# Project Checkpoint 1

## We started our exploratory data analysis by examining the raw data from the Kaggle dataset, which consists of three files: (1) ADHD diagnosis and sex labels, (2) functional MRI connectome matrices, and (3) socio-demographic information. Our objective is to leverage the functional MRI connectome matrices and socio-demographic data to predict ADHD diagnosis and sex. As a preprocessing step, we will remove all missing values from the datasets.

## Functional MRI Connectome Matrices

Each row represents an individual participant, while the columns correspond to connectivity values between pairs of brain regions. With 19,900 connectivity pairs, the data suggests a 200-region parcellation, as calculated by 200(200-1)/2 = 19,900.

Next, we examine the distribution of the fMRI data. The values are centered around zero and range between ±1, which aligns with typical fMRI connectivity representations based on our brief research.

A diagram of a distribution of function

AI-generated content may be incorrect.It is centered around zero and ranges between +/- 1. From brief research, this is how a typical FMRI would be represented.

If our assumption about the 200-region parcellation is correct, the fMRI data is stored as a vectorized upper triangular part of a 200×200 connectivity matrix. This allows us to reconstruct the full matrix, enabling the visualization of connectivity patterns through heatmaps and distribution plots for each participant.

A graph of a function

AI-generated content may be incorrect.

## To prepare the dataset for our machine learning models, we will apply feature selection using SelectKBest in Python to retain the most relevant features. Alternatively, we can use all 19,900 connectivity values as input to train a classifier, such as Random Forest or Logistic Regression.

## Combined Data

## We combined the remaining two datasets (data\_train\_quantitative and data\_train\_categorical) into a single data frame and generated a correlation heatmap to identify weak relationships or potential multicollinearity (Figure 1).

Additionally, we explored the frequency distribution of numerical features using a histogram (figure 2) and examined the class distributions of ADHD diagnosis and sex in the target variables (figure 3 and figure 4).

## Summary

To optimize model performance, we plan to experiment with both balanced and imbalanced class distributions. We will also split the training dataset using 80/20 split for training and validation of our model performance. We are currently examining the feasibility of other machine learning models such as deep learning (especially for the FMRI) or using summary statistics to represent the key network properties of the FMRI. Finally, we may build a GNN model (but might be outside our expertise).

A screen shot of a graph

AI-generated content may be incorrect.

Figure 1

A screenshot of a graph

AI-generated content may be incorrect.

Figure 2

A graph of a number of blue squares

AI-generated content may be incorrect.A graph of a distribution of adhd outcomes

AI-generated content may be incorrect.

Figure 4

Figure 3