ITBA
Instituto Tecnológico de Buenos Aires

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)

ITBA

AGENDA

01
Context

02
Content of the data set

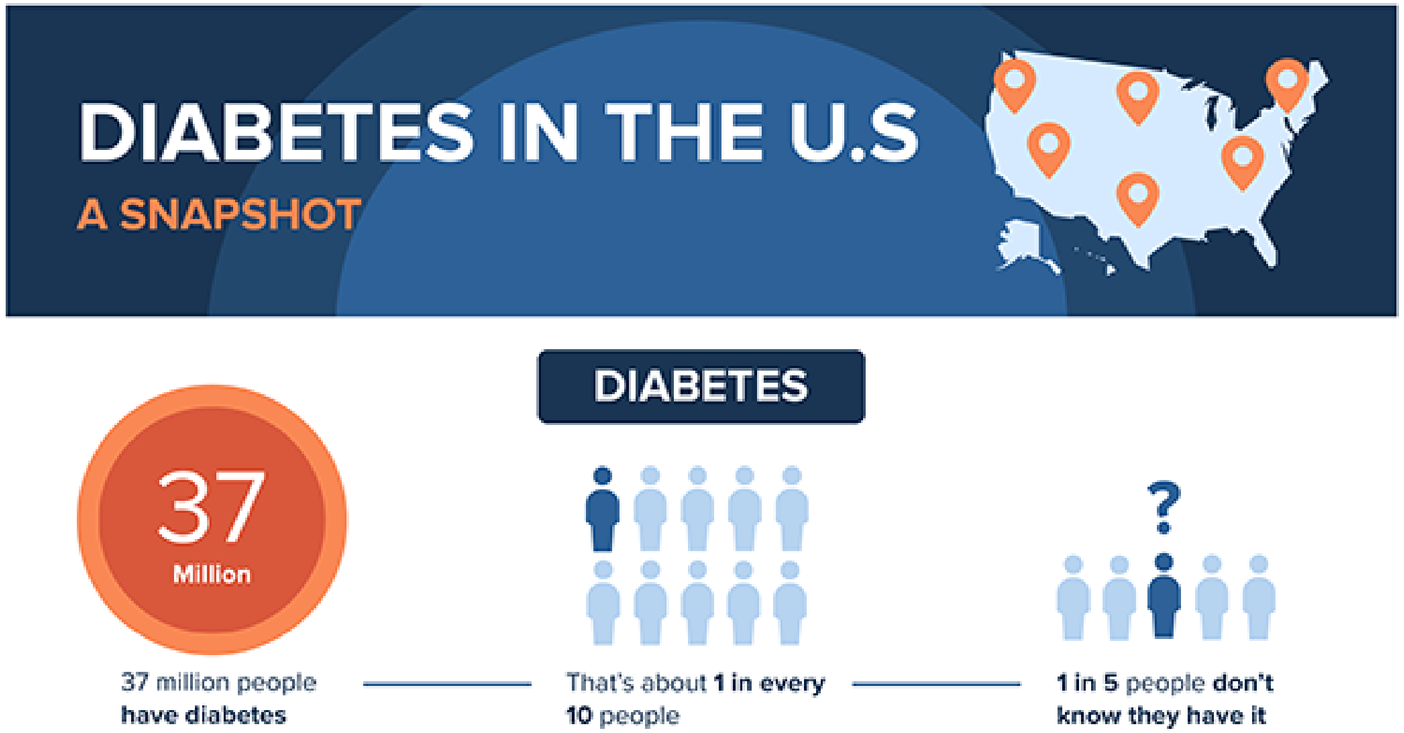
03
Data mining

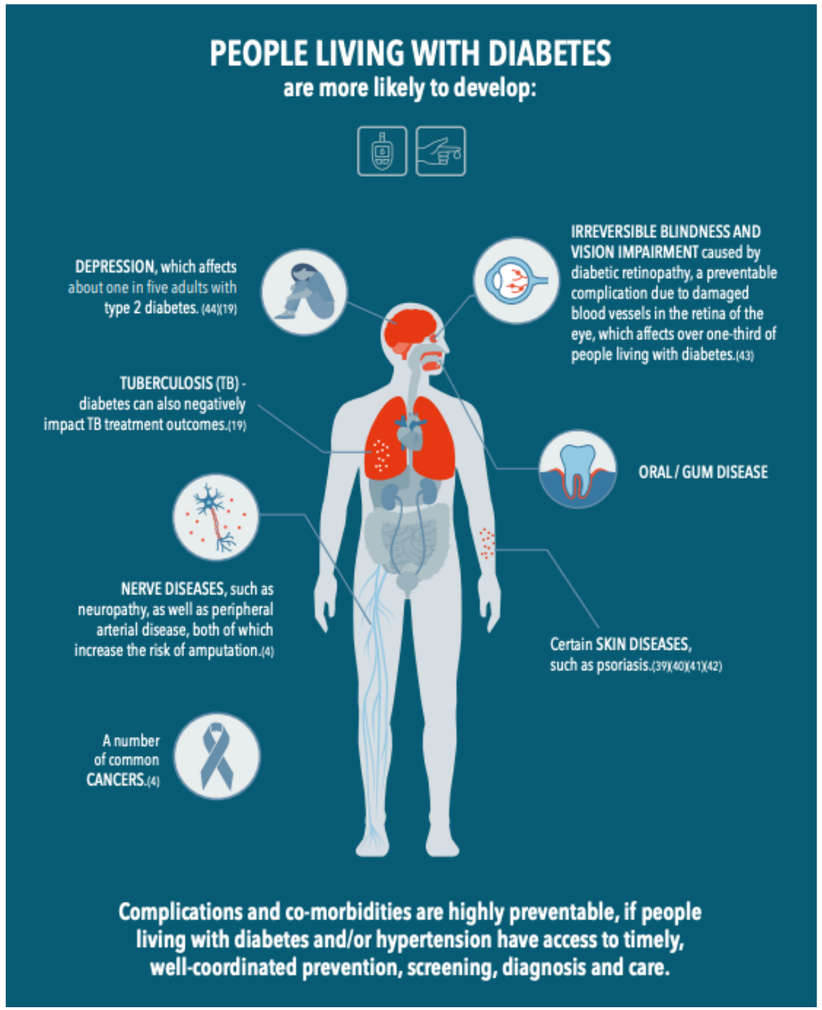
04
Correlations

05
Conclusion

ITBA

CONTEXT





ITBA

CONTENT

kaggle → CDC Centers for Disease Control and Prevention — BRFSS™ Behavioral Risk Factor Surveillance System

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

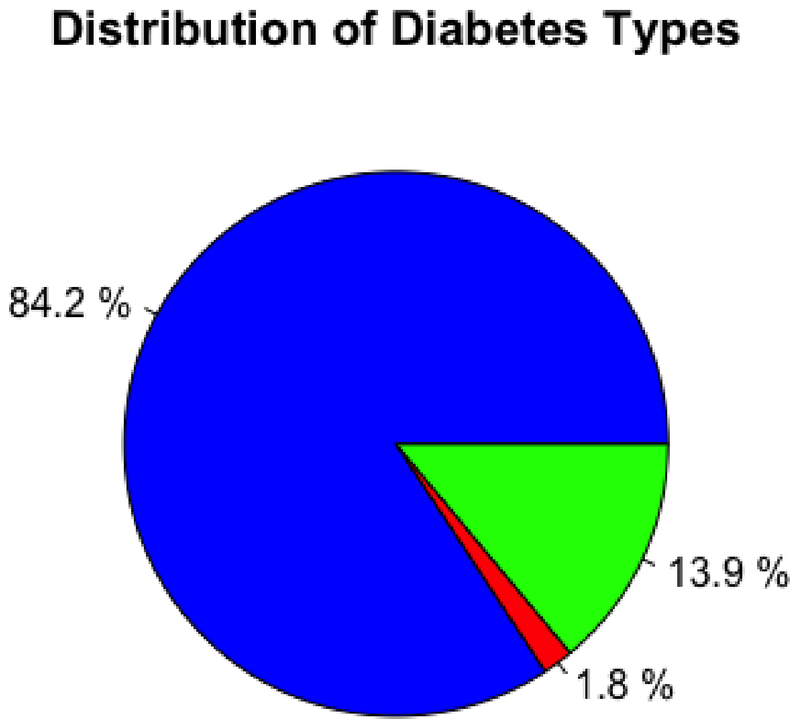
Numerical :

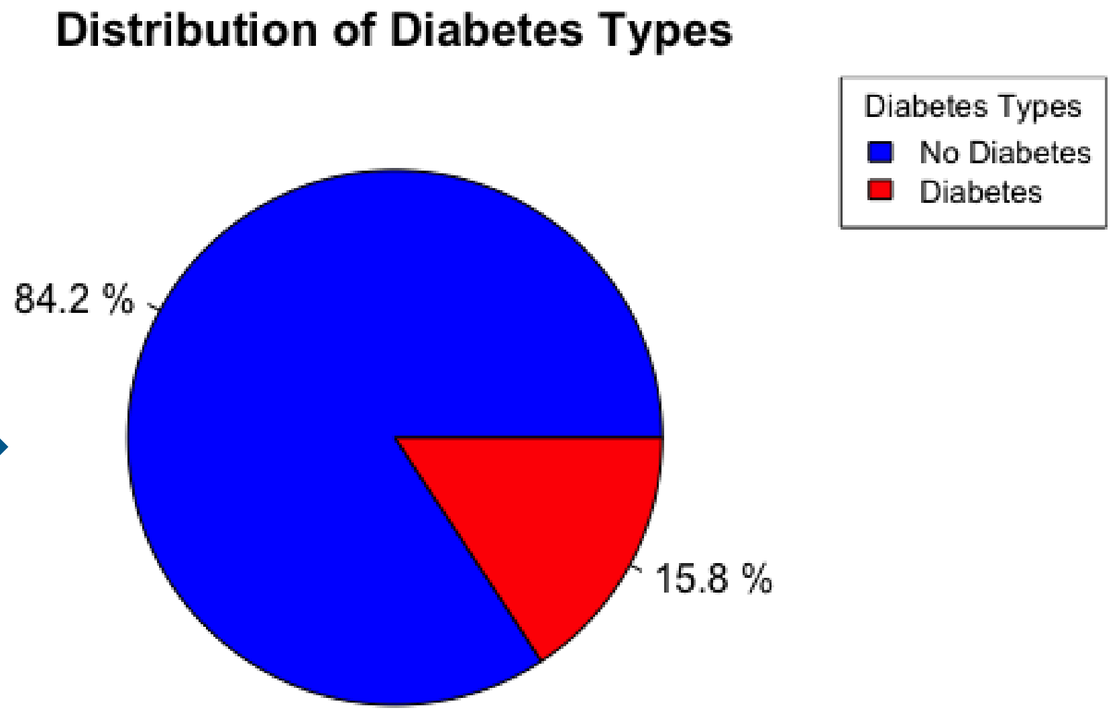
- MentHlth
- PhysHlth
- BMI

ITBA

CONTENT

Response variable





CONTENT

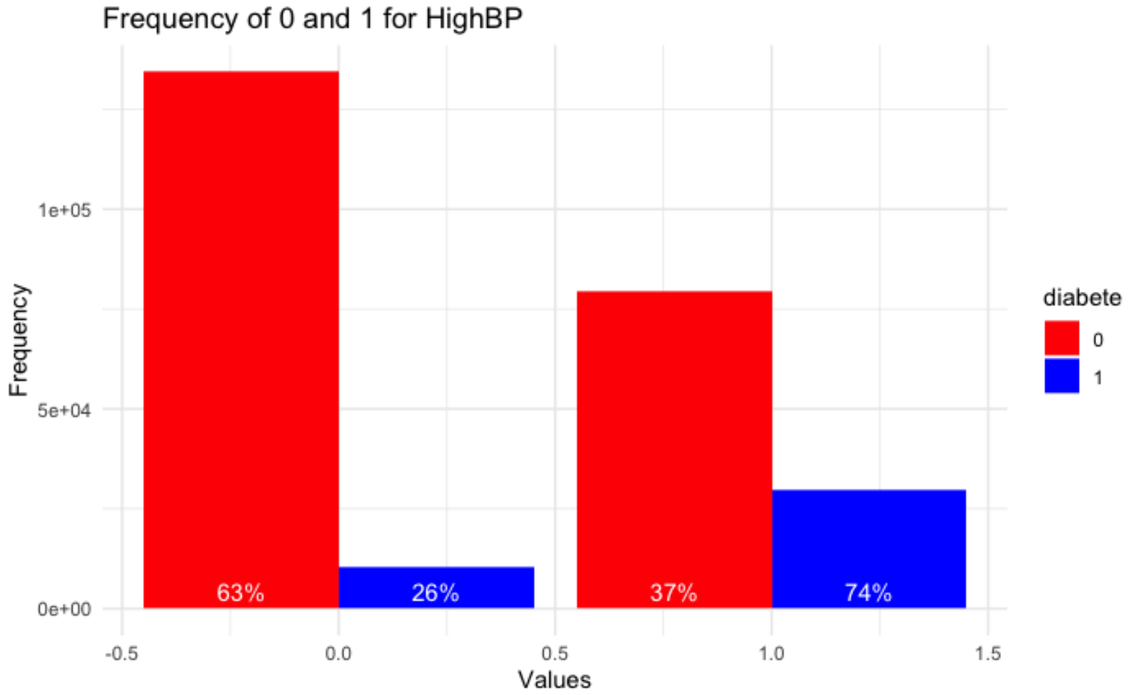
No missing !

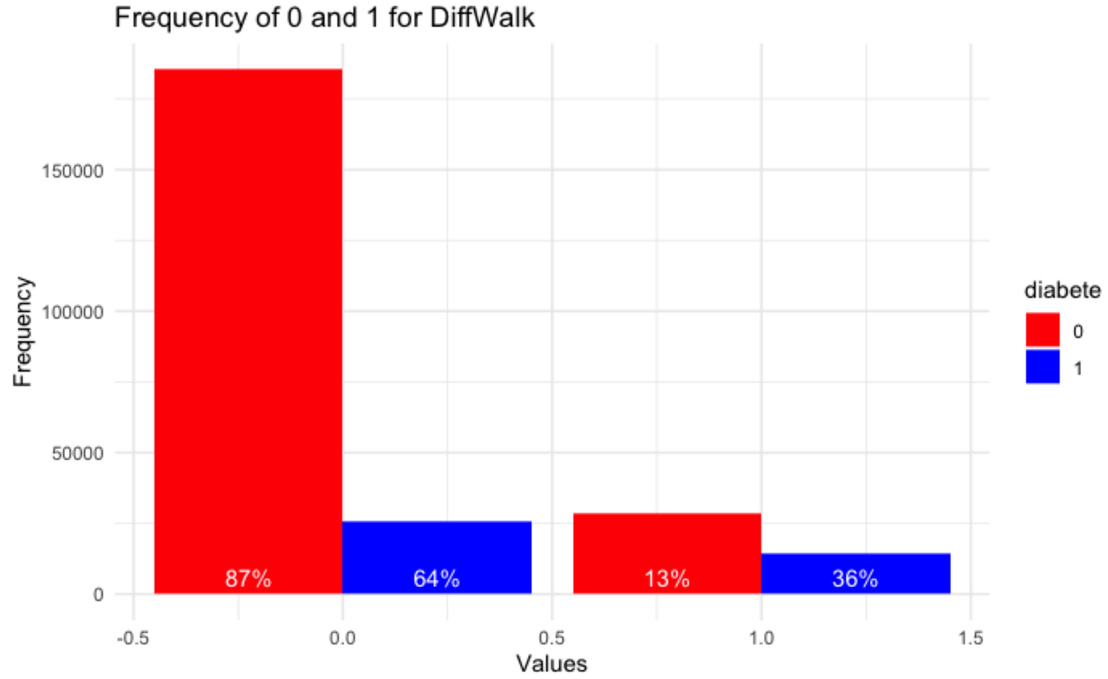
```{r}
na_counts <- colSums(is.na(data))
print(na_counts)
```

```
            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
```

ITBA

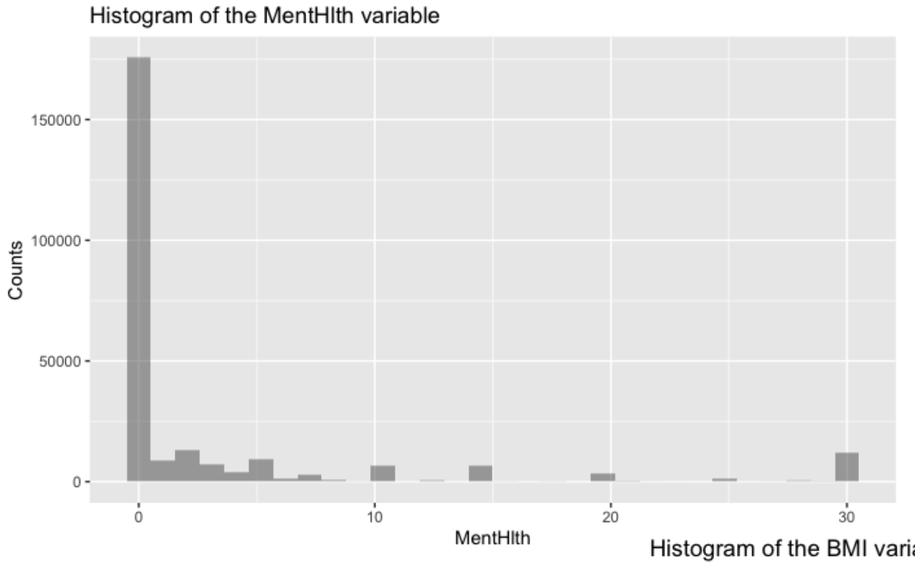
DATA MINING

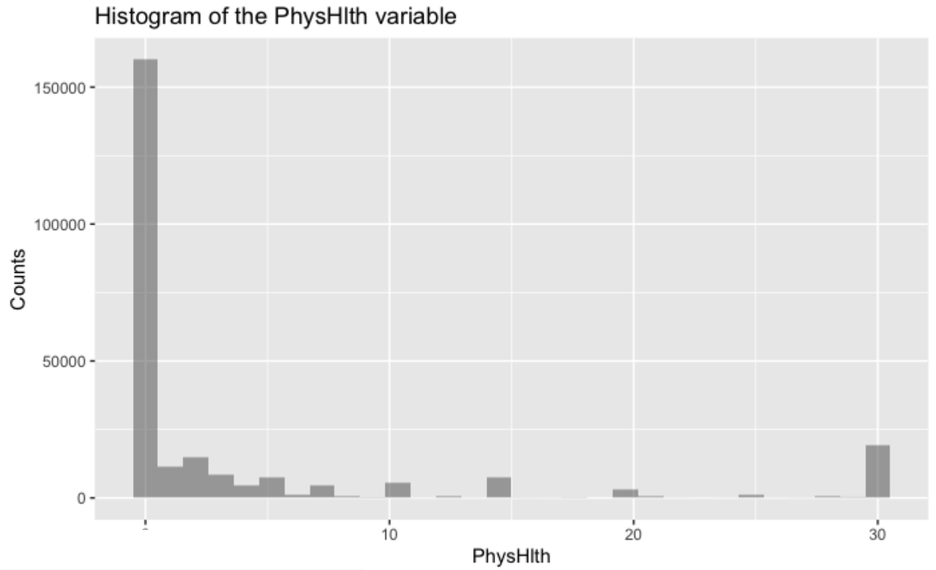


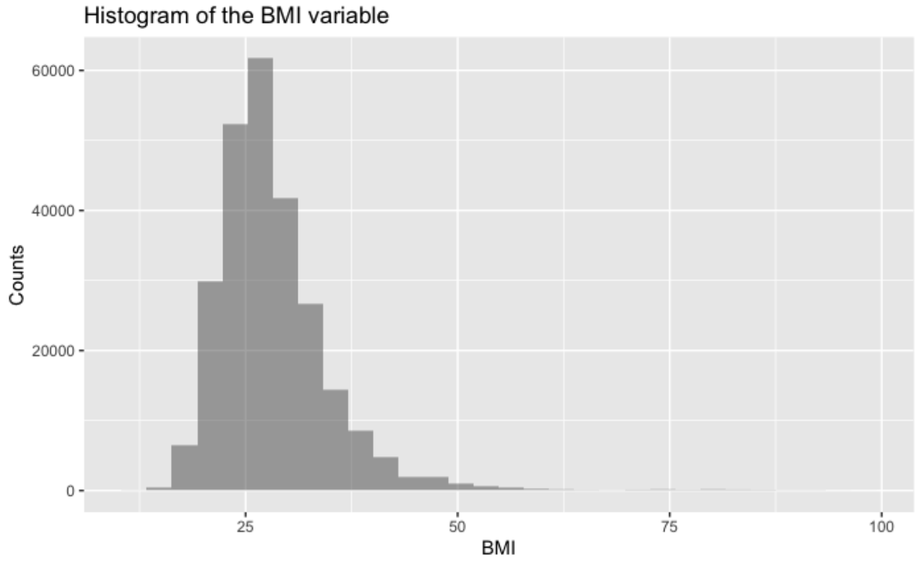


ITBA

DATA MINING

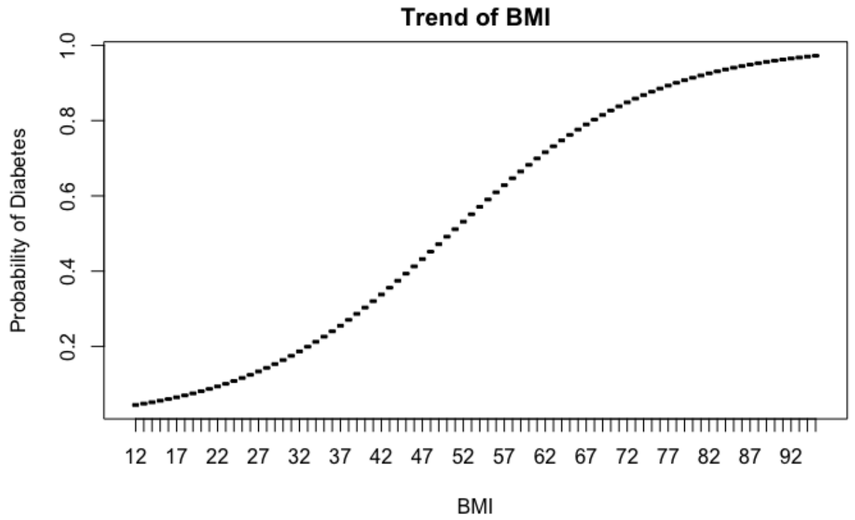


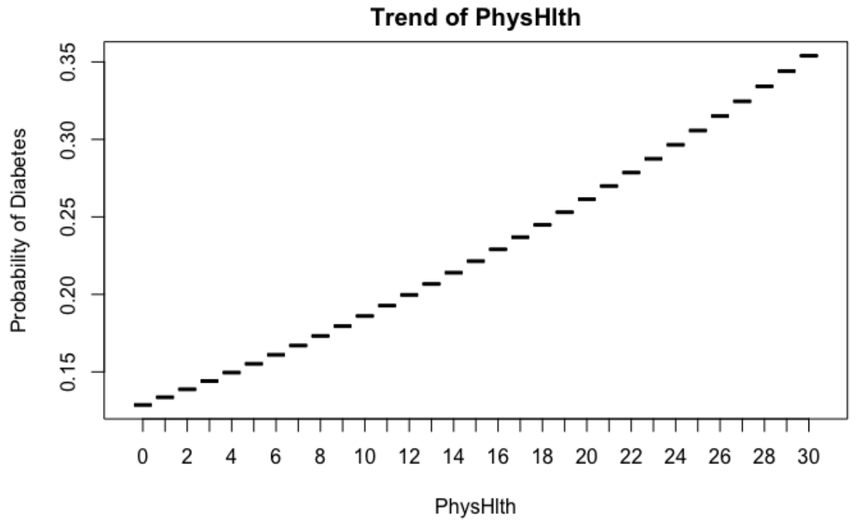


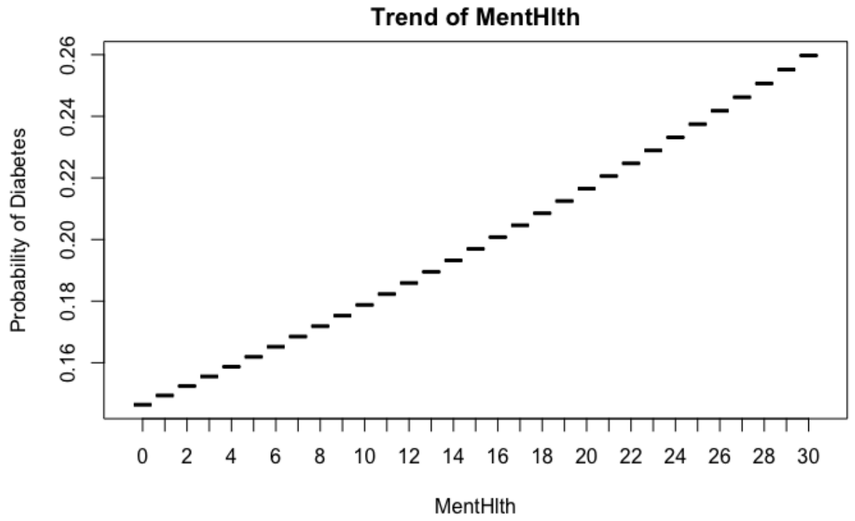


ITBA

DATA MINING



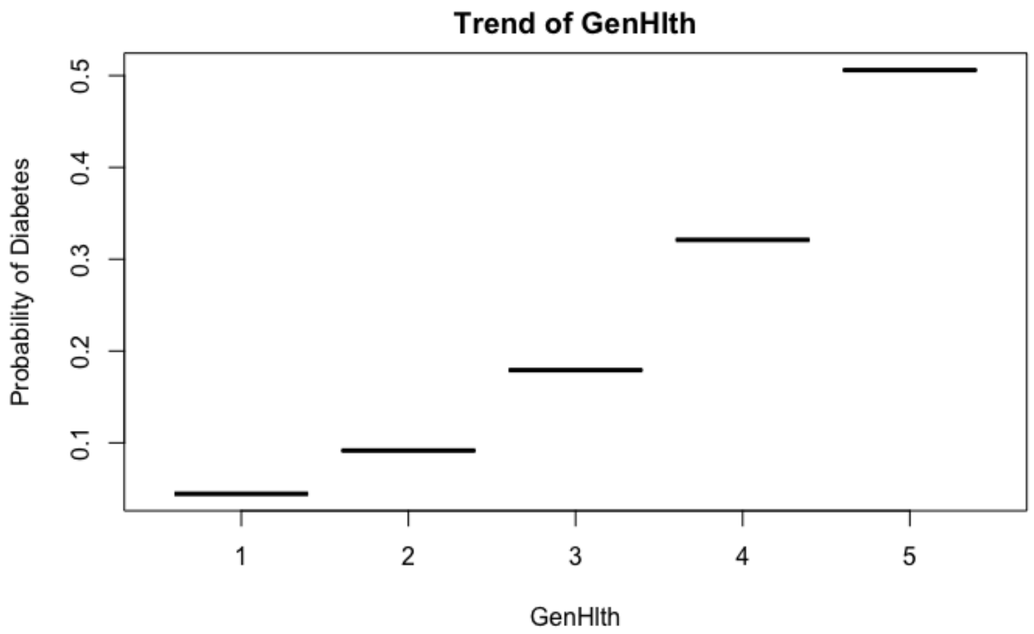


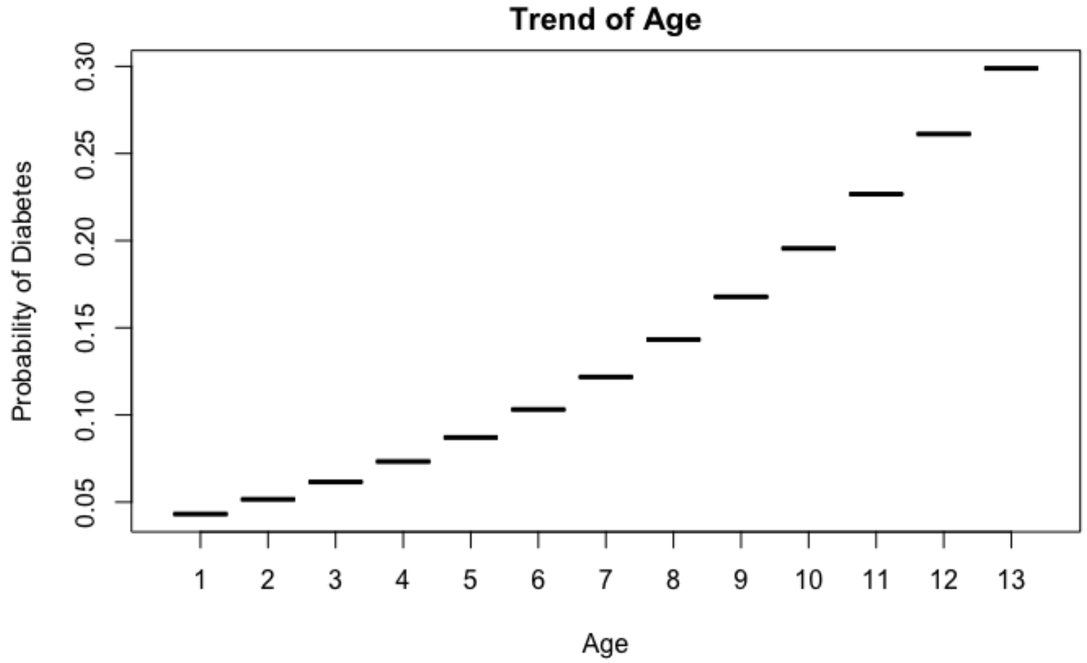


Variables proportional
to diabetes rate

ITBA

DATA MINING

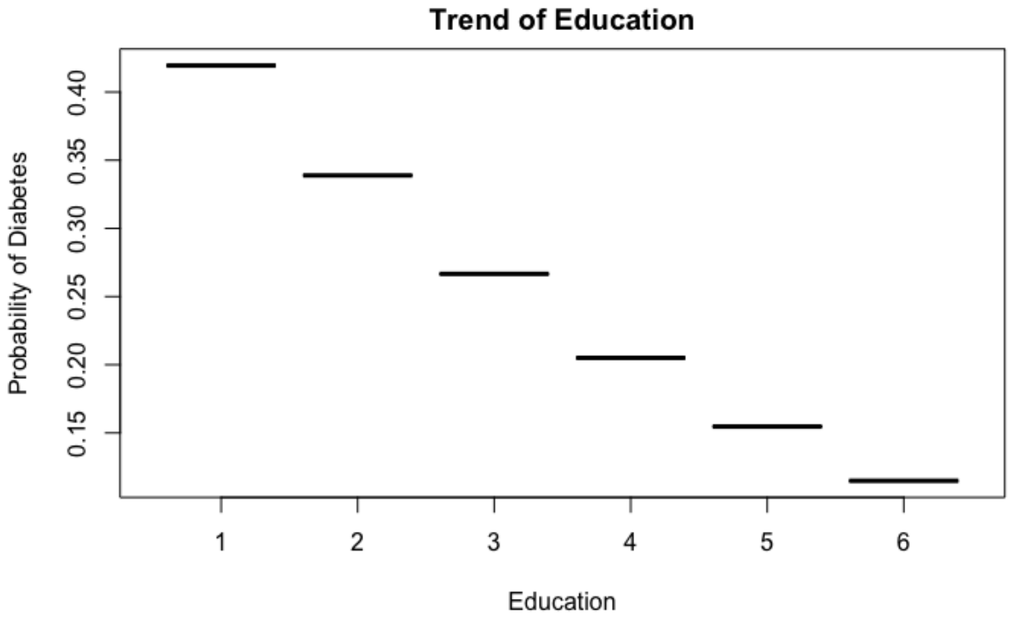


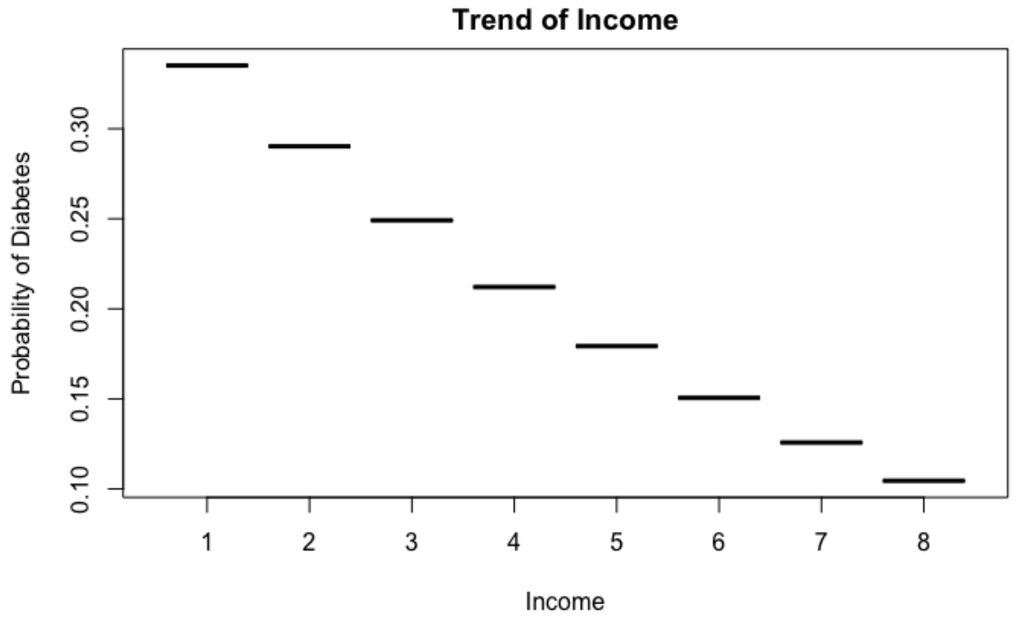


Variables proportional
to diabetes rate

ITBA

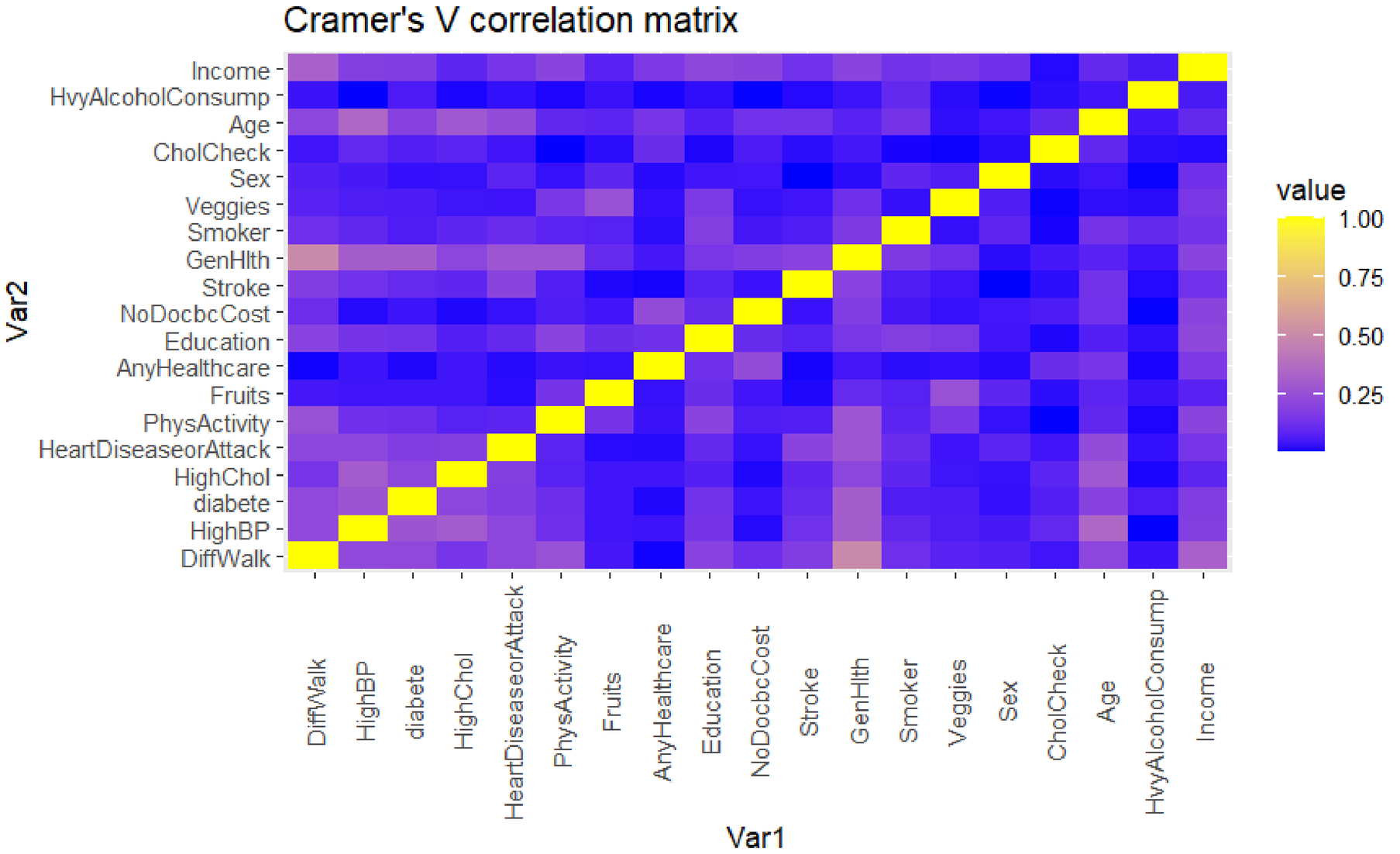
DATA MINING





Variables inversely proportional
to diabetes rate

CORRELATIONS



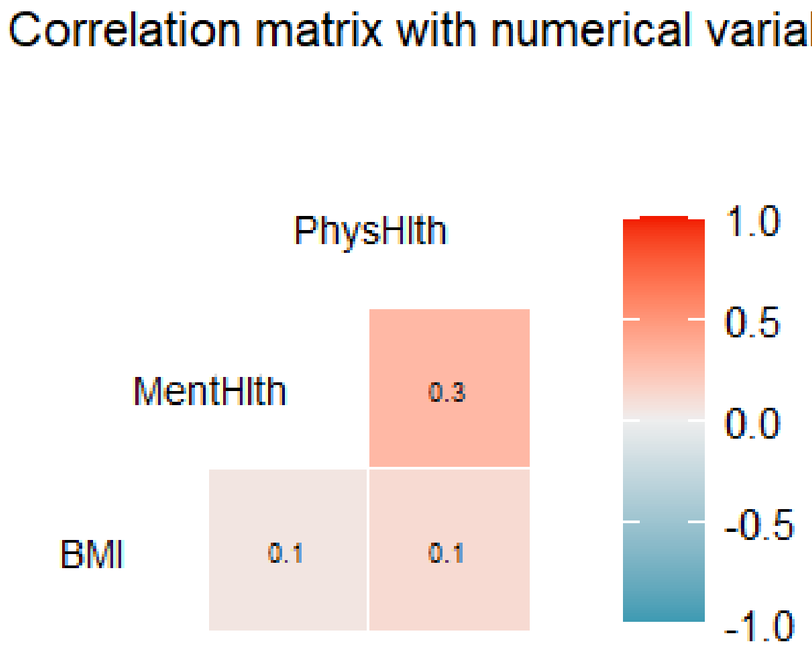
CORRELATIONS

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

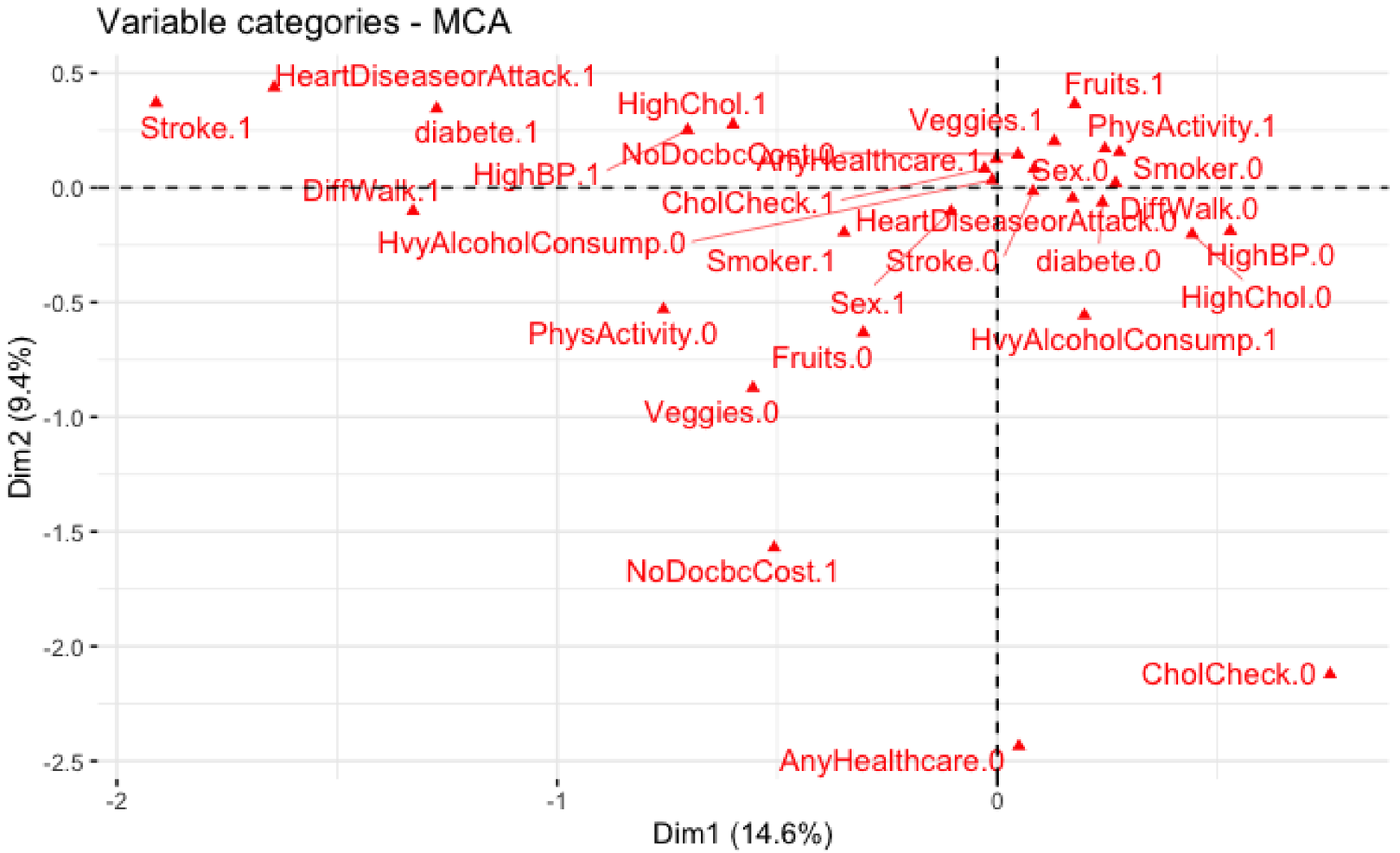
Var1 <chr>	Var2 <chr>	cramer_v <dbl>
diabete	GenHlth	0.30624079
HighBP	diabete	0.27032321
DiffWalk	diabete	0.22214043
HighChol	diabete	0.21027872
diabete	Age	0.19391780

ITBA

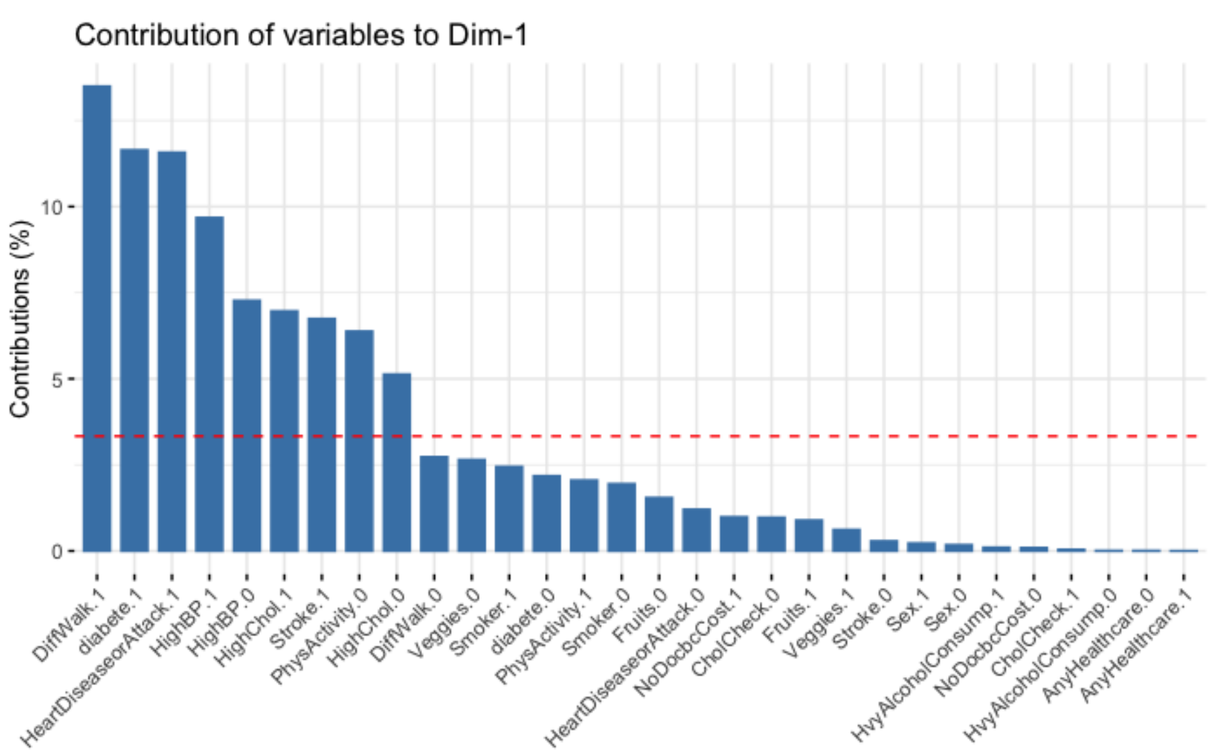
CORRELATIONS

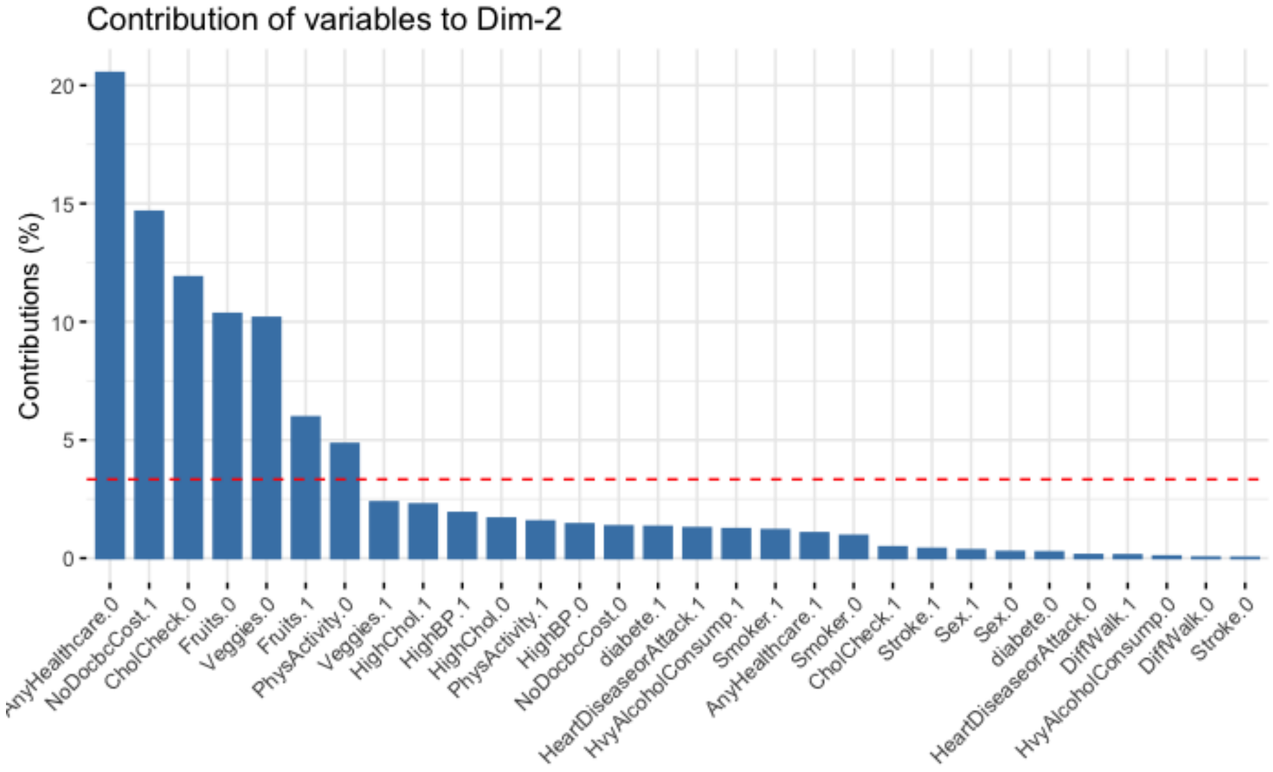


CORRELATIONS



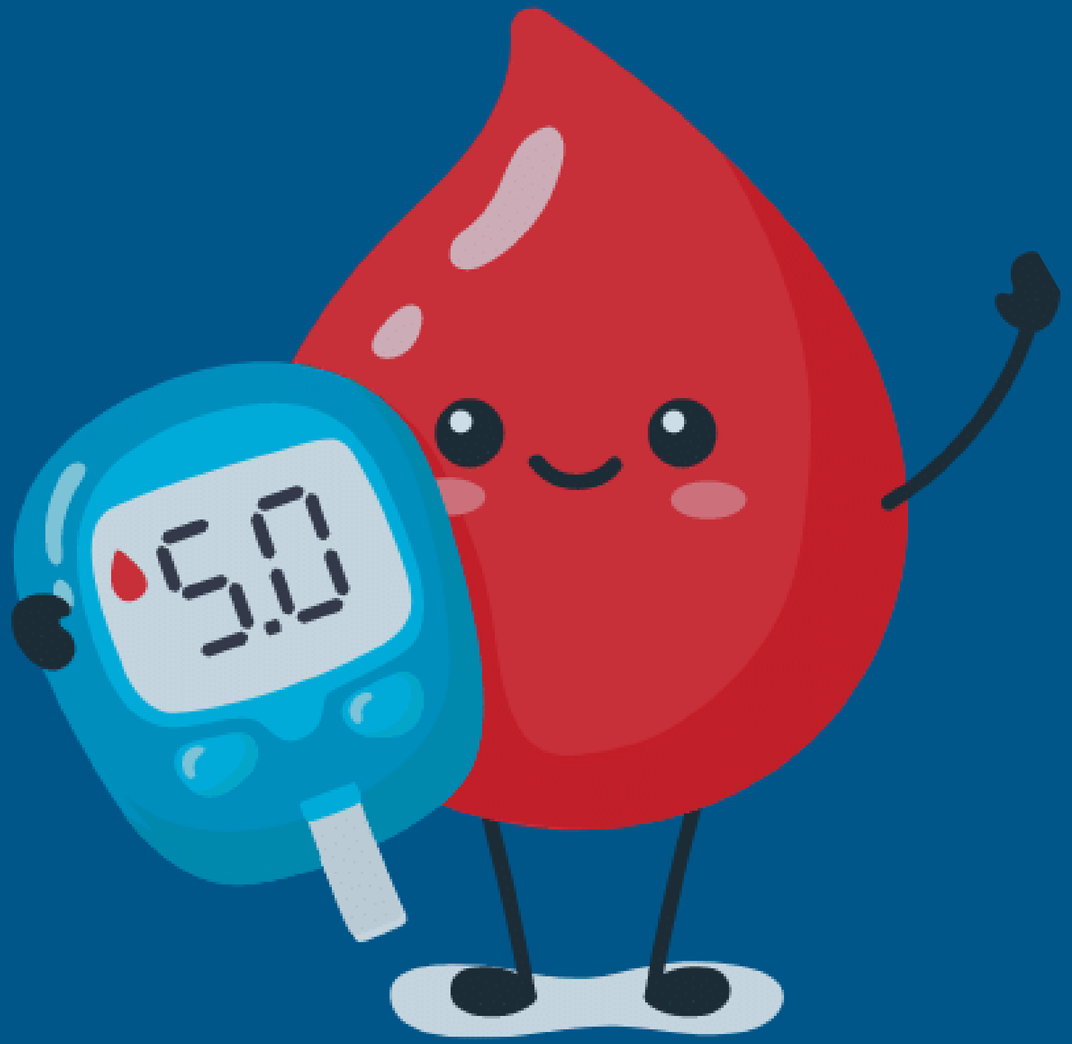
CORRELATIONS





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 !