Rene Guerra - Final Project Data Memo

Sunday, October 2th 2022

Final Project Data Memo

The dataset I selected includes information regarding the percentage of poverty in the United States population since 1959. Data is distributed in decimals according to poverty rates based on the categories: By age group, By family type, By race/ethnicity, By Region and By State.

"Percent of people in poverty" is obtained from the the source USAFacts found through the Google Dataset Search engine. The page offers graphs visualizing the information provided in an Excel document with more than 3,000 observations. Predictors are the 50 states, 3 age groups, 5 family types, 5 ethnic groups, and 4 regions.

There is missing data for all state poverty rates from 1959 to 1979 and some data by region from 1959 to 1970. I believe the most effective way to handle this task is to consider observations starting on 1980 since all observations are complete and focus on contemporary circumstances.

Link to source: https://usafacts.org/data/topics/people-society/poverty/poverty- measures/poverty-rate-of-all-persons/

With this project I would like to predict what combination of the categories mentioned above is at risk of staying in poverty for the next 20 years. Also, this approach will determine the highest and lowest poverty rates in the future. The response to this prediction can be "Particular geographical, and demographic communities at risk of having significant poverty in 20 years." The main question can be answered by a regression approach because it will evaluate quantitative observations and will set a ranking system. The most useful predictors can be racial background and age since they can create a general perspective of multiple communities across the United States. This outcome means that we'll be able to predict communities that will remain affected by poverty in the future and can be a call to action.

I'll start transferring data during the upcoming week, data analysis exploration the following weeks, and have the project done two weeks prior to the deadline to focus on small formatting and additional details for context.

The main concerns I currently have are that I wanted to make sure basing data from an Excel document is okay to use in R, and also what do you recommend to do with the missing data. No other questions at the moment, but would appreciate any feedback.