DATASCI 200: Project 2 Report

Cancer Deep Dive: Mortality and Incidence Rates Across the U.S.

Team: Joe Uren, Kevin Yi, Akshay Sharathchandra, Emily Zhang
Github repository:
https://github.com/UC-Berkeley-I-School/Project2_Yi_Zhang_Sharathchandra_Uren

Introduction

Recent press coverage of cancer rates in America has been rather bleak. You see news articles sounding alarms that younger Americans are being diagnosed with cancer at a higher rate than ever before. The passing of famous actor Chadwick Boseman, known for his Black Panther role, in his 40s to colon cancer surely did not help. Our project team decided to analyze data from the U.S Department of Health & Human Services Centers for Disease Control and Prevention data to investigate the data for ourselves. The report is a summary of our findings.

For our analysis, we investigated the Centers for Disease Control and Prevention data on U.S Chronic Disease Indicators CDI, 2023 Release[1]. On initial glance, the data consisted of over a million rows of self reported data separated by state. Looking through the data, we decided to concentrate our efforts on two of the top cancers in America, lung and colon cancer[2].

Research Question

After briefly looking at the available information from our data, we formulated the following question**: how do the mortality and incidence rates for colon and lung cancer differ amongst location, racial groups, and sex within the US from 2008-19?** To answer the question, we broke up our analysis into three main areas:

1. Overall incidence and mortality rate for lung and colon cancer by state from the time period in the data
2. Sex differences for lung and colon cancer incidence and mortality rate in the U.S.
3. Race and ethnicity differences for lung and colon cancer incidence and mortality rate in the U.S.

Data Cleaning and Preprocessing

1. We began by checking the shape of the original data, which showed that it included 1,185,676 rows and 37 columns in total.
2. As we began to narrow down the scope of our analysis, we ensured that our data had enough information on lung and colon cancer. Our data had 9984 rows of data each for incidence and mortality rate for lung and colon cancer:

[1] Data from: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi
[2] 'Top' cancer was defined as the highest average annual number of the incidence during the period of 2015-2019 https://www.cancer.org/research/acs-research-news/facts-and-figures-2024.html

3. We went through all 37 columns to determine what columns held relevant information. (Appendix - Figure 1). We found that 10 columns had all null values and dropped them (Appendix - Figure 2).

Assumptions:

While working with this dataset, we conducted our analysis under the following assumptions related to the collection, reporting, and compilation of the data. Since the dataset is made up of data collected at the state level, we assumed:

- The data is collected in a similar manner across states.
- The data is collected in a similar manner across regions (urban vs rural), and is representative of the state as a whole.
- The data collected regarding race/ethnicity is accurate (self-reported vs state-reported).

Of note, if the data provider (state) decides to suppress data for quality or confidentiality reasons, the CDC does not report this data, so we also assume this data has surpassed some sort of QC-check at the state level.

Exploratory Data Analysis:

After removing columns and filtering our rows, we now moved onto further exploring our data to ensure that we had the right data to answer our questions.

1. We began to investigate the rest of the columns:
    - Checked that all states were included in our data. Checked the lat/long coordinates and realized they were all the same for each state. Since it was not unique, we decided to drop it since we only need the column for the state.
    - Checked data equally (generally) represented male, female, and overall. (Appendix 3)
    - Checked data had information on race/ethnicity. (Appendix 3)
    - Checked our Datavalue type. There were several different value types (Appendix 4). We decided to pick Age-Adjusted Rate since it looked to be the health industry standard in representing cancer rate across state lines so that the numbers can be compared to each other. An aging population will have higher cancer incidence and mortality rate and you want to control that when comparing to another state that might have a younger population.[3] Age -Adjusted Rate tries to remove the confounding age factor from the equation.
    - We looked at our data values and they were from 2001 to 2021 (Appendix 5). Our stratification columns showed that our data values were an average value out of a population of 100,000 for a 4 year time period.

Highlights:

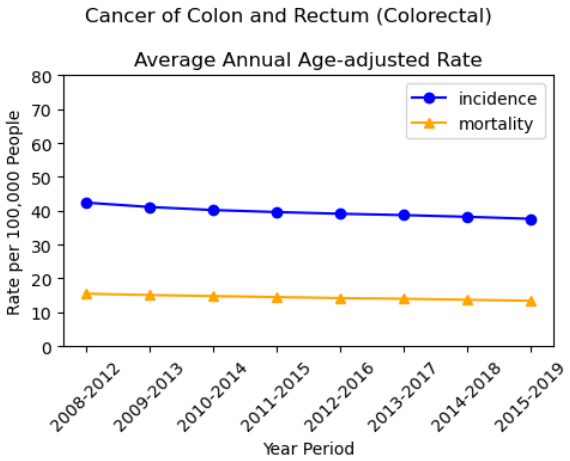
- Our analysis of the CDC data on lung and colon cancer reveals declining age-adjusted incidence and mortality rates across sex and ethnicities.
- Males show higher rates, with steeper declines compared to females.

[3] https://www.health.ny.gov/statistics/cancer/registry/age.htm

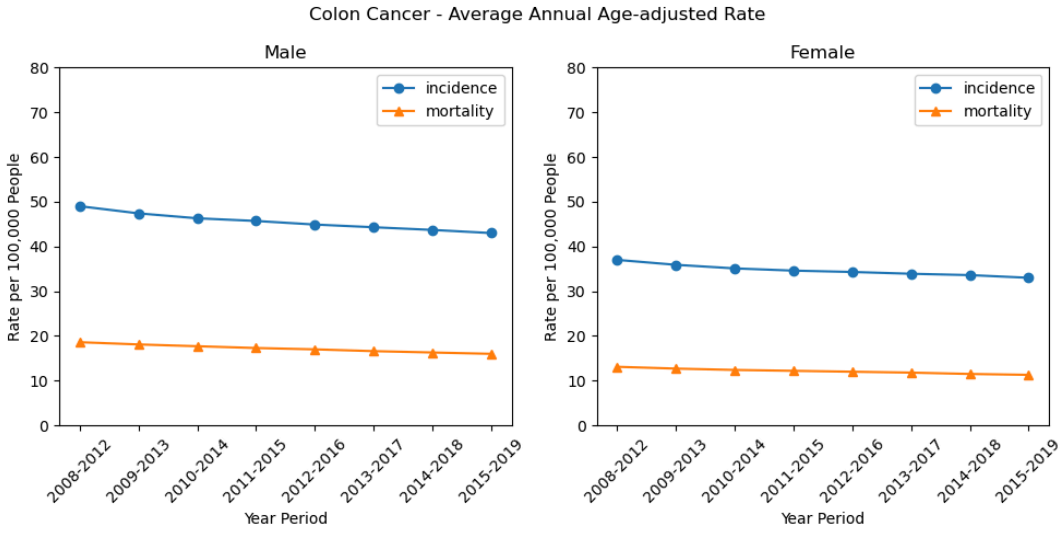
- Regionally, the Southeastern and Appalachian areas exhibit higher lung cancer rates, necessitating targeted public health interventions.
- Asian populations have the lowest colon cancer mortality, while Black non-Hispanics face the highest.
- Future efforts should focus on enhancing screening, improving healthcare access, and tailored public health campaigns to further reduce cancer rates and address disparities.

Key Insights for Colon Cancer

1. **Colon Cancer Decreasing Age-adjusted Rates**: Both the incidence and mortality age-adjusted rates are declining across all groups (male, female, and overall), reflecting a combination of population growth, demographic shifts towards an older population, and improvements in healthcare, early detection, and treatments for colorectal cancer. We need further deepdive to attribute the impact from each element.



2. **Colon Cancer Sex Differences**: Males have higher incidence and mortality rates compared to females, with the trends showing that male mortality numbers align with incidence numbers (both increasing over time) while female mortality numbers trend opposite to incidence numbers (female incidence increases slightly but mortality decreases). This could be attributed to a combination of behavioral, biological, and healthcare-related factors:
    a. **Males**: Higher incidence and mortality, with both increasing, may be due to continued engagement in high-risk behaviors(such as smoking, heavy alcohol consumption, and poor diet), lower screening rates, later stage diagnosis, and potentially more aggressive tumor biology. These factors lead to both higher incidence and higher mortality rates.
    b. **Females**: While incidence is slightly increasing (potentially due to better detection), mortality is decreasing due to higher screening rates, early detection, better adherence to treatment, and healthier lifestyle choices. Advances in treatment and effective public health initiatives also play a significant role in reducing mortality despite a slight increase in incidence.



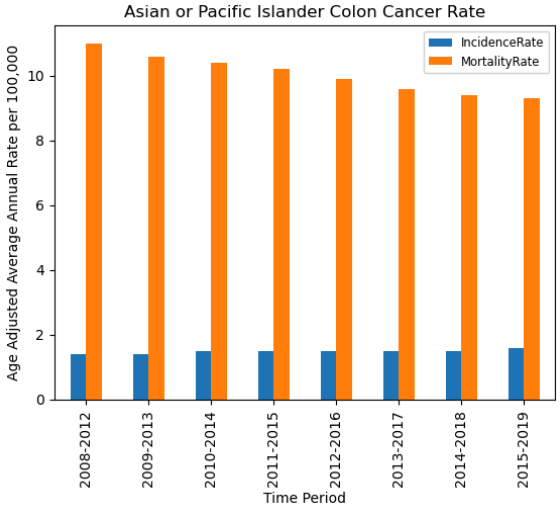
3. **Colon Cancer Race/Ethnicity Differences:**
   Mortality and Incidence rank. 1 being the lowest rate.

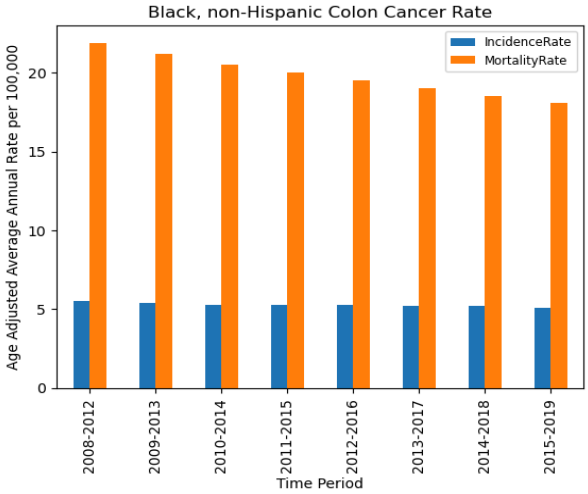
Race/Ethnicity	**Mortality Rank**	**Incidence Rank**
Asian or Pacific Islander:	**1**	**2**
U.S American Indian or Alaska Native	**2**	**1**
Black, Non Hispanic	**5**	**4**
Hispanic	**2**	**5**
White, Non Hispanic	**4**	**2**

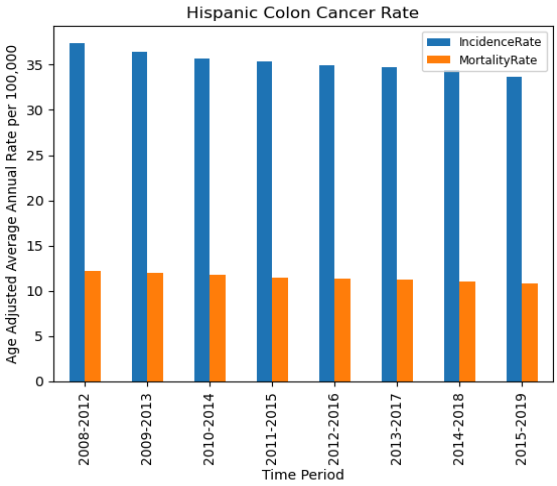
   a. Asians or Pacific Islander have the lowest mortality rate, which might be linked to dietary habits. Traditional Asian diets, rich in vegetables, fruits, and fish, may contribute to lower colon cancer mortality rates. Studies suggest that diet plays a significant role in colon cancer prevention.
   b. U.S American Indian or Alaska Native: The lower incidence rate may result from underreporting due to limited access to healthcare facilities and screenings in rural areas. Increased efforts in healthcare access and screening could provide more accurate data and improve early detection.
   c. Black, Non Hispanic: This group has the highest mortality rate. Factors could include disparities in healthcare access, socioeconomic status, and genetic predispositions. Efforts to improve healthcare access and early screening in this community are crucial.

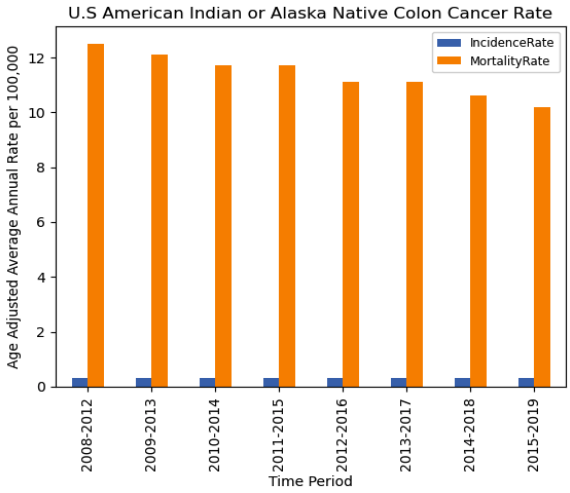
d. Hispanic: Hispanics show one of the highest incidence rates. This could be influenced by lifestyle factors, diet, and healthcare access. Culturally tailored health education and improved access to preventive care can help address this issue.
e. White, Non Hispanic: This group also has high incidence rates. Lifestyle factors, such as diet and physical activity, along with genetic factors, could contribute. Promoting healthy lifestyles and regular screenings can aid in reducing incidence rates.

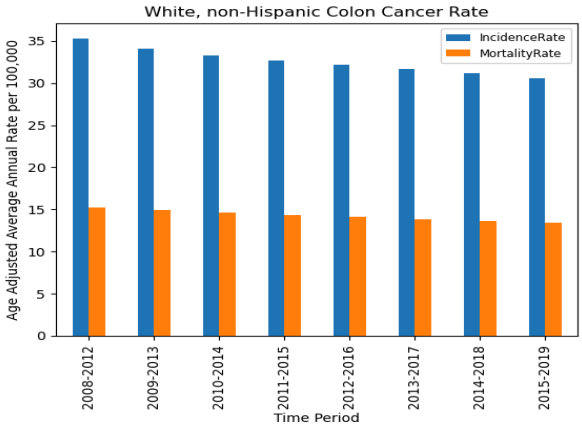
Overall, the rate of incidence and mortality has seen ups and downs but have been relatively stable. Since this reflects an age adjusted rate, it does match up with media coverage on how rates have gone down in the older population but have been steadily increasing in younger populations.



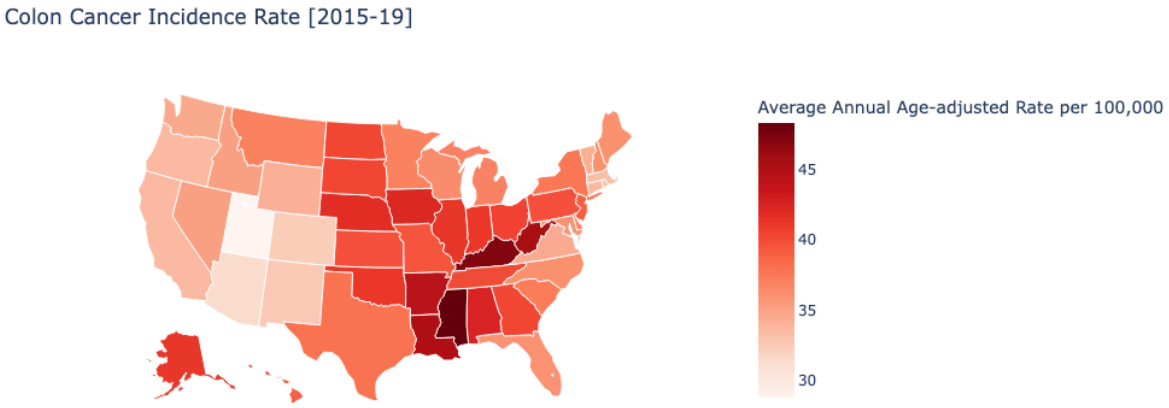


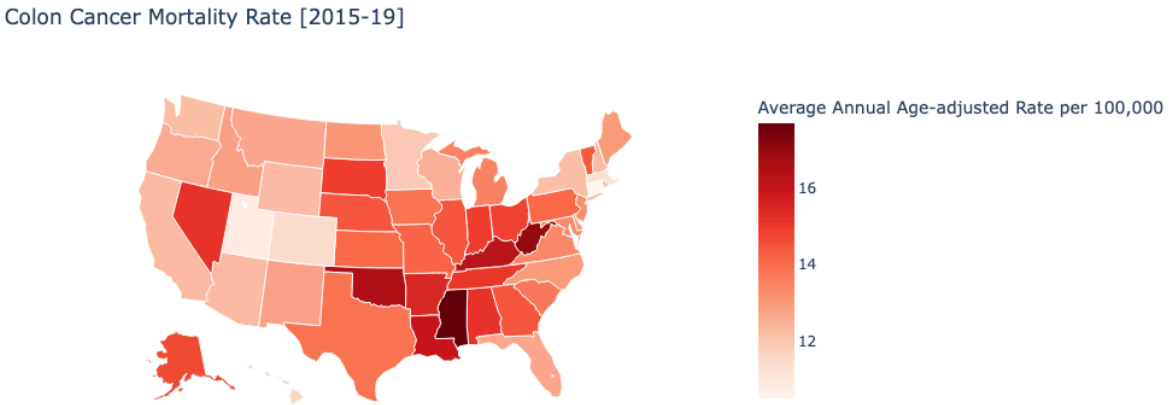






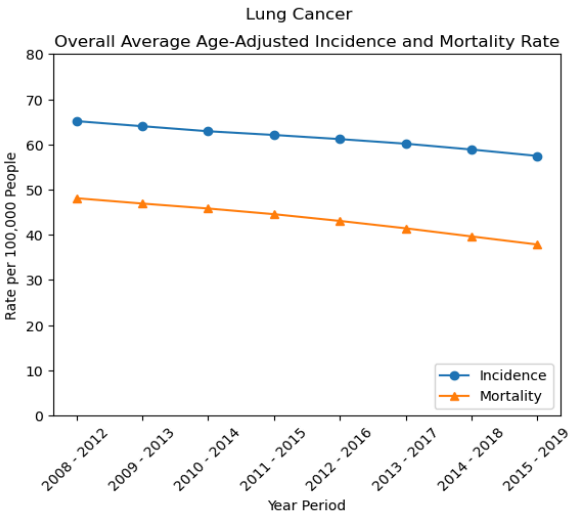
4. **Colon Cancer Overall Mortality and Incidence Heat Map.**
    a. Generally it looks as if colon cancer mortality and Incidence is the highest in the middle of the country both north and south. The states with the highest rates are the southern states: Mississippi, then Kentucky, Louisiana, and Arkansas.
    b. The higher mortality rates in these regions could be associated with lower access to healthcare and screening services, leading to later-stage diagnosis and lower survival rates.
    c. States with lower incidence and mortality might benefit from better preventive healthcare measures, including screening programs and public health initiatives focused on diet and lifestyle.



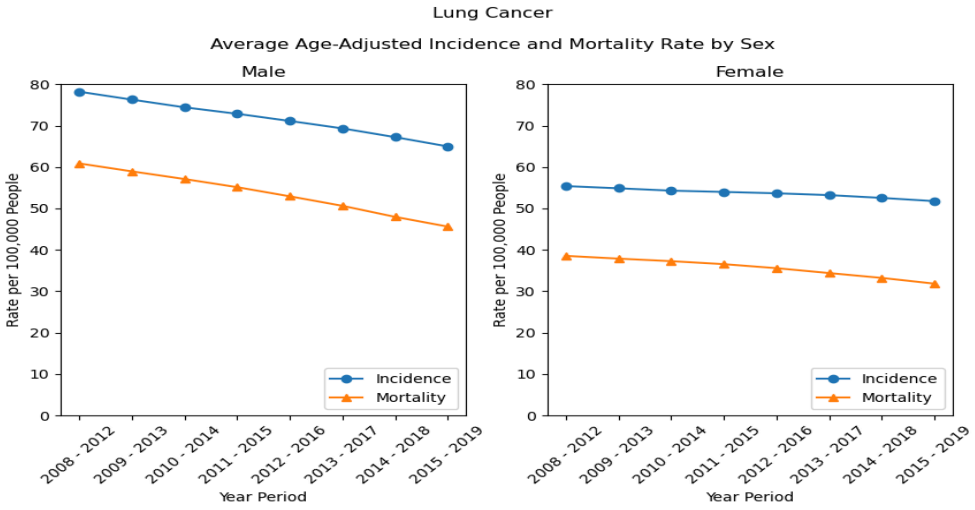


Key Insights for Lung Cancer

1. **Lung Cancer Average Overall Age-Adjusted Rates**:
    a. The graph showing the 'Overall Average Age-Adjusted Incidence and Mortality Rate' shows a decrease in the average incidence and mortality rates for lung cancer. This decrease could be related to changes in lifestyles, early detection, or developments in treating lung cancer. Further investigation would be required to determine the cause of this decline.
    b. Both the incidence and mortality age-adjusted rates are declining across all groups. Additionally the gap between incidence and mortality widens slightly, showing that the survival rate has improved. This improvement is likely due to improved medical technology, early detection, and better access to healthcare.

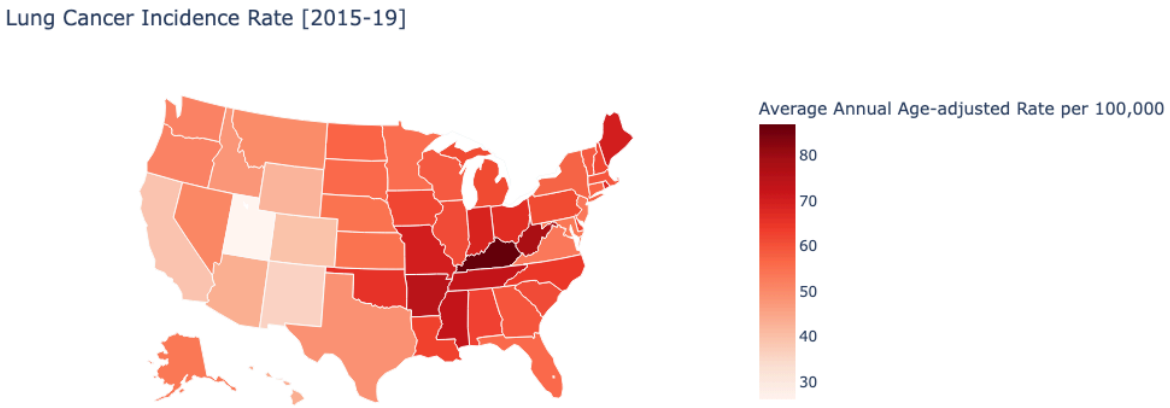


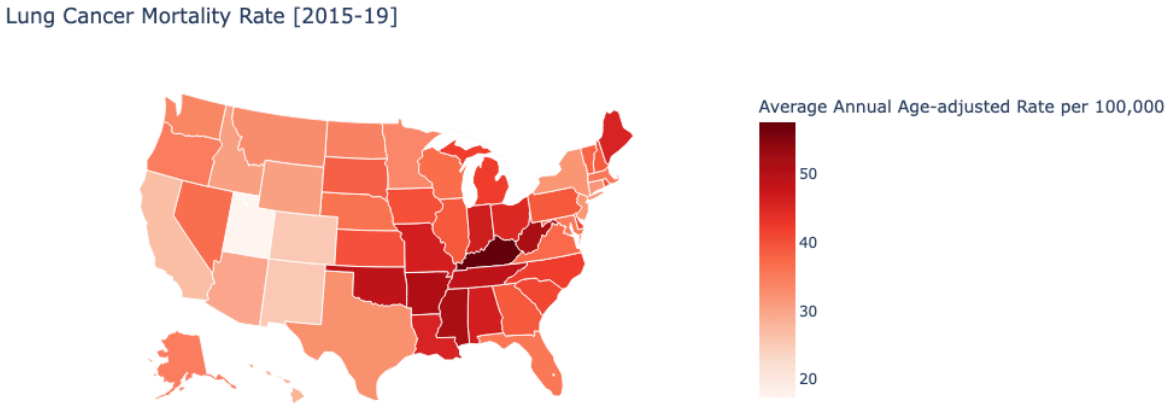
2. **Lung Cancer Sex Differences in Average Overall Age-adjusted Rates**:
    a. More males are developing lung cancer which leads to a higher incidence rate than females. This higher rate leads to males also having a higher mortality rate.
    b. While the incidence and mortality age-adjusted rates are declining across all groups, when broken out by sex the male incidence and mortality rates show a steeper decrease than female rates.
    c. This could potentially be due to factors like differences in lifestyles, smoking, healthcare, occupations, or other factors.



3. **Lung Cancer State Heat Maps**:
    a. Regional Patterns: There is a clear regional pattern where the Southeastern and Appalachian regions show both high incidence and mortality rates for lung cancer. This suggests common risk factors such as higher smoking rates, lower socioeconomic status, and possibly less access to healthcare services.
    b. Public Health Implications: The consistent high rates of incidence and mortality in these regions indicate a need for targeted public health interventions, including smoking cessation programs, increased access to early screening and treatment facilities, and education on lung cancer prevention.

c. Healthcare Access: States with lower incidence and mortality rates may benefit from better healthcare infrastructure, higher rates of early screening, and effective public health campaigns.



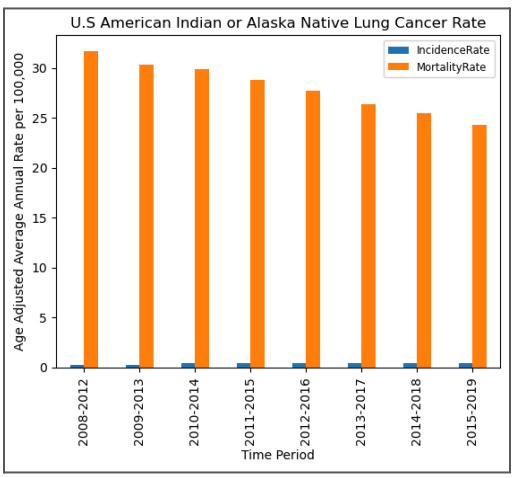


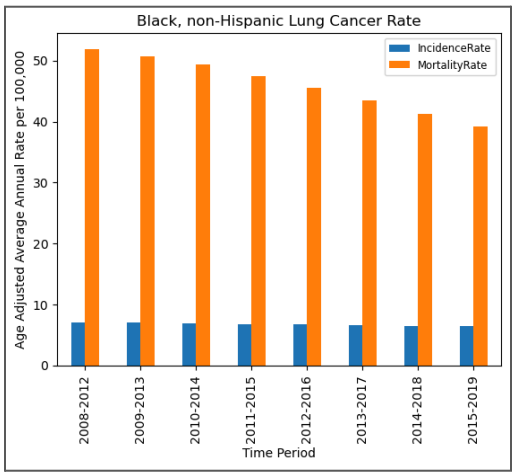
4. **Lung Cancer Race/Ethnicity Differences:**
   Mortality and Incidence rank. 1 being the lowest rate.

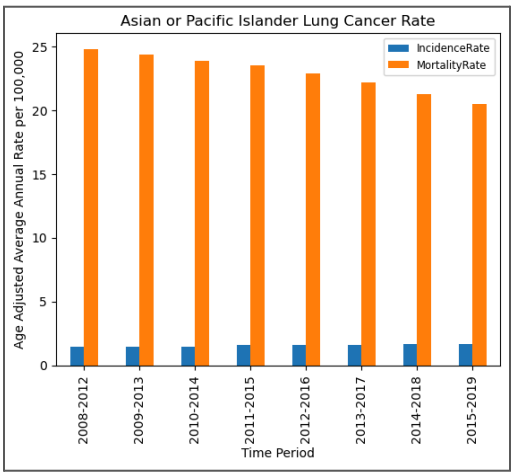
Race/Ethnicity	**Mortality Rank**	**Incidence Rank**
Asian or Pacific Islander:	**4**	**2**
U.S American Indian or Alaska Native	**3**	**1**
Black, Non Hispanic	**1**	**3**
Hispanic	**5**	**4**

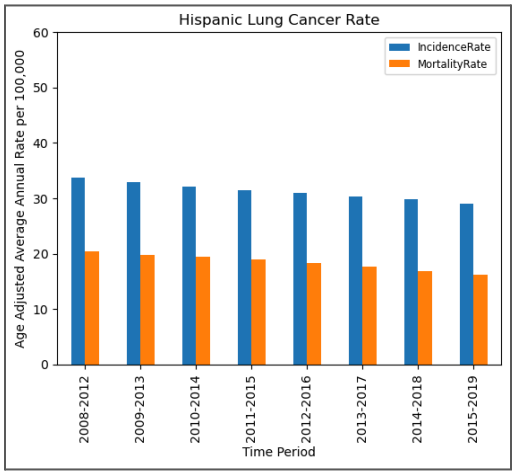
White, Non Hispanic	**2**	**5**

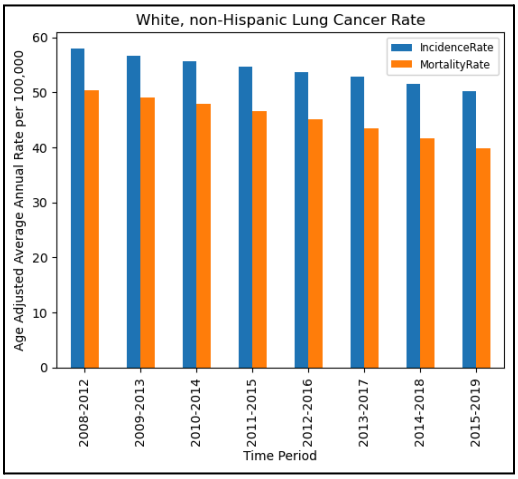
a. U.S. American Indian or Alaska Native: Mortality rates remain much higher than incidence rates, indicating potential issues with late-stage diagnosis and treatment access. This could suggest lack of access to health care including early detection. The mortality rate decreases over time, and the incidence rate seems to slightly increase.
b. U.S. Black: As with American Indian/Alaska Natives, mortality rates are much higher than incidence rates. This could suggest lack of access to health care including early detection. The mortality rate decreases over time, while the incidence rate remains roughly the same.
c. U.S. Asian or Pacific Islander: Mortality rates are lower compared to other groups, incidence rate is lower than all groups other than Hispanics. These data Indicates better overall health outcomes, possibly due to lifestyle factors and healthcare access. Mortality rate decreases over time while incidence rate slightly increases.
d. U.S. Hispanic: Incidence rates are higher than mortality rate. Both rates decrease over time, suggesting improvements in early detection and treatment over time. Unclear answer as to why incidence rate is higher than mortality rate unlike with other minority groups.
e. U.S. White: Higher incidence rates with a declining trend in mortality rates. Shows progress in early detection and effective treatment options.











Appendix:

Figure 1:

```
Index(['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'DataSource',
       'Topic', 'Question', 'Response', 'DataValueUnit', 'DataValueType',
       'DataValue', 'DataValueAlt', 'DataValueFootnoteSymbol',
       'DatavalueFootnote', 'LowConfidenceLimit', 'HighConfidenceLimit',
       'StratificationCategory1', 'Stratification1', 'StratificationCategory2',
       'Stratification2', 'StratificationCategory3', 'Stratification3',
       'GeoLocation', 'ResponseID', 'LocationID', 'TopicID', 'QuestionID',
       'DataValueTypeID', 'StratificationCategoryID1', 'StratificationID1',
       'StratificationCategoryID2', 'StratificationID2',
       'StratificationCategoryID3', 'StratificationID3', 'longitude',
       'latitude', 'YearPeriod'],
      dtype='object')
```

Figure 2:

```
YearStart                       0
YearEnd                         0
LocationAbbr                    0
LocationDesc                    0
DataSource                      0
Topic                           0
Question                        0
Response                  1185676
DataValueUnit              152123
DataValueType                   0
DataValue                  378734
DataValueAlt               381098
DataValueFootnoteSymbol    791966
DatavalueFootnote          791966
LowConfidenceLimit         503296
HighConfidenceLimit        503296
StratificationCategory1         0
Stratification1                 0
StratificationCategory2   1185676
Stratification2           1185676
StratificationCategory3   1185676
Stratification3           1185676
GeoLocation                 10166
ResponseID                1185676
LocationID                      0
TopicID                         0
QuestionID                      0
DataValueTypeID                 0
StratificationCategoryID1       0
StratificationID1               0
StratificationCategoryID2 1185676
StratificationID2         1185676
StratificationCategoryID3 1185676
StratificationID3         1185676
longitude                   10166
latitude                    10166
YearPeriod                      0
dtype: int64
```

Dropping the columns that only contain 'NaN' values.

Figure 3:

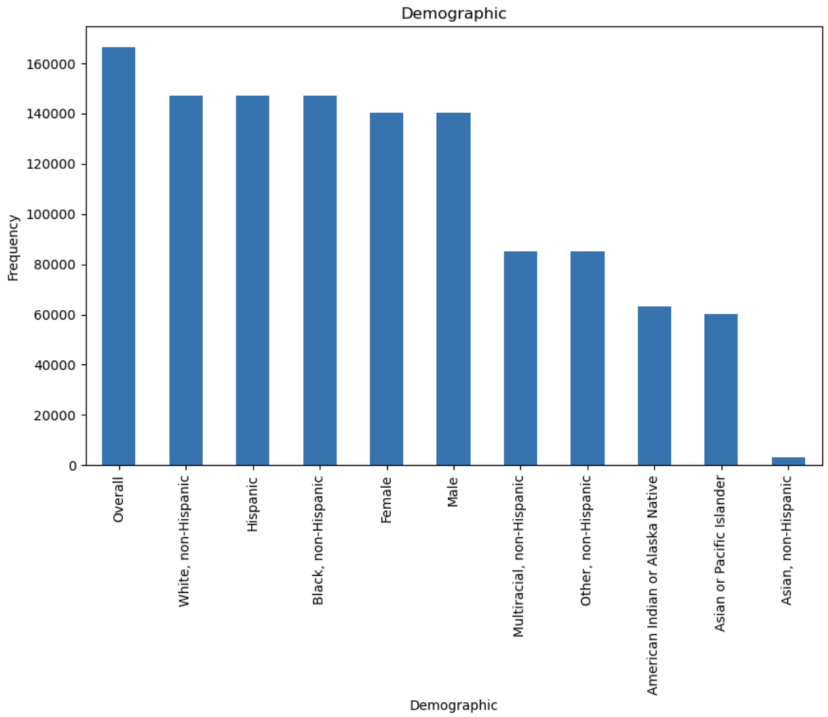


Figure 4:

```
Crude Prevalence                                                    370903
Age-adjusted Prevalence                                             282080
Number                                                              105407
Age-adjusted Rate                                                   102457
Crude Rate                                                          102457
Average Annual Number                                                53248
Average Annual Crude Rate                                            53248
Average Annual Age-adjusted Rate                                     53248
Mean                                                                 24145
Age-adjusted Mean                                                    23925
Median                                                                7200
Percent                                                               2741
Yes/No                                                                2102
US Dollars                                                             972
Adjusted by age, sex, race and ethnicity                               832
Per capita alcohol consumption                                         330
Local control of the regulation of alcohol outlet density              165
Commercial host (dram shop) liability status for alcohol service       164
Prevalence                                                              52
Name: DataValueType, dtype: int64
```

Figure 5:

```
count   1,185,676.00
mean        2,015.10
std             3.32
min         2,001.00
25%         2,013.00
50%         2,015.00
75%         2,018.00
max         2,021.00
Name: YearStart, dtype: float64
```

```
chronic['YearEnd'].describe()
```

```
count   1,185,676.00
mean        2,015.64
std             3.00
min         2,001.00
25%         2,013.00
50%         2,016.00
75%         2,018.00
max         2,021.00
Name: YearEnd, dtype: float64
```