Exploratory Data Analysis

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, 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.

From Figure 2, 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.

- Categorical Data:

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

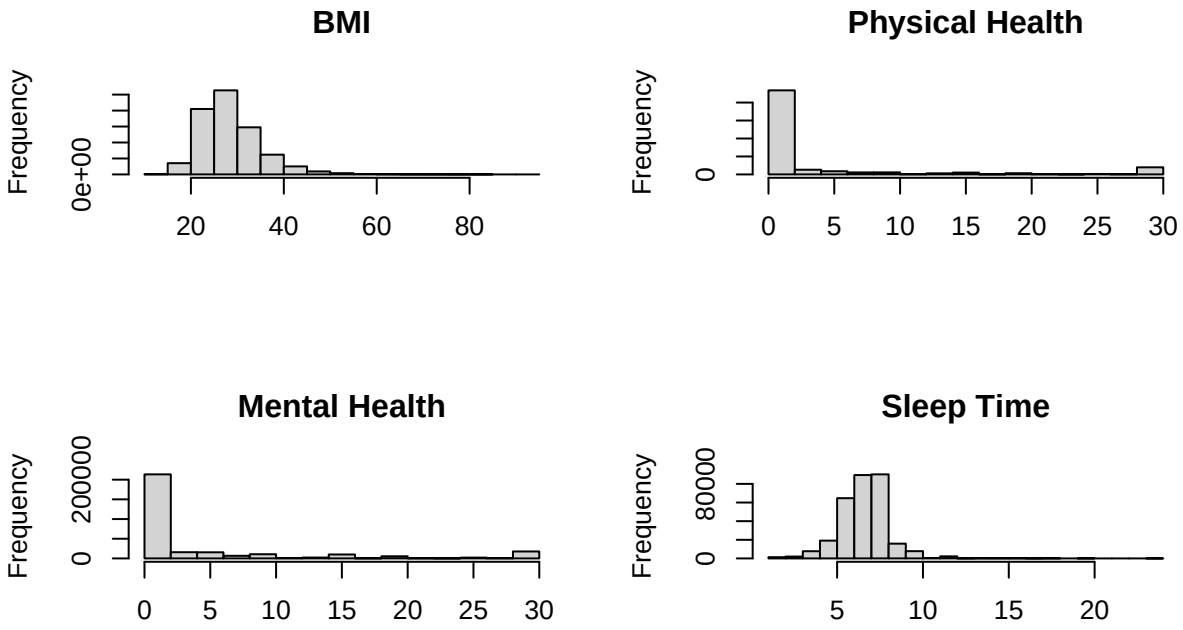


Figure 1: Histogram of Numerical Variables

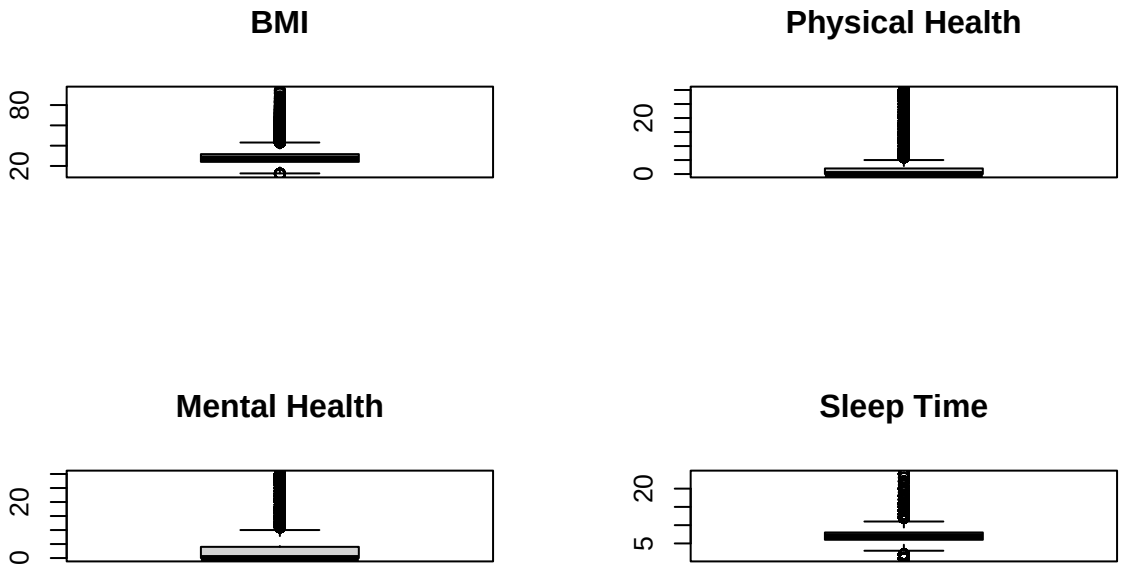


Figure 2: Boxplot of Numerical Variables

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 perform χ^2 test 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.

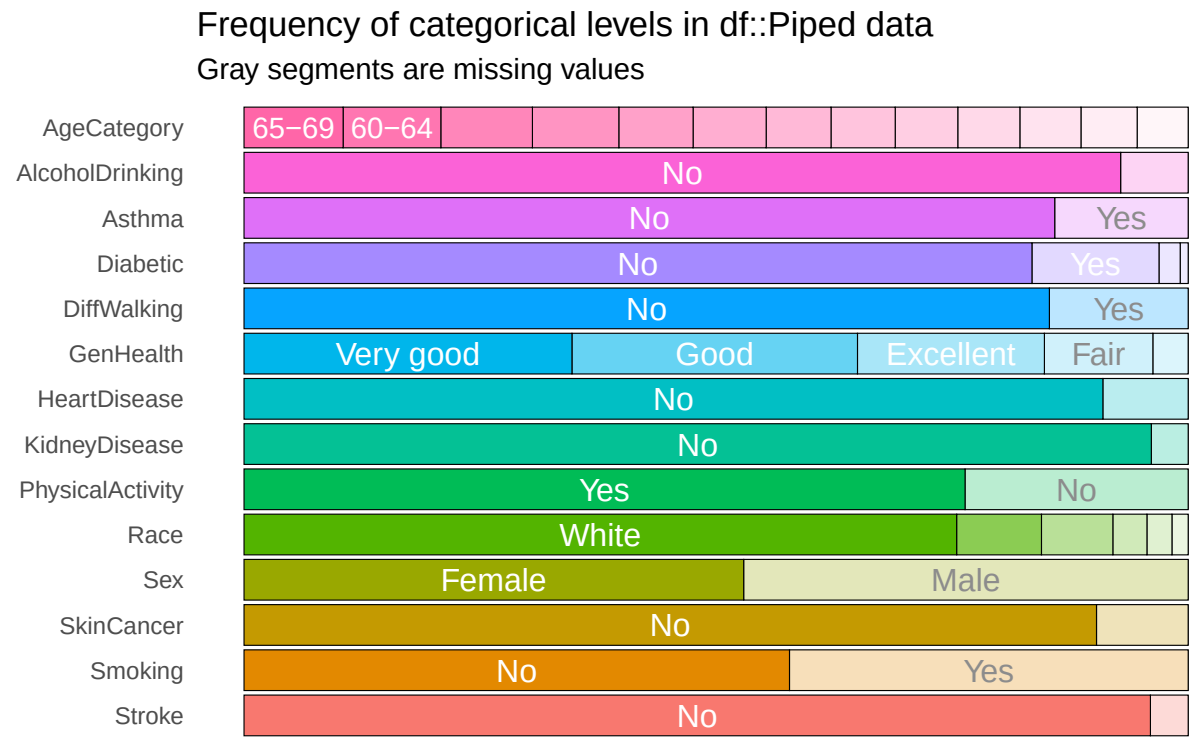


Figure 3: Frequency of Categorical Variables

Bivariate Analysis

- Categorical Data

We perform χ^2 test for correlation between *HeartDisease* and *AgeGroup* and *Sex* and *Race*:

Ho: *HeartDisease* is not correlated to *AgeGroup*.
Ha: They are correlated.
$\alpha = 0.05$
The p-value is much lower than 0.05, so we reject Ho, which suggests different groups has different probabilities to get heart disease.

```
##
##  Pearson's Chi-squared test
##
```

Table 2: Correlation Table

	BMI	PhysicalHealth	MentalHealth	SleepTime
BMI	1.0000000	0.1038125	0.0567245	-0.0486534
PhysicalHealth	0.1038125	1.0000000	0.2796575	-0.0584055
MentalHealth	0.0567245	0.2796575	1.0000000	-0.1170783
SleepTime	-0.0486534	-0.0584055	-0.1170783	1.0000000

```
## data:  xtabs(~data$HeartDisease + data$AgeCategory)
## X-squared = 18912, df = 12, p-value < 2.2e-16
```

Ho: *HeartDisease* is not correlated to *Sex*.
Ha: They are correlated.
$\alpha = 0.05$
The p-value is much lower than 0.05, so we reject Ho, which suggests the probabilities to get heart disease for male and female are different.

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~data$HeartDisease + data$Sex)
## X-squared = 1671.7, df = 1, p-value < 2.2e-16
```

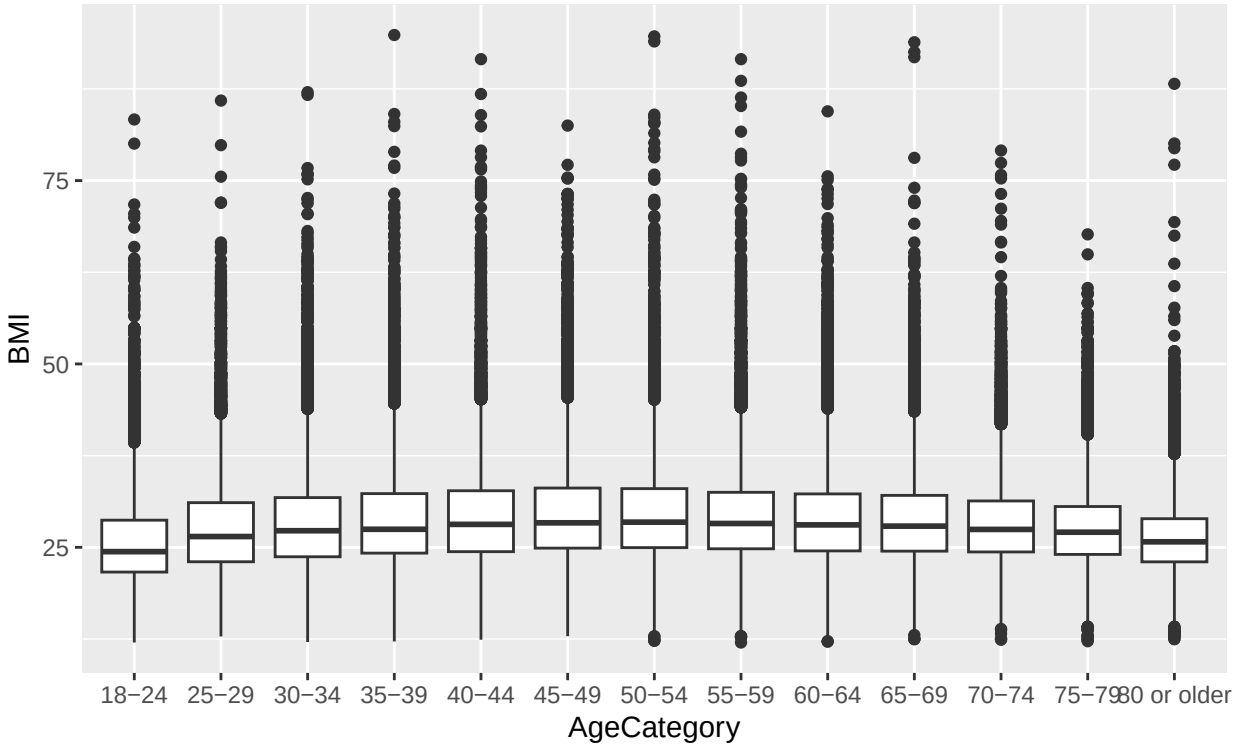
Ho: *HeartDisease* is not correlated to *Race*.
Ha: They are correlated.
$\alpha = 0.05$
The p-value is much lower than 0.05, so we reject Ho, which suggests the probabilities to get heart disease for white and non-white are different.

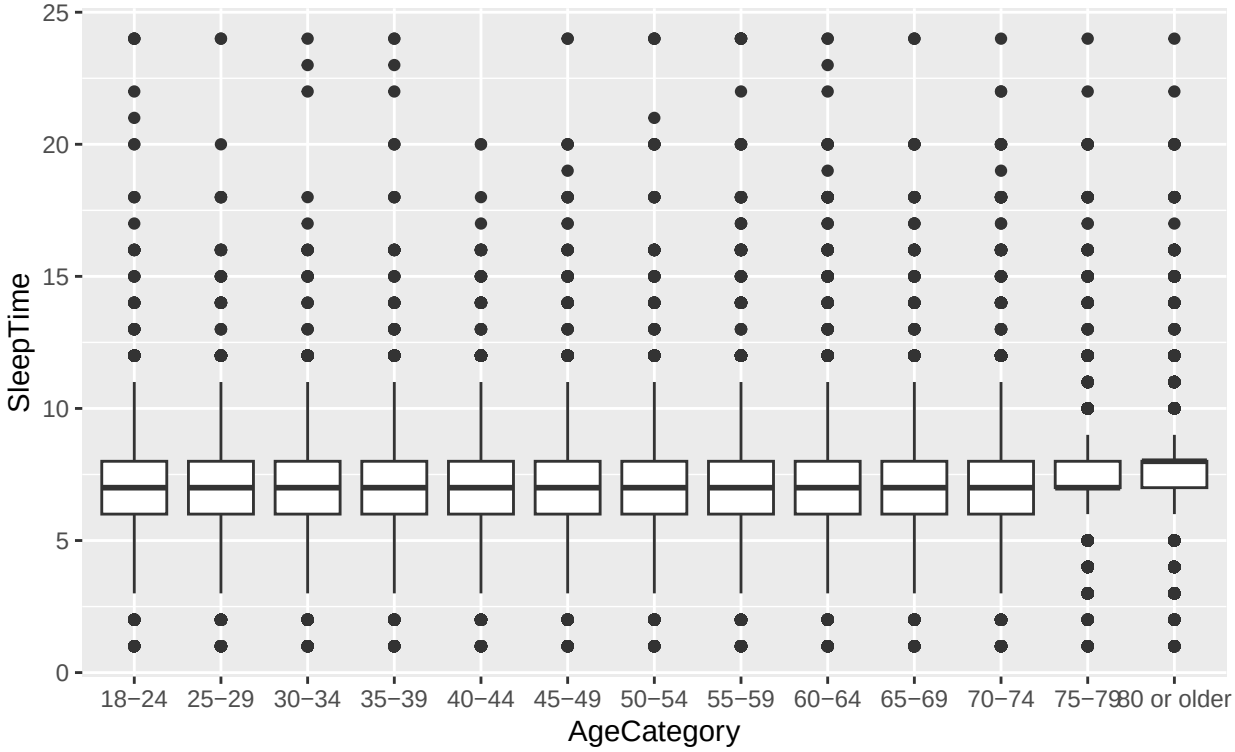
```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~data$HeartDisease + data$Race)
## X-squared = 721.24, df = 1, p-value < 2.2e-16
```

- Numerical Data

According to Table 2, there is no strong linear correlation among the numeric variables.

We then further analyze them in different age groups because we would love to reduce the number of age groups. Here is the summary for Figure 4. *BMI* goes higher as people get older and then after 65 it gets lower. *SleepTime* reveals an uncommon phenomona that people who are 80 or older sleeps more. Also, it indicates that younger generation suffer more from mental issues and older generation suffer more on physical problems.


BMI
75
50
25
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


SleepTime
25
20
15
10
5
0
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

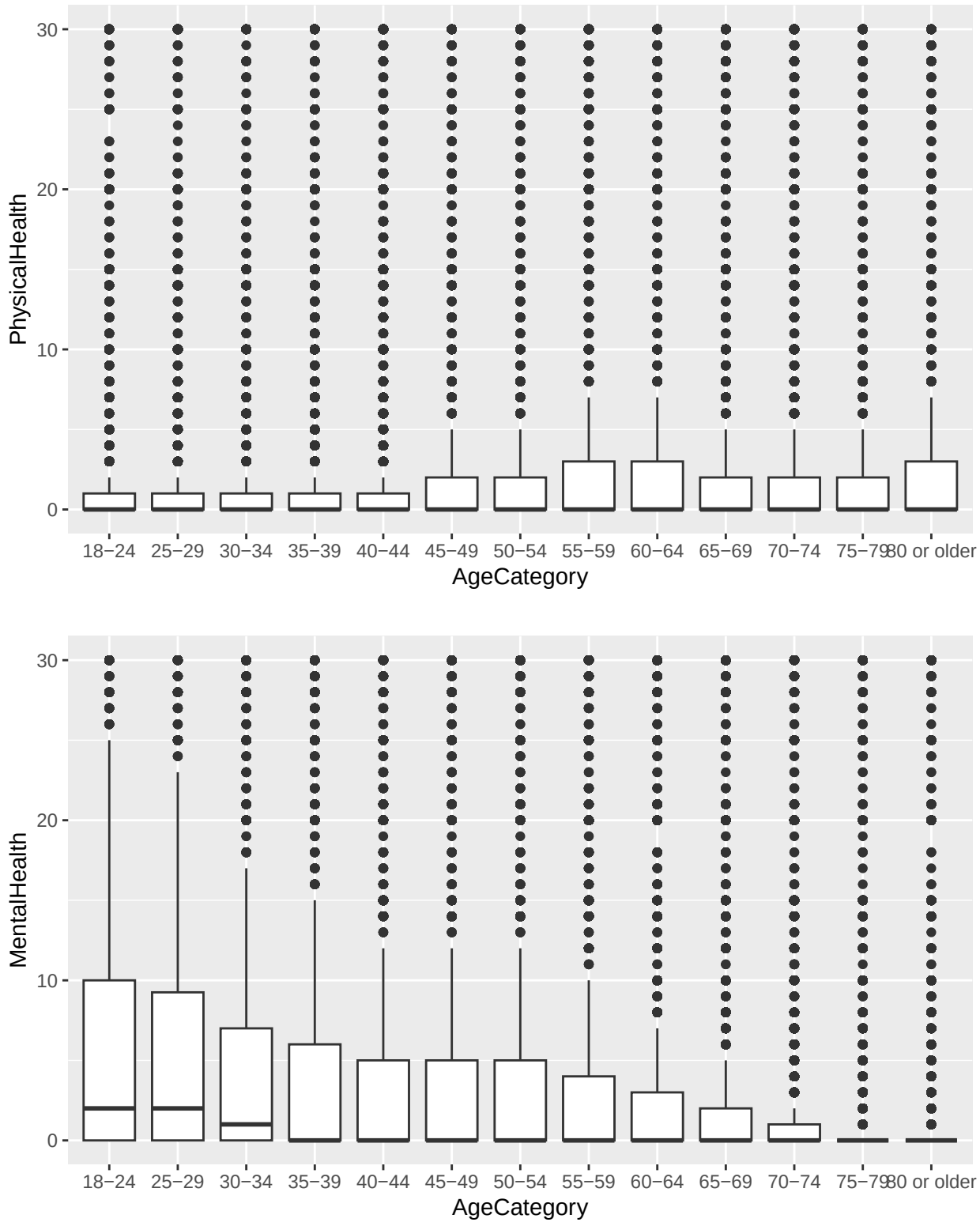


Figure 4: Numerical Variables among Different Age Groups

- Logistic Model

By fitting logistic regression, we can see that almost all the variables are significant for predicting *HeartDisease*. Surprisingly, *AlchoholDrinking* has negative association with *HeartDisease*.

Combine the result from the model and the Figure 4, we might consider to regroup the *AgeGroup* into only 4 groups (< 40, 40-59, 60-79, >80). Also, to find the vulnerable groups of getting heart disease, we may use decision tree, following we would introduce the models we plan to use for this project.

```
##
## Call:
## glm(formula = as.factor(HeartDisease) ~ ., family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1047  -0.4293  -0.2540  -0.1326   3.5881
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -6.4830608  0.1034943 -62.642  < 2e-16 ***
## BMI                             0.0081347  0.0011373   7.153 8.52e-13 ***
## SmokingYes                      0.3482008  0.0143796  24.215  < 2e-16 ***
## AlcoholDrinkingYes             -0.2716103  0.0334590  -8.118 4.75e-16 ***
## StrokeYes                       1.0304371  0.0225307  45.735  < 2e-16 ***
## PhysicalHealth                  0.0029129  0.0008598   3.388 0.000704 ***
## MentalHealth                    0.0040747  0.0008799   4.631 3.64e-06 ***
## DiffWalkingYes                  0.2074626  0.0180683  11.482  < 2e-16 ***
## SexMale                         0.7083199  0.0145780  48.588  < 2e-16 ***
## AgeCategory25-29                0.1236543  0.1241782   0.996 0.319357
## AgeCategory30-34                0.4917089  0.1110833   4.426 9.58e-06 ***
## AgeCategory35-39                0.6084139  0.1063690   5.720 1.07e-08 ***
## AgeCategory40-44                1.0164925  0.1000598  10.159  < 2e-16 ***
## AgeCategory45-49                1.3409679  0.0964953  13.897  < 2e-16 ***
## AgeCategory50-54                1.7561326  0.0931489  18.853  < 2e-16 ***
## AgeCategory55-59                1.9948168  0.0916947  21.755  < 2e-16 ***
## AgeCategory60-64                2.2575566  0.0908500  24.849  < 2e-16 ***
## AgeCategory65-69                2.4930843  0.0905818  27.523  < 2e-16 ***
## AgeCategory70-74                2.7692245  0.0905100  30.596  < 2e-16 ***
## AgeCategory75-79                2.9576256  0.0910387  32.488  < 2e-16 ***
## AgeCategory80 or older          3.2136495  0.0907803  35.400  < 2e-16 ***
## RaceWhite                       0.2009655  0.0185868  10.812  < 2e-16 ***
## DiabeticNo, borderline diabetes 0.0967987  0.0416679   2.323 0.020174 *
## DiabeticYes                     0.4549081  0.0166775  27.277  < 2e-16 ***
## DiabeticYes (during pregnancy)  0.0927621  0.1047582   0.885 0.375894
## PhysicalActivityYes             0.0354841  0.0160041   2.217 0.026610 *
## GenHealthFair                   1.4481710  0.0328406  44.097  < 2e-16 ***
## GenHealthGood                   0.9747177  0.0296685  32.854  < 2e-16 ***
## GenHealthPoor                   1.8457914  0.0408931  45.137  < 2e-16 ***
```

```
## GenHealthVery good               0.4479533  0.0305362  14.670  < 2e-16 ***
## SleepTime                       -0.0234114  0.0043126  -5.429 5.68e-08 ***
## AsthmaYes                        0.2596453  0.0191388  13.566  < 2e-16 ***
## KidneyDiseaseYes                 0.5572447  0.0243079  22.924  < 2e-16 ***
## SkinCancerYes                    0.0931997  0.0194867   4.783 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 183054  on 301716  degrees of freedom
## Residual deviance: 143259  on 301683  degrees of freedom
## AIC: 143327
##
## Number of Fisher Scoring iterations: 7
```

What scientific questions you will try to answer:

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

The statistical analysis techniques you will use to answer those questions (with justification)

The scientific questions above are more focused on statistical inference rather than performing predictions. To understand the relationship between those emphasized variables, we will both perform classification and regression on this dataset.

Classification: The classification is aimed at answering the vulnerable groups and how the variable will contribute to the diagnosis. The statistical method we choose is `Decision Tree Classification`, reasons are:

- **Model interpretation:** Though the decision tree may not provide the most accurate prediction, the boundary it gives is interpretable and fits well with our purpose of studying the contribution to heart disease.
- **Built-in variable selection:** The decision tree itself can help us with variable selection, which is important cause we want to know the most related variables
- **Capable of dealing with mixed variable type**: Our data set is a combination of numerical and categorical data, and `Decision Trees` handle those well.

- **A non-parametric aspect**: It will provide a non-parametric way aside from our regression method to be compared.
- **Perform well on imbalanced dataset**: The proportion of the heart disease in our data set is not balanced, but `Decision tree`'s hierarchical structure allows it to learn signals from both classes.

Regression: Here we will study our data with `Logistic Regression`, which is a pretty intuitive way.

- **Intuitive Regression Method**: Logistic regression is a widely accepted method in terms of building linear relationships between binary response and other predictors, with a range of well-developed statistical testing and evaluation techniques.
- **Marginal Interpretation**: The advantage of using a linear combination of the coefficients is interpretable and we could compare their coefficients to understand how each variable is counting for the heart disease diagnosis.
- **Probabilistic interpretation**: Rather than just give a classification result, the response is actually probabilistic, which will enhance our statistical argumentation.
- **Correlation Calibration**: By examining potential correlations and multilinearity among those variables, we could perform multiple trials on regression with different interaction terms.