Income of Adults

Project

Meryl Garcia, Liv Dudley, Collin Mirkovic, Chayse Windedahl

Intro to our data-set

- The Adult Data Set covers census data of age, marital status, race, gender, education, and country of origin

- They take all collected data to predict whether the person makes either over or under $50,000 a year

- We focused on allowed data in the wage column as it doesn’t communicate much information and is only predicted in USD despite collecting data globally

- This data set also does not define whether the data was collected from only the United States or globally, so due to the different nationalities listed we are operating under the assumption that this census data is global

Intended purpose of data-set

- It is intended to take the collected factors into account and predict whether the person would have a yearly salary of either over or under 50,000 USD

- Looks at the aspects of their identity to make this conclusion

- Looks at the advantages and disadvantages of their identity within their respective countries in order to predict their salary

Reading of Data Sets - Poirier

Connotative Reading

- Connotative Reading allows for looking at the cultural means behind the data collection, which is clear for this data set.

- Collection was set to be globally with a large majority of U.S. adults within the data

- Data shows the ways that people can make similar amounts of money (by using USD) within the data set, though doesn't allow for any views as to the way that person is actually living within different countries making the same amount of money.

Denotative Reading

- Denotative looks at the basis of data, and this data is set to show how different social categories someone is in affects their yearly salary

- Data dictionary shows clear and few categories they allowed for their data collection to make the conclusions clear

- Each column represents a different social category collected and each row is a person's collective information

Readings of the Data Sets - Koopman Main

Koopman's Macrostructure

- The standard to present data for currency values is usually in the US dollar system which does not take into account other countries and the value of a dollar
- Education systems vary vastly in different parts of the world so a high school education might not be the same in the US as everywhere else
- Marital status can differ as well. In the US most women work full time while in other countries they might not be permitted to do so.
- Our data set groups different countries together under the US standards when the value of money and societal norms vary throughout the world so class status cannot be determined under the same umbrella. For example, you might be middle class in the US but upper class somewhere else based on income

Reading of Data Sets - Koopman

Microstructure

- Data set Includes variables such as wage, race, gender, country, etc.
- Missing values in the country column, “?” is put as a placeholder
- Data Type for the wage column is binary, stands out since other variables are not stored in this way
- Binary does not allow for comprehensiveness, since above or below 50k is too wide of a spectrum

Mesostructure

- The design of the data set is very defined and clear in the way that it is structured

- Main focuses on each individual in each row and then focus on each of the identities they hold within each column

- Does not allow for further nuances of the data points but rather acts as very black and white with the data collected

- This data would not be easy for a ML engine to read and comprehend due to the falsehoods of lack of information and care to global standards

Why the lack of comprehensiveness

- Too limited of a salary range to be reported as 51K vs a 1M salary is an extremely large difference to just be ignored while 49k vs 51k is separated

- Only allowing for USD as well makes it incomprehensive for other countries

    - Doesn’t allow for currency rates and change

- Difficult to find true takeaways from the data and therefore makes it confusing as to why it was collected in the first place

What they should've done instead

- The salary range on the data set should have been under less broad parameters and allowed for other currencies.

- Poverty lines should have been evaluated in order to make a more comprehensive data set that's useful under global means.

- Bias should not be evident and more people from other countries should have been collected on if they had wanted a global data set.

Thanks for listening