Data clening

Xinghan Liu

2023/06

"Section 1: Description of the data. This data I used is a comma-separated values (CSV) file containing health information for individuals' HeartDisease conditions.The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS). Factors like BMI, Physical Health,and Mental Health are contineous variables. Other lifestyle habits like smoking, alcoholm Diabetic,Physical Activity and General Health are binary variables.Age is category variable. The data is in a flat file that is delimited, each data point within a record is separated by a specific character- comma (',')."

## [1] "Section 1: Description of the data. This data I used is a comma-separated values (CSV) file containing health information for individuals' HeartDisease conditions.The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS). Factors like BMI, Physical Health,and Mental Health are contineous variables. Other lifestyle habits like smoking, alcoholm Diabetic,Physical Activity and General Health are binary variables.Age is category variable. The data is in a flat file that is delimited, each data point within a record is separated by a specific character- comma (',')."

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#Section 2: Reading the data into R. This next line is reading data by using base R command - read.csv. And then I used dplyr to create a subset called 'dropped' of the original data, it drooped the last two variables.   
  
datanew = read.csv('/Users/hankhan688/Documents/heart\_2020\_cleaned.csv')  
dropped <- datanew %>% select(-KidneyDisease, -SkinCancer)

#Section 3: Clean the data. I changed the value of variables "HeartDisease", "Smoking", "AlcoholDrinking", "Stroke" from " Yes/No" to "1 or 0". And also renamed the two vriables.  
  
dropped[, c("HeartDisease", "Smoking", "AlcoholDrinking", "Stroke")] <-   
 lapply(dropped[, c("HeartDisease", "Smoking", "AlcoholDrinking", "Stroke")],   
 function(x) ifelse(x == "Yes", 1, 0))  
renamed <- dropped %>%  
 rename("Alcohol" = "AlcoholDrinking",  
 "Age" = "AgeCategory")

#Section 4: Characteristics of the data.   
  
cat("The dataset has", nrow(renamed), "rows and", ncol(renamed), "columns. The names of the columns and a brief description of each are in the table below:\n")

## The dataset has 319795 rows and 16 columns. The names of the columns and a brief description of each are in the table below:

column\_descriptions <- c(  
 "HeartDisease - Have heart disease(1) or not(0) ",  
 "BMI- Body Mass Index",  
 "Smoking - Smoking status Yes(1)/ No(0)",  
 "Alcohol -Alcohol consumption Yes(1)/ No(0)",  
 "Stroke - Stroke history Yes(1)/ No(0)",  
 "PhysicalHealth - Days of poor physical health in a month",  
 "MentalHealth - Days of poor mental health in a month",  
 "DiffWalking - Difficulty climbing stairs",  
 "Sex - Gender (Male/Female)",  
 "age - Age category",  
 "Race - Different race",  
 "Diabetic - Diabetic status (Yes/No)",  
 "PhysicalActivity - Physical activity in past 30 days (Yes/No)",  
 "GenHealth - General health perception",  
 "SleepTime - Number of hours of sleep",  
 "Asthma - Asthma history (Yes/No)"  
)  
  
data\_table <- data.frame(  
 Variable = names(renamed),  
 Description = column\_descriptions  
)  
  
library(knitr)  
  
kable(data\_table, caption = "Data Dictionary")

Data Dictionary

| Variable | Description |
| --- | --- |
| HeartDisease | HeartDisease - Have heart disease(1) or not(0) |
| BMI | BMI- Body Mass Index |
| Smoking | Smoking - Smoking status Yes(1)/ No(0) |
| Alcohol | Alcohol -Alcohol consumption Yes(1)/ No(0) |
| Stroke | Stroke - Stroke history Yes(1)/ No(0) |
| PhysicalHealth | PhysicalHealth - Days of poor physical health in a month |
| MentalHealth | MentalHealth - Days of poor mental health in a month |
| DiffWalking | DiffWalking - Difficulty climbing stairs |
| Sex | Sex - Gender (Male/Female) |
| Age | age - Age category |
| Race | Race - Different race |
| Diabetic | Diabetic - Diabetic status (Yes/No) |
| PhysicalActivity | PhysicalActivity - Physical activity in past 30 days (Yes/No) |
| GenHealth | GenHealth - General health perception |
| SleepTime | SleepTime - Number of hours of sleep |
| Asthma | Asthma - Asthma history (Yes/No) |

#Section 5: Summary statistics. I coshed "BMI", "PhysicalHealth", "SleepTime" these three variables:  
choosedcol <- c("BMI", "PhysicalHealth", "SleepTime")  
subset\_data <- renamed[, choosedcol]  
  
# Calculate summary statistics (excluding missing values)  
summary\_results <- summary(subset\_data, na.rm = TRUE)  
print(summary\_results)

## BMI PhysicalHealth SleepTime   
## Min. :12.02 Min. : 0.000 Min. : 1.000   
## 1st Qu.:24.03 1st Qu.: 0.000 1st Qu.: 6.000   
## Median :27.34 Median : 0.000 Median : 7.000   
## Mean :28.33 Mean : 3.372 Mean : 7.097   
## 3rd Qu.:31.42 3rd Qu.: 2.000 3rd Qu.: 8.000   
## Max. :94.85 Max. :30.000 Max. :24.000