

CASE STUDIES

-

ANALYSIS OF A DIABETES HEALTH INDICATORS DATA SET

06/09/2023

82.05 - Análisis Predictivo - Primer Parcial

Fanny LATRON (65998)

Elsa DOYEN (65990)

Thomas SIMON (65931)



AGENDA

01

Context

02

Content of the data set

03

Data mining

04

Correlations

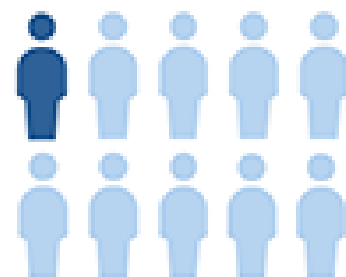
05

Conclusion



37 million people have diabetes

DIABETES



That's about **1** in every **10** people



1 in **5** people **don't** know they have it

PEOPLE LIVING WITH DIABETES are more likely to develop:



DEPRESSION, which affects about one in five adults with type 2 diabetes. (44)(19)



IRREVERSIBLE BLINDNESS AND VISION IMPAIRMENT caused by diabetic retinopathy, a preventable complication due to damaged blood vessels in the retina of the eye, which affects over one-third of people living with diabetes. (43)



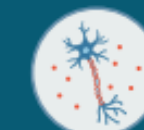
TUBERCULOSIS (TB) - diabetes can also negatively impact TB treatment outcomes. (19)



ORAL / GUM DISEASE



NERVE DISEASES, such as neuropathy, as well as peripheral arterial disease, both of which increase the risk of amputation. (4)



Certain **SKIN DISEASES**, such as psoriasis. (39)(40)(41)(42)



A number of common **CANCERS**. (4)



Complications and co-morbidities are highly preventable, if people living with diabetes and/or hypertension have access to timely, well-coordinated prevention, screening, diagnosis and care.

ITBA

CONTENT

kaggle



Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	1.0	10.0	6.0	8.0
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	14.0	0.0	0.0	9.0	6.0	7.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	11.0	4.0	4.0

Cleaned by Alex Teboul

21 feature variables and 253 680 survey responses



CONTENT

Variables :

Categorical :

- HighBP
- HighChol
- CholCheck
- Smoker
- Stroke
- HeartDiseaseorAttack
- PhysActivity
- Fruits
- Veggies

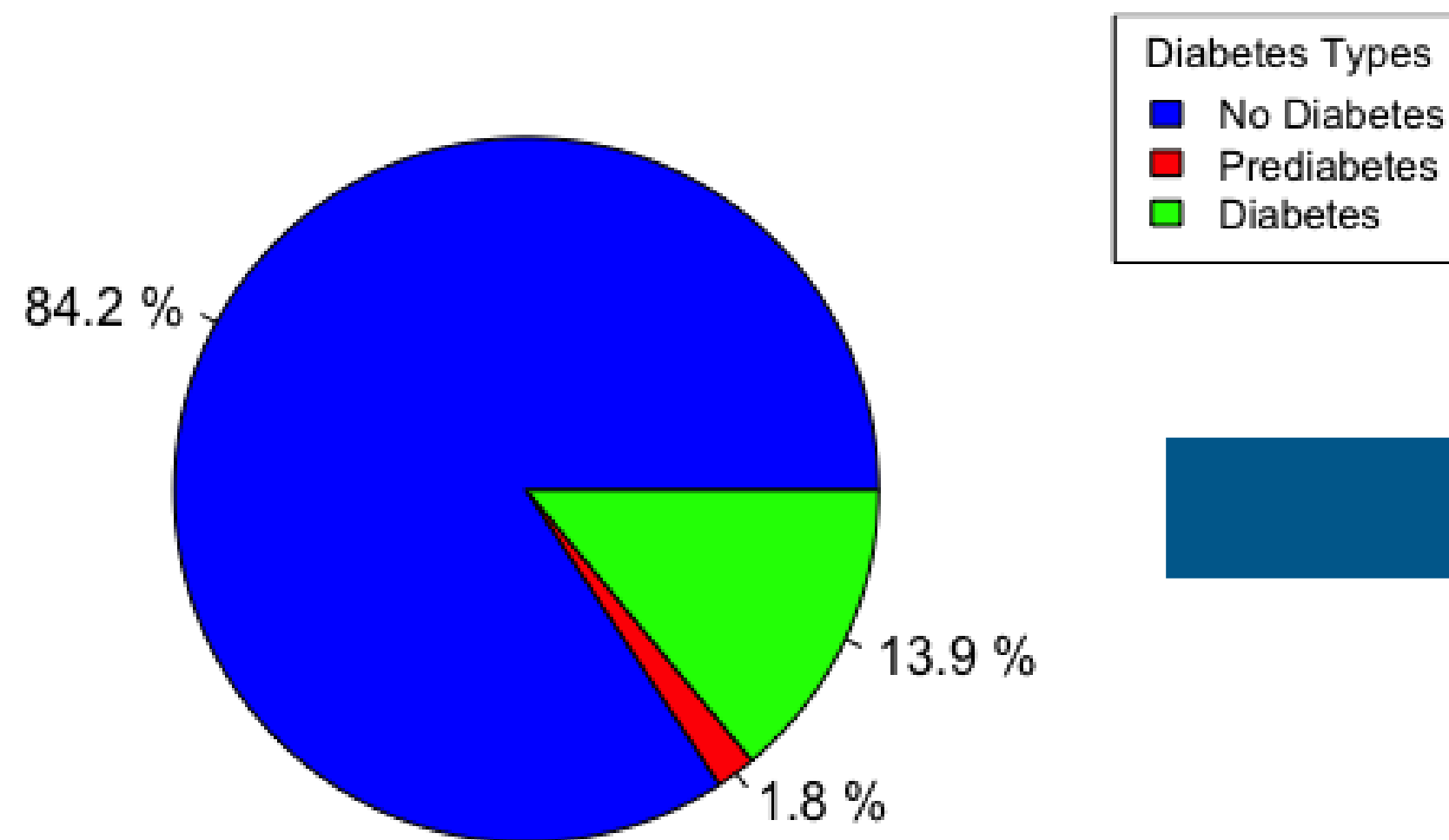
- HvyAlcoholConsump
- AnyHealthcare
- NoDocbcCost
- DiffWalk
- Sex
- GenHlth
- Age
- Education
- Income

Numerical :

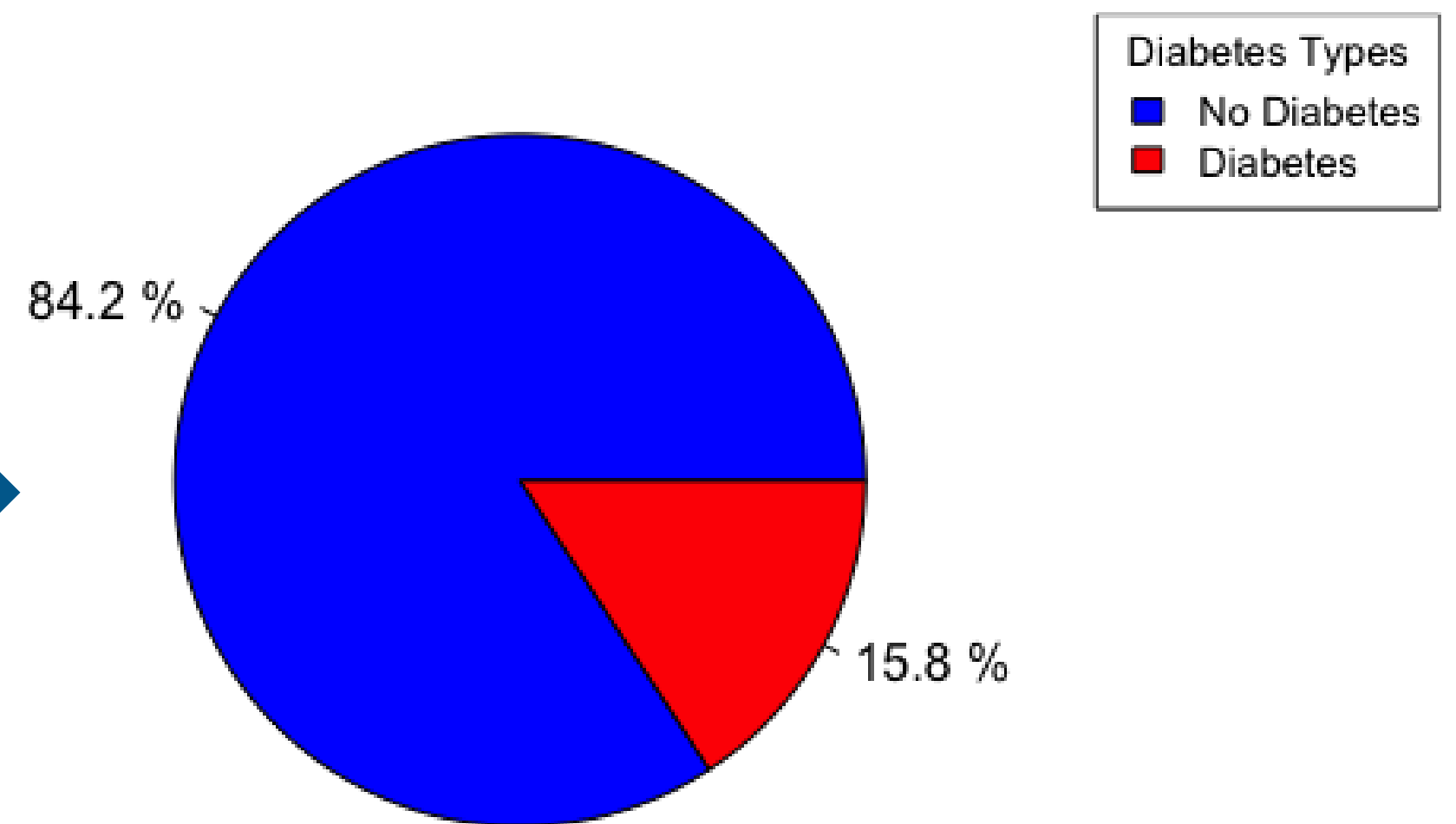
- MentHlth
- PhysHlth
- BMI

Response variable

Distribution of Diabetes Types



Distribution of Diabetes Types





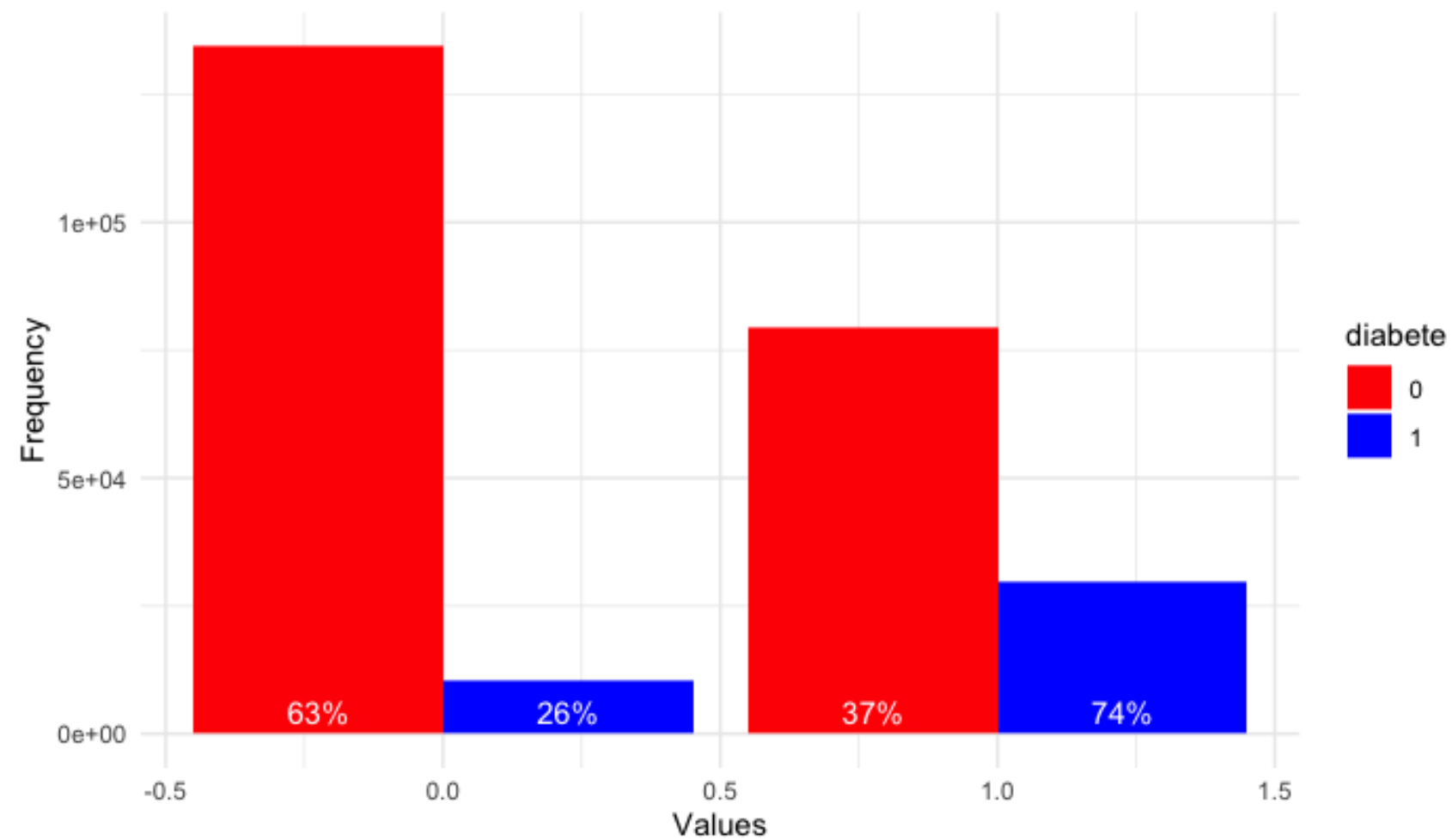
CONTENT

No missing !

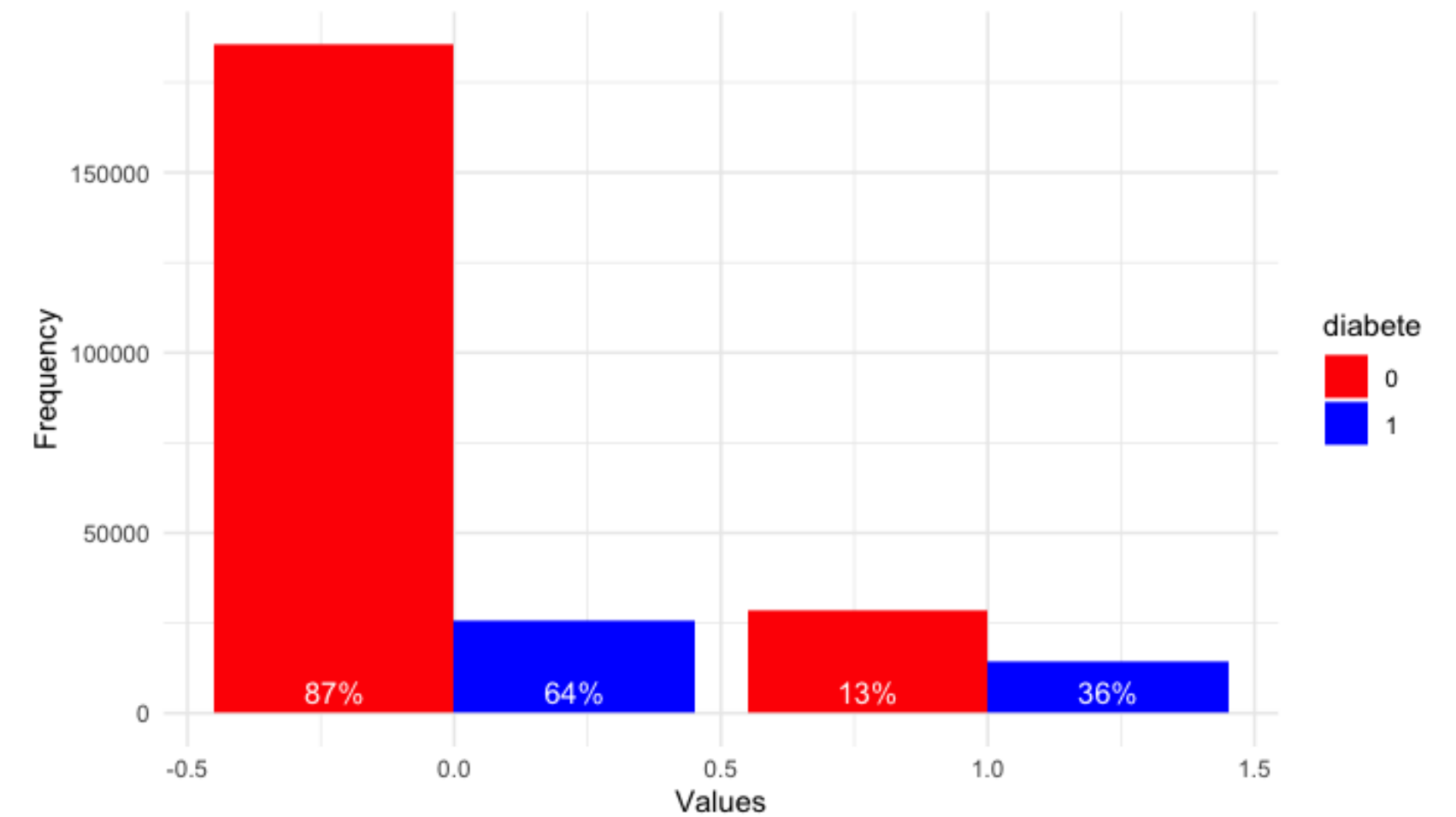
```
'''{r}|
na_counts <- colSums(is.na(data))
print(na_counts)
'''
```

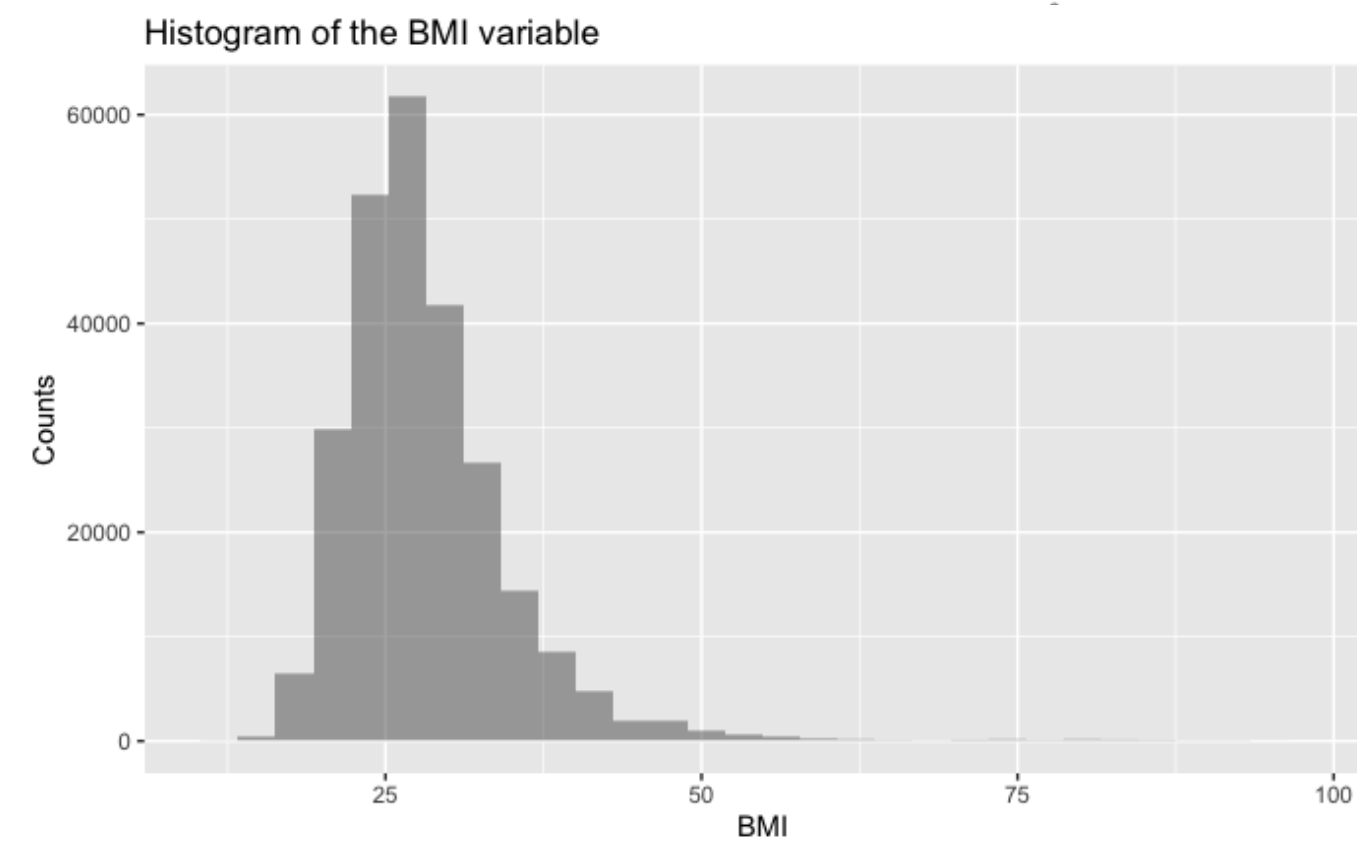
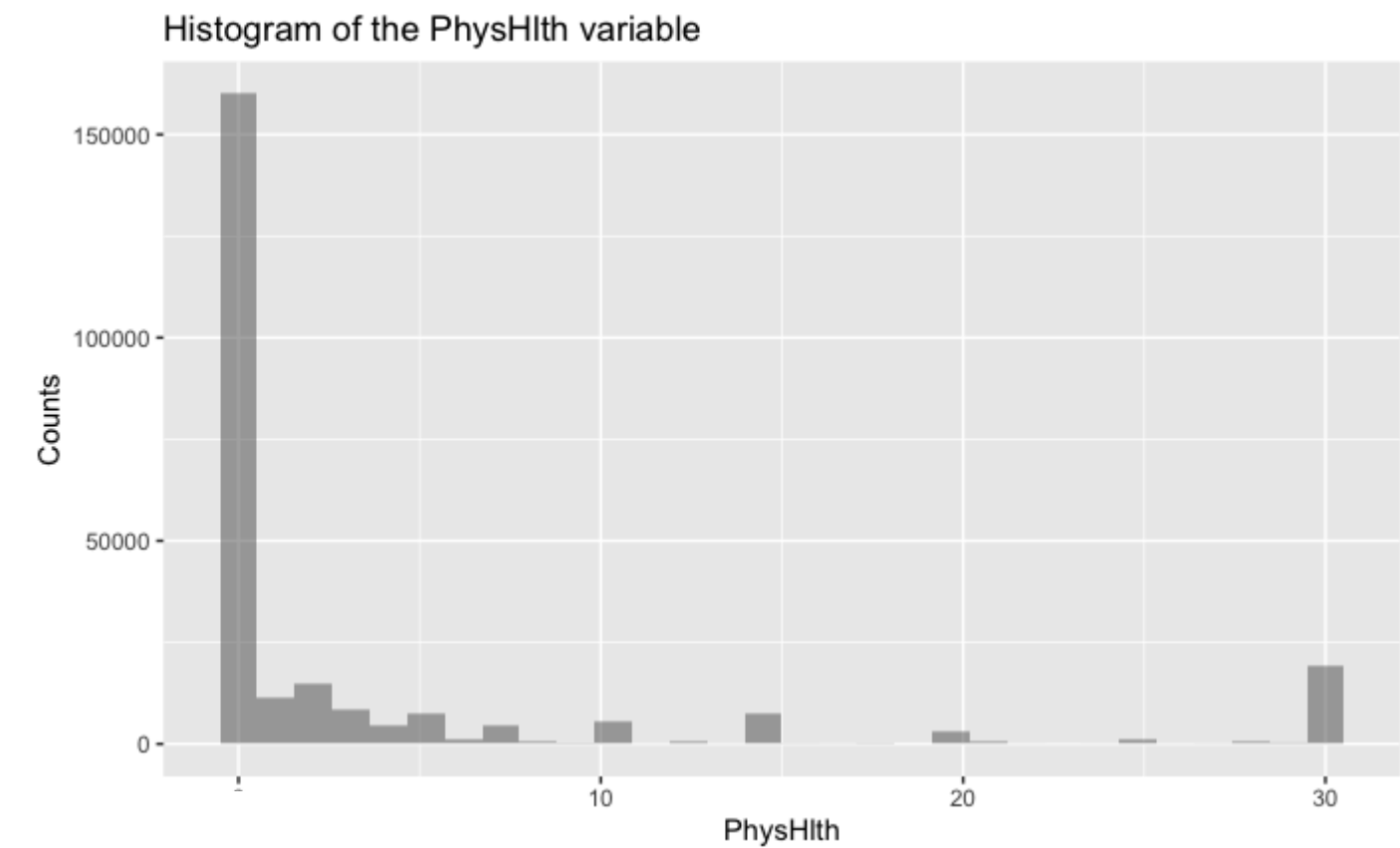
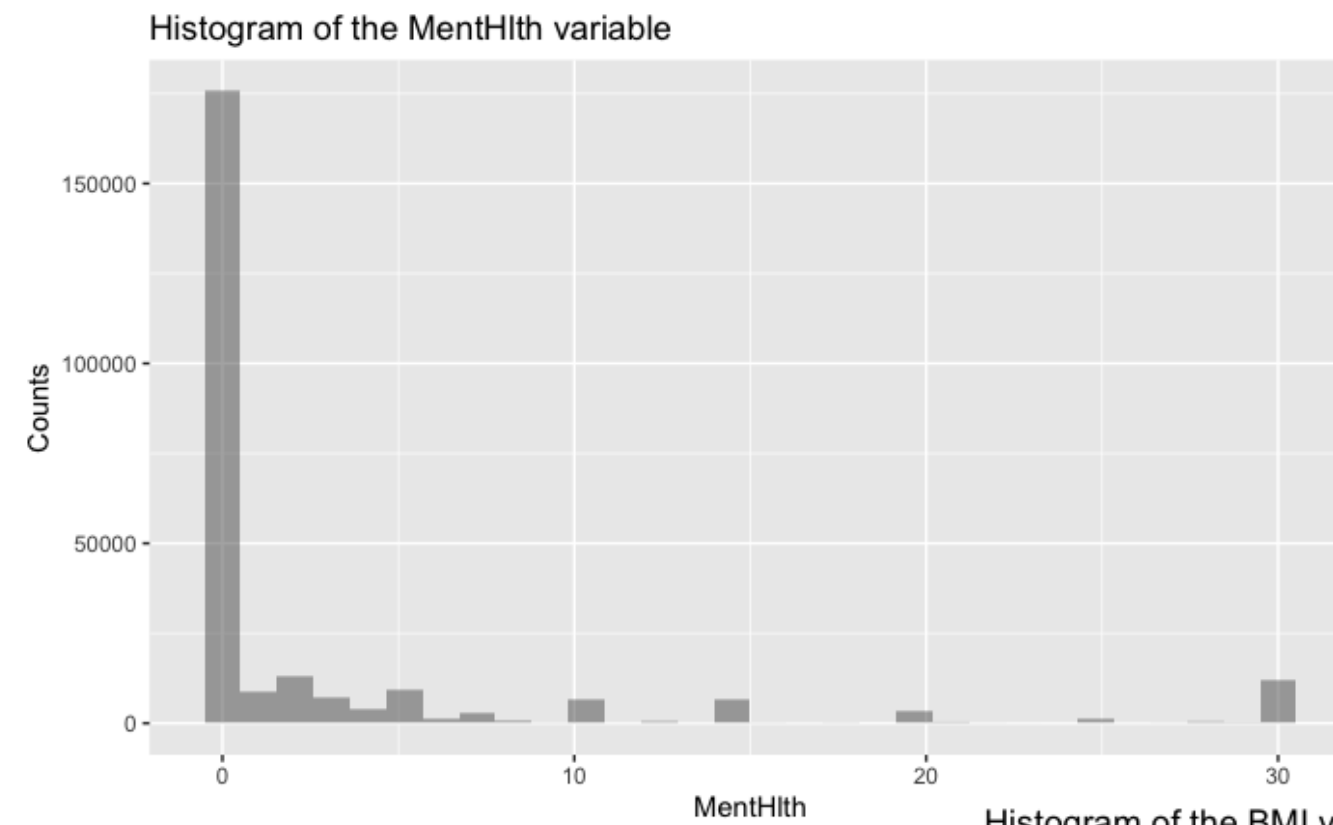
HighBP	HighChol	CholCheck	BMI
0	0	0	0
Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
0	0	0	0
Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare
0	0	0	0
NoDocbcCost	GenHlth	MentHlth	PhysHlth
0	0	0	0
DiffWalk	Sex	Age	Education
0	0	0	0
Income	diabete		
0	0		

Frequency of 0 and 1 for HighBP



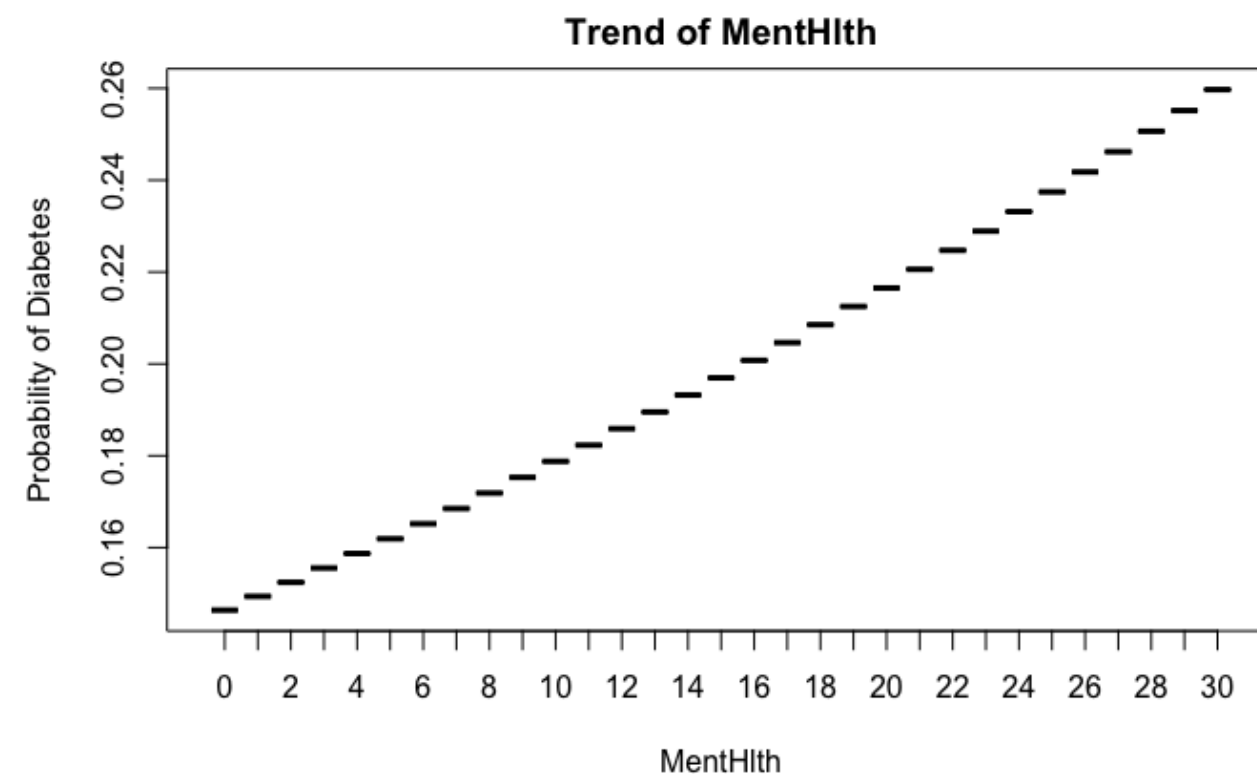
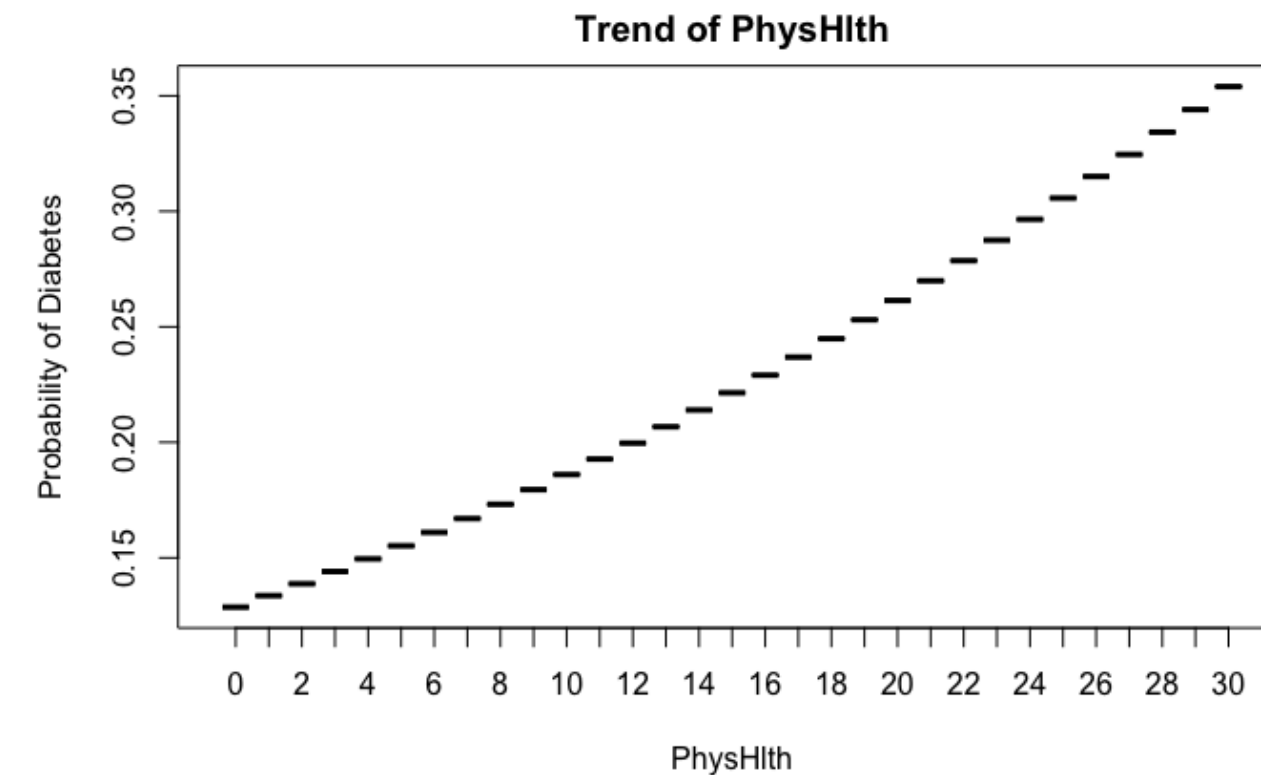
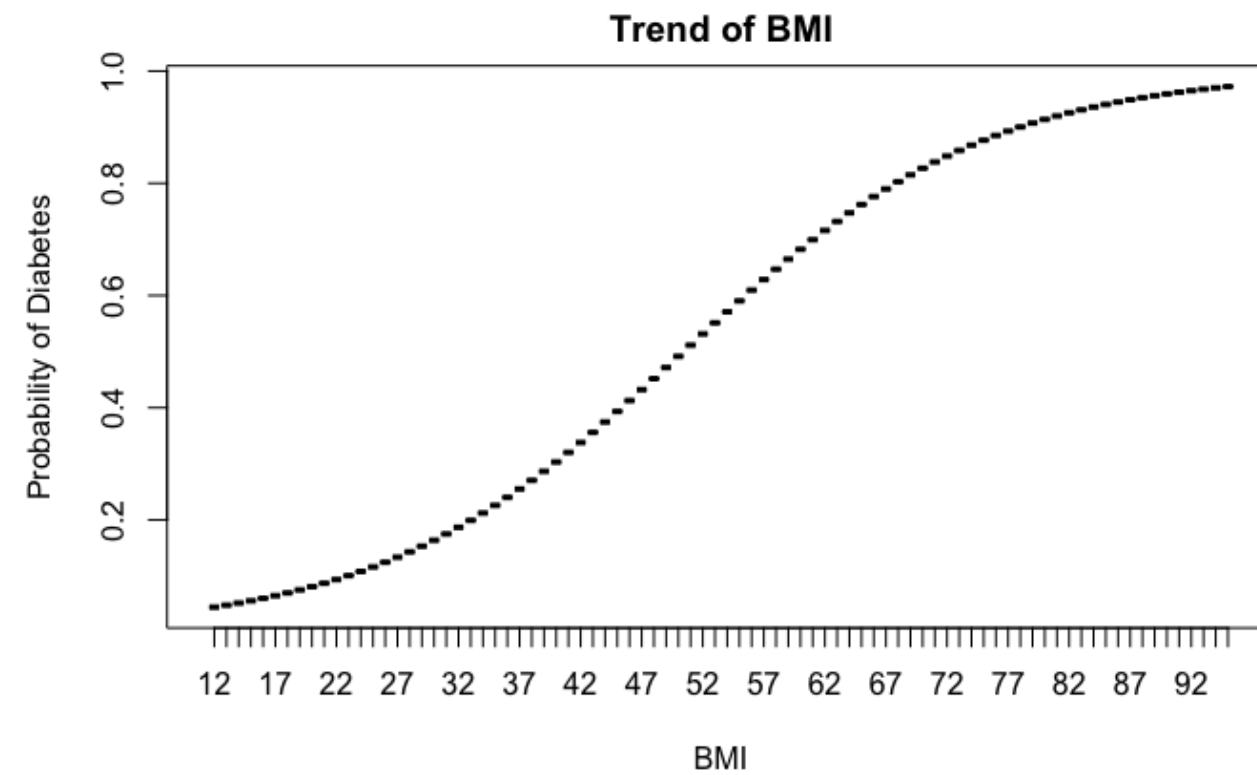
Frequency of 0 and 1 for DiffWalk



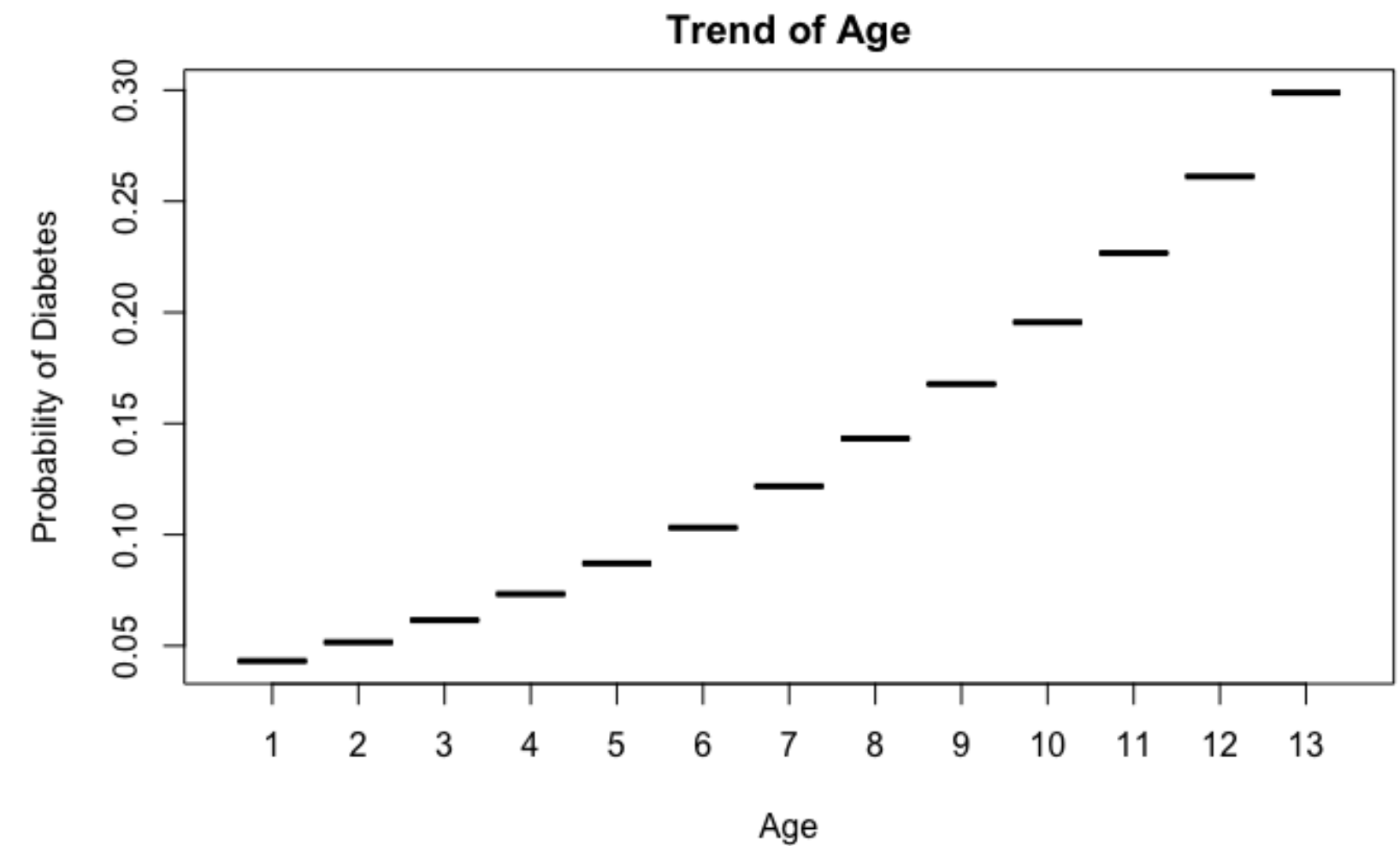
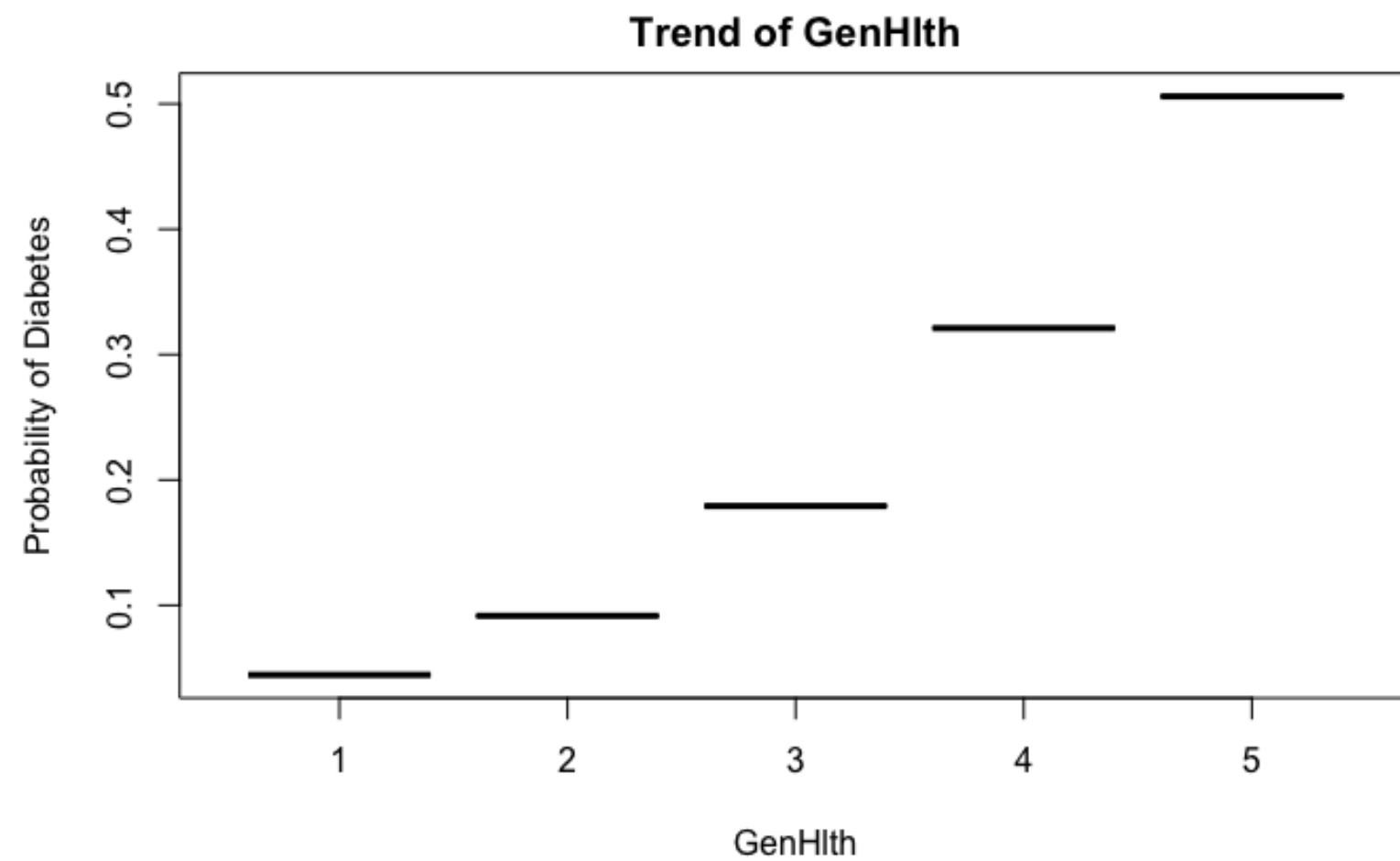




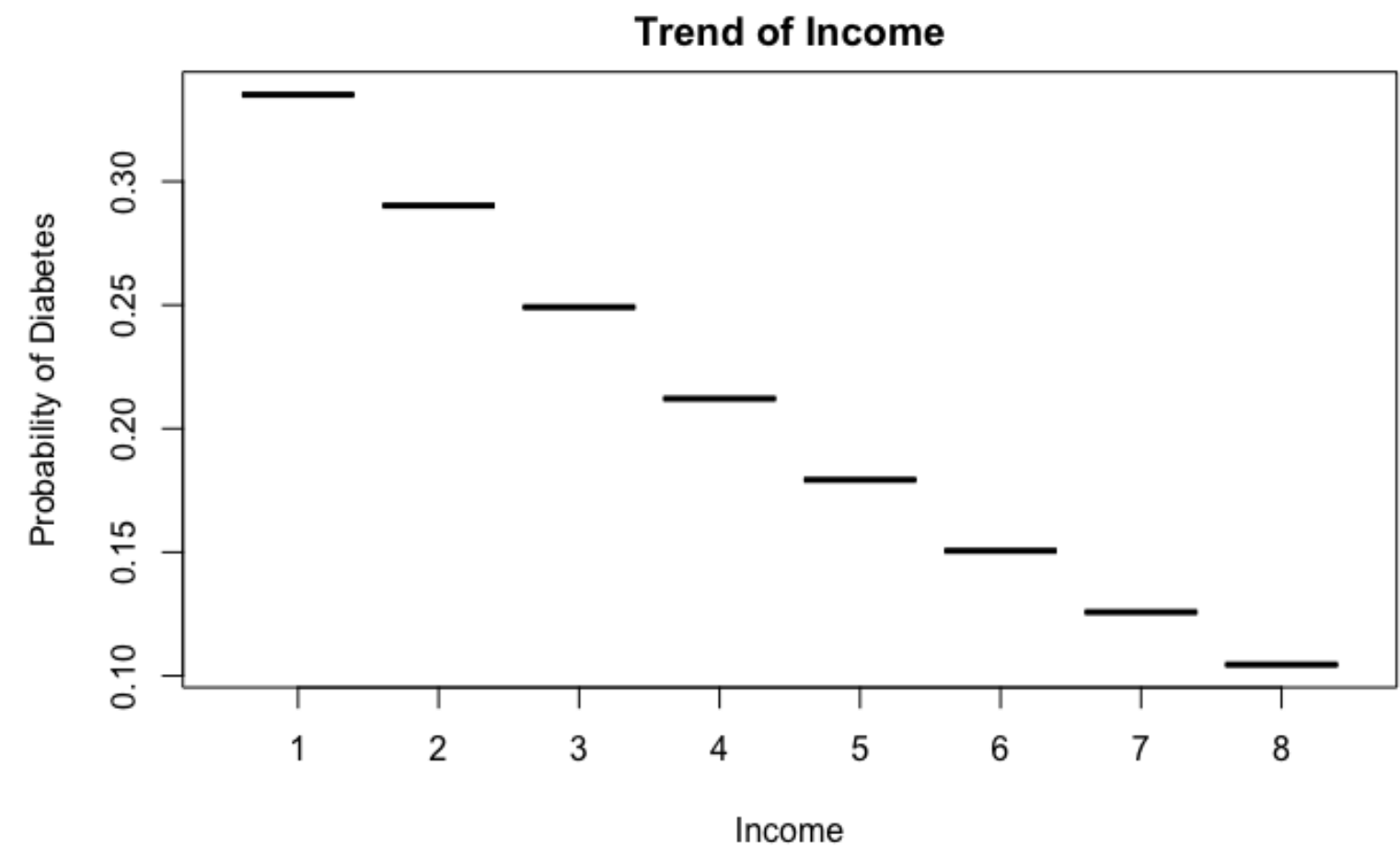
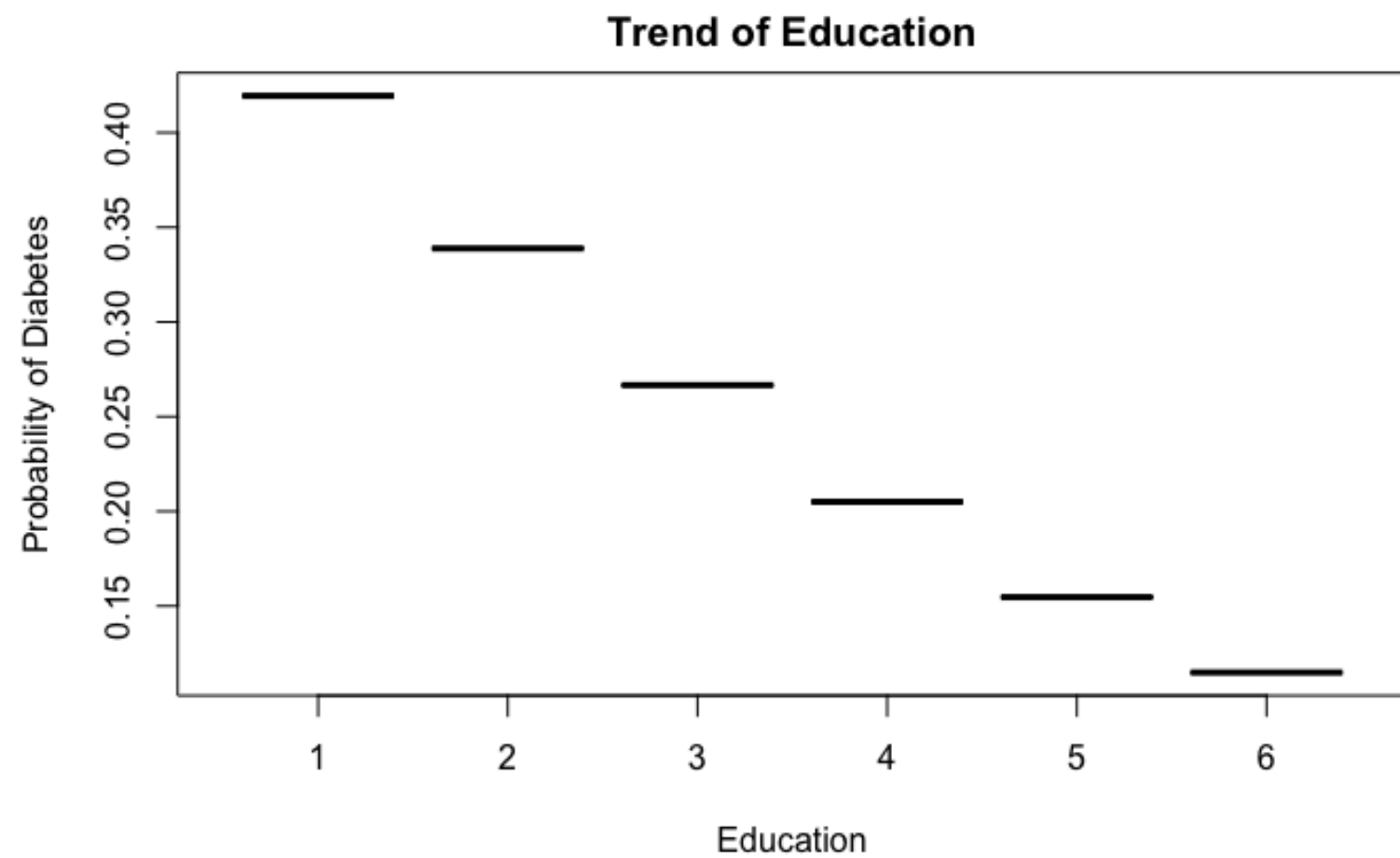
DATA MINING



Variables proportional
to diabetes rate



Variables proportional
to diabetes rate

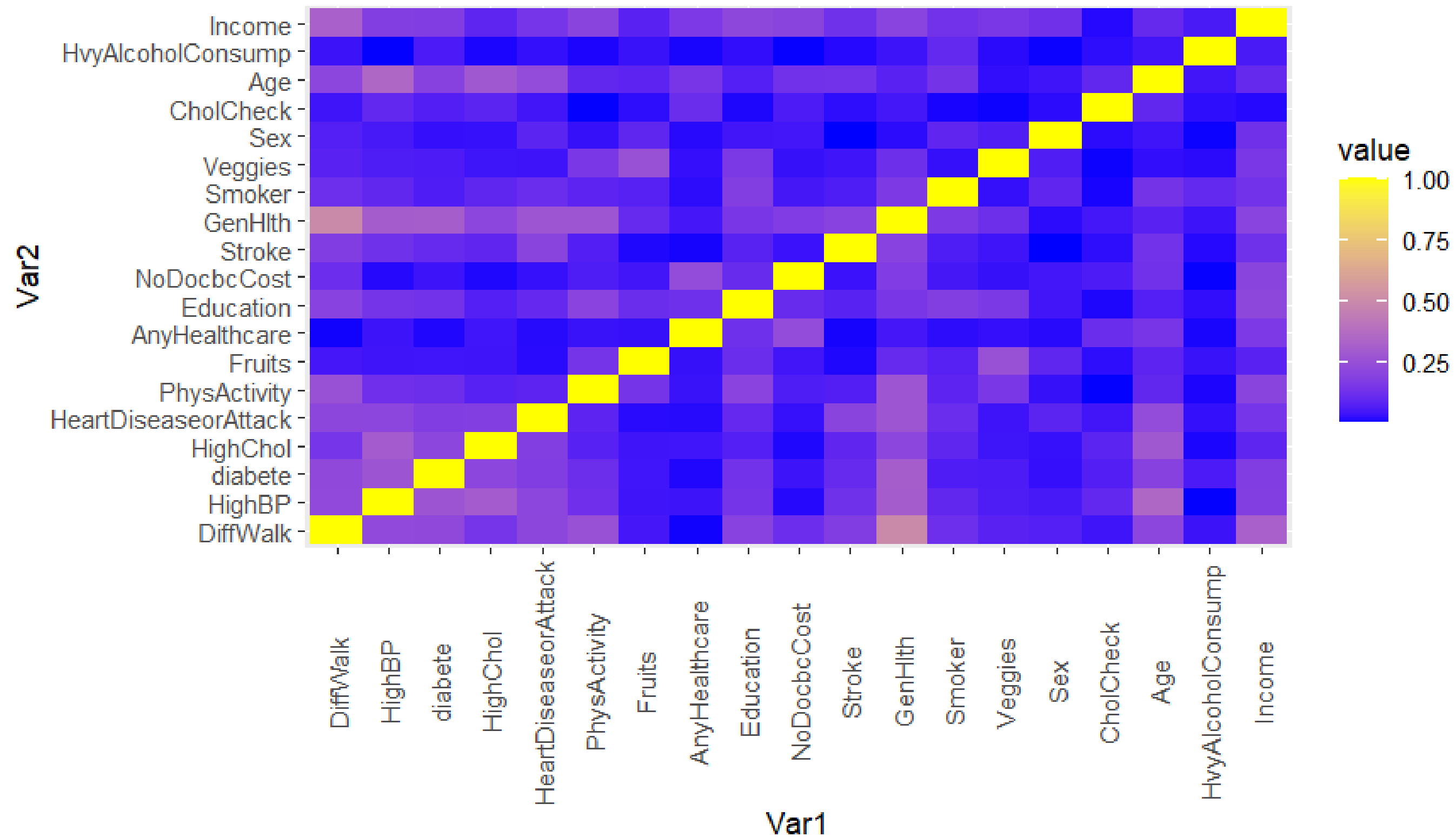


Variables inversely proportional
to diabetes rate



CORRELATIONS

Cramer's V correlation matrix





CORRELATIONS

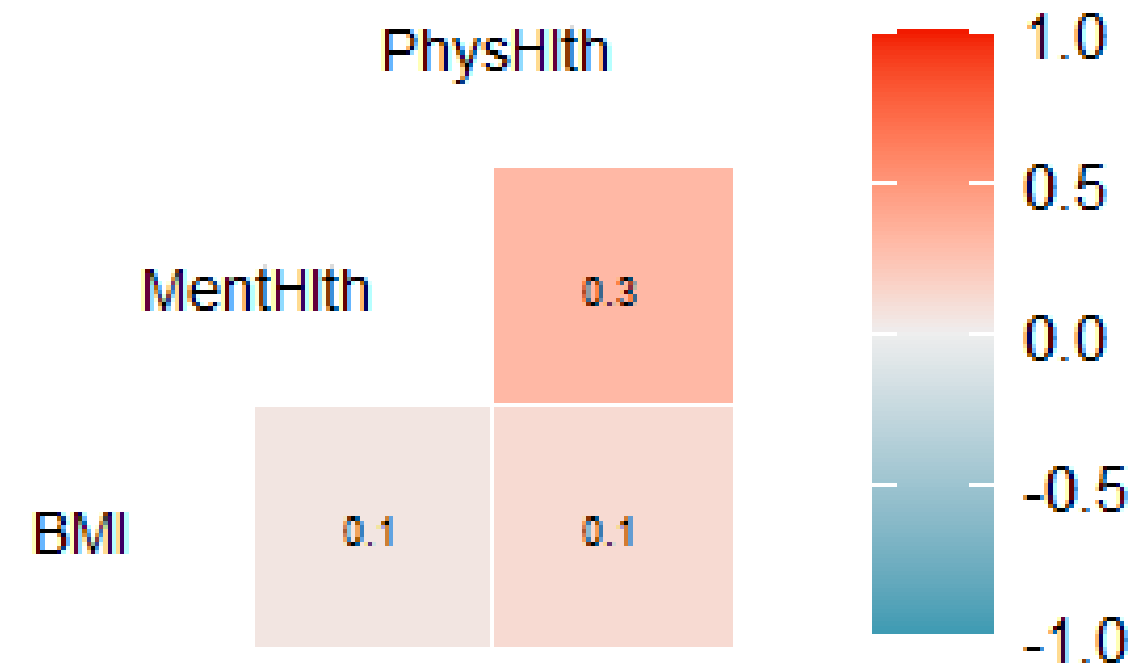
Variables with the highest Cramer's V with the target variable "diabetes"

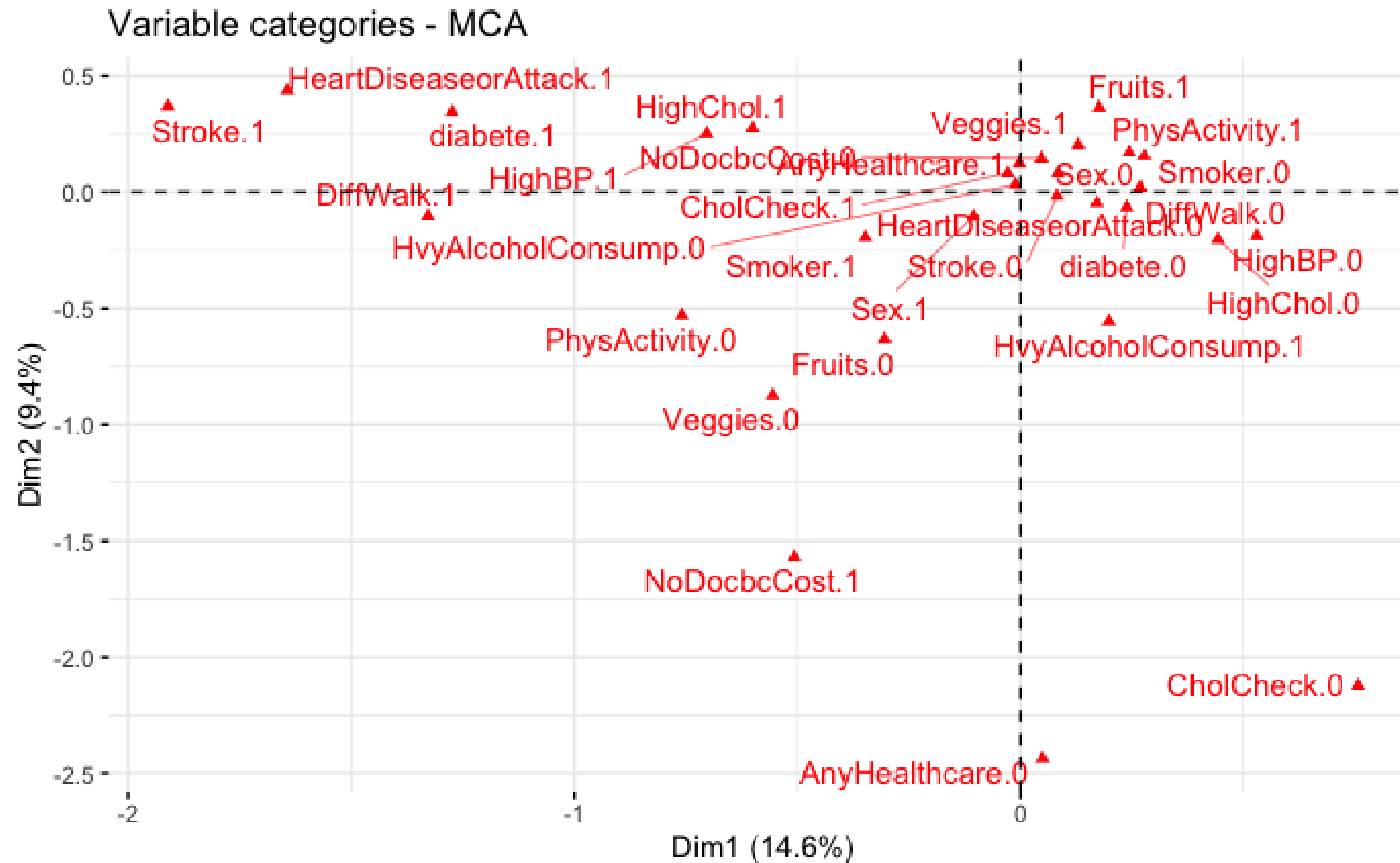
Var1 <chr>	Var2 <chr>	cramer_v <dbl>
diabetes	GenHlth	0.30624079
HighBP	diabetes	0.27032321
DiffWalk	diabetes	0.22214043
HighChol	diabetes	0.21027872
diabetes	Age	0.19391780



CORRELATIONS

Correlation matrix with numerical variables

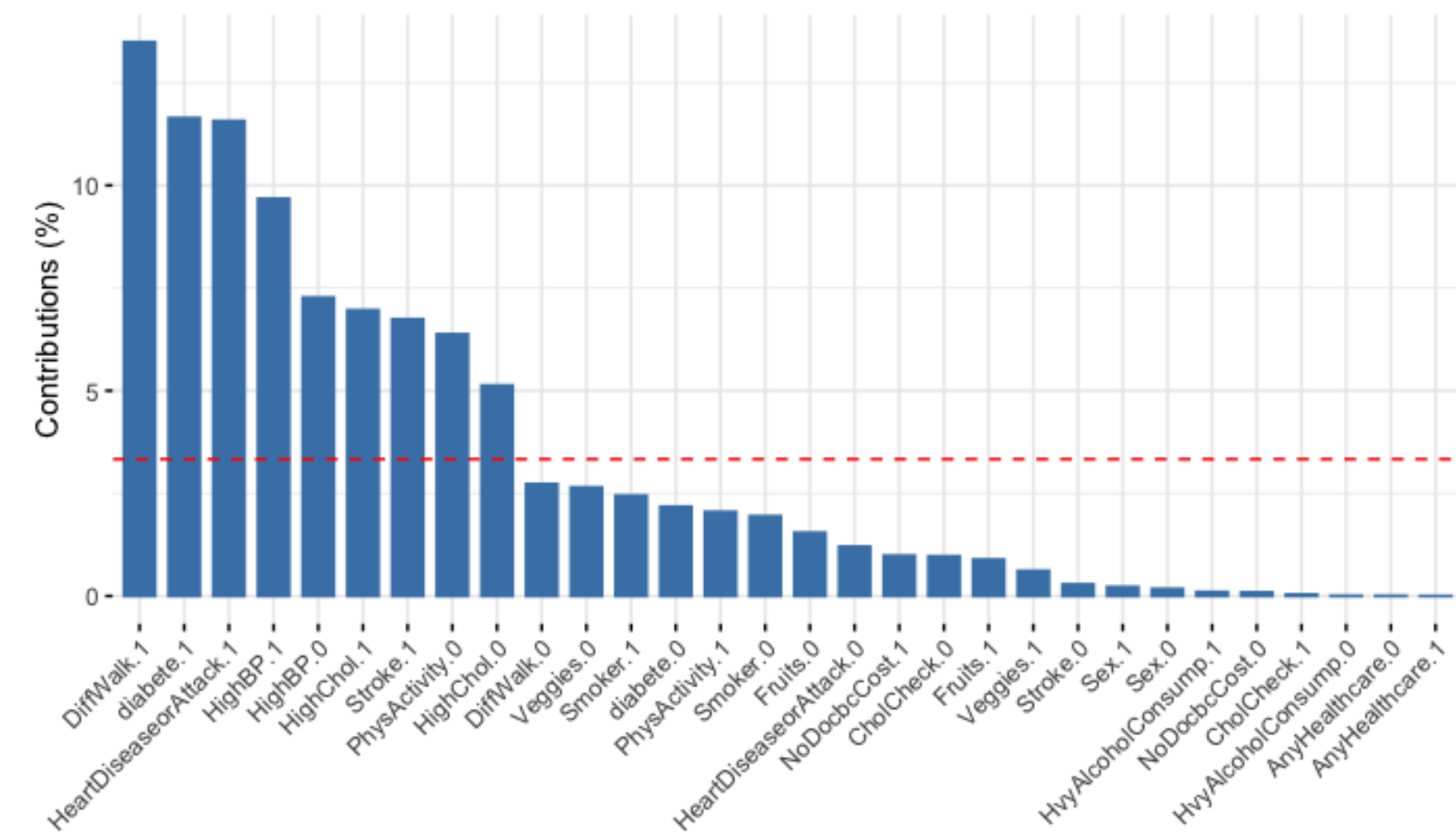




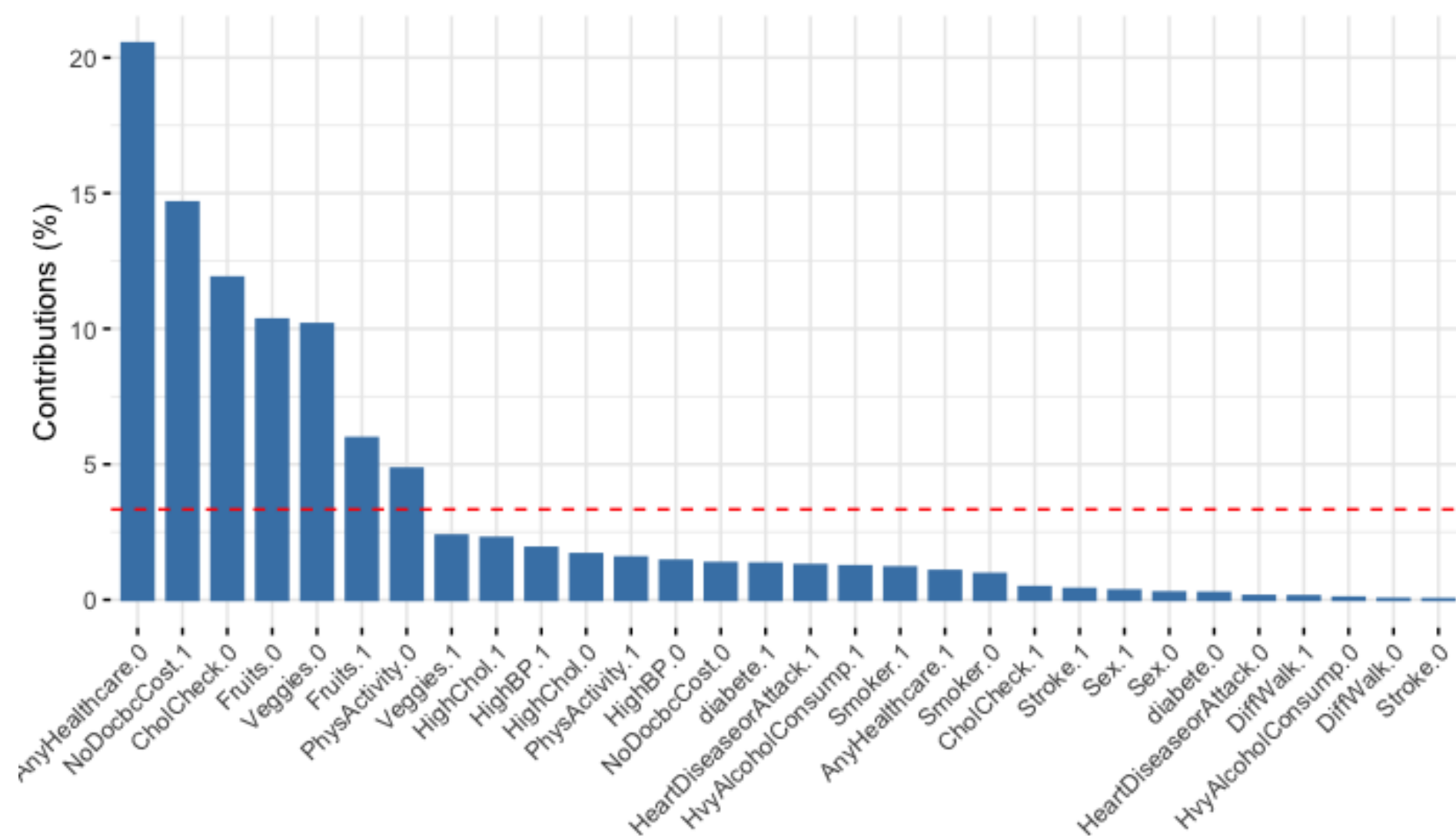


CORRELATIONS

Contribution of variables to Dim-1



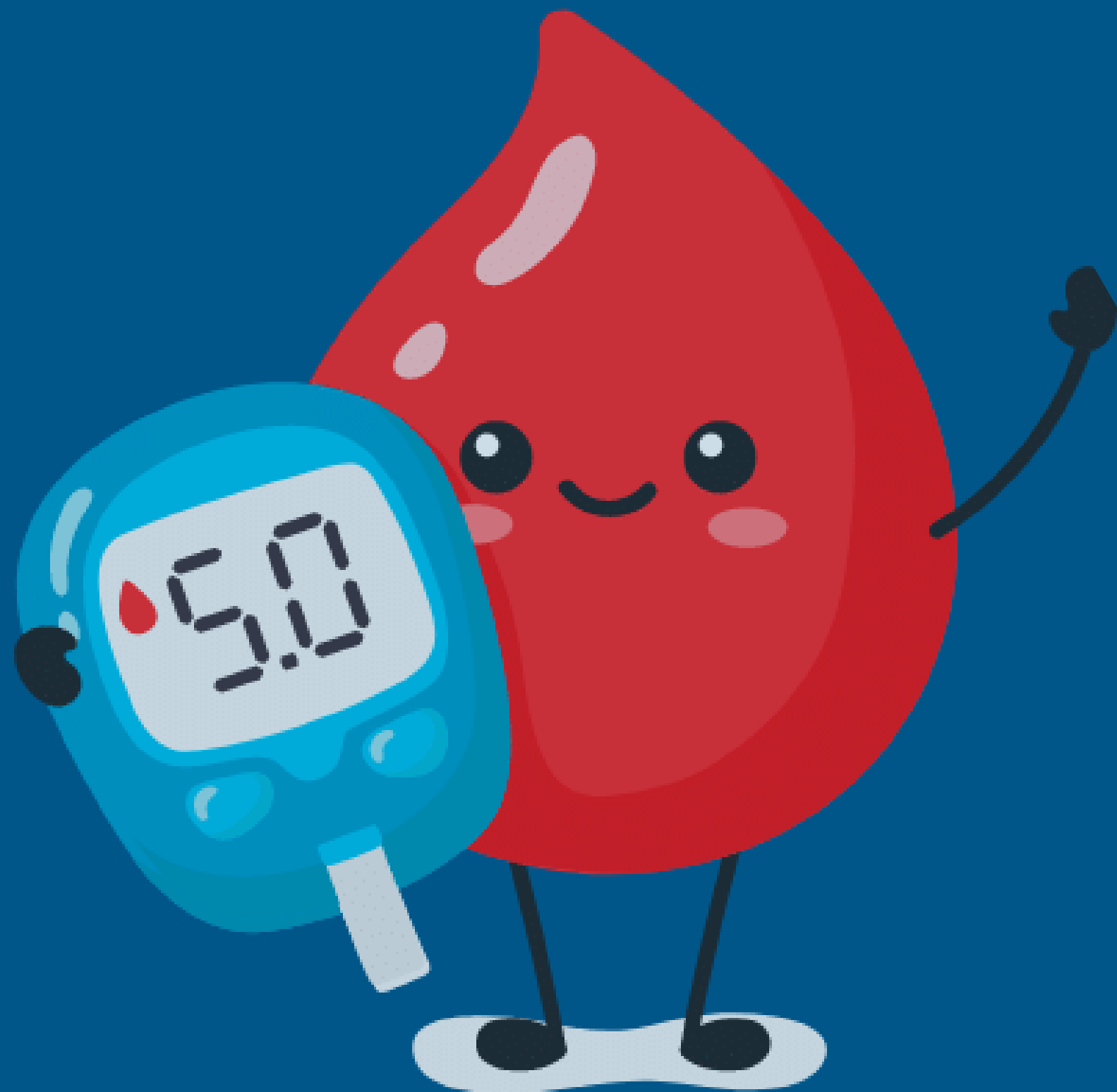
Contribution of variables to Dim-2





CONCLUSION

- mainly categorical variables
- no significant correlation between the variables, so no redundant variables
- some variables have more weight in the explanation of the variable to be explained : GenHlth, HighBP, DiffWalk, HighChol, Age
- possibility of adding a column (for example, individuals with a family member suffering from diabetes)



Thanks !
