**Introduction:**

According to the survey from American Diabetes Association, more than 30 million Americans have diabetes and around 84 million Americans have pre-diabetes. The health care costs of diagnosed diabetes are 327 billion per year according to the 2017 statistics. Compared with those without diabetes, people with diabetes are spending 2.3 times more on their medical expenditure. With National Health and Nutrition Examination Survey on hand, our group are interested in exploring the prevalence of diabetes in terms of demographic features, dietary habits and alcohol use.

In the end of our analysis, we will show how gender, age, nutrient, occupation, alcohol use and sleep quality are associated with the probability of being diagnosed with diabetes. By comparing respective odds ratios, we will come to a conclusion that describes which population are more subject to diabetes.

**Data:**

The response variable is DIQ010 (*Doctor told you have diabetes*) from DIQ\_I.XPT (*Diabetes 2015 - 2016*). We utilized 9 predictors from 6 datasets. The predictor variables include: RIAGENDER (*Gender*) and RIDAGEYR (*Age in years at screening*) from DEMO\_I.XPT (*Demographic Variables and Sample Weights 2015 - 2016*); ALQ130 (*Avg # alcoholic drinks/day – past 12 mons*) from ALQ\_I.XPT (*Alcohol Use 2015 - 2016*); DR1ISUGR (*Total sugars (gm)*) and DR1ITFAT (*Total fat (gm)*) from DR1IFF\_I.XPT (*Dietary Interview – Individual Foods, First Day 2015 - 2016*); DR1TSUGR (*Total sugars (gm)*) and DR1TTFAT (*Total fat (gm)*) from DR1TOT\_I.XPT (*Dietary Interview – Total Nutrients Intakes, First Day 2015 - 2016*); OCQ260 (*Description of job/work situation*) from OCQ\_I.XPT (*Occupation 2015 - 2016*); SLD012 (*Sleep hours*) from SLQ\_I.XPT (*Sleep Disorders 2015 - 2016*).

**Methods:**

We preprocessed and cleaned all the variables beforehand. The rows with missing values are removed and some variables are recoded for consistency. After the data preprocessing step, we fit the dataset with logistic regression. We carried out the above procedures in R, python and STATA. We have set some common rules for performing the data cleaning part. The response variable DIQ010 should be coded as 1: Yes and 0: No and should be treated as a categorical variable. RIAGENDER (*Gender*) should be coded as 1: Males and 2: Female and should be treated as categorical variable. ALQ130 (*Avg # alcoholic drinks/day – past 12 mons*) should range from 1 to 15. The use of two datasets DR1IFF\_I.XPT and DR1TOT\_I.XPT is to confirm the measurement accuracy of total sugar intake and total fat intake. OCQ260 (*Description of job/work situation*) should be recoded as 1: private business employee, 2: government employee and 3: self-employed and should be treated as a categorical variable.