# EDA

Matthew Yau, ZiYing(Sophie) Chen, Xinyu Dong

2023-03-09

## Statistical Description

There are 14 categorical and 4 numerical variables in the dataset, and our target variable is "Heart-Disease". This is a clean dataset without any missing data. Among the 319,795 observations, we removed 18,078 duplicates. Therefore, the following explanatory data analysis would only perform on 301,717 observations.

## Univariate Analysis

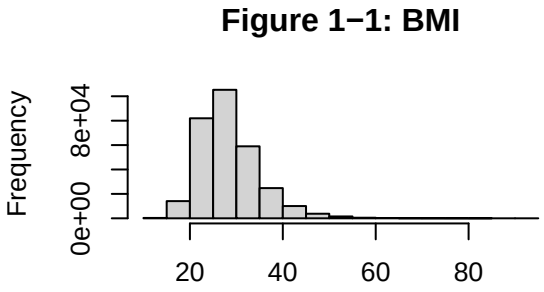
- Numerical Data:

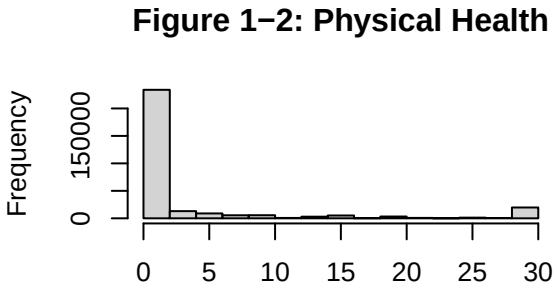
From Table1 we can see that the means and medians of *BMI* and *SleepTime* are close, which suggests that there are no obvious skewness of the data. On the other hand, means of both *PhysicalHealth* and *MentalHealth* are bigger than their medians, which means that they skew to the right. That makes sense because most people in a survey would rather not claiming they have health issues.

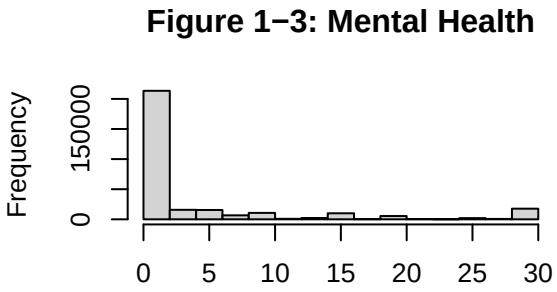
We can use the range and the standard deviation to see the spread of the data, but using histograms would be a better option. From Figure 1-1 to Figure 1-4, we can see that most people scored their physical and mental health to be 0 which means most people did not feel bad in past month. Contrastly, some of them claimed they had physical and mental health problem every day in past month.The spreads of *PhysicalHealth* and *MentalHealth* are similar according to the charts and that makes sense because the range of scores are the same and these two variables might relate to each other.

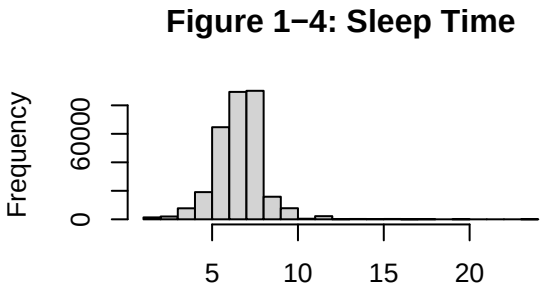
Table 1: Statistics for Numerical Variables

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BMI | 1 | 301717 | 28.441970 | 6.468134 | 27.41 | 27.822767 | 5.574576 | 12.02 | 94.85 | 82.83 | 1.2977241 | 3.690549 | 0.0117755 |
| PhysicalHealth | 2 | 301717 | 3.572298 | 8.140657 | 0.00 | 1.192249 | 0.000000 | 0.00 | 30.00 | 30.00 | 2.5009749 | 4.971089 | 0.0148204 |
| MentalHealth | 3 | 301717 | 4.121475 | 8.128288 | 0.00 | 1.929218 | 0.000000 | 0.00 | 30.00 | 30.00 | 2.2374477 | 3.938407 | 0.0147979 |
| SleepTime | 4 | 301717 | 7.084559 | 1.467122 | 7.00 | 7.093788 | 1.482600 | 1.00 | 24.00 | 23.00 | 0.6972167 | 7.571991 | 0.0026710 |




From Figure 2-1 to 2-4, we detected many outliers existing in all the variables, but we would not remove them for now because maybe these outliers have higher chance to get heart disease.

**Figure 2−1: BMI**

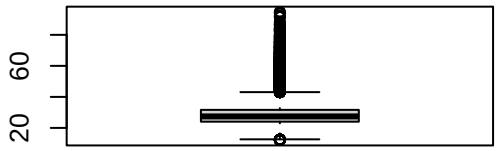



**Figure 2−2: Physical Health**

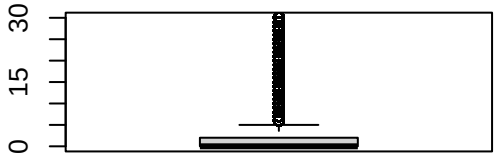



**Figure 2−3: Mental Health**

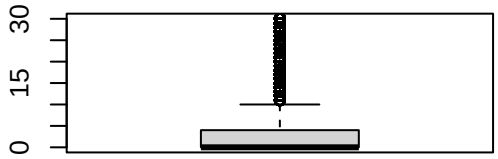
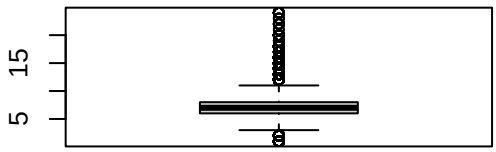
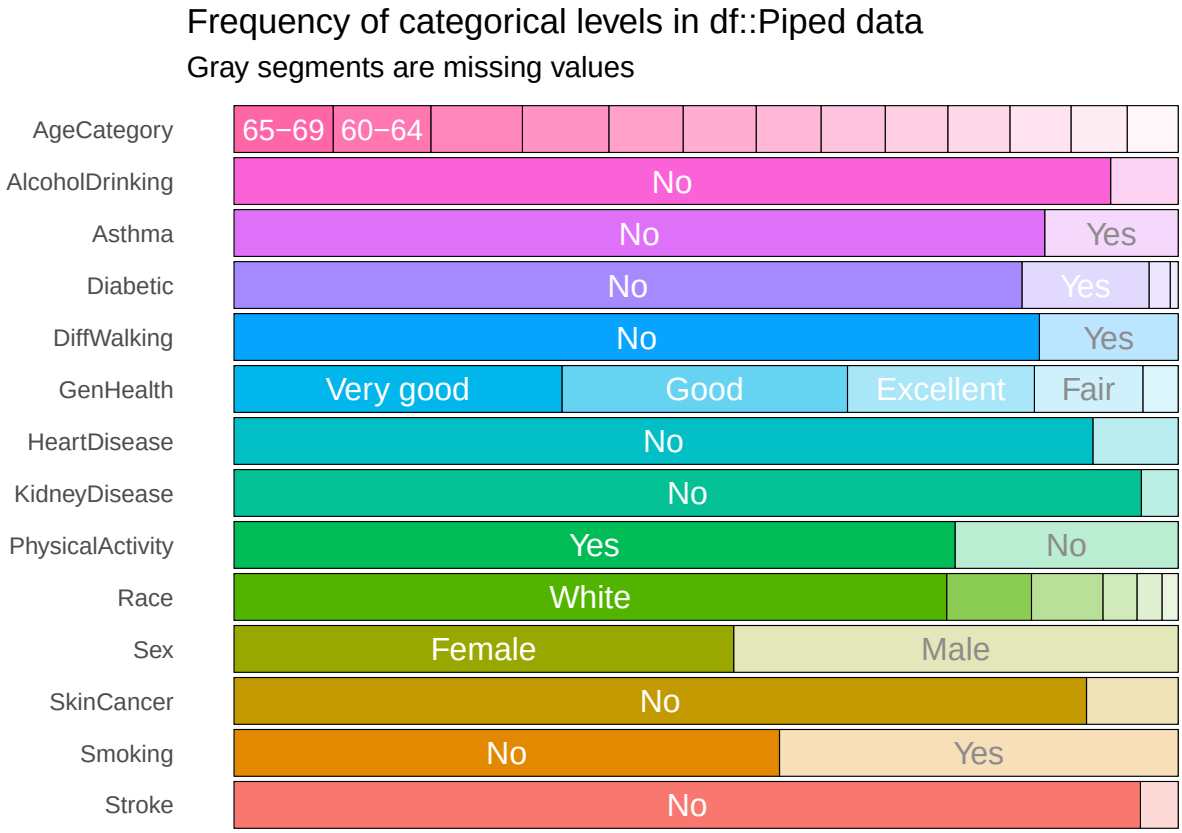



**Figure 2−4: Sleep Time**

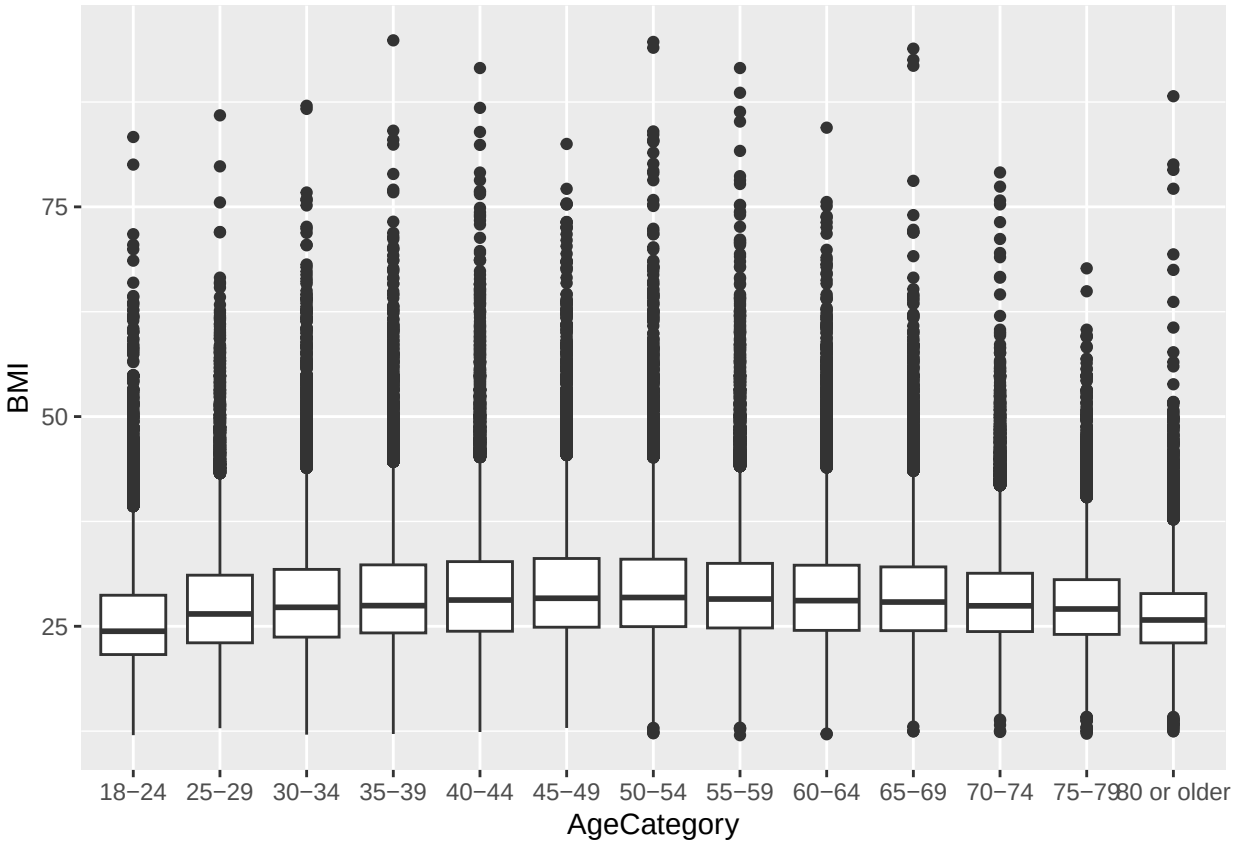



- Categorical Data:

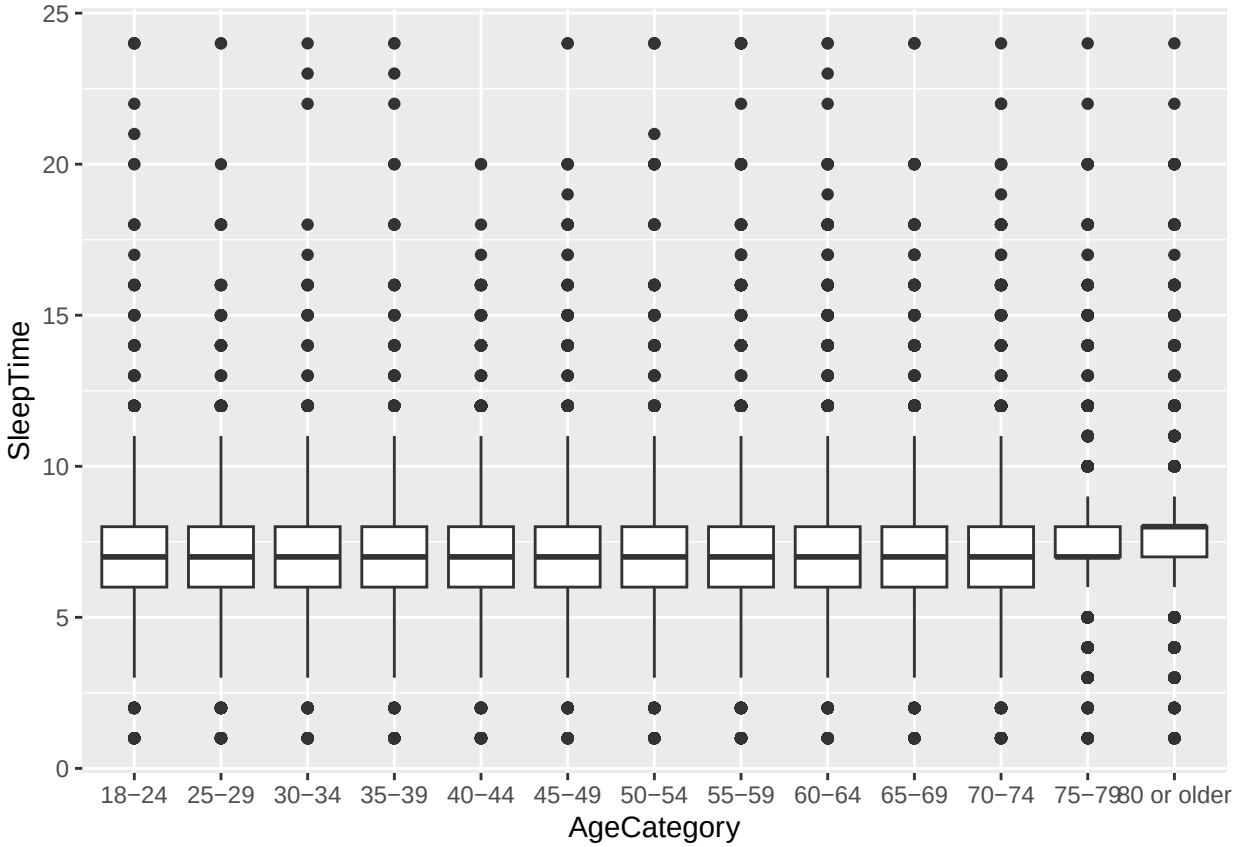
From Figure 3, We can that our target variable *HeartDisease* is unbalanced which most of the cases do not have heart disease. The *AgeCategory* and *Sex* looks balanced here, we would like to see if there is any significant difference of the probabilities of getting heart disease among age groups and sex later. "White" category dominates the *Race*, we would proceed the analysis by convert it to binary data with "non-white" as the other. *GenHealth* indicates that most people feeling good which match the majority of the observations do not have heart disease. The rest of the variables are binary and all of them incline to one side.


Frequency of categorical levels in df::Piped data
Gray segments are missing values
AgeCategory
65−69
60−64
AlcoholDrinking
No
Asthma
No
Yes
Diabetic
No
Yes
DiffWalking
No
Yes
GenHealth
Very good
Good
Excellent
Fair
HeartDisease
No
KidneyDisease
No
PhysicalActivity
Yes
No
Race
White
Sex
Female
Male
SkinCancer
No
Smoking
No
Yes
Stroke
No

**Bivariate Analysis**

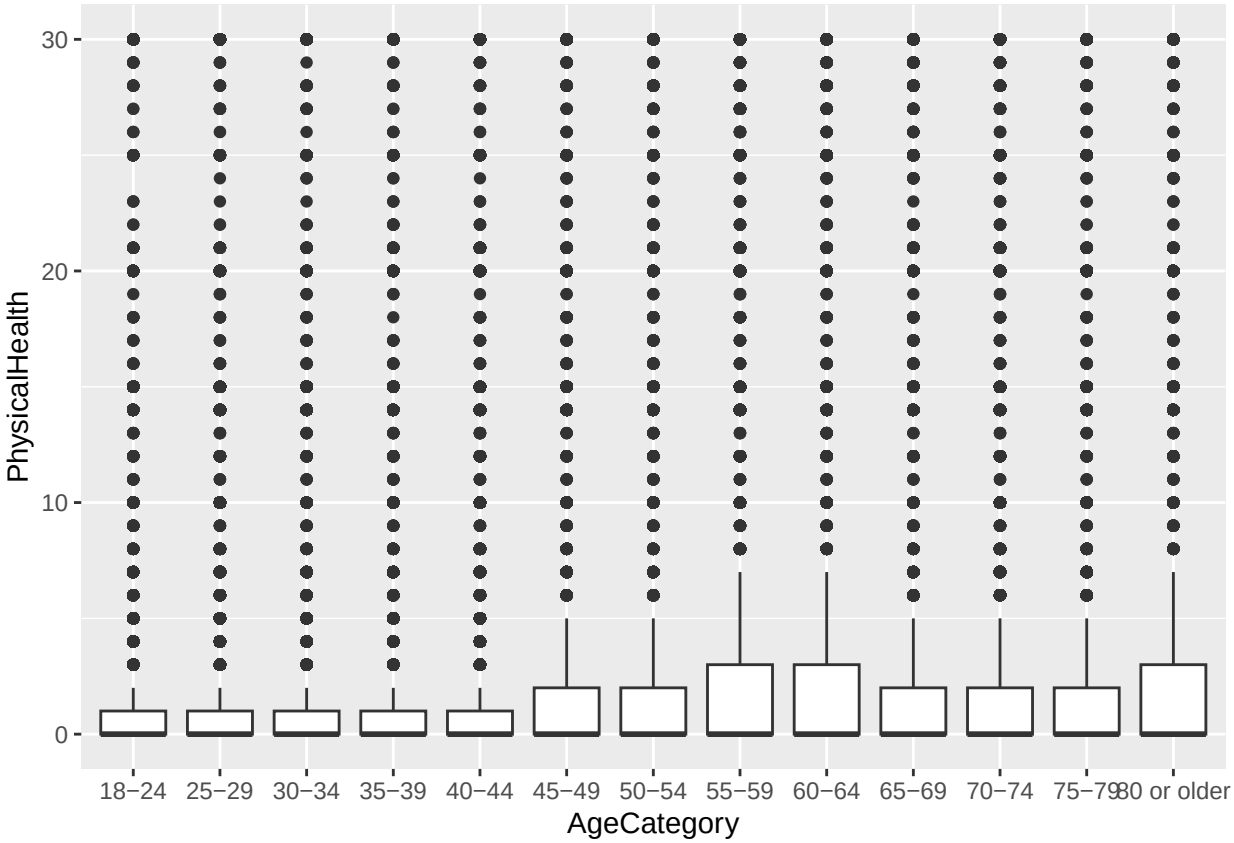


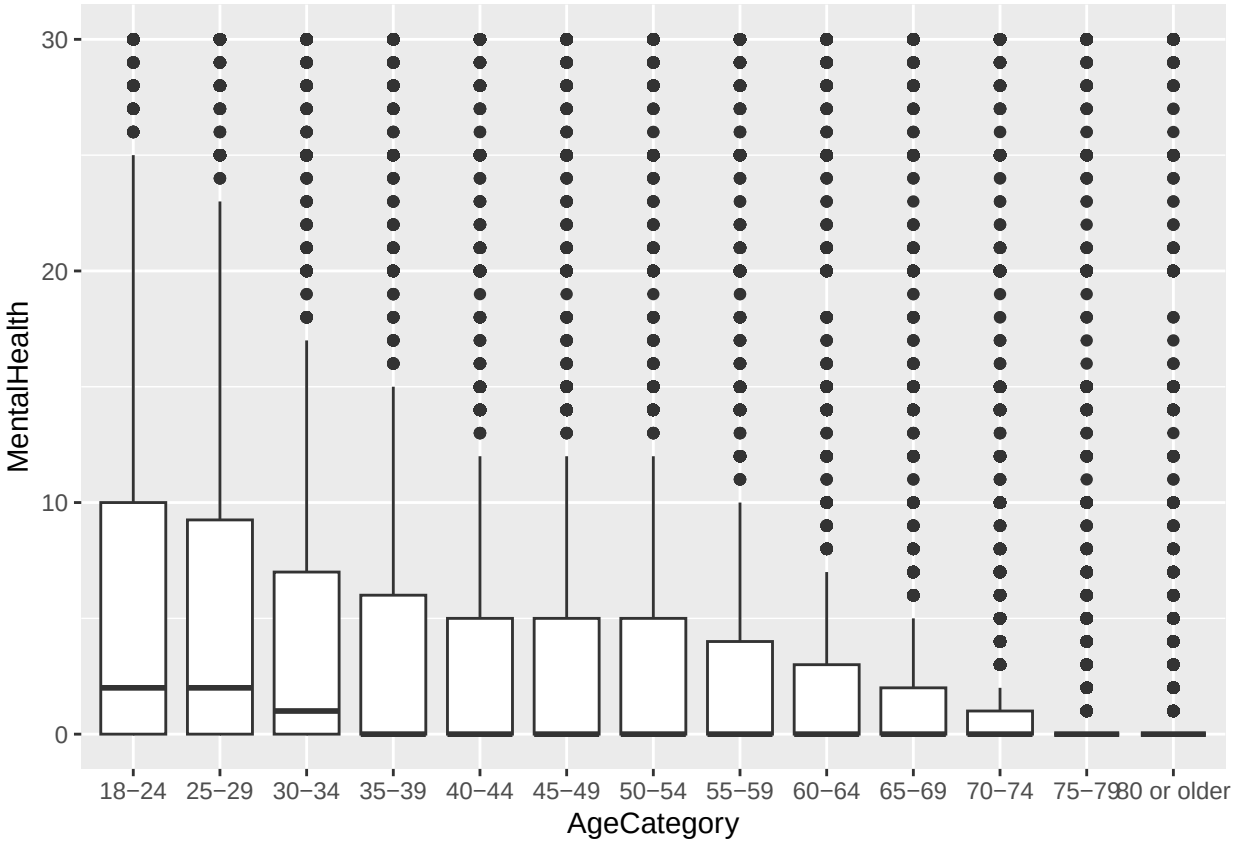

The boxplot of *SleepTime* reveals an uncommon phenomona that people who are 80 or older sleeps




more.

The boxplots indicates that younger generation suffer more from mental issues and older generation suffer more on physical problems.


30
20
10
0
PhysicalHealth
18−24
25−29
30−34
35−39
40−44
45−49
50−54
55−59
60−64
65−69
70−74
75−79
80 or older
AgeCategory




- Numerical and Categorical