ANLY 501 PROJECT PART 2

Where Should You Live For Your Health?

Melissa Collier, Norman Hong, Jingjing Lin
Yeqing Liu, Xiner Ning, Arshia Singh

Introduction to Project:

This project aims to investigate how an individual's choice in residence may be impacting their health. To do this, cancer rates will be the main parameter to determine the overall health of each county in the United States. Cancer is currently the second leading cause of death in the nation, the causes of which have been linked to many poor health choices, including environmental factors, making it a very good indicator of the health of populations. A dataset from the NIH that includes cancer type, demographic information, and location for each across the United States was used for this analysis. Two additional datasets were gathered that help to describe the environmental status of each county in the United States. Air quality datasets were gathered from the Environmental Protection Agency that shows air quality indexes for common air pollutants such as NO2, O3, and CO. Additionally, a water quality dataset was gathered from the Center for Disease Control (CDC) to observe the heavy metal contaminants in the drinking water sources for each county. These three datasets can describe a basic picture of each US County's environment, and how this environment might be correlated with cancer rates.

Section 1.0: Finalizing Data Cleaning

In the first part of the project, a data cleaning metric score was described for each data set. The issues with the data was described in part one, and in this project the data was completely cleaned for analysis. Additionally, outliers in numerical data for each data set were identified, and the outlying records were kept or removed based on their importance to the analysis. To check for missing values, a built-in statistical function was created in python to identify any missing column, e.g isna(). Each dataset dealt with missing values differently and the results are below.

1.1 Cancer Dataset Cleaning

*Cleaning:*The data cleaning score for the cancer data was 98.39%. Symbols were converted to nan values, the index column was removed, and the "County" column was broken into four separate columns: State, County, SEER Registry Number, and NPCR Registry Number. Additionally, all the numerical columns that were imported as strings were removed and converted to numeric.

Checking for Outliers: The histogram below (Figure 1) 'Average Annual Count' shows a few hundred observations having average annual count of more than 2500, which are possibly outliers. In order to confirm whether these observations are outliers, a subset from cancer data with an average annual count of more than 2500 was created. A large number of such observations are from California, New York and so on. All other variables were normal. Therefore, these variables were not removed as they could contain valuable information in analysis.

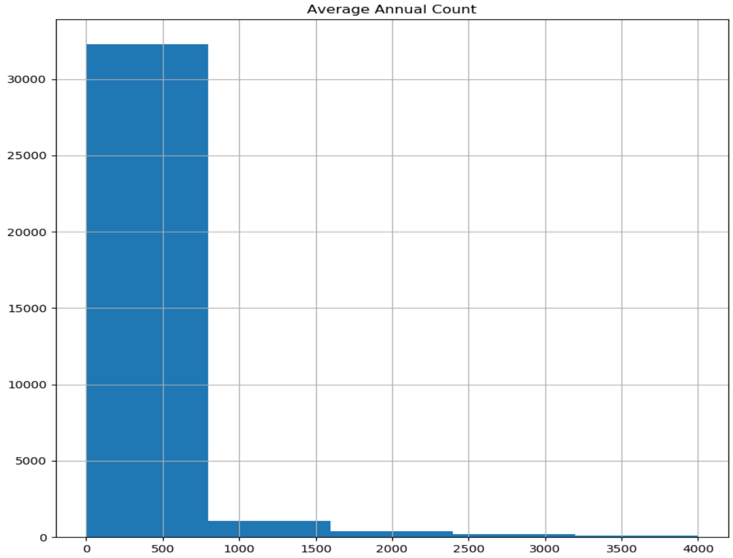


Figure 1: Histogram of Average Annual Count in cancer dataset. Average Annual Count measures the average number of cancer cases from year 2011-2015.

1.2 Air Dataset Cleaning

Cleaning: The data cleaning metric score for the Air Quality Data was 98.23%. In order to join this dataset with other datasets, the FIPS county and state codes were added to this data frame. To do this, each county string was cleaned to match with the county string in the FIPS county code data frame so that the county and state codes could be easily added. There were no null values that needed to be removed, although outliers were present and discussed below.

Next, the numerical values were normalized into proportions. Since the Air Quality Index was not recorded for the same amount of days for every county, the “Good”, “Moderate”, “Unclean”, etc days columns needed to be divided by the number of days AQI was recorded (“Days with AQI”), so that these records could be compared to each other. This also applied to the “Days CO2”, “Days O3” etc. columns, in which the pollutant mentioned was the primary pollutant for the number of days specified. These columns were added to the data frame as “Previouscolumnname_Norm”. For example: “Good Days_Norm”.

Checking for Outliers: When checking for outliers in the numerical data of the air dataset, many outliers were found by using boxplots. The 'Good Days_Norm', 'Moderate Days_Norm', ‘Unhealthy for Sensitive Groups Days_Norm’, Unhealthy Days_Norm”, ‘Very Unhealthy Days_Norm’, ‘Hazardous Days_Norm’ and 'Max AQI' all showed outliers, but for each column except ‘Max AQI” these outliers were left and deemed “extremes” for each column. However, as the “Max AQI columns contained values above 500, which is the max AQI score, these records were completely removed from the data frame (Figure 2).

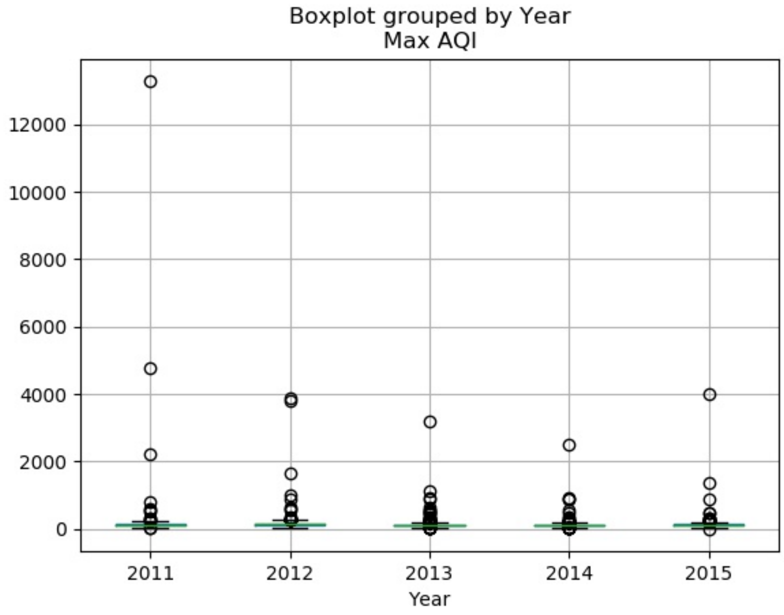


Figure 2: The maximum AQI value for each county for each year. The y axis represents the AQI value for each county (dot), for each year (x axis). AQI measurements do not exceed 500 therefore; values above 500 were errors and removed from the data frame.

1.3 Water Dataset Cleaning

*Cleaning:*The Water quality dataset had a cleaning metric score of 97.46%. The "Location" column was split into a "county" and a "state" column. Additionally, columns such as "geoAbbreviation" and "geoID" were redundant and only one was kept. The water quality mean was also binned into 3 categories: "No pollution detected", "detected but no harm", and "detected and harmful." After cleaning, the "result" data frame was used for analysis and contained only 6 columns with pertinent information for analysis. Overall, there were no missing values found after checking each column with .isna(), value_counts and unique() functions in the original dataset during Part 1 data processing stage.

Checking for Outliers: For the water quality dataset, there is only one numeric column: 'value' (level of Arsenic concentration). According to the CDC's documents, lowering the minimum contaminant load (MCL) from 50 to 10 ppb statistically reduces bladder and lung cancer mortality and morbidity by 37-56 cancers a year in the U.S. (EPA 2001b). Therefore, outliers can not be examined and defined only by applying boxplot function (left in Figure 3). Though there are some points out side of the IQR box, but they should not be considered outliers, because there are many zeros in the dataset and these zeros cause the IQR lie at the bottom. Therefore, these points should be considered normal points within a reasonable range. As a second way to test for outliers, 'Values' were binned into three groups, and a histogram created to show the frequency of each group. The results of the two figures indicate that there were no outliers in the water quality dataset.

Binning strategy: As mentioned before, the level of Arsenic concentration (water quality indicator) was binned into three groups: “No pollution detected”, “detected but no harm”, and “detected and harmful” by considering the CDC official guide (right in Figure 3). For examining this new binned column, “sum([the binned column]).isna())” code was applied, showing a result of “0”. Therefore, there were no missing values binned.

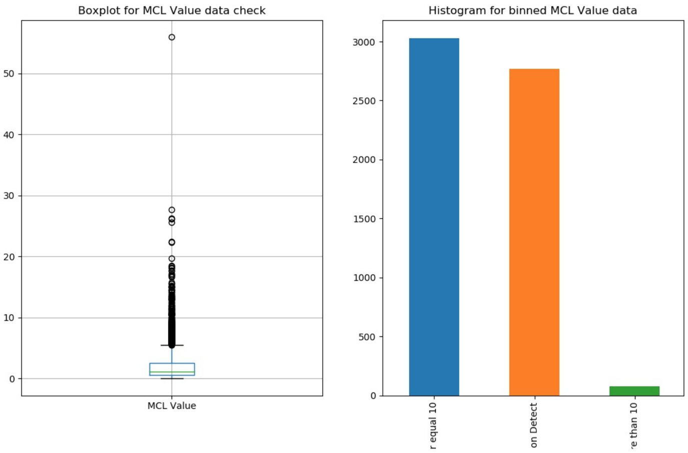


Figure 3: Box plot and histogram check for outliers in the water quality dataset.

Section 2.0 Means, Medians or Modes for 10 attributes

<u>2.1 Air quality dataset attributes</u>

For the air quality dataset the five attributes Max AQI (the maximum air quality value for each county, every year), percentage of “good days” (AQI=0-50) recorded, percentage of “moderate days” (AQI=51-100) recorded, percentage of “unhealthy for sensitive groups days” (AQI=101-150) and percentage of “hazardous days” (AQI=251-300) recorded, were chosen to compute means, medians and standard deviations.

- The average Max AQI Value across all counties for all years is: 117.28 with a standard deviation of: 41.17
  The median is: 112.0
  - This shows that the average “worst” day in America from 2011-2015 was in the “Unhealthy for Sensitive Groups” category. Therefore, people not in sensitive groups are not, on average, subjected to unhealthy air quality.
- The average percentage of Good Days across all counties for all years is: 77.29 %, with a standard deviation of: 16.38 %

The median percentage is: 80.27 %

  - On average, in 2011-2015, America was experiencing 77% of their year with AQIs between 0-50. This means less than ¼ of the year was above a 50 AQI for the average American.

- The average percentage of Moderate Days across all counties for all years is: 20.81 %, with a standard deviation of: 14.04 %

  The median percentage is: 18.63 %

  - In 2011-2015, the average American was experiencing 21% of their year with an AQI between 51-100, which is still not unhealthy.

- The average percentage of Unhealthy for Sensitive Groups Days across all counties for all years is: 1.61 %, with a standard deviation of: 3.39 %.

  The median percentage is: 0.41 %

  - In 2011-2015, the average American was experiencing 1.6% of their year with an air quality between 101-150. On average, sensitive groups in America should be concerned about their air quality for 1.6% of their year (less than 1 week).

- The average percentage of Hazardous Days across all counties for all years is: 0.0014871049716460598 %, with a standard deviation of: 0.03324011895860236 %

  The median percentage is: 0.0 %

  - In 2011-2015, the average American did not experience hazardous air conditions with AQIs between 301-500. Less that 1% of the year was considered hazardous on average in the US.

2.2 Cancer dataset attributes

- The average age adjusted incidence rate across all counties for all years is: 439.88 with a standard deviation of: 57.68

  The median is: 445.6

  - The average cancer rate for all counties between 2011-2015 was 439.88 individuals. The standard deviation is much lower than the following variable (an average annual count of new cases) because this data has been normalized by age and population.

- The The average annual count of new cancer cases across all counties for all years is: 525.64 with a standard deviation of: 1473.51

  The median is: 152.0

  - On average, the number of new cases each year in each county is 526 individuals. There is a huge standard deviation suggesting the need to standardize this by the number of individuals in each county, as is done with the incidence variable.

- The average trend in incidence rates across all counties for all years is: -0.77 %, with a standard deviation of: 4.22 %

The median percentage is: -0.7 %

  - In general, the average trend in cancer rates is decreasing slightly over the years of 2011-2015. Whether or not this is significant with the standard deviation of 4.22%, or correlated with better air or water quality scores, will need to be further analyzed.

2.3 Water quality dataset attributes

- The average Arsenic_content in water across all counties for all years is: 1.88 with a standard deviation of: 1.13
  The median is: 2.39
  - The average American's drinking water is in the "less than or equal to MCL" category, meaning that on average even with a standard deviation of 1.1, the drinking water in America is safe for consumption.

- The mode of water quality level is Less than or equal MCL
  - This statistic furthers the above result, that the average American can expect to consume safe drinking water that has less than the minimum contaminant load of arsenic.

visualization: histogram,correlation

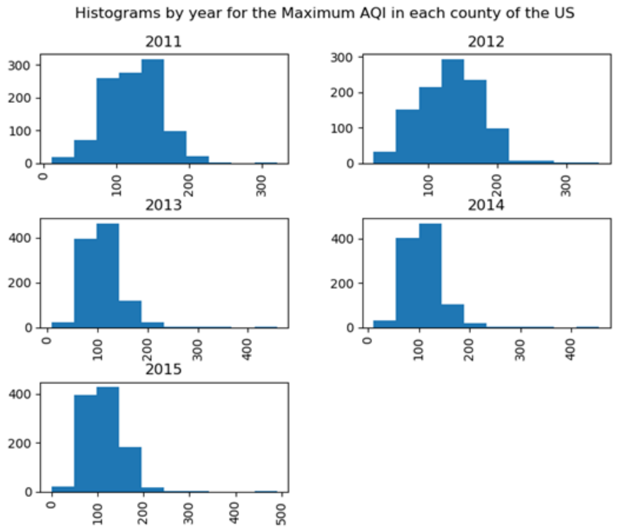
Section 3.0: Histograms and Correlations

3.1 Histograms

One histogram for each dataset was created to get a visual idea of what each dataset shows.

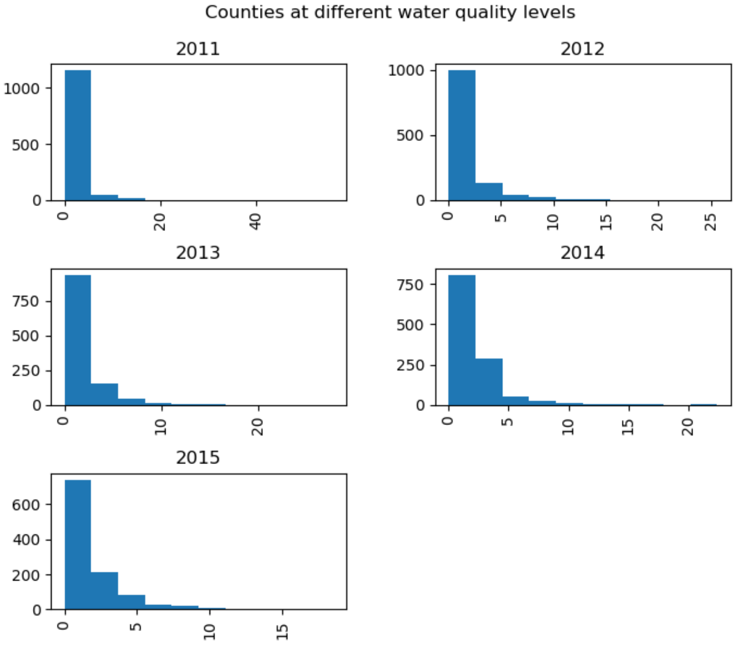
Air Quality

The histogram below shows the Maximum AQI value for each county in the United States, for the years 2011-2015. The histograms show a relatively normal distribution, with the average max AQI value for each county around 150 in 2011 and 2012, reducing to around 100 for 2013, 2014 and 2015. Since the max AQI value is the worst air quality that the county had that year, this suggests that at its worst, the country averages a "Unhealthy for Sensitive Groups" metric.



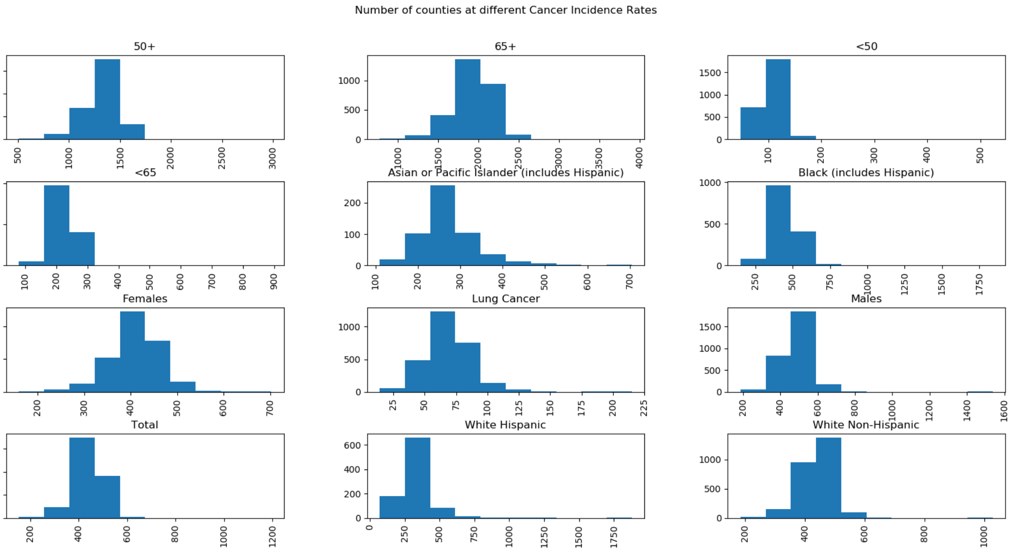
Water Quality

The following histogram shows the frequency of counties at different water quality levels over the years 2011-2015. The distributions are not normal and resemble a Poisson distribution due to high number of zeros, or non-detectable data. The majority of counties have safe drinking water in the US. This is an early indicator that water quality might not be the best predictor of cancer rates in the United States. However, there are still counties above the safe drinking water maximum (10), meaning that analysis should still be conducted to see if these rates might have any indication of health in those particular counties.



Cancer

The histogram below shows the average annual count for cancer divided by different groups based on age and race. Histograms of ages 50+, 65+, <50, <65 indicate that as age increases, the number of counties with less that 500 average annual counts also increases. Race and gender do not seem to have differences among the average frequencies of cancer incidence over all.



3.2 Correlations

One interesting idea for correlation analysis is to see how air quality might be correlated with the water quality of certain counties. Furthermore, certain pollutants might be correlated with water quality scores, as well as AQIs in certain counties which could be of interest in analysis. Therefore, correlations were analyzed for the median AQI for each county, water quality score for each county (Value in Figure 4), and the number of days that Ozone and atmospheric particulate matter were the primary pollutant for each county. Table 1 and Figures 4 and 5 show the correlation results.

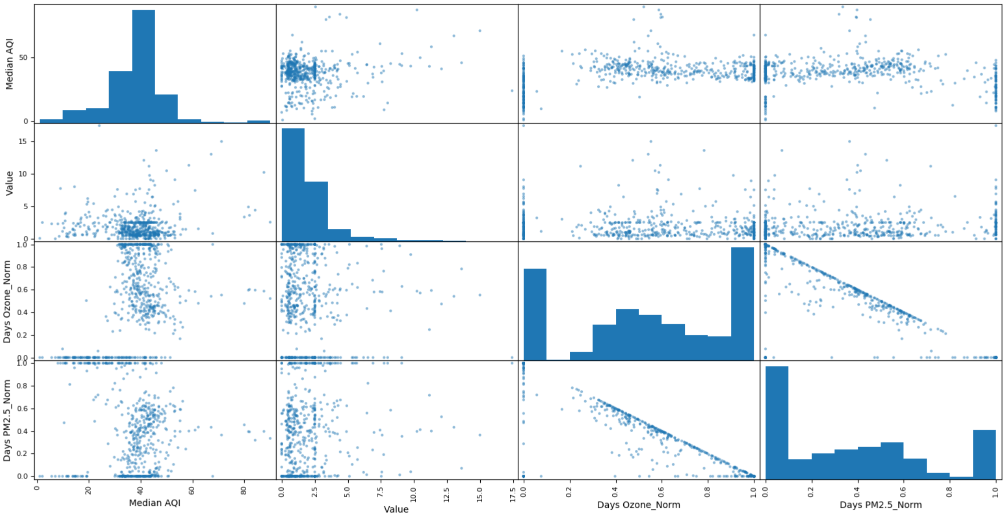


Figure 4: Plots that show correlations between median AQI scores, water quality scores and the number of days in which atmospheric particulate matter and ozone were the primary pollutant in each county.

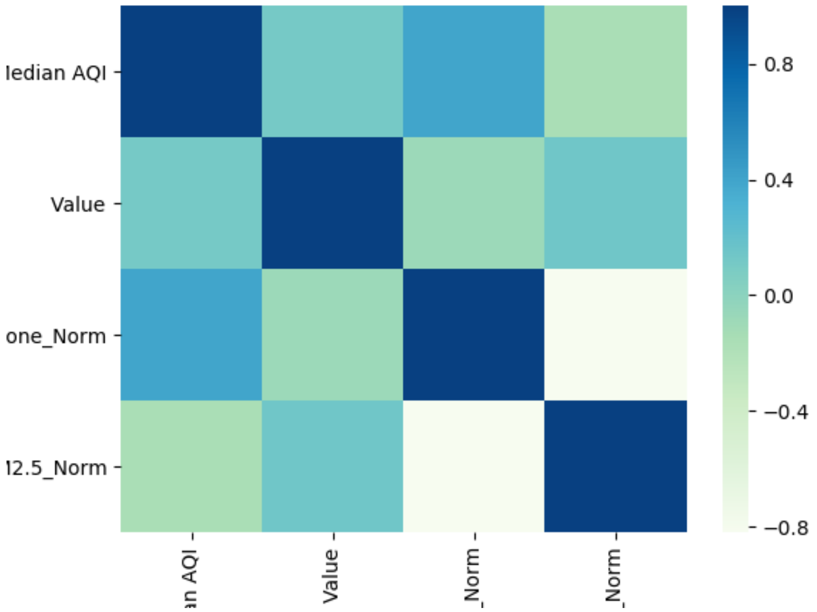


Figure 5: A heat map of the correlation coefficients for each test. Exact values can be found in Table 1.

Table 1: Correlation coefficients for each test ran

Index	Median AQI	Value (Water Quality)	Days Ozone	Days PM2.5
Median AQI	1	0.114	0.397	-0.150
Value (Water Quality)	0.114	1	-0.079	0.143
Days Ozone	0.397	-0.079	1	-0.821
Days PM2.5	-0.150	0.143	-0.821	1

Perhaps the most significant correlation seen here is the slight correlation between Ozone and Median AQI. It appears that as the number of days in which ozone is the primary pollutant increases, the AQI score also increases (gets worse). The same is not true for atmospheric particulate matter suggesting that ozone may play a bigger role in air quality scores, and could be a good pollutant to focus on for future analysis. The strong negative correlation between the number of days atmospheric particulate matter was the primary pollutant and the number of days ozone was the primary pollutant is also an interesting, albeit expected, result. This negative correlation suggests that certain counties AQI scores are being impacted by a specific pollutant, and is an interesting route for further analysis. Perhaps by looking at power plants in each county, a correlation may be found between the type of air pollutant most prevalent in each county and the number or type of plants. It would also be interesting to see how cancer rates are correlated with specific pollutant types.

The other correlations were not significant, suggesting that air quality and water quality are not correlated in the United States.

Section 4.0 Cluster Analysis

A cluster analysis was conducted to explore groups of counties that may share similar overall cancer rates, water qualities and air qualities. The cluster algorithm was applied in a dataset with three variables: the annual counts for all types of cancer, the arsenic concentration value in water, and the standardized time each year that a county had a “high” AQI score (unhealthy for sensitive groups and above). The data spans the years 2011-2015. The dataset had 562 observations after dropping the missing values, with each observation representing a unique county.

4.1 Hierarchical Clustering

Hierarchical clustering was conducted first as it does not require the specification of a number of clusters needed for analysis. Moreover, the dendrogram helps visualize allocations of objects in clusters. In this case, the dendrogram is cut at height 2, which formed 4 clusters. The average silhouette score is 0.4308. In this case, 372 observations are clustered in group 1, 18 in group 2, 24 in group 3, and 148 in group 4. The cluster plot shows that each group has a similar variation on cancer rate (cancer rate axis has lower scale). Group 1 and 4 have similar water quality ranges, although group 1 has better air quality than group 4. Groups 2 and 3 also have similar water quality scores, which are worse than groups 1 and 4. Group 3’s air quality is better than group 2’s. Finally, the plot shows that counties with similar air and water quality are likely to have different cancer rates, while counties with different air and water quality are likely to have same cancer rate. This suggests that there may be other factors other than air and water quality that may be affecting cancer rates in the US. However, the low silhouette score suggests that this clustering method may not be the best comparison tool for this data. Figure 6 shows the dendrogram, and Figure 7 shows a visual interpretation of the clusters.

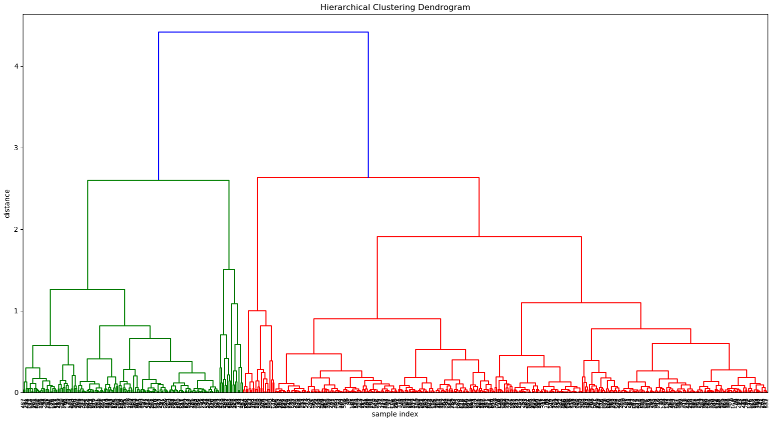


Figure 6: Dendrogram showing the results of hierarchical clustering.

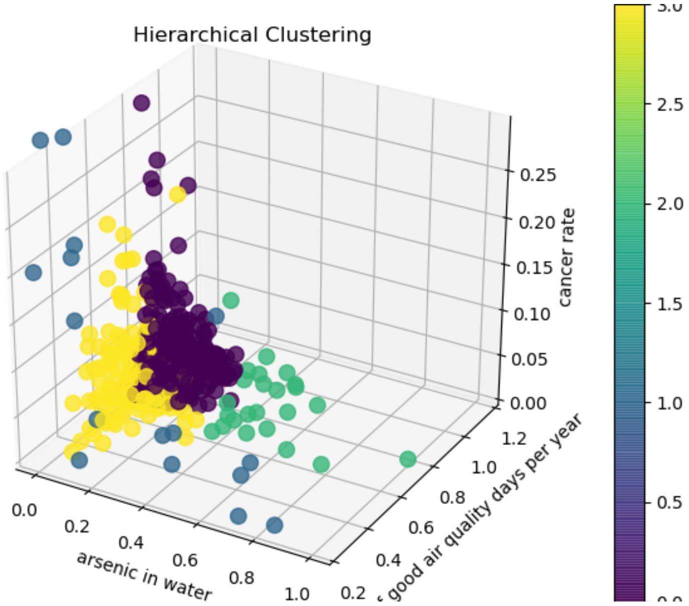


Figure 7: Visual depiction of the 4 clusters created in hierarchical cluster analysis.

4.2 Partition Clustering

The number of clusters (4) from hierarchical clustering was used in KMeans clustering. The average silhouette score is 0.3613. In this method, 23 observations were clustered in group 1, 259 in group 2, 54 in group 3, and 226 in group 4. Unlike the hierarchical clustering, KMeans clustering clustered counties with a relatively low cancer rate together. However, the silhouette score was even lower than the hierarchal, which suggests that KMeans is not a good algorithm to use for this data. Figure 8 visually shows these clusters.

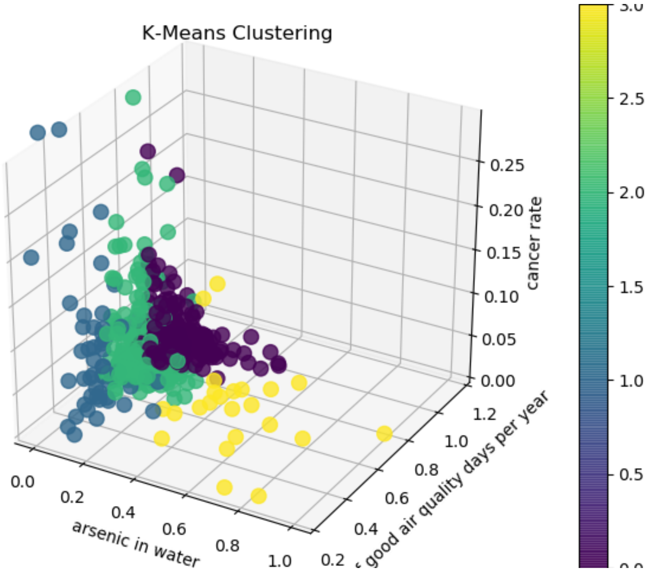


Figure 8: A visual depiction of KMeans cluster analysis

4.3 dbscan Clustering

The average silhouette score of conducting DBSCAN clustering is 0.5022. There are 77 observations clustered in group 1 and 485 in group 2. Although the silhouette score is higher than the previous two clustering methods, DBSCAN in general does not perform well on this dataset. Although this method has indeed detected a pattern that the other two have not detected, there is no apparent meaningful interpretation for this pattern. Figure 9 shows the cluster results.

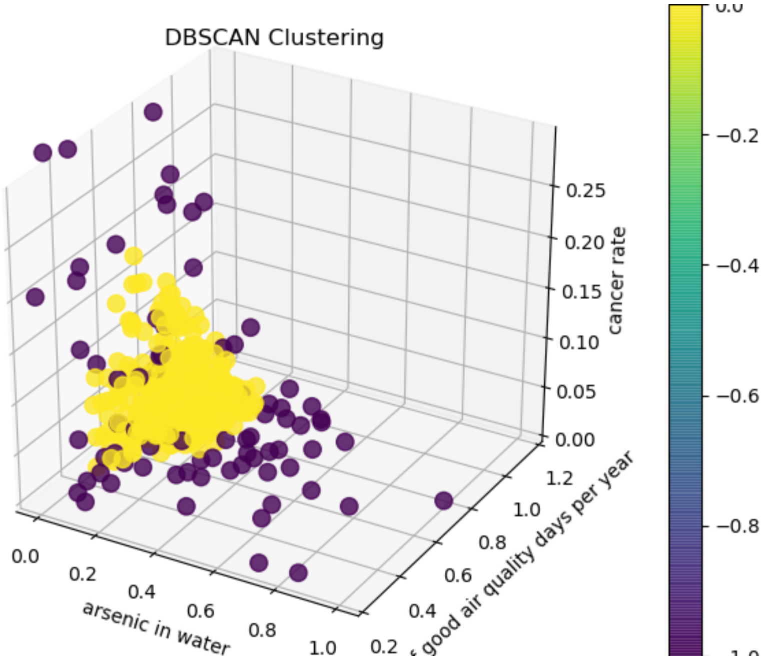


Figure 9: A visual depiction of dbScan cluster analysis.

Section 5.0 Association Rule Mining Analysis

Association rule mining was used to see if “rules” could be established for how often overall cancer rates were associated with certain air quality and water quality statistics. The three factors considered were: the quartile of cancer rates, the 90th percentile of AQIs and the minimum contaminant load in water for each county. For cancer rates, a county could be in the “0-25%”, the “25-50%”, the “50-75%” or the “75-100%” cancer rate group. This says what quartile a county falls in relative to all other counties based on cancer rates normalized by age and population; counties that fall in the first quartile (0-25%) have the lowest relative cancer rates, and those that fall in the fourth (75-100%) have the highest. For air quality scores a county could be in the “Good 90th”, “Moderate 90th”, “Unhealthy for Sensitive Groups 90th”, “Unhealthy 90th”, “Very Unhealthy 90th”, or “Hazardous 90th”. These bins show what the average 90th percentile air quality value is for each county. For water quality, a county could either be “More than MCL” or “Less than or equal to MCL”. This shows whether the water in that county had more than the minimum contaminant load, or less than or equal to the minimum contaminant load for drinking water, on average. The following shows the results of the ARM analysis and the top 11 rules identified using a minimum support of 0.001, a minimum confidence of 0.3, a minimum lift of 2, and a minimum rule set of 2 items. Prior to that ARM was run with supports of 0.0003 and 0.0005 and resulted in 14 and 12 rules, respectively.

Top 11 rules:

Rules 1, 2, & 3:

- Rule: More than MCL --> 25-50%
  Support: 0.012
  Confidence: 0.778
  Lift: 2.824

- Rule: Unhealthy 90th --> 25-50%
  Support: 0.005
  Confidence: 1.0
  Lift: 3.630

- Rule: Unhealthy for Sensitive Groups 90th --> 25-50%
  Support: 0.018
  Confidence: 0.733
  Lift: 2.662

These three rules demonstrate that counties with a 25-50% cancer rate are associated with high water contamination, and unhealthy AQIS (151-250). This being the second quartile of cancer rates suggests that there may be some sort of correlation between cancer rates and water quality, or cancer rates and AQIs and these rules should be analyzed further, perhaps by looking at the incomes of people with cancer in each county. Rule 3 is the most frequent, with the most support.

Rules 4 & 5

- Rule: Unhealthy 90th --> More than MCL
  Support: 0.002
  Confidence: 0.333
  Lift: 22.185

- Rule: More than MCL --> Unhealthy for Sensitive Groups 90th
  Support: 0.005
  Confidence: 0.333
  Lift: 13.311

These rules are not surprising results, and demonstrate that counties with high water contamination levels are also associated with unhealthy AQI scores (151-250). This is an exciting result that suggests that the source of water and air pollution might be related, and it may be possible to correlate poor water and air quality with the presence of structures such as power plants or landfills.

Rules 6 & 7

- Rule: Less than or equal MCL, Unhealthy 90th --> 25-50%
  Support: 0.012
  Confidence: 0.778
  Lift: 2.824

- Rule: Less than or equal MCL, Unhealthy for Sensitive Groups 90th --> 25-50%
  Support: 0.003
  Confidence: 1.0
  Lift: 3.630

These two rules demonstrate that counties with good water quality and poor air quality score are associated with 25-50% cancer rates. This suggests that cancer might be more directly correlated with air quality in these counties, rather than water quality. Perhaps, having both poor air quality and water quality will demonstrate higher cancer rates, and is a potential idea for further analysis. Rule 7, along with Rule 2, both have the highest confidence (1.0).

Rule 8

- Rule: Moderate 90th More than MCL --> 25-50%
  Support: 0.005
  Confidence: 0.6
  Lift: 2.178

This rule demonstrates that counties with moderate air quality and poor water quality are associated with 25-50% cancer rates. This suggests that water quality might also have an effect of cancer rates in certain counties, contradictory to the above two rules. In order to better understand this, looking at the county demographics or cancer type may help to explain some of these contradictions. For example, wealthy people might drink more bottled water, but are still subjected to the same air quality. Perhaps the the counties in this rule are less wealthy.

Rule 9 & 10

- Rule: 25-50% Unhealthy 90th --> More than MCL
  Support: 0.002
  Confidence: 0.333
  Lift: 22.185

- Rule: 25-50% More than MCL --> Unhealthy for Sensitive Groups 90th
  Support: 0.005
  Confidence: 0.429
  Lift: 17.114

These rules demonstrate that counties with 25-50% cancer rates and poor air quality are associated with poor water quality. These rules support rules 4 & 5 which suggests that counties with poor water quality, also have poor air quality, and should be further analyzed.

Rule 11

- Rule: 75-100% Less than or equal MCL --> Good 90th
  Support: 0.005
  Confidence: 0.5
  Lift: 2.340

This is perhaps the most surprising rule, as it demonstrates that counties with the highest cancer rates and good water quality are associated also with good air quality. Therefore, the cancer rates in these counties might not be explained by environmental factors and it is important to take things such as smoking rates, obesity, family histories, socioeconomic statuses or even cancer type into account when evaluating the counties with high cancer rates.

Section 6.0 Hypothesis Testing

The results from clustering analysis and Association Rule Mining provided mixed results on the impacts that air quality and water quality have on cancer rates in the United States. Therefore, hypothesis tests were conducted to find significant relationships between these factors, in order to understand what direction the analysis should go. The hypothesis tested were as follows:

1) As the water quality scores increase (poor water quality), cancer rates will also increase in each county.
2) The air and water quality scores in each county can be used as predictors for that specific county's cancer rate (the average number of individuals affected with cancer in each county over a five year period).
3) The air and water quality scores in each county can be used as predictors for that specific county's cancer trend (rising or falling cancer rates each year).

6.1 Hypothesis 1

As the water quality scores increase (poor water quality), cancer rates will also increase in each county.

To test this hypothesis, both a t-test and a linear regression were performed on the water and cancer datasets. A t-test was chosen as an analysis tool for this hypothesis in order to understand the difference between two populations (no arsenic in water and arsenic present in water). Additionally, a linear regression was conducted following the use of the t-test in order to see the relationship between one predictor variable (water quality) on cancer rates.

The t-test checked if there was a significant difference in cancer incidence between two levels of contamination in water. Cancer rates in counties where arsenic was "undetectable" were compared to cancer rates in counties where arsenic was detectable. The H0 was that counties with arsenic at 'non-detected' level have the same cancer rate with arsenic detected. The results of the t test comparison show p=0.000283 meaning that the cancer rates were significantly different in counties where arsenic was detected in water versus not detected.

Linear regression was used to find how water quality affects cancer incidence. Considering cancer incidence as dependent variable 'y' and arsenic content in water as independent variable 'x', the linear regression was set to be: y= a*x +b. The output was b=456.85 and a=-4.91, with a formula of y= -4.91*x +456.85, shown in Figure 10. This regression shows that as water quality gets worse, the cancer rates actually decrease, which was opposite from the initial hypothesis. There could be many explanations for this, including socioeconomic factors that may be confounding these results. Counties with poor water quality may be drinking more bottled water, thus being less subjected to the negative effects of arsenic. The wealth of an individual in these counties may be a better predictor of cancer rates, since wealthier people can afford better water filters and bottled water. Therefore, census data should be pulled in to run this comparison. Overall the hypothesis that as water quality decreases, cancer rates will increase is not supported, although it brings along ideas for exciting future analysis, and helps to explain some of the mixed results from the ARM and clustering analysis.

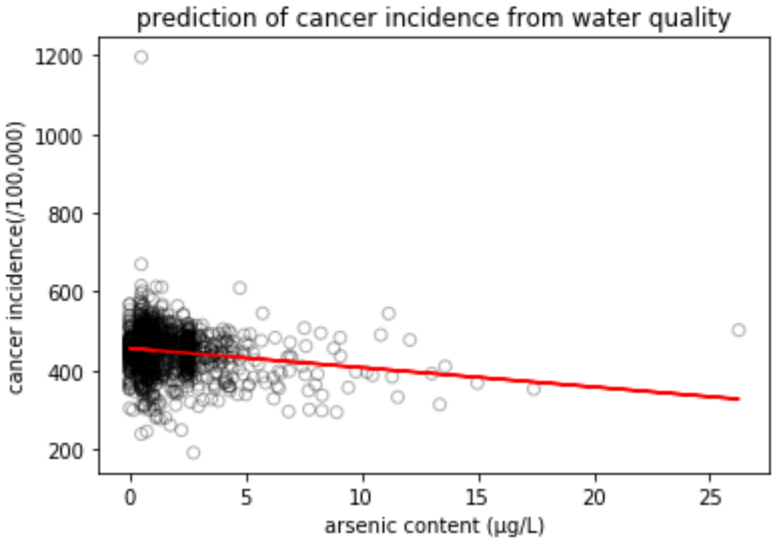


Figure 10:. The regression analysis shows an overall negative trend.

6.2 Hypothesis 2

The air and water quality scores in each county can be used as predictors for that specific county's cancer rate (the average number of individuals affected with cancer in each county over a five year period).

To test this hypothesis, the classifier tests K Nearest Neighbor (KNN) and Decision Tree were used. These tests were chosen because they are excellent machine learning measures that can examine how well multiple variables might predict a certain "label", in this case, the cancer rates for each county: (0-25%, 25-50%, 50-75%, and 75-100%).

The KNN algorithm used the three datasets and classified "new" data points based on a similarity measure. The classification was done by a majority vote to its neighbors. The accuracy results of using air and water quality scores as a predictor for the average number of individuals with cancer in each county for KNN was: 0.66.

A decision tree algorithm created a tree where each node were the features "water quality" and "air quality". The accuracy of the Decision Tree test was: 0.59.

The decision tree classifier did not perform well in the this analysis as it could only predict 59% of the counties cancer rates. However, KNN performed a little better by predicting 66% of the cancer rates. Air and water quality together do not seem to be extremely accurate predictors of the rate of cancer. However, based on the results of the hypothesis 1, water may not be a reliable factor. Future analysis could focus on just using air as a predictor, or bringing in more variables such as socioeconomic status and number of power plants per county. Additionally, looking at the specific type of cancer may also assist in this analysis. For example, skin cancer could be factored out of the overall cancer rate as the main cause of skin cancer is sun exposure and not linked to air or water quality.

6.3 Hypothesis 3

The air and water quality scores in each county can be used as predictors for that specific county's cancer trend (rising, falling or stable cancer rates each year)

To test this hypothesis, cancer trends were classified as "rising", "falling" or "stable". The classifier tests Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) were performed using air and water quality as predictors of the above classifiers.

The Naive Bayes algorithm predicts the class of a new item by assuming that the features being classified are independent of all other features (Gates, 2018). In this case, there is no indicator of the relationship between air quality and water quality due to two different sources. In other words, the two features used to predict the future trend are independent. In addition, NB is a robust algorithm which ignores any irrelevant values or missing values, which will help with excluding the outliers in the air quality data.

Support Vector Machine is another suitable algorithm for predicting the cancer rate trends in this hypothesis. Just looking at the values of water quality and air quality when plotted in a 2-D system, there are many records that are very close (slight differences in water and air quality values) but have different cancer rates. This is likely because they may not be the only two factors to influence the cancer rate trend. SVM maximizes the margin of the category of those points to avoid the bias. Both SVM and NB algorithms classifies the data but in different angles; they are tested and verified by each other.

Finally, Random Forest was used due to the size of the datasets. Performing bootstrap sampling and the Random Forest algorithm for this hypothesis that has about 6000 records in only three categories, will effectively output the importance of each variable.

NB performed the best with a 0.68 accuracy, followed by SVM (accuracy= 0.65). NB could successfully predict 68% of the counties cancer trends based on their air and water quality scores each year. Similarly to hypothesis 2, this accuracy might be increased by eliminating water quality as a predictor and implementing other variables. Random forest was able to predict 112 of the results, with the majority result being a "stable" cancer rate. This suggests that the water and air quality variables together are good predictors for counties with stable cancer rates. This may be due to air quality and water quality also being stable through the years, therefore there may be other factors affecting the cancer rates in each county that must be considered to gather a better idea of cancer trends in the United States. Overall, similar conclusions from hypothesis 2 can be drawn for this final hypothesis. By analyzing the socioeconomic status of each county, it may be possible to predict cancer rates at certain levels of wealth by their water and air quality scores.

Section 7.0 Conclusion

The initial analysis of the water quality, air quality and cancer data in the United States provided substantial, important information to be used in further and final analysis. The questions "Where should you live for your health?" can clearly not be answered with just air and water quality data. The results of the tests performed indicate that air quality might be a better predictor of cancer rates than water, but further analysis on specific type of pollutant, as well as cause of pollution (i.e. power plants, landfills, etc) should be conducted. Furthermore, socioeconomic status needs to be taken into account for each county. Lastly, looking at how air quality might affect specific types of cancer (i.e. lung and breast cancer) might also lead to more interesting results.

It is no secret that environmental factors are not the only cause of cancer in the US, therefore mixed results like the ones we see in this initial analysis are not surprising when air and water quality are the only factors considered. Careful and thoughtful analysis of environmental effects affecting cancer rates will be considered in the final analysis portion of this project, and additional data will need to be brought in to garner significant results.

References:

(1) Kampa, M., Castanas, E. (2007). Human health effects of air pollution. *Environmental Pollution* 151(2):362-367.

(2) Hodges, J. (2018). Air pollution kills 7 million people a year, WHO reports. *Bloomberg Business.* Web.https://www.bloomberg.com/news/articles/2018-05-01/air-pollution-kills-7-million-people-a-year-who-reports Accessed: 1Oct2018.

(3) Herceg, Z. (2007). Epigenetics and cancer: towards an evaluations of the impact of environmental and dietary factors. *Mutagenesis*. 22(2):91-103.

(4) Mokdad, A., Dwyer-Lindgren, L., Fitzmaurice, C., Stubs, R., Bertozzi-Villa, A., Morozoff, C., Charara, R., Allen, C., Naghavi, M., Murray, C. (2017). Trends and patterns of disparities in cancer mortalities among US Counties, 1980-2014. *JAMA* 317(4):388-406.

(5) U.S. EPA 2001b. Quick Reference Guide for Arsenic. EPA 816-F-01-004. Available at http://www.epa.gov/safewater/arsenic/pdfs/quickguide.pdf

(6) Gates, A. (2018). Lecture 7: Machine Learning and Evaluation Techniques.