The Correlation Between Air Quality and Trees

Kezia Regis

656188739

ABSTRACT

This study investigates The Correlation Between Air Quality and Trees. I will use three datasets from the New York City data webpage to explore air quality: the 2015 tree dataset and the 2005 dataset. Based on the analysis of each dataset, there is not a strong correlation between air quality and trees. To conclude, based on the tests and distributions of data values, I would say the null hypothesis can not be rejected.

RESEARCH QUESTION

By investigating datasets that show air quality in different neighborhoods, is there a trend with trees' impact on air quality in New York City?

BACKGROUND & PRIOR WORK

A few street trees were requested in my neighborhood through the NYC Street Tree Planting program. My area tends to have traffic due to a two-way street with passing cars and city buses; additionally, there is a subway station nearby and warehouses with trucks moving shipments. These quirks within my neighborhood have made me wonder whether the many trees planted reduce the pollution in the air.

HYPOTHESIS:

The Null Hypothesis states that trees do not have a noticeable positive effect on the air quality in NYC neighborhoods. Whereas the Alt Hypothesis states trees have a positive correlation on NYC neighborhood's air quality

This project consisted of three datasets that examined Air Quality: New York City's 2015 tree data and New York City's 2005 tree data, with a total of 628,420 entries. For the air quality dataset, I used three features: Data Value, Time Period, and Geo Place Name. The air quality dataset was ideal for providing the air quality data value throughout New York City neighborhoods for 2015 and 2005.

The following dataset used was NYC's 2015 Tree data. There was no direct relationship between both datasets, but exploring the relationship trees have on air quality would be intriguing. I used the features nta_name, which serves as neighborhoods found throughout NYC, a column that shows the address where the trees are located, and status, which ranged from poor to sound, to determine the relationship between a tree's status and air quality data value.

Moreover, I used NYC's 2005 Tree dataset, which had a stronger correlation to the 2015 tree dataset than the air quality dataset because the collected data was a census of the number of trees in each NYC borough, so the number of trees may not change. Similar to the 2015 tree data, it filters out the nta_name for the tree's neighborhood, address, and health, which ranges from dead to good. Overall, my decision to have three datasets was to incorporate a range of data that would allow me to explore the relationship between trees and air quality.

ETHICS & PRIVACY

I accessed these datasets through the NYC Open Data website, so I did not need to request permission. Each dataset was not private or confidential because New York City provides information about air quality through weather apps or phone alerts. The tree data could be found

by manually counting each tree in every borough. New York City contains many neighborhoods, so some data may have been left out, but overall, there is no noticeable bias towards a certain population.

DATA CLEANING

As I analyzed the data, I included many packages to simplify the data manipulation and visualization process. Libraries that I have used are pandas, numpy, and matplotlib.pylab, seaborn, category_encoders, and sklearn. To gather the information from each CSV file, I used pandas's read_csv function, and then it was placed into a variable that would be used to filter out needed columns. During my data cleaning process, I used the panda's library to keep my data as a data frame and to start filtering the air quality dataset. For this dataset, I wanted to merge all three datasets so the process of finding correlations or hypothesis testing would occur smoothly, so I changed the column's name from Geo Place name to nta_name. Then, I replaced the duplicated and missing values from the 2015 tree dataset with the most frequent value using sklearn's Simple Imputer library.

DATA ANALYSIS & RESULTS

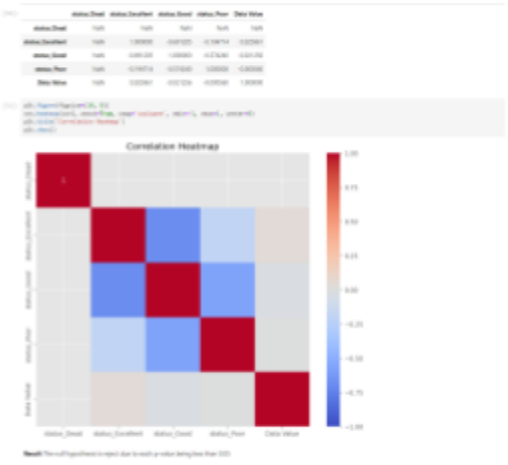
After cleaning the data, I could move into the EDA process, which started by finding outliers. The air quality dataset had many outliers, so I used the IQR method to find where they were placed, and then I removed them using the upper limit and lower limit.

Above the maximum line there are multiple outliers pictured above, so using the IQR method the outliers will be removed

Most of my data from the tree data were objects or categories, so I used One Hot Encoding instead of integer encoding because it would reduce the amount of bias, which could lead to inaccurate predictions when the correlation is calculated. After hot encoding, I could merge each tree's data with the air quality data so I could move on to the Kolmogorov-Smirnov test. I chose this non-parametric test because I did not want to make assumptions about the underlying distribution of the data. Instead, I wanted to determine whether the two distributions were different from each other. After completing this test, I got the result of -54, significantly less than the significance level of 0.05. Since the Kolmogorov-Smirnov test was the only approach I took to find the relationship between each dataset, I continued with pandas's correlation function to determine if there was a relationship between the three datasets.

The merged air quality and tree 2015 dataset results show 0.05 and -0.05 for the health_Fair and health_Good columns compared with the Data Value column. As a result, I concluded that there is not a linear relationship between both datasets. The results were similar between the merged air quality and tree 2005 dataset, where the correlation values were very close to zero.

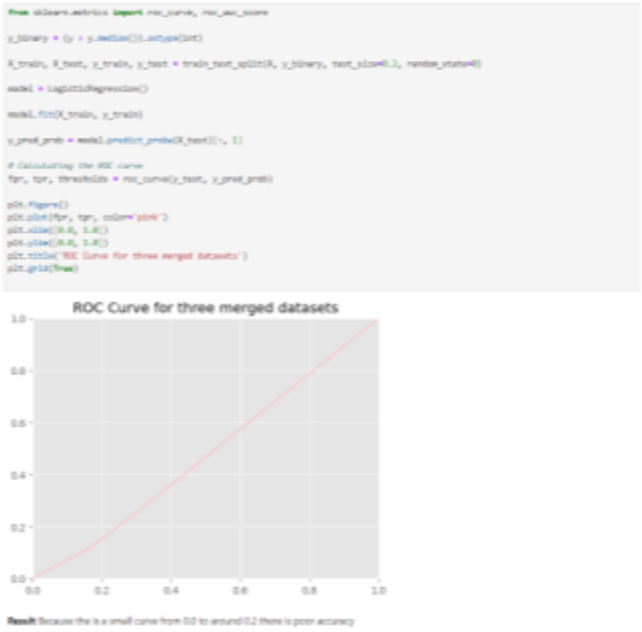
The correlation between health_Fair, health_Good and Data Value shows a no linear relationship



The correlation between status_Excellent, status_Good, status_Poor and Data Value shows a no linear relationship

The last step of the process was model validation. I wanted to know whether the model fit the data properly, so I used linear regression to find the Mean Squared Error and R-squared values. After calculating both values, I found the mean square error to be larger than the R-squared value, so the null hypothesis was not rejected.

To visualize the data, I used logistic regression to plot the ROC curve; however, a majority of the model was perfectly diagonal, so the plot is classified as not accurate and is closer to a false positive rate.



The dataset shows a linear ROC Curve. A diagonal curve mean the plot is not accurate.

CONCLUSION

The null hypothesis could not be rejected from the tests and distributions of values. Before concluding that trees do not have a positive correlation with air quality, I believed the sun

provided nutrients to plants but also contributed to human well-being, so trees receive support from the environment and assist other lifeforms. The lack of correlation between the three datasets makes me question if the amount of variables I had in my datasets was enough to reach a fruitful conclusion. If I redo this project, I would make my hypothesis explore how the number of trees affects the air quality and the health of NYC's population.

REFERENCES

- Author links open overlay panelRodney H. Matsuoka, AbstractHigh school students today are experiencing unprecedented levels of school-related stress. At the same time, Berto, R., Gidlöf-Gunnarsson, A., Hartig, T., Herrington, S., Jackson, L. E., Kaplan, R., Kaplan, S., Lee, S.-W., Owens, P. E., Ozdemir, A., Staats, H., Tennessen, C. M., Tzoulas, K., Ulrich, R. S., Berg, A. E. V. den, Ainslie, R. C., Chambel, M. J., … Hanushek, E. A. (2010, July 24). *Student performance and high school landscapes: Examining the links*. Landscape and Urban Planning. https://www.sciencedirect.com/science/article/abs/pii/S0169204610001465?via%3Dihub

- Department of Health and Mental Hygiene (DOHMH). (2024, April 10). *Air Quality: NYC Open Data*. Air Quality | NYC Open Data. https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r/about_data

- *2015 street tree census - tree data*. NYC Open Data. (n.d.). https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35

- Choi-schagrin, W. (2023, January 19). *How New York City's trees and shrubs help clear its air*. The New York Times. https://www.nytimes.com/2023/01/19/nyregion/trees-plants-air-quality-nyc.html