Name: Fangle Xi **UNI:** fx2180

According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). I hope to use the data and statistical knowledge to find the key indicators of heart disease.Originally, the dataset comes from the CDC in 2020 and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. In this dataset I will use in my final project, the kaggle data provider, Kamil Pylak, selected the most relevant variables from the original one and do some cleaning so that it would be usable for his machine learning projects. The dataset contains 18 variables (9 booleans, 5 strings and 4 decimals) from over 300,000 objects. **Link to the data:**

www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Initial analysis plan: After cleaning and transforming the dataset, I would like to use the graph to show the distribution of each variables and use the clustering models like the KNN, K-means, SVM and random forest to find out the main character of heart disease people. In addition, I will use the hypothesis testing to test the result of clustering is good or not. What's more, I want to use the logistic regression model to show which indicator is the most important one in our regression model.

Features name: HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer