ANLY 501 PROJECT PART1

Where Should You Live For Your Health?

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

Data Science Problem

When considering our health, or what it means to make healthy decisions, we usually think of factors such as what we choose to eat, how much we exercise, and how much alcohol we drink. However, some commonly under looked factors affecting our health are the quality of the air we breathe, our proximity to landfills, power plants and waste facilities, and the quality of the water that we drink. These factors are not ones we can easily modify or change, as they are defined by the place in which we live, and often times our socioeconomic status.

Although possibly forgotten by the general population, science has not overlooked these factors, as shown by the abundance of current research that discusses how our surrounding environment affects our health. For example, one study shows how air pollutants can have both acute and chronic effects on human health, resulting in health problems ranging from upper respiratory infections to lung cancer (Kampa and Castanas, 2008). Additionally, the World Health Organization (WHO) recently reported that air pollution is the cause of 7 million deaths worldwide, every year (Hodges, 2018).

More specifically, ongoing research is looking at how non-genetic factors, such as the environment, will affect your likelihood of being diagnosed with cancer (Herceg, 2007). One study in particular found that where you live may determine your likelihood of dying from cancer, insinuating that there is a relationship between geographic location and cancer incidence rate (Mokdad et al 2017). Although this study suggests reasons such as smoking, obesity, and diet, it does not take into account the environmental factors of the counties with particularly high rates of cancer.

Our data science problem aims to investigate how your choice in residence may be impacting your health. To do this, we plan to use cancer rates as our 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, and its causes have been linked to many poor health choices, including environmental factors, making it a very good indicator of the health of populations. We will use a dataset from the NIH that includes cancer type, demographic information, and location for each

across the United States. We will also gather two additional datasets that will help to describe the environmental status of each county in the United States. We will use air quality datasets from the Environmental Protection Agency that shows air quality indexes for common air pollutants such as NO2, O3, and CO. We can also use a water quality dataset gathered from the Center for Disease Control (CDC) to observe the heavy metal contaminants in the drinking water sources for each county. By using these three datasets we can gather a basic idea of each county's environment, and how it is potentially correlated with their reported cancer rates.

We recognize that this may not be enough data to truly understand the environment and health of each county, or the corresponding cancer rates. Therefore, we plan to eventually implement data on nuclear power plant sights, UV index data, landfill data, agricultural farmland data, and wind data as additional environmental inputs. We are also considering looking at childhood cancer rates, and reproductive birth outcomes to consider other health effects that can occur due to the environment. Additionally, we can look at the "wealth" of a particular county to see how this may affect both the environment and cancer rates. Wealthier populations are likely farther from landfills and plants, can afford bottled water, water filters and air filters, can afford better prenatal/postnatal care, and therefore far less likely to reflect the negative health implications of their environment.

Potential Analyses that can be Conducted Using the Data

From this data we want to investigate several questions, all under the umbrella question of: Is your health being impacted by where you live?

What areas of the country offer the lowest chance of being diagnosed with any long-term health issues?

Can we predict your chances of getting cancer based on where you live?

Can we predict the chances of your child getting cancer based on where you live?

Do your demographics (race, gender, etc) affect your ability to withstand negative health impacts from environmental factors?

Does your income decrease your risk of getting cancer?

Are certain environmental indicators highly correlated with particular health issues?

Is the average household income of your county a predictor of the environment? Of the cancer rates?
Does living near agricultural farms affect the quality of your water, and will this impact your health?
What cities are going to kill you?

We are hypothesizing that counties with poor air quality and water quality will have higher cancer rates and health issues overall. In addition, we believe that your proximity to landfills, power plants, nuclear plants, and agriculture farms will all have an effect on your chance of a cancer diagnosis, and that wind directions may also affect this outcome. Currently, we lack data on landfills, power plants, nuclear plants, and agriculture farms. However, we plan on collecting and analyzing additional data as we move forward with the project. Finally, we hypothesize that the wealth of a particular county will impact these outcomes, in that wealthy people living in wealthier counties may have lower cancer rates than people of lower socioeconomic status in the same county.

We hope to use these data to predict the potential health risks for popular cities in the United States. This information can prove highly useful for people considering a move, as they can better understand the health implications of where they choose to live. We hope to bring to light the importance of considering environmental factors when moving, or at the very least, being aware of the dangers your current or future city may pose to your health.

Data Issues

For the Air Quality data, there are errors in the Entries for State and County attributes. From looking at the output of the script, there seems to be unnecessary white spaces contained within certain values in the County attribute. In the States column, Country of Mexico, Puerto Rico, and Virgin Islands are not states and should be removed. In addition, there seems to be an error

in the District of Columbia data entries. The total number of messy states is 4, and the total number of messy counties is 247. Lastly, there are no invalid entries for the numerical attributes, so no cleaning needs to be done on those columns.

For the cancer data, there are a few hundred rows of missing values, denoted by ¶ or ¶¶, and a whole column with only asterisks. There is also an index column at the beginning which is unnecessary. Additionally, the County column has both county and state names along with the cancer registry numbers that the data came from, which should be split into four columns: State, County, SEER Registry Number, and NPCR Registry Number. Additionally, all the columns besides FIPS are imported as strings even though they are numbers, because of the special characters used to denote missing values. These values will need to be removed and the columns converted to numeric before we can continue with our analysis. Still, other than missing values, there are no invalid entries in this dataset. The only data changes will be for some individual data points which need to be reformatted to exclude marks. For example some values are followed by a hash symbol "#" that denote a footnote about the data, but need to be removed for our analysis.

Overall, the Water Quality data is clean and well organized although there are a few duplicated data columns and unstructured data types. From checking the output of the script by using the .isna() function, there are no missing values, and there are no incorrect data shown by applying value_counts() and unique(). However, the dataset has some repeated columns that state exactly same contents by examining the overview. For example, columns "geoAbbreviation" and "geoId" shows same geographic ID numbers, shown in Water.fig1.

geoAbbreviation	geoId
6001	6001
6001	6001
6001	6001
6001	6001
6003	6003
6003	6003
6003	6003
6005	6005

Water.fig1

Data Checks Methodology

A function called “checkType” was created to gather summary statistics on each dataset. The summary statistics were written to an outfile that corresponds to a specific dataset. Furthermore, this function is general in the sense that it was applied to each of the 3 datasets.

Another function called “dataInfo” was created to determine the size and shape of each dataset. In the context of this project, shape is referring to the dimensions of the dataset and size is referring to the number of elements in the dataset. Lastly, this function analyzed each attribute and determined the type. This helped us determine possible invalid data entries in different attributes. For instance, if a attribute type is an object instead of numeric, this points to the possibility of having strings in a row somewhere.

The “nullCount” function determined the number of rows with missing values and the number of missing values in each attribute. This function helped quantify the amount of dirt found in a dataset.

A function called “get_Univalue” was created to gather all unique values for each attribute. This is a general function that helped determine any data input errors in each column. In the cancer dataset, this function helped us discover strings and symbols in columns that should only contain numerical entries.

The “stateCountyChecker” function compared states and counties from the air quality dataset to a reference. State and County attributes are of type object and each value is a string. The function compared each string to a reference dataset to check for possible data input errors and erroneous white spaces. Lastly, it ensured that each value is in the form of a title.

The "numericColumnChecker" function checked if the numerical attributes in the air quality dataset were within the correct range. For example, we would expect values within the "years" attribute to between 2011 and 2017, and values within the "Good Days" attribute would between 0 and 366. The reason the max number of days is 366 and not 365 is because leap years contain 366 days.

Data Cleaning Metric

We have adopted "6 dimensions of data quality assessment (6D-rubrics)" as rubrics to exam data quality for each dataset (Dama, 2013). The 6D-rubrics examines

1) Completeness - The proportion of missing values in the stored data.
2) Uniqueness - The level of duplication present in the data.
3) Consistency - The degree to which the data diverges from its description or definition.
4) Validity - The syntax of data's definition (e.g. as it relates to the format, type, and range specified in the metadata and data dictionary).
5) Accuracy - The degree to which the data describes its meaning in real world.
6) Timeliness - The degree to which time-related measures are accurate in their collection with respect to their definition. The degree to which the data conform to the time period specified.

Then we calculated the proportion of error values under each dimensions to generate scores for each dataset.

Table1: Score for Data Quality by 6 Dimensions

	Completeness (16.67%)	Uniqueness (16.67%)	Consistency (16.67%)	Validity (16.67%)	Accuracy (16.67%)	Timeliness (16.67%)	**Total (100%)**
Cancer	15.48%	16.67%	16.25%	16.67%	16.67%	16.67%	**98.39%**
Air Pollution	16.67%	16.67%	16.67%	14.88%	16.67%	16.67%	**98.23%**

Water Quality	16.67%	(1-2/26)* 16.67%	16.67%	(1-2/26)* 16.67%	16.67%	16.67%	**97.46%**

Source from: (Dama, 2013)

In the cancer dataset, 224 counties have missing data because either the counties are in states where legislation and regulations don't allow them to share county-level data, or they are in the state of Nevada, which is also missing. There were also a few cases where the data was suppressed to ensure patient confidentiality (this occurred when there were fewer than 16 cases in a county). All the rows are unique, but the incidence statistic isn't totally consistent throughout because there are 80 counties where the data excludes cases which were diagnosed in other states due to a data exchange agreement that prohibited the release of data to third parties. The data also come from different cancer registries depending on the county location, but since the collection methods appear to be the same and the same calculation is used across all counties to determine incidence rates and trends, we think this shouldn't impact the data's accuracy or validity. The data was collected annually for five years and then averaged to produce the incidence rate, so we think this, along with the recent trends (e.g. rising or falling for a given confidence interval), is sufficient from a timeliness perspective to produce accurate data. This produces a score of 98.39%.

There are 2 flaws on aspects of uniqueness and validity in water dataset. For uniqueness, there are 2 pairs of columns ("geoAbbreviation" & "geoId", "dataValue" & "displayValue") are duplicated. Since water quality dataset has 26 columns, the dataset gets (1-2/26)*16.67% for Uniqueness. In addition, when we check the validity of the dataset, we find that 2 columns (dataValue, displayValue) should be in type of float64 but it shows object in the raw dataset. Hence, these two inappropriate data type leads to a (1-2/26)*16.67% off. This produce a score of 97.46% for water data.

The errors in the air quality dataset are a validity problem. The problem comes from the fact that there are 4 states that are either not states or are messy entries. In addition, there are 247 values under the "County" attribute that contain mistakes. As a result, the validity score is [1-(4/54)-(247/7455)]*16.67%, which equals 14.88%. The 54 represents the total number of unique state

entries found in the dataset, and 7455 is the total number of rows in the dataset. There are no null values, therefore, it is logically sound for 7455 to represent the total number of elements in any given column.

Data Features

For the water quality dataset, we generated two new data features. Our third feature was generated with the air quality dataset.

- In the dataset, one county has multiple water suppliers, every water supplier has different values of Arsenic concentration. To modify a unique value of Arsenic concentration to every county, we took the average of the values in one county. At this step, the output data frame has 953 rows and 2 columns - average and title, where title has two indexes: county and state. Hence, we also split "title" into "county" and "state".
- It is necessary to define the water quality by the given average value of Arsenic concentration for each county. The CDC website states that "Lowering the 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).". Here, website gives a pivot number of 10 to define if the concentration of Arsenic is statistically related to the cancer rate. Binning the average value would be obvious to see if the concentration is under or above 10μg/L. Hence, we set the interval for the water quality: [-1,1] as "Non detect", (1,10] as "Less than or equal 10", and (10, 50] as "More than 10".
- Fine particles can come from various sources that include power plants, motor vehicles, airplanes, forest fires, agricultural burnings, and volcanic eruptions. Fine particles refer to atmospheric particulate matter that is smaller than 2.5 micrometers, denoted as PM2.5. These small particles are believed to penetrate deep into the lungs and enter the circulatory system. Furthermore, it is believed that these particles contribute to adverse health effects. In the air quality dataset, we choose to create a new binary feature that determines if each county had any days out of the year with PM2.5. The "Days PM2.5" attribute measures the number of days where PM2.5 was the most significant pollutant, and the new binary feature is called "hasPM2.5". A value of 0 was assigned to counties

that had 0 days in the year, and a value of 1 was assigned to counties that had 1 day or more.

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) Dama. (2013). THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT. [online] Available at: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf [Accessed 6 Oct. 2018].