Dataset Description

The [National Health and Nutrition Examination Survey (NHANES)](https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm) data set contains information to assess the health and nutritional status of adults and children in the United States. It was created by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC).

The data collection for NHANES began in the early 1960s. The survey data examines a nationally representative sample of about 5,000 persons each year. The interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. We were able to get our hands on the 2014 data from Kaggle.

We decided we wanted to create a model that predicts any individual’s blood pressure by using a multitude of variables concerning lifestyle, diet, demographics, and health conditions. After doing some research we narrowed down the variables available in the NHANES data set. The website of the Mayo Clinic was a main resource in determining what variables might influence blood pressure ([www.mayoclinic.org](http://www.mayoclinic.org)).

We learned that the more blood your heart pumps and the narrower your arteries, the higher your blood pressure. Primary (essential) hypertension develops gradually over many years while various conditions and medications can lead to secondary hypertension. Known risk factors include, age, race, family history, being overweight, not being physically active, use of tobacco, too much salt in diet, too little potassium, drinking too much alcohol, stress, some chronic conditions, sleep apnea and pregnancy.

Armed with this information we filtered the data dictionary on these known risk factors to narrow down the full data set to the variables we would be interested in for our model. We also kept in more of the lab results and some survey participant demographic data like income. We have copied information from the NHANES data dictionary and included them in our appendix. The data dictionary was also helpful in determining if the variables we selected had a large enough sample size. For our training we wanted to make sure that we were able to keep as many complete rows as we could and not allow any single variable to limit our sample because it had too many NULLs. The final five categories of data that we pulled from are, demographic, diet, examination, lab and survey data.

The demographic data includes: RIAGENDR(Gender), RIDRETH1(Race/Hispanic origin), RIDEXPRG(Pregnancy status at exam), DMDHHSZA, DMDHHSZB, DMDHHSZE and INDHHIN2(Annual household Income).

Diet data includes: DRQSPREP(Salt used), DRQSDIET(Special Diet, DR1TCALC(Calcium), DR1TMAGN(Magnesium), DR1TSODI(Sodium), DR1TPOTA (Potassium), and DR1TALCO(Alcohol)

Examination data includes: BMXWT(Weight), BMXBMI(Body Mass Index), BMXWAIST(Waist Circumference), as well as the systolic and diastolic blood pressure readings.

Lab data includes: LBDSCASI(Total Calcium), LBDSCHSI(Cholesterol), LBDSGLSI(Glucose), LBDSTRSI(Triglycerides), and LBDHDD(Direct HDL-Cholesterol).

Survey data includes: ALQ130(Avg # alcoholic drinks), DIQ010 (Doctor told you have diabetes), DIQ160(Ever been told you have prediabetes), DBD895(# of meals not home prepared), BPQ030(Told had high blood pressure – 2+ times), MCQ080(Doctor ever said you were overweight), MCQ160b(Ever told had congestive heart failure), MCQ160c(Ever told you had coronary heart disease), MCQ160f(Ever told you had a stroke), PAD680(Minutes sedentary activity) and SMQ020(Smoked at least 100 cigarettes in life).