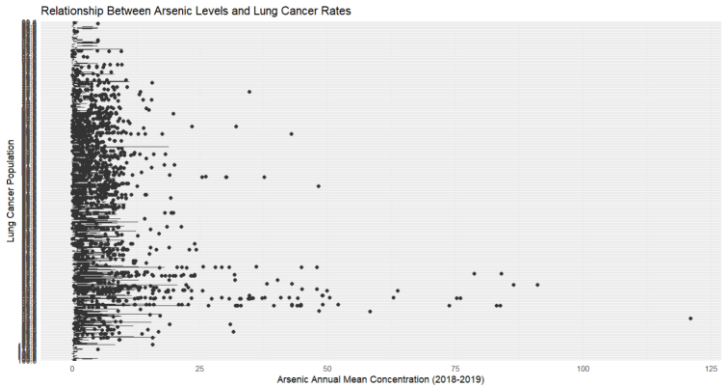



The dataset is named "lungcancer_inc_per100k_pop_2015_2019_race.csv." It contains information on lung cancer incidence rates per 100,000 population from 2015 to 2019, categorized by race/ethnicity. My hypothesis is that race does not affect the prevalence of lung cancer. I loaded the data into R, cleaned it by removing missing values and suppressed data, and

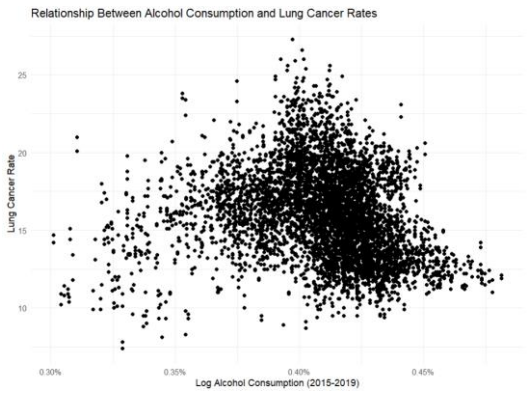
selected relevant columns (State, County, Race/Ethnicity, and Value), converting the "Value" column to numeric. Then I created a bar plot using ggplot2 to visualize lung cancer incidence rates by race/ethnicity. The plot shows a comparison of incidence rates among different racial/ethnic groups. I conclusion based on the plot is that white people have the highest lung cancer prevalence rates, while Asians have the lowest, leading to the conclusion that white people are more likely to get lung cancer than other races.



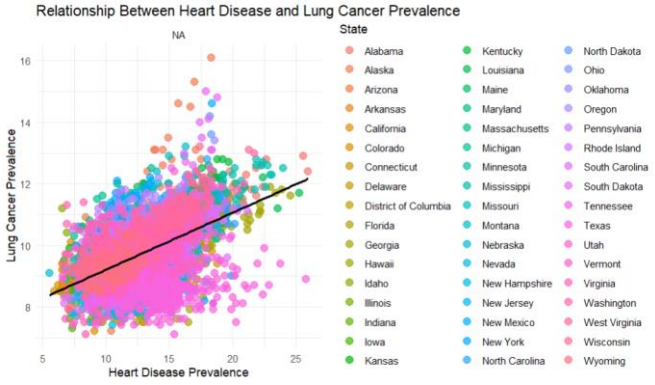

I am using two datasets: “lungcancer_inc_per100k_pop_2015_2019.csv” and “air_quality_pm2.5annualavg_bycounty_2018_2019.csv” for this one. They contain information about lung cancer incidence per 100k population from 2015 to 2019, and annual average PM2.5 levels by county for the years 2018 and 2019. My hypothesis is that the worse the air quality, the higher the risk of lung cancer. First, duplicates are removed from both datasets based on common columns like StateFIPS, CountyFIPS, and County. Then, the datasets are merged using the merge function based on the common columns "StateFIPS," "CountyFIPS," and "County." The merged dataset (merged_data0) is used for plotting using ggplot2. The x-axis represents air quality, and the y-axis represents lung cancer incidence per 100k population. A scatter plot is created to visualize the relationship between air quality and lung cancer incidence. I conclude that, based on the observation of the plot, there does not seem to be a direct correlation between air quality and the prevalence of lung cancer. U.S. medical government organizations indicate that people exposed to PM2.5 have only an 8% higher risk of developing lung cancer compared to people not exposed to PM2.5. This is consistent with the data; air quality is not a major factor in lung cancer.

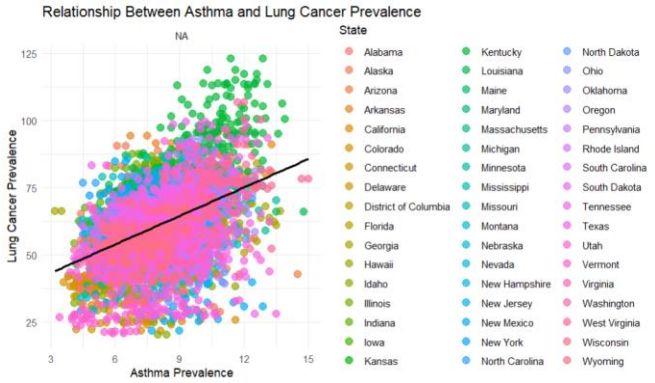


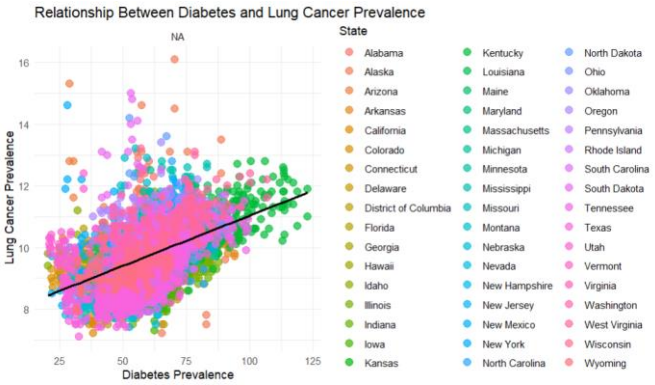
I am using two datasets: “lungcancer_inc_per100k_pop_2015_2019.csv” and “arsenic_annual_mean_conc_2018_2019.csv.” They contain information about lung cancer incidence per 100k population from 2015 to 2019 and arsenic. Merging is done based on common columns, specifically "StateFIPS," "CountyFIPS," and "County." The primary hypothesis being tested is whether there is a significant association between arsenic levels and lung cancer rates. The null hypothesis assumes independence between arsenic levels and lung cancer rates. I am using a chi-squared test for independence to assess the relationship between arsenic levels and lung cancer rates. The chi-squared test involves creating a contingency table using the table function on the merged data. The p-value from the chi-squared test is the primary criterion for evaluation. A p-value less than the significance level suggests rejecting the null hypothesis. The chi-squared test results show a highly significant p-value (< 2.2e-16), indicating a statistically significant association between arsenic levels and lung cancer rates. The chi-squared statistic (X-squared = 609398) is substantial, indicating a strong deviation from independence. The conclusion rejects the null hypothesis, suggesting evidence of a relationship between arsenic levels and lung cancer incidence in the dataset.




I am analyzing two datasets: one on lung cancer incidence “lungcancer_inc_per100k_pop_2015_2019.csv” and another on binge drinking alcohol among adults “binge_drinking_alcohol_adults_per100k_pop_2018_2019.csv." The specific columns being used for evaluation include "StateFIPS," "CountyFIPS," "County," "Value.y", and "Value.x". The hypothesis I propose is a positive correlation, suggesting that higher alcohol consumption is associated with higher rates of lung cancer. The R code sets up a scatter plot using ggplot2 to visualize the relationship. It includes log transformation for alcohol consumption and presents the data in a minimal theme. “The evidence for [alcohol causing] lung cancer is inconsistent and is considered limited,” says Marji McCullough, a registered dietitian and senior scientific director of epidemiology research for the American Cancer Society. The conclusion acknowledges the importance of visual representations of relationships and correlation coefficients. Between 0.4% and 0.45%, the scatter distribution will be denser. This means that within this range, there are more cancer patients. But this does not necessarily mean alcohol is the leading cause of lung cancer, and more information is needed.

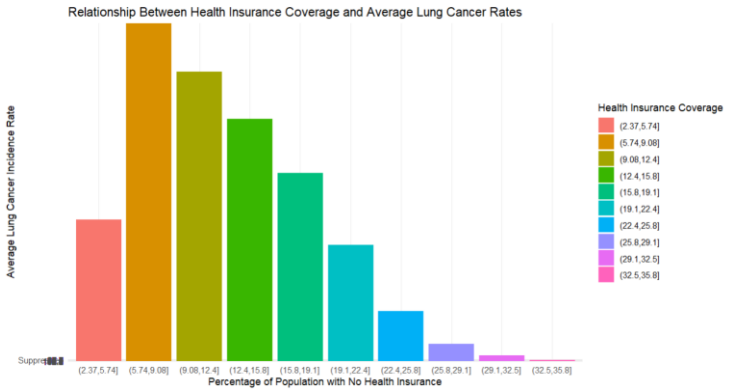


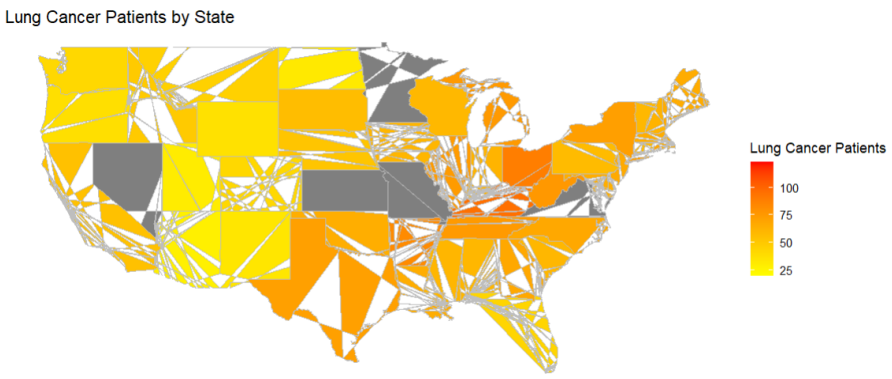




The datasets include information on lung cancer incidence, heart disease prevalence, asthma prevalence, and diabetes prevalence at the county level. The analysis uses linear regression models (lm function in R) to fit regression lines for the relationships between the prevalence of each health condition and lung cancer. Correlation coefficients are calculated to quantify the strength and direction of these relationships. The hypotheses are heart disease-The hypothesis tested is whether there is a significant correlation between heart disease prevalence and lung cancer incidence. Asthma-The analysis explores if there is a significant correlation between asthma prevalence and lung cancer incidence. Diabetes-The hypothesis examines whether there is a significant correlation between diabetes prevalence and lung cancer incidence. The experimental setup is that the datasets are merged based on common fields (StateFIPS, CountyFIPS, County, Year). Linear regression models are then fitted to the merged datasets for each health condition. The relationships are visualized using scatter plots with regression lines. Correlation coefficients are used as external evaluation criteria. These coefficients quantify the strength and direction of the linear relationships between pairs of variables. My R code focuses

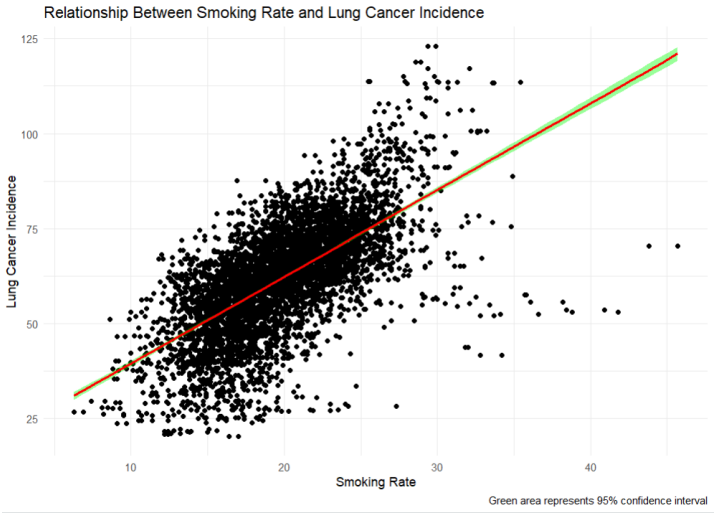
on linear regression analysis and correlation coefficients to explore relationships between health conditions and lung cancer. At the end, the analysis concludes that there is a moderate positive correlation between the prevalence of heart disease, asthma, and diabetes with lung cancer. Dr. Amina Wali noted that heart disease and other cardiovascular diseases are often associated with comorbidities, which are conditions that occur together. One such comorbidity is lung cancer, one of the most common forms of cancer. This implies that counties with higher rates of these health conditions tend to have higher rates of lung cancer, and vice versa.




I am analyzing two datasets, namely "lungcancer_inc_per100k_pop_2015_2019.csv" and "county_noHealthIns_perc_pop_2018_2019.csv," using R code. It focuses on the relationship between health insurance coverage and average lung cancer rates at the county level. The key steps include data loading, merging, visualization, conversion of columns to numeric values, calculation of the correlation coefficient, and a conclusion based on the analysis. The analysis uses a bar plot to visually represent the relationship between health insurance coverage and average lung cancer incidence rates. It calculates the correlation coefficient to quantify the strength and direction of the relationship. The hypothesis is that there is a correlation between the percentage of the population without health insurance and lung cancer incidence rates. The experimental setup is merging datasets using the inner_join function based on common columns ("StateFIPS", "CountyFIPS", "County"). Plotting the relationship with a bar plot and calculating the correlation coefficient. The code uses the correlation coefficient as an external evaluation criterion to measure the strength and direction of the relationship. A correlation coefficient of -0.2011321 suggests a weak to moderate negative linear relationship between health insurance coverage and lung cancer rates. This means that areas with a higher percentage of the population without health insurance tend to have, on average, slightly lower lung cancer incidence rates. The code concludes that there is a relationship between the percentage of the population without health insurance and lung cancer incidence rates. A positive correlation suggests that areas with higher percentages of uninsured population tend to have higher lung cancer incidence rates, and vice versa.



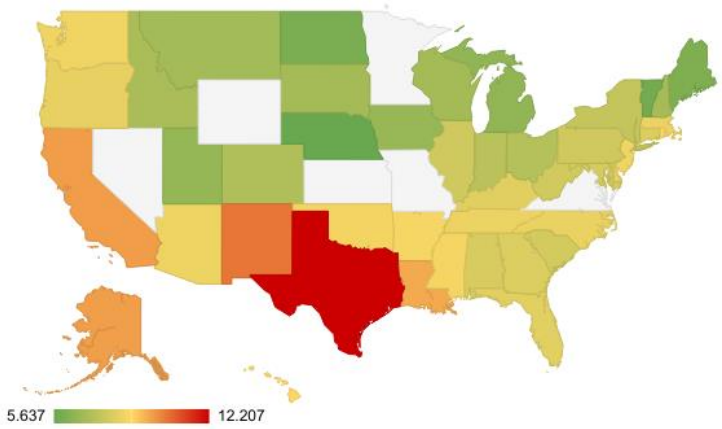

The dataset being analyzed is named "lungcancer_inc_per100k_pop_2015_2019.csv". It contains information about lung cancer incidence rates per 100,000 population for the years 2015 to 2019, broken down by U.S. states. The R code reads the lung cancer dataset and merges it with U.S. state map data to create a choropleth map. The evaluation involves visualizing lung cancer incidence rates across states. The hypotheses I made is that there are variations in lung cancer incidence rates across different U.S. states. The experimental setup involves loading the dataset, merging it with map data, and plotting the choropleth map using ggplot2. The setup does not include explicit statistical testing or hypothesis testing in the provided code. For evaluation criteria, it includes finding the maximum and minimum values of lung cancer incidence rates. This information serves as a basic evaluation criterion, giving insights into the range and extremities of the data. Based on analyzing the map plot and performing data calculations, I concluded that the maximum number of lung cancer patients is 123 in Kentucky. Lowest lung cancer patient number: 20.3 in Idaho.



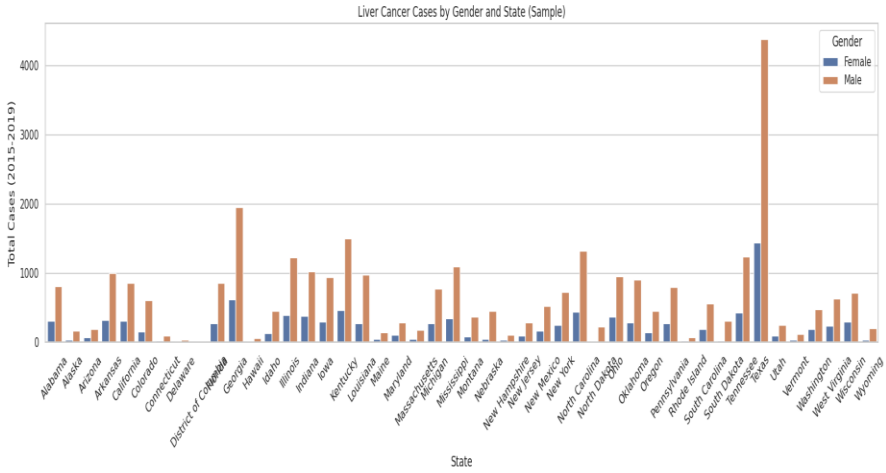

I am using 2 datasets, lung cancer: Contains information on lung cancer incidence per 100,000 population for different counties. Smoke: Includes data on smoking rates per 100,000 adults for the same set of counties. The hypothesis I made is that the analysis is focused on exploring the relationship between smoking rates and lung cancer incidence. The hypothesis is that there is a positive correlation between these two variables. For experimental setup, data is loaded, merged,

and non-finite or missing values are removed. A scatter plot is created to visually inspect the relationship between smoking rates and lung cancer incidence. The correlation coefficient is calculated to quantify the strength and direction of the relationship. The correlation coefficient is used as an external evaluation criterion to measure the degree of linear association between smoking rates and lung cancer incidence. The R code provides a conclusion based on the correlation coefficient and the visual inspection of the scatter plot. It suggests a positive association between smoking rates and lung cancer incidence, reinforcing the well-established link between smoking and lung cancer.

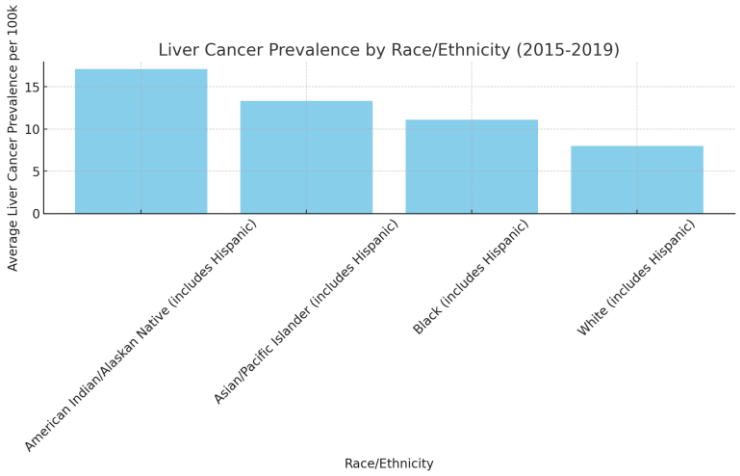
- **Liver Cancer**



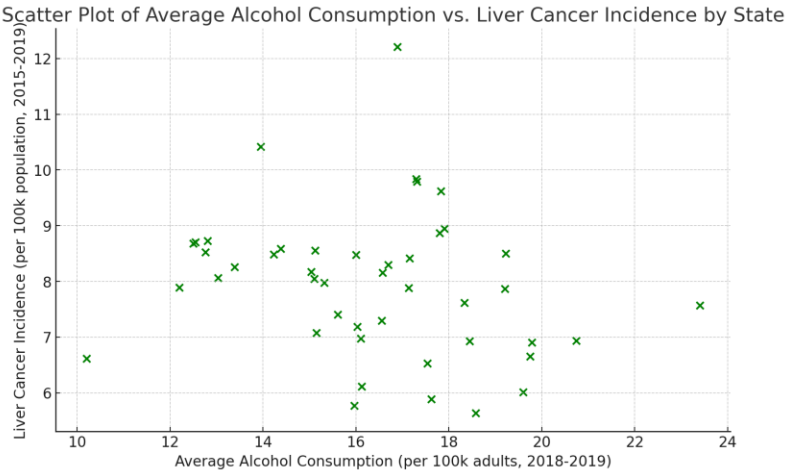

This is a map chart for liver cancer population per state. The data set "livercancer_inc_per100k_pop_2015_2019" was used to make this map chart. We first calculated the average liver cancer patient by state and exported the results into a separate csv file and then used google charts to generate a map chart. By the result, Texas is the state with the highest of 12 liver cancer patients per 100k population and Nebraska is the lowest with 5 liver cancer per 100k population.



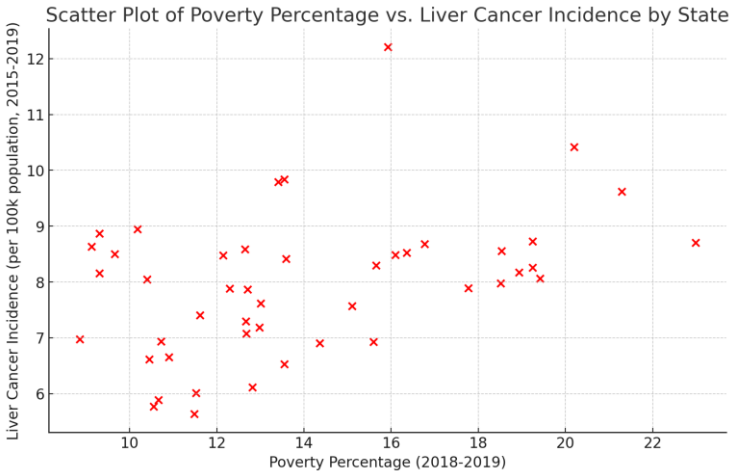

The dataset used for analyzing the relationship between liver cancer and gender is "livercancer_inc_per100k_pop_2015_2019_gender.csv." This file contains information about male and female population with liver cancer per 100k population from each county and state. The value column contains missing data, and they were represented as "suppressed." So, the first thing is to convert the data from string to numeric data type to calculate male and female with liver cancer by state.




The dataset used for analyzing how race is affected by liver cancer is "livercancer_inc_per100k_pop_2015_2019_race.csv." This file contains the information of liver cancer per 100k population by different races in each county and states. We first converted the value from non-numeric data to numeric data then made the bar chart of liver cancer prevalence by race. We can conclude that the native American population has the highest rate of getting liver cancer. Which lines up with previous study findings of liver disease number among the native population is increasing.

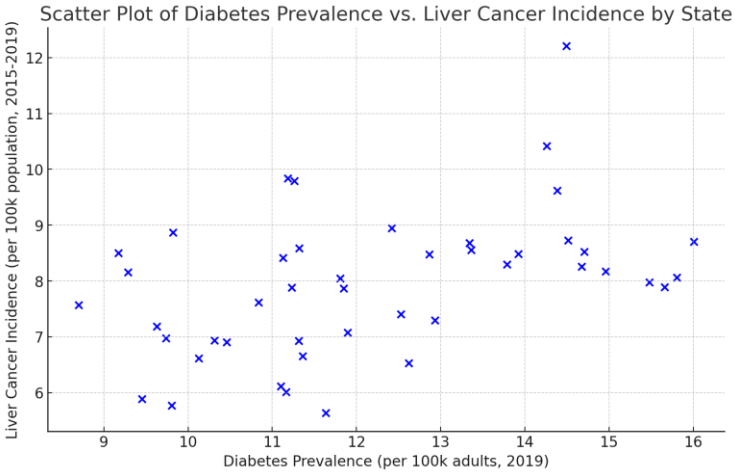



Since alcohol is known to cause damage to liver, average alcohol consumption is also analyzed to determine whether it is a contributing factor for liver cancer. The data set used for this is “binge_drinking_alcohol_adults_per100k_pop_2018_2019.csv” and “liverungcancer_inc_per100k_pop_2015_2019.csv.” The resulting scatter plot revealed a visual representation of the relationship between these two factors. While this analysis does not establish a direct causative relationship, it does provide an important perspective on the potential impact of alcohol consumption on liver cancer prevalence across different states. More research needs to be done to further analyze the relationship between alcohol consumption and liver cancer.

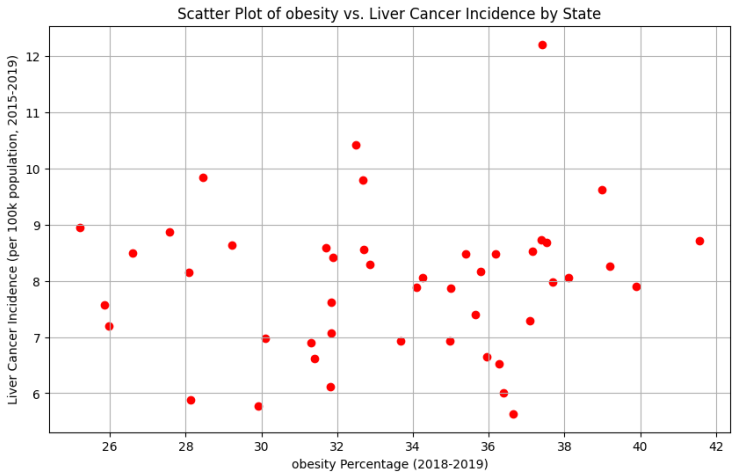



In examining the relationship between poverty and liver cancer, the analysis utilized data from "pop_perc_below_poverty_2018_2019.csv" and "livercancer_inc_per100k_pop_2015_2019.csv". This study focused on the percentage of the population living below the poverty line in each state in comparison to the state-wise liver cancer incidence rates. The scatter plot created from this data provided a visual correlation between

poverty levels and liver cancer. The scatterplot forms a linear pattern suggesting poverty may lead to getting no treatment for liver cancer hence more population in liver cancer patients.




The investigation into the association between diabetes and liver cancer, the data sets "diabetes_adults_per100k_pop_2018_2019.csv" and "livercancer_inc_per100k_pop_2015_2019.csv" is used. This study focused on comparing state-level diabetes prevalence rates with liver cancer incidence rates. The scatterplot shows a linear pattern that suggests diabetes and liver cancer are related. It might be easier for people with diabetes to have liver cancer.




As for obesity vs liver cancer prevalence, the dataset “obesity_adults_per100k_pop_2018_2019.csv” was used. The resulting scatter plot displays a correlation between obesity percentage (2018-2019) and liver cancer incidence (per 100,000 population on 2015-2019) by state. The x-axis represents the obesity percentage ranging from 26% to 42%, and the y-axis represents the liver cancer incidence rate. The distribution of data points suggests a positive correlation between the two variables, indicating that states with higher obesity percentages tend to have higher incidences of liver cancer. However, the relationship does not appear to be strictly linear.

## Results and Discussion

- **Lung Cancer**
  - Male vs. Female Lung Cancer Rates by State:
    - Results: Bar plot comparing lung cancer rates between males and females in different states.
    - Conclusion: In every state, males have higher lung cancer rates than females.
  - Lung Cancer Incidence Rates by Race/Ethnicity:
    - Results: Bar plot showing lung cancer incidence rates for different racial/ethnic groups.
    - Conclusion: White people have the highest lung cancer prevalence rates, while Asians have the lowest. The conclusion is that white people are more likely to get lung cancer than other races.
  - Air Quality vs. Population:
    - Results: The scatter plot does not show a clear trend between air quality and lung cancer incidence.
    - Conclusion: The level of air quality does not directly affect lung cancer prevalence.
  - Arsenic Levels vs. Lung Cancer Rates:
    - Results: Highly significant chi-squared test ($p < 2.2e-16$), strong deviation from independence.
    - Conclusion: Rejecting the null hypothesis; there is a statistically significant association between arsenic levels and lung cancer rates.
  - Alcohol Consumption vs. Lung Cancer Rates:
    - Results: Positive correlation between alcohol consumption and lung cancer rates.
    - Conclusion: While a positive correlation is observed, it does not imply causation. More information is needed.
  - Heart Disease, Asthma, Diabetes vs. Lung Cancer Prevalence:
    - Results: Moderate positive correlations between heart disease, asthma, diabetes, and lung cancer.
    - Conclusion: Counties with higher rates of these diseases tend to have higher rates of lung cancer.
  - Health Insurance Coverage vs. Lung Cancer Rates:
    - Results: Positive correlation between the percentage of the population with no health insurance and lung cancer incidence.
    - Conclusion: Areas with higher percentages of the uninsured population tend to have higher lung cancer incidence rates.
  - Smoking Rate vs. Lung Cancer Incidence:
    - Results: Positive correlation between smoking rate and lung cancer incidence.
    - Conclusion: Smoking is a major preventable cause of lung cancer.

    - Lung Cancer Patients by State:
        - Results: Choropleth map showing the number of lung cancer patients by state.
        - Conclusion: No quantitative analysis presented, but it visually displays the geographic distribution of lung cancer cases.
    - Strengths:

A variety of statistical methods were used, including chi-squared tests, correlation coefficients, and linear regression. Effective use of visualization tools, such as scatter plots, box plots, bar plots, and choropleth maps.

    - Weaknesses:

While correlations are identified, the analyses do not establish causation. Some conclusions lack nuance and emphasize the need for more information. The analysis would benefit from a more comprehensive statistical modeling approach, considering confounding variables and potential interactions.

- **Liver Cancer**
    - Male vs. Female liver Cancer Rates by State
        - Result: The bar chart comparing the population of liver cancer patients by state.
        - Conclusion: Men are easier to get liver cancer than women
    - liver Cancer Rates by race
        - Result: The bar chart showing the population of liver cancer patients by different races
        - Conclusion: Native American population has the most liver cancer patients per 100k population with the average of 17.09 and white population has the least of 7.97
    - liver Cancer Rates vs alcohol consumption
        - Result: A scatter plot with alcohol consumption as x value and liver cancer patients per 100k population as y value.
        - Conclusion: There is not a linear pattern in the scatter plot which means that further research needs to be done.
    - Liver Cancer Rate vs poverty
        - Result: A scatter plot with poverty population as x value and liver cancer patients per 100k population as y value that shows a linear pattern.
        - Conclusion: The pattern indicates that people who live in poverty are likely to get liver cancer since they cannot afford the treatment.

## Related Work

While the presented analysis focuses on understanding the factors influencing lung cancer and liver cancer incidence, related work often explores specific risk factors independently. The unique aspect of this study lies in its comprehensive approach, examining a diverse set of

variables such as air quality, arsenic levels, alcohol consumption, heart disease, asthma, diabetes, health insurance coverage, alcohol consumption, obesity, poverty, and smoking rates. By merging datasets and employing various statistical methods, this research seeks to identify potential interconnections and correlations among these factors and cancer prevalence. Many existing studies tend to concentrate on singular risk factors, whereas this work aims to provide a more comprehensive and nuanced understanding by considering multiple variables simultaneously. Additionally, the incorporation of visualizations, regression analyses, and statistical tests contributes to a richer interpretation of the complex relationships between these factors and lung cancer and liver cancer, distinguishing this approach from more traditional univariate analyses in related works.

## Conclusion

Analysis of several factors influencing lung cancer incidence revealed several important results. The study found that arsenic levels and smoking rates were significantly associated with increased incidence of lung cancer, strengthening the established link. Additionally, heart disease, asthma, and diabetes were moderately positively associated with lung cancer, suggesting potential interconnected health effects. The study also highlights that areas with higher proportions of the uninsured population tend to have higher lung cancer rates. Additionally, this project highlights the importance of considering multiple factors, such as air quality, alcohol consumption, and health insurance coverage, when understanding the complex picture of lung cancer prevalence. In the future, this work could be expanded by incorporating additional socioeconomic and environmental variables, considering temporal trends, and employing more advanced predictive modeling techniques to gain a comprehensive understanding of lung cancer risk factors. By analyzing the factors that potentially cause liver cancer, we have concluded that it is easier for men to get liver cancer than women and among different races, it is more likely for the native American population to get liver cancer. Also due to lack of access to proper treatment, there are more people who live in poverty getting liver cancer. Both obesity vs liver cancer and diabetes vs liver cancer's scatter plot show a positive relation while obesity vs liver cancer is not strictly linear. While alcohol is known to be bad for livers, drinking alcohol and getting liver cancer does not seem to have strong connections by analyzing the scatterplot, however an article on cancer.org suggests that long term use of alcohol is a factor of causing liver cancer, so the result in our research might be due to inconsistent data.

## Bibliography


Relationship between exposure to PM2.5 and lung cancer incidence and mortality: A meta-analysis - PMC (nih.gov)

How Alcohol Affects Lung Cancer (webmd.com)

Lung Cancer and Heart Disease: Understanding the Connection | MyHeartDiseaseTeam

Alcohol Use and Cancer

Racial disparities in liver cancer: Evidence for a role of environmental contaminants and the epigenome