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. A 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, the 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)?

Any questions pertaining to future versions of this dataset should be directed to research@adioscovid.org.

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

There are no current plans to remove older versions of this dataset. For any questions, please reach out to research@adioscovid.org.

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.

This data was collected in the context of a government agency attempting to better understand its population.

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 Latinos 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 Washington. 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.

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, which has 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 an 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 some did not report accurate numbers due to reporting concerns or 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 an XLSX file.

10. Missing information: This dataset is nearly fully complete, however, the “ANNUAL” feature is missing data for nearly every row:

- ANNUAL: 3559

This is a flag that identifies a given row as an annual salary, as opposed to an hourly rate. These rows will also have empty (but not null) values for fields specific to hourly positions.

11. What data might be out of date: Occupation data does shift over time, but this data set will remain accurate to the time it was collected.

12. Validity issues: As these values are reported directly by the Bureau of Labor Statistics, there should not be any validity issues.

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.

OccupationDataWA-NonMet-cleaned.csv

GENERAL INFORMATION

21. Name: Occupation Data Eastern Washington (Non-Metropolitan Statistical Areas)
22. Version number or date: May 01, 2021 (accessed Feb 03, 2023)
23. Place of origin: https://www.bls.gov/oes/current/oes_5300007.htm
24. Access: Anyone
25. 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 an XLSX file.

10. Missing information: This dataset is nearly fully complete, however, the "ANNUAL" feature is missing data for nearly every row:

- ANNUAL: 705

This is a flag that identifies a given row as an annual salary, as opposed to an hourly rate. These rows will also have empty (but not null) values for fields specific to hourly positions.

11. What data might be out of date: Occupation data does shift over time, but this data set will remain accurate to the time it was collected.

12. Validity issues: As these values are reported directly by the Bureau of Labor Statistics, there should not be any validity issues.

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.

EasternWaIncome-cleaned.csv

GENERAL INFORMATION

26. Name: Eastern Washington Income Data
27. Version number or date: July 01, 2021 (accessed Feb 01, 2023)
28. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/PST045222?
29. Access: Anyone
30. 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 income of residents of Eastern Washington.

7. This data should be representative of how income brackets in Eastern Washington are distributed within this demographic group. It was collected by the U.S. Census, which 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 income data.

Limitations are possible. The Census, as a government agency, may not adequately meet the needs of all Americans, and some people may not wish to answer personal questions such as income level.

8. No demographic groups are identified within this data set, as it only reports aggregated income brackets per county, with no disaggregation factors.

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: None.

11. What data might be out of date: Income changes over time, but the data will remain accurate to the time period it was collected in.

12. Validity issues: As these values are reported directly by the Census, there should not be any validity issues.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was relatively straightforward and mainly consisted of transposing the data to be consistent with the other data sets, and renaming columns to match our standard.

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 income brackets only held by a single person.

LatinoHealthInsuranceCoverage.csv

GENERAL INFORMATION

31. Name: Eastern Washington Health Insurance Coverage for Latinos
32. Version number or date: July 01, 2021 (accessed Feb 01, 2023)
33. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/PST045222?
34. Access: Anyone
35. 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 health insurance coverage of Latinos in Eastern Washington.

7. This data should be fairly representative of the health insurance status of Latino people in Eastern Washington. It was collected by the U.S. Census, which 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 income data.

Limitations are possible. The Census, as a government agency, may not adequately meet the needs of all Americans, and some people may not wish to answer personal questions such as income level. Additionally, there may have been language barriers or other cultural factors that led to an inadequate data collection process.

8. The demographic group of Latino is specifically 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: None.

11. What data might be out of date: Health insurance is a difficult subject in the US, and could change rapidly between administrations who may hold different policies, or during health crises like the COVID-19 pandemic. This data could possibly become quickly out of date due to these factors.

12. Validity issues: As these values are reported directly by the Census, there should not be any validity issues, though there are potential issues in the collection process which have been previously discussed.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was fairly involved for this data, as the format did not align well with the existing data sets. Each age bracket was collapsed into one row as the original format did not allow for meaningful comparisons to be made, as the rows did not have any links to each other. Counties were renamed to match other data sets and added a new column to indicate health insurance status.

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)?

This data does not contain any individually-identifying information, but might still be considered somewhat sensitive due to the combination of insurance information and ethnicity.

LatinoHouseholdTypeEasternWA-cleaned.csv

GENERAL INFORMATION

36. Name: Eastern Washington Household Members for Latinos
37. Version number or date: July 01, 2021 (accessed Feb 01, 2023)
38. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/PST045222?
39. Access: Anyone
40. 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 members of households for Latinos in Eastern Washington.

7. This data should be fairly representative of the household composition of Latino people in Eastern Washington. It was collected by the U.S. Census, which 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 income data.

Limitations are not only possible but likely. The Census, as a government agency, may not adequately meet the needs of all Americans, and some people may not wish to answer personal questions such as household status. Additionally, there may have been language barriers or other cultural factors that led to an inadequate data collection process. The questionnaire may not have been adequately informed by the cultural needs and preferences of Latino people, leading to inaccurate data.

8. The demographic group of Latino is specifically 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: None.

11. What data might be out of date: Household composition can rapidly shift due to various socioeconomic factors, such as health crises like the COVID-19 pandemic. This data could possibly become quickly out of date due to these factors.

12. Validity issues: As these values are reported directly by the Census, there should not be any validity issues, though there are potential issues in the collection process which have been previously discussed.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was fairly involved. The aggregation columns as they were very unclear due to labels such as “family”, “non-family”, and “other family”. Transposed dataset to a longer format that matched the indexing of other data sets, and finally adjusted column names to match existing sets as well.

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)?

This data does not contain any individually-identifying information, but might still be considered somewhat sensitive due to the combination of household composition and ethnicity.

WorkTransportLatinoEasternWA-cleaned.csv

GENERAL INFORMATION

41. Name: Eastern Washington Work Transportation for Latinos
42. Version number or date: July 01, 2021 (accessed Feb 01, 2023)
43. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/PST045222?
44. Access: Anyone
45. 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 transportation methods used to access employment for Latinos in Eastern Washington.

7. This data should be fairly representative of the methods of transportation used to access employment by Latino people in Eastern Washington. It was collected by the U.S. Census, which 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 transportation data.

Limitations are not only possible but likely. The Census, as a government agency, may not adequately meet the needs of all Americans, and some people may not wish to answer personal questions such as household status. Additionally, there may have been language barriers or other cultural factors that led to an inadequate data collection process. The questionnaire may not have been adequately informed by the cultural needs and preferences of Latino people, leading to inaccurate data.

8. The demographic group of Latino is specifically 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: None.

11. What data might be out of date: Transportation methods are likely to remain stagnant in most decades, but can rapidly shift due to various socioeconomic factors, such as health crises like the COVID-19 pandemic or new availability of carpool options. This data could possibly become quickly out of date due to these factors.

12. Validity issues: As these values are reported directly by the Census, there should not be any validity issues, though there are potential issues in the collection process which have been previously discussed.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was minimal. The dataset was transposed to have indexing matching the preferred format, and column names were adjusted.

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)?

This data does not contain any individually-identifying information. It does contain demographic information, but the axis of work transportation is not in any way sensitive.

LatinoHouseholdTypeEasternWA-cleaned.csv

GENERAL INFORMATION

46. Name: Eastern Washington Household Members for Latinos
47. Version number or date: July 01, 2021 (accessed Feb 01, 2023)
48. Place of origin: https://www.census.gov/quickfacts/fact/table/WA/PST045222?
49. Access: Anyone
50. 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 members of households for Latinos in Eastern Washington.

7. This data should be fairly representative of the household composition of Latino people in Eastern Washington. It was collected by the U.S. Census, which 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 income data.

Limitations are not only possible but likely. The Census, as a government agency, may not adequately meet the needs of all Americans, and some people may not wish to answer personal questions such as household status. Additionally, there may have been language barriers or other cultural factors that led to an inadequate data collection process. The questionnaire may not have been adequately informed by the cultural needs and preferences of Latino people, leading to inaccurate data.

8. The demographic group of Latino is specifically 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: None.

11. What data might be out of date: Household composition can rapidly shift due to various socioeconomic factors, such as health crises like the COVID-19 pandemic. This data could possibly become quickly out of date due to these factors.

12. Validity issues: As these values are reported directly by the Census, there should not be any validity issues, though there are potential issues in the collection process which have been previously discussed.

PROCESSING, CLEANING, AND LABELING

13. This data was manually collected from the URL above. Processing was fairly involved. The aggregation columns as they were very unclear due to labels such as “family”, “non-family”, and “other family”. Transposed dataset to a longer format that matched the indexing of other data sets, and finally adjusted column names to match existing sets as well.

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)?

This data does not contain any individually-identifying information, but might still be considered somewhat sensitive due to the combination of household composition and ethnicity.

GovernmentTrustBy*.csv

GENERAL INFORMATION

51. Name: Government Trust By (Party Affiliation, President, Race and Ethnicity, Overall)
52. Version number or date: June 06, 2022 (accessed Feb 17, 2023)
53. Place of origin: Pew Research Center
54. Access: Anyone
55. 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. These datasets contains data about Americans' general trust in government. These files were all processed the same way and scraped from the same source, so the data documentation has been gathered together for brevity.

7. This data should be fairly representative of the general trust in government across various demographic groups. The Pew Research Center is a notable agency, and the sources they have listed are all high-quality: Pew Research Center, National Election Studies, Gallup, ABC/Washington Post, CBS/New York Times, and CNN Polls.

The data was collected in the context of a research group attempting to assess the general public's trust in government.

Limitations are extensive. These polls may not have reached a diverse group of Americans, as many of the news organizations in question tend to have similar audiences in terms of ethnicity, education level, and socioeconomic status. They themselves include the caveat that the responses for Asian-Americans are only representative of the English-speaking section of the population. Finally, there is a strong self-selection bias in who will respond to a survey about trust in government. These factors, in combination, can lead to a strong limitation in how representative this data is.

8. Race and ethnicity is used as a feature in one dataset.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was scraped using a python script above-listed URL into a spreadsheet, which was then exported to a CSV file.

10. Missing information: In the "GovernmentTrustByRaceandEthnicity" file, there are a small number of missing data (hispanic: 125, black: 46, asian: 146). Each other file is complete.

11. What data might be out of date: Trust in government is a fast-shifting metric. It can become inaccurate days after it's collected due to policy changes, statements made by politicians. This data should only be considered accurate to the date it was collected.

12. Validity issues: As these values are reported directly by a notable non-partisan organization, there should not be any validity issues, though there are potential issues in the collection process which have been previously discussed.

PROCESSING, CLEANING, AND LABELING

13. This data was scraped from the URL above. No processing was needed due to the collection process.

14. Who did the pre-processing, cleaning, and/or labeling: Jacqueline Hsu

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

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)?

This data does not contain any individually-identifying information. It does contain demographic and political leaning information, which is sensitive in and of itself.

Physicians_County_WA_Cleaned.csv

GENERAL INFORMATION

56. Name: Physicians per County in Washington
57. Version number or date: November 2020 (accessed Mar 07, 2023)
58. Place of origin: Office of Financial Management Health Care Research Center
59. Access: Anyone
60. 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 information about the number of physicians for each county in Washington.

7. This data should be representative of the number of physicians for each county. The sources used, such as physician license databases, are reputable and complete.

The data was collected in the context of a government agency attempting to establish important healthcare metrics.

Limitations are not likely. Physicians are held to high licensing standards, so it is unlikely that the data is incomplete.

8. No demographic data appears in the dataset.

DATA COLLECTION AND QUALITY

9. How the data was collected: The data was downloaded from the given URL.

10. Missing information: None.

11. What data might be out of date: This data comes from November 2020, after nearly one full year of the COVID-19 pandemic. This would have had significant impacts on the medical community, and there were likely unpredictable shifts in the physician numbers since then.

12. Validity issues: As these values are reported directly through licensing databases, validity concerns are minimal.

PROCESSING, CLEANING, AND LABELING

13. This data was minimally processed. There were two empty columns left as artifacts, which were removed.

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 SCRIPT]

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)?

This data does not contain any individually-identifying information or sensitive data.

TEMPLATE

DATASET NAME

GENERAL INFORMATION

61. Dataset name
62. Dataset version number or date
63. Dataset place of origin
64. Who can access this dataset (e.g., team only, internal to the company, external to the company)?
65. 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.