Name : Ayachit Madhukar   
Course : DSC530  
Instructor : Mr. Shankar Parajulee  
Date : 8 Mar 2021

* Outcome of your EDA

Exercise has following outcomes:

In this exercise, you’ll focus on one variable in this dataset, 'year', which represents the year each respondent was interviewed.

variable 'age' contains respondents’ age in years

The distribution of income in almost every country is long tailed; that is, there are a small number of people with very high incomes.

In the U.S, 12 years of education usually means the respondent has completed high school (secondary education). A respondent with 14 years of education has probably completed an associate degree (two years of college); someone with 16 years has probably completed a bachelor’s degree (four years of college).

It might not be surprising that people with more education have higher incomes, but looking at these distributions, we can see where the differences are.

In many datasets, the distribution of income is approximately lognormal, which means that the logarithms of the incomes fit a normal distribution.

The lognormal model is a pretty good fit for the data, but clearly not a perfect match. That’s what real data is like; sometimes it doesn’t fit the model.

* What do you feel was missed during the analysis?

I had most of the categorical variables in dataset. I wish to include more continuous variables

* Were there any variables you felt could have helped in the analysis?

GSS provided more than enough data and variable for analysis, I did not feel need of any additional variable for my analysis.

* Were there any assumptions made you felt were incorrect?

none

* What challenges did you face, what did you not fully understand?

It was quiet challenging to get the data in appropriate format. Dataset was in SPSS format and it took a while to get it converted into .csv format. I did use IBM tool to load the data and transform it into .csv file.

Another challenge was to strop the data to acceptable size and narrow down variables to 8 important fields.