DATA NUTRITION PROJECT

What is this label?

NYC Borough Air Quality and Crime Report Statistics

Description

Dataset is a cumulative data frame that consists of statistics regarding both air quality and crime reports of the Boroughs in the City of New York, separated by each year between 2009-2020. The purpose of the dataset is to provide a comprehensive look at possible relationships between air pollution and number/types of crime committed in an urban setting. Content of the dataset that could help with this analysis includes, the average value of four types of air pollution (O_3, NO_2, SO_2, $PM_{2.5}$) within each Borough throughout the years, the number of total crimes committed categorized by age and gender groups, the number of certain types of crime committed (assault, larceny, drugs), etc.

Keywords

NYC Boroughs | Air Pollution | Annual Statistics | NYPD Crime Reports | NAAOS Standards

Use cases (potential real-world application of the dataset):

1. What is the effect of Ozone pollution on person's mind?
2. Which pollutant type is most correlated with increase in crime?
3. Is there a pattern between crime and season?
4. Which demographic group is most affected by air quality?
5. Which Borough has the highest crime rate vs air quality?

How to use it?

Intended Use

- **Intended Domain.** Environmental research
- **Intended Domain.** Policy making
- **Intended Domain.** Human psychology
- **Intended Use.** Analyzing trends and correlations between air pollution level and crime rates.
- **Other Responsible Uses.** Making policies on air quality standards and policing based on trends or correlations found.

Restrictions on Use

- **NOT** used for making prejudiced inferences for certain Boroughs: Data does not provide a clear cut picture of what each neighborhood is actually like. **NOT** used to make general assumption about air quality of the effect for different individuals: Data does not take into account many individual factors of a particular person.

Known Uses

- **Research/analysis** To better understand what is needed for well-being of humanity.

Do Not Use

- **Domain.** Any in which data is not used for collective well-being of society.
- **For unethical reasons.**

About the dataset

People

Owned by	NYC OpenData
Created by	Group BB6
Maintained by	City of New York
Funding	City of New York
Management	City of New York

Technical information

Publish Date	2023-11-24
Format	csv
Instances	340 recorded data for each Borough in each given year
Version	1
License	Public Domain
Collection timeframe	2009-2020

Collection process

Air pollutant data is collected by the NYCCAS real-time air quality monitor network reporting the air quality of the NYC Boroughs at different years. Crime rate data is collected by the NYPD Police Department consisting of arrest reports made across each NYC Boroughs throughout the year including demographic and type details.

Useful links

Dataset access point

https://github.com/INFO-201-Fall-2023-Final/final-projects-sgiang1

Link to the data dictionary for the rows of the dataframe.

https://github.com/INFO-201-Fall-2023-Final/final-projects-sgiang1/blob/main/data_wrangling/data_dictionary.md

Representation

Subpopulation: Borough, Pollutant, Age, Gender

Concerns about using data to make decisions/predictions at the individual level:
The data does not account for many individual factors such as current health status and other environmental exposure, thus the data should not be used to make predictions at the individual level.

Mitigation strategy:
Grouping by location, gender, or age group may allow us to see some general trends and insights among each demographic group but will still not be representative of each individual instance.

Other potential representation issues:
For crime reports, there may be an overrepresentation of reports due to potential practices of predictive policing, which ethically is unfair and biased since it disproportionately targets low-income neighborhoods[1] and high minority areas.[1] This may also be caused by different strictness in enforcement between different demographics. For air quality, only areas with sensors installed within are correctly represented, thus not all of New York City is represented.

*See sentences marked with corresponding footnote indicators (i.e. 1 or 1)

Row Count	**340**
Alerts	4
Upsteam: Air Quality dataframe	2
Underrepresentation	1*
Inaccurate Prediction	2*
Upstream: NYPD dataframe	2
Racial bias	1*
Socioeconomic Bias	1*
8 Human Rights Principles	
Privacy	R
Accountability	R
Safety and Security	R
Transparency and Explainability	R
Fairness and Non-discrimination	NR
Human Control of Technology	NR
Professional Responsibility	R
Promotion of Human Values	R

Ingredients: Borough, Year, Name, Measure.Info, avg_value, male, female, felony, misdemeanor, violation, <18, 18-24, 25-44, 45-64, 65+, drug_use, larceny, DUI, assault, total_crime, crime_per_value, above_NAAQS_standard, start_season

*Refer to data_dictionary for more details

Privacy: Crime reports and Air Quality do not report any private infromation. Information are formatted as general descriptions.

Accountability: Data are monitored by NYC government entity and data are not intended to be used for unethical practices.

Safety and Security: Data is safely secured by the city government.

Transparency and Explainability: Collected through air monitors and police reports and is reviewed/mainted by government. Data is presented in comphrehensible format.

Fairness and Non-discrimination: Air Quality and Crime data may not fully be representative of corresponding groups in dataset due to reasons mentioned above thus may introduce some bias within.

Human Control of Technology: Other than checking for outliers, it is hard to know by a human checking if the value recorded by the monitors are accurate or not.

Professional Responsibility: Data is curated by professionals and by following professional value and practices.

Promotion of Human Values: Dataset may be leveraged for research that benifits well-beinig of society as a whole.

Inference risks

At a Glance

About humans	Upstream sources	Technical review	Ethical review	Update frequency
Yes	Yes	Yes	Yes	Yes
Looks at effect of air quality on the psychology of human minds	NYC OpenData	NYC Health and Police Department	Reviewed before reported onto official city website	Every year in NYC annual report

Data values

What values are in each column?

Collection and Labeling Protocols

Police reports including descriptions and details are manually written by the individual NYPD officers. Air quality monitors are trained to automatically feeds back information about each of the air pollutant levels.

Data Imputation Protocols

There are no missing data for the air quality of each individual Borough within the year 2009-2020. Missing data about the type of crime from police report are sometime omitted and thus not accounted for in column that stores the respective cumulative count of that crime type.

Data Manipulation Protocols

New numerical columns storing the count each demographic group, crime type, and annual average pollutant level is created for each Borough during each given year from 2009-2020. New categorical variable such as the Boolean above_NAAQS_standard stores TRUE or FALSE based on whether the given average pollutant value given its pollutant type is above or below government standards.

Missing Data

Some description for crime were not included within report. Air quality sensors may not be the most accurate.

Raw Data

https://catalog.data.gov/dataset/

Feature selection

Which columns were chosen and why?

Cultural or Domain Assumptions

The Boroughs of New York City are the five major governmental districts that compose New York City. Air pollutant values are generally measured using the unit 'ppm' or 'ppb' representing how many parts a certain molecule makes up within the one million parts of the whole solution.

Proxy Characteristics

Person's age may have influence on likelihood to commit crimes. Air quality may affect the crime rate within a given Borough.

Planning Representation

Reports were based on an observational study as opposed to a survey.

Domain Knowledge

The Boroughs of New York City are the five major governmental districts that compose New York City. Air pollutant values are generally measured using the unit 'ppm' or 'ppb' representing how many parts a certain molecule makes up within the one million/billion parts of the whole solution.

Upstream sources

Are there known risks in datasets upstream?

Intended Use Familiarity

- NYC OpenData: Very familiar. Data are collected through observation and are intended to be used by government officials such as policy makers or researchers along with the general public to analyze.

Data Collection Familiarity

- NYC OpenData: Very familiar. Representation due to oversampling in certain neighborhoods may cause bias within the data. Areas marked with predictive policing may consist of overrepresentation of a certain demographic while other areas may be underrepresented. Areas with less air quality sensor networks such as more remote places are subject to underrepresentation as the air there is not recorded as often.[1]

General risks

Any additional risks?

Individual Information

Demographic of criminals are reported when arrested but no personal information is shared.

Consent

Yes. Individual privacy is respected and private information is secured in government records.

Generalized Inferences

Data is only from NYC, thus data will reflect mostly the trends present in said state. Rural neighborhoods may not be represented as much as urban ones due to lack of resources installed there to record data on.

Generalized Inferences - Mitigation

Data should only be used to make inferences on the mentioned Boroughs within New York.

Sensitive Content

The data might incorrectly or disproportionally represent a certain demographic group in a bad light.

Documented Known Issues

https://catalog.data.gov/dataset

Other Known Issues

The data frame only encapsulate, on the grand scale, a small time frame around only a decade of year, thus the data may not be representative of the actual causal or effect between air quality and crime.[2]