Obj2_1

Reilly Maw

6/8/2018

DOA,

I wanted to send a quick email regarding the PSID data you sent three days ago. First off, we understand and appreciate the effort required to complete this task, so thank you for taking the time to compile and anonymize the data for us. I wanted to summarize our findings, bring up a few issues we ran into, and propose a few recommendations that would help us in the future.

After digging into the files and conducting three main tests, we found that the data is pretty strong overall. Before I get into the tests we performed, I would like to say that your team did a great job on data entry. Throughout our tests, we found only a couple of provable mis-typed entries. We understand that this step is tedious and there are bound to be mistakes. We first wanted to look at the reliability of the data and the underlying survey population. We looked at the age distribution of interviewees to see if the surveys were being deployed to a population that represents our nation as a whole. After taking care of unknown data, we found that the population does a good job of representing our country. One issue we did come across was that 0-2 year old's seem to be overrepresented in the survey data. We then tackled the issue of confirmation bias. We not only looked at the survey, but how the engineers entered the information. We looked at the amount of manually entered/altered data, and it seemed to be relatively low. We also took a look at unknown information, more specifically variables related to income. We found that about 95% of information in these categories ER34144A,ER34144C,ER34144D, and ER34144E were labeled as "0", representing the value zero or an unknown value. On our end, it is hard to draw conclusions with this ambiguity in the data. Finally, we wanted to verify the data transformation, specifically if the two separate survey instances were zipped up correctly. To do so, we looked at the "relationship to head", birth month, and birth year variables for every data point that had interviewees present for both surveys. From what we could tell with anonymized data, the majority of observations seem to be correct.

The data overall seemed to be strong and helpful for our study. The first recommendation we would make to further strengthen our conclusions would be to dedicate a separate bucket for unknown information and actual zero values. From our end we can't be certain as to what a "0" represents without having been at the interview or having the non-anonymized data. We feel this may significantly cut down the number of observations we have to discard. Our second recommendation would be to continue to try to improve on data entry. As I'm sure you know, each input usually maps to a classification, so a mistyped integer will completely change the story of that data point. Again, we really appreciate all the work your team has done and we look forward to working with you in the future.

Best,

Reilly Maw