Course Project - QMSS GR 5069 - Applied Data Science

Meredith Meeks, Hao Liu, Shilpa Sure, Kai Huo, Steve Andriyishen

I. Project description

Our project deals with the variables that affect marital age at the time of the first marriage (variable name is 'agewed'). Through this endeavor, we intend to find predictor variables that will inform us of characteristics of individuals who get married earlier or later. We will aim to create a model or tool that predicts the age at which someone will get married.

II. Insight

What predictor variables affect the age at which people first get married? Can we predict marital age for our classmates? We think this would be an interesting project for our classmates. We think this because individuals tend to like these types of tasks, because they like to know where they fall amongst the ages of marriage trends. Furthermore, if we find interesting and novel statistically significant results, it would also inform researchers in the field on future patterns within social science.

In addition, we think this project can eventually become a useful tool for marketers and social scientists. Weddings in the United States are an approximately 55 billion dollar industry. For businesses in the wedding industry, a tool that can inform them of the average age someone gets married can help target marketing. For example, if a bridal boutique wants to open a new store in a new market, they can use our tool to inform decisions about which age group to target and what advertising channels to use as a result.

III. Research strategy

We will be using data from the 2006 General Social Survey (most recent GSS with our outcome variable of interest). We will be gathering a handful of predictor variables, and will report on the variables that have a very high predictive power. To build our predictive model, we will test different predictive techniques, including some machine learning techniques.

IV. Data

The data we are using is from the General Social Survey 2006, which is the last edition of this survey that contained the variable we are interested in. This is a very comprehensive dataset that has thousands of responses. The limitation of this dataset is that the data is old, and therefore may not be as applicable modern-day. Still, we feel confident in the dataset because it is a randomized nationwide sample and from a reputable organization.

V. Output

The intended output for this project is to show which predictor variables are significantly correlated with marital age.