LONG COVID RESEARCH DATA DOCUMENTATION

Introduction

Data is central to any research project, but where it comes from is just as important as what it contains. As an open-source research project, it is our priority to ensure that our data sources, processing practices, and related tools are clear and replicable. The Project Data Information contains guidelines for adding new data to the project. This documentation details where data came from originally, when it was collected, how it was processed, and what types of access are possible. The table of contents view in Google Documents is the most effective way to navigate through the document.

This template was adapted from Microsoft's Aether Data Documentation Template. More information about this template can be found here.

Project Data Information

Canonical Variable Names

This is a reference as to the preferred format for variable names within the dataset. New data added to the set should have variable names updated to match.

County

Example: `Adams`

Refers to a census-determined geographic area. Variable should be the name of the county as a string only, not "County, State" or "Adams County". If there are multiple states in a dataset, state should be contained in a separate variable.

Population

Example: `123000`

Refers to a numeric population for a given area. Should be expressed as an integer, as the whole number (i.e. not in millions).

Distribution & Access

1. How can dataset users receive information if this dataset is updated (e.g., corrections, additions, removals)?

To be filled out

2. For static datasets: What will happen to older versions of the dataset? Will they continue to be maintained?

To be filled out

censusdata.csv

GENERAL INFORMATION

1. Name: US Census Data
2. Version number or date: July 1, 2021 (accessed Jan 24, 2023)
3. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/RHI725221
4. Access: Anyone
5. How can the dataset be accessed? Visiting the link, or accessing the internal S3 bucket until it is moved to the public repository.

CONTENTS

6. This dataset contains Latino population estimates for each county in eastern Washington. There are 20 rows, each representing one county. The data has 4 variables: County, Population Estimate, Pct Latino, and Estimated Latino Pop.

7. This data should be fairly representative of the population of eastern Washington. It comes from the US census, which aims to accurately understand the demographics of the US.

The context of the data is that collected from regular people by government officials, which is likely to direct responses toward certain values.

Limitations include the fact that some people, particularly those in rural settings, may not have access to the census for reasons of language or location barriers. Additionally, those with access may not wish to participate in it due to mistrust of the government, concerns about data privacy, or other reasons.

8. The demographic group of Latino is self-identified in this data set.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was manually inputted from the above-listed URL into a spreadsheet, which was then exported to a CSV file.

10. Missing information: There is no missing information in this dataset.

11. What data might be out of date: Population data changes rapidly, so as this data set gets older it will become more out of date.

12. Validity issues: Census data is manually collected, so there may have been gaps in the collection process that has led to inaccurate data about the population of Latinos in eastern WA. Still, this is likely the source that is closest to the truth.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Features were selected for their applicability to our research.

14. Who did the pre-processing, cleaning, and/or labeling: Raina Scherer.

15. Provide a link to the code used to preprocess/clean/label the data, if available: No code used.

PRIVACY AND PII

16. Is it possible to identify individuals either directly or indirectly (i.e., in combination with other data) from the dataset, OR, does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

No, the data is completely anonymous and there is no way to link it to a specific person. It is also not sensitive despite providing information related to ethnicity, as it can't be linked to a specific person.

HealthcareMetricsWA-cleaned.csv

GENERAL INFORMATION

6. Name: Healthcare Metrics Washington
7. Version number or date: Feb 03, 2023 (accessed Feb 03, 2023)
8. Place of origin: https://doh.wa.gov/emergencies/covid-19/data-dashboard#tables
9. Access: Anyone
10. How can the dataset be accessed? Visiting the link, or accessing the internal S3 bucket until it is moved to the public repository.

CONTENTS

6. This dataset contains hospital admission rates for Washington State over time. It contains a date range, 7-day admission count, 7-day admission rate per 100,000, and the population at the time of the date range. There are 858 rows of data.

7. This data should be representative of the hospital admission rates for Washington in the given time frame. The data is compiled by the Washington State Department of Health, who have primary access to these figures from the hospitals themselves.

The context of the data is that it represents a view of the Washington Hospital system for a given time.

Limitations may be that older hospitals or overworked employees may not be able to fully record hospital admissions, but this is unlikely given the high degree of regulation surrounding the healthcare system.

8. No demographic groups are identified within this data set, as it only reports a single value for admissions, with no disaggregation.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was downloaded from the above-listed URL as a XLSX file.

10. Missing information: There is no missing information in this dataset.

11. What data might be out of date: Hospital admission data changes weekly, but given that this data only represents the time frames it lists, it will never be out of date.

12. Validity issues: As these values are reported directly by the hospitals, there should not be any issues of validity. There could be an issue if hospitals were not reporting accurate numbers

due to the pressures of the COVID-19 pandemic, but this is difficult if not impossible to ascertain.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above, and moved whole to the database. The only processing done was conversion to a CSV file.

14. Who did the pre-processing, cleaning, and/or labeling: Raina Scherer.

15. Provide a link to the code used to preprocess/clean/label the data, if available: N/A

PRIVACY AND PII

16. Is it possible to identify individuals either directly or indirectly (i.e., in combination with other data) from the dataset, OR, does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

No, the data is completely anonymous and there is no way to link it to a specific person.

cases-cleaned.csv

GENERAL INFORMATION

11. Name: COVID-19 Cases Washington
12. Version number or date: Feb 03, 2023 (accessed Feb 03, 2023)
13. Place of origin: https://doh.wa.gov/emergencies/covid-19/data-dashboard#tables
14. Access: Anyone
15. How can the dataset be accessed? Visiting the link, or accessing the internal S3 bucket until it is moved to the public repository.

CONTENTS

6. This dataset contains data about COVID-19 cases in Washington state. It covers: Earliest Specimen Collection Date, County, Total Cases, Confirmed Cases, Probable Cases, Total Cases (7-Day Average), 7-Day Case Count, 7-Day Case Rate, 14-Day Case Count, 14-Day Case Rate, and Date/Time Updated. There are 63726 rows of data.

7. This data should be representative of the COVID-19 rates for Washington in the given time frame. The data is compiled by the Washington State Department of Health, who have had the best possible access to testing data.

The data was collected in the context of an emerging pandemic, where high-qualty data was needed to assess danger to residents.

Limitations are substantial. COVID-19 testing was not always available in all areas, particularly rural ones. There may have also been issues with government trust that prevented people from wanting to access testing resources. However, this is still the best data source on the subject, and represents the most comprehensive view of the situation.

8. No demographic groups are identified within this data set, as it only reports aggregated case counts, with no disaggregation factors.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was downloaded from the above-listed URL as a XLSX file.

10. Missing information: This dataset is nearly fully complete, with a few missing data:

- Total Cases (7-Day Average): 6
- 7-Day Case Count: 6
- 7-Day Case Rate: 3360
- 14-Day Case Count: 13
- 14-Day Case Rate: 3367

11. What data might be out of date: Case counts change daily, but given that this data only represents the time frames it lists, it will never be out of date.

12. Validity issues: As these values are reported directly by the department of health, there should not be issues of invalid data. The previously-discussed issues of incompleteness still remain, but nothing here should be invalid.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. The processing was minimal: the format was changed to CSV, and the county variables were formatted to match the existing data structure.

14. Who did the pre-processing, cleaning, and/or labeling: Raina Scherer.

15. Provide a link to the code used to preprocess/clean/label the data, if available: [ADD GITHUB LINK TO DATA CLEANING NOTEBOOK]

PRIVACY AND PII

16. Is it possible to identify individuals either directly or indirectly (i.e., in combination with other data) from the dataset, OR, does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

No, the data is completely anonymous and there is no way to link it to a specific person.

OccupationDataWA-MSA-cleaned.csv

GENERAL INFORMATION

16. Name: Occupation Data Eastern Washington (Metropolitan Statistical Areas)
17. Version number or date: May 01, 2021 (accessed Feb 03, 2023)
18. Place of origin: https://www.bls.gov/oes/current/oes_5300007.htm
19. Access: Anyone
20. How can the dataset be accessed? Visiting the link, or accessing the internal S3 bucket until it is moved to the public repository.

CONTENTS

6. This dataset contains data about the occupations of residents of Eastern Washington state.

7. This data should be representative of the prevalence of various occupations in Eastern Washington. It was collected by the U.S. Bureau of Labor Statistics, who would have the access needed to collect representative and accurate data.

The data was collected in the context of a government agency intending to update its overview of the state's occupational data.

Limitations are possible. The BLS website indicates that it was collected by asking employers to report their own data. It's possible some employers were not contacted, or that some did not report accurate numbers either due to concerns in reporting or due to a lack of care.

8. No demographic groups are identified within this data set, as it only reports aggregated occupation numbers, with no disaggregation factors.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was downloaded from the above-listed URL as a XLSX file.

10. Missing information: This dataset is nearly fully complete, with a few missing data:

- Total Cases (7-Day Average): 6
- 7-Day Case Count: 6
- 7-Day Case Rate: 3360
- 14-Day Case Count: 13
- 14-Day Case Rate: 3367

11. What data might be out of date: Case counts change daily, but given that this data only represents the time frames it lists, it will never be out of date.

12. Validity issues: As these values are reported directly by the department of health, there should not be issues of invalid data. The previously-discussed issues of incompleteness still remain, but nothing here should be invalid.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was minimal: removed PCT_TOTAL, PCT_RPT, and HOURLY columns as they were entirely blank.

14. Who did the pre-processing, cleaning, and/or labeling: Raina Scherer.

15. Provide a link to the code used to preprocess/clean/label the data, if available: [ADD GITHUB LINK TO DATA CLEANING NOTEBOOK]

PRIVACY AND PII

16. Is it possible to identify individuals either directly or indirectly (i.e., in combination with other data) from the dataset, OR, does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

No, the data is completely anonymous and there is no way to link it to a specific person. There do not seem to be any occupations only held by a single person.

TEMPLATE

DATASET NAME

GENERAL INFORMATION

21. Dataset name
22. Dataset version number or date
23. Dataset place of origin
24. Who can access this dataset (e.g., team only, internal to the company, external to the company)?
25. How can the dataset be accessed?

CONTENTS

6. What are the contents of this dataset? Please include enough detail that someone unfamiliar with the dataset who might want to use it can understand what is in the dataset. Specifically, be sure to include:

- What does each item/data point represent (e.g., a document, a photo, a person, a country)?
- How many items are in the dataset?
- What data is available about each item (e.g., if the item is a person, available data might include age, gender, device usage, etc.)? Is it raw data (e.g., unprocessed text or images) or features (variables)?
- *For static datasets:* What timeframe does the dataset cover (e.g., tweets from January 2010–December 2020)?

7. How representative is this dataset?

- What population(s), contexts (e.g., scripted vs. conversational speech), conditions (e.g., lighting for images) is it representative of?
- How was representativeness ensured or validated?
- What are known limits to this dataset's representativeness?

8. What demographic groups (e.g., gender, race, age,etc.) are identified in the dataset, if any?

- How were these demographic groups identified (e.g., self-identified, inferred)?
- What is the breakdown of the dataset across demographic groups?
- Consider also reporting intersectional groups (e.g., race x gender) and including proportions, counts, means or other relevant summary statistics.

DATA COLLECTION AND QUALITY

9. How was the data collected (download, scraper)?

10. Is there any missing information in the dataset? If yes, please explain what information is missing and why (e.g., some people did not report their gender).

11. What data might be out of date or no longer available (e.g., broken links in old tweets)?

12. What are potential validity issues a user of this dataset needs to be aware of (e.g., survey answers might not be truthful, age was guessed by a model and might be incorrect, GPA was used to quantify intelligence)?

PROCESSING, CLEANING, AND LABELING

13. What pre-processing, cleaning, and/or labeling was done on this dataset?

- Include information such as: how labels were obtained, treatment of missing values, grouping data into categories (e.g., was gender treated as a binary variable?), and dropping data points.

14. Who did the pre-processing, cleaning, and/or labeling (e.g., were crowd workers involved in labeling?)

15. Provide a link to the code used to preprocess/clean/label the data, if available.

PRIVACY AND PII

16. Is it possible to identify individuals either directly or indirectly (i.e., in combination with other data) from the dataset, OR, does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

- If the answer to either of these questions is yes, please be sure to follow guidelines on handling sensitive data.