Data Analysis, Cleaning and Feature Generation of a Diabetes Dataset

Seif ElAnsary†  
 CSCE Department  
 The American University in Cairo  
 ID:900221511  
 seif\_elansary@aucegypt.edu

MagdElDin AbdalRaaof  
 CSCE Department  
 The American University in Cairo  
 ID: 900211145  
 MagdElDin@aucegypt.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

DATASET

In accordance with the previous report, we chose Alex Teboul’s diabetes dataset due to the fact that it is a large and a highly descriptive dataset. It contains 253,680 complete rows and 22 columns that are not empty for any row. This report aims to analyse each column, and the rationale behind the data processing of each one if it was processed.

ANALYSIS

1. Diabetes

The distribution of the Diabetes column is not uniform. A Diabetes score of zero means that the individual does not have diabetes. A Diabetes score of one entail that the individual is prediabetic and a score of two means that the person has diabetes.

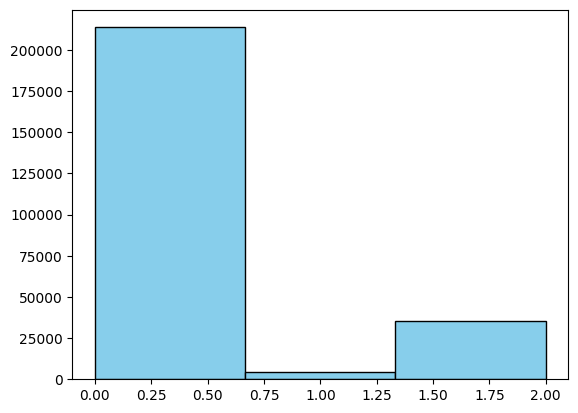


Figure 1

As shown in figure 1, around 80% of the people in the dataset are not diabetic. Even though the data is skewed a bit towards non-diabetic features, looking at the ratios and analysing columns with respect to each diabetes class will provide insightful data to train the model on.

1. High Blood Pressure

The HBP column is a binary classifier that labels people with blood pressure above a certain threshold as having high blood pressure (1), else they are classified as not having high blood pressure (0).

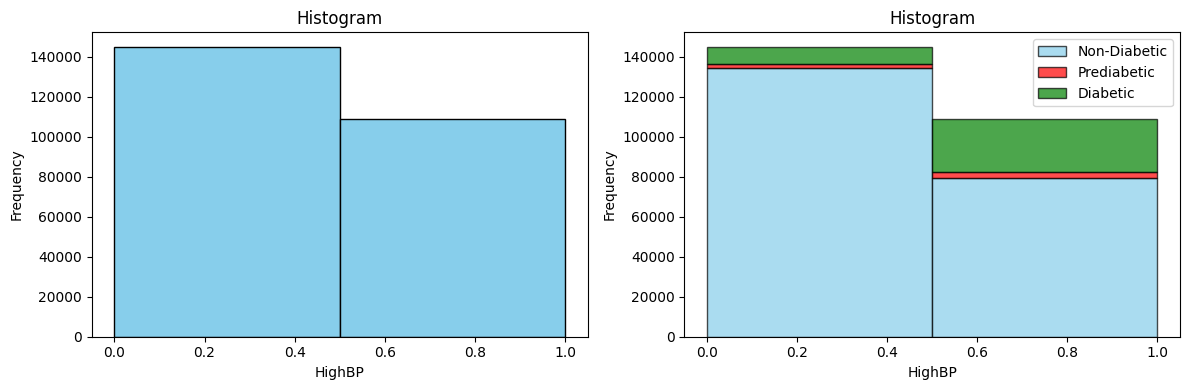


Figure 2

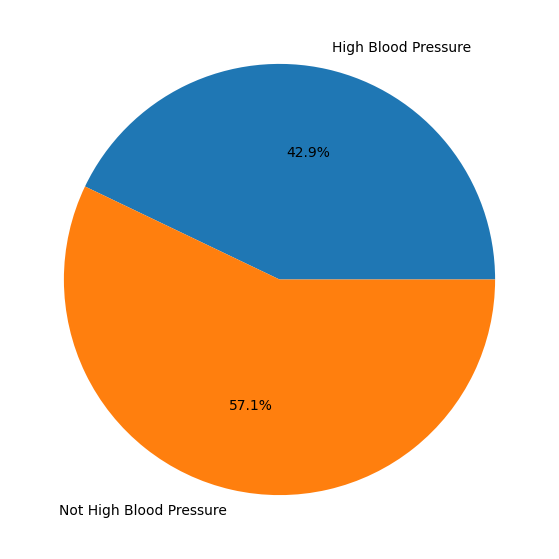


Figure 3

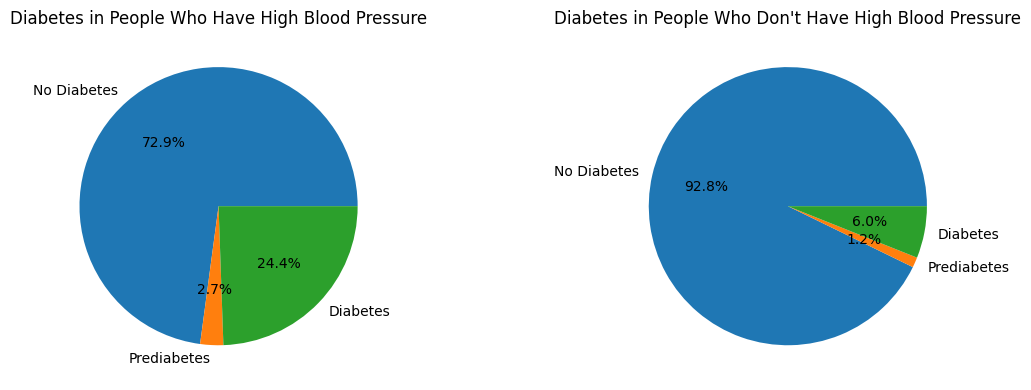


Figure 4

As shown by the pie charts, HPD is a significant indicator of whether a person has diabetes or not.

1. High Cholesterol

This column is also a binary classifier that is 0 if they individual does not have high cholesterol and is 1 if the individual’s cholesterol level is beyond a threshold.

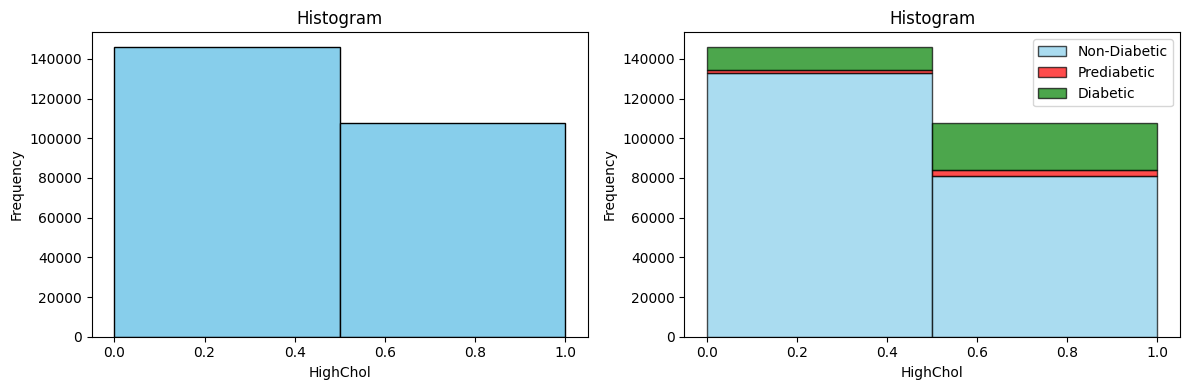


Figure 5



Figure 6

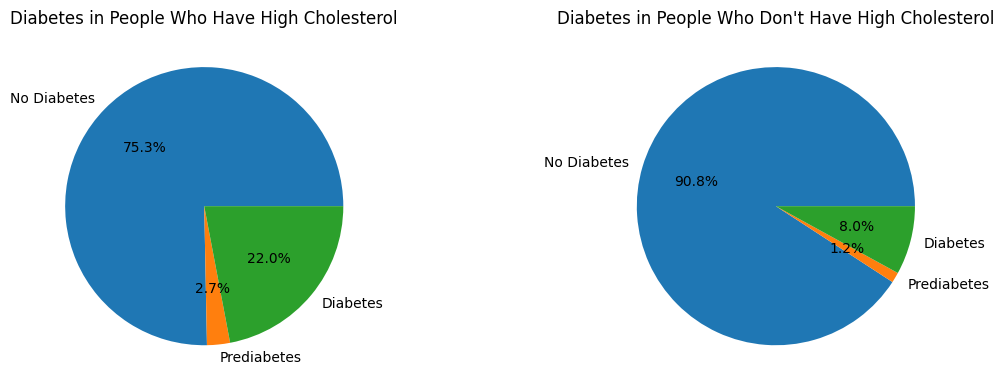


Figure 7

Figure 7 clearly shows that individuals which have high cholesterol levels increases the risks of diabetes and prediabetes.

1. **Cholesterol Check**

This column is a binary classifier that checks if a person checked their cholesterol levels during the last 5 years. Yes is 1 and no is 0.

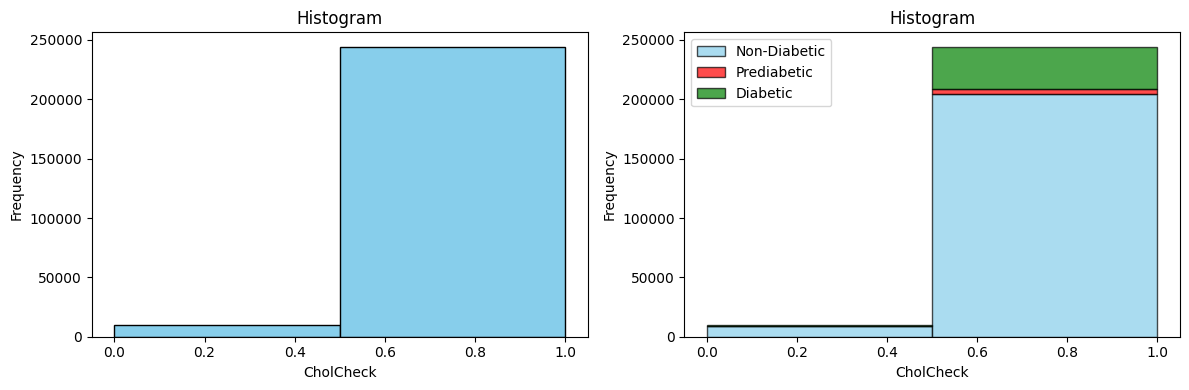


Figure 8

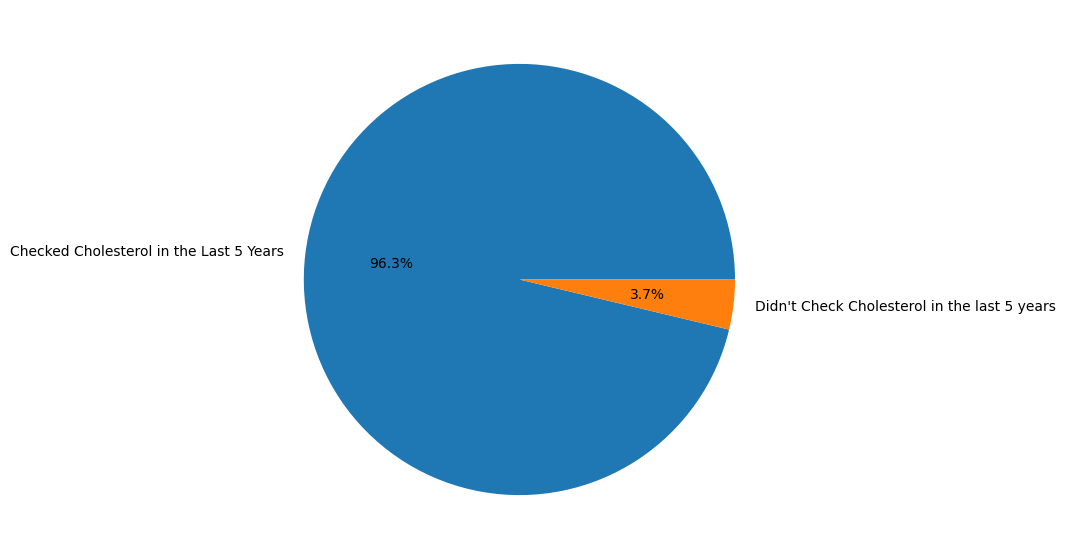


Figure 9

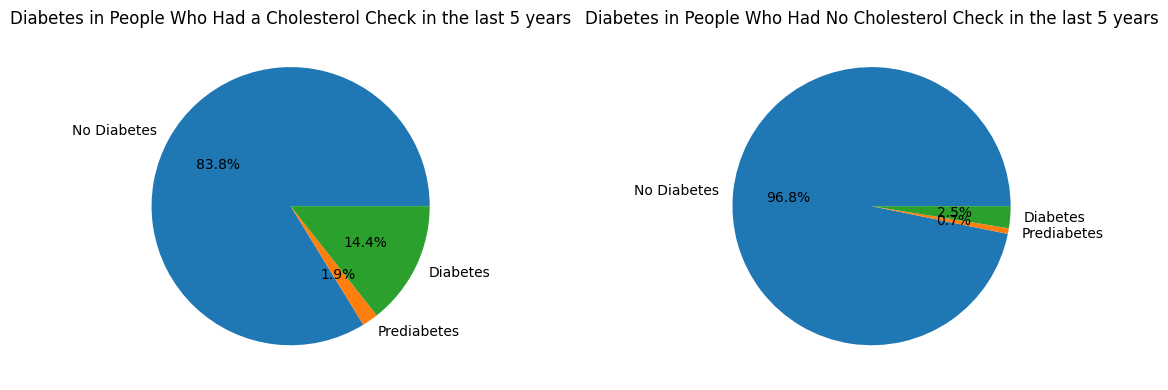


Figure 10

This feature might not seem that useful at first glance. It could imply that people getting their cholesterol checked are people who also suffer from various diseases that could lead them to check their levels more.

1. **Stroke**

Another binary classifier that checks if a person suffered a stroke before (1) or not (0).

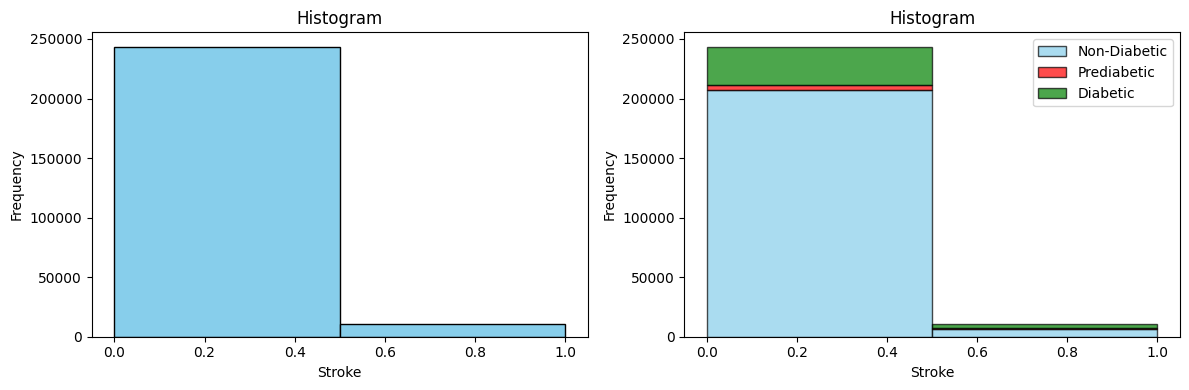


Figure 11

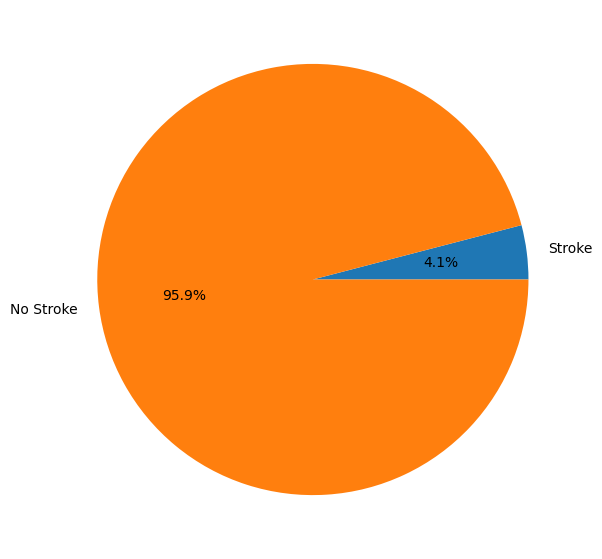


Figure 12

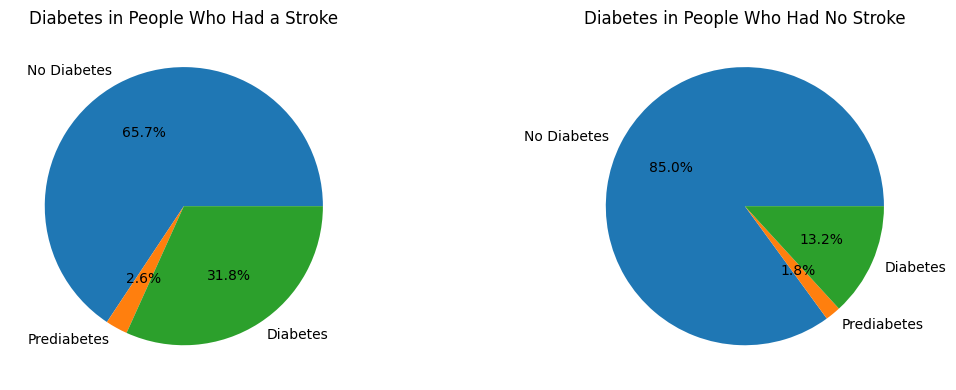


Figure 13

Expectedly, not many individuals have suffered a stroke. This is due to the fact that it is quite a rare event that does have a considerable mortality rate. However, those who had a stroke tend to have a much higher probability of having diabetes than people who did not suffer from a stroke.

1. **Heart Disease or Attack**

This column is also a binary classifier that checks whether an individual had coronary heart disease (CHD) or myocardial infarction (MI) (1) or does not (0).

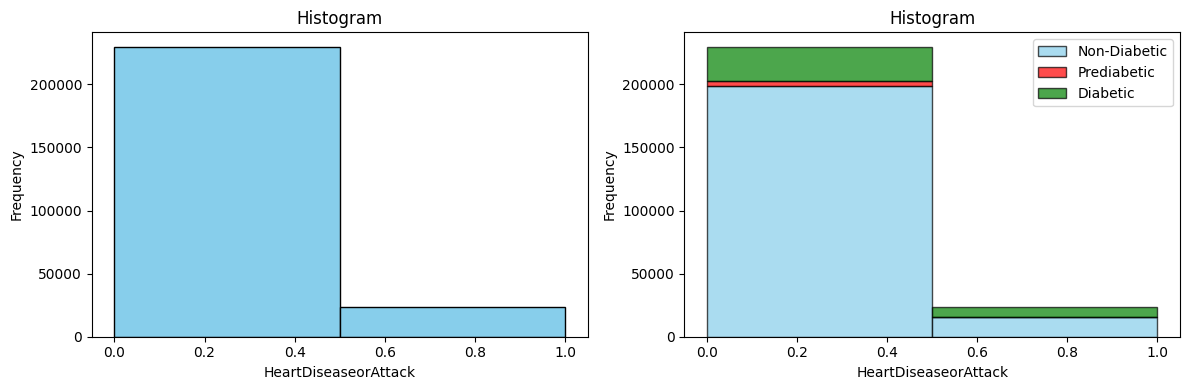


Figure 14

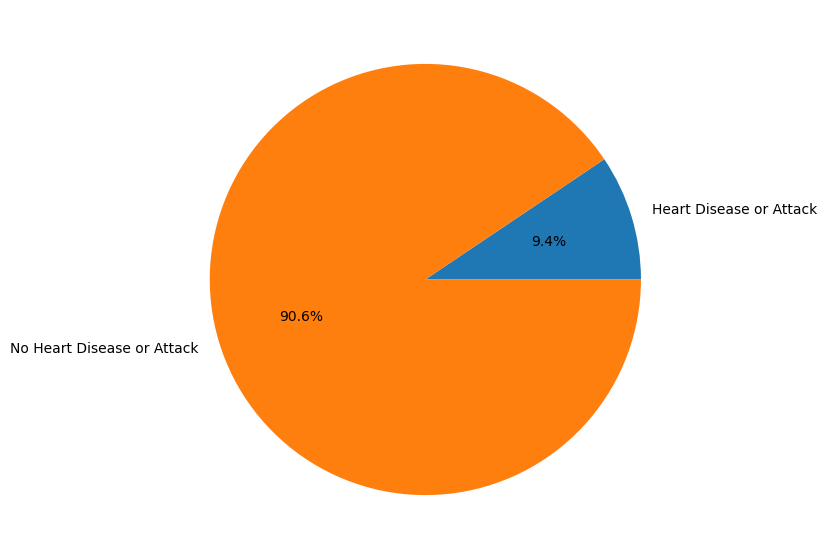


Figure 15

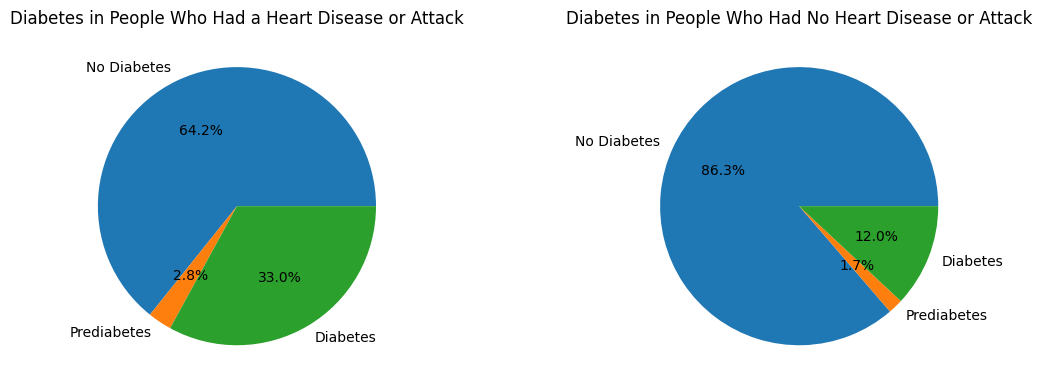


Figure 16

This is a good feature since it shows a correlation between heart disease or attack and having diabetes.

1. **Physical Activity in the Last 30 Days**

This column is a binary classifier that measures if a person did physical activity in the last 30 days. If yes, then it is 1, else 0.

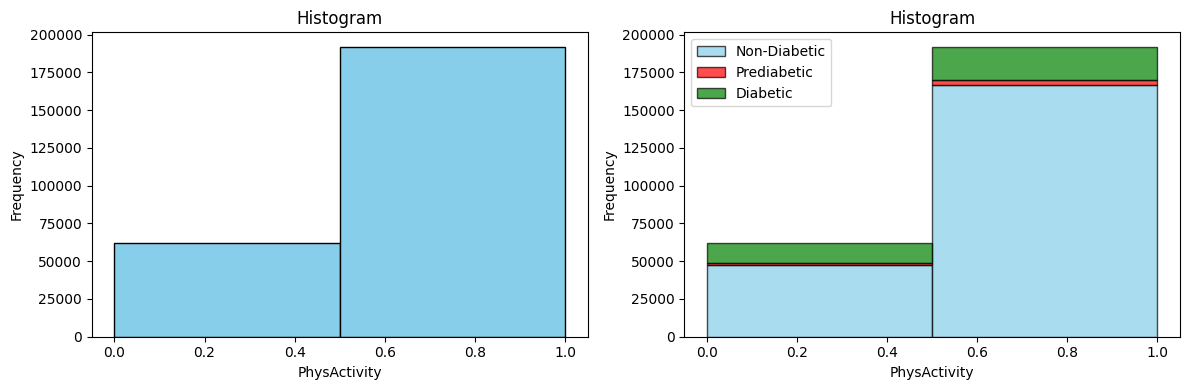


Figure 17

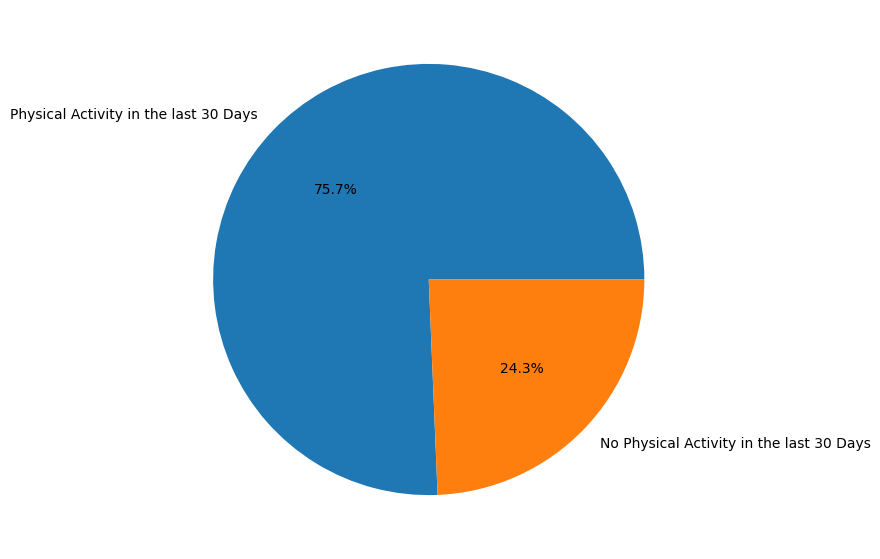


Figure 18

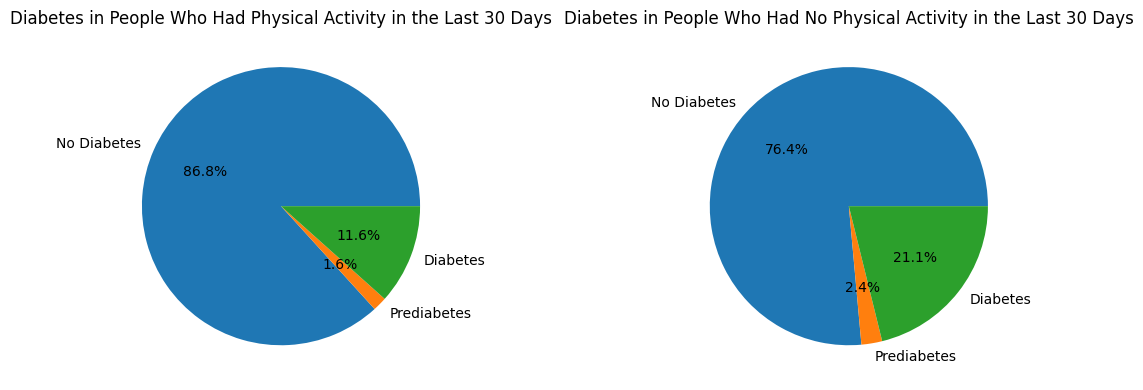


Figure 19

Intuitively, physical activity decreases the chance of a person having diabetes by around 10%. This is a pretty descriptive feature to have.

1. **Consume Fruits One or More Times a Day**

This column is a binary classifier that indicates whether an individual eats fruits daily (1) or not (0).

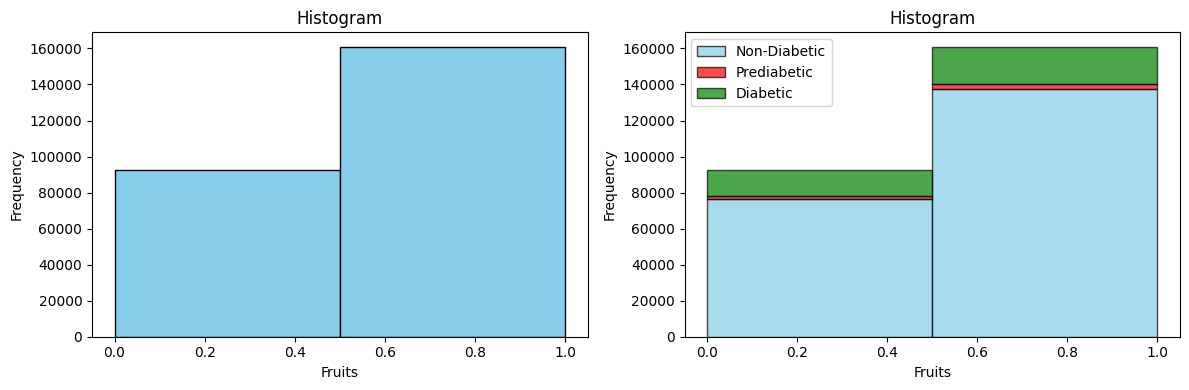


Figure 20

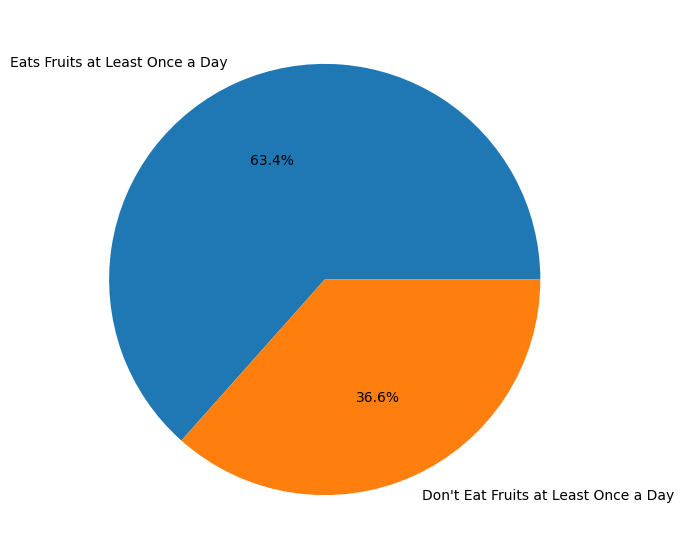


Figure 21

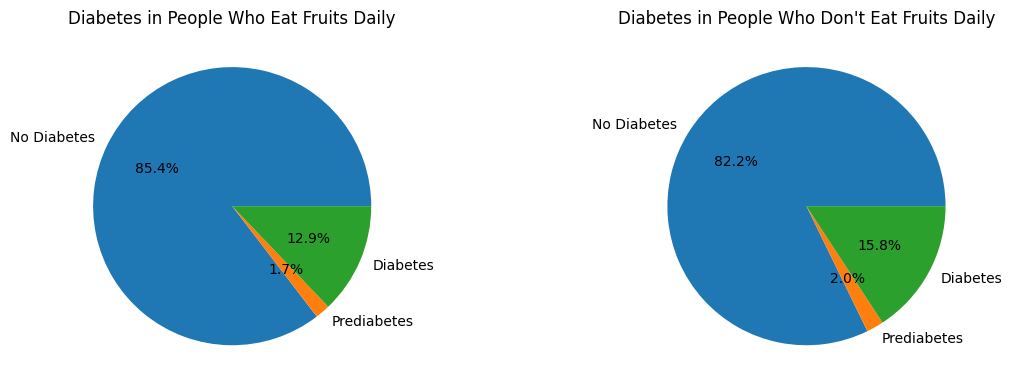


Figure 22

The pie charts in figure 22 shows that diabetes is less likely to occur to individuals who eat fruits daily, albeit the difference is not that considerable.

1. **Consume Vegetables One or More Times a Day**

Similar to the fruit’s column, this is a binary classifier that indicates whether an individual consumes vegetables daily (1) or not (0).

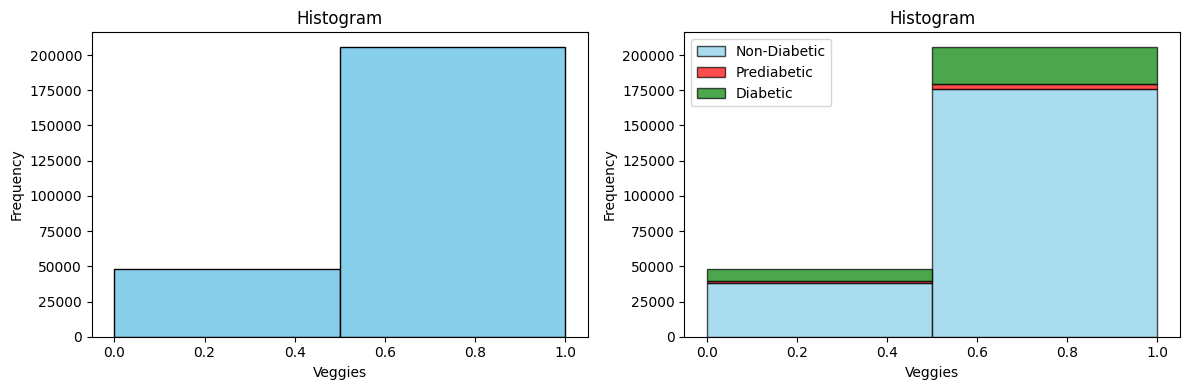


Figure 23

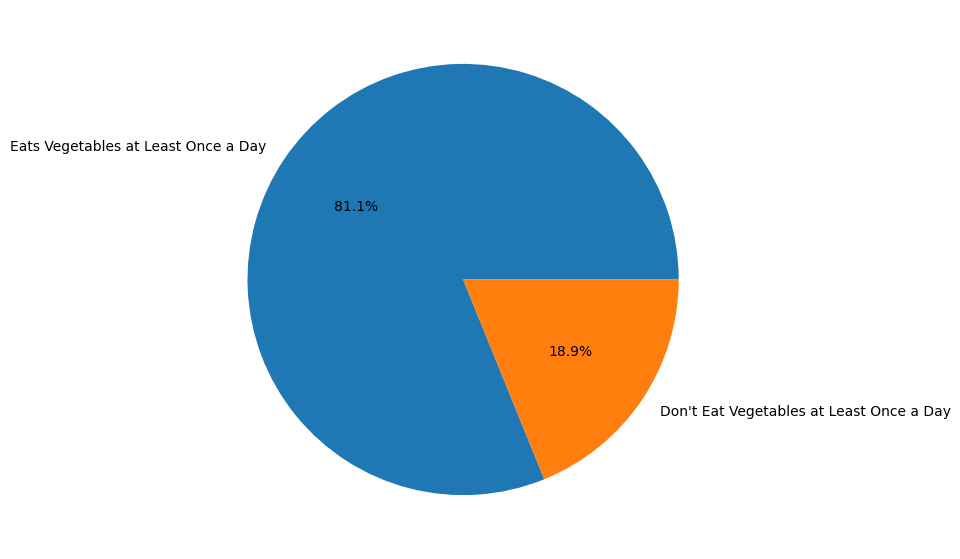


Figure 24

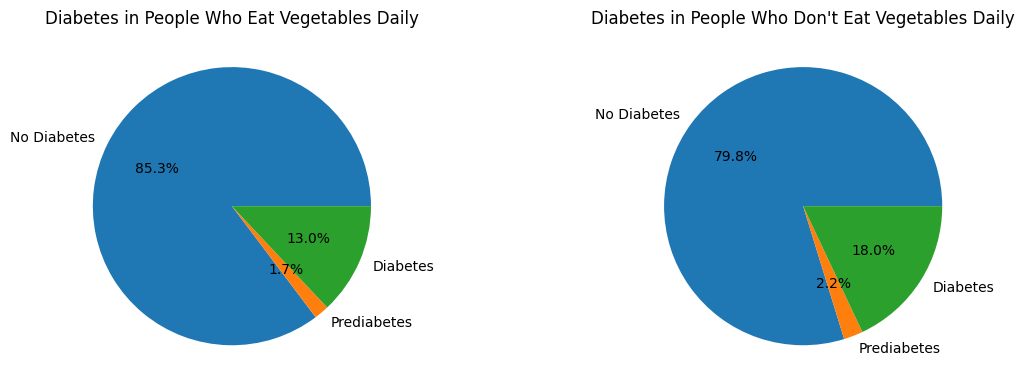


Figure 25

As shown in figure 25, eating vegetables daily reduces the risk of diabetes and prediabetes.

1. **Heavy Drinkers**

Adult men are considered heavy drinkers if they have more than 14 drinks per week, and adult women are considered the same if they have more than 7 drinks per week. This feature is a binary classifier that labels non-heavy drinkers as 0 and heavy drinkers as 1.



Figure 26

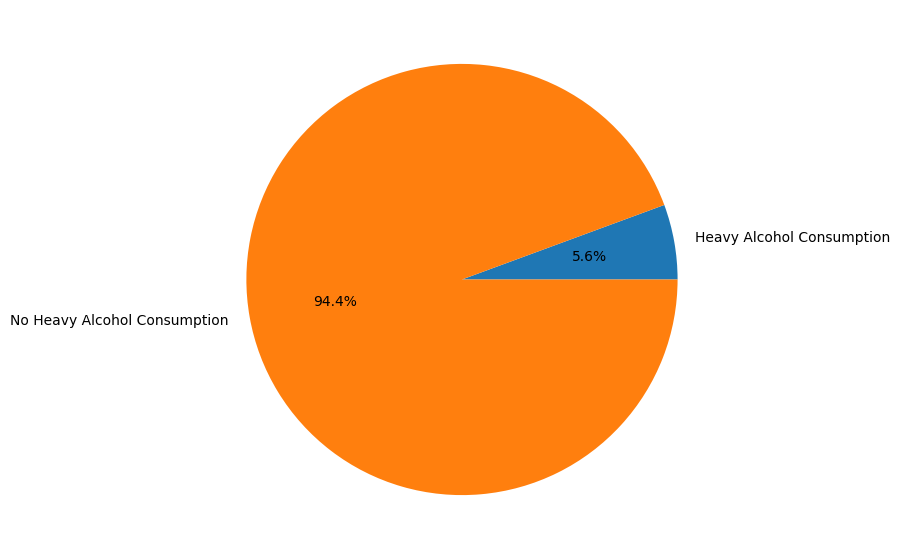


Figure 27

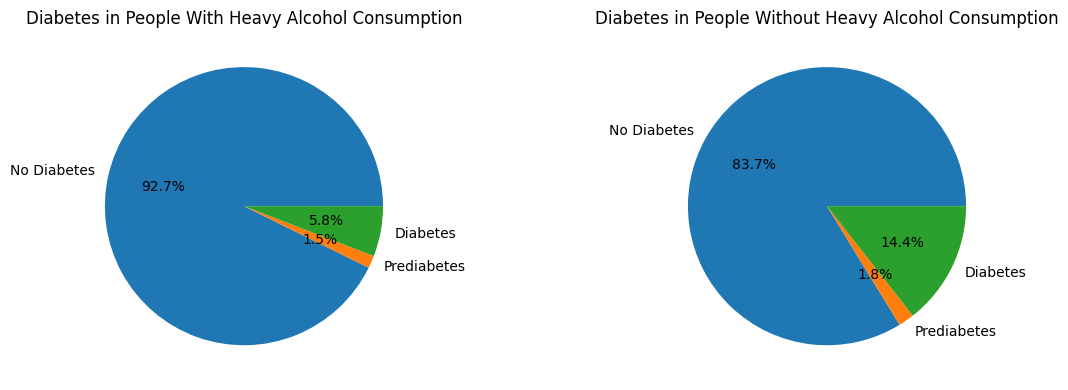


Figure 28

Surprisingly, heavy drinkers are less likely to have diabetes than non-heavy drinkers. We do not know why or how this correlation came to be, but that is how the data has spoken.

1. **Any Healthcare**

This feature is a binary classifier as well. Having any kind of healthcare plan counted as a 1, else 0.

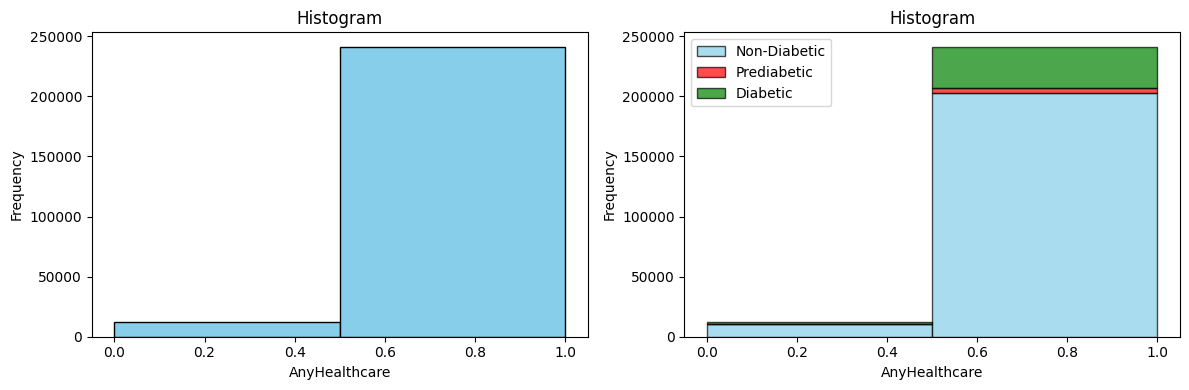


Figure 29

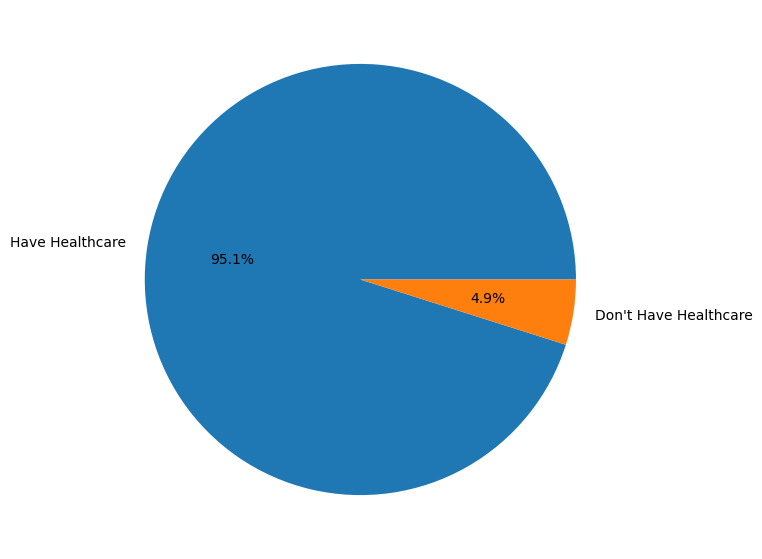


Figure 30

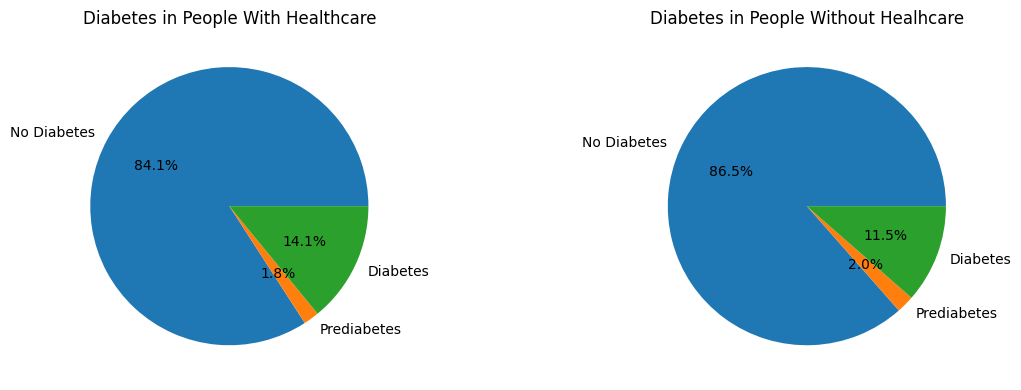


Figure 31

Somehow, people without healthcare were less likely to develop diabetes, but are slightly more likely to have prediabetes. This seemingly nonsensical correlation might be the sign of an anomaly or a deeper relationship between the two.

1. **No Doctor Because of Cost**

This feature records whether an individual who needed to see a doctor in the past 12 months could not because of the cost involved. 0 means no and 1 means yes.

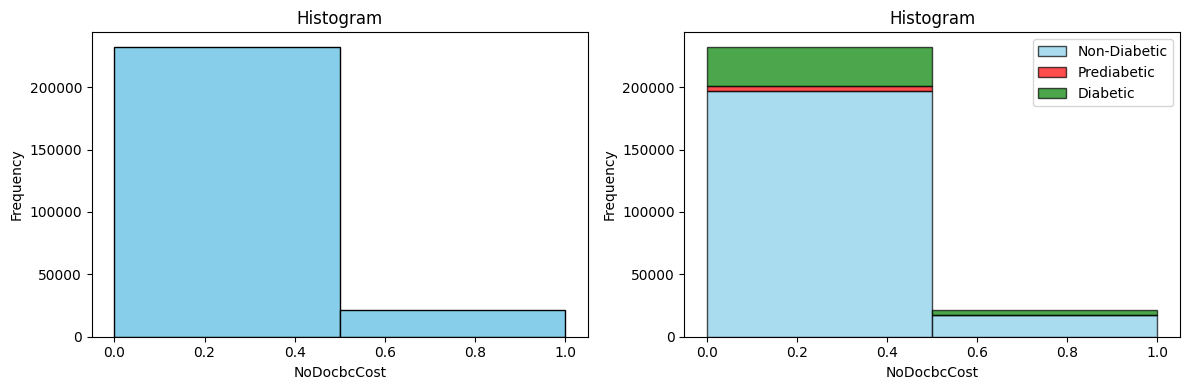


Figure 32

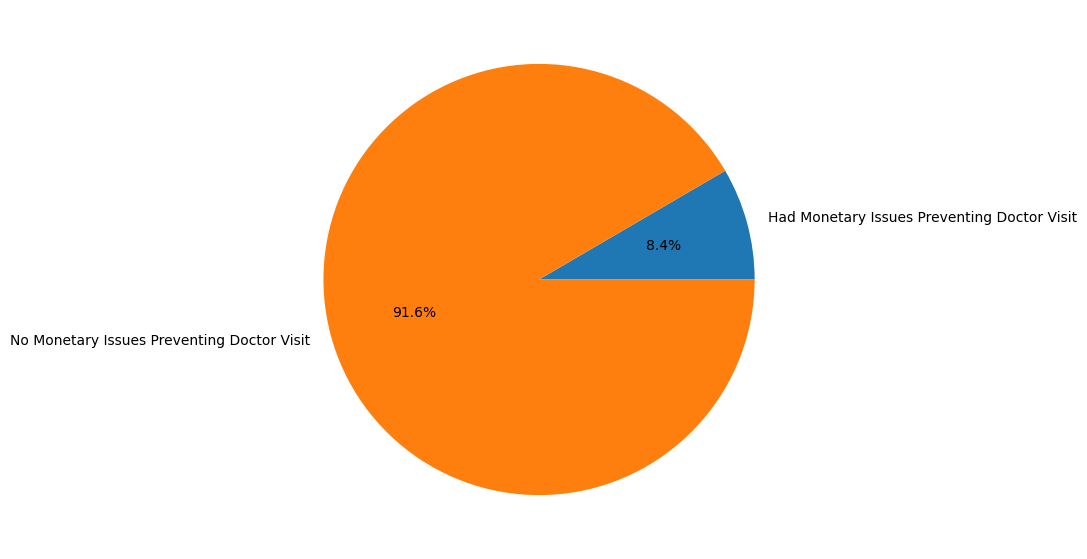


Figure 33

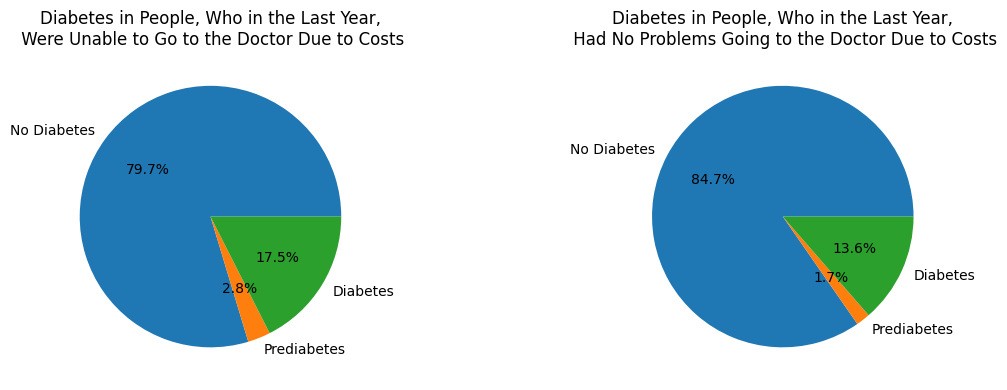


Figure 34

People who were unable to go to a doctor due to costs were more likely to develop diabetes and prediabetes. Maybe their unfavourable financial status could be a cause of diabetes due to the overall poorer quality of life.

1. **Difficulty Walking**

This feature measures whether an individual has serious difficulty walking or climbing stairs. As usual 0 means no and 1 means yes.

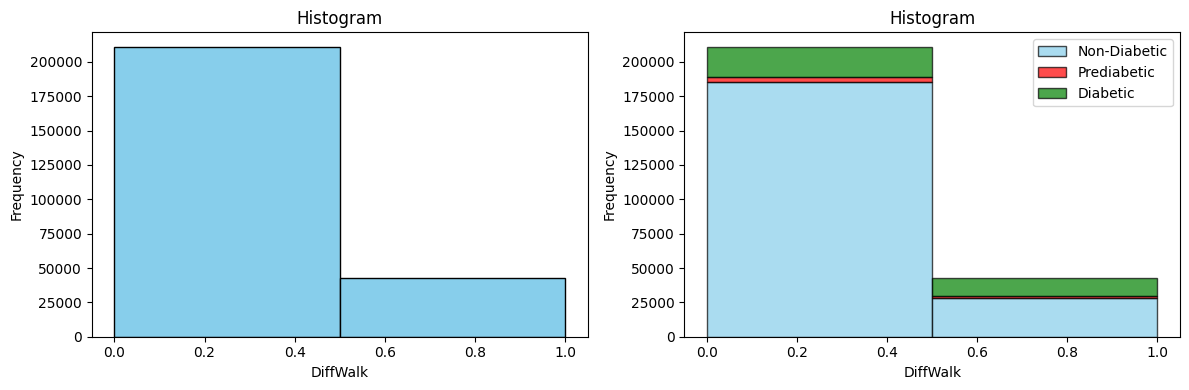


Figure 35

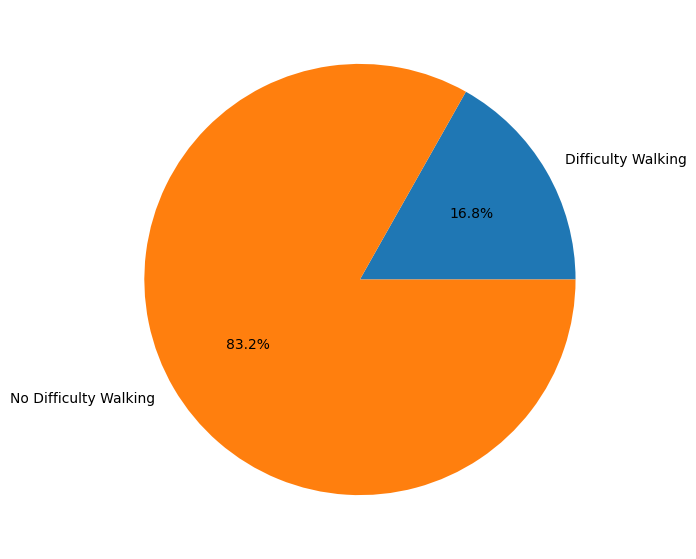


Figure 36

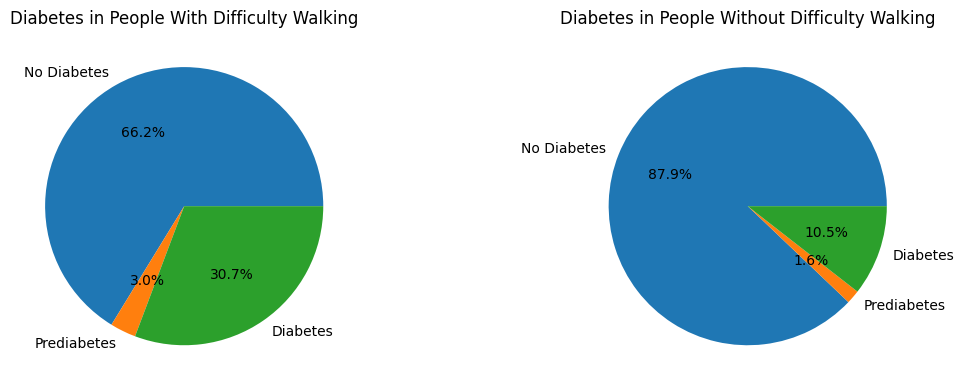


Figure 37

This feature is truly a descriptive one, for it shows a significant correlation between difficulty moving around and diabetes/prediabetes.

1. **Sex**

This feature is a binary classifier that classifies female as 0 and male as 1.

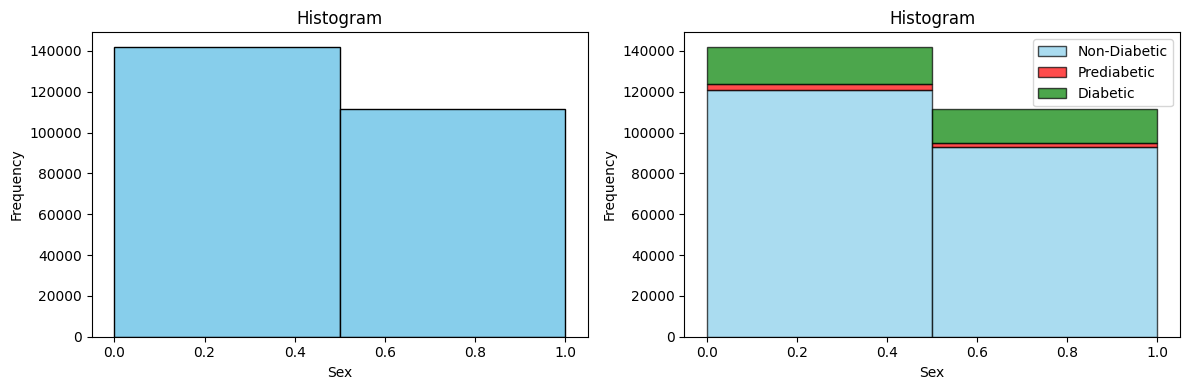


Figure 38

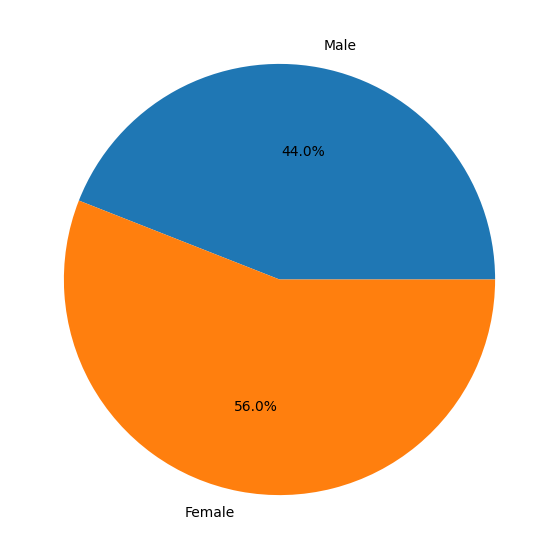


Figure 39

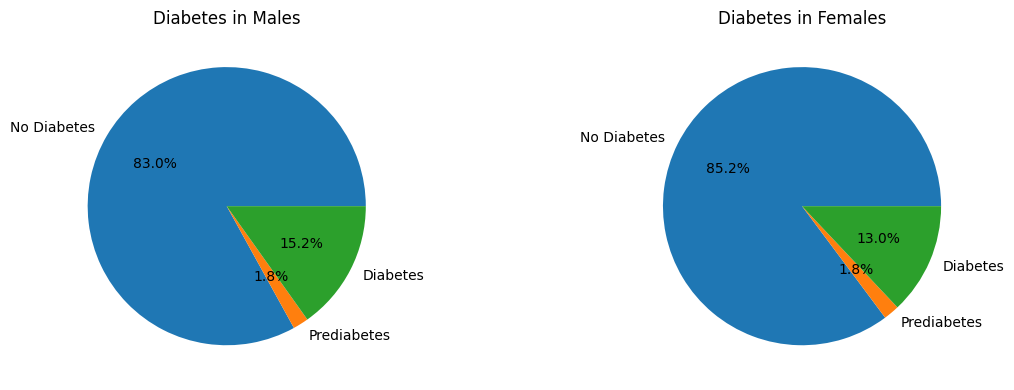


Figure 40

Even though it is not that considerable, being a man increases the risk of you developing diabetes.

1. **BMI**

The BMI feature contains values that when plotted on a histogram against their frequencies gives rise to something that resembles a normal distribution. However, there appears to be many values far away from the mean and median.

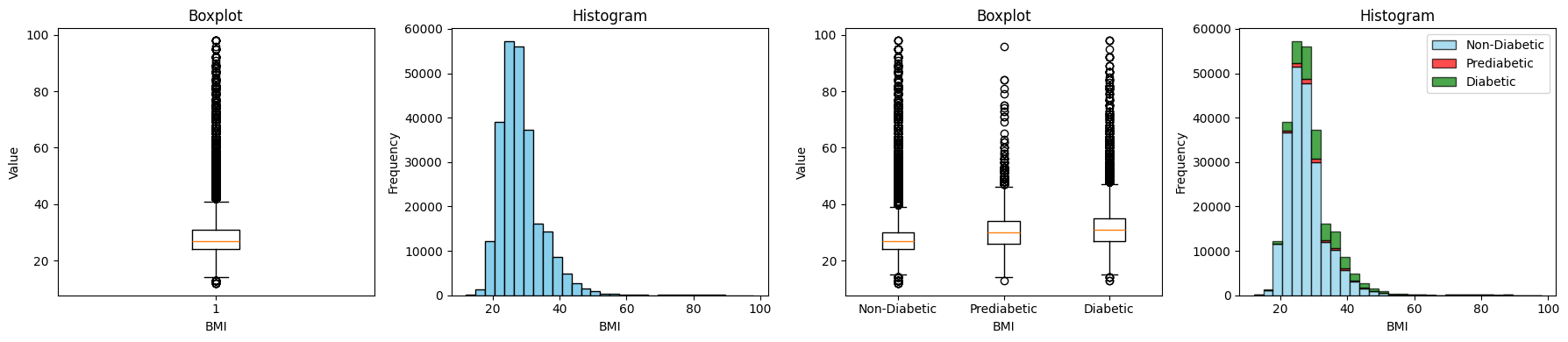


Figure 41

Summary of the BMI column:  
Mean: 28.4

Standard Deviation: 6.6  
Median: 27.0  
Minimum: 12.0  
Maximum: 98.0  
Prediabetics and diabetics tend to have a higher BMI on average than non-diabetics.

The yeo-Johnson transformer was used to decrease the skew of the values. This was followed up by a robust scaler.

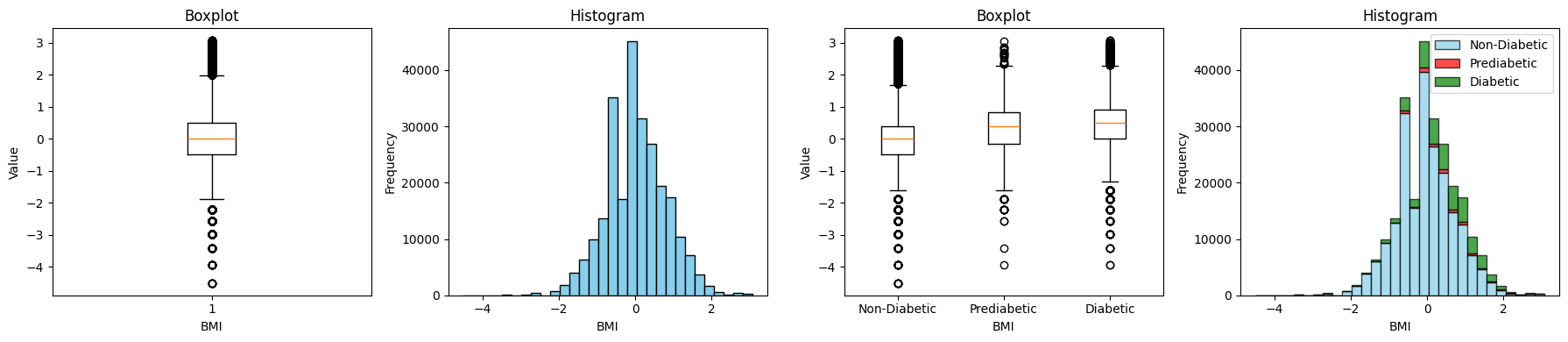


Figure 42

1. **General Health**

The general health feature is defined as a scale from 1-5 which describe the health of an individual. 1 = excellent, 2 = very good, 3 = good, 4 = fair and 5 = poor.



Figure 43

All the individuals in the dataset had a mean score of 2.5 and a median of 2.0. This data indicates a trend that the worse the health of an individual is, the more likely they are to suffer from diabetes or be prediabetic.

In order to process the data better, we decided to one-hot encode the scale into respective 5 bins as shown the following pie chart.

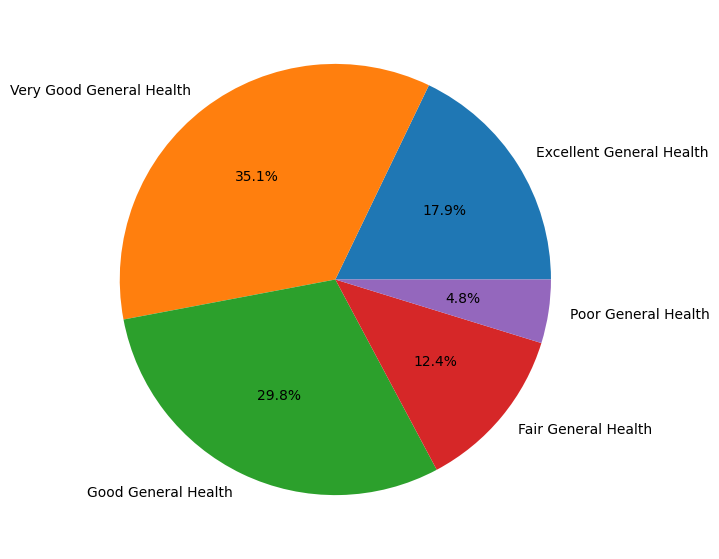


Figure 44

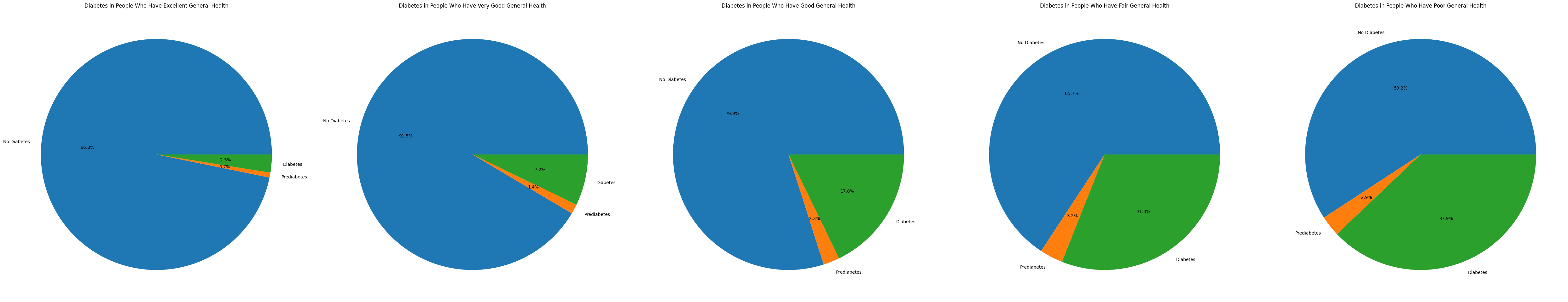


Figure 45

After one-hot encoding, the relationship between perceived general health and diabetes should be clearer.

1. **Mental Health**

This feature in the dataset asks the individuals to rate their mental health, taking into account stress, depression and problems with emotions, from 1-30. This scale is based on how many days an individual has gauged his mental health as subpar. The higher the number, the worse the mental health of an individual overall is.

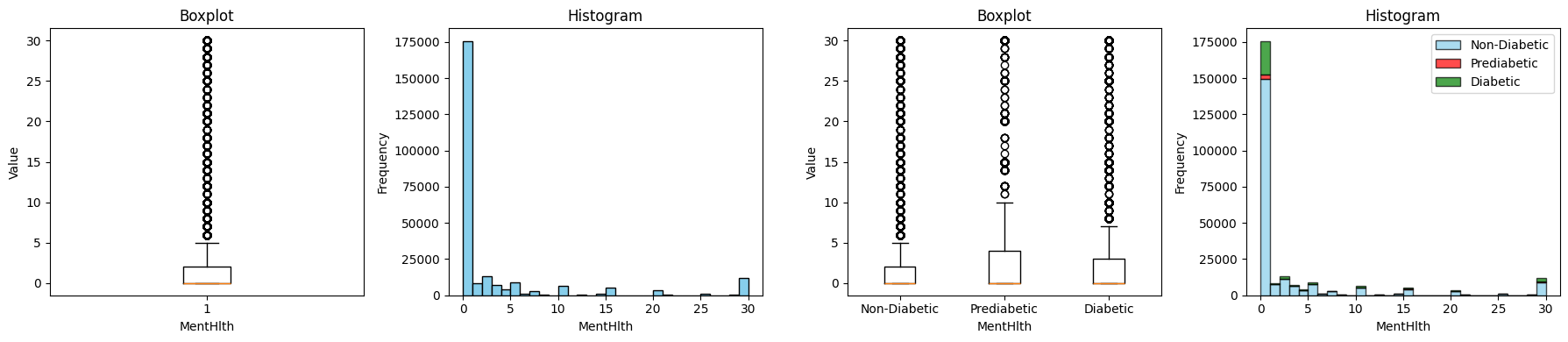


Figure 46

Summary of the mental health feature:

Mean: 3.2

Standard Deviation: 7.4

Median: 0.0

Minimum: 0.0

Maximum: 30.0

Due to the fact that the data was extremely skewed to the left, we decided to turn the feature into a binary classifier, where 0 means mentally unhealthy and 1 means mentally healthy.

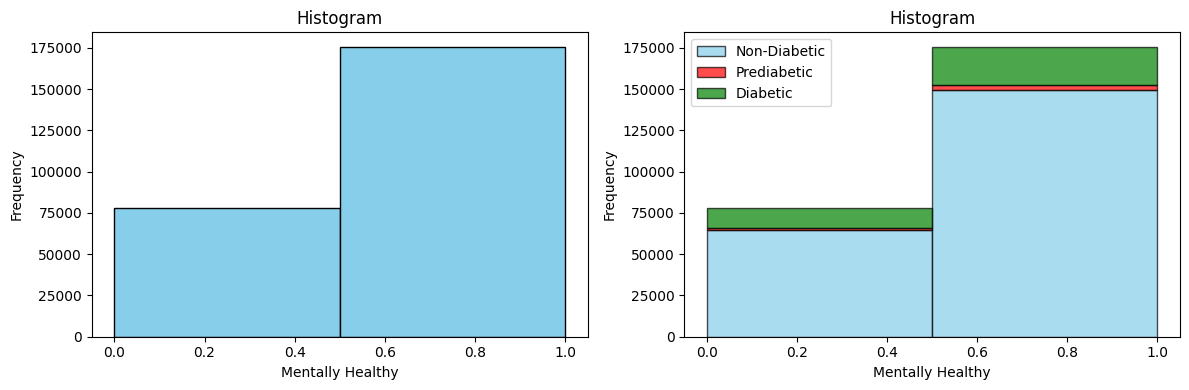
****

Figure 47

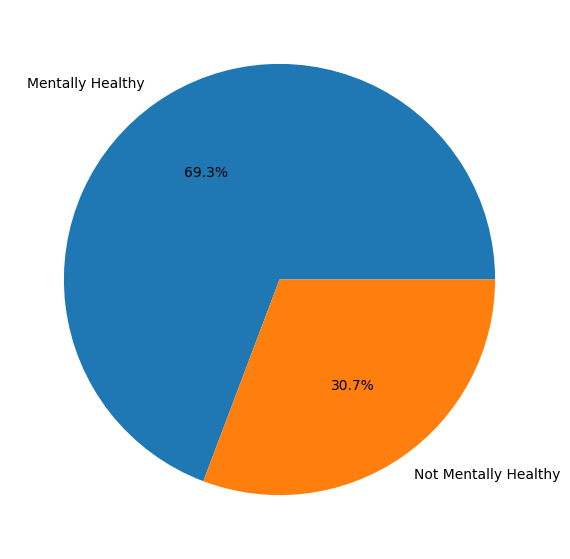


Figure 48

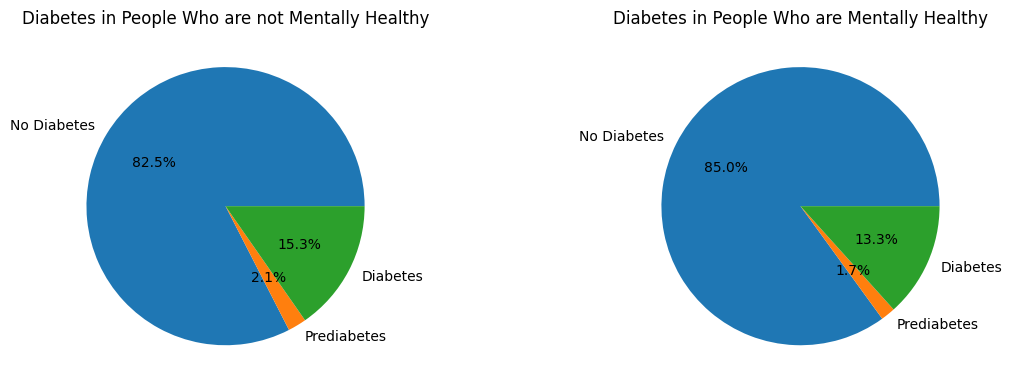
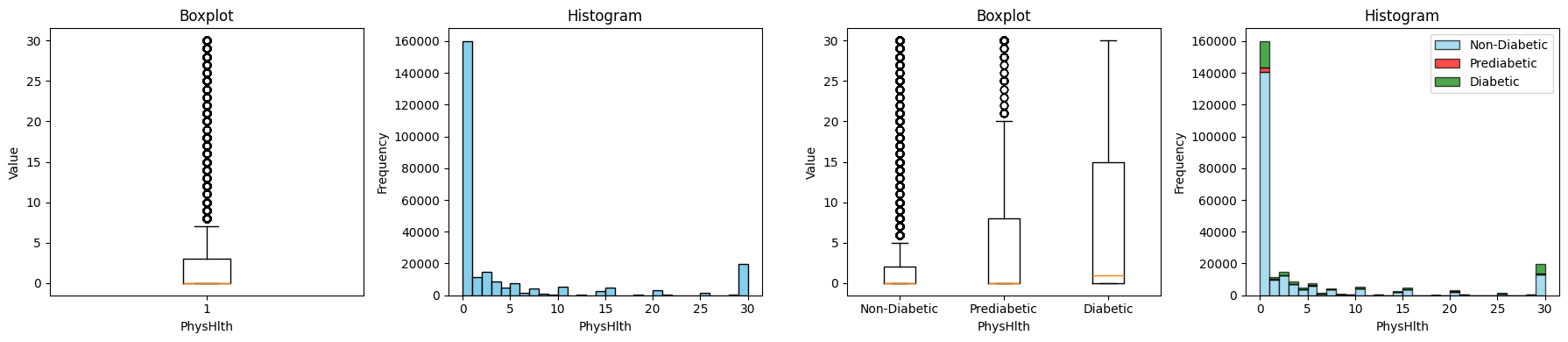


Figure 49

Poor mental health is slightly correlated wither a higher risk of diabetes and prediabetes.

1. **Physical Health**

This feature asks the individuals in the dataset to rate their physical health, taking into account physical illness and injury, from 1-30. This scale, the same as the feature before, is based on how many days the individual felt themselves physically unwell. A higher rating indicates a worse overall physical health.



The summary of this feature is as follows:

Mean: 4.2

Standard Deviation: 8.7

Median: 0.0

Minimum: 0.0

Maximum: 30.0

Just like the feature just before this one, the data is heavily skewed towards the left. This property calls for it to be turned into a binary classifier, where being physically health is 1, while the opposite is 0.

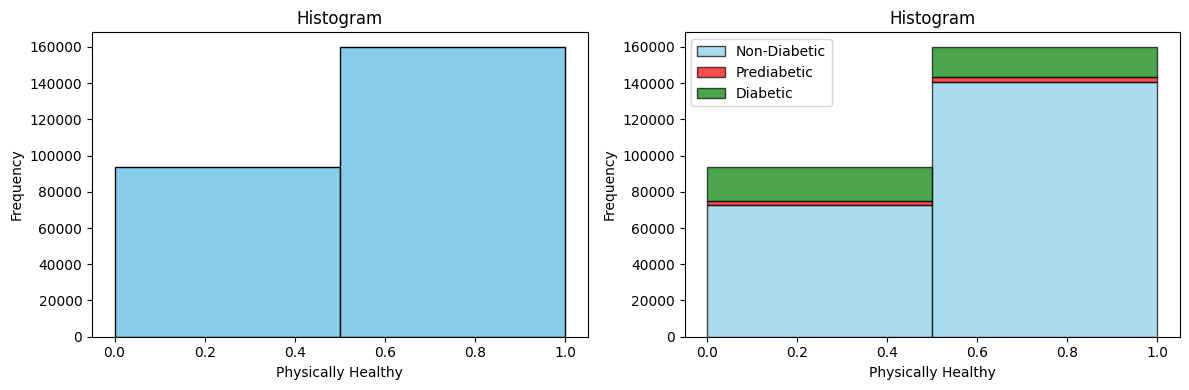


Figure 50

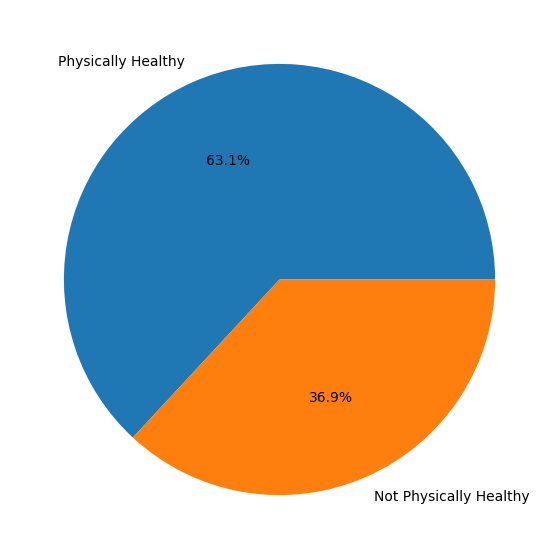


Figure 51

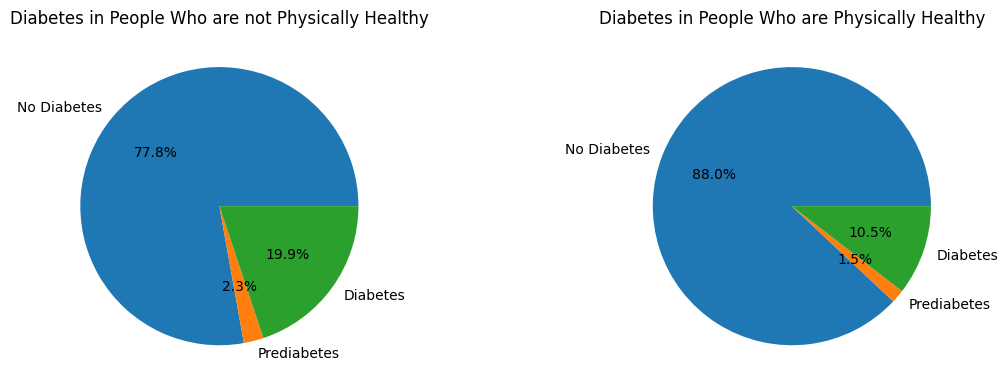


Figure 52

As shown in figure 52, people who are generally not physically well are more likely to have diabetes and prediabetes.

1. **Age**

The age feature is a 13-level category where each 5-year gap is represented by a level. For example, 1 = 18-24, 9 = 60-64 and 13= 80 and above. Detailed ranges can be found in figure

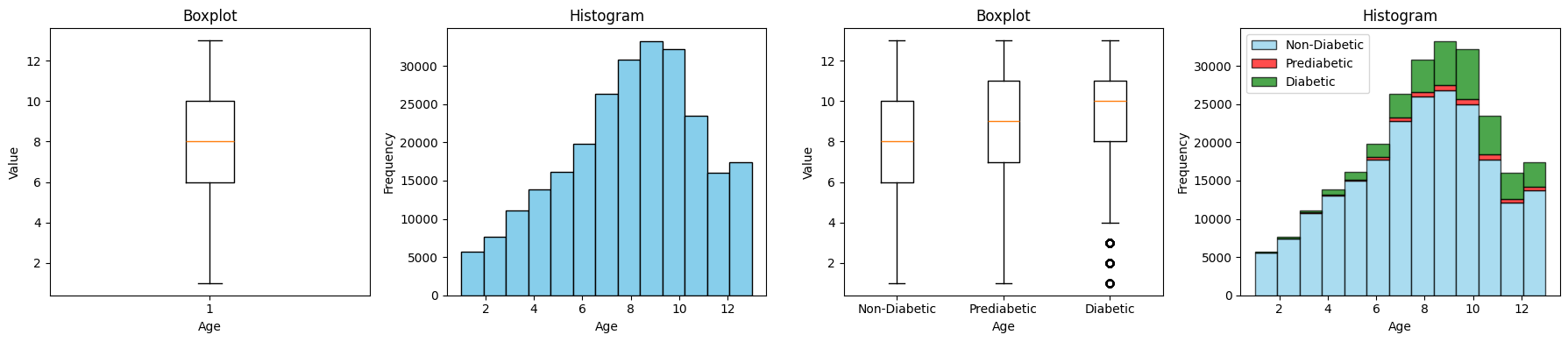


Figure 53

The summary of the data is:

Mean: 8.0

Standard Deviation: 3.1

Median: 8.0

Min: 1.0

Max: 13.0

We decided to one-hot encode the ranges as features for easier processing. The distribution of the frequences are shown in figure 54.

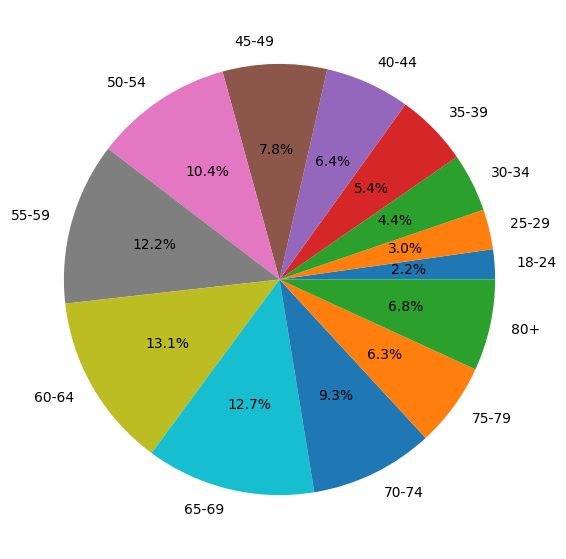


Figure 54

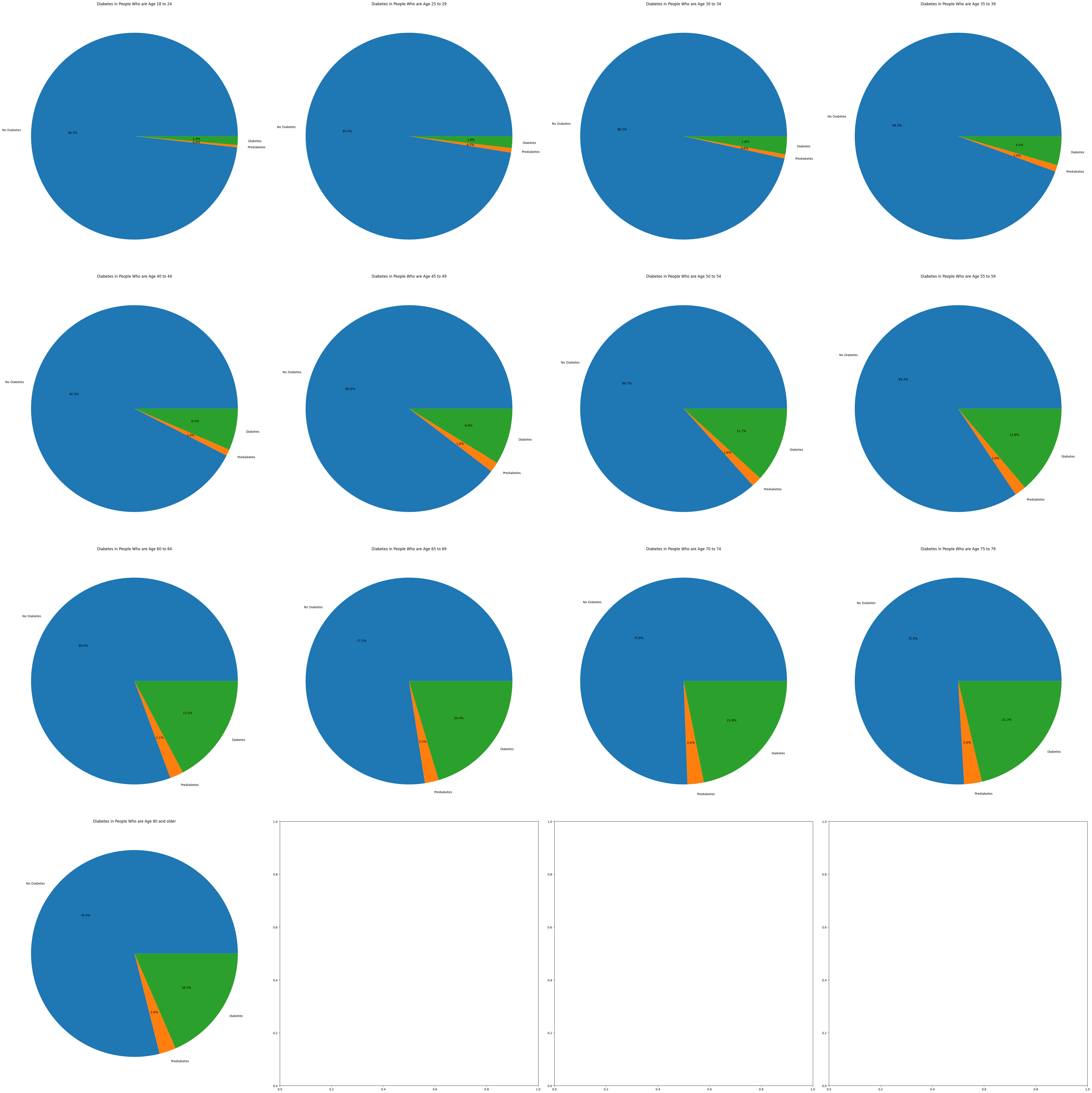


Figure 55

As shown in figure 55, as the age increases, the likelihood of having diabetes or prediabetes increases as well.

1. **Education**

This feature is based on a 1-6 scale that indicates the level of education of an individual in the dataset. 1 = Never attended school or only kindergarten, 2 = Grades 1 through 8 (Elementary), 3 = Grades 9 through 11 (Some high school), 4 = Grade 12 or GED (High school graduate), 5 = College 1 year to 3 years (Some college or technical school) and 6 = College 4 years or more (College graduate).

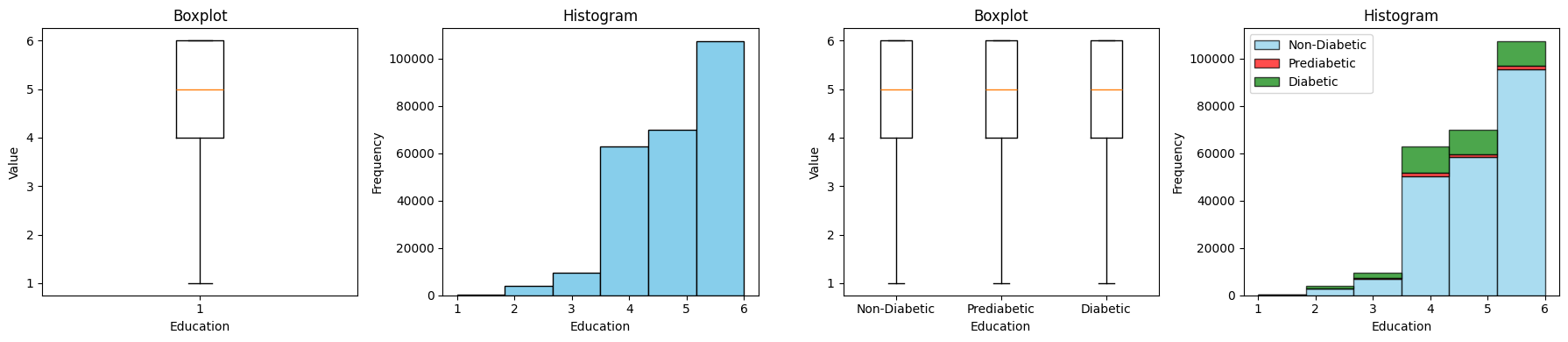


Figure 56

The summary of this feature is:

Mean: 5.0

Standard Deviation: 1.0

Median: 5.0

Min: 1.0

Max: 6.0

To make the processing easier, we decided to one-hot encode these ranges as separate features as shown in figure 57

A pie chart with numbers and text

AI-generated content may be incorrect.

Figure 57

A group of pie charts

AI-generated content may be incorrect.

Figure 58

As shown in the pie charts of figure 58, high school graduates are the least likely to develop diabetes and prediabetes while people who didn’t finish past kindergarten were the least likely to develop prediabetes. Elementary schoolers were the most likely to develop both diabetes and prediabetes.

1. **Income**

This final feature is based on a 1-8 scale. 1 = less than $10,000, 5 = less than $35,000 and 8 = $75,000 or more.

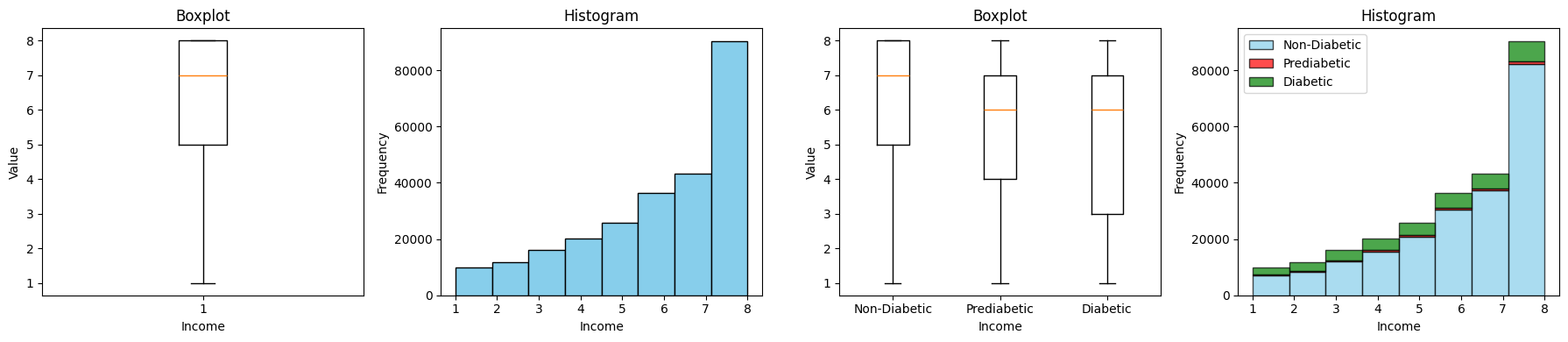


Figure 59

The summary of this feature is:

Mean: 6.1

Standard Deviation: 2.1

Median: 7.0

Min: 1.0

Max: 8.0

As with the previous feature, in order to ease processing, we decided to one-hot encode the different levels as separate features.

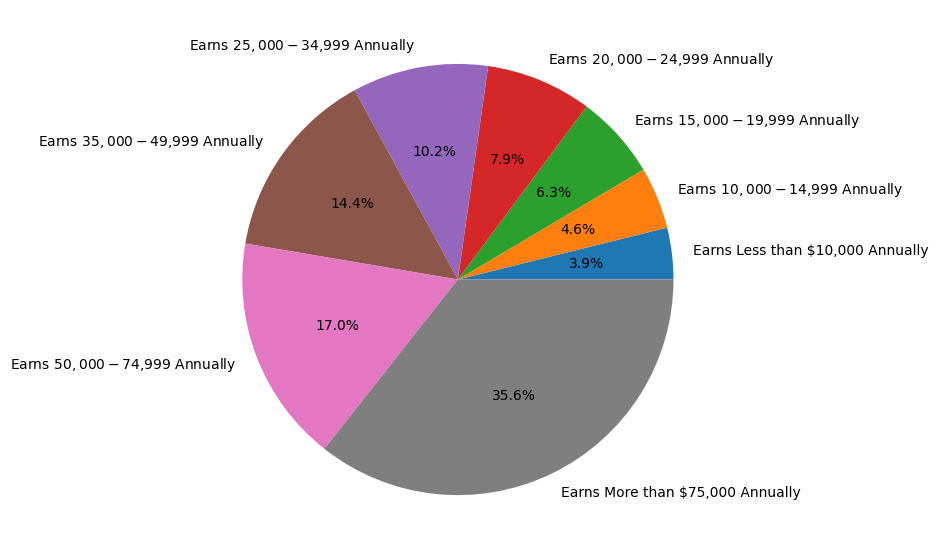


Figure 60

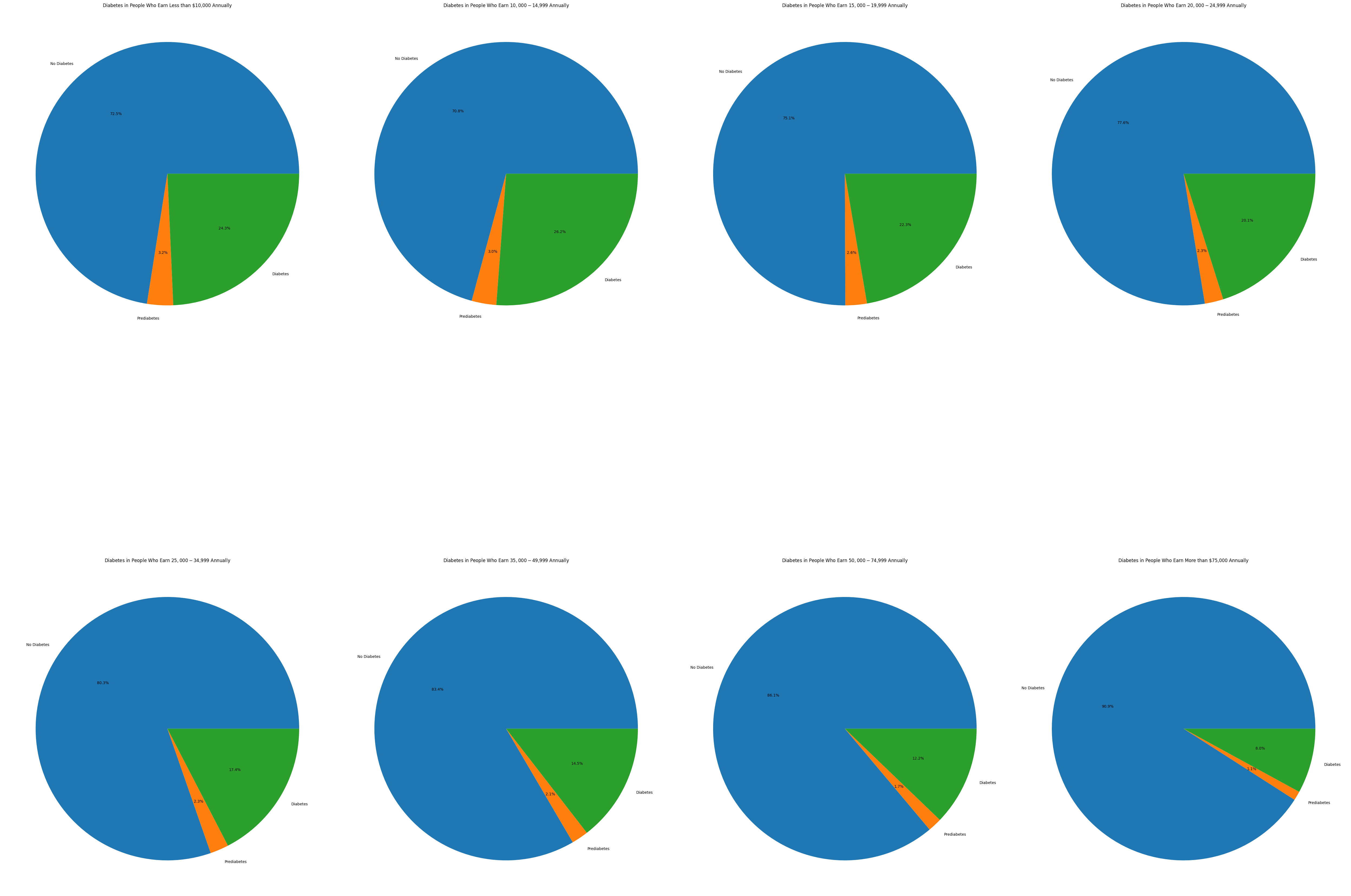


Figure 61

This is a significantly descriptive feature, due to the fact that is shows a clear trend that the higher the income, the lower the risk of diabetes and prediabetes. Figure 61 shows this.

CORRELATION MATRIX

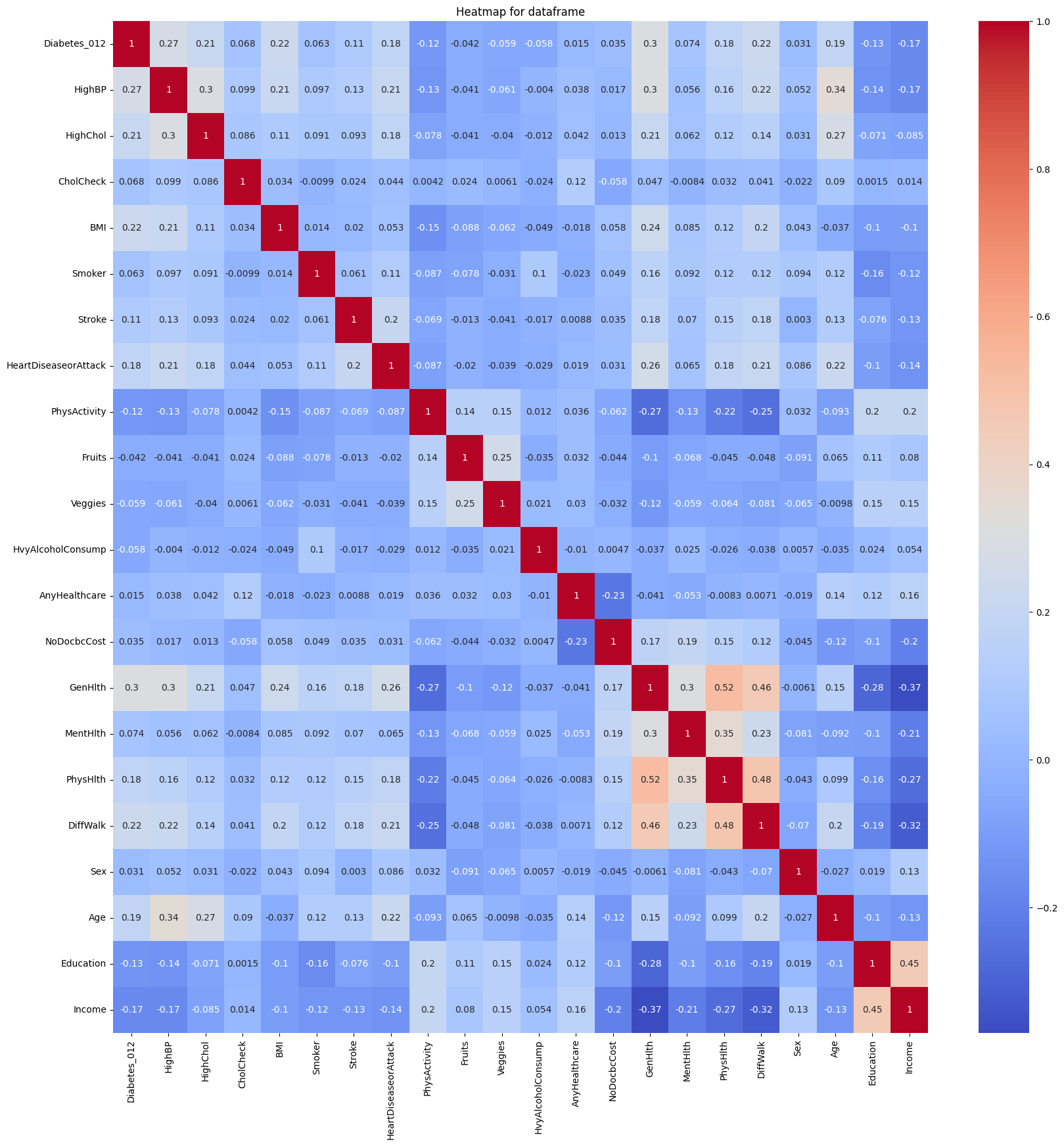


Figure 62

**CONCLUSION**

To conclude, from figure 62, general health has the highest correlation with diabetes, but almost all the other features like BP, cholesterol, income and education play some roles in the likelihood of have diabetes/prediabetes. Furthermore, none of the features have a correlation that is too high as to ruin the training.

Due to these observations, we will keep all the features for training. We dropped most of the numerical features that are not just 1 and 0 in favor of their one-hot encoded counterparts. All rows were maintained and the number of columns increased to 48 due to one-hot encoding.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00