Data Wrangling Project

--Lung Cancer and Air Quality Analysis

By No Error

Introduction

The effects of future climate change on public health are an active and growing area of research, the impacts of climate extremes on future air quality and associated health implications are also under analysis.

Lung cancer is the first cancer killer of both men and women in the United States, meanwhile, overwhelming evidence shows that particle pollution in the outdoor air we breathe—like that coming from vehicle exhaust, coal-fired power plants and other industrial sources—can cause lung cancer.

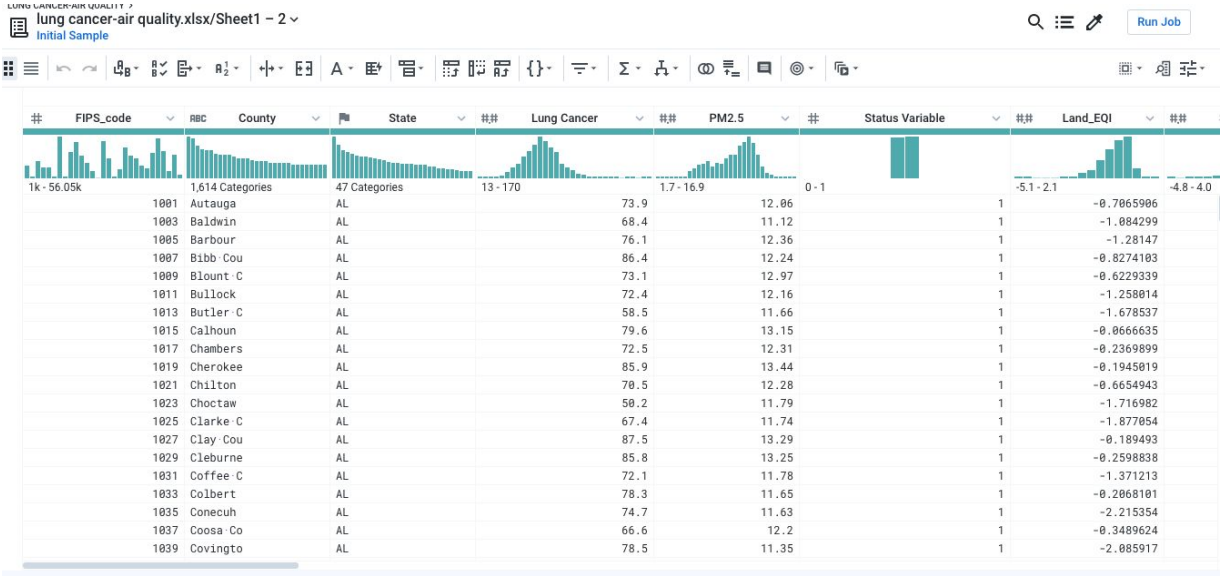
Even though air pollution levels in the U.S. are not really serious, particle pollution (for example, PM 2.5) still exists. Particle pollution increases the risk of dying early, heart disease and asthma attacks, and it can also interfere with the growth and function of the lungs.

However, most people still don't know that particulate pollution is a risk factor for lung cancer. That's why it is important we need to analyze the relationship between air quality and lung cancer.

Data Source and Initial Quality

We found the “Air Quality-Lung Cancer Data” from the Harvard Dataverse website. From its description, we know that the original data comes from two different sources. The Population-based lung cancer incidence rates data were abstracted from the National Cancer Institute state cancer profiles, which is a national county-level database collected by state public health surveillance systems. And the domain-specific county-level environmental quality index (EQI) data were abstracted from the United States Environmental Protection Agency (USEPA) profile. Such data sources are reliable, so we perform our data wrangling process on this data.

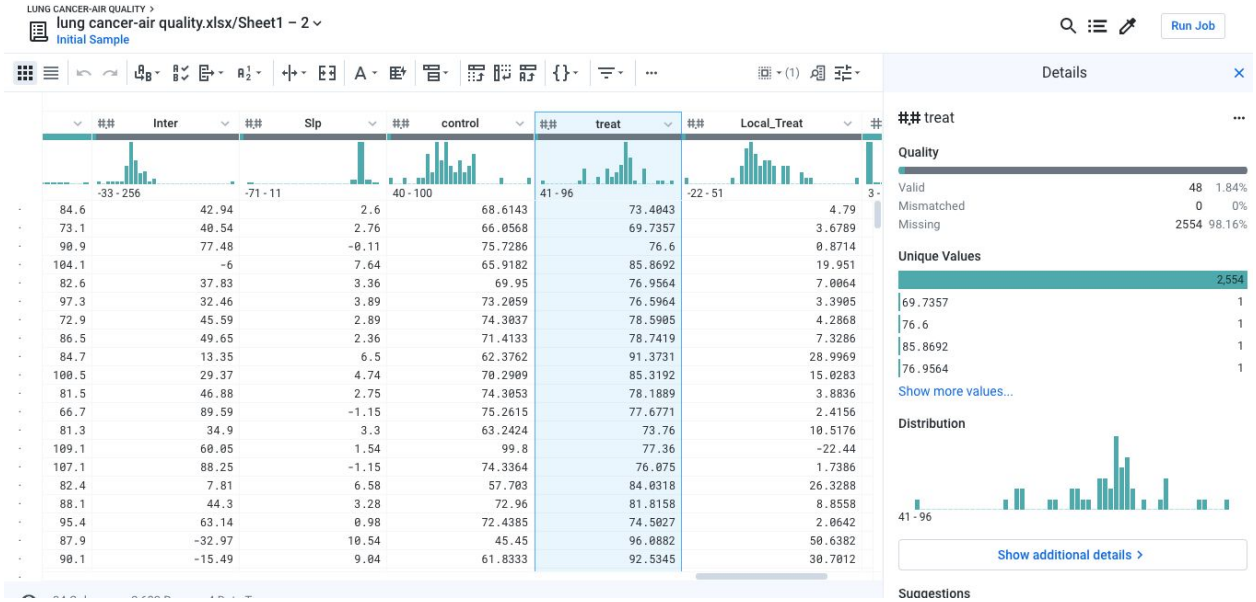
The initial data quality is great, the values are valid and clean, there is no mismatching value, we can see from the histograms in Trifacta, most of them are in green color:



Screenshot 1

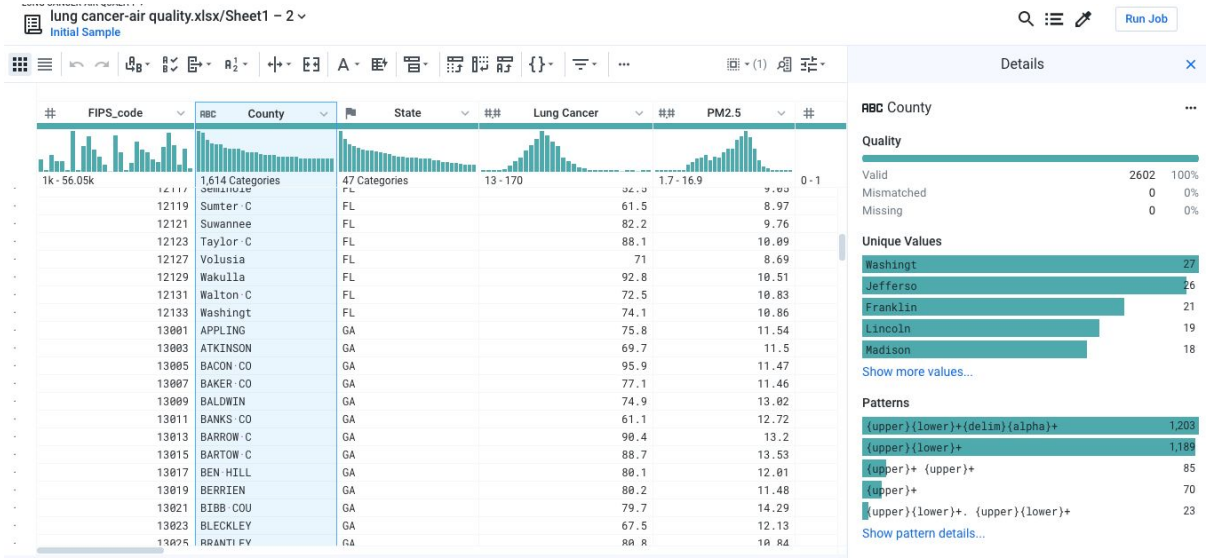
However, out of 3144 counties in the United States this dataset has available information for 2602 counties, some state's data is not available. Besides, some variables have large percentage missing values and are not relevant to our analysis goal, so we may need to remove them.

For instance, value "treatment" only contains 1.84% valid values.



Screenshot 2

Moreover, we notice that some county names are not in the same format(some are uppercase in all letters), which should be modified:



Screenshot 3

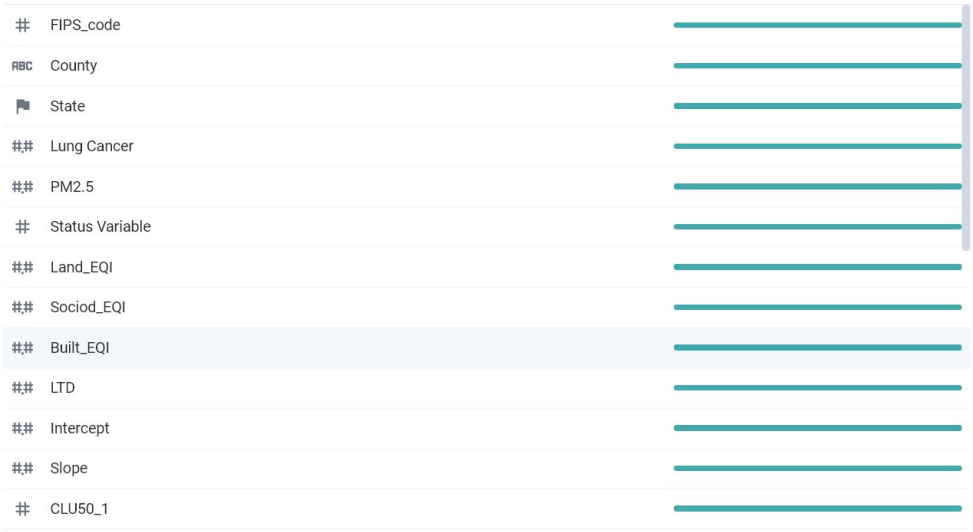
Then we start our data wrangling process.

Data Wrangling using Trifacta Wrangler

Firstly, Lung cancer mortality information in states namely Kansas, Michigan, Minnesota, and Nevada are not available, so we deleted the rows of these states.
Then, we deleted the columns of Inter, Slp, control, treat, and Local_Treat since the valid data in the five columns only occupy less than 2% of the rows and information is not important for our goal.
The Union county, Florida is an outlier in terms of mortality, so we deleted it from the dataset.
Finally, we changed all the county names in uppercase to lowercase with the first letter capitalized.
There are columns such as intercept and slope, which seem to be irrelevant, but they help to divide the dataset into several clusters and are helpful for understanding the distribution of data.

Final Data Quality

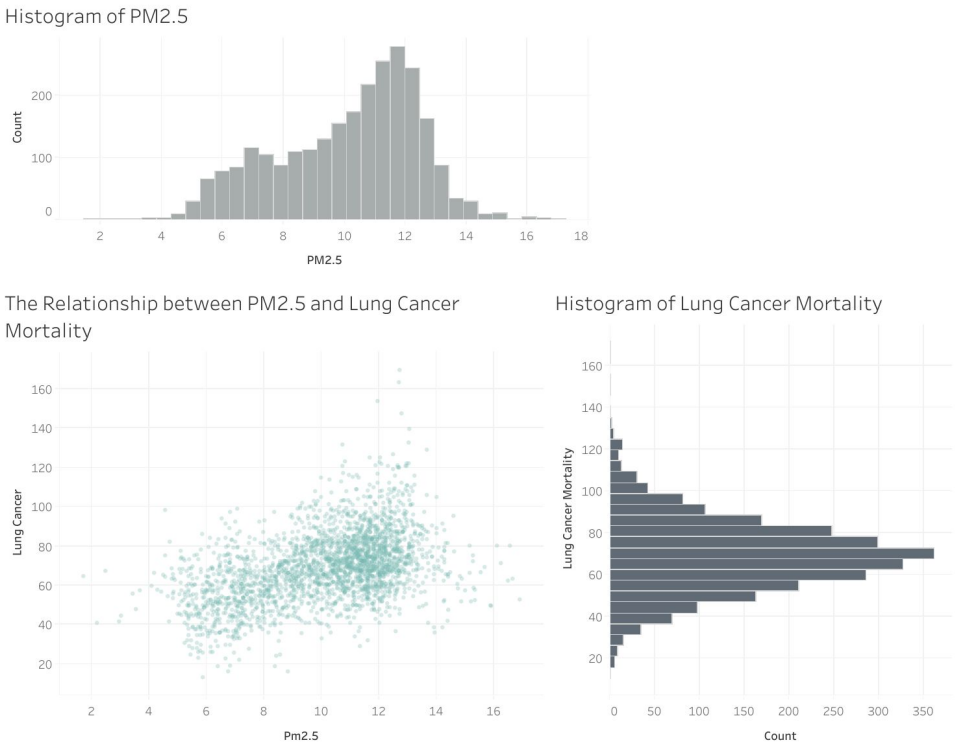
The final dataset has 100% valid values and no mismatching values or missing values. And the values in some columns are more consistent and relevant, and there are fewer outliers. All the values are in the right formats, so the final dataset has better uniformity. Since most of the data are from the official resources, they are likely to be accurate and reliable. Additionally, the dataset concludes the most used air pollution index and air quality index, it is relatively complete.



Screenshot 4

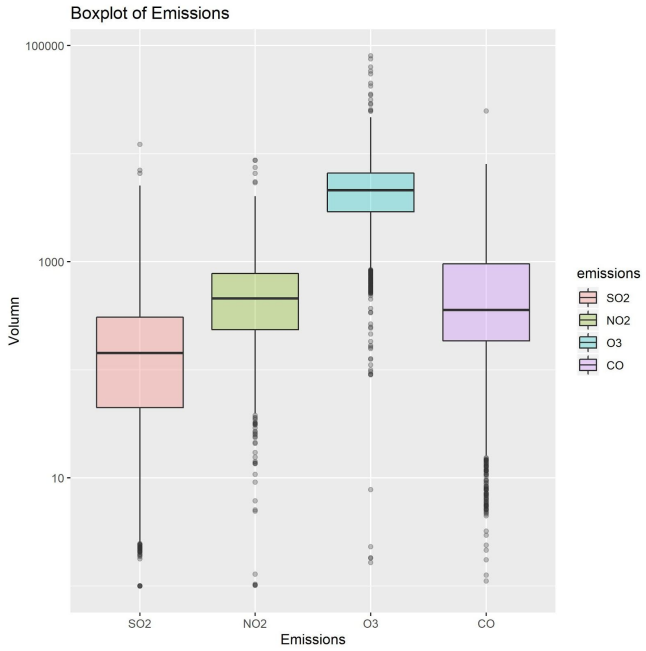
Descriptive Statistics for the Dataset

Taking our project to the next level, we plan to further explore the relationship between lung cancer mortality with environmental factors. Considering this, PM2.5 is regarded as the very first key variable.



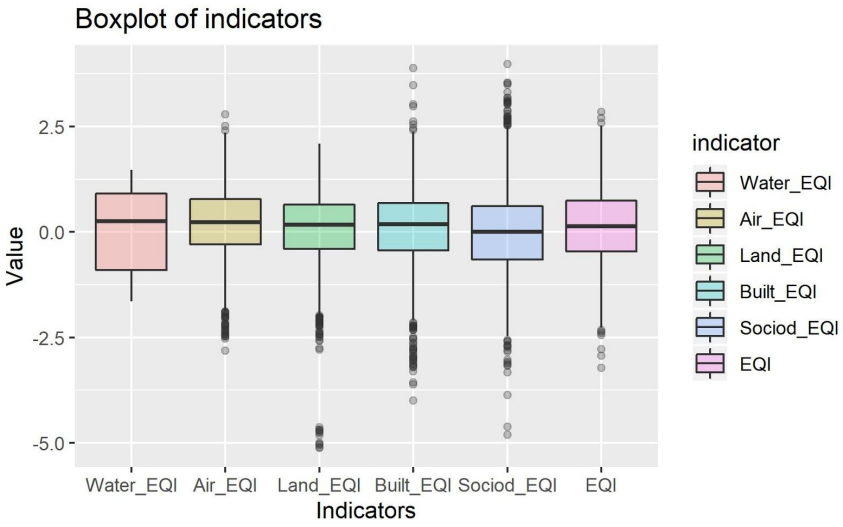
Graph1 - Relationship Between PM2.5 and Lung Cancer Mortality with their Histograms

We plot the histogram for both PM2.5 and lung cancer mortality, with a scatter plot indicating their correlation. The histogram for mortality shows symmetry without outliers while the one for PM2.5 has thick left tail without unusual values as well.



Graph2 - Boxplot of Emissions

From this boxplot, we can tell that the emissions across different areas vary largely from nearly 0 to almost 100000. It cannot conclude that the values which largely deviated from mean are outliers. In this case, we can only say that the air quality differs from place to place.



Graph3 - Boxplot of EQI (Environment Quality Index)

EQI demonstrates environment quality from different perspectives, ranging from -5 to 5. We can see that water quality doesn't have that huge variation as the other indicators.

Summary

By using Trifacta Wrangler, we prepare our data and improve its quality, making it ready for analysis. After the process, we have some reflections on it.

Trifacta Wrangler In Analytical Modeling Process

Trifacta Wrangler can be extremely helpful throughout the entire analytical modeling process. Before modeling, Trifacta allows us to know the dataset better in a more straightforward way. Through the Data quality bar and the Histogram, we are able to directly visualize and detect missing values, unusual values and outliers without writing long paragraphs of codes. Trifacta also makes it convenient to clean and structure the data into an organized format. Taking our dataset for example, if the geographical factor is considered to be a determinant of lung cancer mortality in our further model, the "State" column is required to stay in the same pattern(capitalized abbreviation) first and should then be transformed into multiple columns of dummy variables. By using the Builder, the cleaning and one-hot encoding process can be easily done. And these steps can even be written in the Recipe and be applied to other similar categorical variables. By doing so, it's not necessary to go to the trouble to repeat the data preparation process manually. Moreover, Trifacta is a good helper as well in terms of building the model. For instance, when KNN or neural networks model is going to be utilized in classification, Trifacta will provide "shortcut" to normalize numerical variables, making sure they are at the same scale. To conclude, building analytical models needs data with high quality, preparing which could be a time-consuming and suffering process. However, Trifacta offers a simplified self-service data preparation platform, saving our time and making the entire modeling thing more enjoyable.

Trifacta Wrangler In Combination With Other Software Tools

After data wrangling, we could use what has been produced in Trifacta Wrangler to combine with other software and make further exploration. For example, we could import cleaned data in Tableau TDE format from Trifacta and open the file directly via Tableau. In the past, to make visualization in Tableau, a huge amount of time was needed to be invested in the data cleansing process, though Excel, SQL or R. Sometimes we need to use all of those tools which make things more complicated. But now, with the help of Trifacta, data wrangling becomes a straightforward process and could be done just in one tool. Also, data wrangling and data visualization could be tightly combined, it's much easier to handle data issues.

In addition, Trifacta also included extensive support for deploying Trifacta in Spark or Hadoop environments and integration with various cloud services. In this way, a lot of organizations which are moving to the cloud can apply their wrangling process on a user's desktop, in the cloud, amongst teams with diverse data in various systems and in the Hadoop/big data realm.

That’s to say, a convenient data wrangling process with Trifacta could be applied everywhere without obstacles.

Trifacta Wrangler In A Collaborative Workflow Within Or Across Organizational Boundaries

In most cases, data wrangling is complicated and multivariate. With the “share” function in Trifacta Wrangler, different members in a team can work on different parts and accelerate the progress of data cleaning.

In addition, since data preparation is a company aspect effort that requires collaboration between the business team and IT team, Trifacta could also help improve the efficiency between different functional departments. For example, with collaborative workflow, IT teams can be responsible for maintaining data quality throughout the entirety of the organization, business teams could decide what can be used for analysis and what needs to be refined.

Also, when it comes to issues in a specific field, sometimes the knowledge within a team is not enough. In this case, collaborative workflow across organizational boundaries is necessary. External experts could be invited to the company's workflow to solve the problem, or data from third parties could be connected to the company's workflow to provide solutions.