Code

Emily Mittleman & Julia Rosner

2022-12-05

This file will be used for our initial code while we explore the data, different models, etc. Then we'll compile it into Report.Rmd

Load Data

```
data <- read.csv("Data/diabetes_012.csv", header = TRUE)
```

EDA

```
head(data)
```

```
##   Diabetes_012 HighBP HighChol CholCheck BMI Smoker Stroke HeartDiseaseorAttack
## 1            0      1        1         1  40      1      0                    0
## 2            0      0        0         0  25      1      0                    0
## 3            0      1        1         1  28      0      0                    0
## 4            0      1        0         1  27      0      0                    0
## 5            0      1        1         1  24      0      0                    0
## 6            0      1        1         1  25      1      0                    0
##   PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost
## 1            0      0       1                 0             1           0
## 2            1      0       0                 0             0           1
## 3            0      1       0                 0             1           1
## 4            1      1       1                 0             1           0
## 5            1      1       1                 0             1           0
## 6            1      1       1                 0             1           0
##   GenHlth MentHlth PhysHlth DiffWalk Sex Age Education Income
## 1       5       18       15        1   0   9         4      3
## 2       3        0        0        0   0   7         6      1
## 3       5       30       30        1   0   9         4      8
## 4       2        0        0        0   0  11         3      6
## 5       2        3        0        0   0  11         5      4
## 6       2        0        2        0   1  10         6      8
```

Look at correlations between variables. helps to know which attributes are highy dependent on the prediction variable

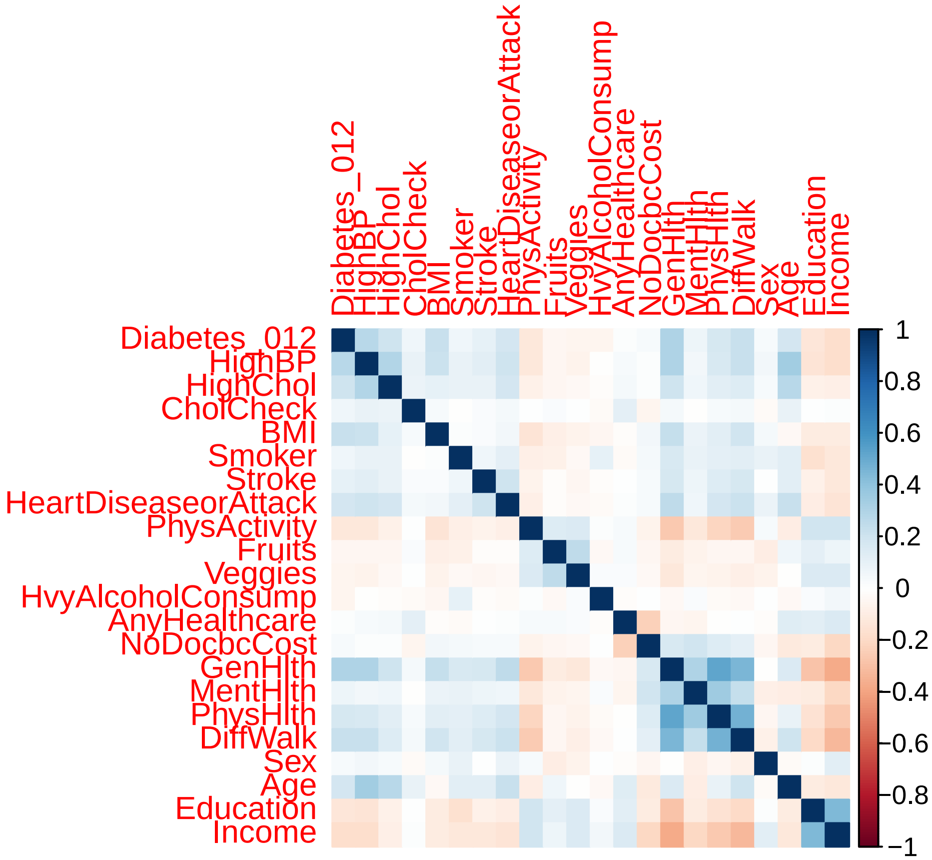
```r
correlations <- cor(data)
correlations
```

```
##                      Diabetes_012       HighBP    HighChol    CholCheck
## Diabetes_012           1.00000000  0.271596424  0.20908491  0.067546476
## HighBP                 0.27159642  1.000000000  0.29819930  0.098508273
## HighChol               0.20908491  0.298199295  1.00000000  0.085642228
## CholCheck              0.06754648  0.098508273  0.08564223  1.000000000
## BMI                    0.22437947  0.213748120  0.10672208  0.034495087
## Smoker                 0.06291410  0.096991467  0.09129936 -0.009928878
## Stroke                 0.10717867  0.129574913  0.09262007  0.024157667
## HeartDiseaseorAttack   0.18027169  0.209361211  0.18076535  0.044205810
## PhysActivity          -0.12194717 -0.125266866 -0.07804619  0.004189617
## Fruits                -0.04219163 -0.040554659 -0.04085908  0.023849406
## Veggies               -0.05897160 -0.061266165 -0.03987361  0.006121010
## HvyAlcoholConsump     -0.05788191 -0.003971574 -0.01154252 -0.023730091
## AnyHealthcare          0.01541038  0.038424769  0.04222986  0.117625625
## NoDocbcCost            0.03543569  0.017357984  0.01331016 -0.058255084
## GenHlth                0.30258662  0.300529631  0.20842555  0.046588865
## MentHlth               0.07350677  0.056455917  0.06206915 -0.008365598
## PhysHlth               0.17628674  0.161211571  0.12175053  0.031774808
## DiffWalk               0.22423912  0.223618466  0.14467154  0.040585057
## Sex                    0.03104016  0.052206961  0.03120533 -0.022115036
## Age                    0.18502579  0.344452330  0.27231823  0.090321114
## Education             -0.13051692 -0.141357934 -0.07080189  0.001510491
## Income                -0.17148304 -0.171234581 -0.08545931  0.014258747
##                              BMI       Smoker       Stroke HeartDiseaseorAttack
## Diabetes_012          0.22437947  0.062914095  0.107178670           0.18027169
## HighBP                0.21374812  0.096991467  0.129574913           0.20936121
## HighChol              0.10672208  0.091299357  0.092620074           0.18076535
## CholCheck             0.03449509 -0.009928878  0.024157667           0.04420581
## BMI                   1.00000000  0.013804467  0.020152661           0.05290426
## Smoker                0.01380447  1.000000000  0.061172675           0.11444122
## Stroke                0.02015266  0.061172675  1.000000000           0.20300194
## HeartDiseaseorAttack  0.05290426  0.114441218  0.203001940           1.00000000
## PhysActivity         -0.14729363 -0.087401163 -0.069151416          -0.08729899
## Fruits               -0.08751812 -0.077665839 -0.013389353          -0.01979035
## Veggies              -0.06227519 -0.030677710 -0.041124225          -0.03916741
## HvyAlcoholConsump    -0.04873628  0.101618687 -0.016950330          -0.02899052
## AnyHealthcare        -0.01847079 -0.023250803  0.008775925           0.01873419
## NoDocbcCost           0.05820629  0.048945823  0.034804106           0.03099970
## GenHlth               0.23918537  0.163143067  0.177942260           0.25838341
## MentHlth              0.08531016  0.092196474  0.070171812           0.06462129
## PhysHlth              0.12114111  0.116459714  0.148944169           0.18169754
## DiffWalk              0.19707776  0.122463215  0.176566917           0.21270870
## Sex                   0.04295030  0.093662361  0.002978288           0.08609551
## Age                  -0.03661764  0.120641084  0.126973699           0.22161763
## Education            -0.10393202 -0.161955255 -0.076008557          -0.09959992
## Income               -0.10006871 -0.123937229 -0.128598578          -0.14101123
##                      PhysActivity      Fruits      Veggies HvyAlcoholConsump
## Diabetes_012         -0.121947167 -0.04219163 -0.058971599      -0.057881912
## HighBP               -0.125266866 -0.04055466 -0.061266165      -0.003971574
## HighChol             -0.078046186 -0.04085908 -0.039873607      -0.011542519
```

```
## CholCheck             0.004189617  0.02384941  0.006121010      -0.023730091
## BMI                  -0.147293634 -0.08751812 -0.062275194      -0.048736275
## Smoker               -0.087401163 -0.07766584 -0.030677710       0.101618687
## Stroke               -0.069151416 -0.01338935 -0.041124225      -0.016950330
## HeartDiseaseorAttack -0.087298987 -0.01979035 -0.039167409      -0.028990516
## PhysActivity          1.000000000  0.14275586  0.153149570       0.012392236
## Fruits                0.142755863  1.00000000  0.254342244      -0.035287733
## Veggies               0.153149570  0.25434224  1.000000000       0.021064481
## HvyAlcoholConsump     0.012392236 -0.03528773  0.021064481       1.000000000
## AnyHealthcare         0.035504737  0.03154392  0.029583817      -0.010488085
## NoDocbcCost          -0.061638387 -0.04424269 -0.032231705       0.004683595
## GenHlth              -0.266185624 -0.10385417 -0.123066330      -0.036723570
## MentHlth             -0.125587088 -0.06821738 -0.058883553       0.024715803
## PhysHlth             -0.219229522 -0.04463332 -0.064290327      -0.026415474
## DiffWalk             -0.253174007 -0.04835167 -0.080505717      -0.037668174
## Sex                   0.032481686 -0.09117487 -0.064765156       0.005740219
## Age                  -0.092510633  0.06454722 -0.009771198      -0.034577637
## Education             0.199658057  0.11018710  0.154329262       0.023996867
## Income                0.198539455  0.07992931  0.151086944       0.053618566
##                      AnyHealthcare  NoDocbcCost      GenHlth     MentHlth
## Diabetes_012           0.015410377  0.035435685  0.302586621  0.073506766
## HighBP                 0.038424769  0.017357984  0.300529631  0.056455917
## HighChol               0.042229862  0.013310163  0.208425550  0.062069154
## CholCheck              0.117625625 -0.058255084  0.046588865 -0.008365598
## BMI                   -0.018470787  0.058206290  0.239185373  0.085310159
## Smoker                -0.023250803  0.048945823  0.163143067  0.092196474
## Stroke                 0.008775925  0.034804106  0.177942260  0.070171812
## HeartDiseaseorAttack   0.018734186  0.030999705  0.258383409  0.064621292
## PhysActivity           0.035504737 -0.061638387 -0.266185624 -0.125587088
## Fruits                 0.031543919 -0.044242689 -0.103854171 -0.068217375
## Veggies                0.029583817 -0.032231705 -0.123066330 -0.058883553
## HvyAlcoholConsump     -0.010488085  0.004683595 -0.036723570  0.024715803
## AnyHealthcare          1.000000000 -0.232532105 -0.040817072 -0.052706597
## NoDocbcCost           -0.232532105  1.000000000  0.166397186  0.192106853
## GenHlth               -0.040817072  0.166397186  1.000000000  0.301674393
## MentHlth              -0.052706597  0.192106853  0.301674393  1.000000000
## PhysHlth              -0.008276167  0.148997564  0.524363644  0.353618868
## DiffWalk               0.007074092  0.118446862  0.456919503  0.233688079
## Sex                   -0.019405465 -0.044931366 -0.006091004 -0.080704863
## Age                    0.138045679 -0.119777068  0.152449830 -0.092068024
## Education              0.122514239 -0.100701002 -0.284911532 -0.101829695
## Income                 0.157999279 -0.203182369 -0.370013734 -0.209806127
##                          PhysHlth     DiffWalk          Sex          Age
## Diabetes_012          0.176286736  0.224239123  0.031040164  0.185025794
## HighBP                0.161211571  0.223618466  0.052206961  0.344452330
## HighChol              0.121750528  0.144671538  0.031205330  0.272318226
## CholCheck             0.031774808  0.040585057 -0.022115036  0.090321114
## BMI                   0.121141107  0.197077760  0.042950303 -0.036617635
## Smoker                0.116459714  0.122463215  0.093662361  0.120641084
## Stroke                0.148944169  0.176566917  0.002978288  0.126973699
## HeartDiseaseorAttack  0.181697536  0.212708695  0.086095508  0.221617632
## PhysActivity         -0.219229522 -0.253174007  0.032481686 -0.092510633
## Fruits               -0.044633325 -0.048351675 -0.091174865  0.064547217
## Veggies              -0.064290327 -0.080505717 -0.064765156 -0.009771198
```

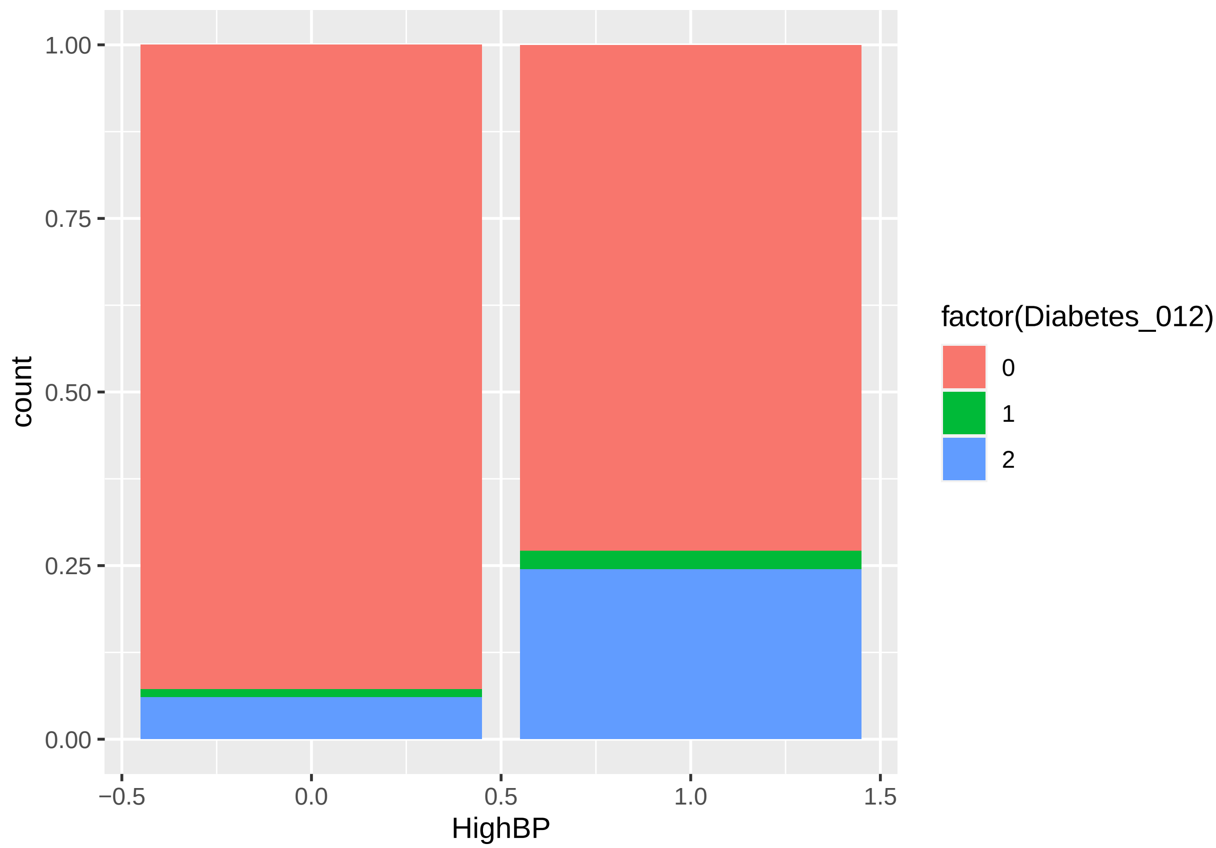
```
## HvyAlcoholConsump    -0.026415474 -0.037668174  0.005740219 -0.034577637
## AnyHealthcare        -0.008276167  0.007074092 -0.019405465  0.138045679
## NoDocbcCost           0.148997564  0.118446862 -0.044931366 -0.119777068
## GenHlth               0.524363644  0.456919503 -0.006091004  0.152449830
## MentHlth              0.353618868  0.233688079 -0.080704863 -0.092068024
## PhysHlth              1.000000000  0.478416619 -0.043136502  0.099129925
## DiffWalk              0.478416619  1.000000000 -0.070298902  0.204450090
## Sex                  -0.043136502 -0.070298902  1.000000000 -0.027340383
## Age                   0.099129925  0.204450090 -0.027340383  1.000000000
## Education            -0.155092517 -0.192642100  0.019479786 -0.101901070
## Income               -0.266798962 -0.320124244  0.127141058 -0.127775278
##                         Education       Income
## Diabetes_012         -0.130516918 -0.17148304
## HighBP               -0.141357934 -0.17123458
## HighChol             -0.070801887 -0.08545931
## CholCheck             0.001510491  0.01425875
## BMI                  -0.103932022 -0.10006871
## Smoker               -0.161955255 -0.12393723
## Stroke               -0.076008557 -0.12859858
## HeartDiseaseorAttack -0.099599915 -0.14101123
## PhysActivity          0.199658057  0.19853946
## Fruits                0.110187097  0.07992931
## Veggies               0.154329262  0.15108694
## HvyAlcoholConsump     0.023996867  0.05361857
## AnyHealthcare         0.122514239  0.15799928
## NoDocbcCost          -0.100701002 -0.20318237
## GenHlth              -0.284911532 -0.37001373
## MentHlth             -0.101829695 -0.20980613
## PhysHlth             -0.155092517 -0.26679896
## DiffWalk             -0.192642100 -0.32012424
## Sex                   0.019479786  0.12714106
## Age                  -0.101901070 -0.12777528
## Education             1.000000000  0.44910642
## Income                0.449106424  1.00000000
```

```
corrplot(correlations, method="color")
```

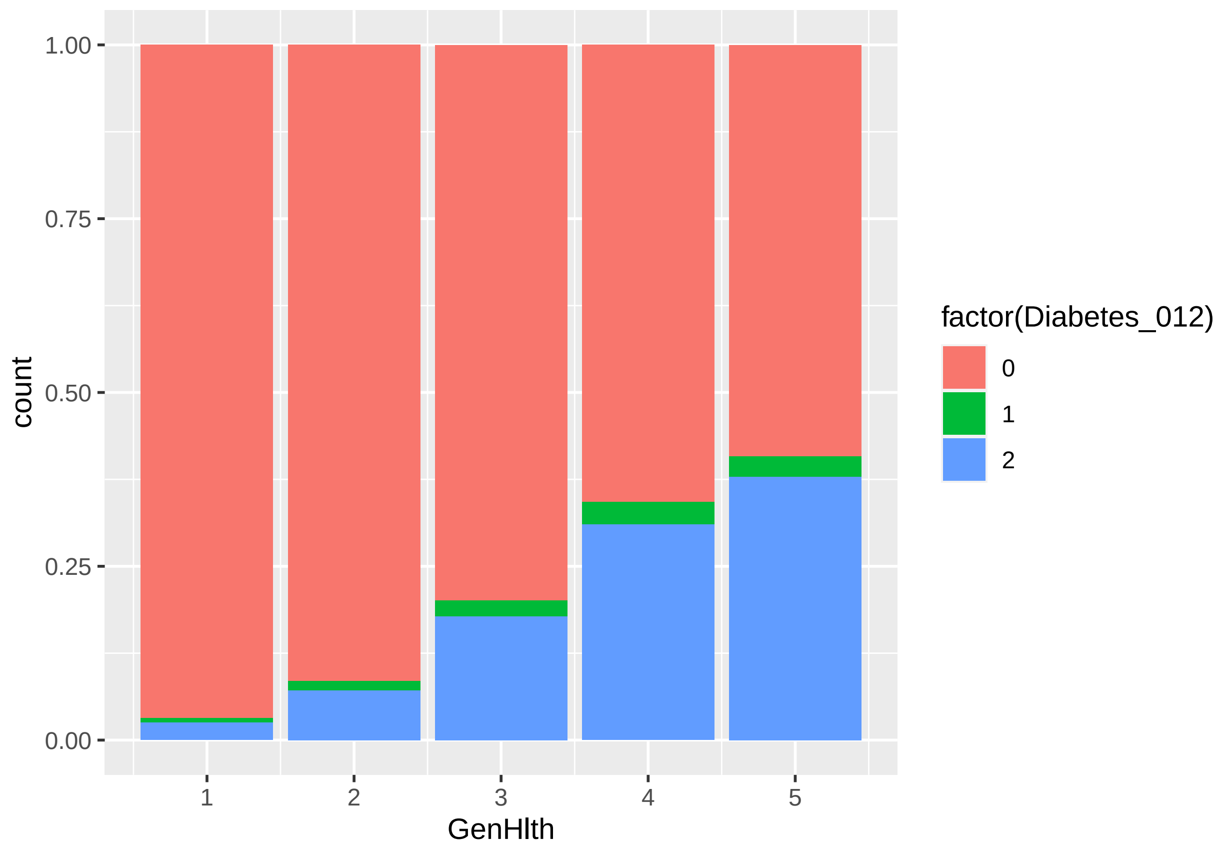


Next, look at box plots of the 2 most correlated predictors and color by outcome.

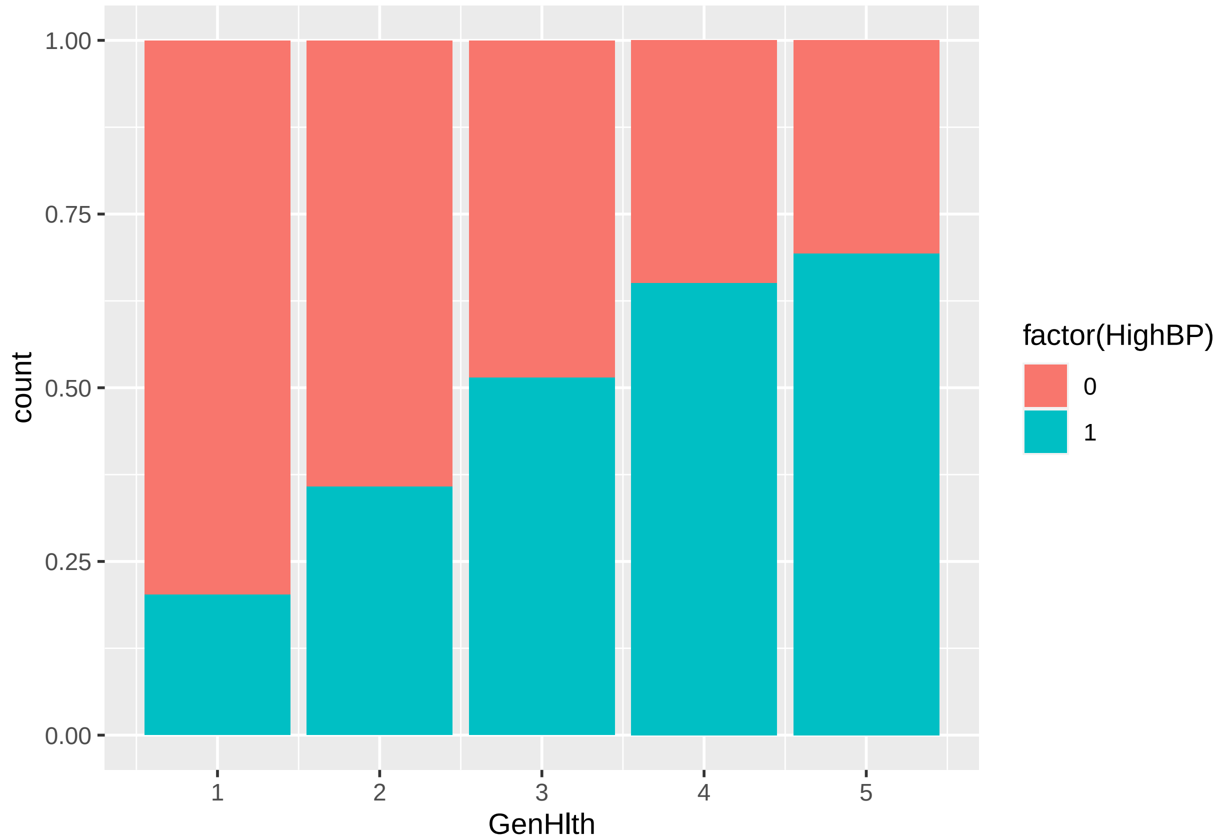
```
ggplot(data, aes(x = HighBP, fill = factor(Diabetes_012))) +
  geom_bar(position="fill")
```




```r
ggplot(data, aes(x = GenHlth, fill = factor(Diabetes_012))) +
  geom_bar(position="fill")
```




```r
ggplot(data, aes(x = GenHlth, fill = factor(HighBP))) +
  geom_bar(position="fill")
```



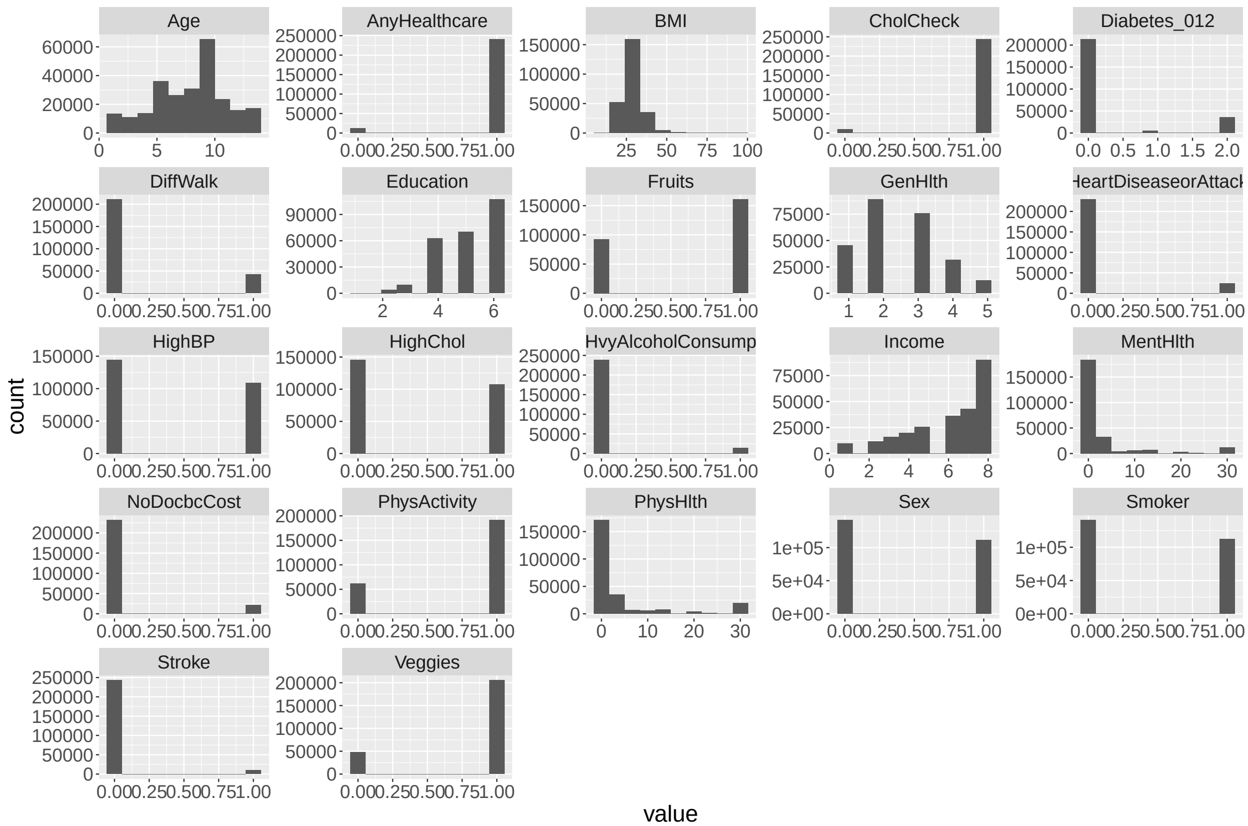
Make pivot table to make historgrams of each variable simpler

```
data_long <- data %>%                          # Apply pivot_longer function
  pivot_longer(colnames(data)) %>%
  as.data.frame()
head(data_long)
```

```
##           name value
## 1 Diabetes_012     0
## 2       HighBP     1
## 3     HighChol     1
## 4    CholCheck     1
## 5          BMI    40
## 6       Smoker     1
```

Visualize predictor variable distrbutions:

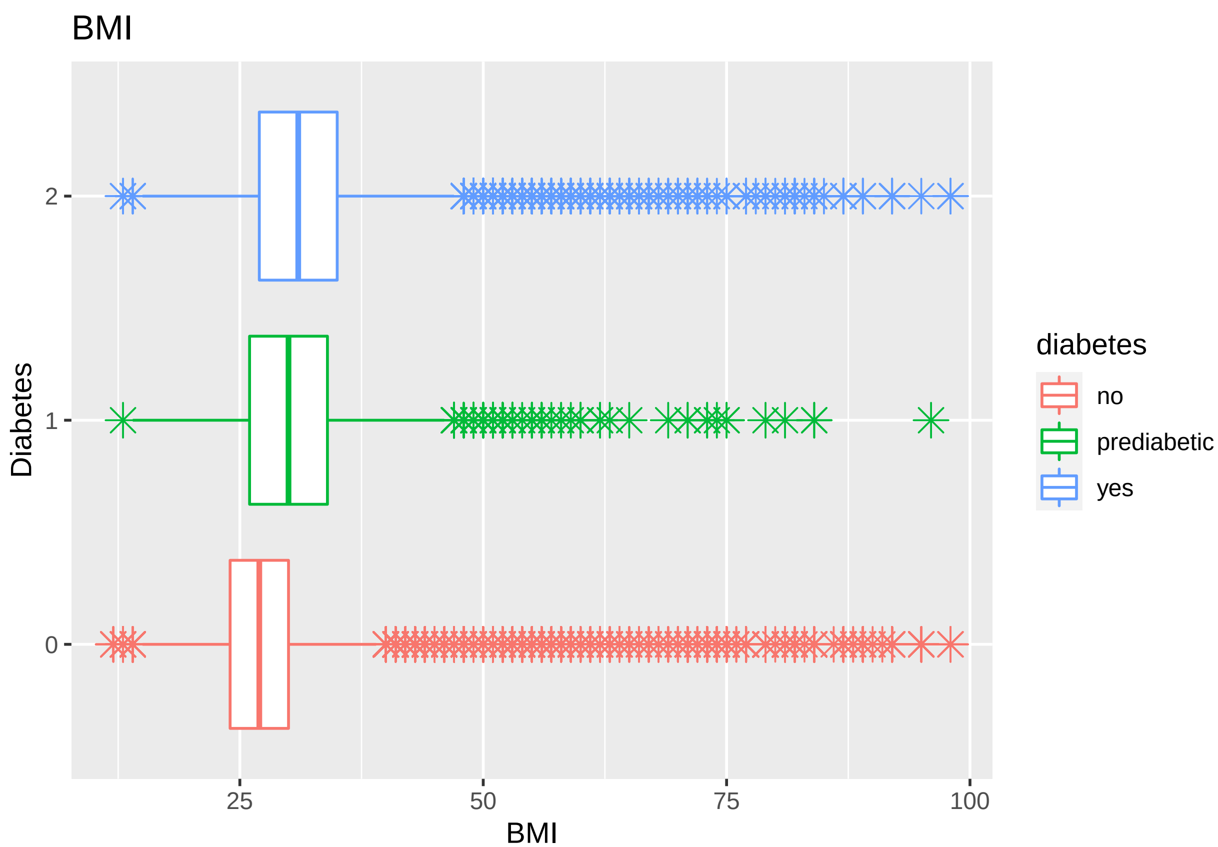
```
ggp1 <- ggplot(data_long, aes(x = value)) +    # Draw each column as histogram
  geom_histogram(bins=10) +
  facet_wrap(~ name, scales = "free")+
  theme(text=element_text(size=20))
ggp1
```



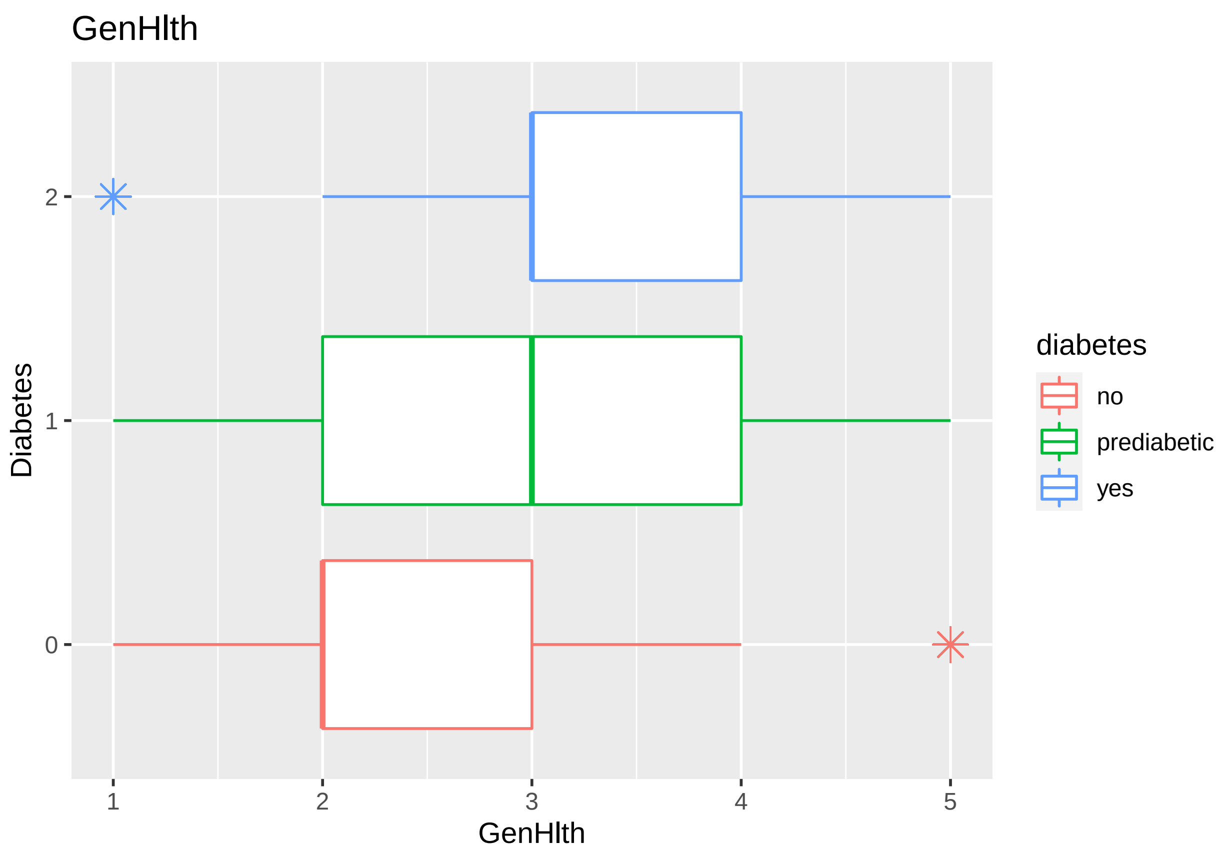
Next, look for outliers in predictors:

```
p1 <- ggplot(data, aes(x = BMI, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="BMI", y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p2 <- ggplot(data, aes(x = GenHlth, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="GenHlth",y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p3 <- ggplot(data, aes(x = MentHlth, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="MentHlth", y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p4 <- ggplot(data, aes(x = PhysHlth, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="PhysHlth",y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p5 <- ggplot(data, aes(x = Age, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="Age",y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p6 <- ggplot(data, aes(x = Education, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="Education", y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p7 <- ggplot(data, aes(x = Income, y=factor(Diabetes_012), color=factor(Diabetes_012))) +
```

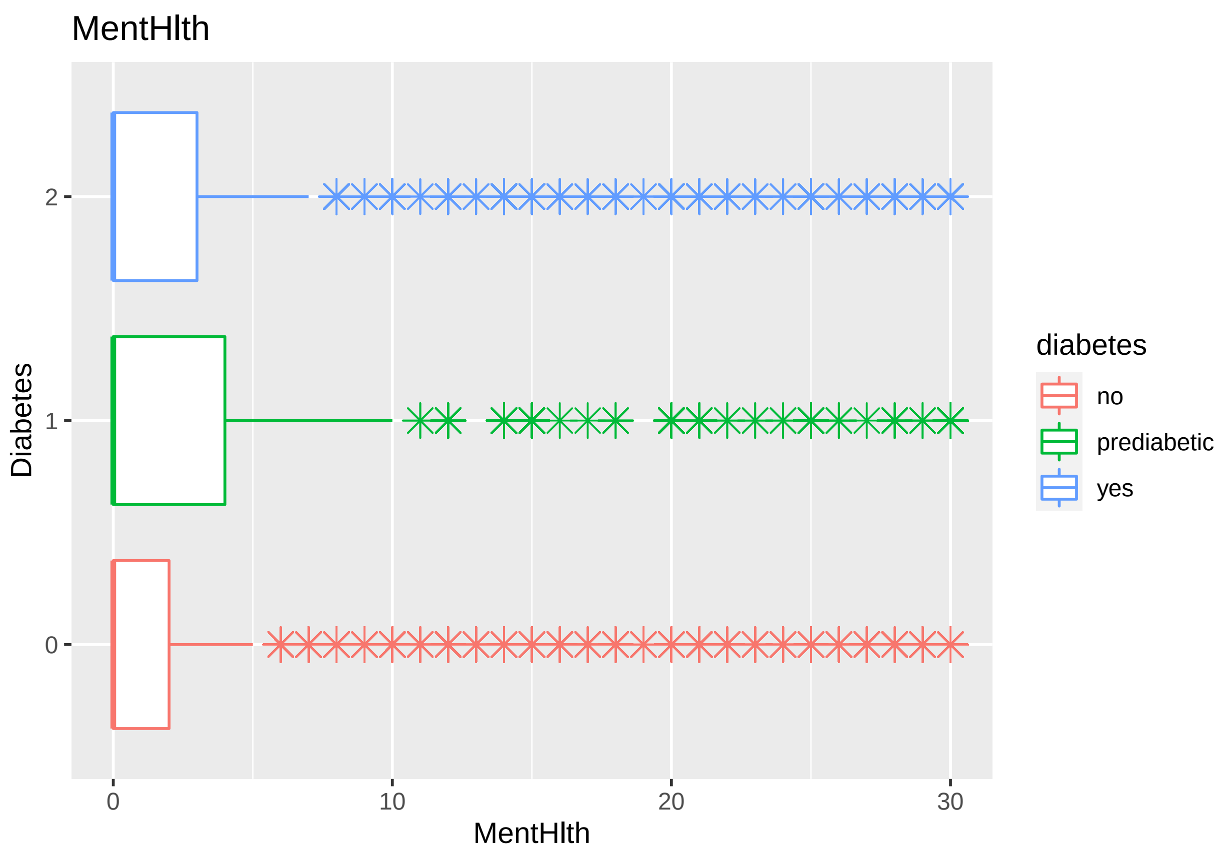
```
  geom_boxplot(outlier.shape=8, outlier.size=4)+
  labs(title="Income", y="Diabetes")+
  scale_color_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
p1
```



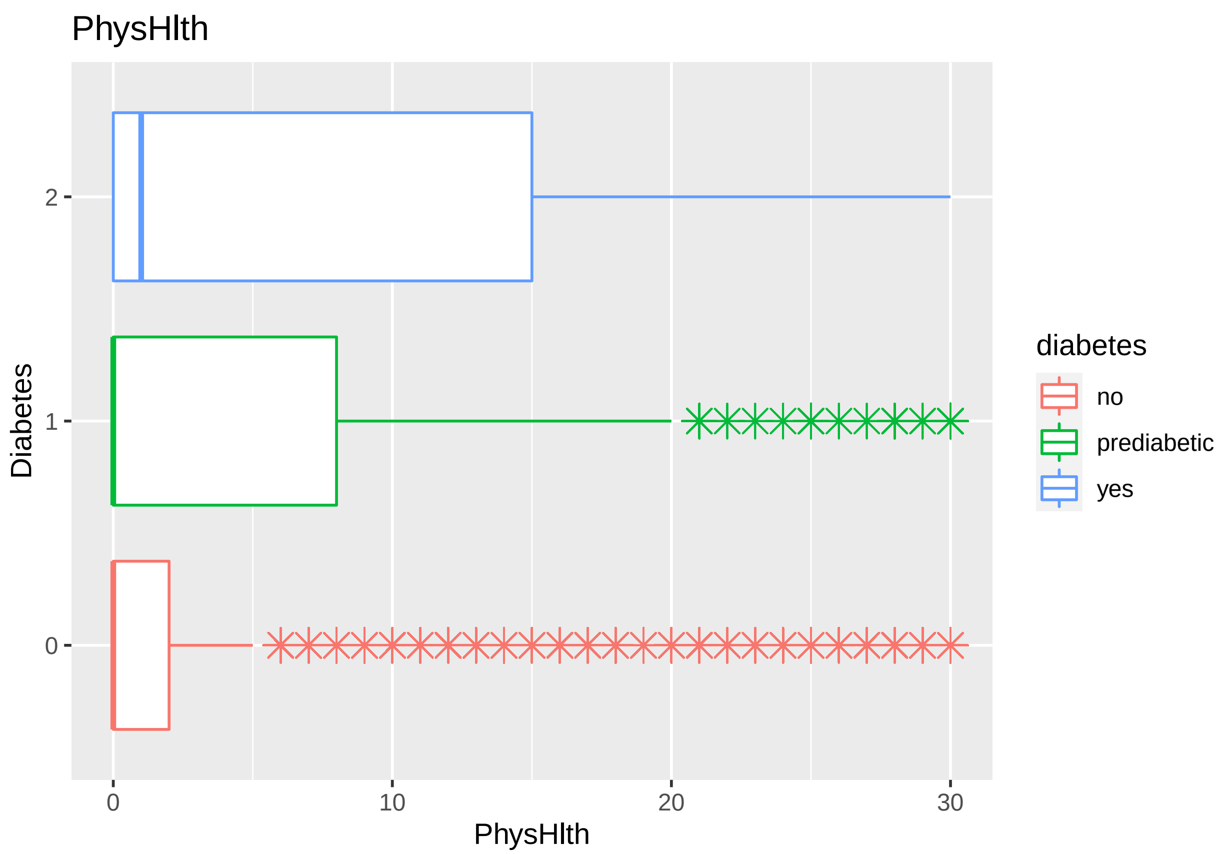

```
p2
```


GenHlth
Diabetes
2
1
0
1
2
3
4
5
GenHlth
diabetes
no
prediabetic
yes

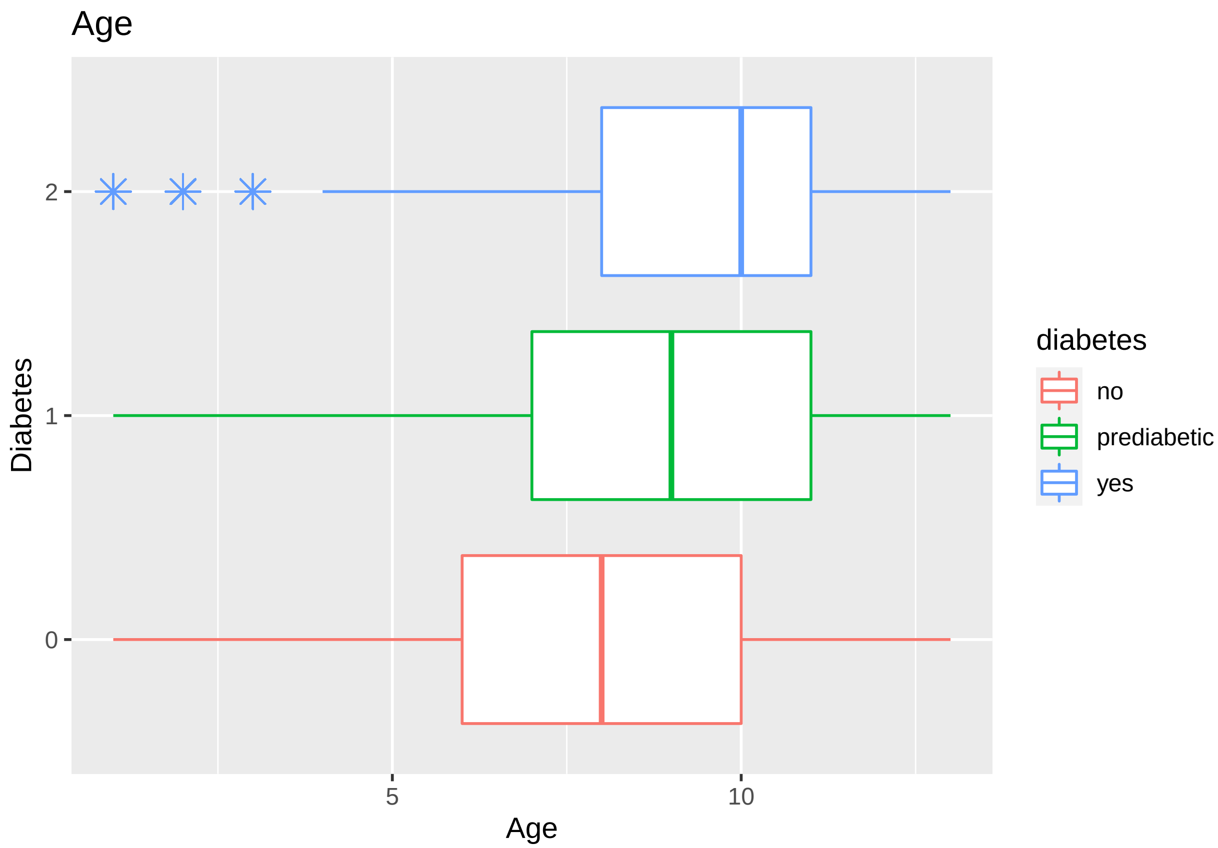
```
p3
```


MentHlth
Diabetes
2
1
0
0
10
20
30
MentHlth
diabetes
no
prediabetic
yes

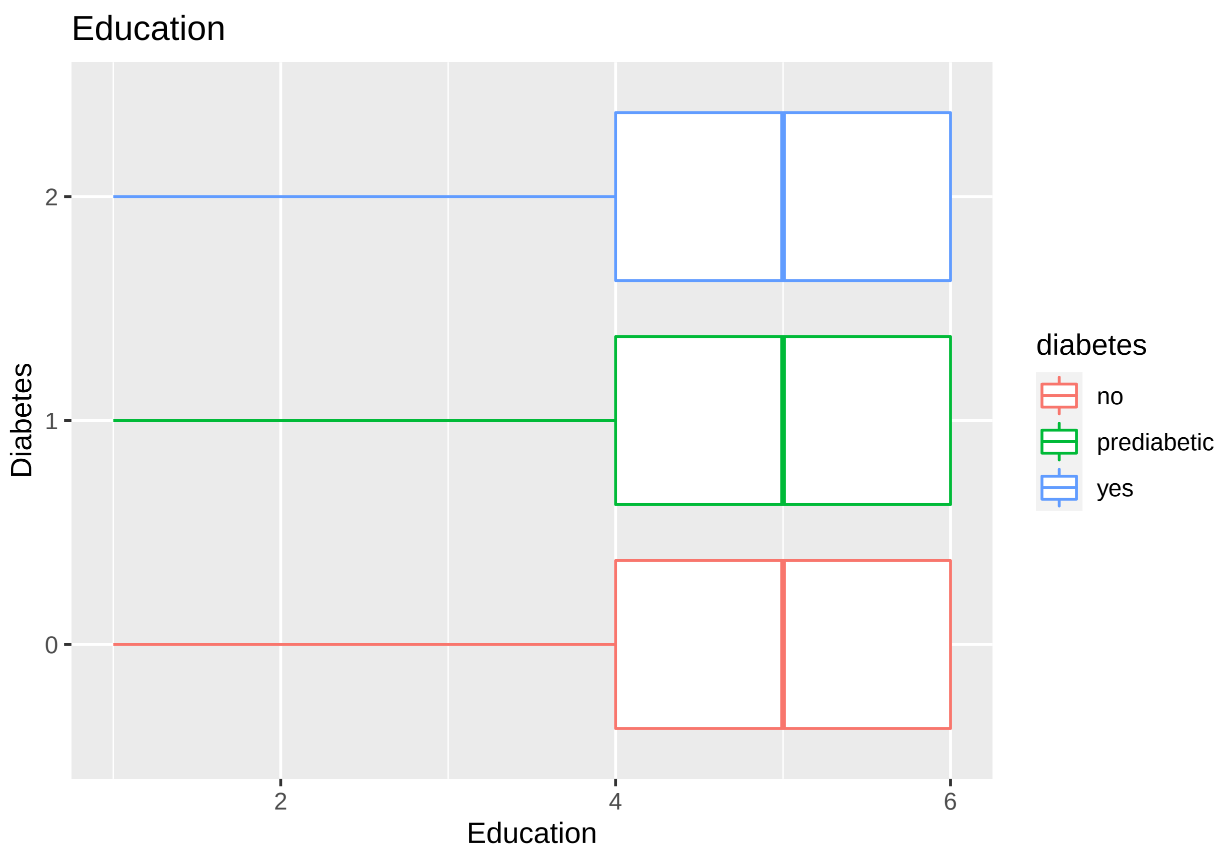
```
p4
```


PhysHlth
Diabetes
2
1
0
0
10
20
30
PhysHlth
diabetes
no
prediabetic
yes

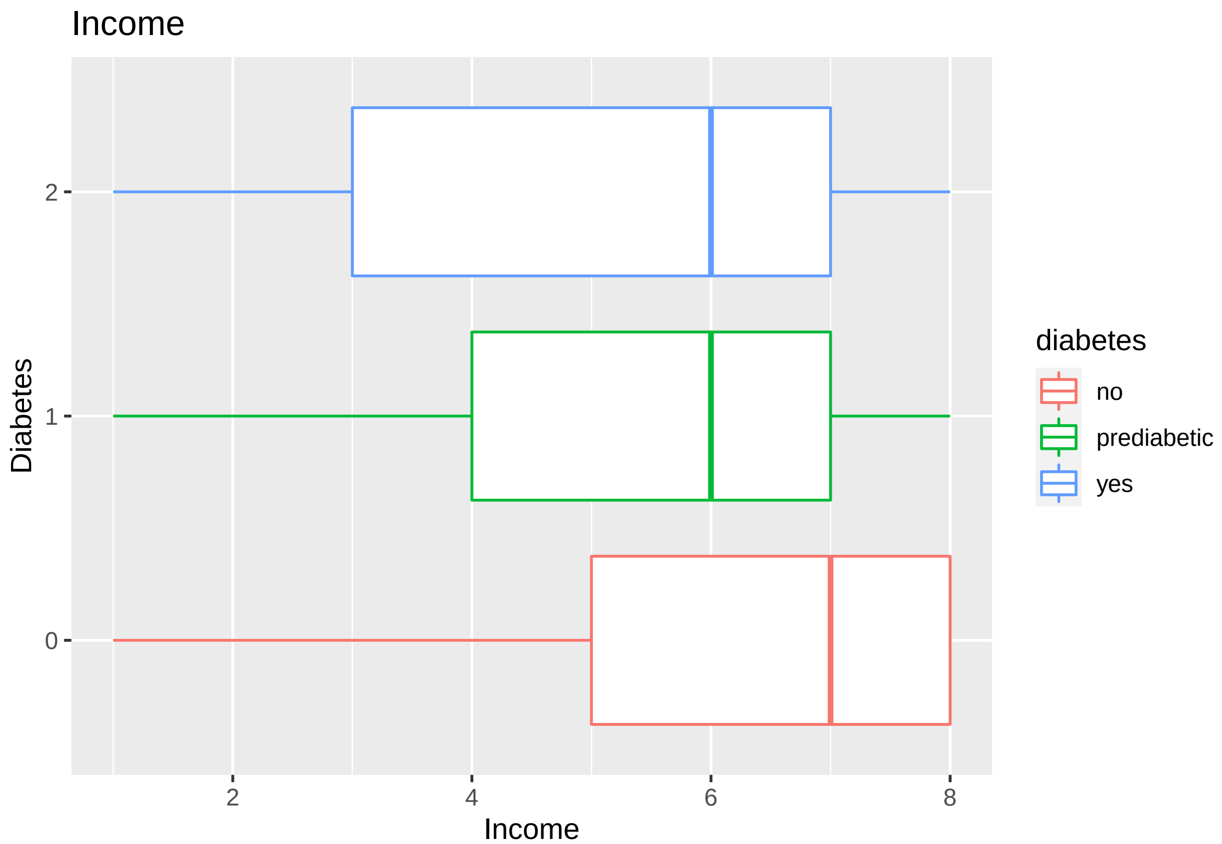
```
p5
```




```
p6
```


Education
2
1
0
Diabetes
2
4
6
Education
diabetes
no
prediabetic
yes

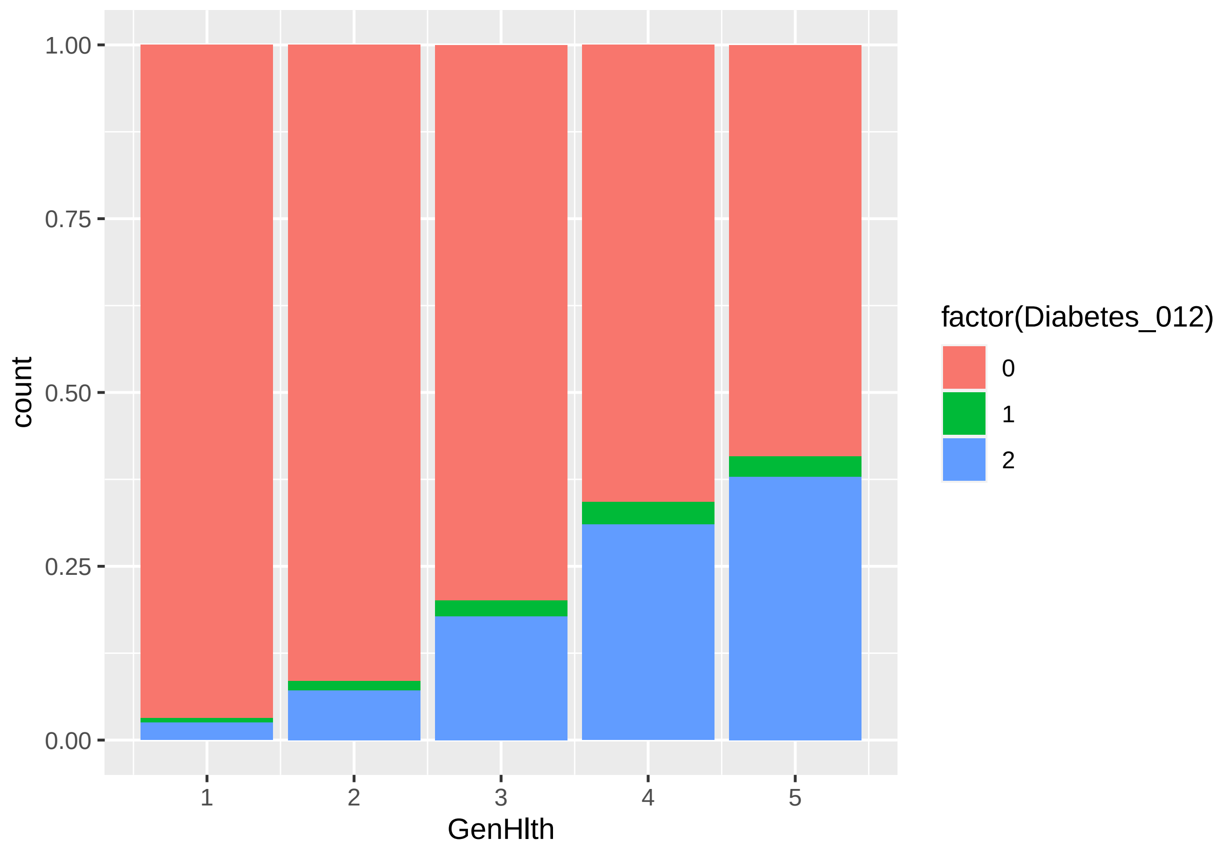
```
p7
```



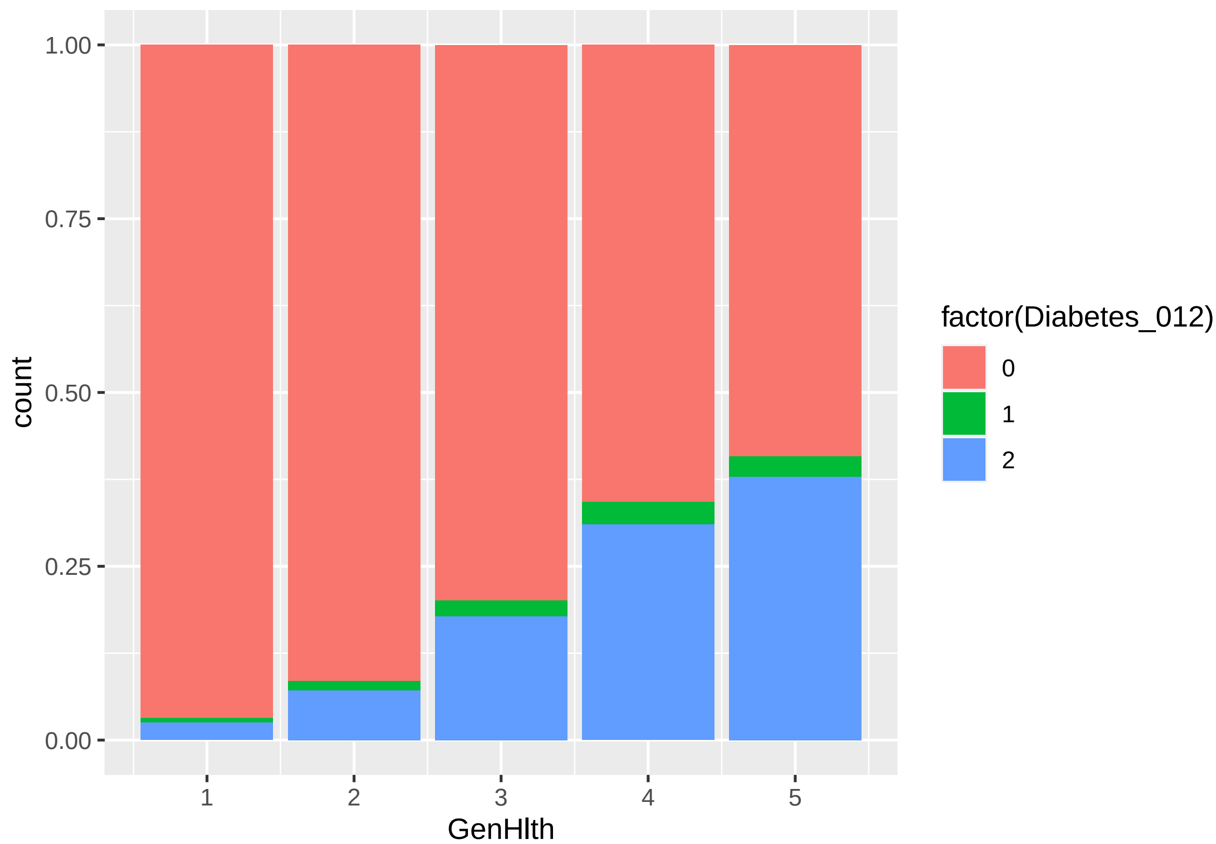
From the boxplots above, we see that the predictors BMI, MntHlth, and PhysHlth have a lot of outliers. All three distributions aee very skewed to the right. GenHlth and Age have only a couple outliers. Education and Income have none.

Now, we visulaize predictor distributions and relation to response.

```
ggplot(data, aes(x = GenHlth, fill = factor(Diabetes_012))) +
  geom_bar(position="fill")
```

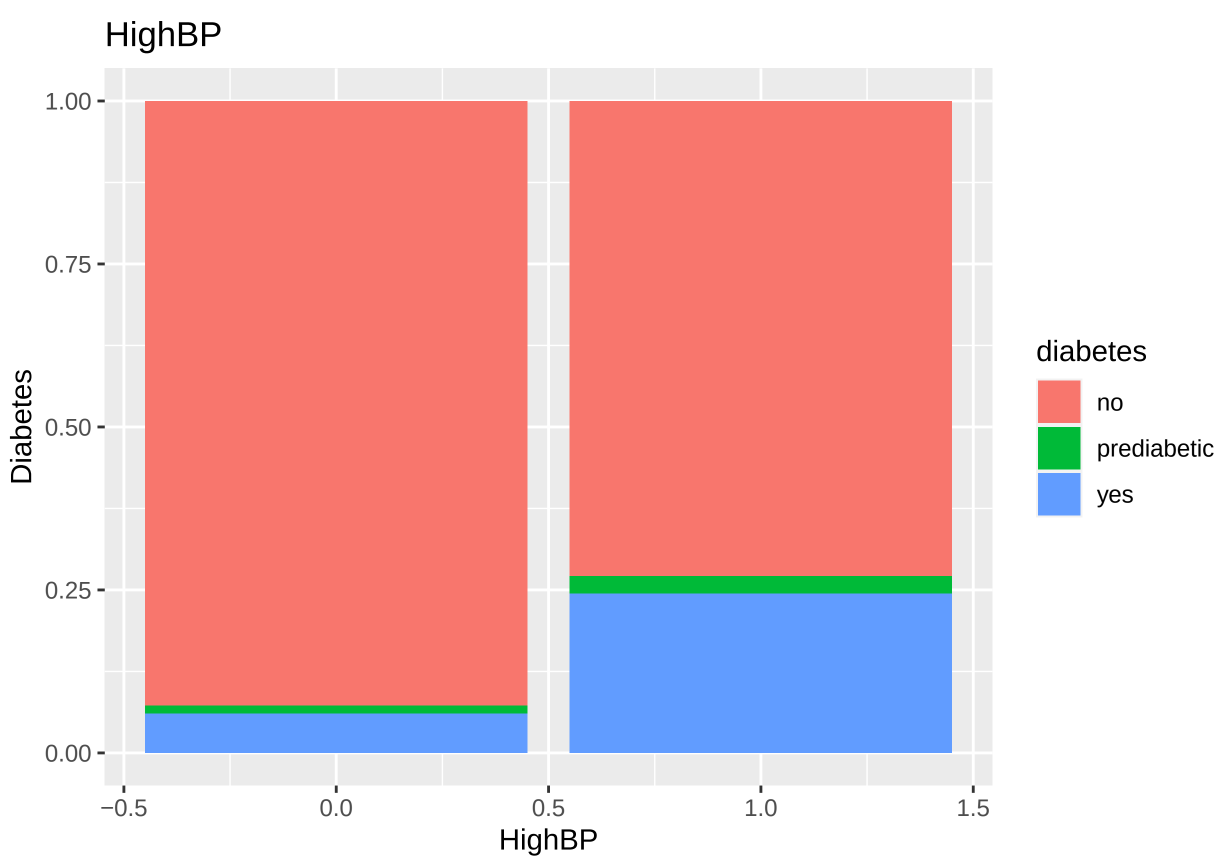



```
ggplot(data, aes(x = GenHlth, fill = factor(Diabetes_012))) +
  geom_bar(position="fill")
```

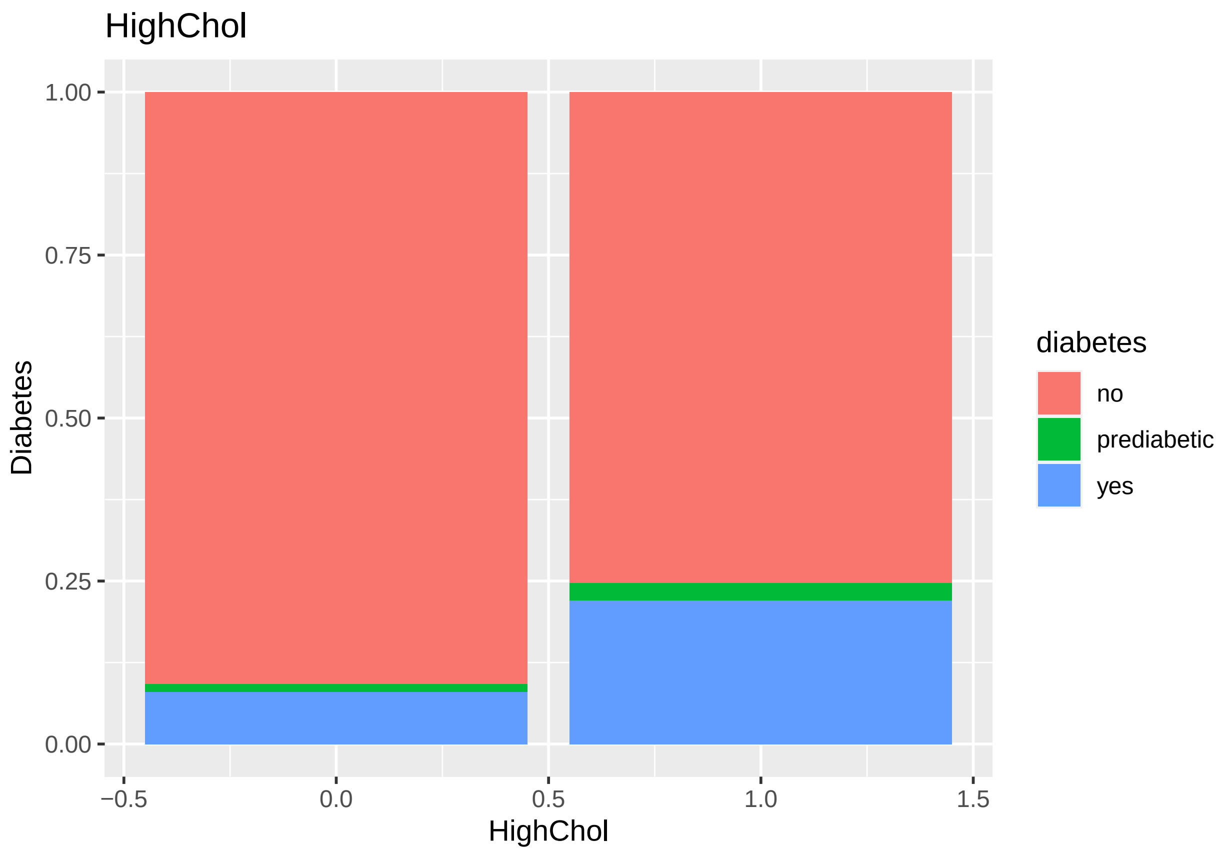



```
pbox1 <- ggplot(data, aes(x = HighBP, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="HighBP", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox2 <- ggplot(data, aes(x = HighChol, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="HighChol", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox3 <- ggplot(data, aes(x = CholCheck, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="CholCheck", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox4 <- ggplot(data, aes(x = Smoker, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="Smoker", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox5 <- ggplot(data, aes(x = Stroke, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="Stroke", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox6 <- ggplot(data, aes(x = HeartDiseaseorAttack, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="HeartDiseaseorAttack", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox7 <- ggplot(data, aes(x = PhysActivity, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
```

```r
  labs(title="PhysActivity", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox8 <- ggplot(data, aes(x = Veggies, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="Veggies", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox9 <- ggplot(data, aes(x = HvyAlcoholConsump, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="HvyAlcoholConsump", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox10 <- ggplot(data, aes(x = AnyHealthcare, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="AnyHealthcare", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox11 <- ggplot(data, aes(x = NoDocbcCost, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="NoDocbcCost", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox12 <- ggplot(data, aes(x = DiffWalk, fill=factor(Diabetes_012))) +
  geom_bar(position="fill")+
  labs(title="DiffWalk", y="Diabetes")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
pbox1
```

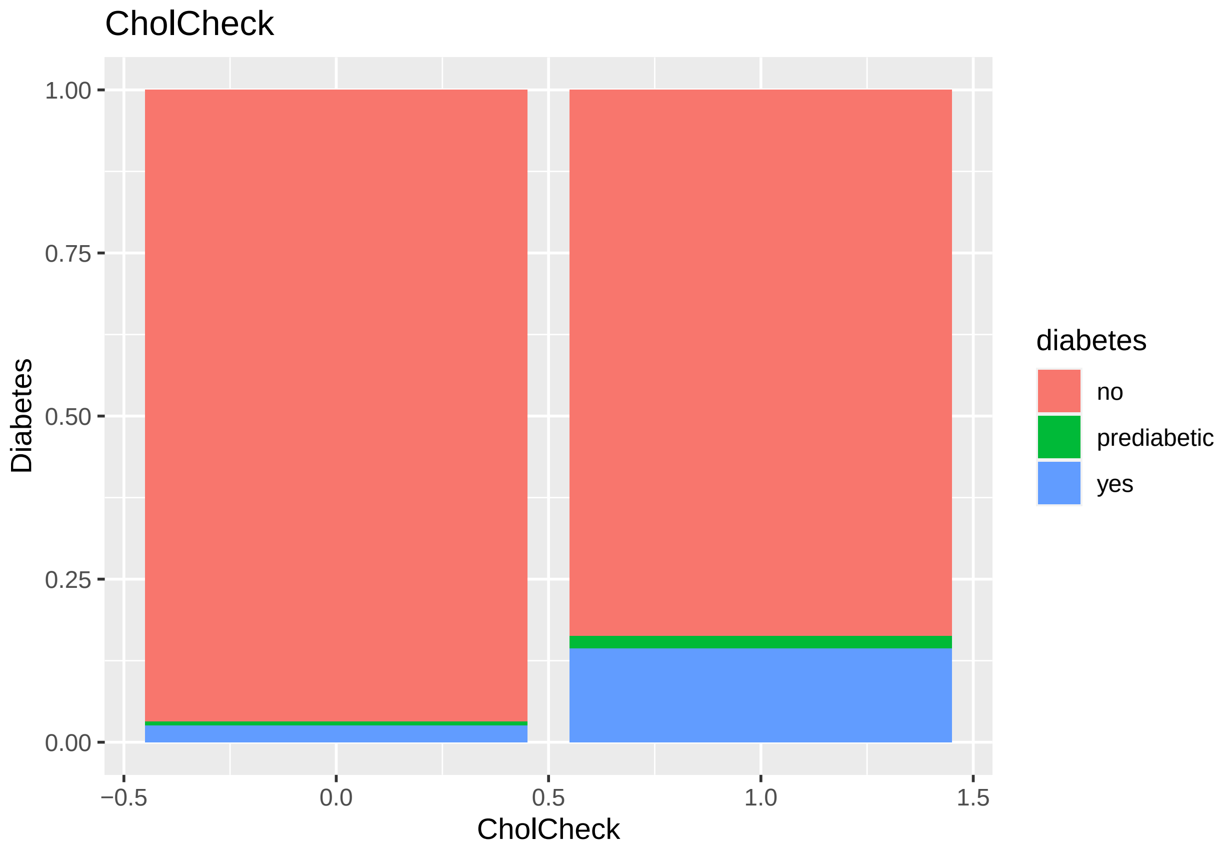


```
pbox2
```

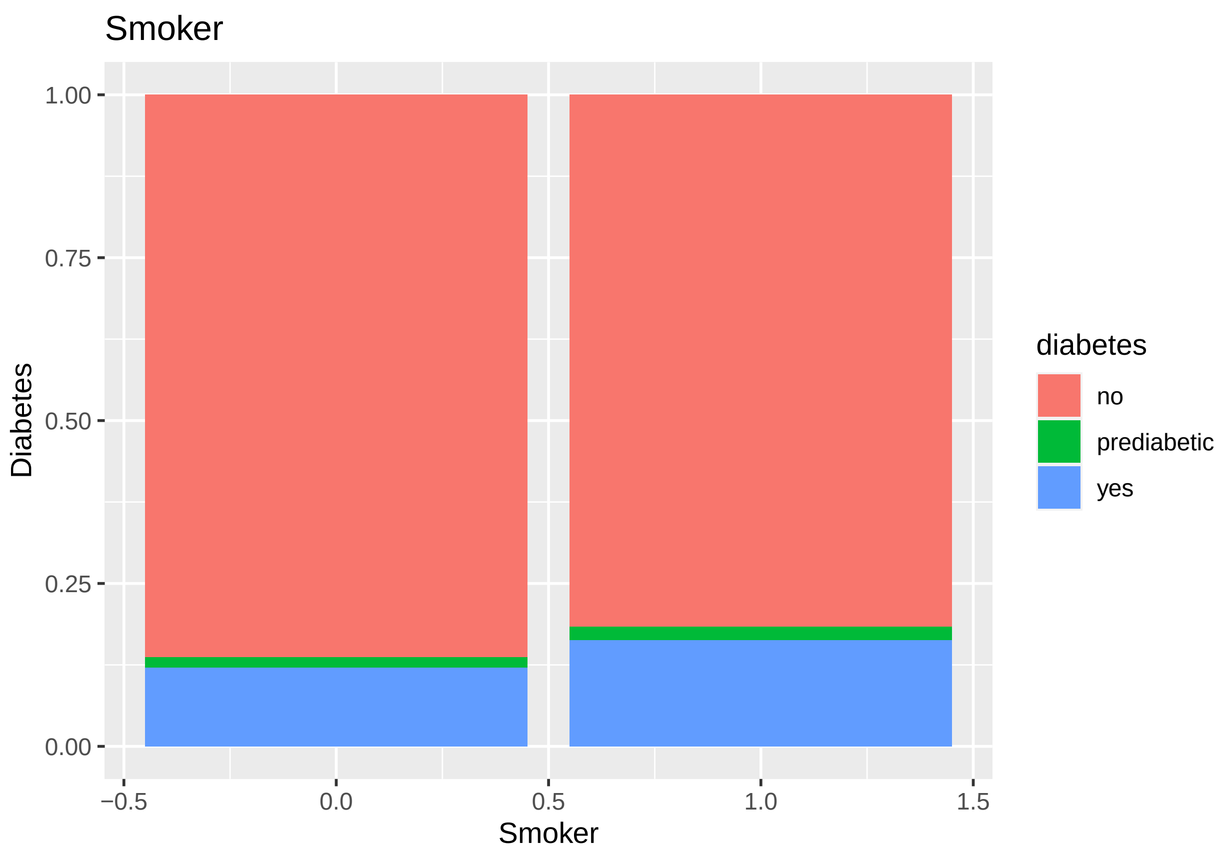


HighChol
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
HighChol
diabetes
no
prediabetic
yes

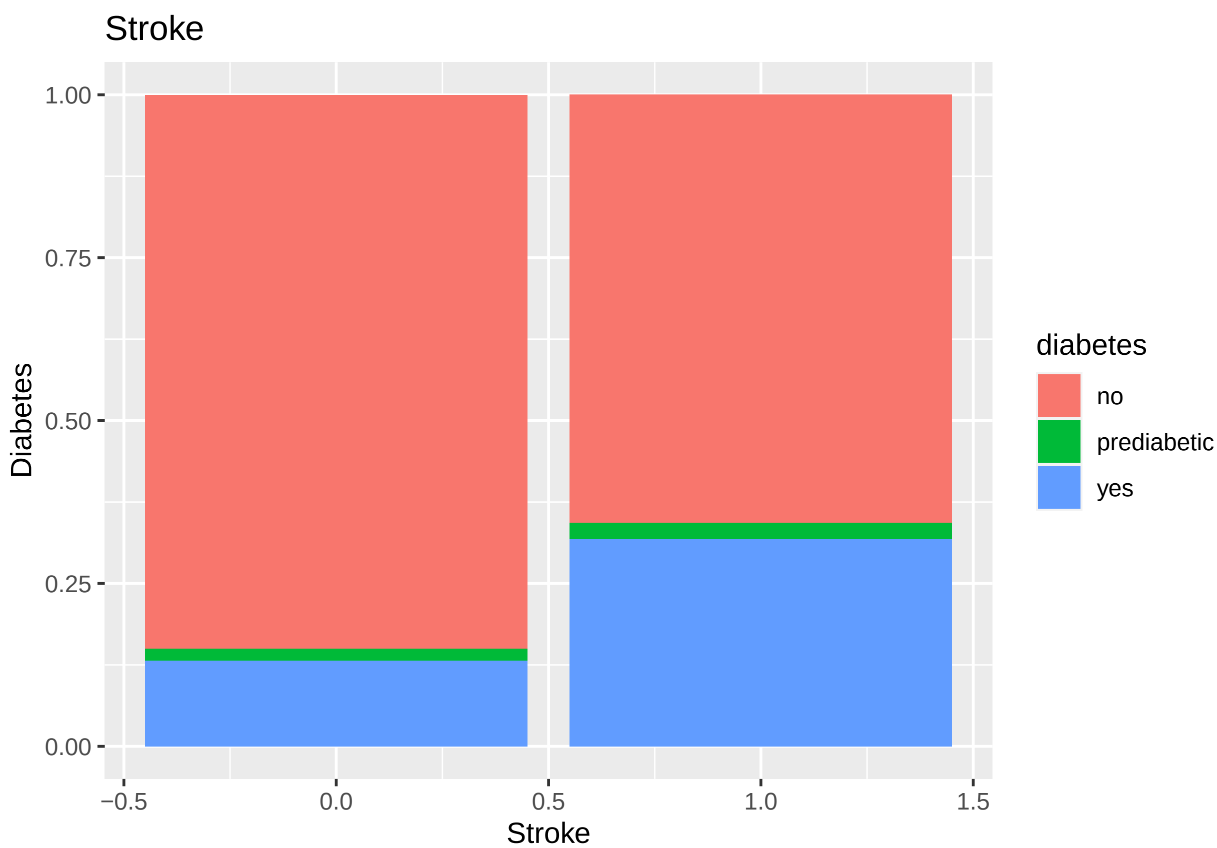

```
pbox3
```


CholCheck
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
CholCheck
diabetes
no
prediabetic
yes

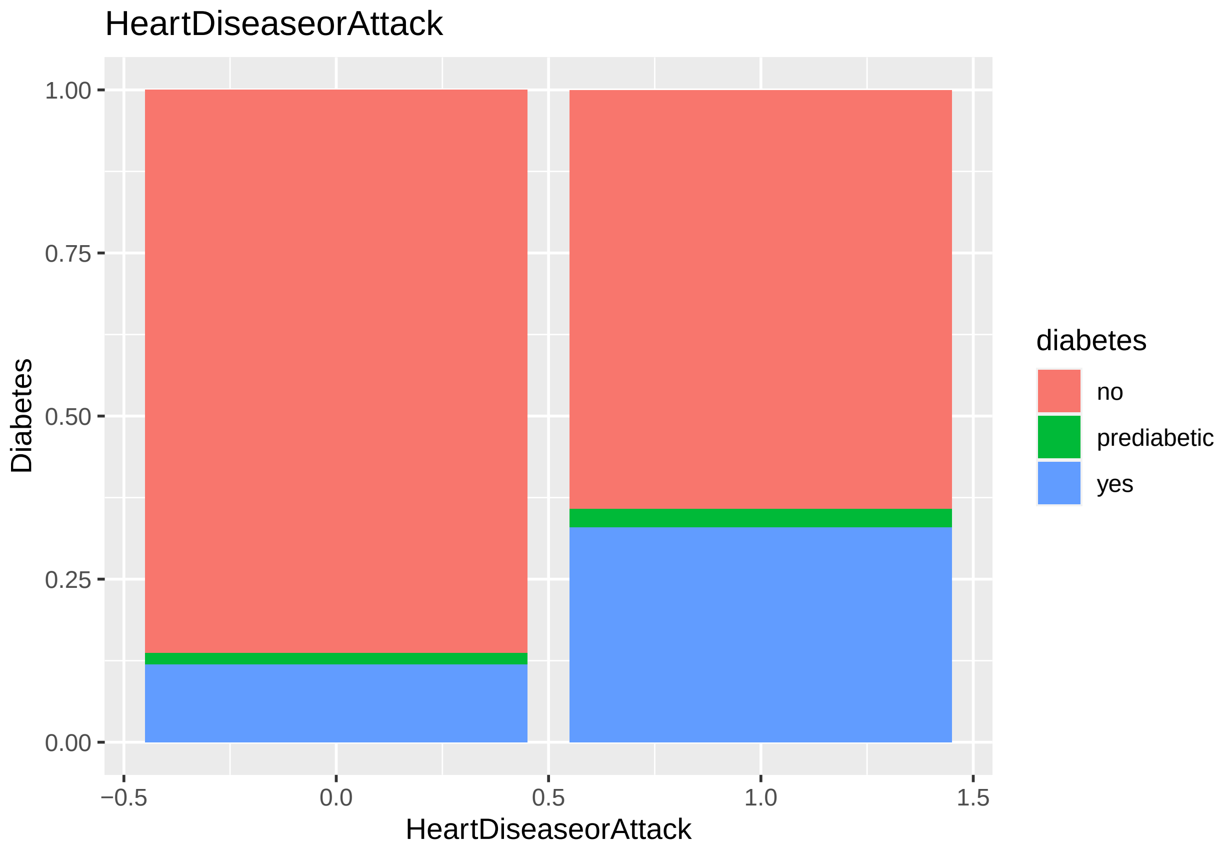
```
pbox4
```


Smoker
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
Smoker
diabetes
no
prediabetic
yes

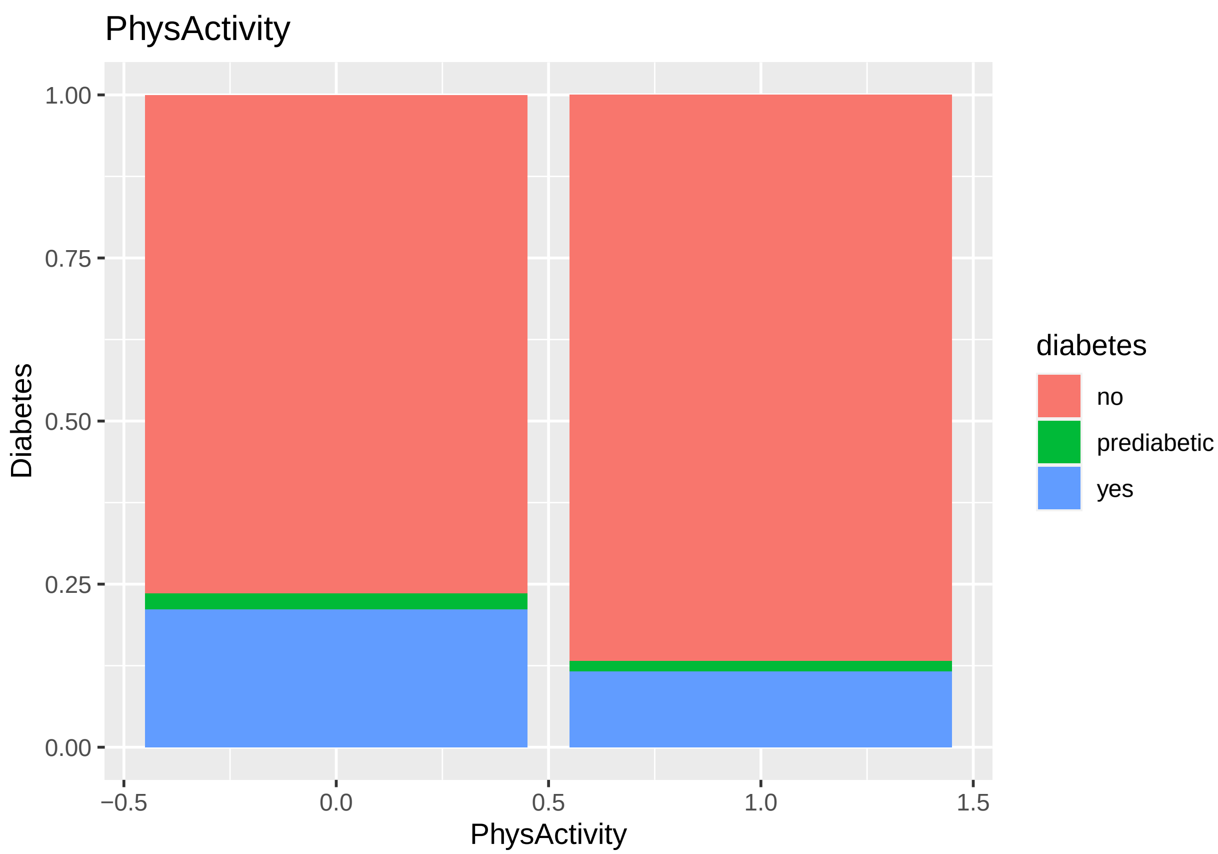
```
pbox5
```


Stroke
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
Stroke
diabetes
no
prediabetic
yes

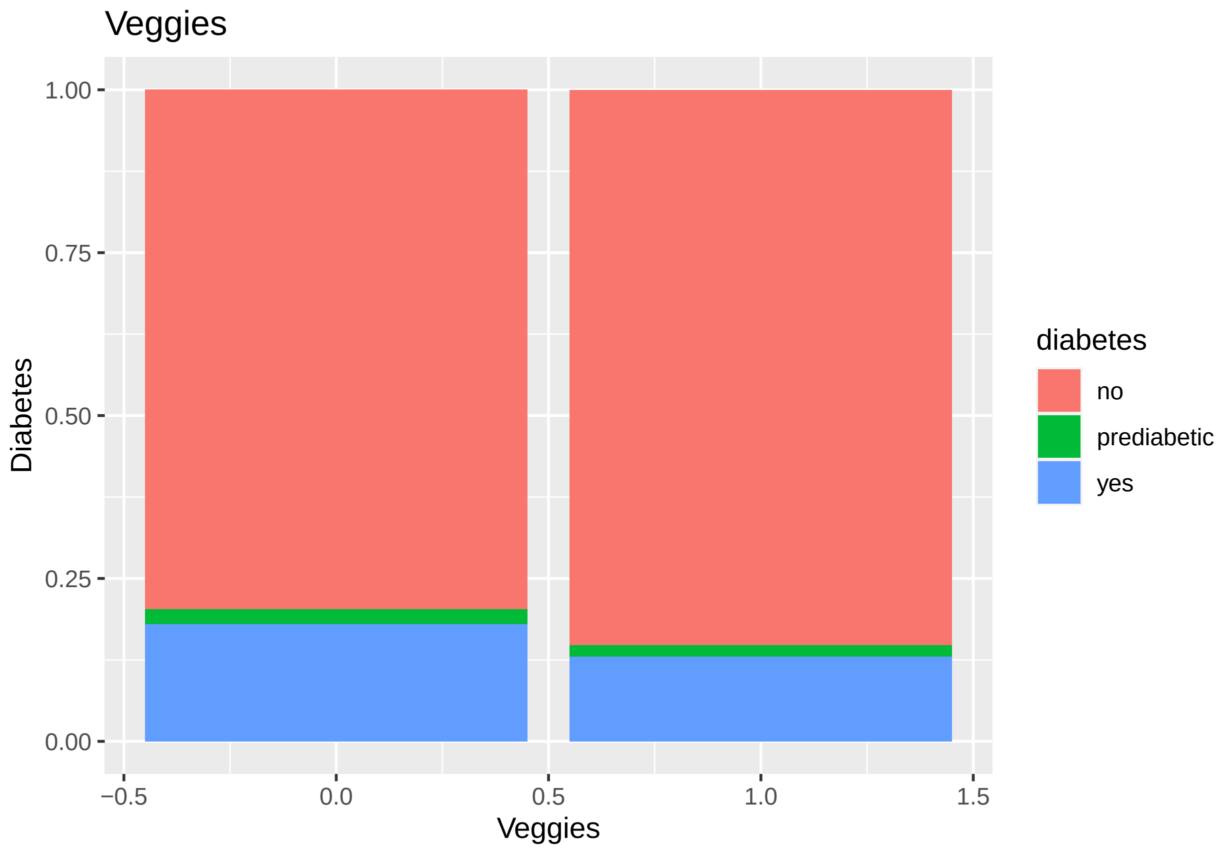
```
pbox6
```


HeartDiseaseorAttack
Diabetes
1.00
0.75
0.50
0.25
0.00
−0.5
0.0
0.5
1.0
1.5
HeartDiseaseorAttack
diabetes
no
prediabetic
yes

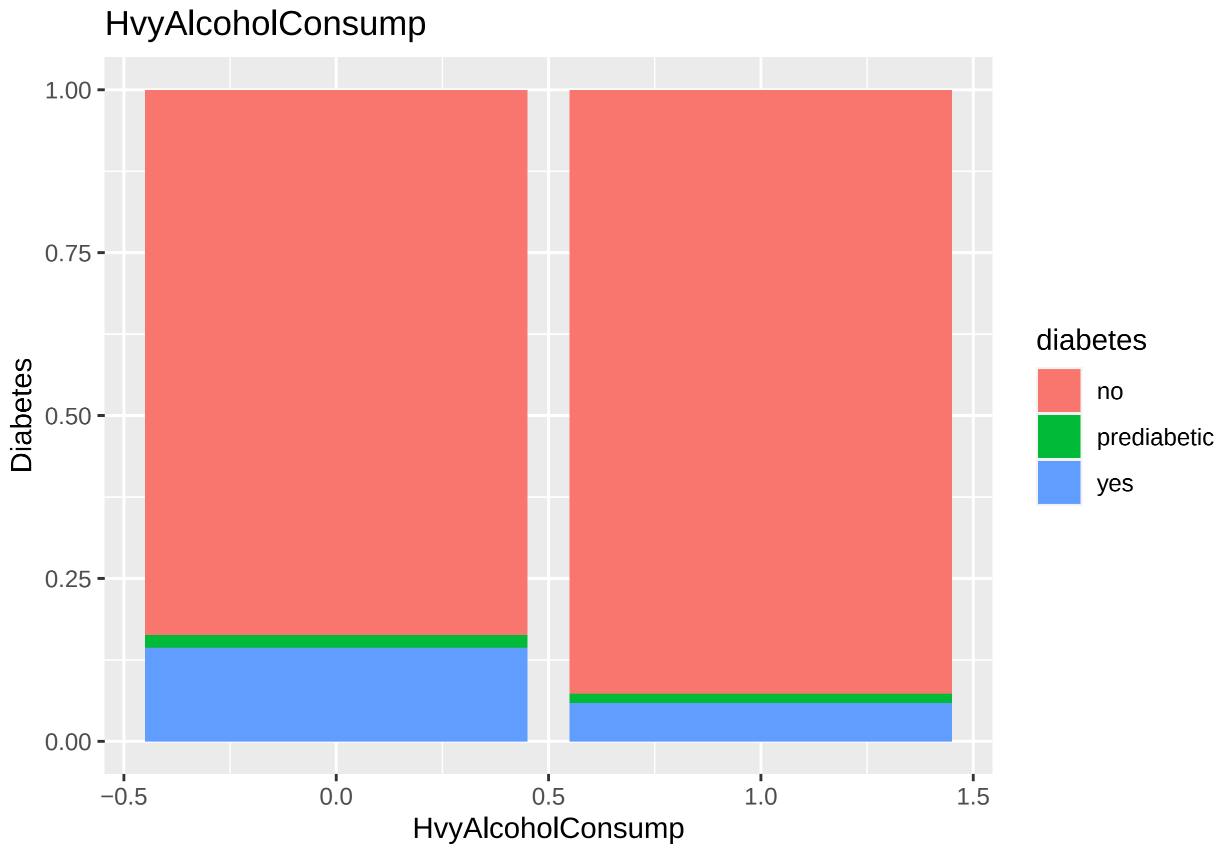
```
pbox7
```


PhysActivity
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
PhysActivity
diabetes
no
prediabetic
yes

```
pbox8
```

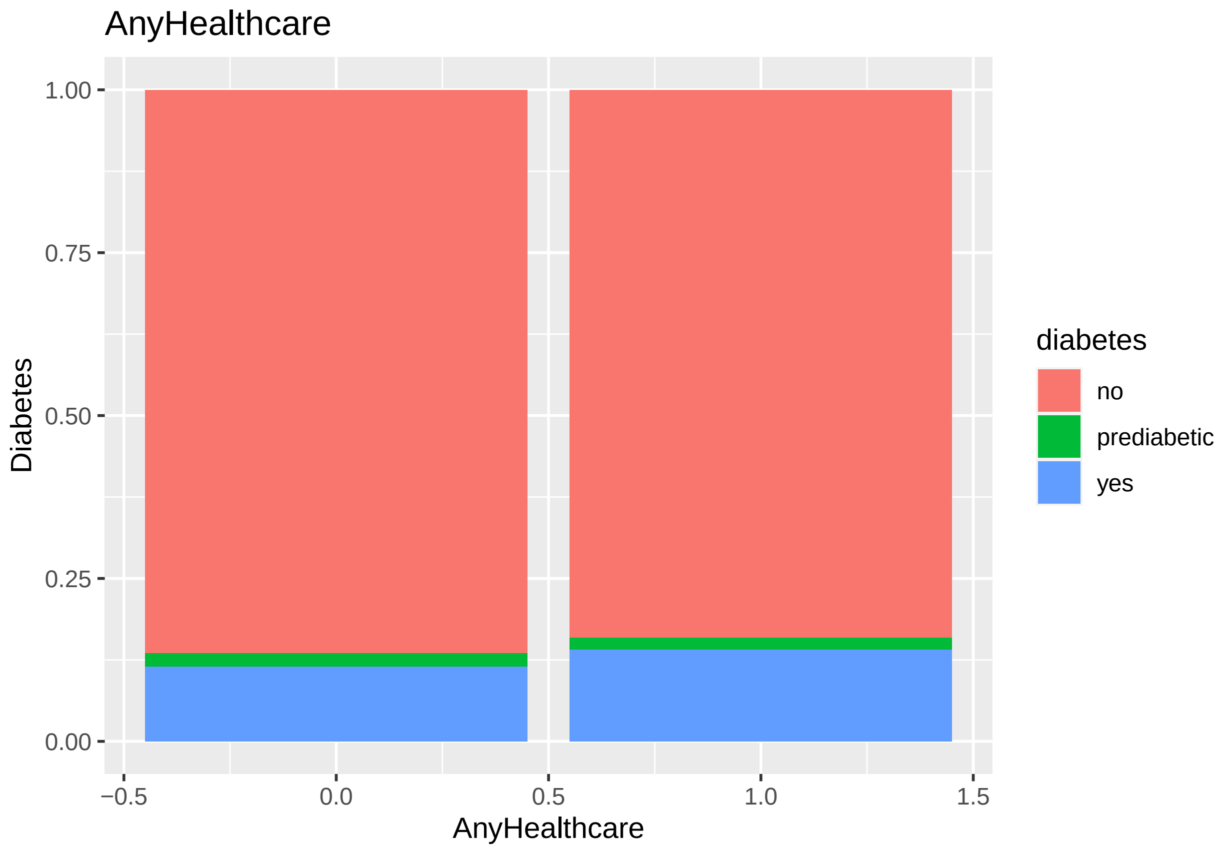

Veggies
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
Veggies
diabetes
no
prediabetic
yes

```
pbox9
```

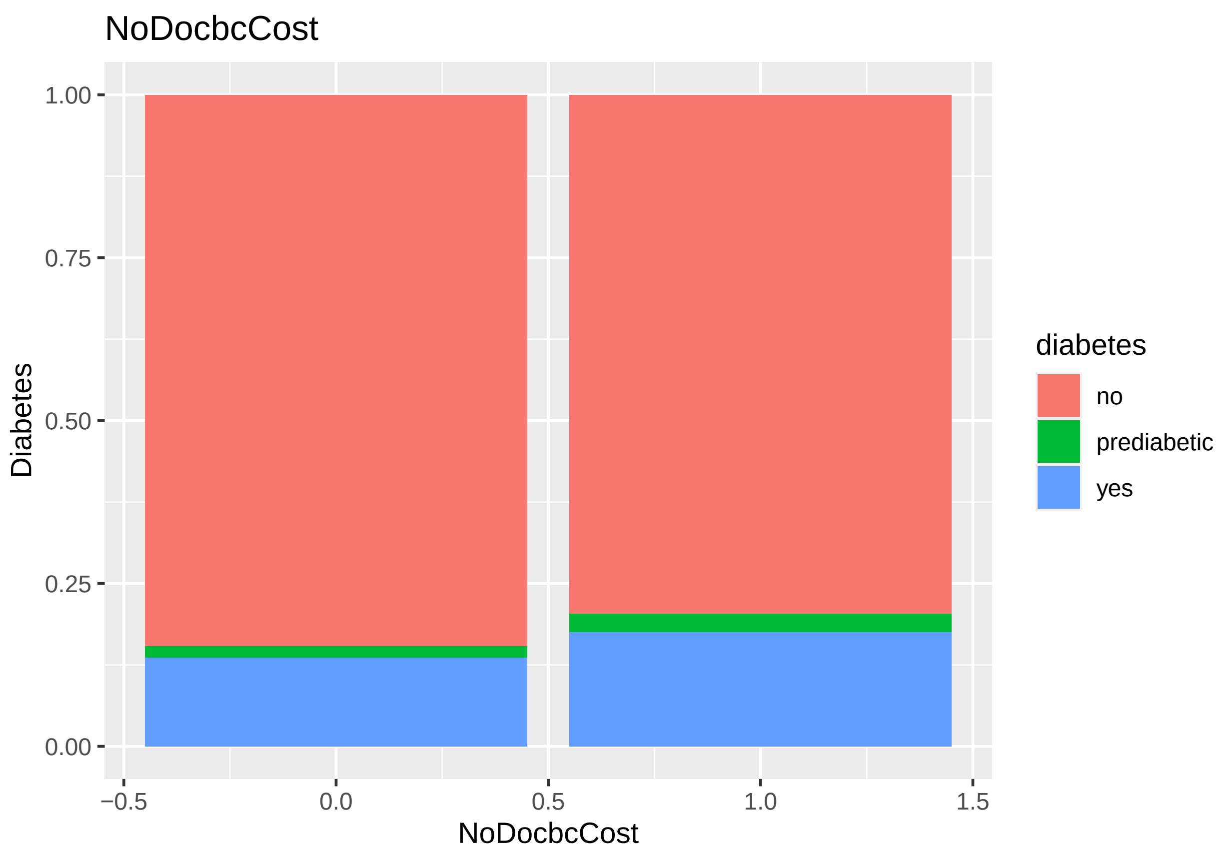


HvyAlcoholConsump
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
HvyAlcoholConsump
diabetes
no
prediabetic
yes


```
pbox10
```

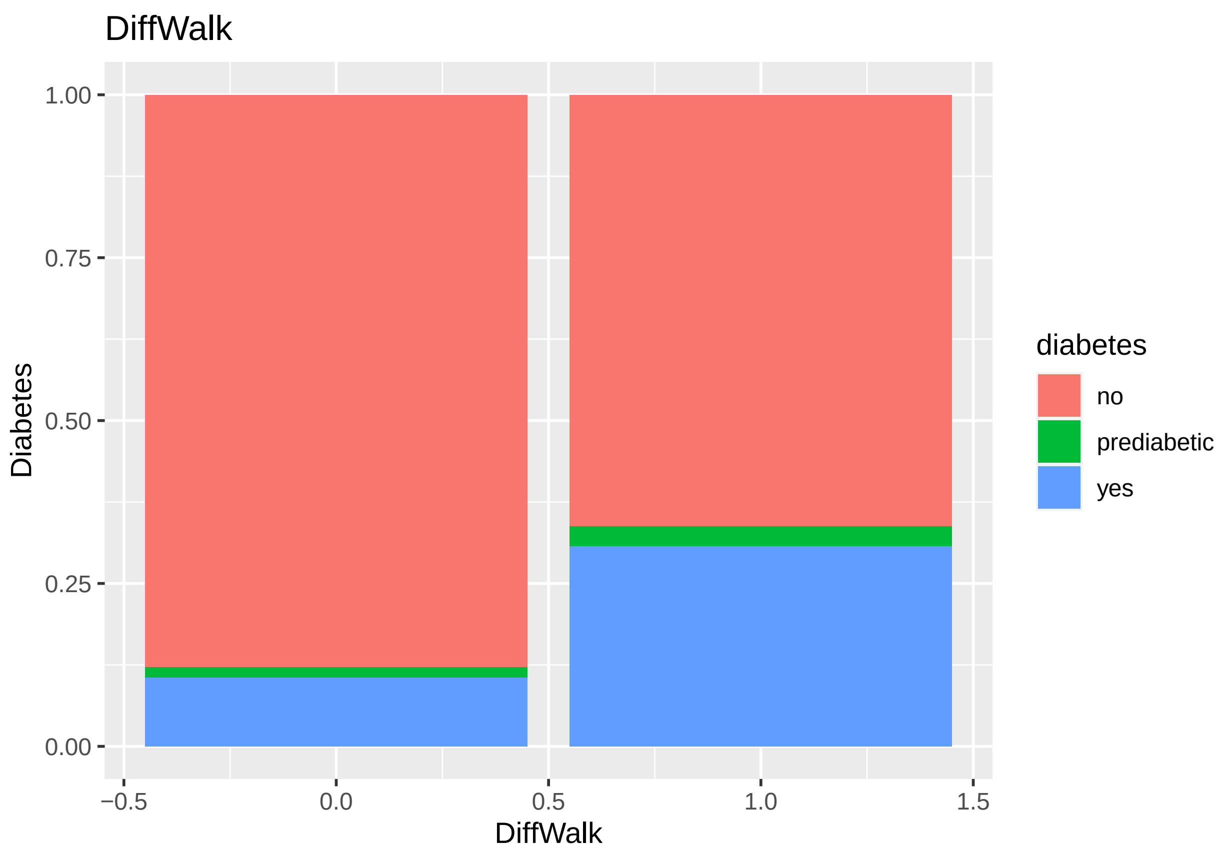

AnyHealthcare
Diabetes
1.00
0.75
0.50
0.25
0.00
−0.5
0.0
0.5
1.0
1.5
AnyHealthcare
diabetes
no
prediabetic
yes

```
pbox11
```



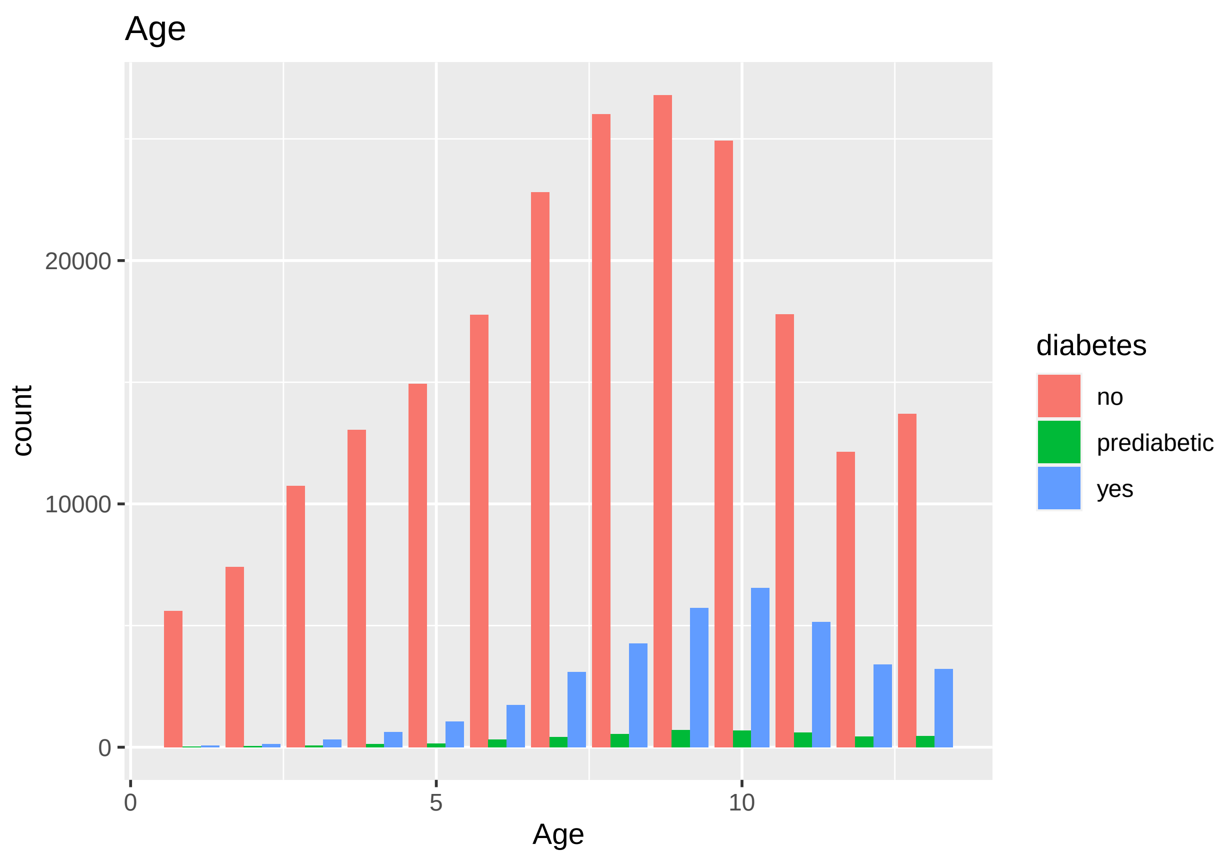
NoDocbcCost
1.00
0.75
0.50
0.25
0.00
Diabetes
−0.5
0.0
0.5
1.0
1.5
NoDocbcCost
diabetes
no
prediabetic
yes


```
pbox12
```



Next, look at "Age" and its relation to response (diabetes diagnosis):

```
ggplot(data, aes(x = Age, fill=factor(Diabetes_012))) +
  geom_bar(position="dodge")+
  labs(title="Age")+
  scale_fill_discrete(name="diabetes", labels=c('no', 'prediabetic', 'yes'))
```



Factor Numeric Variables

```
factored <- data
```

```
factored$Diabetes_012 <- as.factor(factored$Diabetes_012)
factored$HighBP <- as.factor(factored$HighBP)
factored$CholCheck <- as.factor(factored$CholCheck)
factored$Smoker <- as.factor(factored$Smoker)
factored$Stroke <- as.factor(factored$Stroke)
factored$HeartDiseaseorAttack <- as.factor(factored$HeartDiseaseorAttack)
factored$PhysActivity <- as.factor(factored$PhysActivity)
factored$Fruits <- as.factor(factored$Fruits)
factored$Veggies <- as.factor(factored$Veggies)
factored$HvyAlcoholConsump <- as.factor(factored$HvyAlcoholConsump)
factored$AnyHealthcare <- as.factor(factored$AnyHealthcare)
factored$NoDocbcCost <- as.factor(factored$NoDocbcCost)
factored$GenHlth <- as.factor(factored$GenHlth)
factored$MentHlth <- as.factor(factored$MentHlth)
factored$DiffWalk <- as.factor(factored$DiffWalk)
factored$Sex <- as.factor(factored$Sex)
factored$Age <- as.factor(factored$Age)
factored$Education <- as.factor(factored$Education)
factored$Income <- as.factor(factored$Income)
```

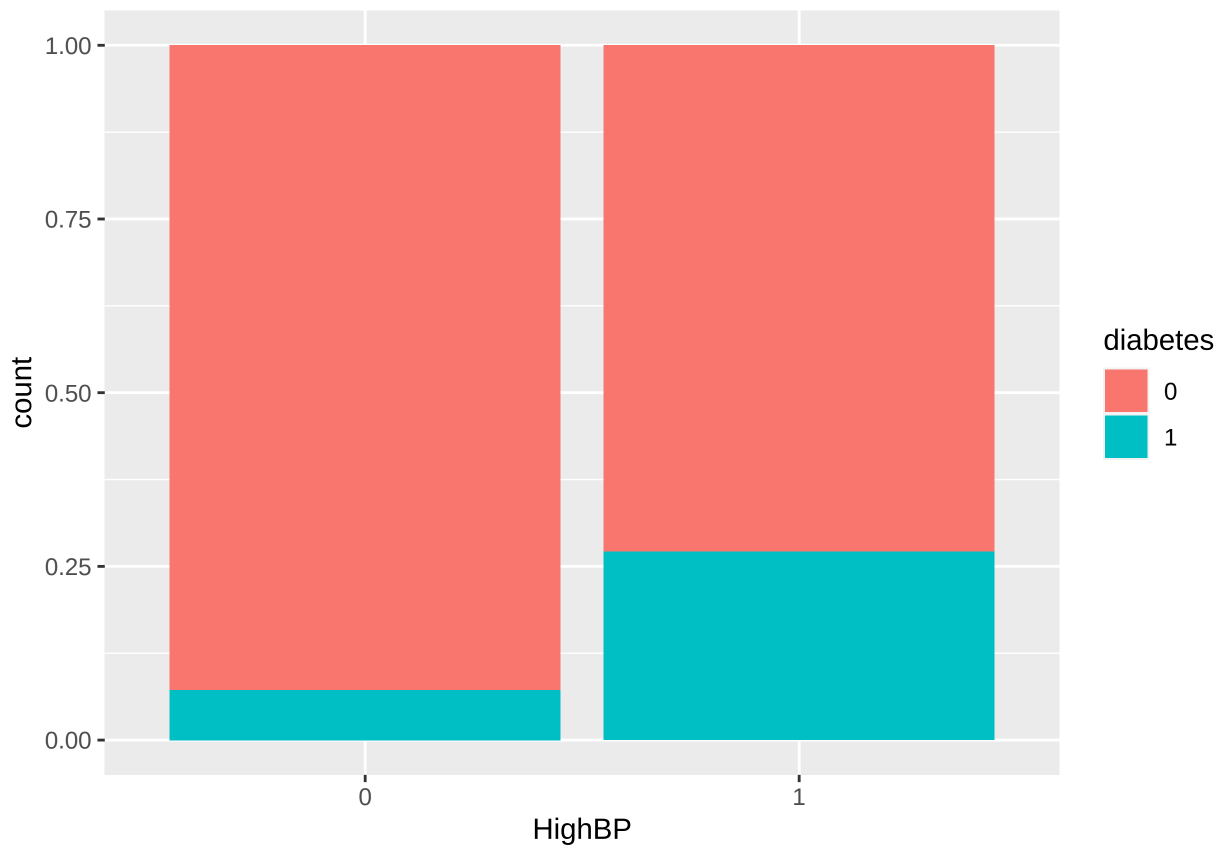
Make diabetes response variable binary

```
factored$diabetes <- ifelse(factored$Diabetes_012 == 0, 0, 1)
factored$diabetes <- as.factor(factored$diabetes)
```

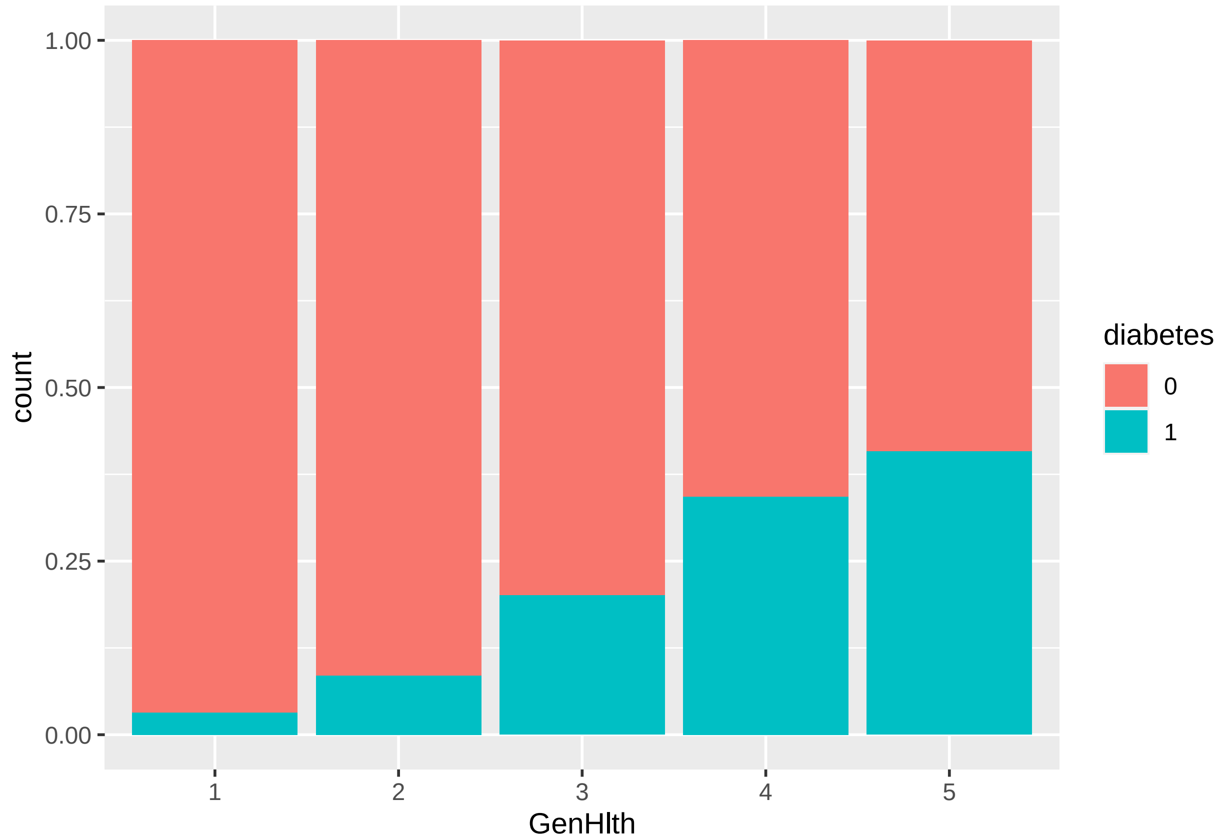
EDA of factored binary outcome dataset

Next, look at plots of 2 most correlated predictors and color by outcome.

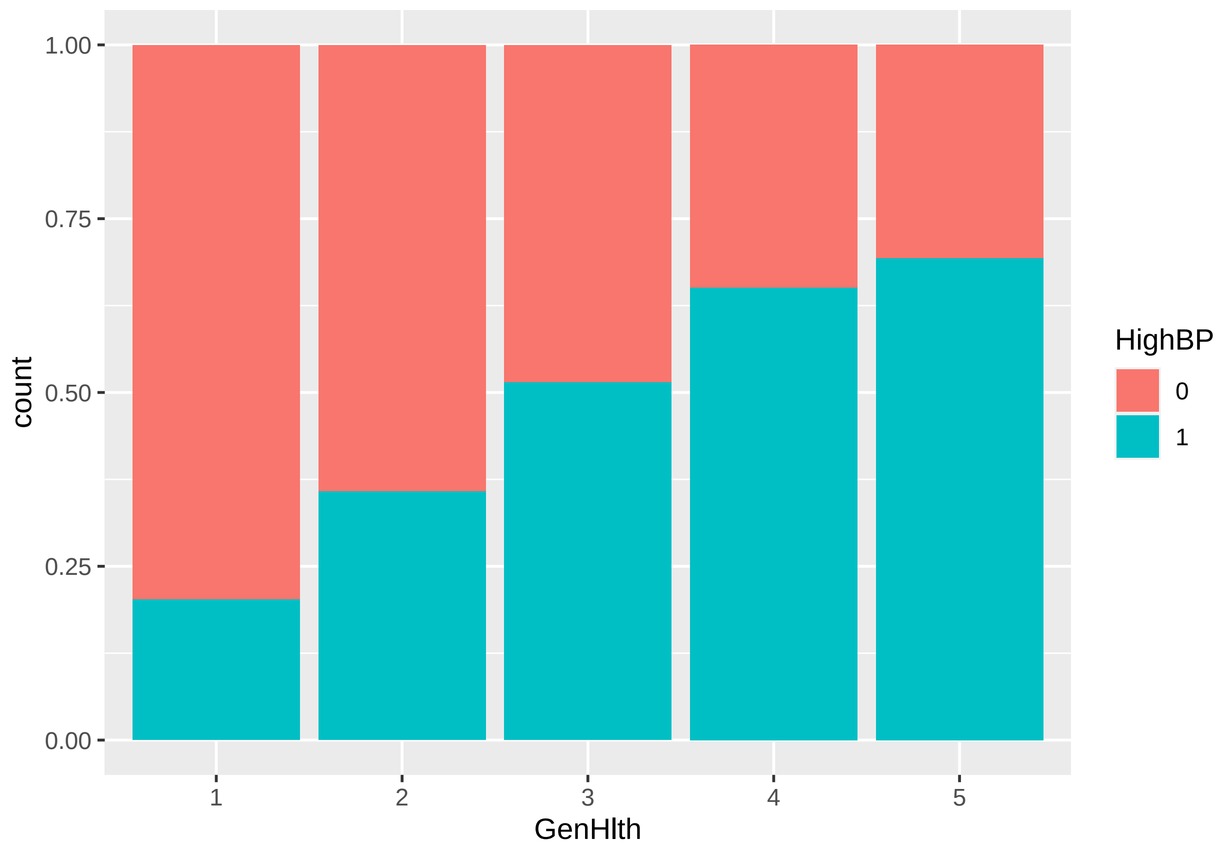
```
ggplot(factored, aes(x = HighBP, fill = diabetes)) +
  geom_bar(position="fill")
```




```
ggplot(factored, aes(x = GenHlth, fill = diabetes)) +
  geom_bar(position="fill")
```




```
ggplot(factored, aes(x = GenHlth, fill = HighBP)) +
  geom_bar(position="fill")
```



Modeling

Split data train and test

```
set.seed(17)
sample <- sample(c(TRUE, FALSE), nrow(factored), replace=TRUE, prob=c(0.7,0.3))
train  <- factored[sample, ]
test   <- factored[!sample, ]
```

Logistic Regression

```
glm.fit.all <- glm(diabetes ~ HighBP+ HighChol + CholCheck + HeartDiseaseorAttack + AnyHealthcare
+ PhysActivity + HvyAlcoholConsump + Fruits + Veggies + GenHlth + DiffWalk + Sex + Income + Education +
               data = factored, family = binomial)
summary(glm.fit.all)
```

```
##
## Call:
## glm(formula = diabetes ~ HighBP + HighChol + CholCheck + HeartDiseaseorAttack +
##     AnyHealthcare + PhysActivity + HvyAlcoholConsump + Fruits +
##     Veggies + GenHlth + DiffWalk + Sex + Income + Education +
##     BMI + PhysHlth, family = binomial, data = factored)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6391  -0.5709  -0.3458  -0.1950   3.3048
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.3521264  0.2029512 -31.299  < 2e-16 ***
## HighBP1               0.8580177  0.0135340  63.397  < 2e-16 ***
## HighChol              0.6622910  0.0127883  51.789  < 2e-16 ***
## CholCheck1            1.2682978  0.0612894  20.694  < 2e-16 ***
## HeartDiseaseorAttack1 0.3486985  0.0169737  20.543  < 2e-16 ***
## AnyHealthcare1        0.2637651  0.0302104   8.731  < 2e-16 ***
## PhysActivity1        -0.0773155  0.0137267  -5.632 1.78e-08 ***
## HvyAlcoholConsump1   -0.7310683  0.0348105 -21.001  < 2e-16 ***
## Fruits1               0.0187151  0.0129260   1.448    0.148
## Veggies1             -0.0317548  0.0151412  -2.097    0.036 *
## GenHlth2              0.6743468  0.0299579  22.510  < 2e-16 ***
## GenHlth3              1.3102135  0.0293286  44.674  < 2e-16 ***
## GenHlth4              1.7089027  0.0320518  53.317  < 2e-16 ***
## GenHlth5              1.8260068  0.0390360  46.778  < 2e-16 ***
## DiffWalk1             0.2331478  0.0161499  14.436  < 2e-16 ***
## Sex1                  0.2248034  0.0126142  17.821  < 2e-16 ***
## Income2               0.0486912  0.0340345   1.431    0.153
## Income3               0.0147855  0.0326853   0.452    0.651
## Income4              -0.0033858  0.0319829  -0.106    0.916
## Income5              -0.0531764  0.0314616  -1.690    0.091 .
## Income6              -0.1323599  0.0309673  -4.274 1.92e-05 ***
## Income7              -0.1921712  0.0313189  -6.136 8.47e-10 ***
## Income8              -0.3975044  0.0309880 -12.828  < 2e-16 ***
## Education2            0.1319623  0.1910435   0.691    0.490
## Education3           -0.0619709  0.1891820  -0.328    0.743
## Education4           -0.1544469  0.1878698  -0.822    0.411
## Education5           -0.1416108  0.1879406  -0.753    0.451
## Education6           -0.2041633  0.1880597  -1.086    0.278
## BMI                   0.0503335  0.0008347  60.300  < 2e-16 ***
## PhysHlth             -0.0038936  0.0007620  -5.110 3.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221031  on 253679  degrees of freedom
## Residual deviance: 177461  on 253650  degrees of freedom
## AIC: 177521
##
## Number of Fisher Scoring iterations: 6
```

```r
glm.probs.all <- predict(glm.fit.all, type = "response")
glm.probs.all[1:10]
```

```
##          1          2          3          4          5          6          7
## 0.63249102 0.01687247 0.38194366 0.09855522 0.16073669 0.13680568 0.16900813
##          8          9         10
```

```
## 0.31757030 0.58580367 0.05293381
```

```r
glm.pred.all <- rep(0, length(factored$diabetes))
glm.pred.all[glm.probs.all > 0.5] <- 1
```

```r
table(glm.pred.all, factored$diabetes)
```

```
##
## glm.pred.all      0      1
##            0 208152  32874
##            1   5551   7103
```

```r
accuracy <- sum(diag(table(glm.pred.all, factored$diabetes)))/nrow(factored)
accuracy
```

```
## [1] 0.8485296
```

Now make model based off of training data:

```r
glm.fit.trainall <- glm(diabetes ~ HighBP+ HighChol + CholCheck + HeartDiseaseorAttack + AnyHealthcare
+ PhysActivity + HvyAlcoholConsump + Fruits + Veggies + GenHlth + DiffWalk + Sex + Income + Education +
               data = train, family = binomial)
glm.probs.trainall <- predict(glm.fit.trainall, test, type = "response")
```

```r
glm.pred.trainall <- rep(0, length(test))
glm.pred.trainall[glm.probs.trainall > 0.5] <- 1
table(glm.pred.trainall, test$diabetes)
```

```
##
## glm.pred.trainall    0    1
##                 0   17    4
##                 1 1624 2128
```

```r
accuracy <- sum(diag(table(glm.pred.trainall, test$diabetes)))/nrow(test)
accuracy
```

```
## [1] 0.02817697
```

To improve the accuracy we will consider a subset of predictors. Look at correlations to decide. The most correlated to diabetes are GenHlth and HighBP.

```r
glm.fit.cor <- glm(diabetes ~ GenHlth + HighBP, data=train, family = binomial)
glm.probs.cor <- predict(glm.fit.cor, test, type = "response")
glm.pred.cor <- rep("no diabetes", length(test))
glm.pred.cor[glm.probs.cor > 0.5] <- "diabetes"
table(glm.pred.cor, test$diabetes)
```

```
##
## glm.pred.cor   0  1
##   no diabetes 18  5
```

```
Accuracy <- (0+5)/(1+18+5+0)
Accuracy
```

```
## [1] 0.2083333
```

The subset of predictors made our predictive performance worse.

KNN

```
#KNN wont knit but works (just takes a while to run)
#library(class)
#set.seed(1)
#knn.pred <- knn(train, test, train$diabetes, k = 10)
#table(knn.pred, test$diabetes)
```

```
#accuracy <- sum(diag(table(knn.pred, test$diabetes)))/nrow(test)
#accuracy
```

Perform CV to find best k value. . . .?

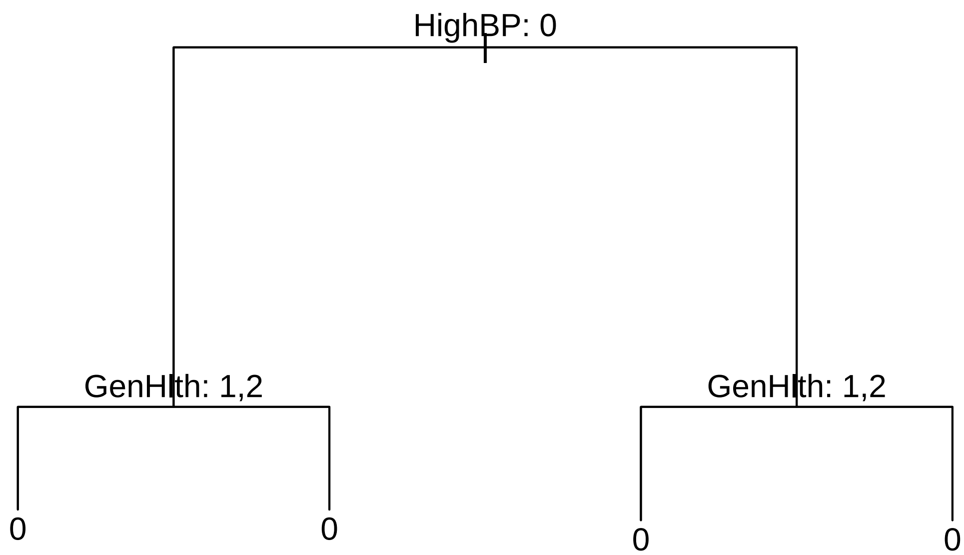
Trees

```
library(tree)
```

```
tree.all <- tree(diabetes ~ HighBP+ HighChol + CholCheck + HeartDiseaseorAttack + AnyHealthcare
+ PhysActivity + HvyAlcoholConsump + Fruits + Veggies + GenHlth + DiffWalk + Sex + Income + Education +
summary(tree.all)
```

```
##
## Classification tree:
## tree(formula = diabetes ~ HighBP + HighChol + CholCheck + HeartDiseaseorAttack +
##     AnyHealthcare + PhysActivity + HvyAlcoholConsump + Fruits +
##     Veggies + GenHlth + DiffWalk + Sex + Income + Education +
##     BMI + PhysHlth, data = factored)
## Variables actually used in tree construction:
## [1] "HighBP"  "GenHlth"
## Number of terminal nodes:  4
## Residual mean deviance:  0.7534 = 191100 / 253700
## Misclassification error rate: 0.1576 = 39977 / 253680
```

```
plot(tree.all)
text(tree.all, pretty = 0)
```



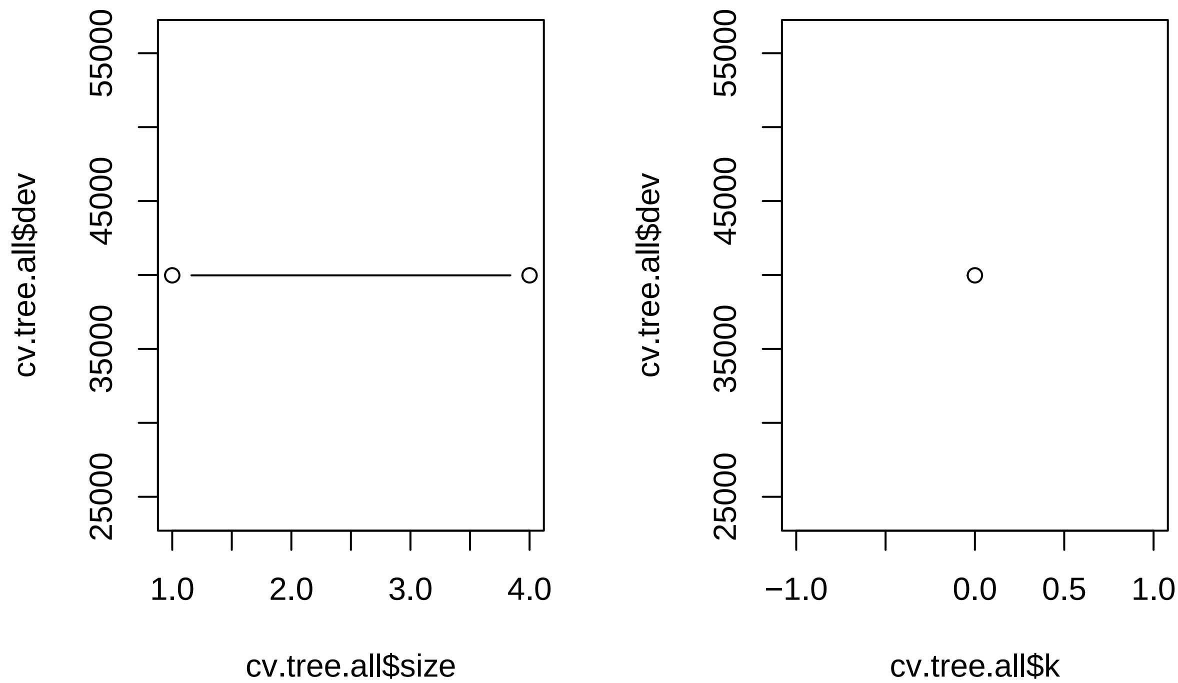

```r
set.seed(3)
cv.tree.all <- cv.tree(tree.all, FUN = prune.misclass)
names(cv.tree.all)
```

```
## [1] "size"   "dev"    "k"      "method"
```

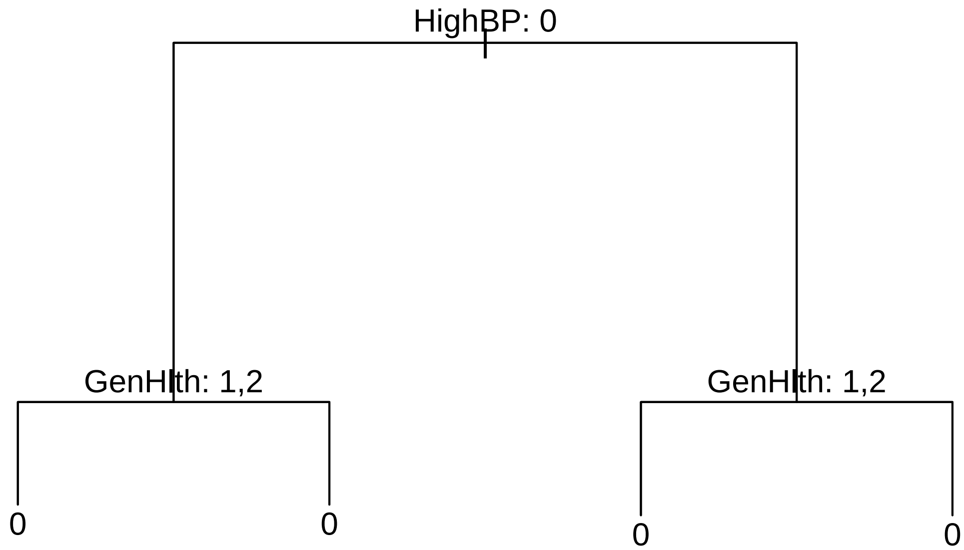
```r
cv.tree.all
```

```
## $size
## [1] 4 1
##
## $dev
## [1] 39977 39977
##
## $k
## [1] -Inf    0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
par(mfrow = c(1,2))
plot(cv.tree.all$size, cv.tree.all$dev, type = "b")
plot(cv.tree.all$k, cv.tree.all$dev, type = "b")
```




```
prune.tree <- prune.misclass(tree.all, best = 4)
plot(prune.tree)
text(prune.tree, pretty = 0)
```


HighBP: 0
GenHlth: 1,2
GenHlth: 1,2
0
0
0
0