Final Paper

STOR 320/520 Group 17

December 02, 2022

INTRODUCTION

With the improvement in people's living standards, more and more people prefer high-calorie food in their diet, resulting in physical problems such as obesity, dyslipidemia, high blood pressure, and elevated blood sugar. All of these are essential risk factors for the occurrence and development of diabetes. November 14 is United Nations Diabetes Day and this year's theme is "Diabetes Health Management for All". In recent years, the increasing incidence of diabetes and the increasingly younger age group of patients are alarming. The United Nations has repeatedly called on countries to focus on improving care for people with diabetes to control the disease and avoid complications. In response to the call, we found an interesting data set on a survey of basic lifestyle and diabetes.

Although no known cure for diabetes is developed today, experts have discovered that some lifestyles can lessen the disease's toll on patients. The most common strategies are losing weight, eating a healthy diet, being physically active, and receiving medical treatment can. Given that this data set is essentially a survey, we begin to consider how to make the survey more effective. Based on this we developed our first question: Can we create a shorter survey that only includes selected risk factors that can accurately predict whether an individual has diabetes or not? With shorter and cleaner questions, people will be more willing to spend time doing the survey.

Other than people's willingness of doing the survey, how to make the survey more precise and effective in predicting diabetes is also essential to our research. Here we developed our second question: What is the rubric of that short survey? In other words, whether a specific habit will lead to a higher or lower risk of having diabetes. By setting up a precise rubric, people can know whether their current lifestyle will bring a high risk of diabetes through the results of the questionnaire, so as to make some targeted adjustments.

Impacting billions of people around the world, diabetes is among the most prevalent chronic diseases not only in the country but all around the world. This seemingly insignificant condition always comes with very serious complications. Even worth, the treatment of diabetes is now aggravating the burden of social medical resources, creating serious burdens on the economy. While attempting to prevent it in advance, our shortened serves and precise rubric would provide people with a clear view of what to do and what not to do. We examined these two crucial questions in detail with an authoritative data set by suggesting some valid models.

DATA

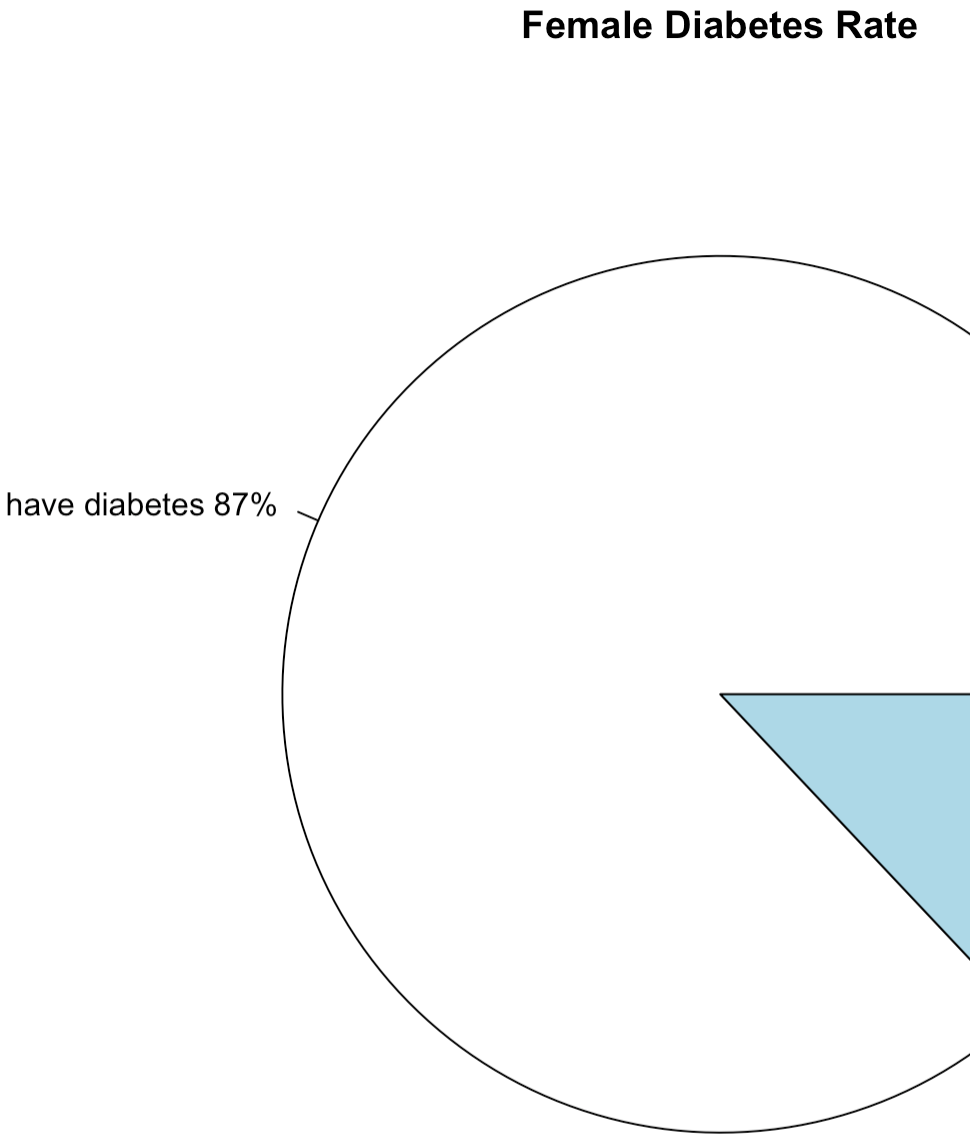
We analyzed the data from the Behavioral Risk Factor Surveillance System (BRFSS), a system of surveys conducted by the CDC. From Kaggle, we selected a dataset from the 2015 survey with 253680 responses from the US population. It combined the binary diabetes target variable, diabetes-risk related-behaviors variables, and demographic variables. The dataset looks as this table:

Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies
1	1	1	1	30	1	0	1	0	1	1
0	0	0	1	24	0	0	0	0	0	1
1	0	0	1	25	1	0	0	1	1	1
0	1	1	1	34	1	0	0	0	1	1

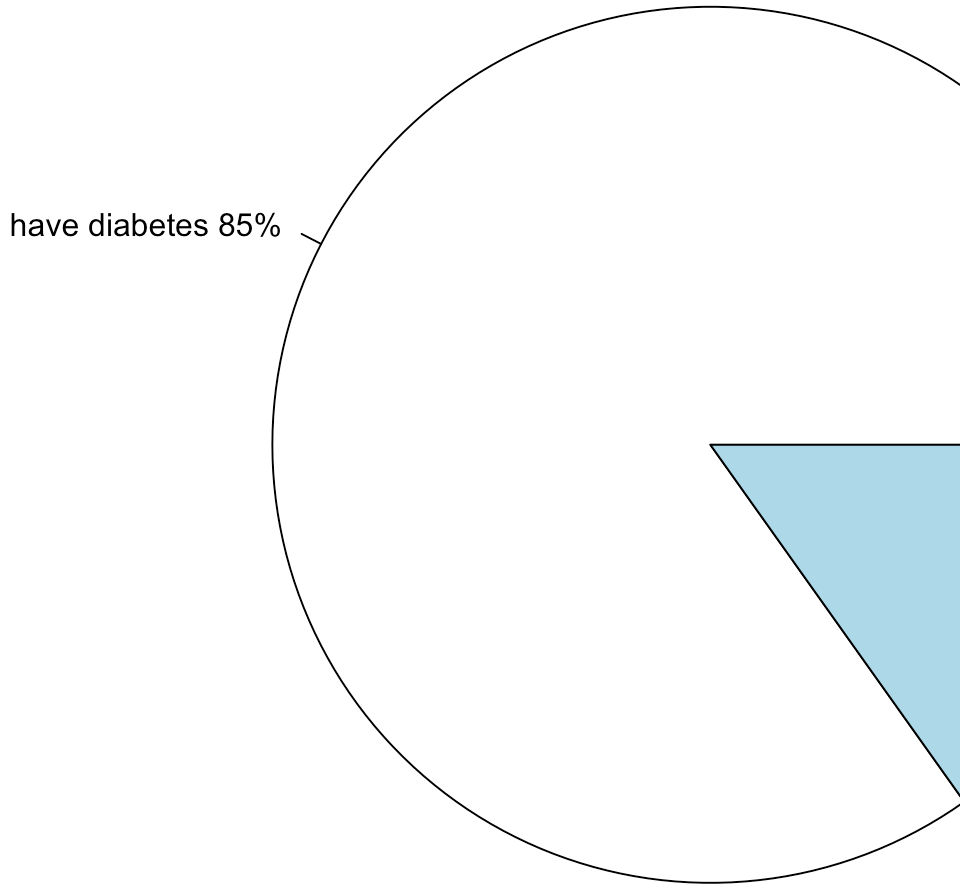
HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	0	5	30	30	1	0	9	5	1
0	1	0	2	0	0	0	1	8	4	3
0	1	0	3	0	0	0	1	13	6	8
0	1	0	3	0	30	1	0	10	5	1

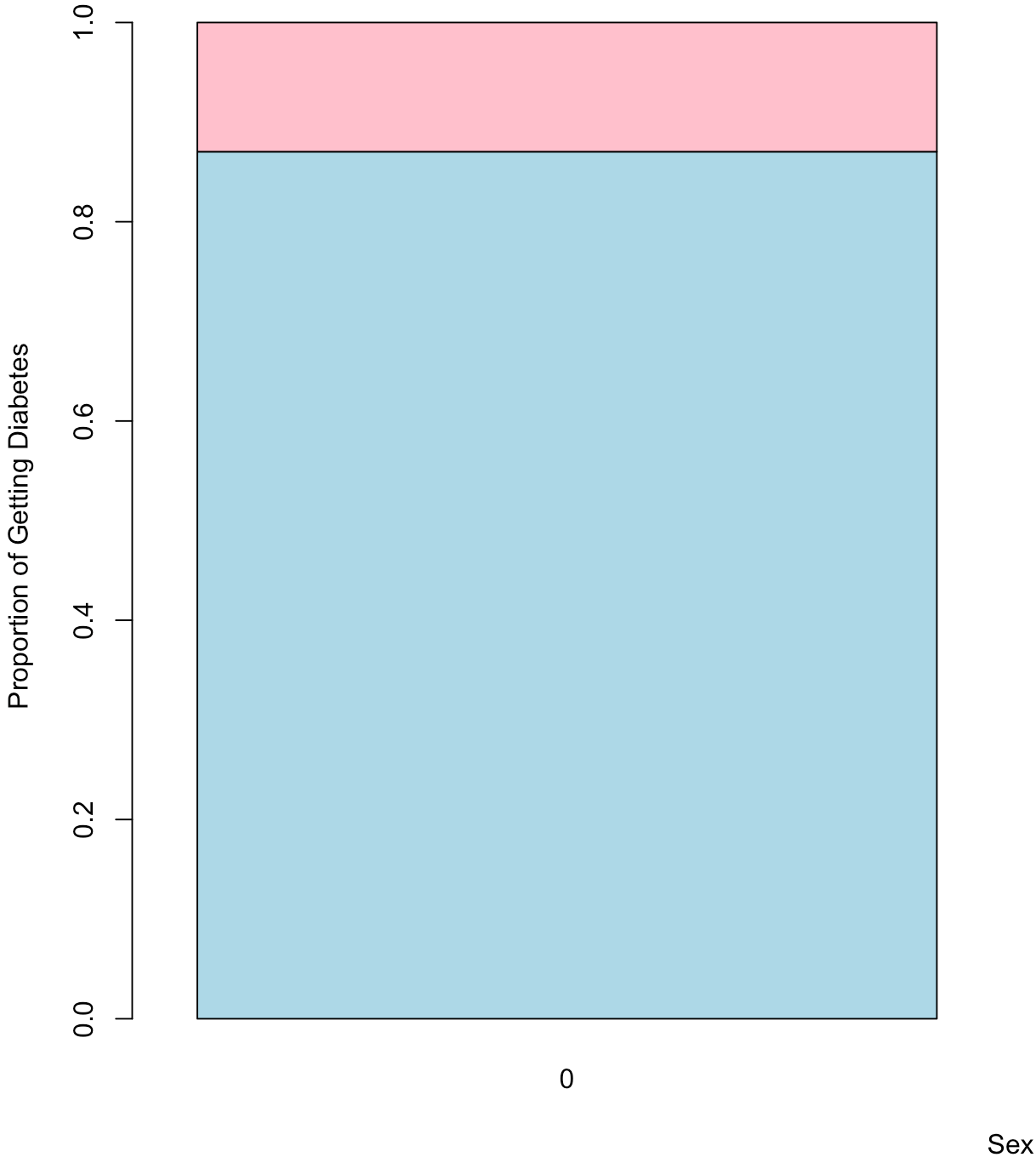
We used all the variables in the original dataset. Most of our response variables are binary(0 incidates no, 1 indicates yes): - High Bloody Pressure & High cholesterol, Stroke, Heart disease, confirmed by health professionals - Smoker(if the respondents smoked at least 100 cigarettes in their entire life) - Obesity(BMI is 30 or greater) - Physical activity(if the respondents do physical activity or exercise during the past 30 days other than their regular job) - Fruit, Veggie(consume fruit/veggie one or more times per day) - Heavy Alcohol Consumption(adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) - Any Healthcare(if the respondents have any kind of health care coverage) - No Doctor Because of Cost(if the respondents have a time in the past 12 months when you needed to see a doctor but could not because of cost) - Difficulty in walking(if the respondents have serious difficulty

walking or climbing stairs) - Sex(0 is female and 1 is male)



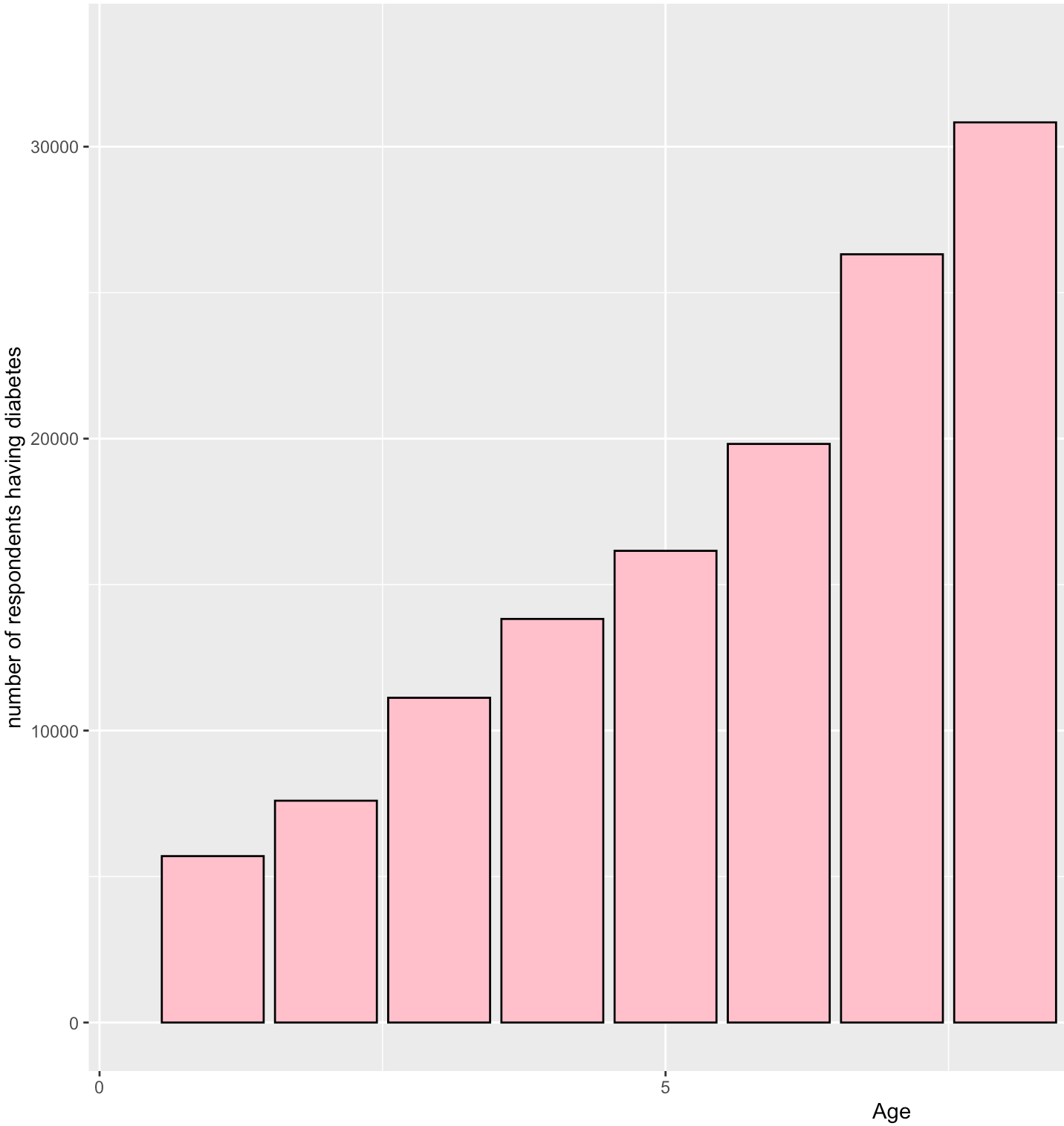
Male Diabetes Rate





As a particular example of what the categorical variables look like in our dataset, according to the pie charts, male have a 2 percent higher chance of getting diabetes than the female. Combining the second bar chart, we conclude that sex does not affect the chance of getting diabetes obviously.

Other variables are non-binary. - BMI - General Health Level(scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor) - Mental Health Level(how many days during the past 30 days was the respondents has a not good mental health condition, including stress, depression, and problems with emotions) - Physical Health Level(how many days during the past 30 days was the respondents has a not good physical health condition, including physical illnesses and injuries) - Education(the highest grade or year of school the respondents completed) - Income(Income scales 1 = less than 10000 dollars, 5 = less than 35000 dollars, 8 = $75000 or more) - Age(in a fourteen-level age category)



As a particular example of what the non-binary variables look like in our dataset, for age value at 9, there are the most people have diabetes. In this data set, at age value 9 means a age range from 60 to 64.

RESULT

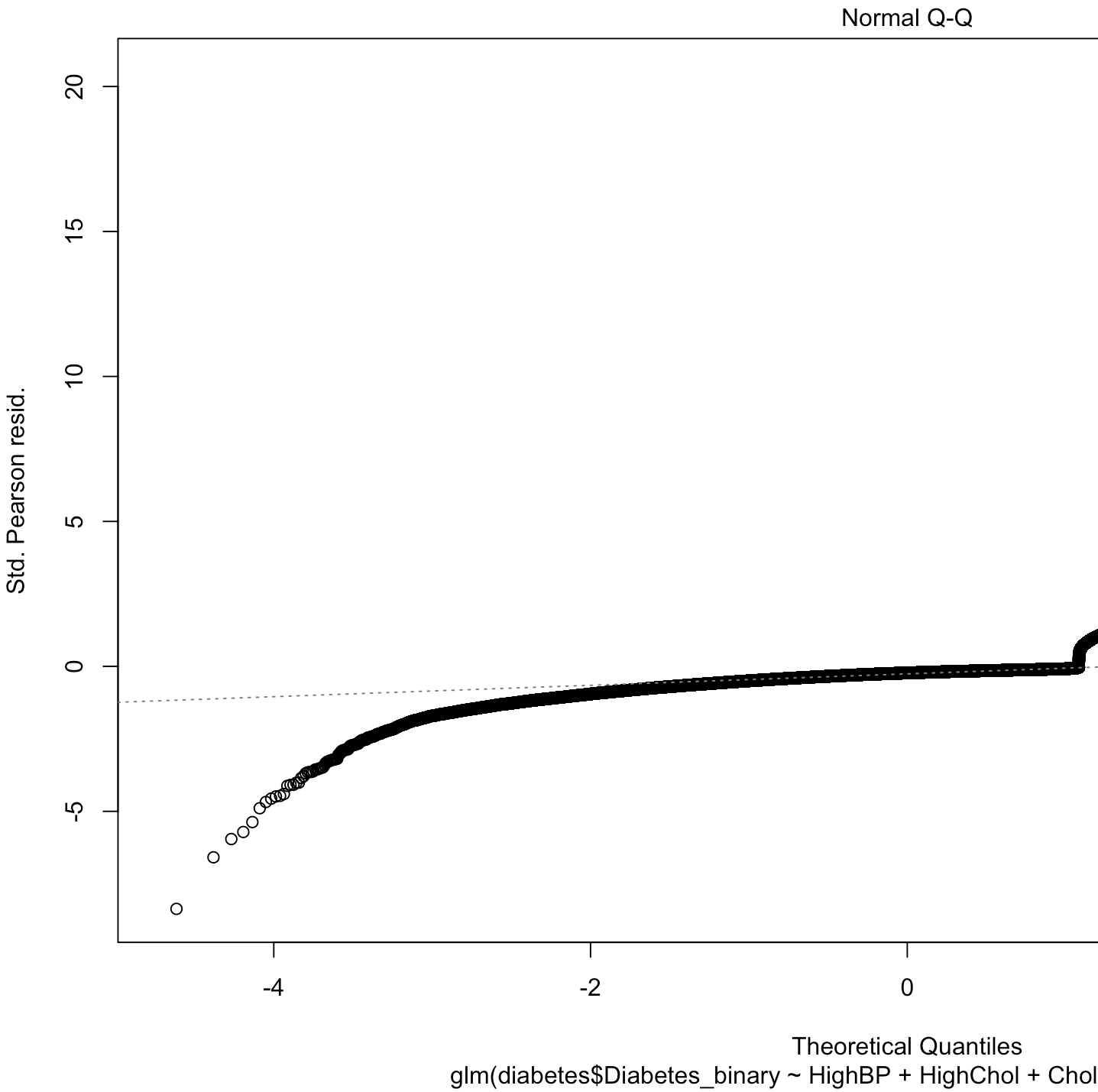
Question 1

To solve the first question, we choose the stepAIC model selection method. Since we are aiming to make a shorter survey to alleviate the time-consuming problem while not greatly influencing the accuracy of the survey, we choose the stepAIC model selection method. Different from other model selection methods, StepAIC does not necessarily mean to improve the model performance, however, it is used to simplify the model without impacting much on the performance. Therefore, the value of AIC (Akaike Information Criteria) quantifies the amount of information loss due to this simplification, which means we need to find a model with minimal AIC.

StepAIC method helped us eliminate two risk factors: NoDocbcCost, which means whether the respondents had a time in the past 12 months when they needed to see a doctor but could not because of cost, and Smokers, which means whether the respondents had smoked at least 100 cigarettes in their entire life.

Degrees.of.Freedom	Total	Residual.Null.Deviance	Residual.Deviance	AIC
253679	253660	204800	162200	162200

After taking off these two predictors, according to the table aborve, our model's AIC value decreases from 162227.6 to 162200. Therefore, our model becomes shorter without huge impacts on the model's performance.



We also draw a normal Q-Q plot to access our model. According to the normal Q-Q plot above, there is large skewness on the right tail of the plot and there is also relatively small skewness on the left tail of the plot. Hence, the normal Q-Q plot shows that the normal distribution condition of our model is not met. Even though our model can roughly predict the probability of

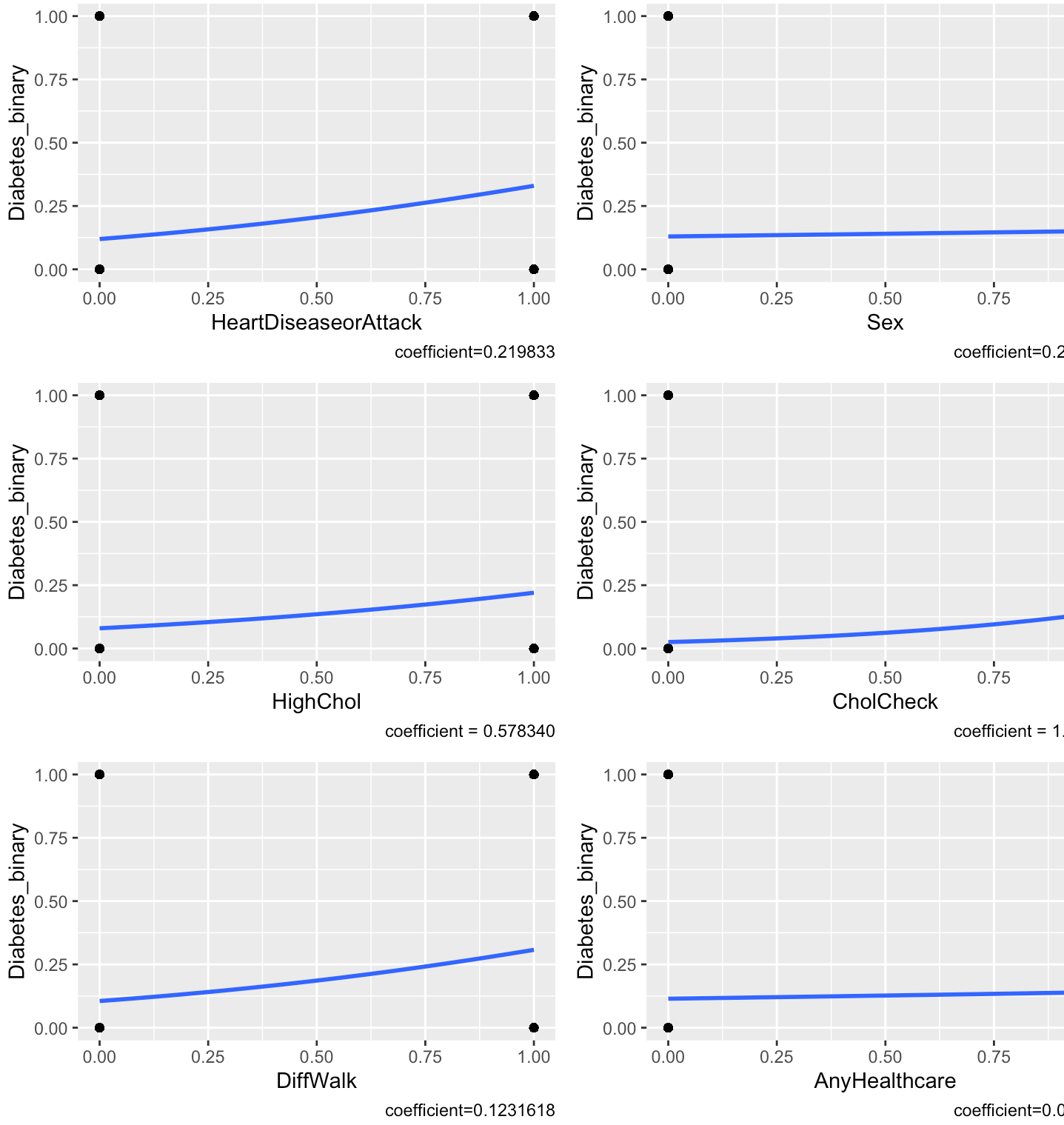
respondents according to their responses on the shorter surveys, there are a lot of other risk factors needed to be considered to precisely predict the odds. Furthermore, the relationship between the chance of getting diabetes and other risk factors is not simply linear. We have tried different ways to transform the model. However, all these methods did not improve the performance of our model. In the future study, we need to do more transformations on the model to improve model's performance better.

Question 2

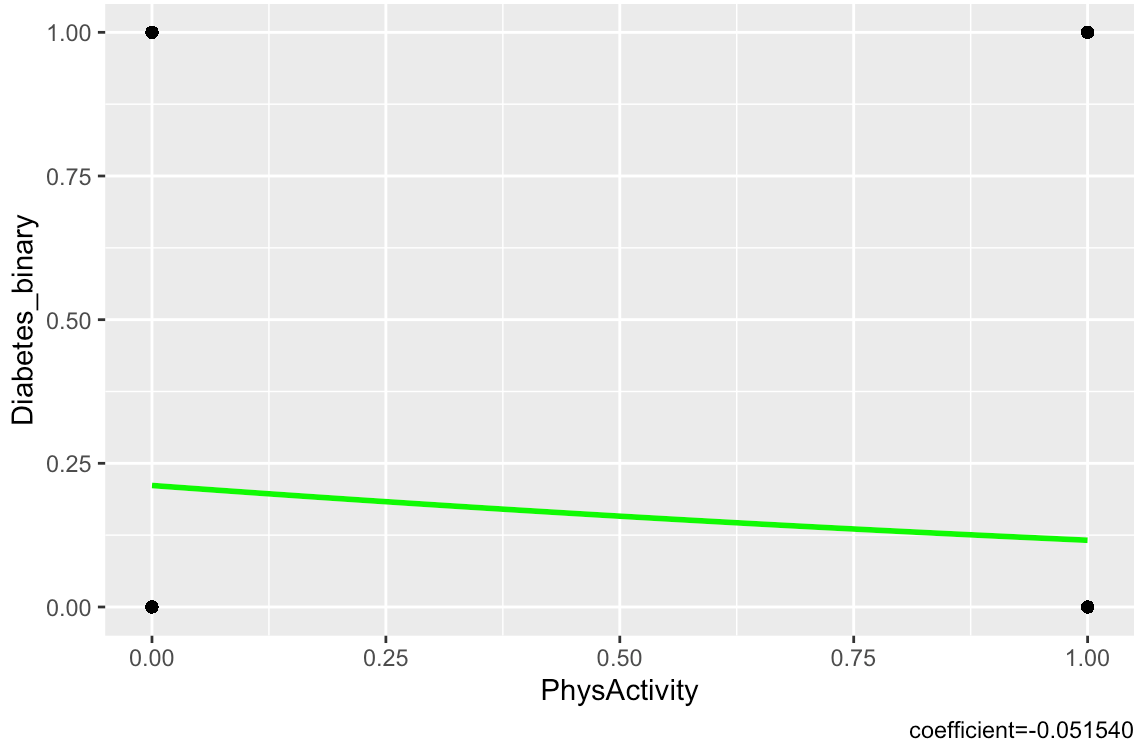
Our second question was, "What is the rubric of our survey?" In the first question, the model selection method was utilized to determine the optimal model, which consists of 19 predictor variables. The objective is to establish a different model between these 19 and our response variable Diabetes_ binary. To facilitate a more intuitive understanding of the link between variables, we produced a total of 19 diagrams to illustrate the relationship between the specific predictor variable and the response variable. After determining the link between 19 predictor variables and responder variables, we may design a diabetes prevention survey based on their slope.

Then we approached this question by dividing the rest of the predictor variables into two groups: quantitative variables and categorical variables. For each categorical variable, we developed a logistic model of the categorical variable in each predictor variable separately from our response variable Diabetes binary and plotted the results using ggplot. The final curves were drawn with ggplot. And it is evident from the curves whether the two variables are positively or negatively correlated.

The combined graph below includes all models with a positive slope. In the diagram, HeartDiseaseorAttack, HighBP, HighChol, CholCheck, Stroke, Sex, and CholCheck are depicted, indicating that the higher their values, the greater the likelihood of acquiring diabetes.

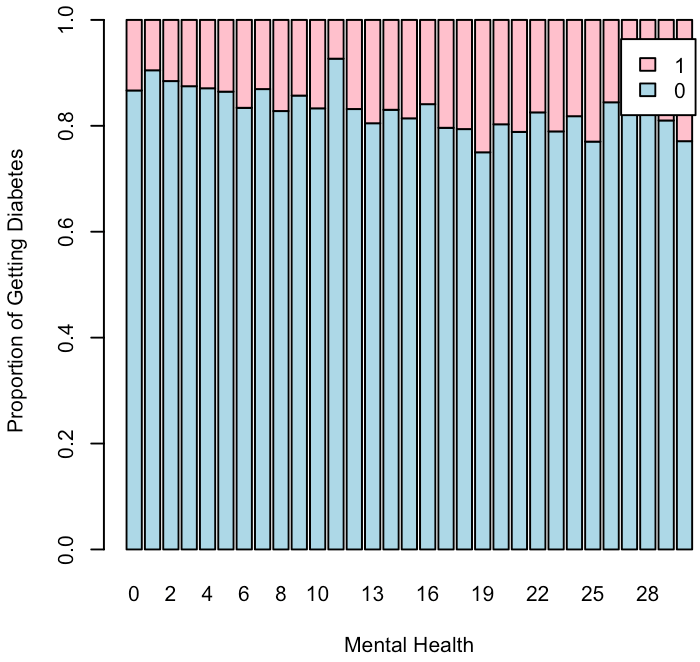


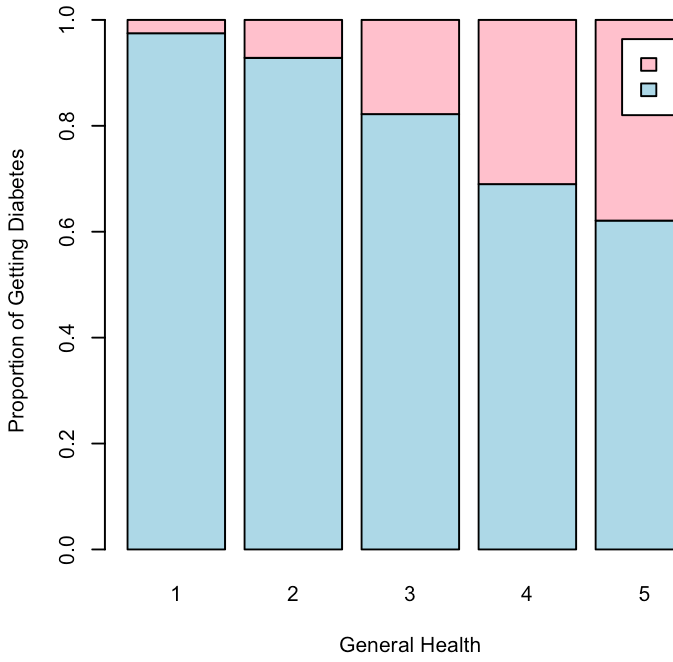
The graph below summarizes all models with negative slopes. As seen in the illustration, Fruits, Vegetables, Physical Activity, and High Alcohol Consumption are the four categorical variables that are negatively correlated with Diabetes_binary. This means that the greater the value of the category factors, the less likely an individual is to get diabetes.

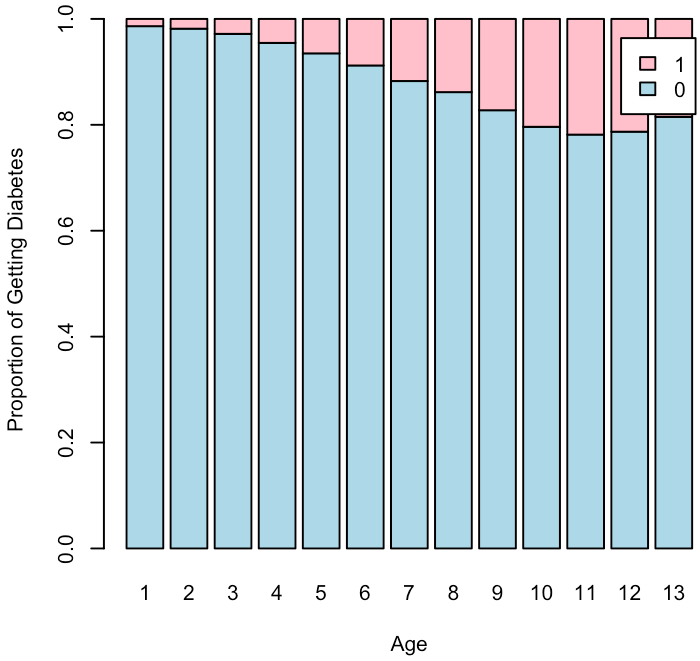


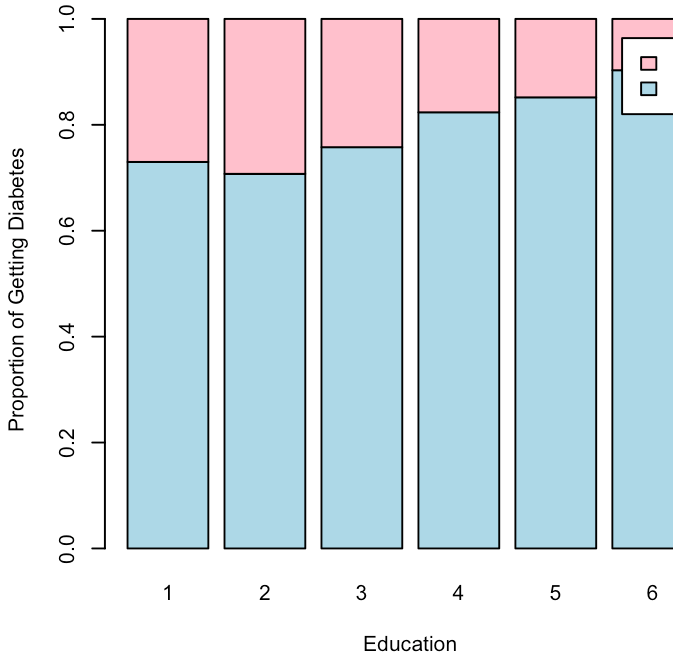
Immediately afterward, we created six separate bar charts for the remaining six quantitative variables in addition to BMI and our response variable. Each bar chart was divided into two regions, with the pink being the percentage of people who had diabetes for each specific condition and the blue being those who did not, and our goal was to identify a broad trend. From

the six bar charts integrated below, we can draw the following conclusions: 1. For GenHlth, as its value gets larger, the proportion of people who have diabetes gets larger. The worse the overall health status, the more likely to get it. 2. For MentHlth, we cannot see a definite trend, but the overall trend is flat and not very suitable for reference. 3. For PhysHlth, as its value increases, the proportion of people with diabetes increases. The worse the state of health, the more likely to get it. 4. For Age, the older the person is, the more likely he or she is to have diabetes. 5. For Education, the more educated people are, the less likely they are to have diabetes. 6. For Income, the higher the income, the less likely people are to have diabetes.

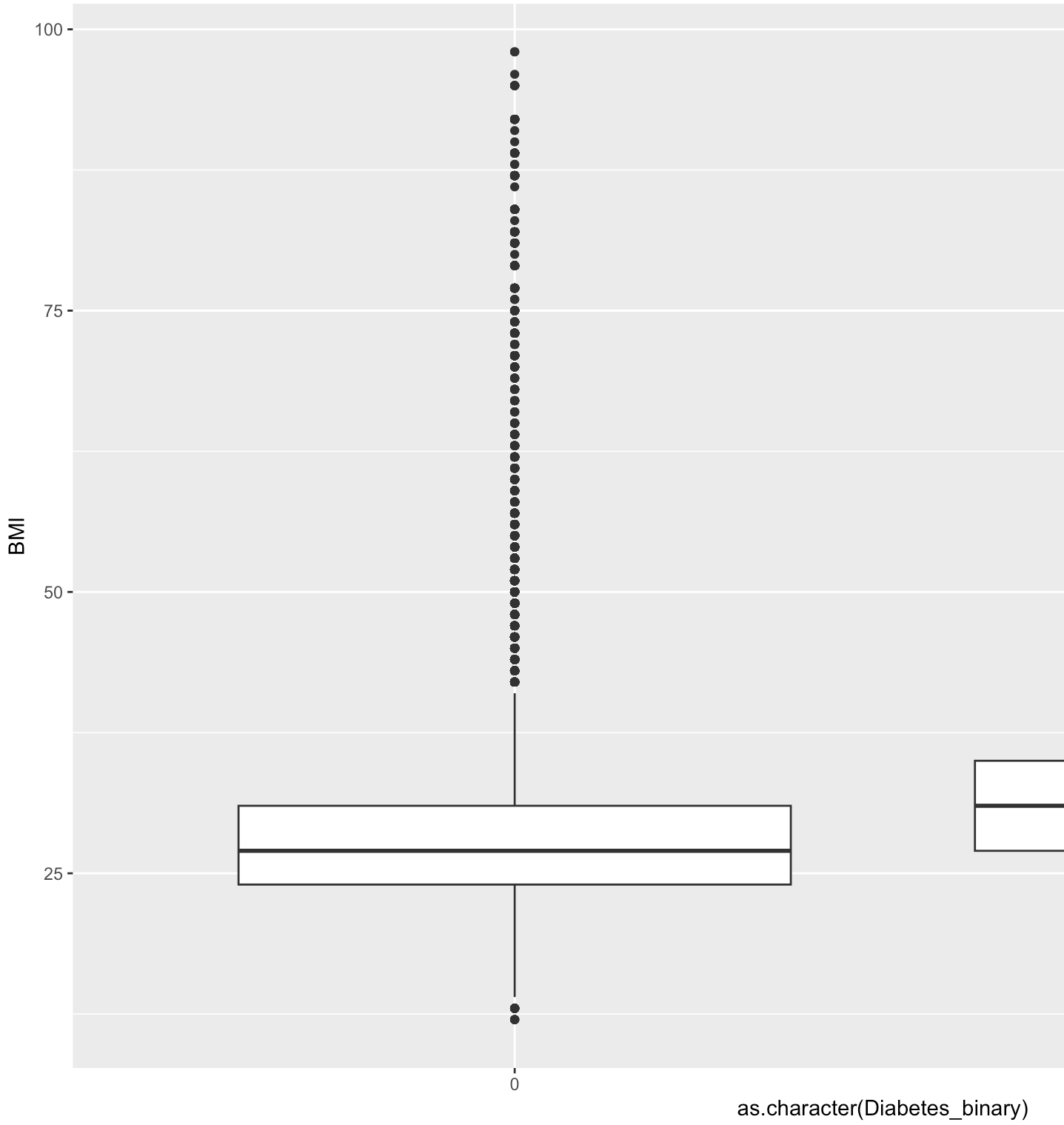








As for the last quantitative variable BMI, we made a box plot for it and our response variable Diabetes_binary separately because the range of its data is too large. And 25% of the lower quartile and 75% of the higher quartile also have larger BMI values. So we can see from this sample that the greater the BMI, the greater the probability of having diabetes.



CONCLUSION

When we answered our first question, we built a stepAIC model to select variables that significantly impact the possibility of getting diabetes. We could conclude from the model that whether someone had time in the past 12 months when they needed to see a doctor but could not because of cost and whether someone smokes do not play a significant role when we try to create a shorter survey to predict the chance of getting diabetes more accurately. This does not mean that these two factors are entirely uncorrelated with diabetes, but it means that there does not exist a direct causal relationship between these factors and diabetes. This improved model could show a greater accuracy with decreased AIC value, but still need further improvement, such as transformation on normality according to the Q-Q plot.

Moreover, we further explore the detailed rubric of the short survey we created in the first question to provide more information about the survey. We found that having heart disease, high BP, high cholesterol, having a cholesterol check in 5 years, having a higher BMI value, having a stroke, being a female, having better general health, better physical health, and being older would lead to a positive impact on the possibility of getting diabetes. In contrast, eating fruits every day, doing less physical activity, having higher alcohol consumption, and having a higher education level and higher income would lead to a negative impact on the possibility of getting diabetes. At first, we might think that it is unusual that drinking more alcohol leads to a lower possibility of getting diabetes, as we commonly believe that alcohol harms one's health condition. From research, we found that drinking more alcohol does lead to a lower possibility of getting diabetes because excess alcohol can actually decrease blood sugar levels. However, drinking alcohol can worsen diabetes-related medical complications, such as disturbances in fat metabolism, nerve damage, and eye disease. This result could be significant in proving the medical theory and the role of alcohol plays in the formation of diabetes.

Both of the questions and their answers would be helpful for exploring the causes of diabetes in the medical area in real life and for everyone who is a potential patient of diabetes. The first question helps us to target factors that have a significant relationship with diabetes so that researchers can further investigate the underlying mechanisms of the body and also the social influences on the chance of getting diabetes. Also, the shortened survey is more efficient and less time-consuming for people who want to know their chance of getting diabetes from their lifestyle. The detailed rubric of the survey suggests a relatively healthier lifestyle with fewer possibilities of getting diabetes for people who have concerns about getting diabetes. The importance of preventing diabetes will continue to rise as the increasing incidence of diabetes and the increasingly younger age group of patients.

While the dataset is generally cohesive, we believe a few additions or changes could strengthen the analysis. For example, there is another dataset with three conditions of the variable of diabetes, including no diabetes, prediabetes, and diabetes. Prediabetes is a serious health condition. People with prediabetes have higher blood sugar than normal, but not high enough yet for a diabetes diagnosis. However, being diagnosed with prediabetes does not mean developing type 2 diabetes, especially if you follow a treatment plan and make healthy lifestyle choices. Therefore, instead of only diabetes and no diabetes, the more detailed conditions could illustrate the impact of different factors on diabetes conditions in different stages and give more information for people with prediabetes to improve a healthier lifestyle.