Beyond the Numbers: Dissecting New York City's Leading Causes of Death Across Demographics*

A demographics-based analysis of morbditiy in the Big Apple

Rayan Awad Alim Maria Mangru MD Mubtasim-Fuad

March 16, 2024

We analysed the leading causes of death in New York City, segmented by sex and ethnicity, from 2007 to 2014. Utilizing data derived from NYC death certificates, our research offers insights into public health trends, and the importance of understanding how these causes of death vary across race.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. **?@sec-data**….

2 Data

This analysis utilizes data from the New York City Department of Health and Mental Hygiene, specifically provided by the Bureau of Vital Statistics {New York: Department of Health and (DOHMH) (2023)}. The dataset encompasses records of deaths in New York City since 2007, detailing the leading causes of death categorized by sex and ethnicity. Each entry in this dataset is derived from NYC death certificates which are the official documentation of every death occurring within the city's jurisdiction.

*Code and data are available at: https://github.com/RayanAlim/Analysis-of-Morbidity-in-New-York

2.1 Dataset Context and Broader Implications

The investigation into the leading causes of death within New York City holds significant public health importance. It allows for a nuanced understanding of mortality trends, guiding health policy, and intervention strategies tailored to specific demographics and causes. By examining mortality across different sexes and ethnic groups, this analysis contributes to identifying health disparities and targeting efforts to address them.

2.2 Variables and Data Examination

The dataset comprises several key variables, including: - Year: The year of the recorded death.

- Leading Cause: The primary cause of death as categorized by ICD-10 codes.
- Sex: The sex of the deceased (Male, Female, or Gender Non-Conforming).
- Race Ethnicity: The self-reported ethnicity of the deceased, including categories such as Hispanic, White, Non-Hispanic, Black Non-Hispanic, Asian and Pacific Islander, and others.
- Deaths: The total number of deaths attributed to the leading cause.
- Death Rate and Age-Adjusted Death Rate: Rates per 100,000 population, providing standardized comparisons across different population sizes and age structures.

2.3 Alternative Datasets and Justification

While other mortality databases exist, such as the CDC's National Center for Health Statistics, the chosen dataset offers a detailed and localized perspective specific to New York City. This specificity provides a more precise tool for understanding and addressing urban health dynamics, making it more relevant for city-specific policy-making and health intervention strategies.

2.4 Data Cleaning and Variable Construction

Preliminary data cleaning focused on addressing missing values, ensuring consistency in categorical variables, and suppressing unreliable rates as mentioned. New variables, such as aggregated categories for cause of death or demographic groupings, were constructed to facilitate analysis. For instance, causes of death were grouped into broader categories (e.g., cardiovascular diseases, cancers) to examine trends at a macro level.

2.5 Summary Statistics and Relationships

Initial exploratory data analysis revealed key trends and disparities in mortality rates across different populations within New York City. We provide summary statistics, including mean death rates and distributions of deaths by cause ethnicity, to provide insights into the health landscape of the city. For example, preliminary findings indicate significant differences in heart disease mortality between ethnic groups, warranting further investigation.

2.6 Measurement and Methodology Notes

The measurement of mortality and its causes relies on the accurate classification of death certificates, adhering to ICD-10 standards. This classification ensures comparability with other datasets and robustness in identifying health trends. However, it's important to acknowledge potential limitations in cause-of-death reporting and classification, which may impact the analysis. To ensure privacy and reliability, rates based on small numbers (Relative Standard Error, RSE, > 30) and aggregate counts less than 5 have been suppressed. This suppression safeguards against the identification of individuals in rare categories and ensures statistical reliability.

3 Model

In our research, we investigated the causes of mortality within New York City's diverse racial and ethnic populations for the year 2014. We hypothesize that the total number of deaths by leading cause follows a negative binomial regression. Formally, the model is expressed as:

$y_i|n_i \sim \text{NegBin}(\mu_i, k)$

In the expression above, y_i, signifies the count of deaths for cause i,μ_i is the expected count, and k is the dispersion parameter, which captures over-dispersion. Our model also utilizes a log-link function, modeling the expected log-count as:

$\log(\mu_i) = \beta_0 + \beta_1 \times \text{Cause}_i$

Where β_0 is the intercept, and β_1 represents the vector of coefficients for the leading causes of death. The leading causes are treated as categorical predictors, which results in the need to use dummy coding with one reference category.

For our Bayesian regression framework, we employ the stan_glm() function from the rstanarm package, which assumes default normal priors for the regression coefficients as follows: $\beta_0 \sim$ Normal(0,2.5) $\beta_1 \sim$ Normal(0,2.5)

3.0.1 Model justification

The negative binomial regression was used due to its ability to account for the over-dispersion present in the data. This is a common occurrence with mortality counts due to the heterogeneity of death causes. Through the use of the parameter, we can model the variance separately from the mean which provides a more flexible and accurate representation of the data. This model also accommodates the count nature of the dependent variable. Furthermore, the model's priors are informed by the normative distributions typically used in Bayesian regression analysis. This enables a degree of regularization that reduces the potential for overfitting.

4 Results

Table 1: Count of Deaths by Leading Cause

No.	Cause of Death	Total Death Count
1	**Diseases of Heart (I00-I09, I11, I13, I20-I51)**	147551
2	**Malignant Neoplasms (Cancer: C00-C97)**	106367
3	**All Other Causes**	77999
4	**Influenza (Flu) and Pneumonia (J09-J18)**	18678
5	**Diabetes Mellitus (E10-E14)**	13794
6	**Chronic Lower Respiratory Diseases (J40-J47)**	13214
7	**Cerebrovascular Disease (Stroke: I60-I69)**	12941
8	**Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)**	7467
9	**Essential Hypertension and Renal Diseases (I10, I12)**	6955
10	**Human Immunodeficiency Virus Disease (HIV: B20-B24)**	5436

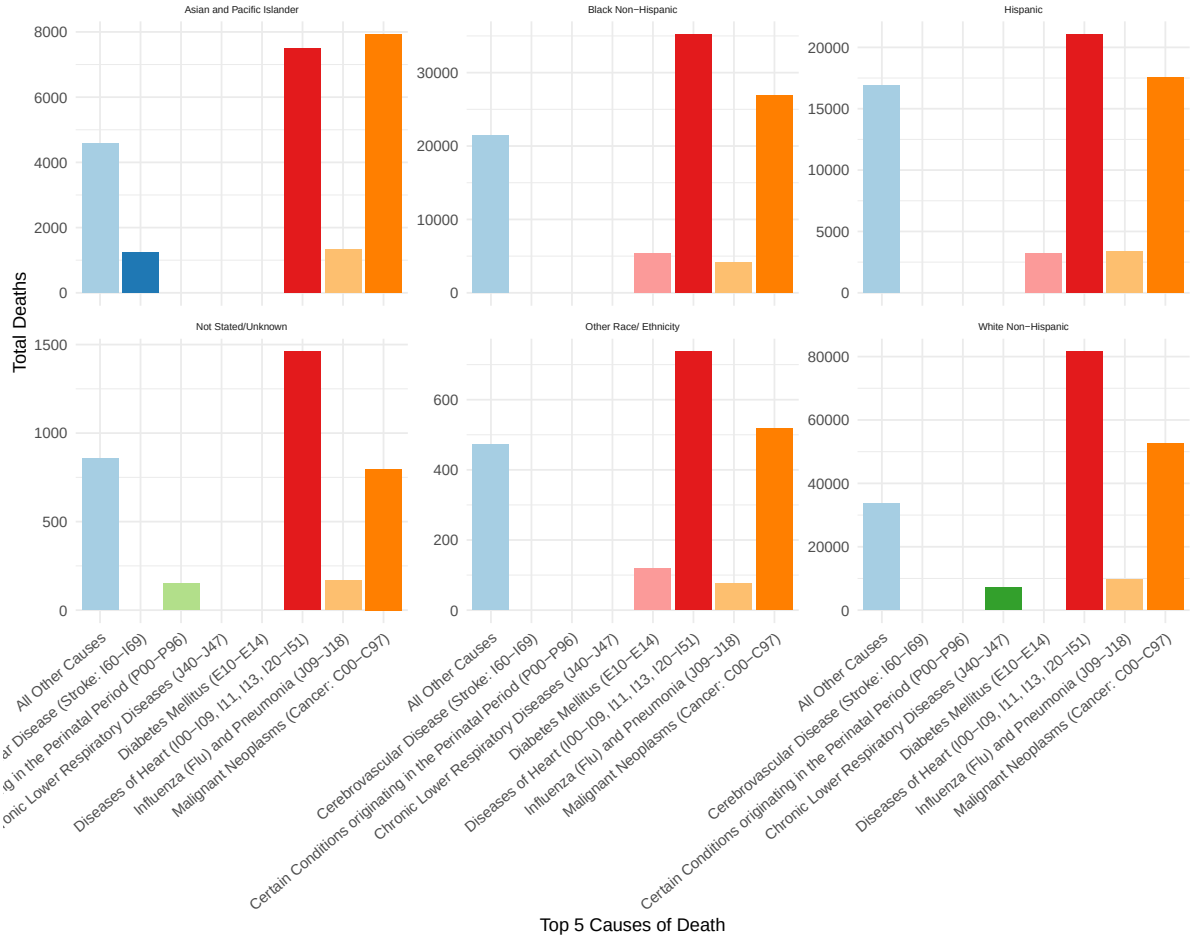


Figure 1: Top 5 Causes of Death by Race. This graph displays the leading causes of death, broken down by race/ethnicity

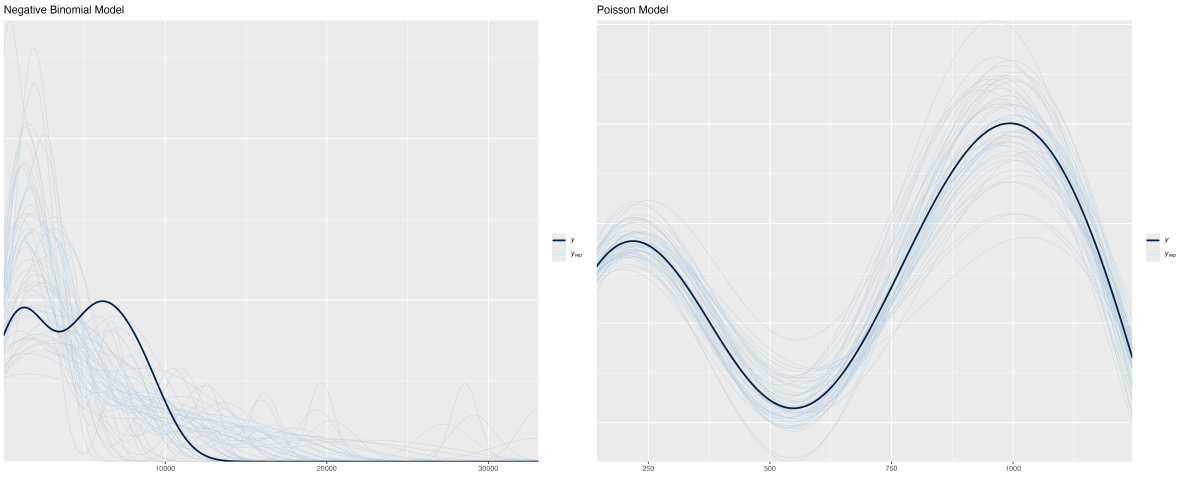
4.1 Model Results

For each population we built a negative binomial and possion regression the model and then use posterior predictive checks to compare the models. We also compare between the models using the resampling method leave-one-out (LOO) cross-validation (CV). The results for the models based on each population is as follows:

4.1.1 White Non-Hispanic Population

Posterior Predictive Check:

Leave-one-out (LOO) cross-validation (CV):



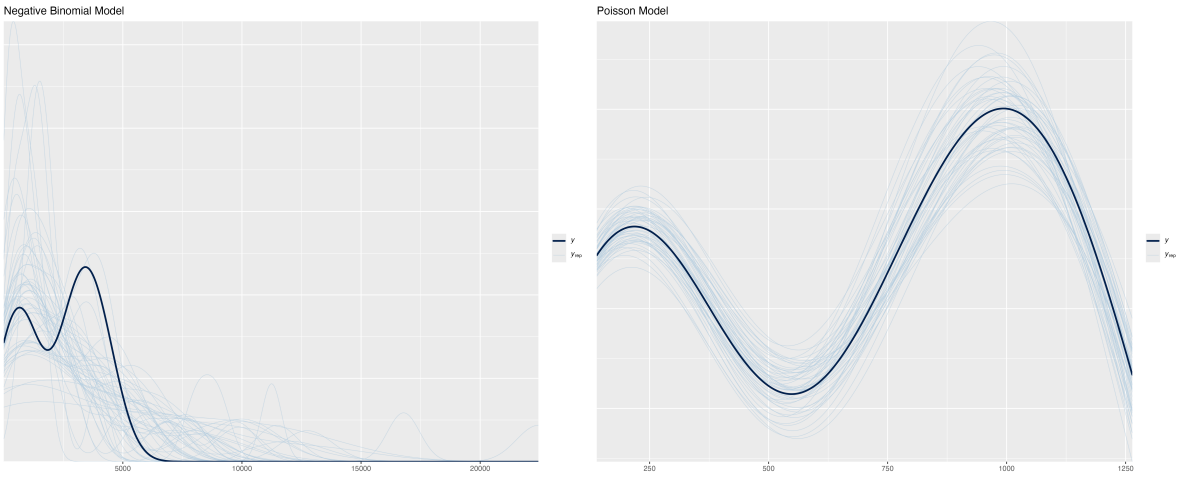
(a) Negative binomial model (b) Poisson model

Figure 2: Comparing posterior prediction checks for Poisson and negative binomial models in Leading cause of death of White Non-Hispanic Population in New York City

Model	elpd_diff	se_diff
Negative Binomial	-43.4	4.0
Poisson	0.0	0.0

4.1.2 Black Non Hispanic Population

Posterior Predictive Check:

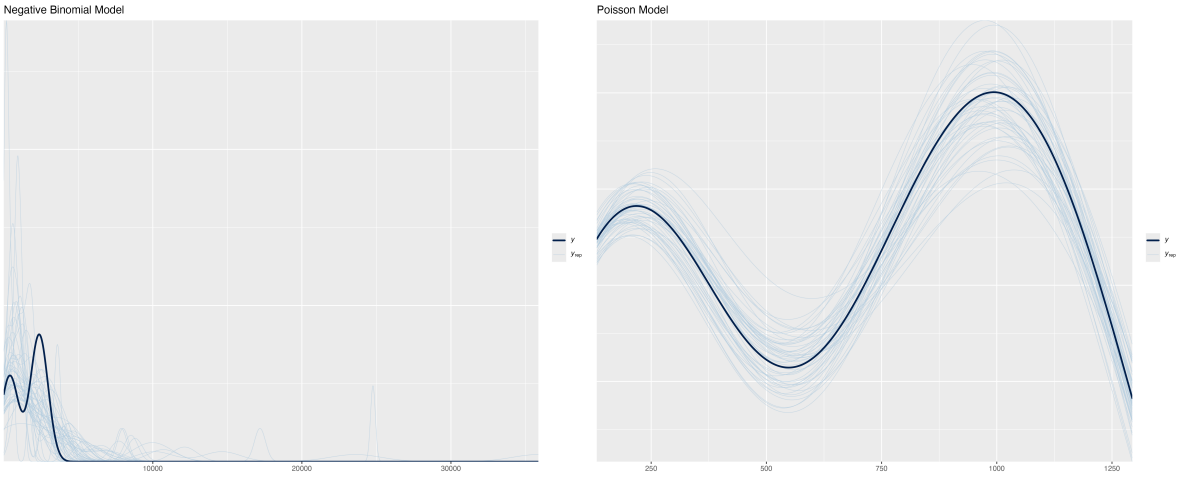


(a) Negative binomial model (b) Poisson model

Figure 3: Comparing posterior prediction checks for Poisson and negative binomial models in Leading cause of death of Black Non-Hispanic Population

Leave-one-out (LOO) cross-validation (CV):

Model	elpd_diff	se_diff
Negative Binomial	-37.6	3.0
Poisson	0.0	0.0



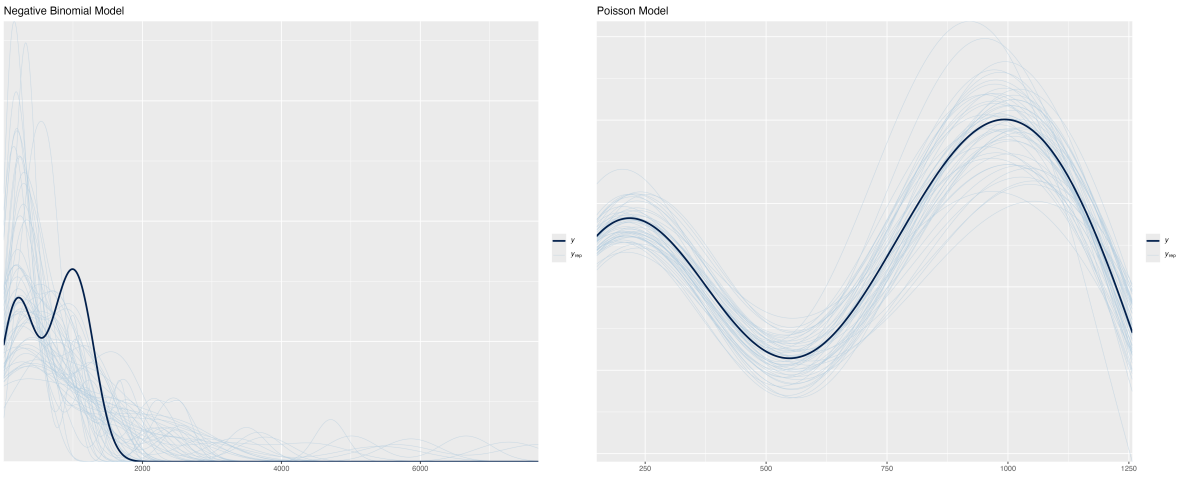
(a) Negative binomial model (b) Poisson model

Figure 4: Comparing posterior prediction checks for Poisson and negative binomial models in Leading cause of death of Hispanic Population in New York City

Model	elpd_diff	se_diff
Negative Binomial	-34.0	2.8
Poisson	0.0	0.0

4.1.4 Asian And Pacific Islanders Population

Posterior Predictive Check:



(a) Negative binomial model (b) Poisson model

Figure 5: Comparing posterior prediction checks for Poisson and negative binomial models in Leading cause of death of Asian And Pacific Islanders Population in New York City

Leave-one-out (LOO) cross-validation (CV): [7]

Model	elpd_diff	se_diff
Negative Binomial	-25.7	1.4
Poisson	0.0	0.0

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

C Datasheet

C.1 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
  - This dataset was compiled to facilitate the analysis of mortality patterns in New York City, providing insights into the leading causes of death across the general population and by race.
- Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
  - This dataset was created by a research team at the New York City Department of Health and Mental Hygiene.
- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
  - Funding was provided as part of the budget of the New York City Department of Health and Mental Hygiene.
- Any other comments?
  - No.

C.2 Composition

- What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  - Each row in the dataset represents an aggregate count of deaths for a specific leading cause, categorized by year, sex, and race/ethnicity.
- How many instances are there in total (of each type, if appropriate)?
  - Over 1000 instances.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
  - This dataset contains all recorded instances of death in New York City between 2007 - 2014.
- What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description. Each instance consists of counts of death including demographic data (race and sex) and year. Is there a label or target associated with each instance? If so, please provide a description.
  - No.
- Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.
  - Some instances may have missing data where the information was not recorded or reported.
- Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
  - Yes through demographic categorization (race and sex) and alignment.

- Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
  - No.
- Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
  - Concerns arise given the potential for inaccuracies in the reporting and classification of death causes.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
  - Self-contained.
- Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
  - Generated from public sources.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  - No.
- Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
  - Yes, sex is identified.
- Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
  - No.

- Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
  - This data does contain race origins and causes of death.
- Any other comments?
  - No.

C.3 Collection process

- How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
  - The data were gathered from death certificates and medical reports filed within New York City.
- What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
  - Data was collected through official city and health databases. R scripts were used to clean and aggregate the data.
- If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
  - Not a sample.
- Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?
  - New York City Department of Health and Mental Hygiene.
- Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
  - No.

- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
  - Collected from a third-party source.
- Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
  - No.
- Any other comments?
  - No.

C.4 Preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
  - Yes, data was cleaned to address missing or incomplete records.
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
  - Yes, both the "raw" and "cleaned" data were saved.
- Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
  - R was used.
- Any other comments?
  - No

C.5 Uses

- Has the dataset been used for any tasks already? If so, please provide a description.
  - Yes, the dataset has been used for annual health reports and studies on mortality trends within New York City.

- Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. - No.
- What (other) tasks could the dataset be used for?
  - The dataset could be used to predict future trends in public health.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
  - No.
- Are there tasks for which the dataset should not be used? If so, please provide a description.
  - Discrimination of insurance rates/policies based on race.
- Any other comments?
  - No.

C.6 Distribution

- Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
  - The dataset is available through the NYC Open Data Profile.
- When will the dataset be distributed?
  - The dataset is uploaded and distributed annually through the NYC Open Data Profile.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
  - None that are known.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  - None that are known.

- Any other comments?
  - No.

C.7 Maintenance

- Who will be supporting/hosting/maintaining the dataset?
  - NYC Department of Health and Mental Hygiene.
- How can the owner/curator/manager of the dataset be contacted (for example, email address)?
  - A request can be sent to the NYC Open Data Portal team through their website (https://opendata.cityofnewyork.us/engage/)
- Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
  - Yes, the dataset is updated annually.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
  - Contributions/suggestions can be made through the NYC Open Data portal.
- Any other comments?
  - No

C.8 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

New York: Department of Health, City of, and Mental Hygiene (DOHMH). 2023. "NYC Morbidity Data." https://data.cityofnewyork.us/Health/New-York-City-Leading-Causes-of-Death/jb7j-dtam/about_data.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.