STATS115_preliminary_analysis

Seiya Uno

Preliminary Analysis

```
Warning: package 'Hmisc' was built under R version 4.3.3
```

```
Warning: package 'corrplot' was built under R version 4.3.3
```

```
df <- read.csv("./data/binary_diabetes.csv")
sample_n(df, 5)
```

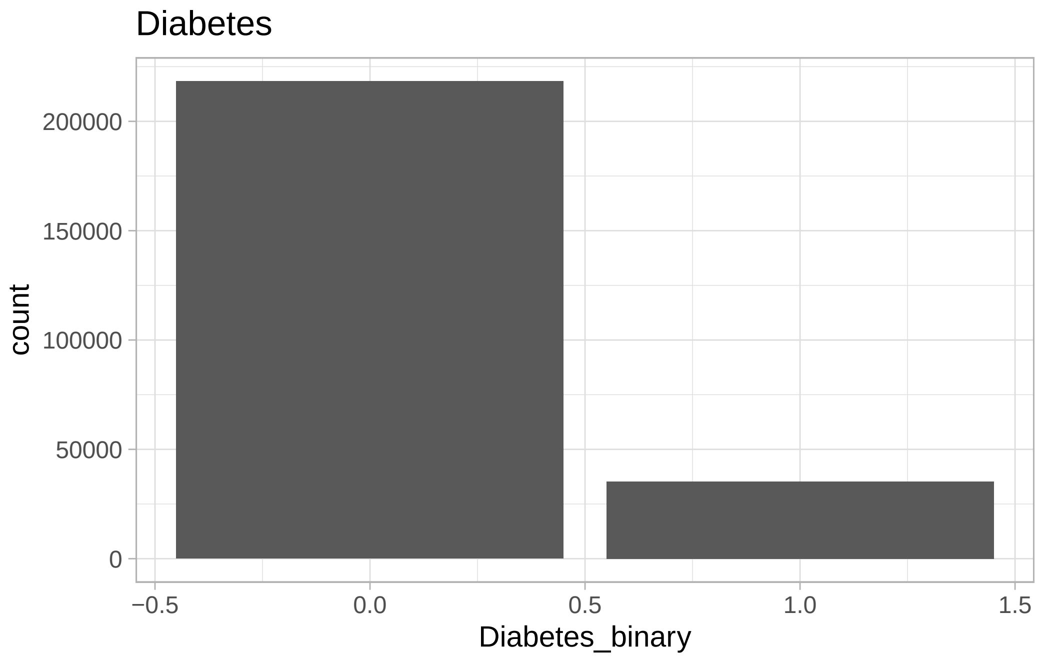
```
  Diabetes_binary HighBP HighChol CholCheck BMI Smoker Stroke
1               0      0        0         0  36      1      0
2               1      1        0         1  36      0      0
3               0      1        0         1  26      1      1
4               1      1        1         1  37      0      0
5               0      0        0         1  25      1      0
  HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
1                    0            1      1       1                 0
2                    0            1      0       1                 0
3                    0            1      1       1                 0
4                    1            0      0       0                 0
5                    0            1      1       1                 0
  AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
1             1           0       2        0        0        0   1   4
2             1           0       3        0        0        0   1  10
3             1           0       3        0        0        0   0  10
4             1           0       4        0       15        1   0   9
5             1           0       2        0        0        0   1   8
  Education Income
1         5      7
2         6      8
```

```
3           5       5
4           5       4
5           5       4
```

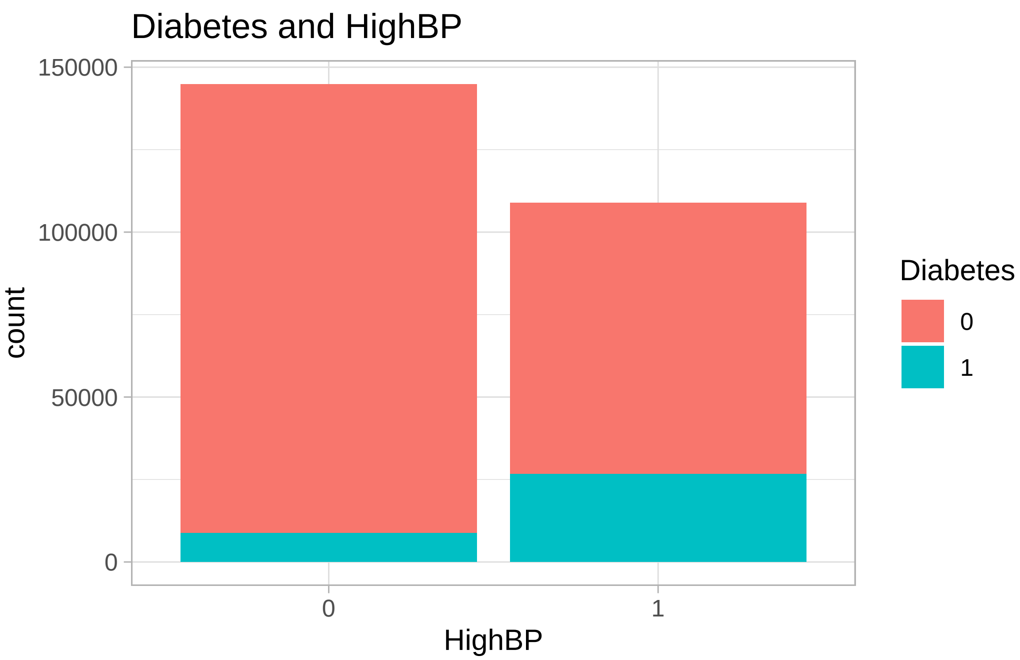
```
nrow(df)
```

```
[1] 253680
```

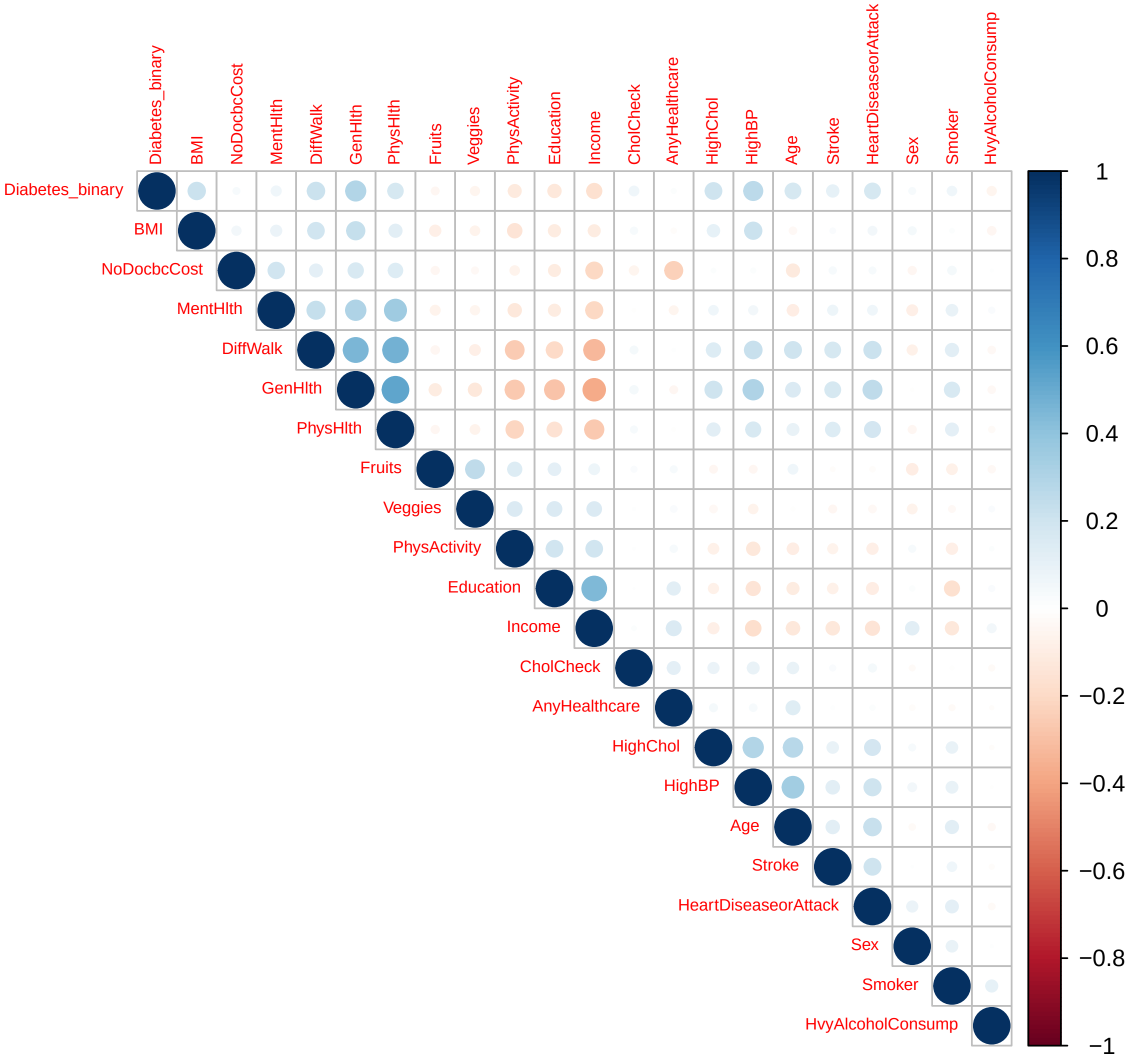
```
ggplot(df, aes(x=Diabetes_binary)) +
  geom_bar() +
  theme_light() +
  labs(title="Diabetes")
```



```
ggplot(df, aes(x=as.factor(HighBP), fill=as.factor(Diabetes_binary), group=Diabetes_binary
  geom_bar() +
  theme_light() +
  labs(title="Diabetes and HighBP", x="HighBP",
       fill="Diabetes")
```



```
cor_mtx <- cor(df)
corrplot(cor_mtx, type="upper", order = "hclust", tl.cex = 0.6, )
```



It seems that BMI, DiffWalk, GenHealth, PhysHealth, HighCol, HighBP, Age, and HeartDiseaseAttack are positively correlated with Diabetes_binary. DiffWalk, GenHealth, and PhysHealth are correlated with each other; hence we may not need all of them.

For negative correlations, PhysActivity, Education, and Income seem more significant than other variables.