Executive Report: 5415 Final Project

Healthiest Hometowns

Bates Crowther, Jeremy Green, Charlie Rudy, Jozie Wille

Prior to choosing a dataset to examine, we wanted to decide on a subject that everyone in the group was enthusiastic about. After many hours researching datasets across the internet, we found our current dataset from Kaggle. We are all interested in health and life expectancy and grew up in different parts of the country, which led to our interest in this dataset.

Business Understanding

The problem that we chose to examine was life expectancy in years and the health status of major American cities. We were motivated to solve this issue because we have strong beliefs in the importance of living in a healthy environment.

Our hypothesis was that there were significant discrepancies in life expectancy between various US cities and we examined which specific variables contributed to these differences.We decided to employ data analytics to solve this issue because it is an objective approach and does not take into account pre existing ideas and opinions about US cities. Data analytics enables us to analyze the effects of the many variables listed below by providing an objective picture of life expectancy and death rates.

Data Understanding

We downloaded the Big City Health dataset, obtained initially from Data World, via Kaggle. The dataset illustrates the health status of 26 of the nation's largest and most urban cities, as captured by 34 health-related indicators. These indicators represent some of the leading causes of mortality in the United States. Public health data was captured in nine overarching categories: HIV/AIDS, cancer, nutrition/physical activity/obesity, food safety, infectious disease, maternal and child health, tobacco, injury/violence, and behavioral health/substance abuse.

The target variable we chose to analyze was the life expectancy value because we believe life expectancy is the most informative metric for evaluating the health of big cities. We dug deeper into the data to find other indicators that might impact life expectancy.

Data Preparation

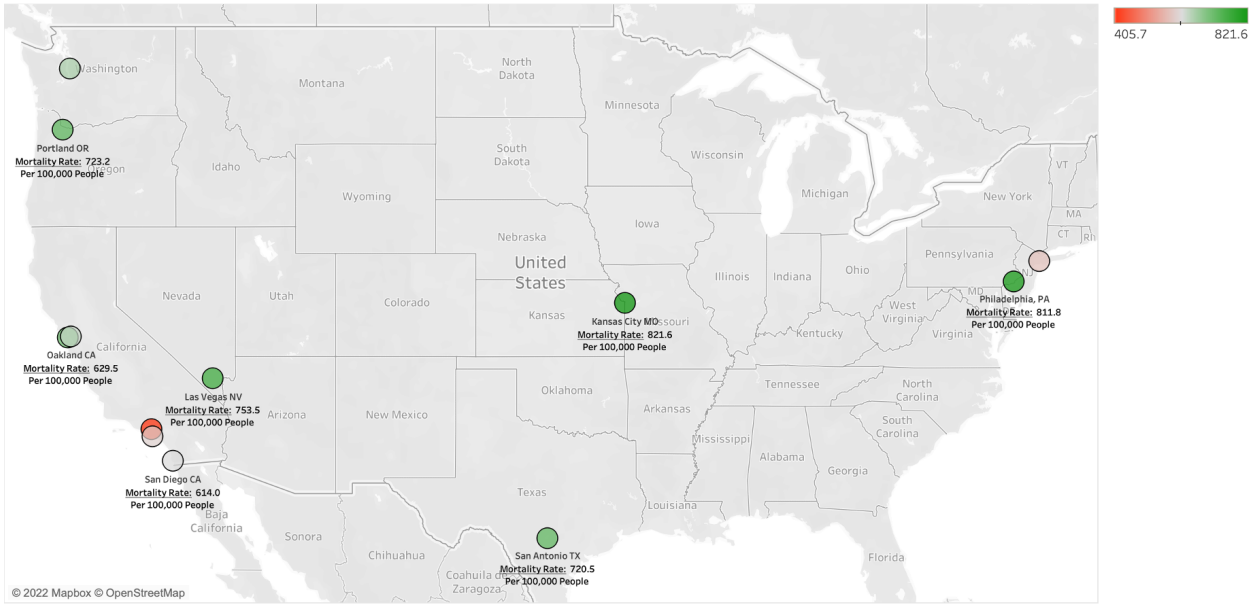
Previously, all the datasets our group worked with were reasonably clean and uniformly formatted, so we did not have to do any preliminary data cleaning work. These experiences led us to the mistake of trying to dive directly into running models and performing analysis. We quickly realized that we could not do any analysis until we addressed the significant inconsistencies in the dataset's format and created subsets to make the data set more manageable. The data in Big City Health spanned ten years, but each year contained a vastly different amount of data. To clean the dataset, we dropped the range values such as '2003-2012' and '2011-2013' and kept only the years formatted as singular values such as '2011'. We maintained the integrity of our data set by keeping all rows with a singular year value and dropping the 21 rows with year ranges. By counting the number of data points in each year and selecting the three years with the most similar amounts of data, it made the year-to-year comparisons as accurate as possible, with 3,498 data points from 2011, 3,947 from 2012, and 3,652 from 2013. These three years had the most similar number of data points, so we decided that for the rest of our analysis, we would be using these 11,097 rows from either 2011, 2012, or 2013.

Another issue that our group faced was that 'All-Cause Mortality Rate (Age-Adjusted; Per 100,000 people)' produced a rate, while 'Life Expectancy at Birth (Years)' produced a value in years, and we could not compare these two metrics because of their differing units. To address this problem, we created subsets with uniform units of measurement, which were necessary to conduct an accurate analysis.

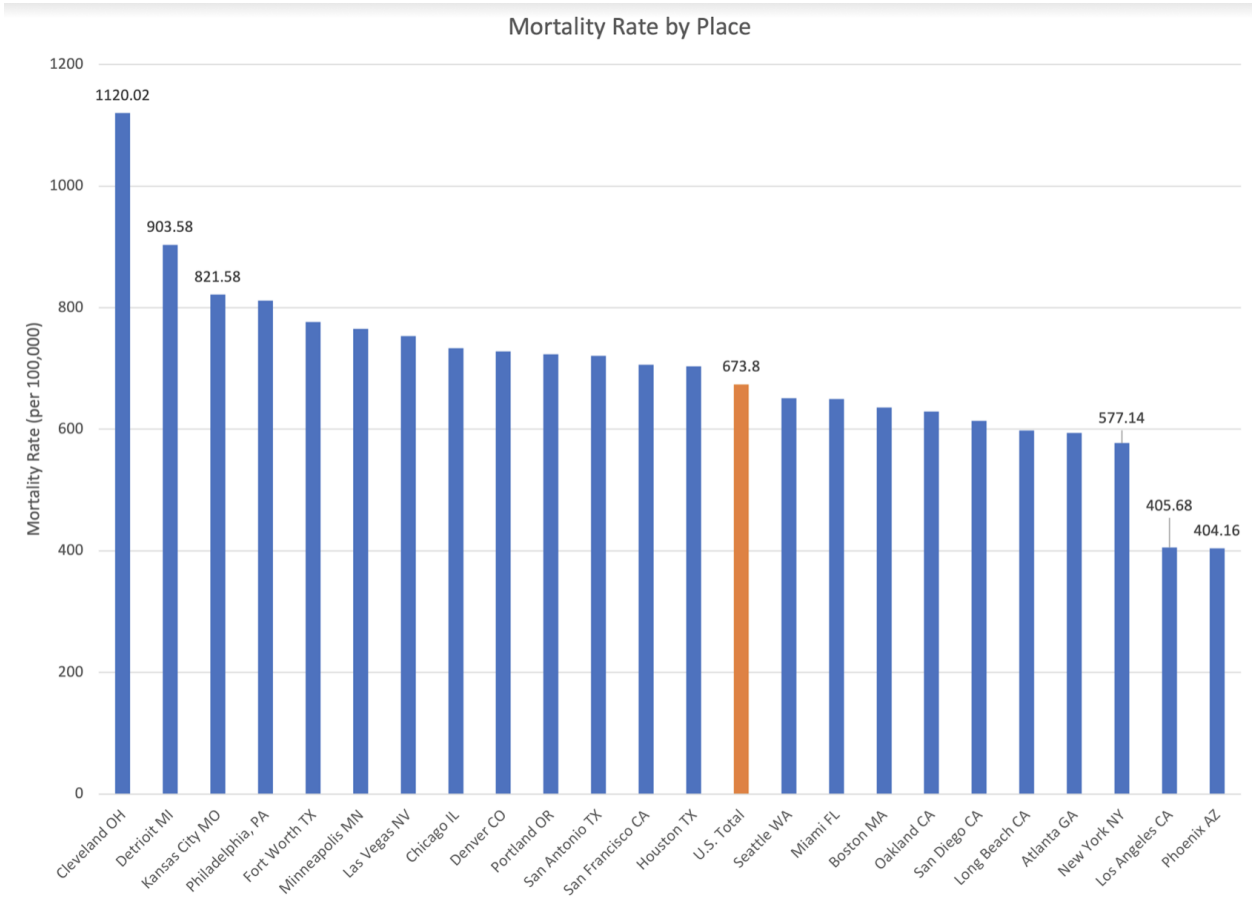
Following that, we eliminated the following variables from our dataset since they exhibited death rates that were unrelated to city health variables. These elements might influence the overall mortality rate but would not aid in our understanding of the healthiest cities to live in. These indicators included Motor Vehicle Mortality Rate, Unemployment Rate, Firearm-Related Mortality, Homicide Rate, and Percent Foreign Born.

After fitting a linear model, the three cities with the highest mortality rates, meaning the most recorded deaths per 100,000 people from 2011-2013, were Cleveland (1120.02), Detroit (903.59), and Kansas City (821.58). The three cities with the lowest mortality rates, meaning the least recorded deaths from 2011-2013, were New York (577.14), Los Angeles (405.68), and Phoenix (404.16). Based on the dataset, the US death rate was 673.80 per 100,000, indicating that the top three cities had mortality rates that were much higher than the national average while the bottom three cities had mortality rates that were lower.

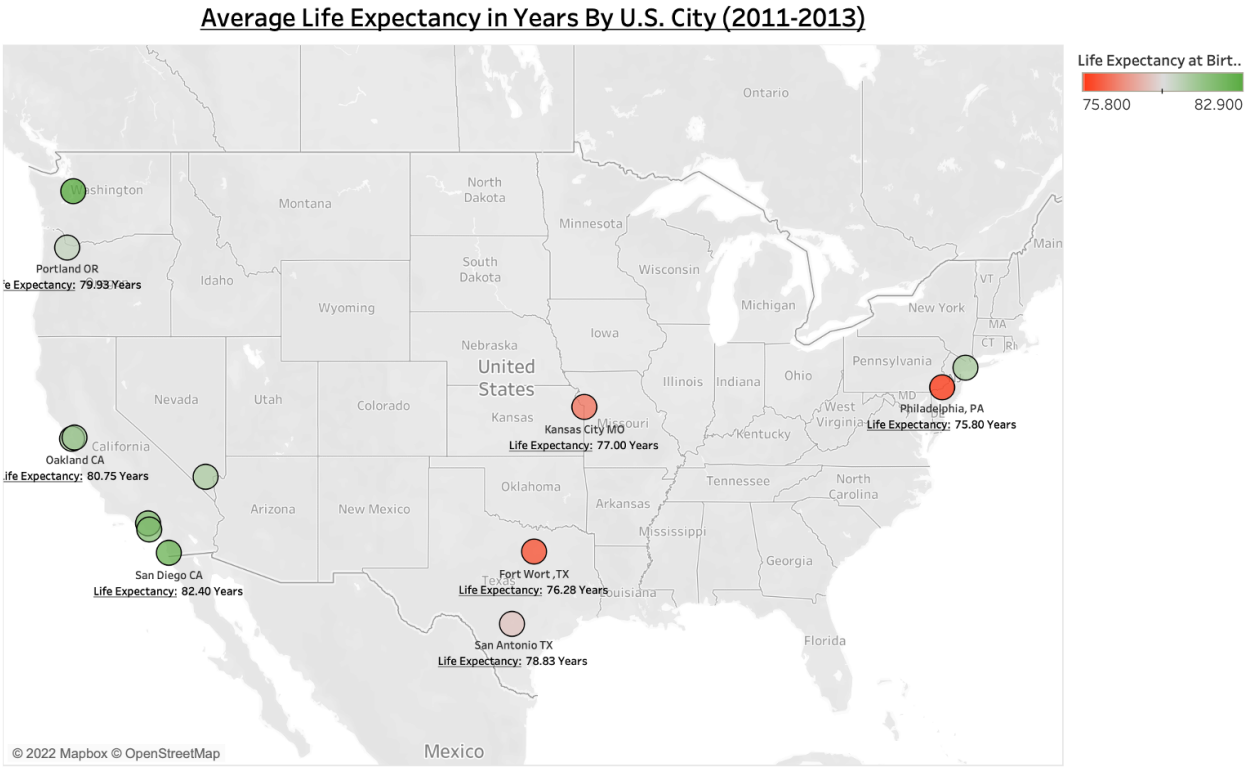
Average Mortality Rate Per 100,000 People By City



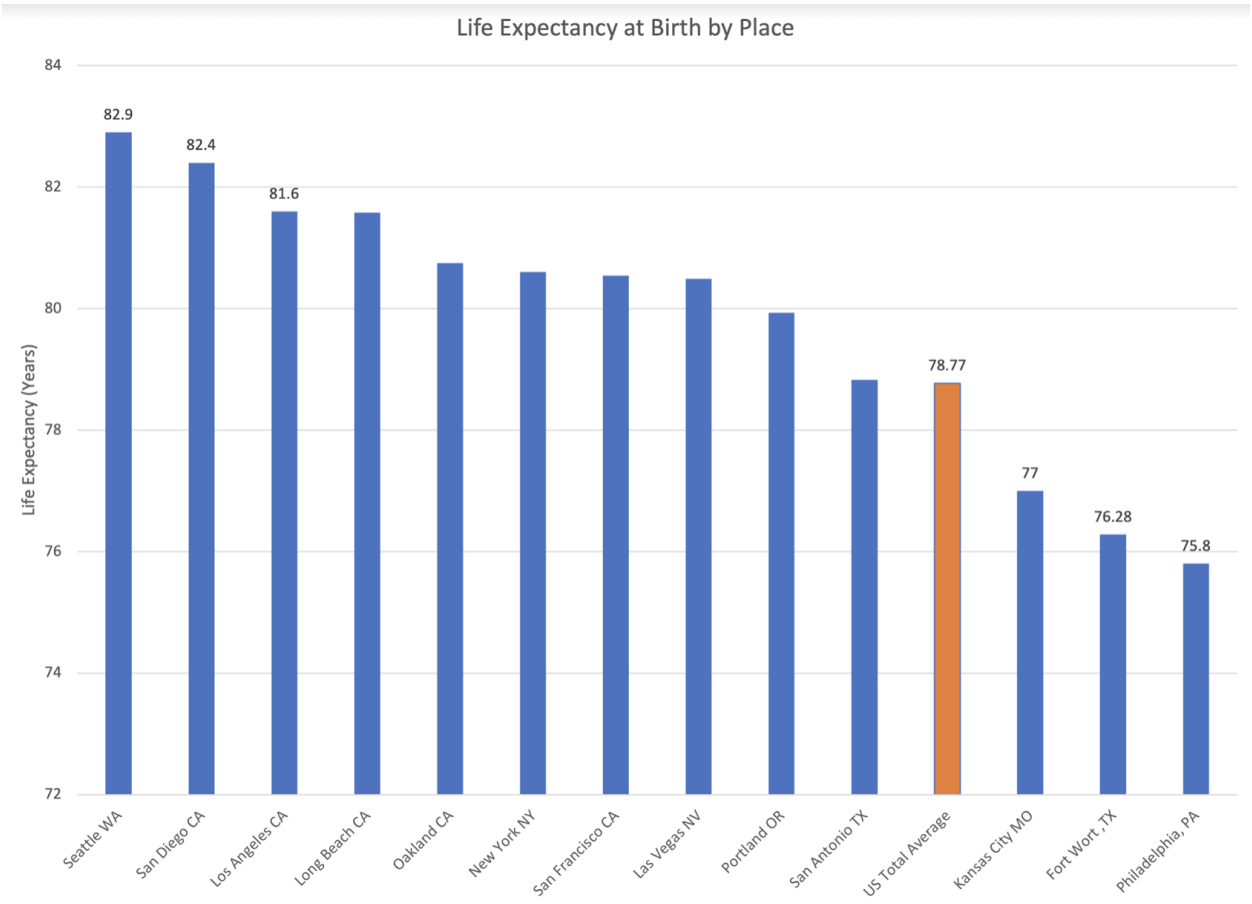
Map based on average of Longitude and average of Latitude. Color shows average of Mortality Rate Per 100,000. The marks are labeled by average of Mortality Rate Per 100,000, City and sum of Mortality Rate Per 100,000. Details are shown for City. The data is filtered on average of Life Expectancy at Birth (Years), which includes everything. The view is filtered on sum of Mortality Rate Per 100,000, which includes everything.



With the indicator of 'Life expectancy at Birth (Years),' our results differed from those obtained by mortality rate in our subset of years 2011-2013. We found that the cities with the highest life expectancy in years were Seattle (82.90 years), San Diego (82.40 years), and Los Angeles (81.60 years). The cities with the lowest life expectancy in years were Kansas City (77.00 years), Fort Worth (76.28 years), and Philadelphia (75.80 years). Based on the dataset, the US average life expectancy was 78.76 years for 2011-2013, which showed that the top three cities were significantly above average and the bottom three cities were slightly below average.

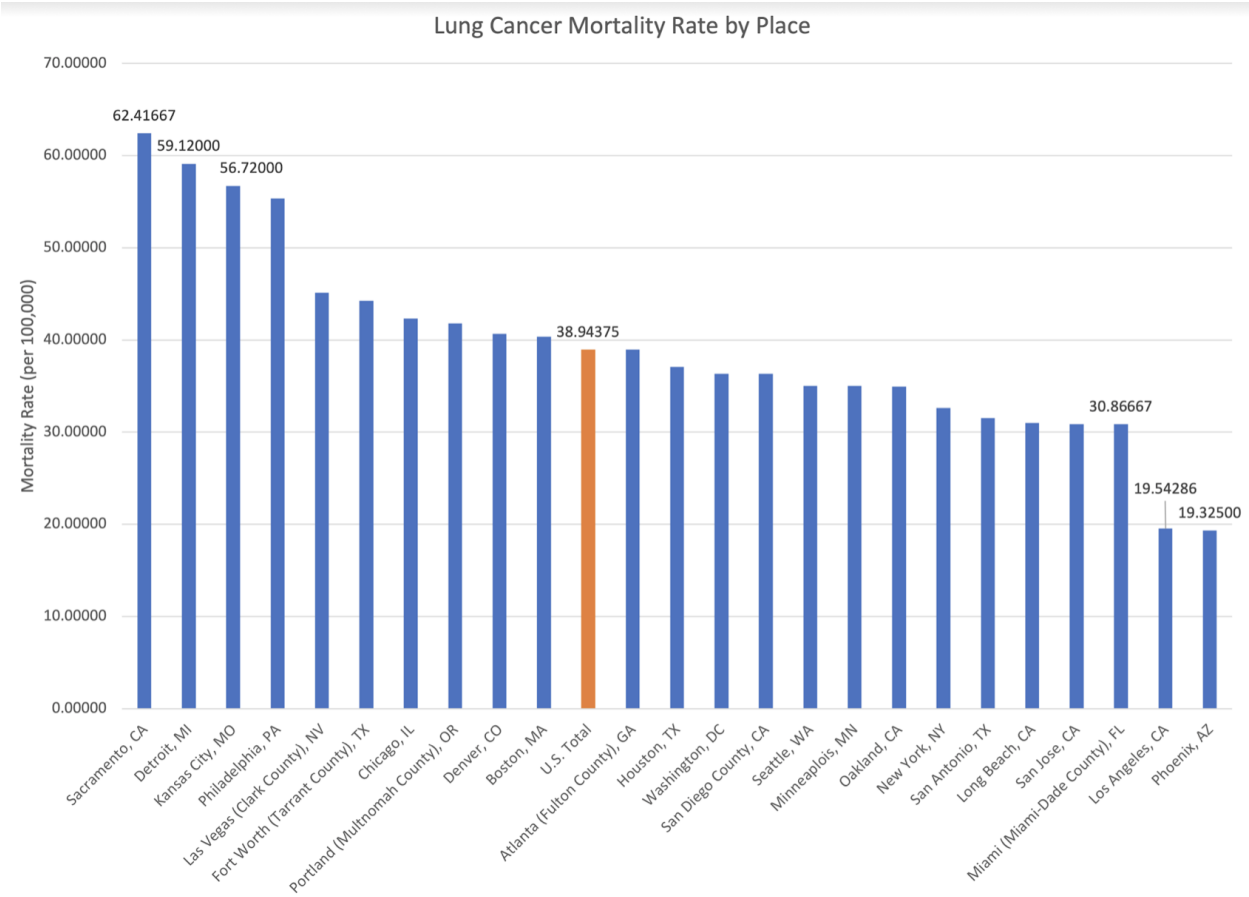


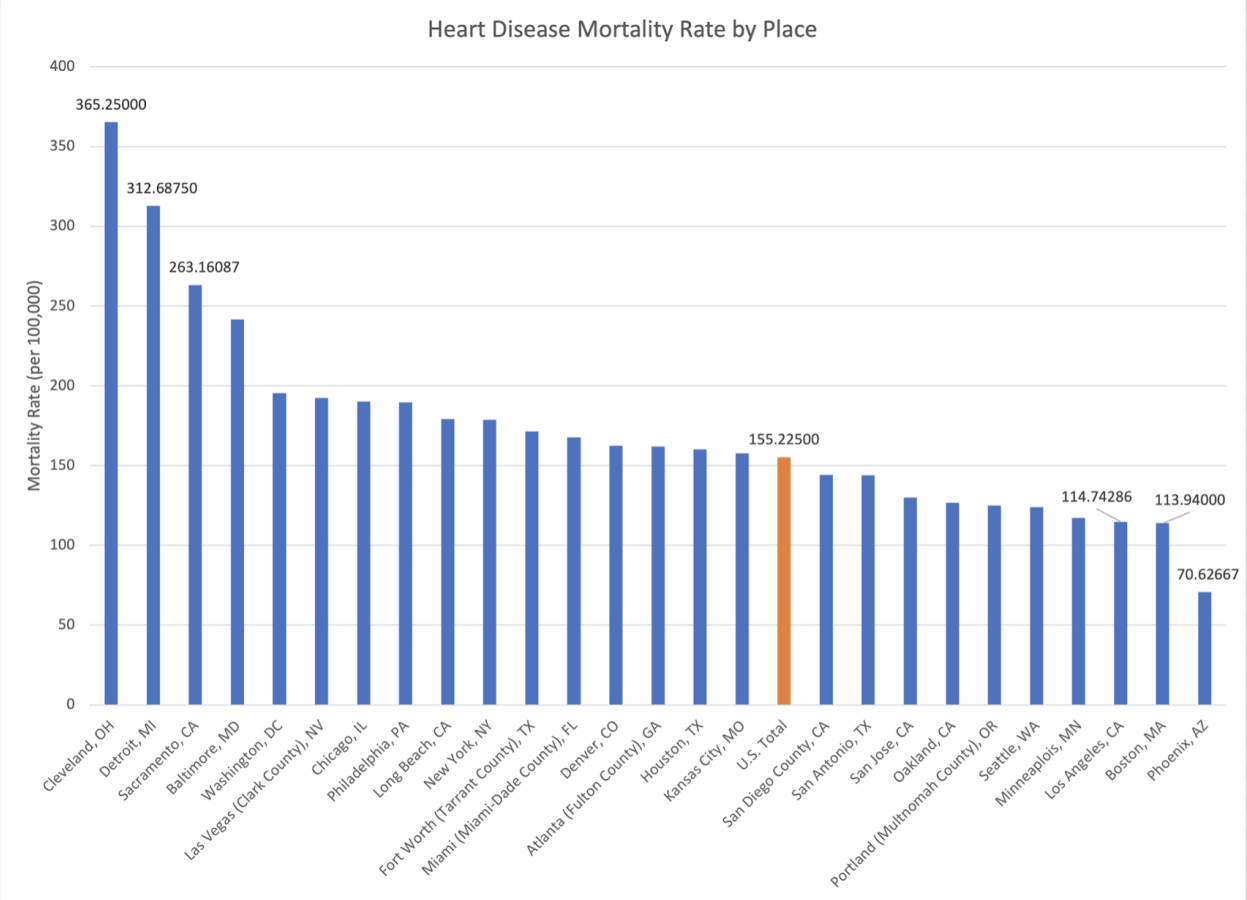
Map based on average of Longitude and average of Latitude. Color shows sum of Life Expectancy at Birth (Years). The marks are labeled by average of Life Expectancy at Birth (Years) and City. Details are shown for City. The view is filtered on average of Life Expectancy at Birth (Years), which includes everything.



Upon finding the cities' averages for both mortality rate and life expectancy, we were curious as to what numbers were a significant factor/influence on these values. To investigate these results, we did a deeper dive into the data to find similar indicators whose values would paint a better picture of the overall landscape of these cities' health. Upon further research, we determined that the significant indicators were all-type cancer, lung cancer, heart disease, pneumonia & influenza, and diabetes mortality rates. We subsetted the data based on these indicators, then reran the numbers to see how they compared to the overall mortality rate and life expectancy. We found cities that were in both the overall mortality rate and the specified mortality rates. From this data, we found that, along with being leaders in overall mortality rate, Detroit and Kansas City were amidst the top three in lung cancer mortality rates, and Cleveland

and Detroit were leaders in heart disease mortality rates. Conversely, Phoenix and Los Angeles were consistently on the low end of the spectrum of these specific mortality rates.



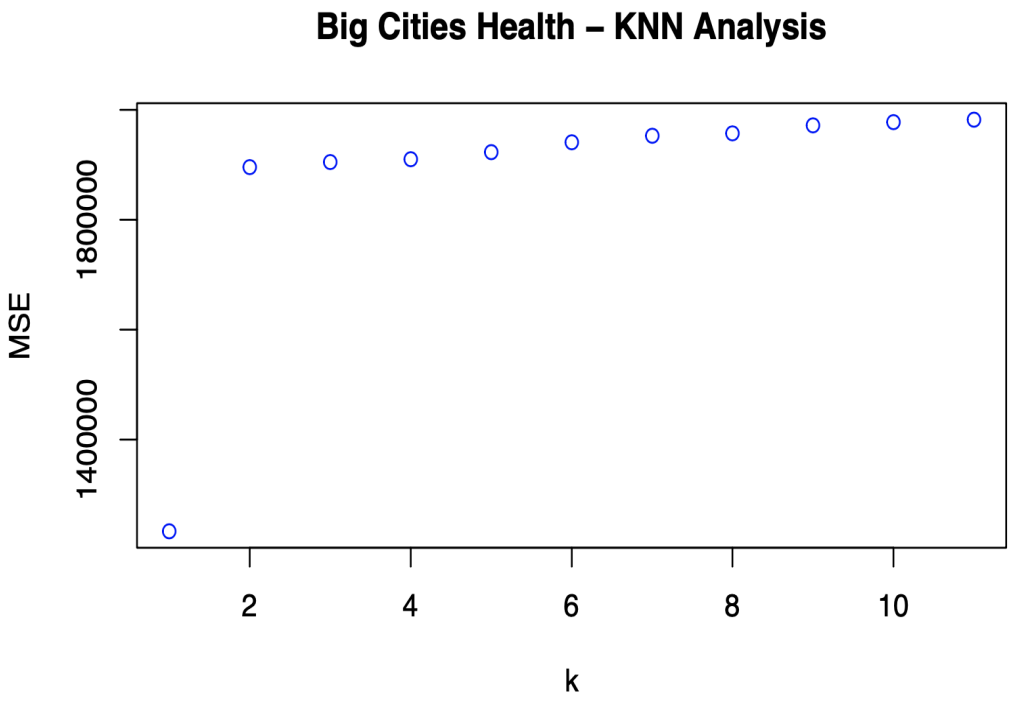


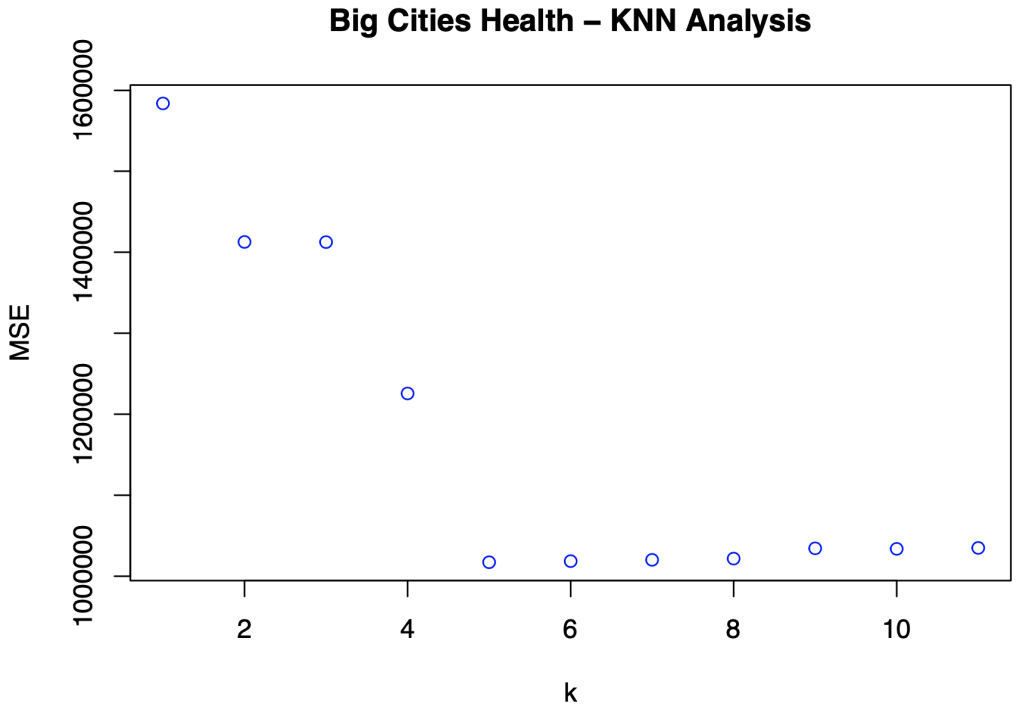
Evaluation

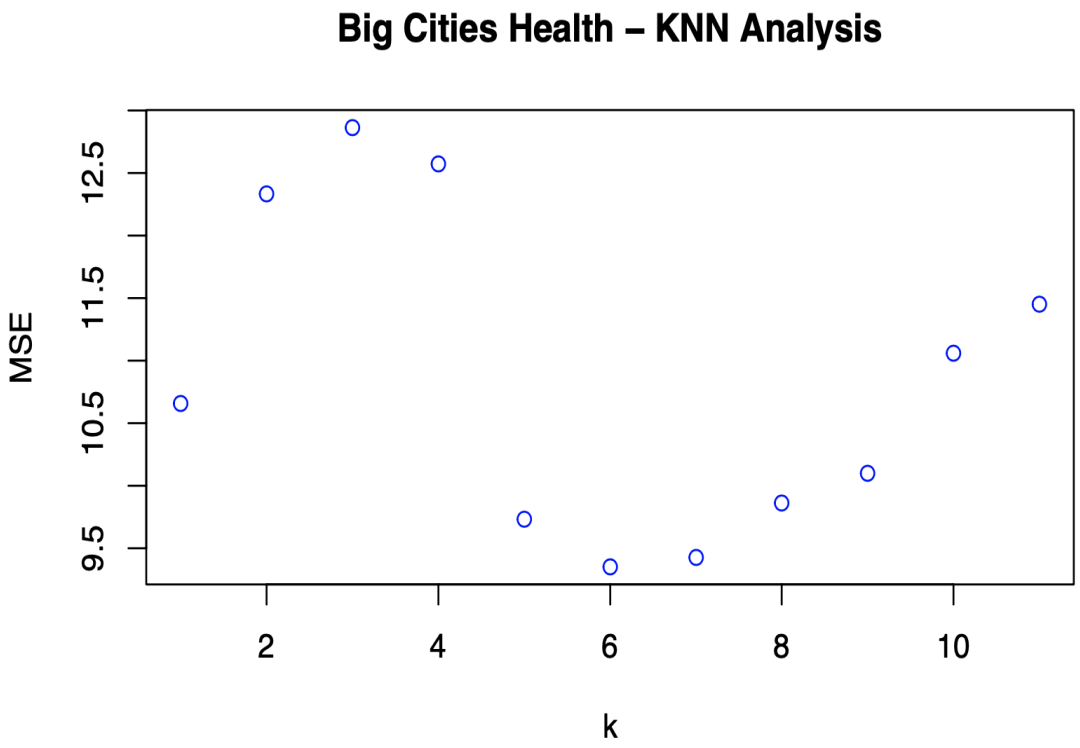
K Nearest Neighbors: For our analysis, we used a K Nearest Neighbors (KNN) model to evaluate the Big Cities Health dataset. The KNN method makes a prediction for an observation based on the average of points among a specified number of K closest observations in the training set. Because this is a non-parametric model, KNN does not assume that the data follow some structure to be discovered. Some advantages of KNN include its simplicity of explanation and implementation and its absence of data assumptions. Some disadvantages to this method include sensitivity to noisy and missing data, it can be very slow and cumbersome, and it requires a large amount of data to use numerous predictors. KNN can be more stable with a larger value of K.

We created a KNN regression model to decide the average target value rather than doing classification, where we would have gotten a majority vote or proportions as probability predictions. To ensure we did not include irrelevant features, we considered variable selection carefully. We did not include Indicator Category due to its varying ways of calculating rates, Source because it does not affect this regression, and Notes and Methods because they were both blank columns. It is often a good idea with KNN to remove any highly correlated variables, but we did not have any with high correlation to remove specifically.

We experienced firsthand how cumbersome a KNN model can be with the extremely long execution time in R. The code to test 11 different K values and calculate the Mean Squared Error (MSE) took over 4 hours to run. Once it was complete, we plotted the results and used the Elbow Method to establish which K value would be most successful for estimating an average target value.




Big Cities Health – KNN Analysis
MSE
1000000
1200000
1400000
1600000
2
4
6
8
10
k


Big Cities Health – KNN Analysis
MSE
9.5
10.5
11.5
12.5
2
4
6
8
10
k

The outputs above resulted from our KNN analysis. The first plot is an output of a KNN regression using the entirety of our subsetted data for 2011 to 2013. The plot showed that using the Elbow Method, the best K for analysis would be one. It also showed an extremely high Mean Squared Error. It took an extremely long time to produce this plot, over 4 hours. We knew that KNN could be slow, but this taught us many important lessons for the future about running KNN and what to look for in an output. The next iteration of our KNN is shown in the second plot. This plot was created using the same subsetted data, but we used feature selection to remove three columns that were not beneficial to our analysis. This iteration ran much faster and gave a more desirable plot with K equalling five and the MSE decreasing. We were still concerned with how high the MSE was and realized we needed to filter the data further to only the indicators that were the correct value calculated by 'Life Expectancy at Birth, ' allowing us to remove all of the outliers created by other indicators. The last plot resulted from our final KNN regression model with the smaller dataset giving us a K of seven and an MSE of 9.42665, which shows the model to be much more accurate than before.

Decision Trees & Random Forests: We began our second analysis method by building a decision tree. We chose this method because it has high interpretability and versatility, and there is less data preparation than in other models. Unfortunately, decision trees have disadvantages such as overfitting, optimization, feature reduction, and data resampling. The main challenge we faced during this process was that some of our columns, such as "Indicator," would disappear when we tried to use a test data set to plot a decision tree. We addressed this issue by learning to create and plot subsets within the decision tree to avoid future problems.

Due to the drawbacks of decision trees, we used our decision tree to construct a random forest, a more robust modeling technique with much higher accuracy than a single tree. Their

algorithm avoids and prevents overfitting by using multiple trees. Random forests are their ability to limit overfitting without substantially increasing error due to bias. Random forests can reduce variance through training on different sample datasets within Big City Health. We began our decision tree by splitting the data into training and test data sets, with 90% in the training set. Once we had our decision tree, we ran a random forest on the top years (2011-2013) data predicting the Value column.

The process of analysis through decision trees and the random forest was instructive. It demonstrated how to build random forests on imperfect real-world data sets, even though our random forest model did not produce the most significant results.

Given our dataset, we wanted to explore life expectancy by city, but we acknowledge that other factors greatly influence life expectancy by city that wasn't included in our data. Ideally, we would find data on factors such as poverty rates, air quality, education level, etc., and incorporate this information into our current models to create more accurate and well-rounded models.

Deployment

We used our findings to determine the life expectancy rankings of cities for effective deployment. A city would need to undergo considerable changes over a long period of time in order to increase its average life expectancy. Since the data we are focusing on is from 2011 to 2013, some of the statistics and top cities may have changed in the nine years since we collected the data.

It is important to note that our data set focuses on a limited number of indicators, which exclude additional factors that may contribute to a city's life expectancy. To create a more holistic understanding of which cities have the highest and lowest life expectancies, we need to incorporate data sources that paint a more complete picture. The economy of each city or the

cultural norms governing the city's inhabitants are some potential factors that might affect the data but are not included in our dataset. These factors could increase unhealthy habits in different aspects of one's life based on how others around them are acting. Our data set focuses on a limited number of variables, which exclude many other factors that contribute to a city's potential life expectancy.

In order to properly convey the purpose of our study, our team feels that it is necessary to discuss the moral propriety of designating a city as the "least healthy" place to live. A city's designation as having the "lowest average life expectancy" or "highest mortality" may not automatically signify that its residents have a very unhealthy environment; rather, it may be a sign of widespread poverty or other unresolved problems. Publishing health-based city rankings constructed from limited data may also be unethical because it is not as accurate as possible and could damage tourism and the regional economy.

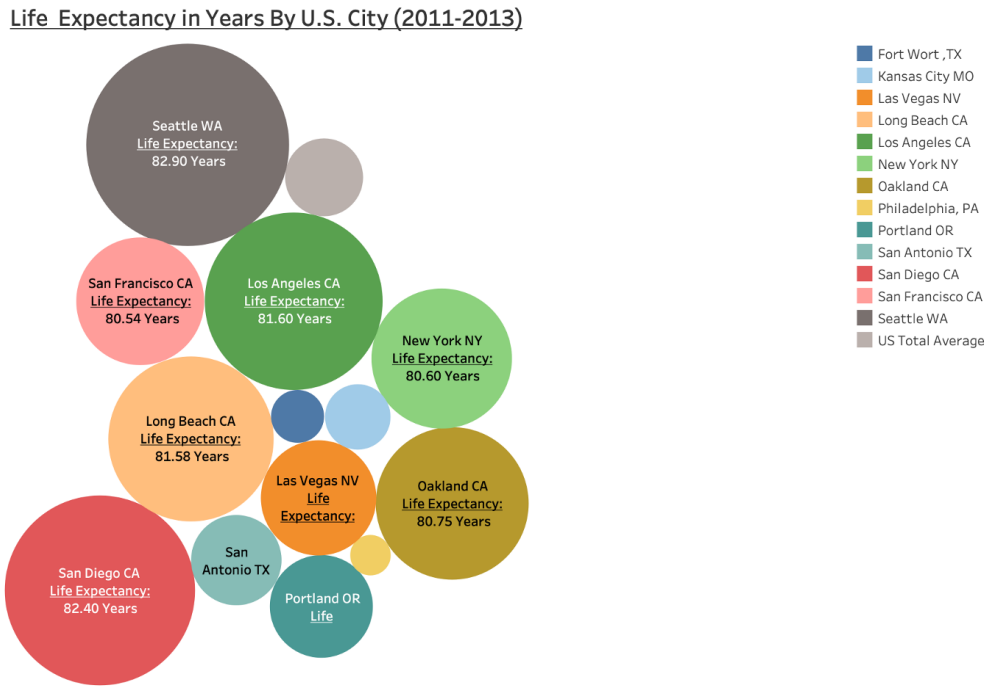
Future Steps

The Big City Health data set and visualizations provide an overview of U.S. life expectancy and a starting point for further analysis. Our next step would be to incorporate outside data sources to improve our models, identify influential variables, and improve our overall understanding of both life expectancy and mortality rates. Potential supplemental data sources could include the U.S. Census Bureau, the Economic Census, and the Department of Health and Environmental Reports, which address broader determinants of health, including food safety, housing standards, health and safety, air quality, noise, and environmental issues, all of which make fundamental contributions to the overall level of public health. We understand that the death rates and life expectancy in each city are impacted by homelessness, but in our research, we could not discover an accurate record of homelessness. We found that the most

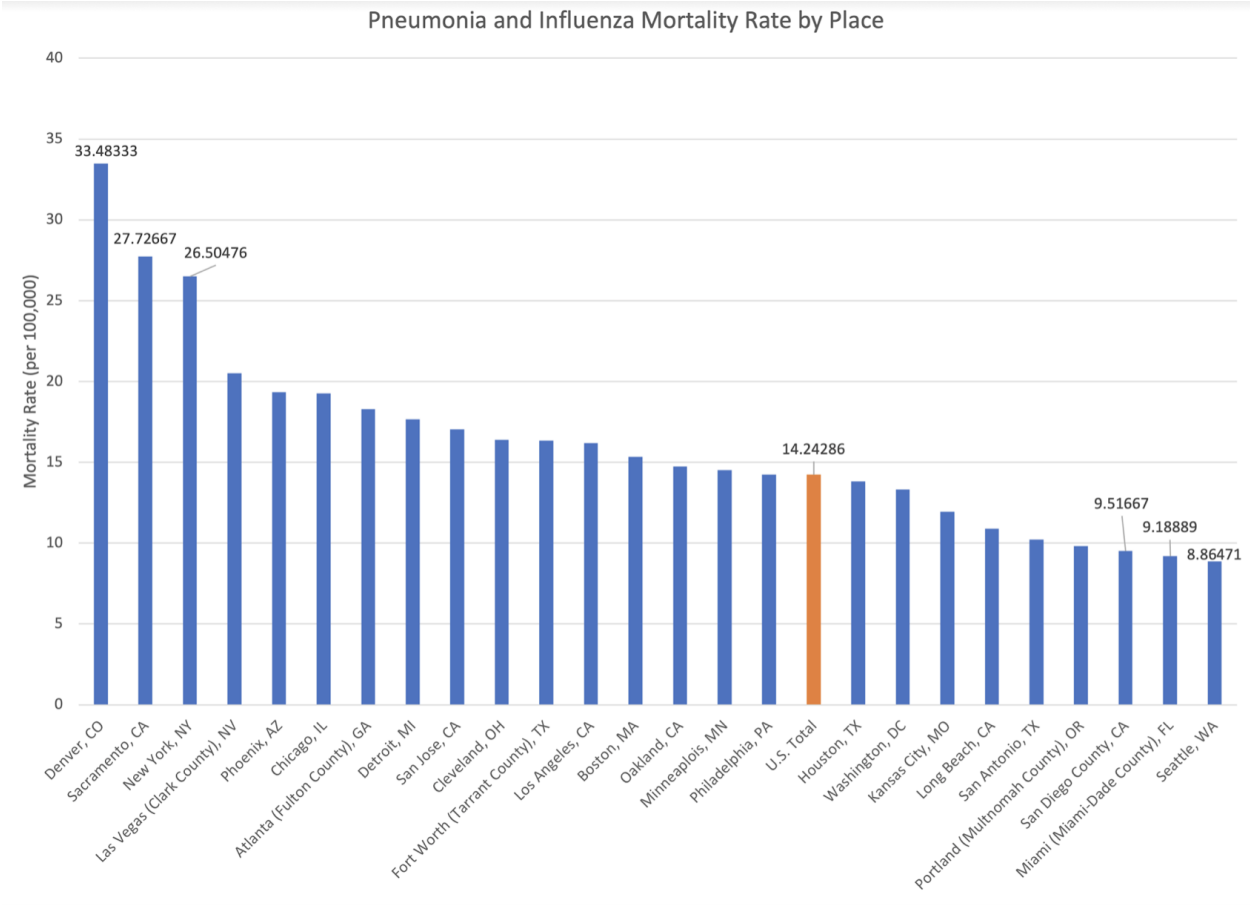
common way of measuring homelessness is through 'point in time' estimates. Additionally, if we wanted to look further into homelessness, we would use cross-sectional analysis. In summary, by incorporating additional data sources, we can build models with much higher accuracy and comprehensively understand current life expectancy and mortality rates in large U.S. cities.

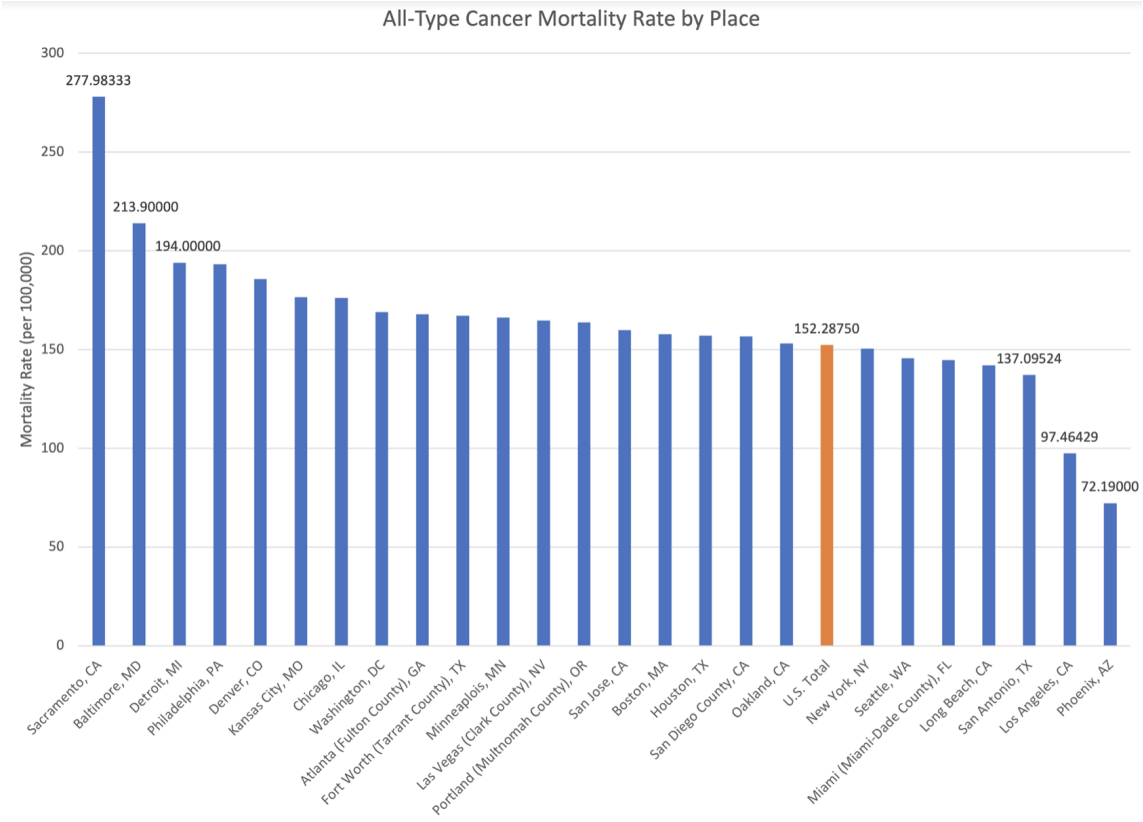
Conclusion

During the course of this project, we strengthened our capacity to prepare, purify, and examine a messy, real-world data set. In order to create usable subsets and learn how to best divide the data set into smaller groups, we had to comprehend the complete set of data. KNN, decision trees, and random forests models were employed in the analysis. Overall, through this project, we were able to hone our already existing skill sets and develop a number of new ones in the areas of data preparation, analysis, and interpretation. Additionally, because our group is passionate about health, our big city health project gave us the chance to explore new ideas and methods while honing our analytical skills.


Life Expectancy in Years By U.S. City (2011-2013)
Seattle WA
Life Expectancy:
82.90 Years
San Francisco CA
Life Expectancy:
80.54 Years
Los Angeles CA
Life Expectancy:
81.60 Years
New York NY
Life Expectancy:
80.60 Years
Long Beach CA
Life Expectancy:
81.58 Years
Las Vegas NV
Life
Expectancy:
Oakland CA
Life Expectancy:
80.75 Years
San Diego CA
Life Expectancy:
82.40 Years
San
Antonio TX
Portland OR
Life
Fort Wort ,TX
Kansas City MO
Las Vegas NV
Long Beach CA
Los Angeles CA
New York NY
Oakland CA
Philadelphia, PA
Portland OR
San Antonio TX
San Diego CA
San Francisco CA
Seattle WA
US Total Average

Appendix


Pneumonia and Influenza Mortality Rate by Place
Mortality Rate (per 100,000)
40
35
30
25
20
15
10
5
0
33.48333
27.72667
26.50476
14.24286
9.51667
9.18889
8.86471
Denver, CO
Sacramento, CA
New York, NY
Las Vegas (Clark County), NV
Phoenix, AZ
Chicago, IL
Atlanta (Fulton County), GA
Detroit, MI
San Jose, CA
Cleveland, OH
Fort Worth (Tarrant County), TX
Los Angeles, CA
Boston, MA
Oakland, CA
Minneapolis, MN
Philadelphia, PA
U.S. Total
Houston, TX
Washington, DC
Kansas City, MO
Long Beach, CA
San Antonio, TX
Portland (Multnomah County), OR
San Diego County, CA
Miami (Miami-Dade County), FL
Seattle, WA


All-Type Cancer Mortality Rate by Place
Mortality Rate (per 100,000)
300
250
200
150
100
50
0
277.98333
213.90000
194.00000
152.28750
137.09524
97.46429
72.19000
Sacramento, CA
Baltimore, MD
Detroit, MI
Philadelphia, PA
Denver, CO
Kansas City, MO
Chicago, IL
Washington, DC
Atlanta (Fulton County), GA
Fort Worth (Tarrant County), TX
Minneaplois, MN
Las Vegas (Clark County), NV
Portland (Multnomah County), OR
San Jose, CA
Boston, MA
Houston, TX
San Diego County, CA
Oakland, CA
U.S. Total
New York, NY
Seattle, WA
Miami (Miami-Dade County), FL
Long Beach, CA
San Antonio, TX
Los Angeles, CA
Phoenix, AZ