Data Proposal - 583

Matthew Yau, ZiYing(Sophie) Chen, Xinyu Dong

2023-02-26

Data Proposal: Personal Key Indicators of Heart Disease

This project is aimed at discovering the association between heart disease and living habits.Not only will we identify and explain important variables, but we will also try to find relationships between different habits and then group the data accordingly. People can tell some habits are bad and others are good without supportive evidence, and this analysis might make those judgments more convincing or help people get rid of those myths.

The following are some challenges for the dataset:
- The data concentrated among older population who has more chance to have landline for telephone survey;
- Subjective answers such as PhysicalHealth and MentalHealth with scale 0-30, the 3rd quantiles of them are only 2 and 3 respectively;
- Most data are collected by race-white, we might consider to convert it to binary races or remove the race variable because performing research based on races might be an issue.

Statistical Description

There are 18 columns in total, including if the patient is diagnosed with heart disease or not and variables regarded as potentially associated with heart disease.The property of each variable is listed below.

Variable Name	Description	Data Type	Potential Distribution
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	Binary	Binary Distribution
BMI	Body Mass Index (BMI)	Continuous	Skew Normal Distribution
Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	Binary	Binary Distribution
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week	Binary	Binary Distribution
Stroke	(Ever told) (you had) a stroke?	Binary	Binary Distribution
PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)	Discrete	Poisson Distribution

Variable Name	Description	Data Type	Potential Distribution
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)	Discrete	Poisson Distribution
DiffWalking	Do you have serious difficulty walking or climbing stairs?	Binary	Binary Distribution
Sex	Are you male or female?	Binary	Binary Distribution
AgeCategory	Fourteen-level age category	Discrete	
Race	Imputed race/ethnicity value		
Diabetic	(Ever told) (you had) diabetes?	Binary	Binary Distribution
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	Binary	Binary Distribution
GenHealth	Would you say that in general your health is...	Discrete	
SleepTime	On average, how many hours of sleep do you get in a 24-hour period?	Continuous	Skew Normal Distribution
Asthma	(Ever told) (you had) asthma?	Binary	Binary Distribution
KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?	Binary	Binary Distribution
SkinCancer	(Ever told) (you had) skin cancer?	Binary	Binary Distribution

Data Collection

Originally, the dataset come from the CDC 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. As the CDC describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.". The most recent dataset (as of February 15, 2022) includes data from 2020. It consists of 401,958 rows and 279 columns. The vast majority of columns are questions asked to respondents about their health status, such as "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]". In this dataset, the author noticed many different factors (questions) that directly or indirectly influence heart disease, so they decided to select the most relevant variables from it and do some cleaning so that it would be usable for machine learning projects.

Question of Intrest

- How those living habit and health condition related variables may affect the possibility of being diagnosed with heart disease in general.
- How each variable contribute to the inference of heart disease, identify mostly related variables.
- How is the correlation among those variables.
- How to identify vulnerable groups of getting heart disease.