Kerah Iyall
LIS 545
27 January 2023
Word Count: 821

Term Project - Final Submission

GitHub Repository: https://github.com/Kerahiyall/Term-Project-

Dataset Profile

Dataset: https://www.eeoc.gov/sites/default/files/2021-12/EEO1%202017%20PUF.xlsx

Appendix: https://www.eeoc.gov/sites/default/files/2020-12/EEOC%20Explore%20FAQs.pdf

Data

I decided to focus my data profile on the US Equal Employment Opportunity Commission. Specifically, I centered my profile on the dataset for Job Patterns for Minorities and Women in Private Industry (EEO-1) Public Use File 2017. The data is extensive and covers numerous demographic areas that impact minority and women employees in private industry. The data includes the sex of the employee, their geographic location, their race, their job category, and the NAICS code (which further narrows their job industry). The geographic location includes the nation, region, division, state, CBSA, and county. An example of this for a specific data entry would be: United States (Nation), West (Region), Pacific (Division), Washington (State), Seattle-Tacoma-Bellevue (CBSA), King (County). Racial data is divided into six categories; White, Black, Hispanic, Asian, American Indian or Alaskan Native, and Native Hawaiian or Pacific Islander. There is an option to select two or more races.

This dataset is comprised of information that is reported to the Equal Employment Opportunity Commision by the employers that indicate the composition of their employees. Specifically, the EEO-1 2017 PUF is gathered from employers of private industry and business who have 100 or more employees.

There are many potential stakeholders for the data. Employers who are reporting their data have an investment in the composition of their employees in order to meet Equal Opportunity guidelines and foster diverse workplaces. Additionally, employees and potential employees of these businesses have an interest in knowing the demographics of their workplace. Equal Opportunity organizations, nonprofits, and federal agencies are also stakeholders in this data as they design initiatives to increase diversity and job accessibility in the workplace.

There is a file for each year from 1996-2020. The files contain the data collected by employers for that year. The data collected is indicative of the fields mentioned above, and are separated by the year. They are in a zip file format and downloadable for public use. There does not appear to be any usage restrictions. The files download as Microsoft Excel sheets from the zip files, it is easiest to view them as a Google Sheet or Microsoft Excel format.

Metadata

The data comes with several metadata fields. For each demographic area (race, sex, job category, geographic location, etc.) there is a corresponding metadata field that denotes the entry for that employee. The metadata is in multiple places as it is the column and column labels. The variable labels are as follows: NATION, REGION, DIVISION, STATE, CBSA, COUNTY, NAICS2, NAICS2_Name, NAICS3, NAICS3_Name. The metadata fields for the racial demographic information have corresponding column names and labels that are abbreviated versions of the racial identities. For example, a person who is a Hispanic female working as a senior official or manager, the metadata would be “HISPF1”. “HISP” refers to Hispanic, “F” for female, “1” referring to the job category. For each year, the metadata is in the same file. The metadata is quite extensive and covers the entries for each demographic field. Different combinations of the metadata fields illustrate whether an individual is mixed race or working a job not included in the job categories through the “aggregate fields”. The metadata encompasses nearly all possible entries for each field.

There was a specific methodology to the metadata structure. According to the appendix, the metadata were aggregated by geographic location, job categories, and industry codes over race/sex demographics. The metadata schema begins with the race entry, sex, then the job category. The geographic location entries are in a separate column. The metadata in this data set is descriptive as it is providing information for each demographic area. With the data is a PUF User Guide. The guide describes the variable labels and how to understand the names. As the variables combine with one another, the guide is helpful in illustrating what variables are included.

To enrich the data, it could be helpful to include age, disability status, and gender identity. These could be pertinent information fields to Equal Opportunity employers and federal agencies. For the current data, it could be useful to include the geographic location with the metadata schema for the racial and job entries so if someone wanted to quickly find the total hispanic employee population within certain geographic location, that could be easily done.

The EEO-1 was used to inform the publication *Evaluation of Compensation Data Collected Through the EEO-1 Form.* This publication used the datasets to draw conclusions regarding various disparities amongst minority and women employees. There are no publications referenced within the dataset itself, although the purpose of this dataset is to support diversity and Equal Opportunity efforts. Through my searches, I was unable to find publications that referenced the specific dataset I used, although there were several that referenced the EEO-1 at large.

Repository Profile

Repository: https://www.icpsr.umich.edu/web/pages/RCMD/index.html

Policies & Procedures

I chose the Resource Center for Minority Data (RCMD) as my repository because I felt it complimented the data set I chose for the first portion of this project. My data set from the US

Equal Employment Opportunity Commission was centered on minority and women employees in private industry. This repository has the potential to provide expansion and insight on the data set. The RCMD seeks to assist in the public dissemination and preservation of quality data to generate contributions to scientific research and study. Further, the RCMD was created to participate in a community of professionals who are interested and involved in minority related issues and inquiries in order to conduct the broadest scope of research endeavors and examinations. The RCMD is a fantastic repository for equality and ethics based research.

The RCMD is open for accepting data deposits from the public. There are procedures to ensure the data is prepared for deposit. The RCMD encourages contributors to follow their protocols which include the submission of an acceptable file type, omission of certain identifiers to preserve confidentiality, and proper documentation and description of the data. Preparing the data deposit maintains the integrity of the collection. SAS, SPSS, or Stata files are encouraged, however, ASCII files are allowed if there are data definition statements. The RCMD provides details for how metadata should be organized. Each variable in the data collection should have a set of exhaustive, mutually-exclusive codes. Variable labels and value labels should clearly describe the information or question recorded in that variable. When applicable, all identifying information should be removed from the records to ensure confidentiality. There is no requirement for a metadata structure or standard for the data to be deposited, appropriate labels and mutually-exclusive codes are emphasized.

The RCMD is quite open to accepting various data sets, the institution provides details as to how the data should be presented for each file type. For each type of plain text file - column, comma, and tab-delimited, the RCMD explains how the files are imported for analysis. For column and comma plain text files, the RCMD uses setups to read them into SAS, SPSS, and Stata files. The setups assign variable labels. For comma delimited files, the values are separated with commas, but the RCMD uses setups as well. For tab-delimited files, the files are usually easily imported to Excel and rarely require setups. Files can also be converted to “R” due to its extensive, open-source language. ICPSR data is easily converted to R. SAS datasets can be read by any SAS command, they are easily distributed by ICPSR. SPSS data is distributed in two forms: SPSS SAV files written by the SPSS save command and SPSS portable files written by the SPSS export command. They require two different commands to be loaded. SPSS SAV files into SPSS use the SPSS get command, and SPSS portable files into SPSS use the SPSS import command. Stata files utilize the Stata use command, as they are platform independent.

After data is deposited, it is reviewed by an archivist who performs quality-checks, builds a study description, archives the data for long-term preservation, and approves the data for distribution in the repository. Although there are no exemptions clearly indicated that would disqualify a data deposit from being approved, it is implied that only data relating to the mission of the RCMD will be included. The data must be centered on contributing to minority and race/ethnicity based inferences.

The RCMD is a private, secure repository. A login is required to access and contribute to the collection. To create an account, users must provide their personal information as well as an institution affiliation. Users are not required to be members of partner institutions, however, some data is only available to member affiliates. The full repository is available for members of the Institution for Social Research at the University of Michigan (ICPSR) and their partner

institutions. For ICPSR members from an affiliate institution, a campus Official Representative (OR) serves as a point of contact between the ICPSR and its users.

The majority of the data downloads as a zip file for the user, even non-members are able to download zip files. Members are able to download the data files as other file types, as well as using online analysis tools. Affiliate members are also able to generate utilization reports and download statistics. It is very beneficial as a user to utilize the repository through a partner institution.

There does not appear to be a specific metadata structure. The metadata fields are subject, series, geography, format, type of analysis, time period, restriction type, mode of collection, object type, archive, thematic collection, and investigator affiliation. Each title has a corresponding ICPSR number.

Additional Information

For the data from the EEO-1, I would follow the recommended data citation from DataCite as follows:

United States Equal Employment Opportunity Commission (2022-10-17). Job Patterns for Minorities and Women in Private Industry (EEO-1): Count of Employees by Race/Ethnicity and Gender in Private Industry | 22 | Senior Officials and Managers | White | Black | Male | Female, 2007 - 2017. Sage Data. Sage Publishing Ltd. (Dataset). Dataset-ID: 084-001-001. https://doi.org/10.6068/DP17972718EFC11

For long term preservation, the files are owned by their respective publishers. Should ownership change, I do not believe there is a concern for the files becoming obsolete provided they are reorganized and published under new ownership with reference to their original publisher. There is no specific software required to open the files, many are available for download in multiple formats. An appropriate copyright license would be an open copyright license as it is publically collected and published data that is meant to be used to draw conclusions and inferences on job patterns for women and minorities in private industry. An open license would allow the data to be easily shared and increase discoverability. There are human subject considerations with this data as it contains personally identifiable data such as race, age, occupation, gender, etc.. However, personal information was redacted to ensure anonymity for participants. The demographic information collected is not published with any personal information attached.

References

National Academies of Sciences, Engineering, and Medicine. 2022. Evaluation of Compensation Data Collected Through the EEO-1 Form. Washington, DC: The National Academies Press. https://doi.org/10.17226/26581.