Ania Shaheed, Riya Parikh, Adhya Hoskote
DS 110
7 November 2022

# Final Project Proposal

**Describe in your proposal how you will get your data.**

Our group is proposing to conduct a qualitative study in which we utilize Google Forms to distribute a survey to answer our research questions. The study participants of this experiment include a minimum of 100 students (a large enough sample to run statistical tests and assume that the participants are representative). If we find that the survey does not provide us with sufficient data to test our hypotheses, we plan to send out another round of the survey and gather as many responses before performing statistical analysis. The source population for these subjects is Boston University. Other inclusion criteria include full-time undergraduate students. We will send out the survey online to BU students. After obtaining results, we will clean the data before performing statistical analysis and visualization in Python. For example, we may have to standardize the responses so that multiple responses are not associated with the same key terms (for example “Yes” and “Yeah” both mean “Yes”). Additionally, we will have to identify any potential outliers.

**Decide on two hypotheses.**

1. Does the amount of perceived stress depend on a student’s major?
    - Null hypothesis: There is no statistically significant difference between perceived stress levels among different majors at Boston University.
    - Alternative hypothesis: There is a statistically significant difference between perceived stress levels among different majors at Boston University.
2. Question 2: Does the average amount of sleep depend on a student’s major?
    - Null hypothesis: There is no statistically significant difference between the amount of sleep students in a STEM major get as compared to students in a non-STEM major.
    - Alternative hypothesis: There is a statistically significant difference between the amount of sleep students in a STEM major get as compared to students in a non-STEM major.

**Describe how you plan to evaluate your hypotheses with statistics and/or visualization.**

Our data will be multivariate. Predictor variables will be categorical, including major as well as type of major (STEM vs. non-STEM). Response variables will be discrete data (which is also categorical), including stress levels, average amount of sleep, and anything else that we may consider relevant to our research questions (such as homework level per week). As such, a fitting statistical test is Pearsons’s chi square analysis, which can be used to determine whether this is a difference in distribution of categorical variables between two or more independent groups. This test will directly answer our research questions, which are concerned with statistically significant differences, based on the p-value output. As for visualization, we believe that mosaic plots, balloon plots, and/or bar charts may be useful. If possible, we would also like to go beyond by implementing machine learning through logistic regression, a predictive analysis technique.