12/15/2022
Math 448: Project Proposal
Gabrielle Salamanca

I. Team

There is only one team member: Gabrielle Salamanca.

II. Description of the Problem

There have been studies and small test-runs of names having an impact on your life, especially concerning the job aspect. The objective of this project is inference. This is because I wish to explore the relationship between the names and other resume factors, such as job experience and ethnicity, and if they have impacted the possibility of a call back.

III. Description of the Data

The title of the dataset is "Are Emily and Greg More Employable Than Lakisha and Jamal?" from Vincent Arel-Bundock's Github projects, which it had taken from Stock and Watson (2007). This dataset was from a randomized, controlled experiment conducted by Bertrand and Mullainathan (2004). It is 4870 by 27, and out of 27 variables, 2 are numerical. Those 2 are: Jobs and Experience. The categorical variables are: Name, Gender, Ethnicity, Quality, Call, City, Honors, Volunteer, Military, Holes, School, Email, Computer, Special, College, Minimum, Equal, Wanted, Requirements, Reqexp, Reqcomm, Reqeduc, Reqcomp, Reqorg, and Industry. The 4870 rows are of fictitious applicants, who applied to employment advertisements in Chicago and Boston in 2001.

IV. Supervised or Unsupervised

The closest output would be the Call variable ("Was the applicant called back?"), but the goal isn't to predict, it is to infer. I specifically want to learn the structure and patterns within this dataset, so this project will use unsupervised learning.

V. Additional Comments

There are two concerns I have with this dataset. One is a category called minimum, a factor indicating minimum experience requirement of the employer. This variable is partially numerical, where the employer wanted exactly how many years, and partially categorical, where the employer was vague. This may provide a little difficulty in trying to explore the data and trying to code this category into something R can use to make graphs and such. The second concern is with categories having vague answers. There are two categories that have this: Wanted, a factor indicating the type of the position wanted by the employer; and Industry, a factor indicating the type of employer industry. I have noticed some cells have "other" in Wanted and some have "unknown" in Industry. This could possibly skew the data when made into plots and graphs. I'm unsure if I will remove the Industry and/or Wanted variable because of this