Exploratory Data Analysis(EDA) for Alzheimer disease data

Chaeeun Shin

Introduction

The purpose of this document is to get insight of the data before assumption, focusing on the patient's demographic details.

1. Package Install

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
```

1. *tidyverse*: This package is used for data manipulation.
2. *dplyr*: This package is used for data manipulation.
3. *ggplot2*: This package is used for data visualization.

2. Reading data

```
data <- read.csv("alzheimers_disease_data.csv")
head(data)
```

```
##   PatientID Age Gender Ethnicity EducationLevel      BMI Smoking
## 1      4751  73      0         0              2 22.92775       0
## 2      4752  89      0         0              0 26.82768       0
## 3      4753  73      0         3              1 17.79588       0
## 4      4754  74      1         0              1 33.80082       1
## 5      4755  89      0         0              0 20.71697       0
## 6      4756  86      1         1              1 30.62689       0
##   AlcoholConsumption PhysicalActivity DietQuality SleepQuality
## 1          13.297218        6.3271125   1.3472143     9.025679
## 2           4.542524        7.6198845   0.5187671     7.151293
```

```
## 3          19.555085        7.8449878   1.8263347     9.673574
## 4          12.209266        8.4280014   7.4356041     8.392554
## 5          18.454356        6.3104607   0.7954975     5.597238
## 6           4.140144        0.2110616   1.5849220     7.261953
##   FamilyHistoryAlzheimers CardiovascularDisease Diabetes Depression HeadInjury
## 1                       0                     0        1          1          0
## 2                       0                     0        0          0          0
## 3                       1                     0        0          0          0
## 4                       0                     0        0          0          0
## 5                       0                     0        0          0          0
## 6                       0                     0        1          0          0
##   Hypertension SystolicBP DiastolicBP CholesterolTotal CholesterolLDL
## 1            0        142          72         242.3668       56.15090
## 2            0        115          64         231.1626      193.40800
## 3            0         99         116         284.1819      153.32276
## 4            0        118         115         159.5822       65.36664
## 5            0         94         117         237.6022       92.86970
## 6            0        168          62         280.7125      198.33463
##   CholesterolHDL CholesterolTriglycerides      MMSE FunctionalAssessment
## 1       33.68256                162.18914 21.463532             6.518877
## 2       79.02848                294.63091 20.613267             7.118696
## 3       69.77229                 83.63832  7.356249             5.895077
## 4       68.45749                277.57736 13.991127             8.965106
## 5       56.87430                291.19878 13.517609             6.045039
## 6       79.08050                263.94365 27.517529             5.510144
##   MemoryComplaints BehavioralProblems        ADL Confusion Disorientation
## 1                0                  0 1.72588346         0              0
## 2                0                  0 2.59242413         0              0
## 3                0                  0 7.11954774         0              1
## 4                0                  1 6.48122586         0              0
## 5                0                  0 0.01469122         0              0
## 6                0                  0 9.01568628         1              0
##   PersonalityChanges DifficultyCompletingTasks Forgetfulness Diagnosis
## 1                  0                         1             0         0
## 2                  0                         0             1         0
## 3                  0                         1             0         0
## 4                  0                         0             0         0
## 5                  1                         1             0         0
## 6                  0                         0             0         0
##   DoctorInCharge
## 1      XXXConfid
## 2      XXXConfid
## 3      XXXConfid
## 4      XXXConfid
## 5      XXXConfid
## 6      XXXConfid
```

3. Conversion to categorical data

```
data$Gender <- as.factor(data$Gender)
data$Ethnicity <- as.factor(data$Ethnicity)
data$EducationLevel <- as.factor(data$EducationLevel)
data$Smoking <- as.factor(data$Smoking)
```

```r
data$FamilyHistoryAlzheimers <- as.factor(data$FamilyHistoryAlzheimers)
data$CardiovascularDisease <- as.factor(data$CardiovascularDisease)
data$Diabetes <- as.factor(data$Diabetes)
data$Depression <- as.factor(data$Depression)
data$HeadInjury <- as.factor(data$HeadInjury)
data$Hypertension <- as.factor(data$Hypertension)
data$MemoryComplaints <- as.factor(data$MemoryComplaints)
data$BehavioralProblems <- as.factor(data$BehavioralProblems)
data$Confusion <- as.factor(data$Confusion)
data$Disorientation <- as.factor(data$Disorientation)
data$PersonalityChanges <- as.factor(data$PersonalityChanges)
data$DifficultyCompletingTasks <- as.factor(data$DifficultyCompletingTasks)
data$Forgetfulness <- as.factor(data$Forgetfulness)
data$Diagnosis <- as.factor(data$Diagnosis)

data <- data %>% dplyr::select(-c(PatientID, DoctorInCharge))
```

There are some categorical variables mis-classified as numeric data. They are assigned as factor. Also, the identification numbers of patients and doctors are removed since they are not used for any investigations for this project.

4. Missing data

```r
missing_summary <- sapply(data, function(x) sum(is.na(x)))
cat("\nMissing Data Summary\n")
```

```
##
## Missing Data Summary
```

```r
print(missing_summary)
```

```
##                         Age                      Gender                   Ethnicity
##                           0                           0                           0
##             EducationLevel                         BMI                     Smoking
##                           0                           0                           0
##         AlcoholConsumption           PhysicalActivity                DietQuality
##                           0                           0                           0
##               SleepQuality     FamilyHistoryAlzheimers      CardiovascularDisease
##                           0                           0                           0
##                   Diabetes                  Depression                  HeadInjury
##                           0                           0                           0
##               Hypertension                 SystolicBP                 DiastolicBP
##                           0                           0                           0
##           CholesterolTotal             CholesterolLDL             CholesterolHDL
##                           0                           0                           0
##   CholesterolTriglycerides                        MMSE       FunctionalAssessment
##                           0                           0                           0
##           MemoryComplaints          BehavioralProblems                         ADL
##                           0                           0                           0
##                  Confusion              Disorientation          PersonalityChanges
##                           0                           0                           0
## DifficultyCompletingTasks               Forgetfulness                   Diagnosis
##                           0                           0                           0
```

```r
saveRDS(data,'data.rds')
```

There is no missing values in the dataset.

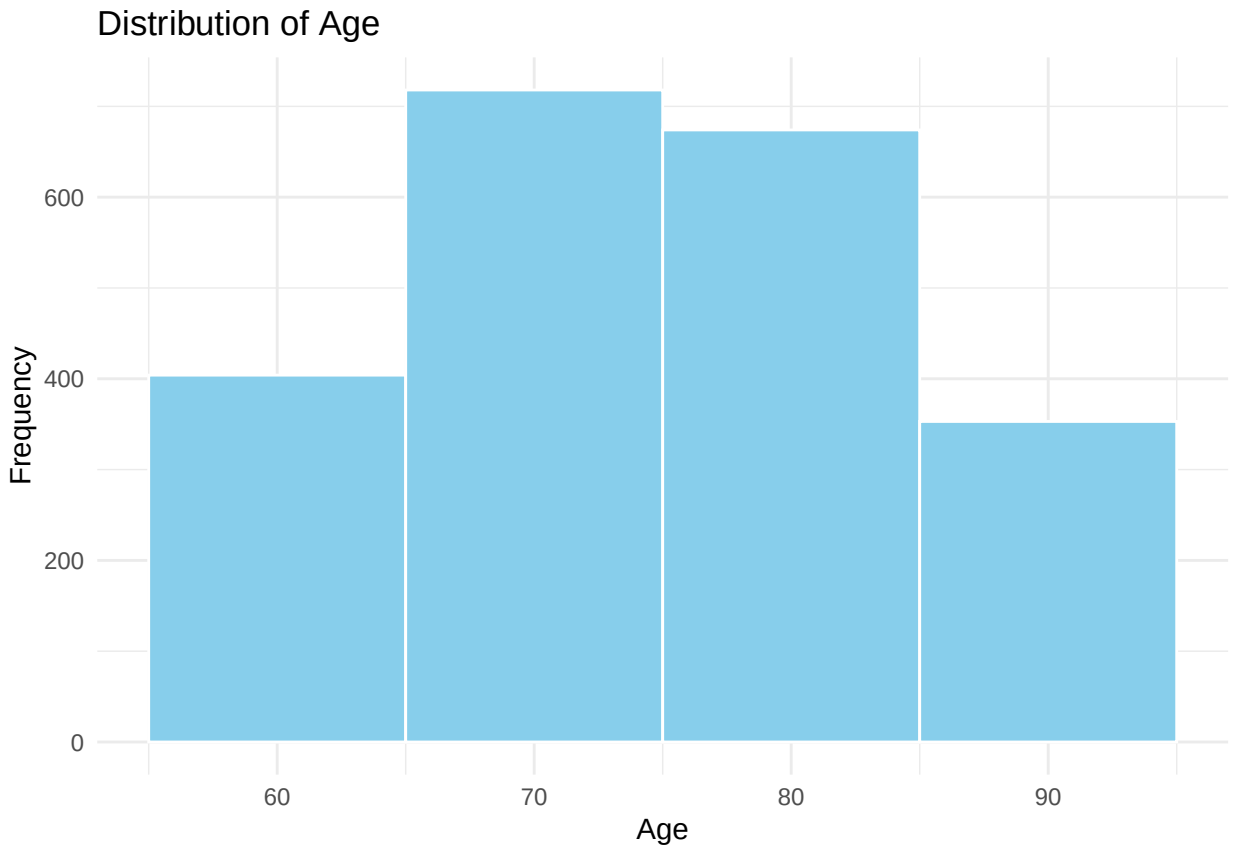
5-1. Distribution of numeric variables

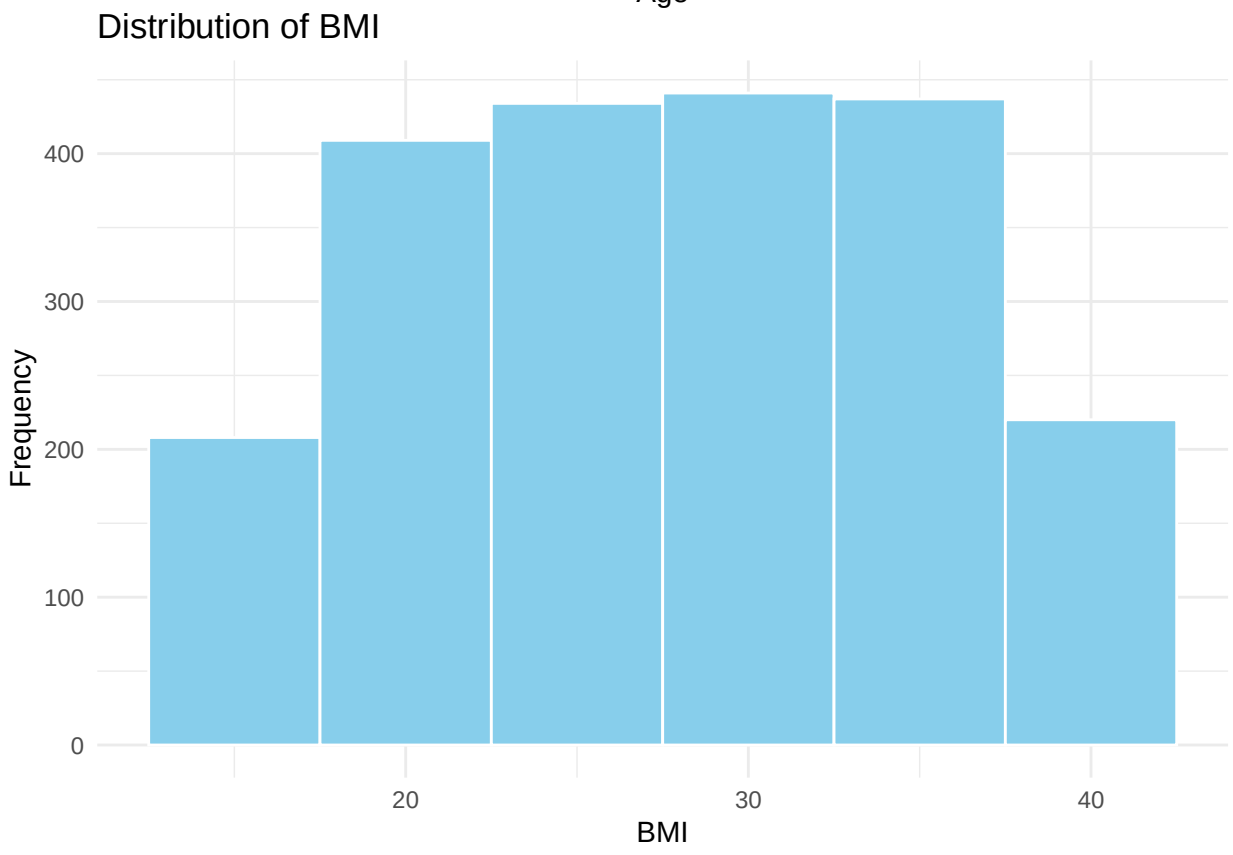
```r
cat("\nHistograms of Continuous Variables:\n")
```

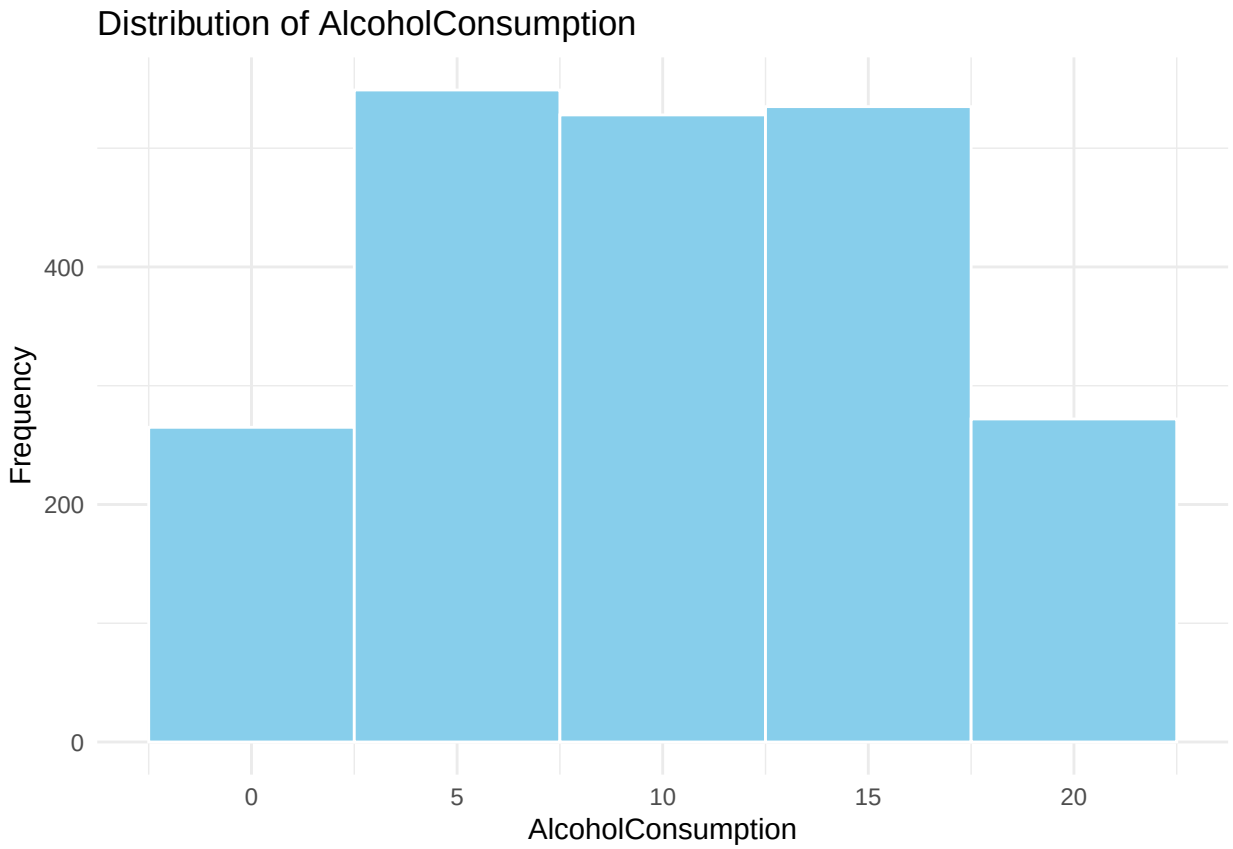
```
##
## Histograms of Continuous Variables:
```

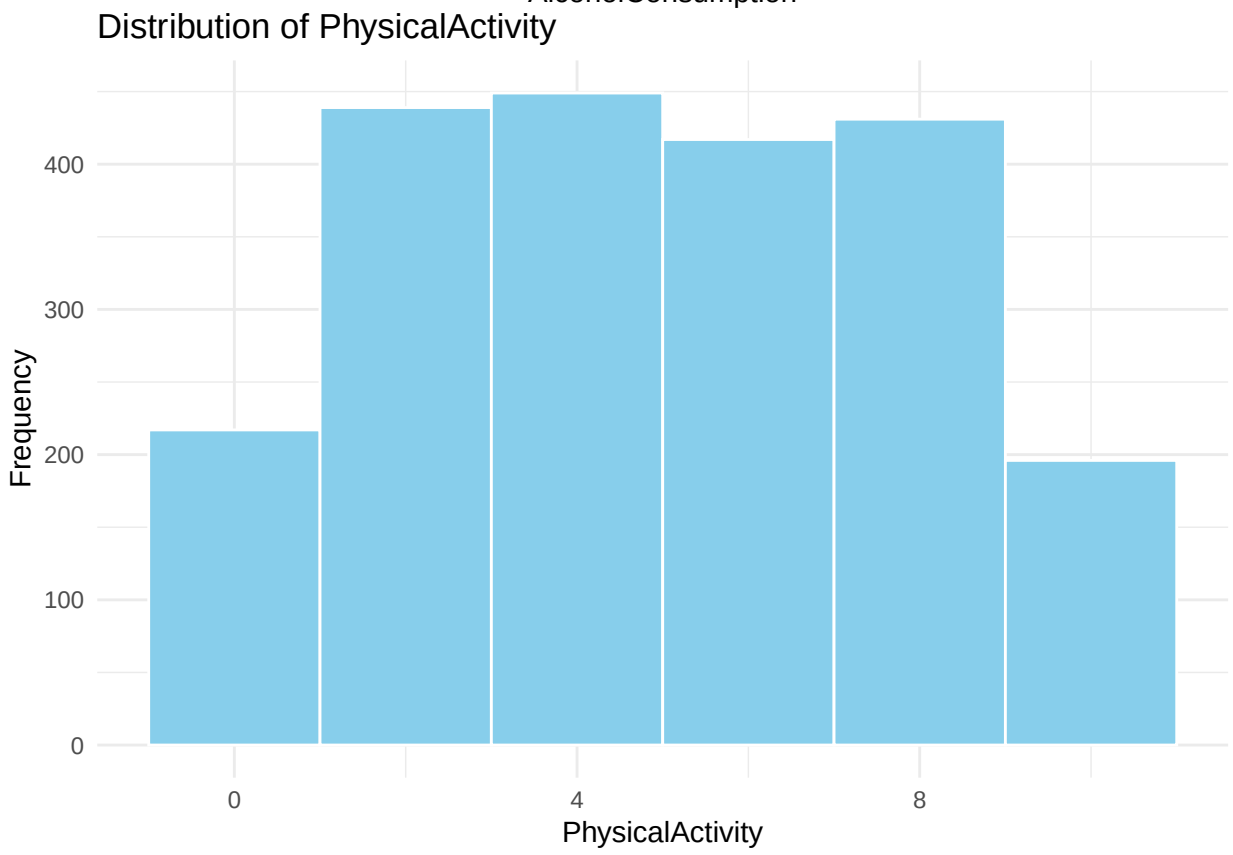
```r
continuous_columns <- data[, sapply(data, is.numeric)]
width <- c(Age = 10,BMI = 5,AlcoholConsumption = 5,PhysicalActivity = 2,DietQuality = 2,SleepQuality = 1
for (col in colnames(continuous_columns)) {
  binwidth <- width[[col]]
 p <- ggplot(data, aes_string(x=col)) +
   geom_histogram(binwidth = binwidth, fill = 'skyblue', color = 'white') +
   labs(title = paste('Distribution of', col),
        x = col,
        y = "Frequency") +
   theme_minimal()
 print(p)
}
```

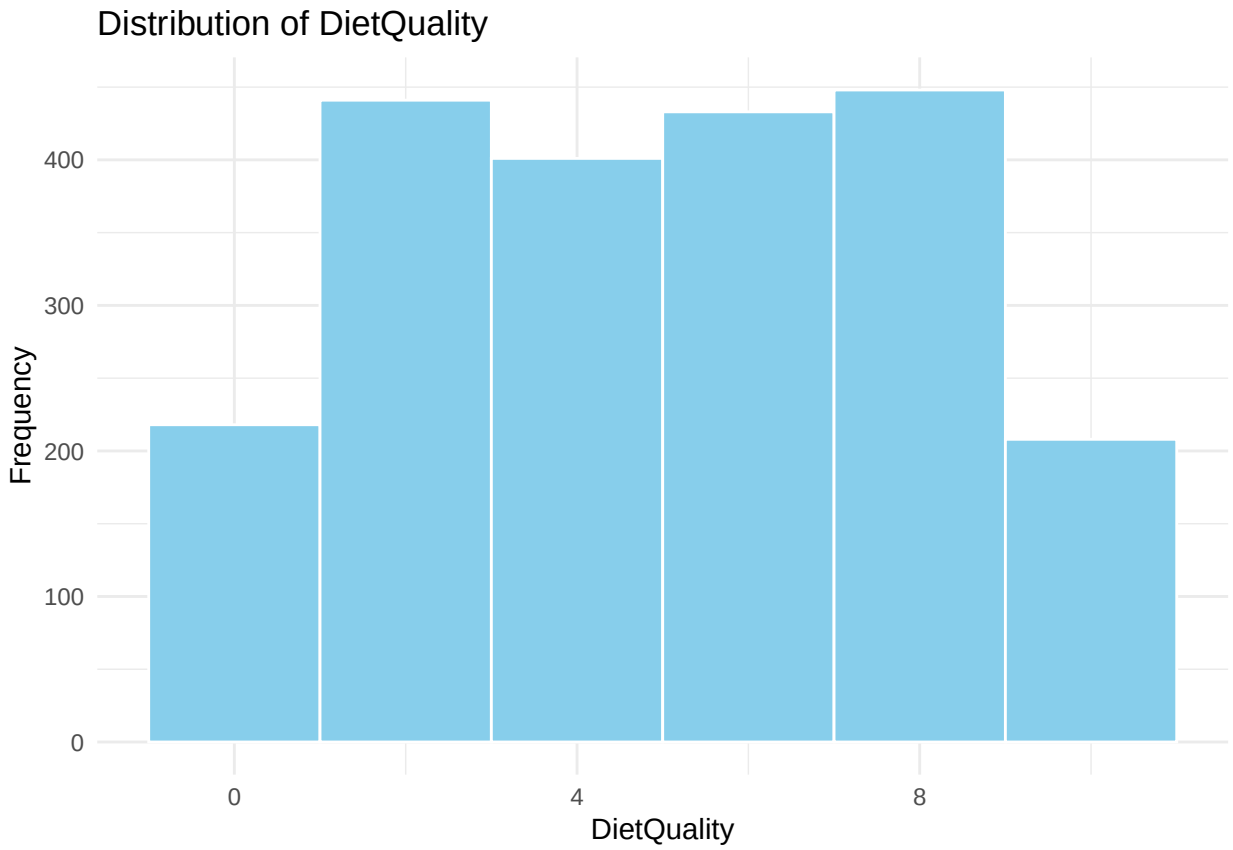
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

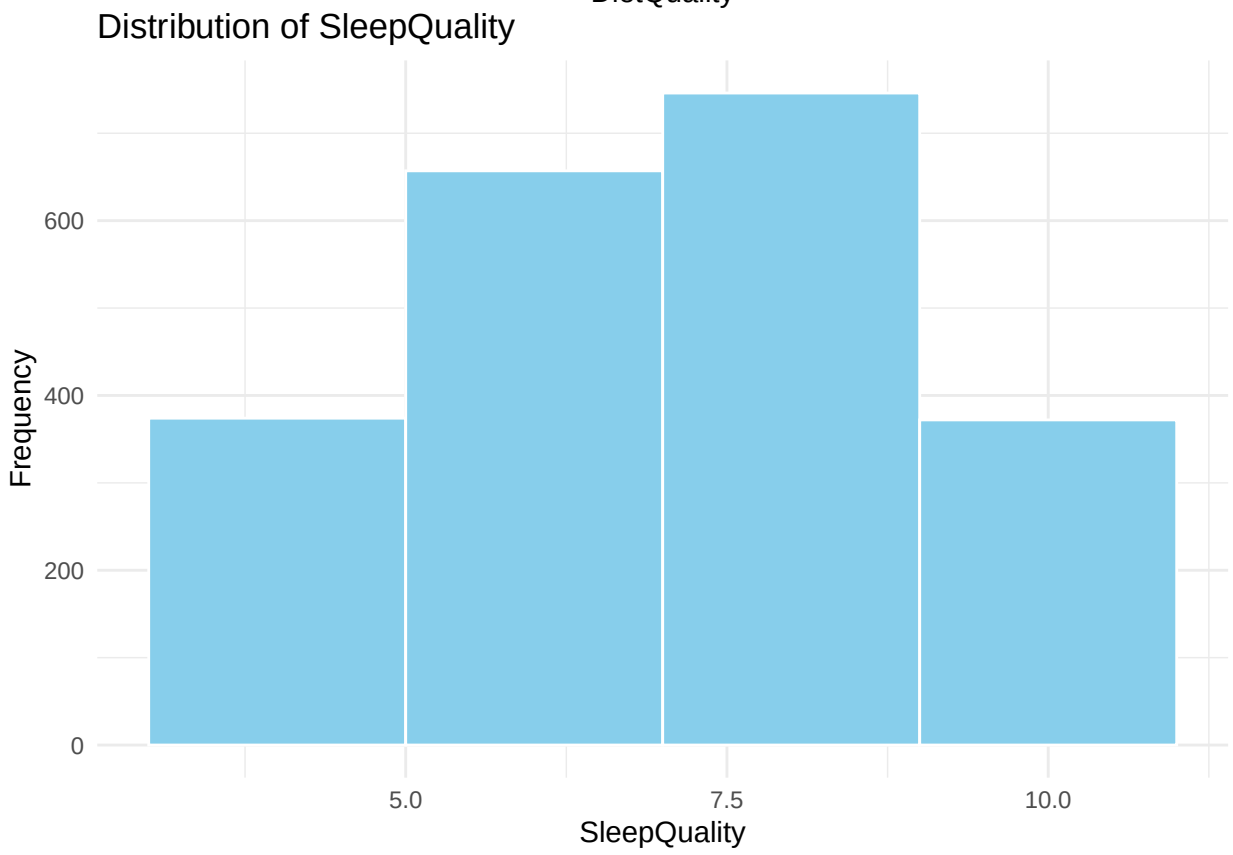

Distribution of Age
Frequency
600
400
200
0
60
70
80
90
Age

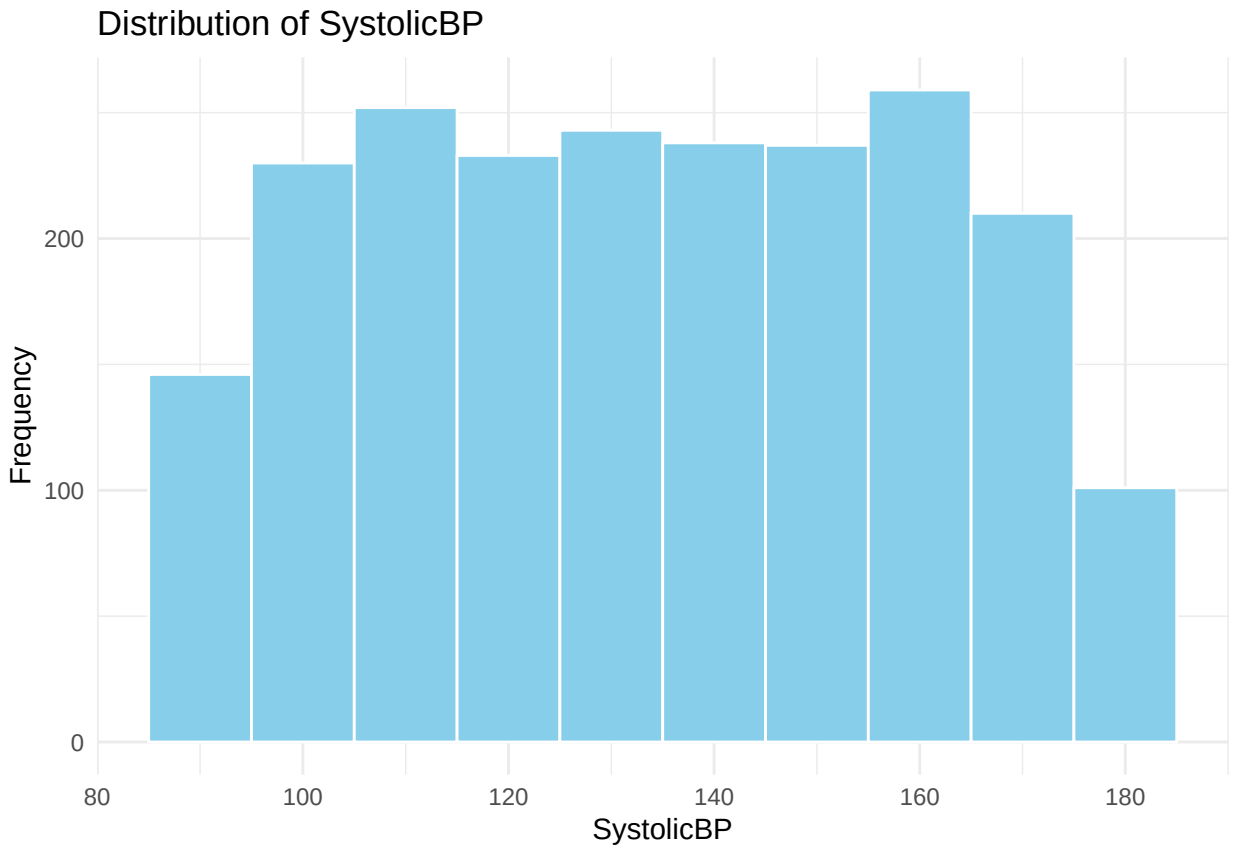

Distribution of BMI
Frequency
400
300
200
100
0
20
30
40
BMI

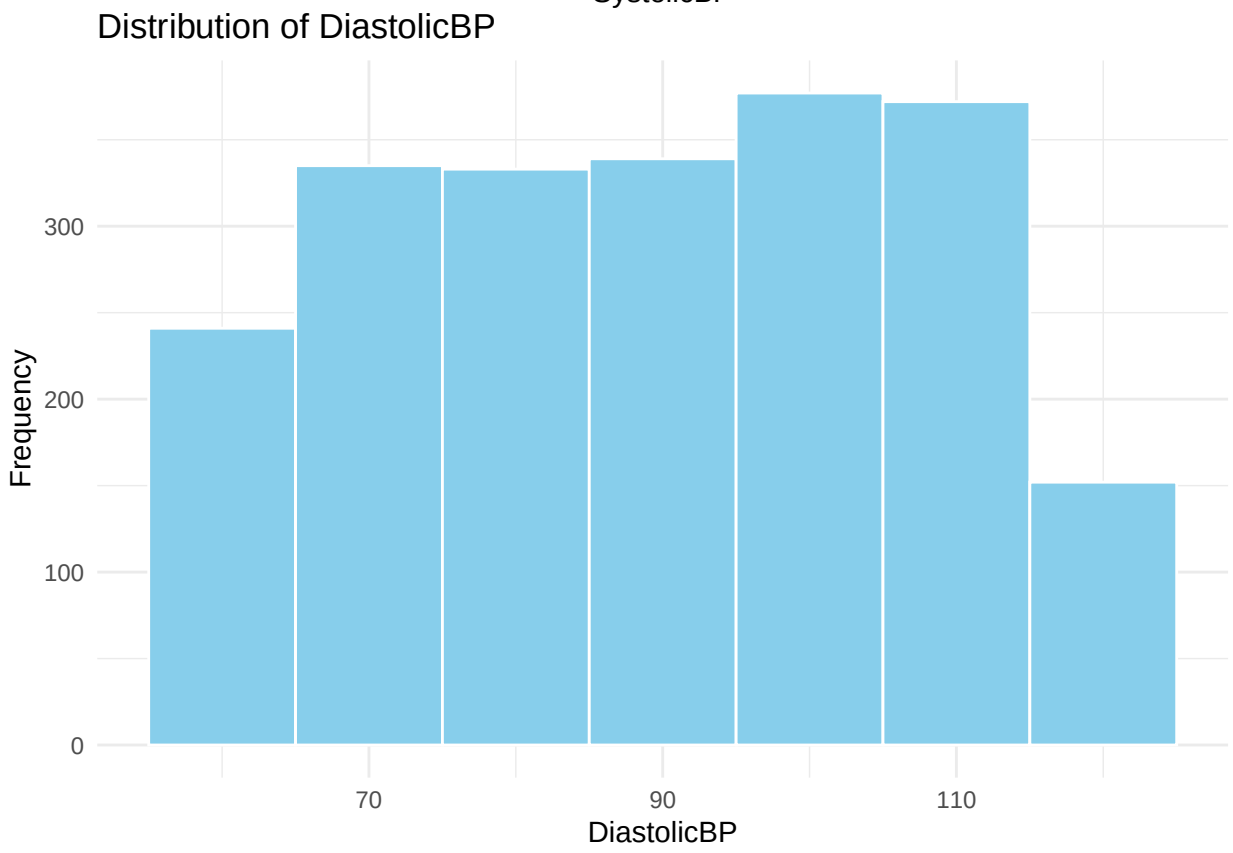

Distribution of AlcoholConsumption
Frequency
400
200
0
0
5
10
15
20
AlcoholConsumption

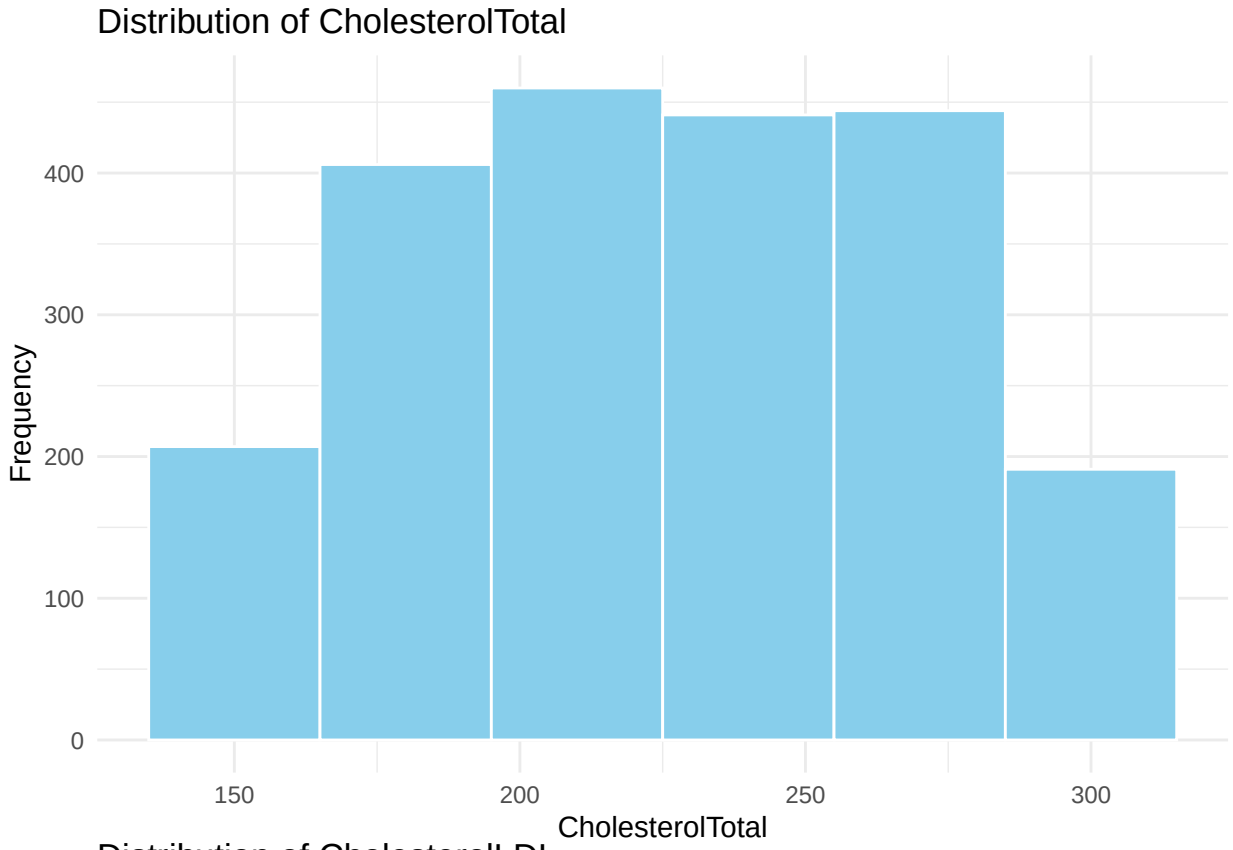

Distribution of PhysicalActivity
Frequency
400
300
200
100
0
0
4
8
PhysicalActivity

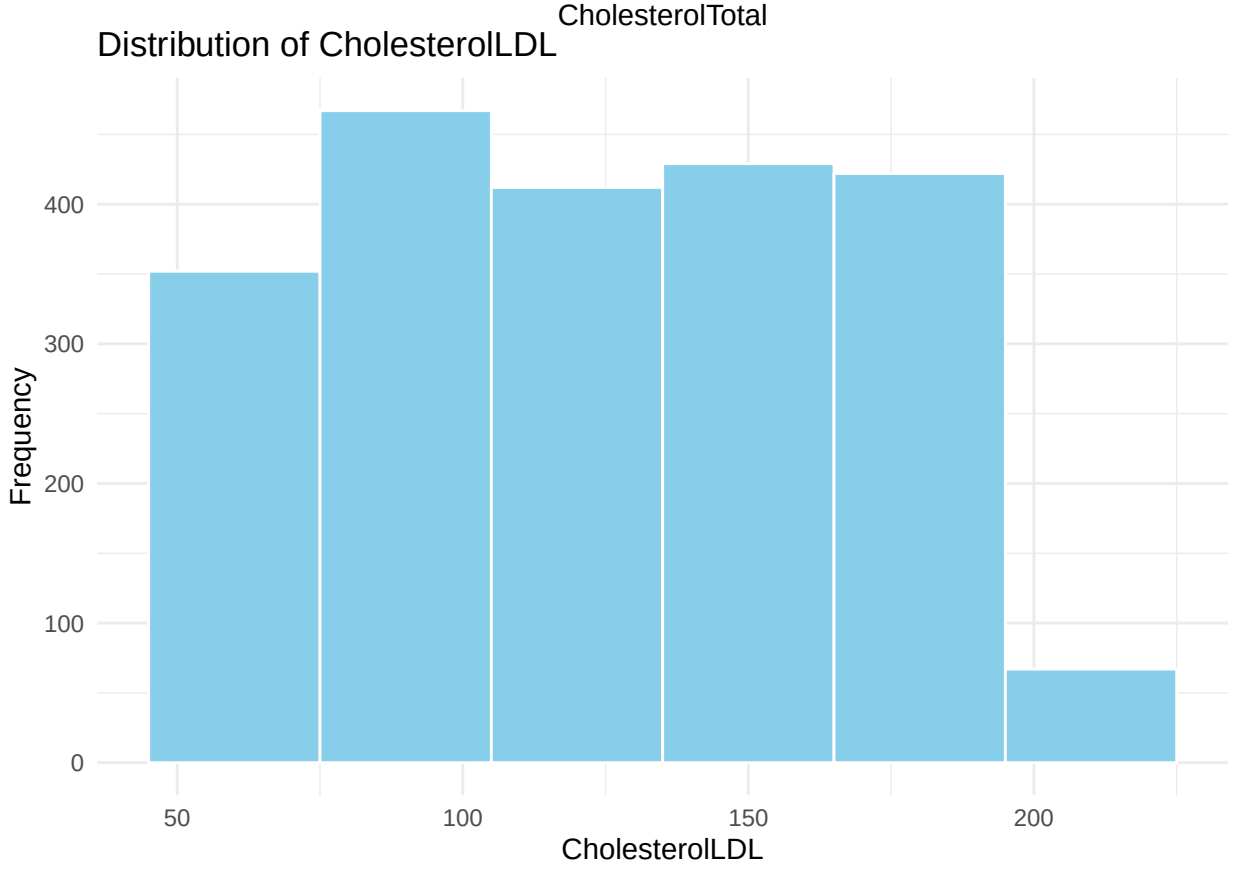

Distribution of DietQuality
Frequency
400
300
200
100
0
0
4
8
DietQuality

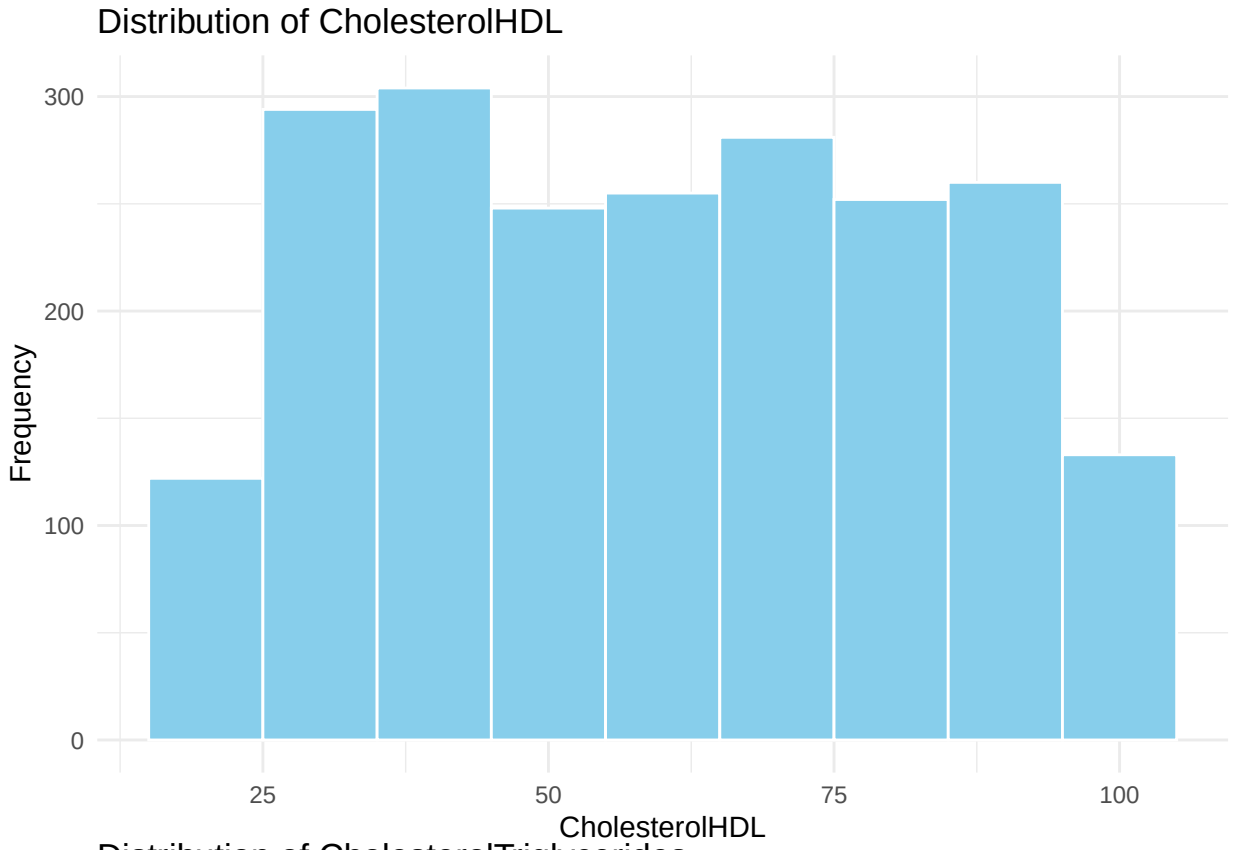

Distribution of SleepQuality
Frequency
600
400
200
0
5.0
7.5
10.0
SleepQuality

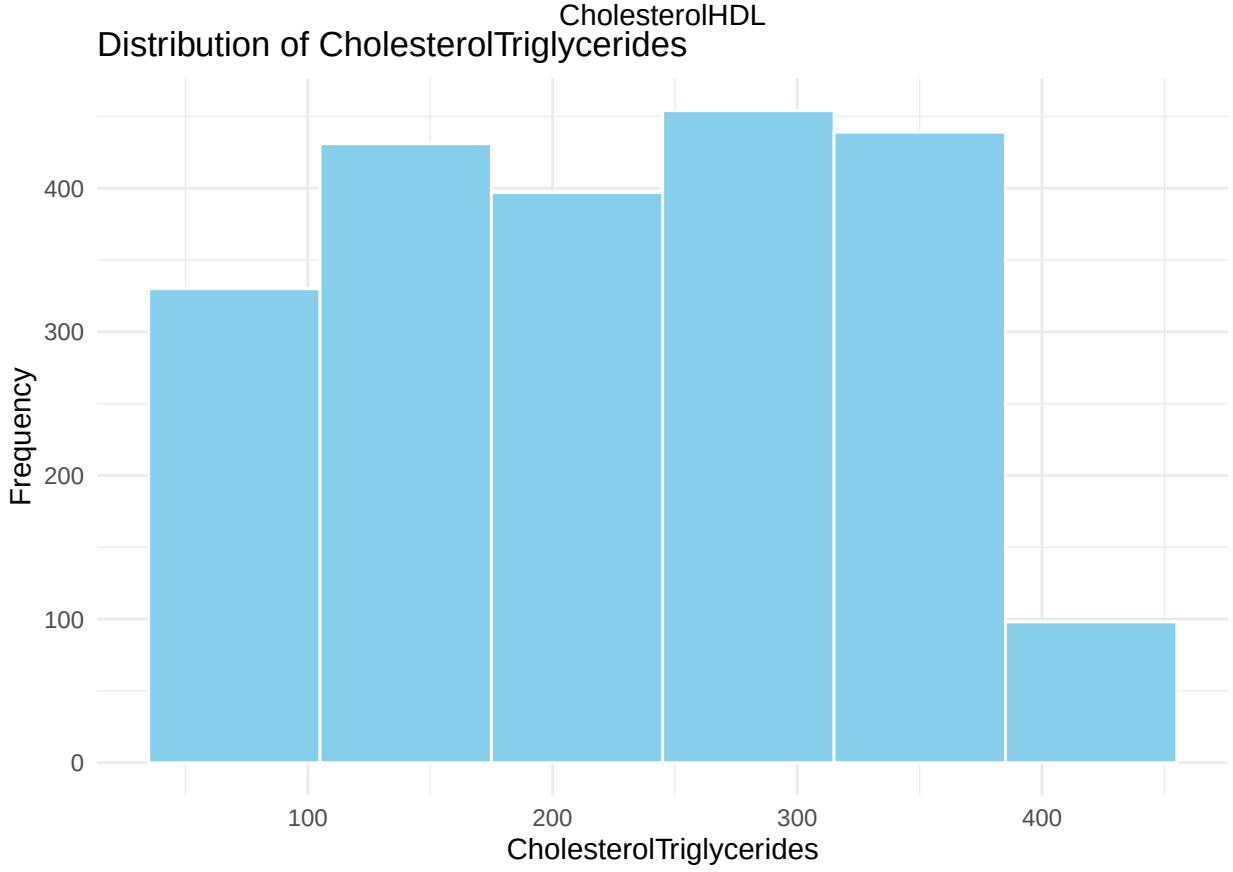

Distribution of SystolicBP
Frequency
0
100
200
80
100
120
140
160
180
SystolicBP

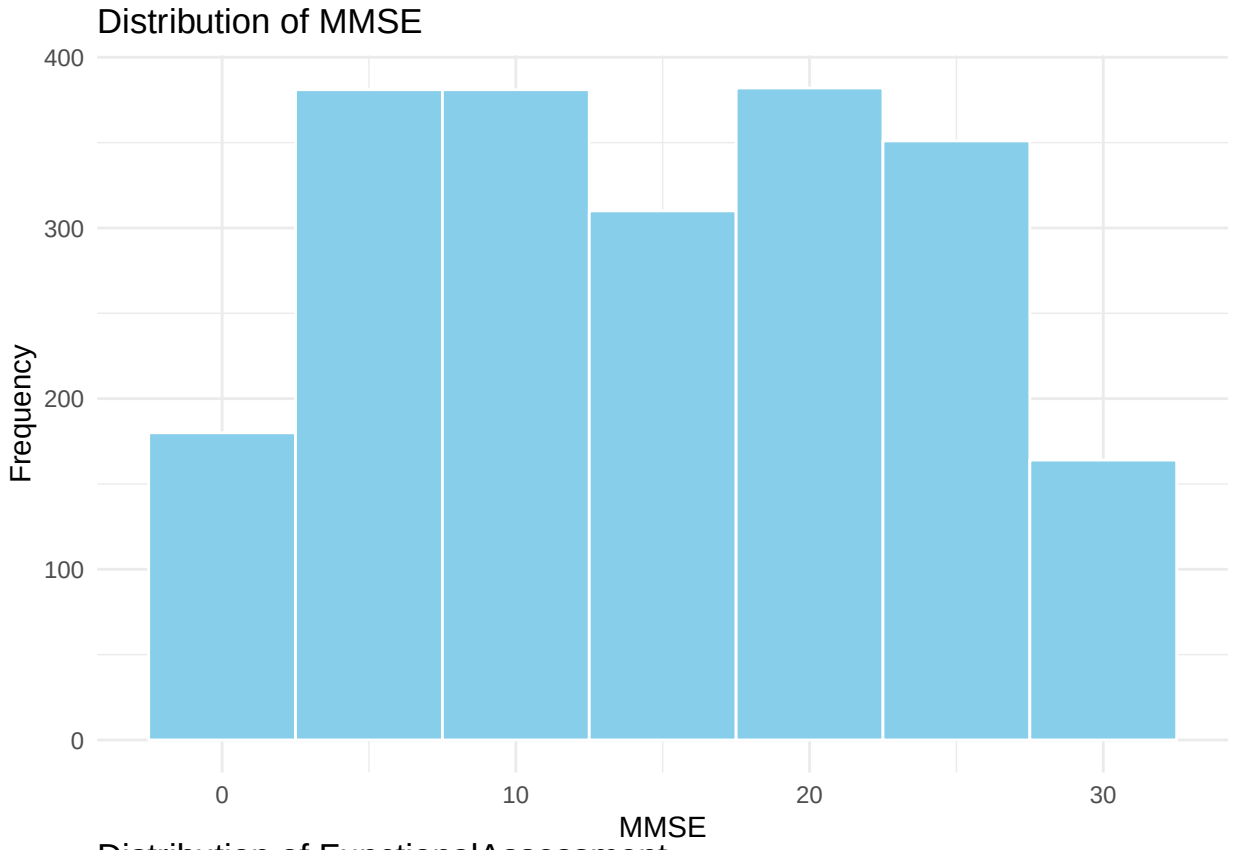

Distribution of DiastolicBP
Frequency
0
100
200
300
70
90
110
DiastolicBP

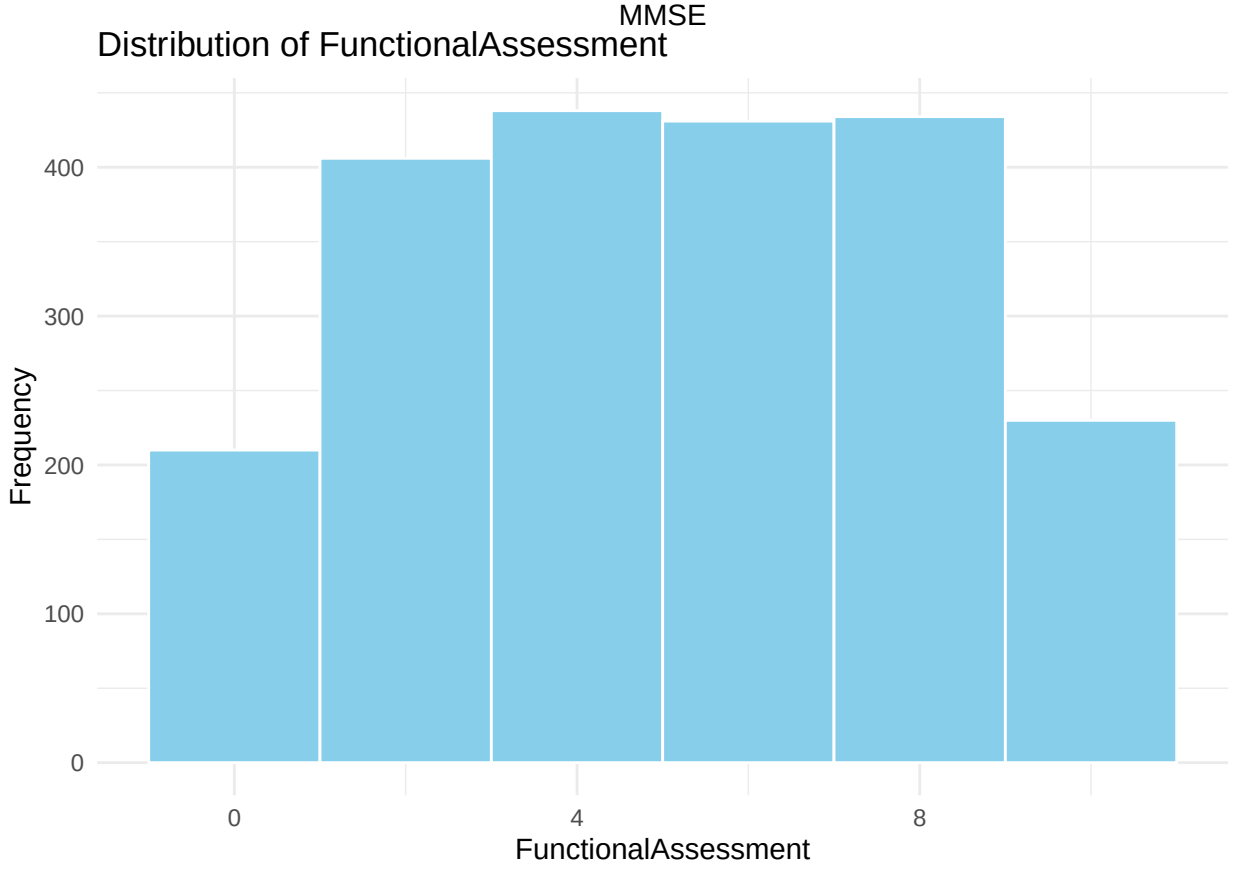

Distribution of CholesterolTotal
Frequency
0
100
200
300
400
150
200
250
300
CholesterolTotal

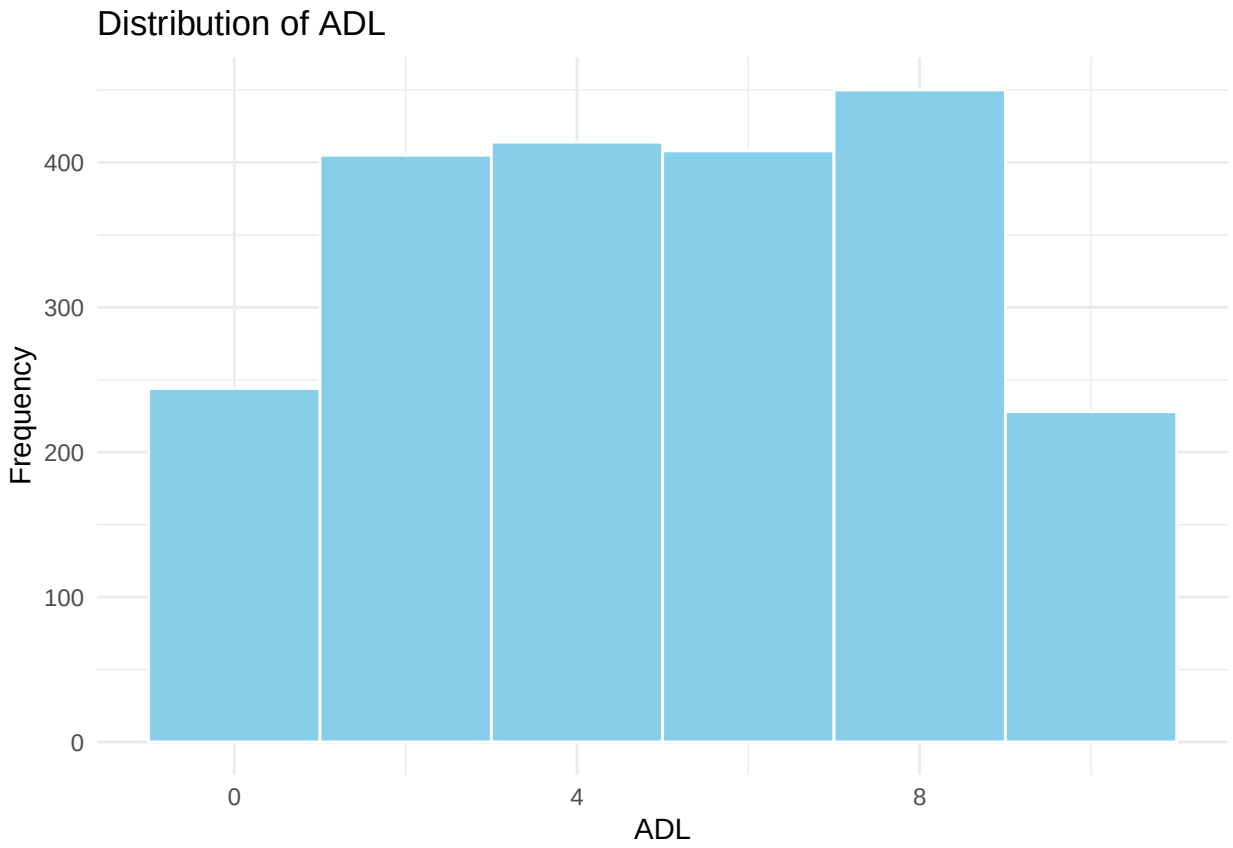

Distribution of CholesterolLDL
Frequency
0
100
200
300
400
50
100
150
200
CholesterolLDL


Distribution of CholesterolHDL
Frequency
300
200
100
0
25
50
75
100
CholesterolHDL


Distribution of CholesterolTriglycerides
Frequency
400
300
200
100
0
100
200
300
400
CholesterolTriglycerides


Distribution of MMSE
Frequency
400
300
200
100
0
0
10
20
30
MMSE


Distribution of FunctionalAssessment
Frequency
400
300
200
100
0
0
4
8
FunctionalAssessment



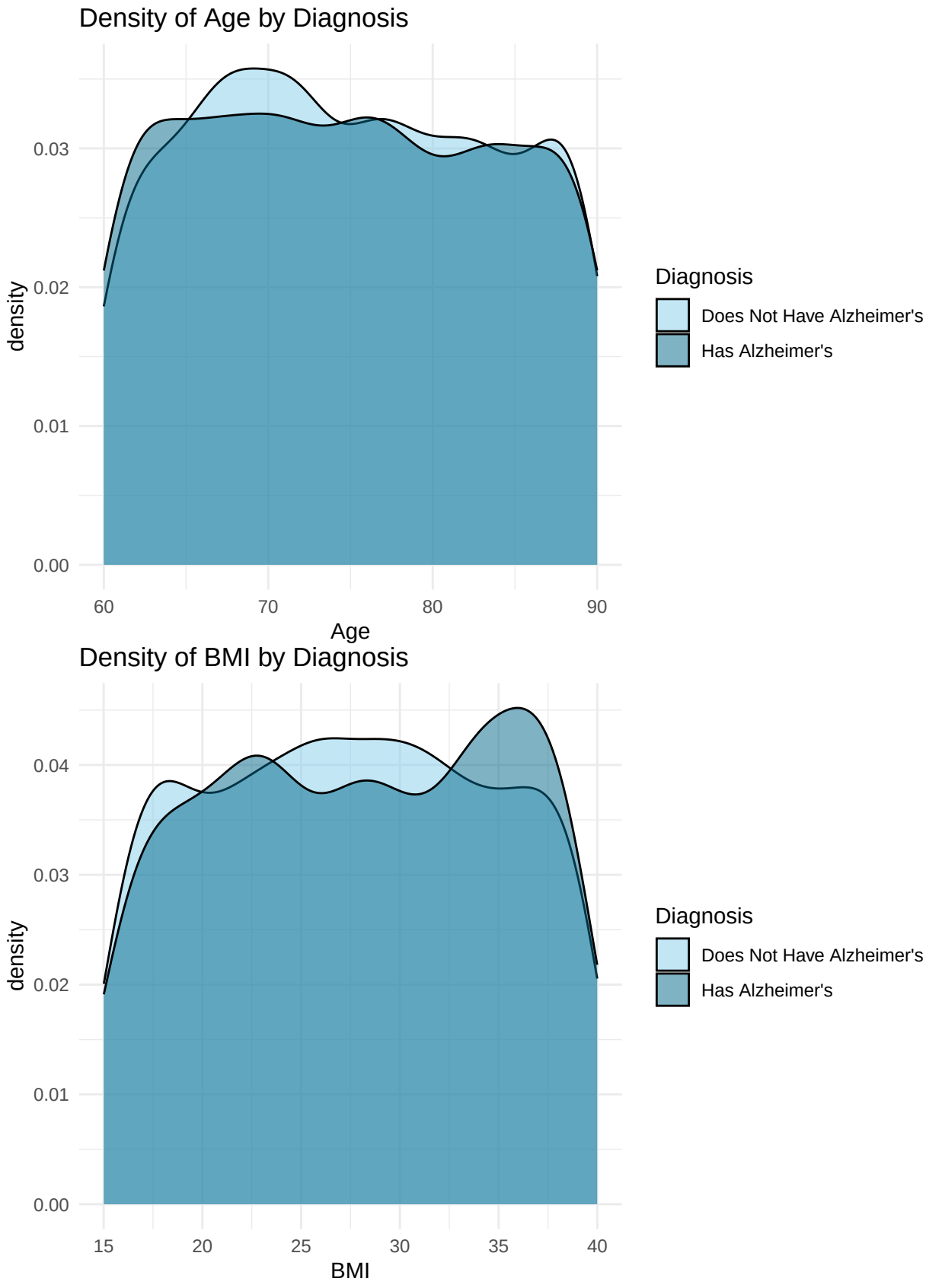
Histograms of numeric data are generated.

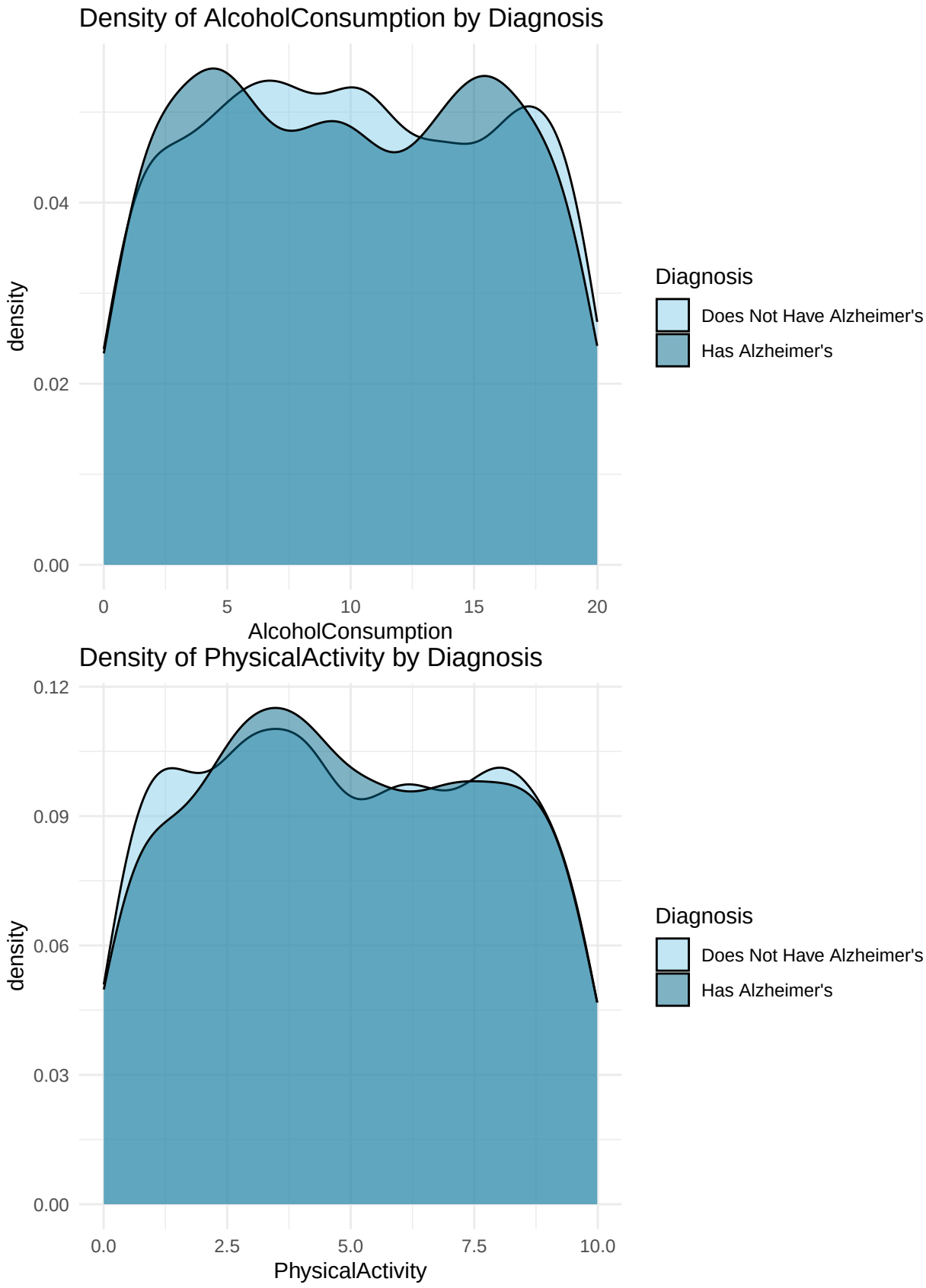
5-2. Distribution of numeric variables, based on diagnosis status

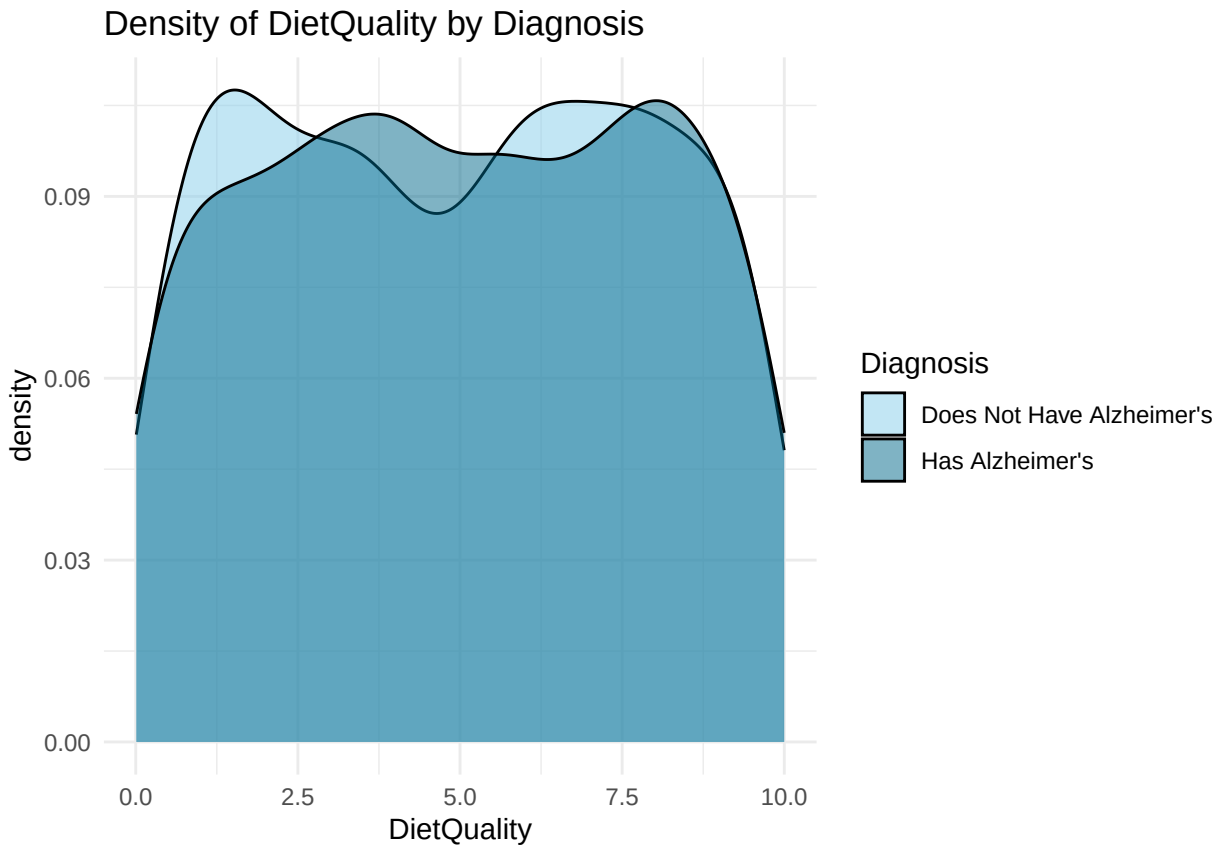
```
data_vis <- data %>%
  mutate(
    Diagnosis = factor(
      Diagnosis,
      levels = c(0, 1),
      labels = c("Does Not Have Alzheimer's", "Has Alzheimer's")
    )
  )

cat("\nDensity Plots of Continuous Variables by Diagnosis:\n")
```

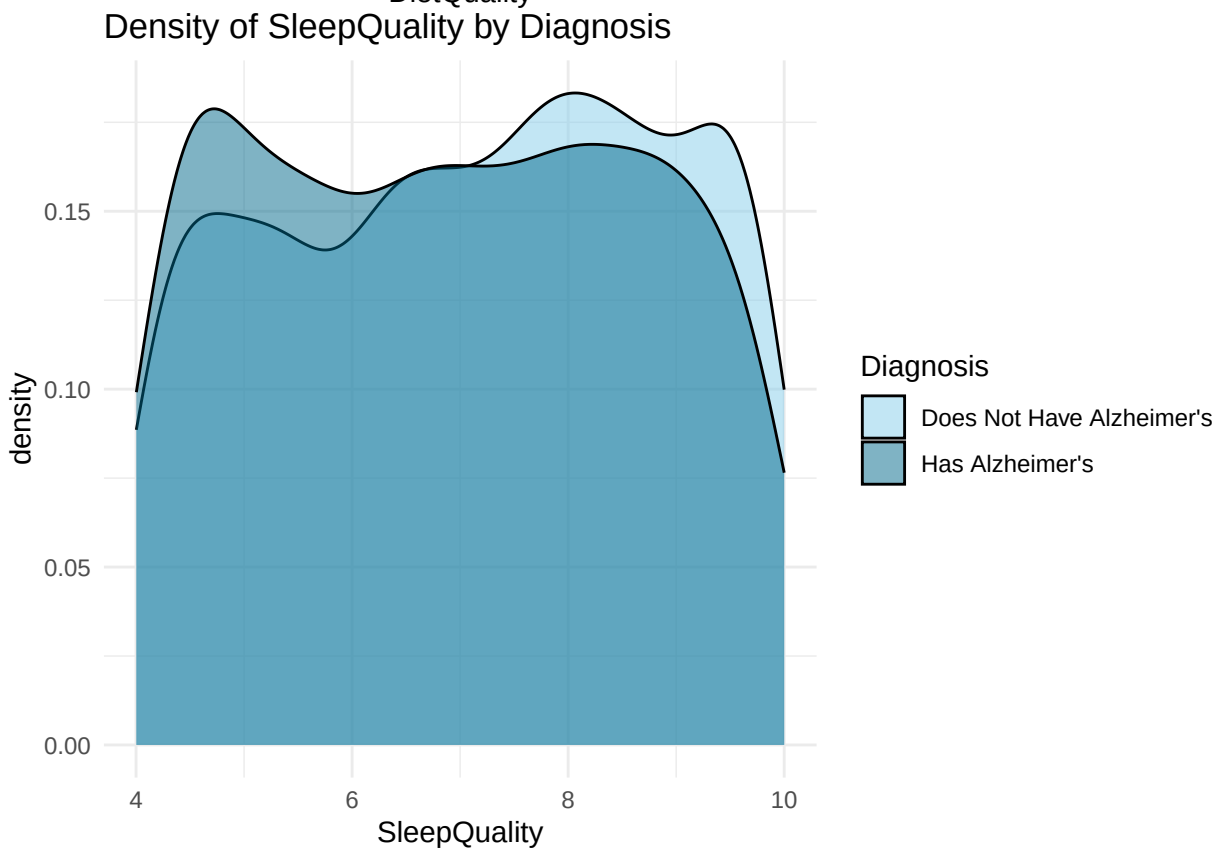
```
##
## Density Plots of Continuous Variables by Diagnosis:
```

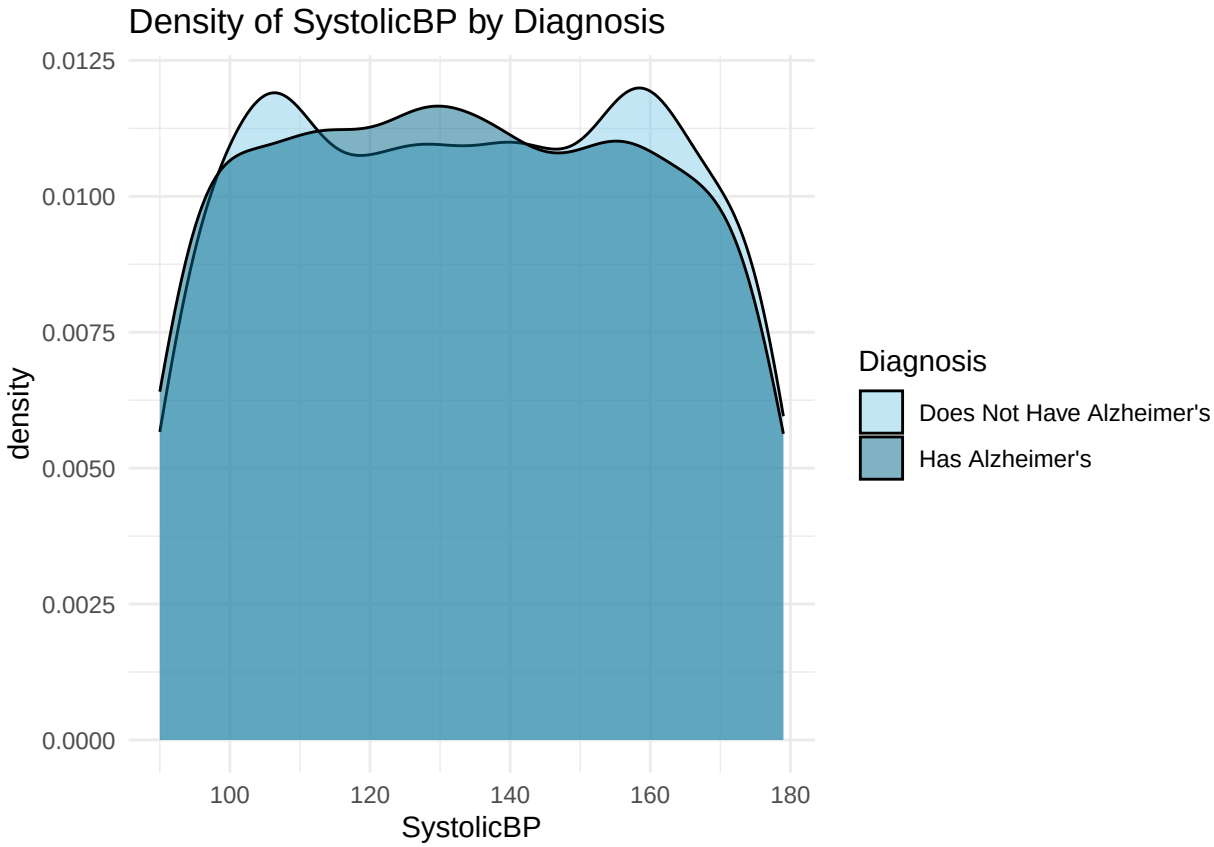
```
for (col in colnames(continuous_columns)) {
  p <- ggplot(data_vis, aes(x = .data[[col]], fill = Diagnosis)) +
    geom_density(alpha = 0.5) +
    labs(title = paste("Density of", col, "by Diagnosis"), x = col) +
    scale_fill_manual(values = c("skyblue", "deepskyblue4")) +
    theme_minimal()
  print(p)
}
```

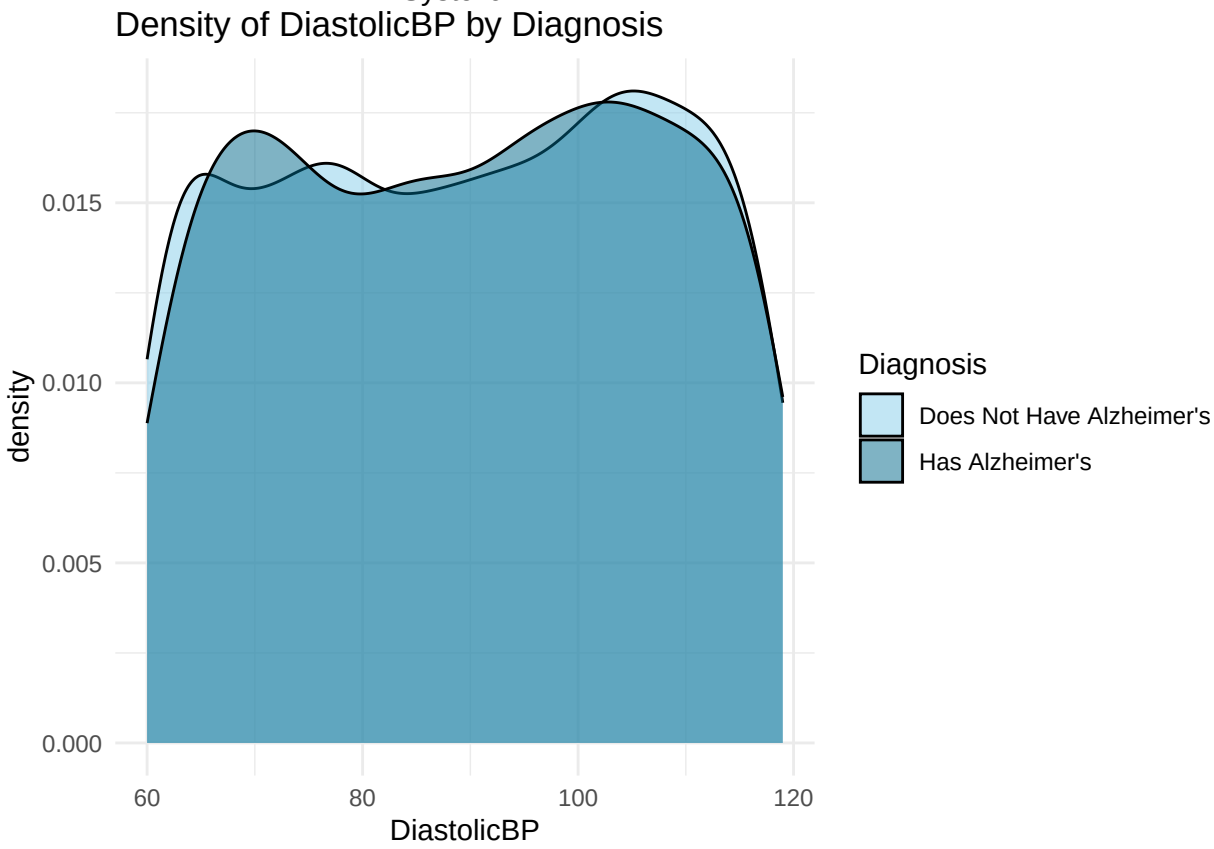

Density of Age by Diagnosis
density
0.03
0.02
0.01
0.00
60
70
80
90
Age
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's
Density of BMI by Diagnosis
density
0.04
0.03
0.02
0.01
0.00
15
20
25
30
35
40
BMI
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

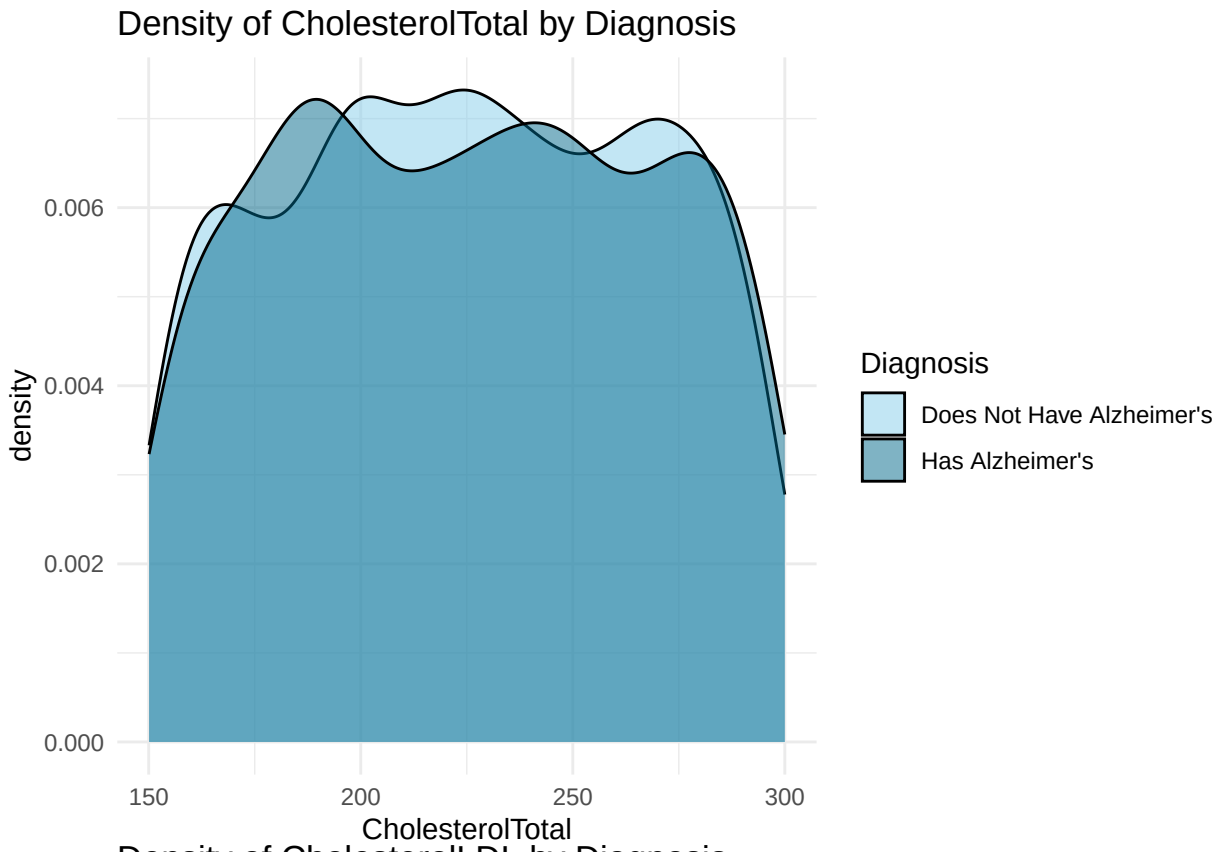

Density of AlcoholConsumption by Diagnosis
density
0.04
0.02
0.00
0
5
10
15
20
AlcoholConsumption
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's
Density of PhysicalActivity by Diagnosis
density
0.12
0.09
0.06
0.03
0.00
0.0
2.5
5.0
7.5
10.0
PhysicalActivity
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

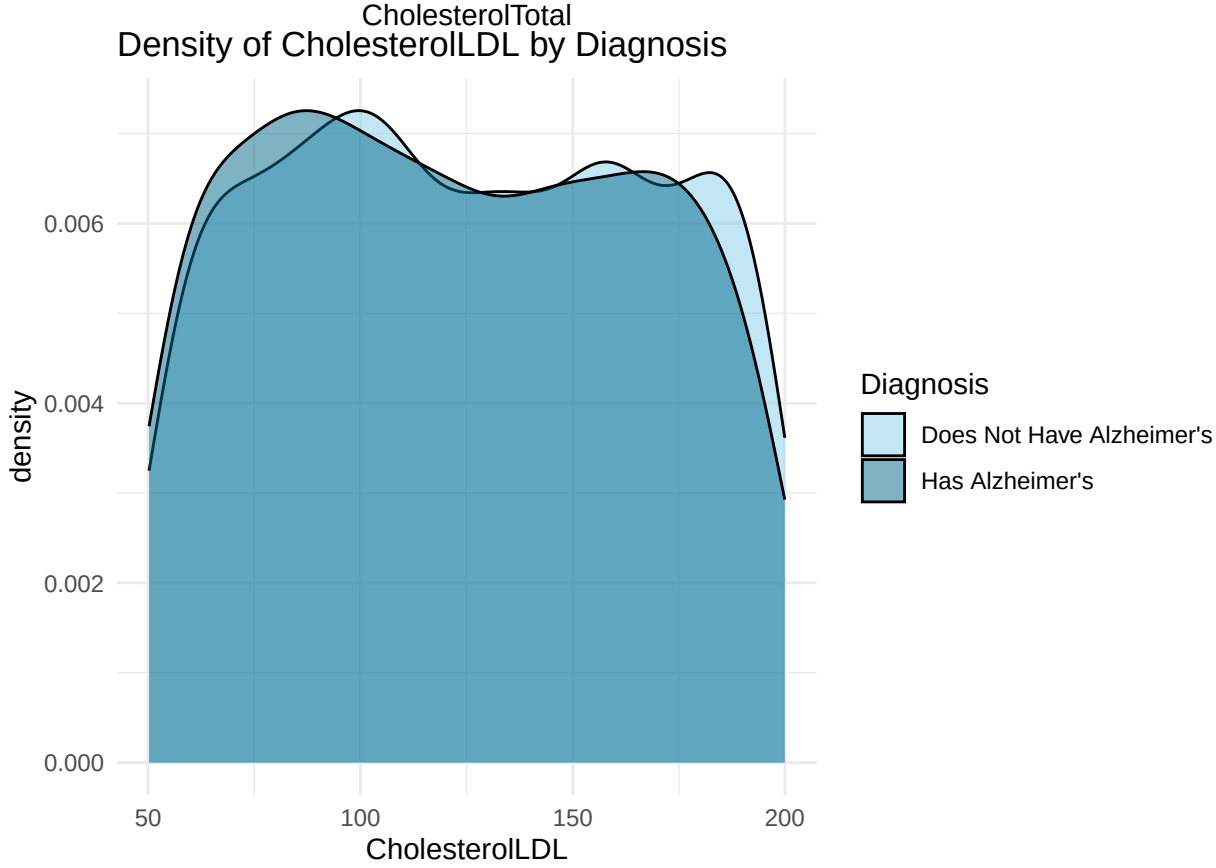

Density of DietQuality by Diagnosis
density
0.09
0.06
0.03
0.00
0.0
2.5
5.0
7.5
10.0
DietQuality
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

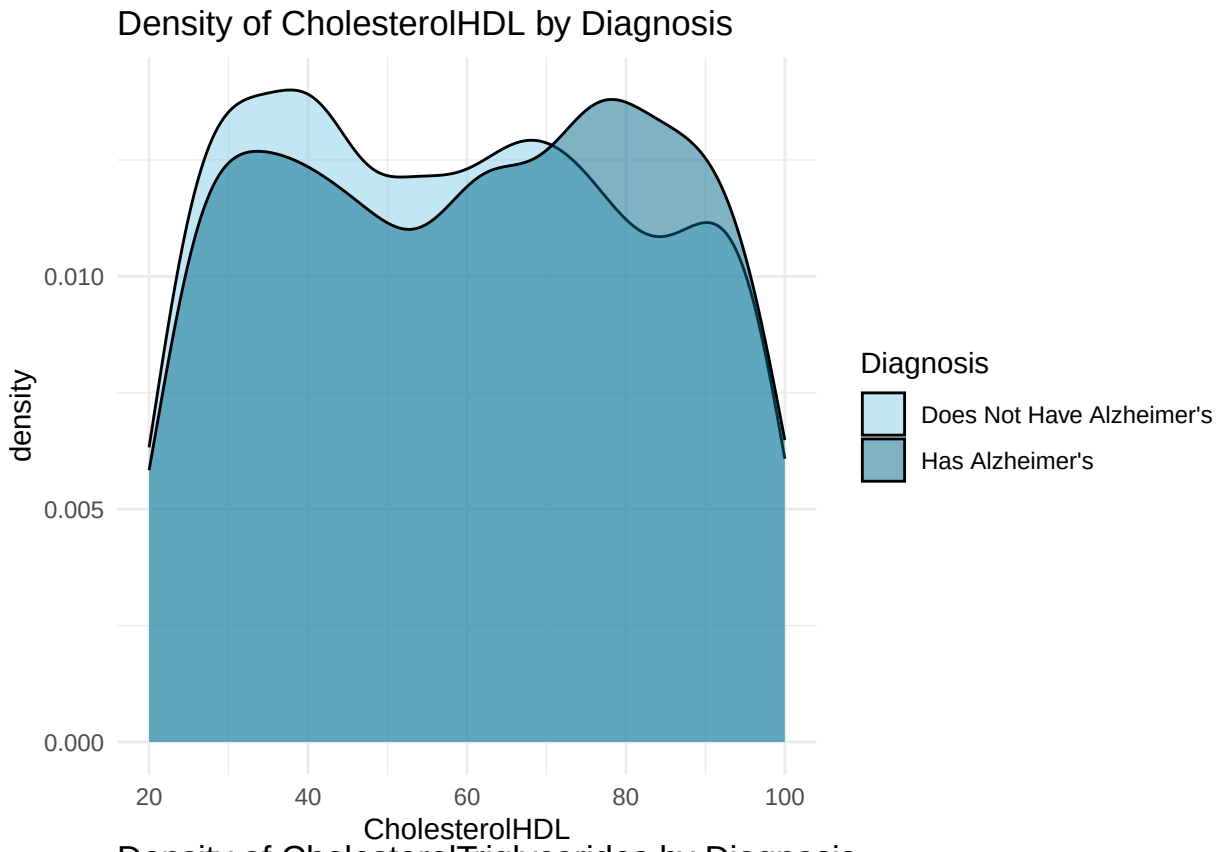

Density of SleepQuality by Diagnosis
density
0.15
0.10
0.05
0.00
4
6
8
10
SleepQuality
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

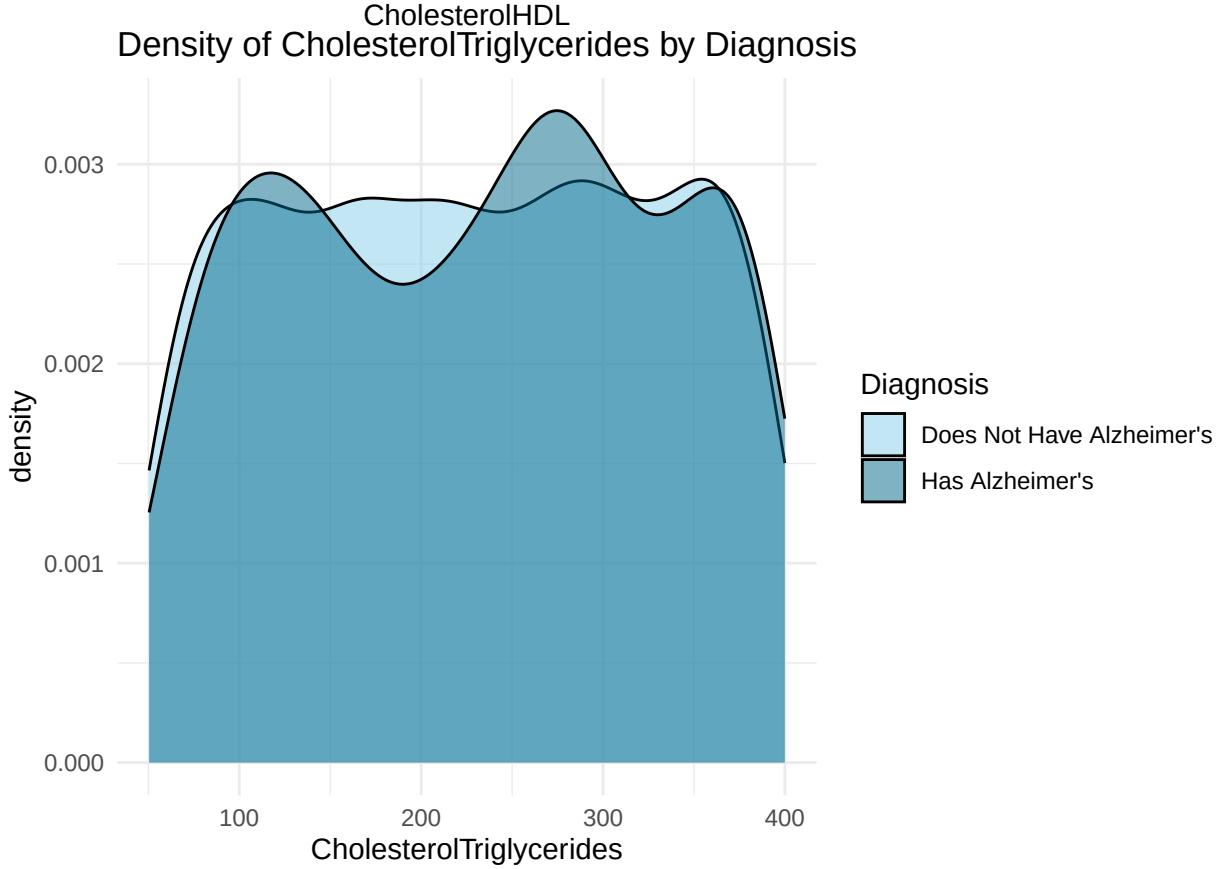

Density of SystolicBP by Diagnosis
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
density
100
120
140
160
180
SystolicBP
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

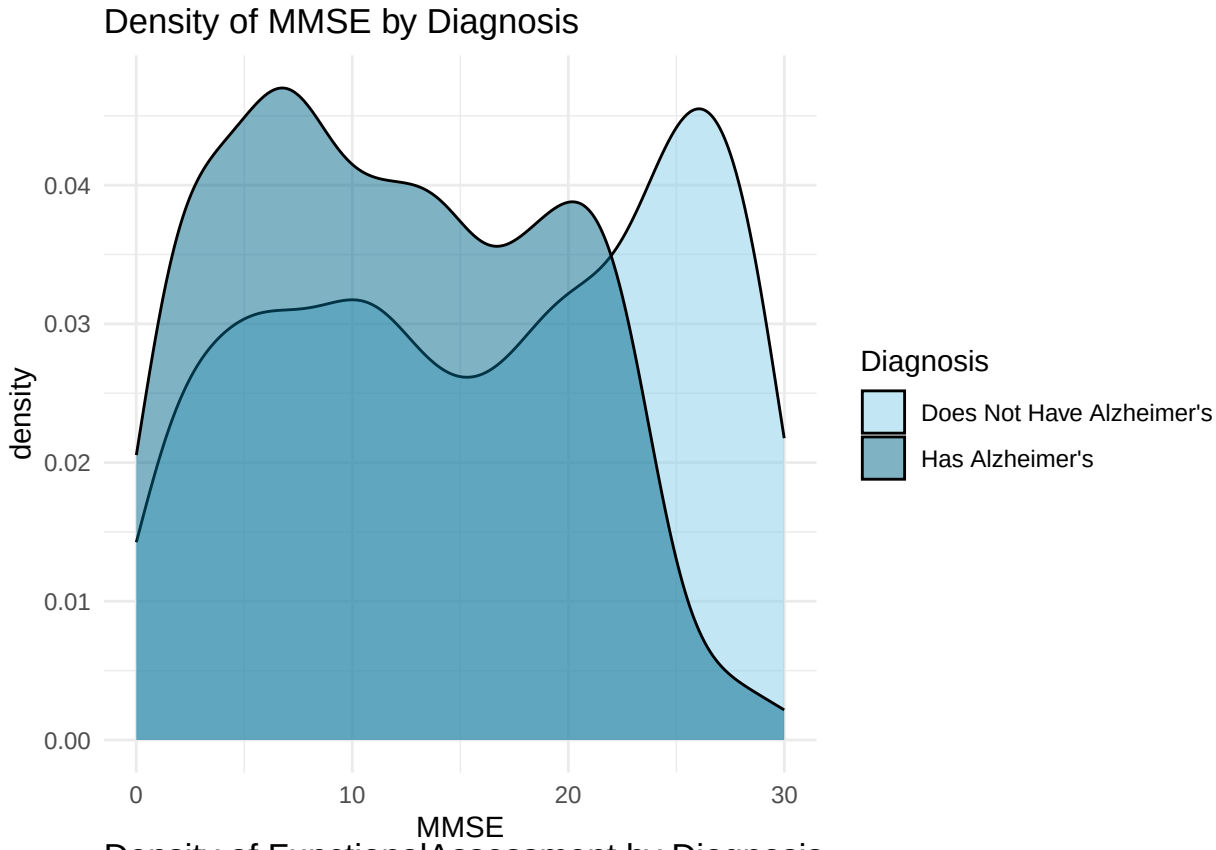

Density of DiastolicBP by Diagnosis
0.015
0.010
0.005
0.000
density
60
80
100
120
DiastolicBP
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

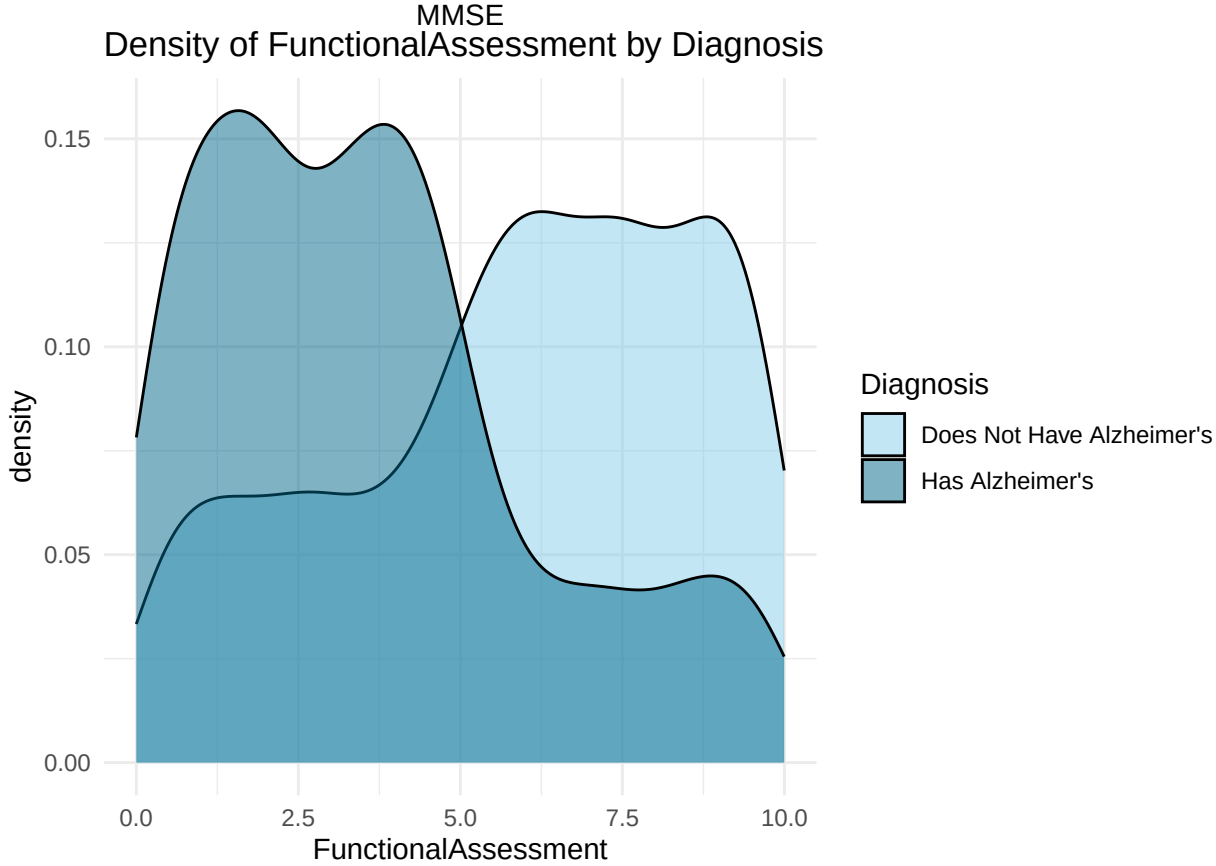

Density of CholesterolTotal by Diagnosis
density
0.006
0.004
0.002
0.000
150
200
250
300
CholesterolTotal
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's

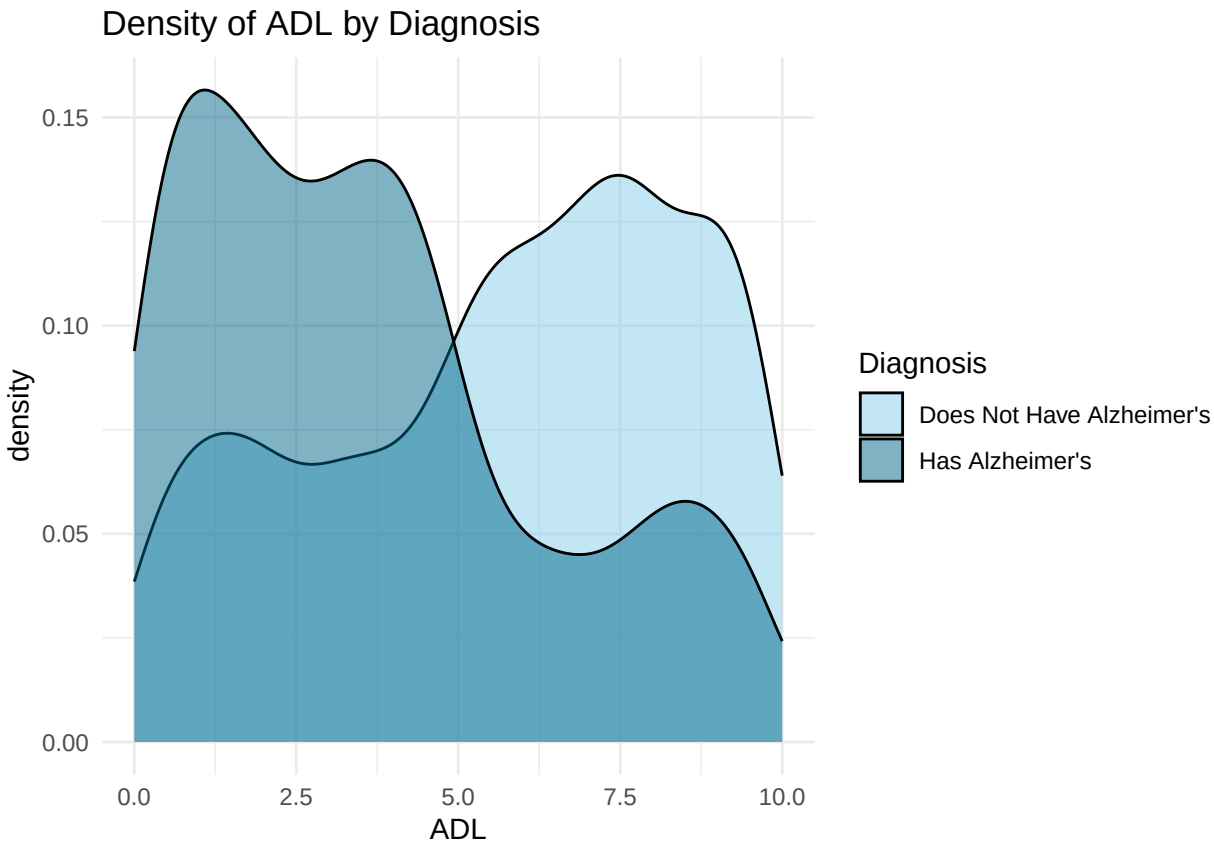

Density of CholesterolLDL by Diagnosis
density
0.006
0.004
0.002
0.000
50
100
150
200
CholesterolLDL
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's


Density of CholesterolHDL by Diagnosis
density
0.010
0.005
0.000
20
40
60
80
100
CholesterolHDL
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's


Density of CholesterolTriglycerides by Diagnosis
density
0.003
0.002
0.001
0.000
100
200
300
400
CholesterolTriglycerides
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's


Density of MMSE by Diagnosis
density
0.04
0.03
0.02
0.01
0.00
0
10
20
30
MMSE
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's


Density of FunctionalAssessment by Diagnosis
density
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
FunctionalAssessment
Diagnosis
Does Not Have Alzheimer's
Has Alzheimer's



For better visualization, distribution of numeric variables are re-generated based on diagnosis status of Alzheimer's disease.

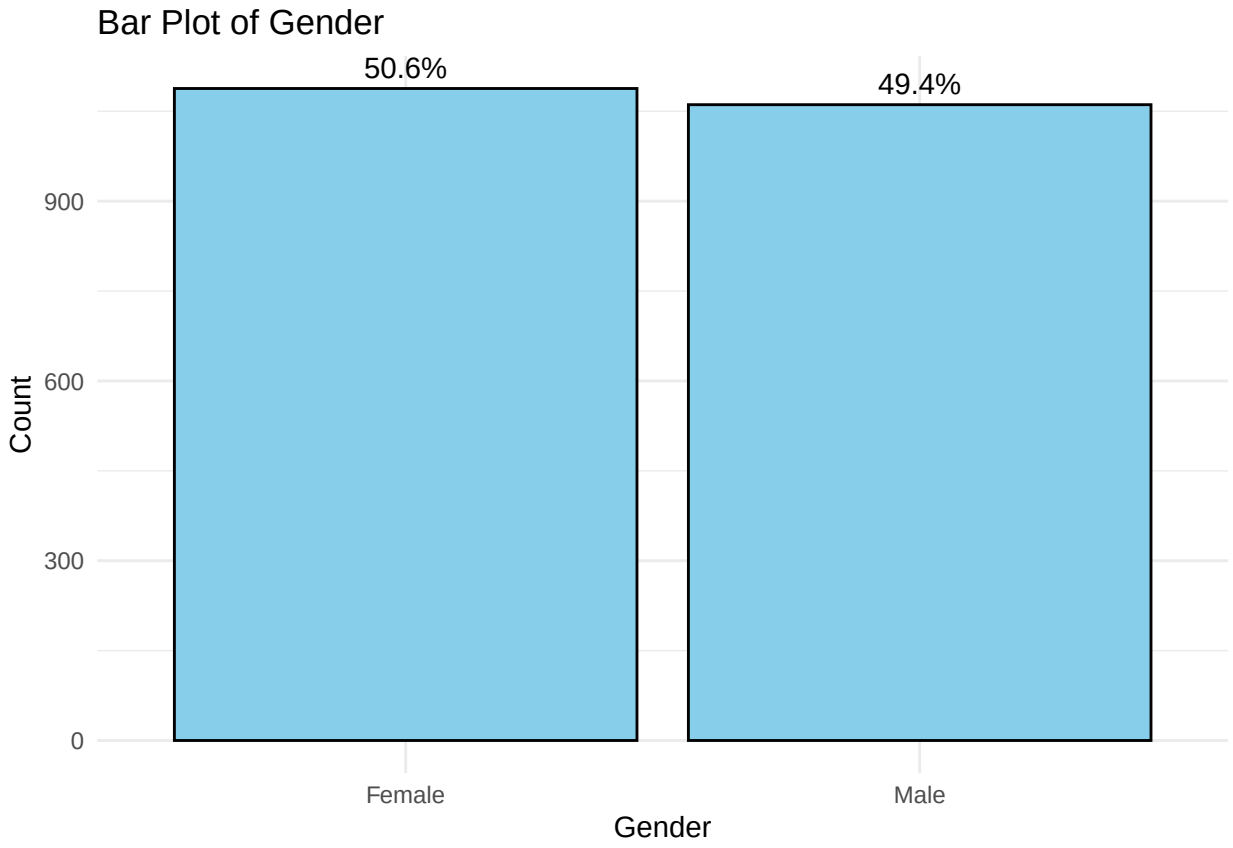
6-1. Distribution of categorical variables

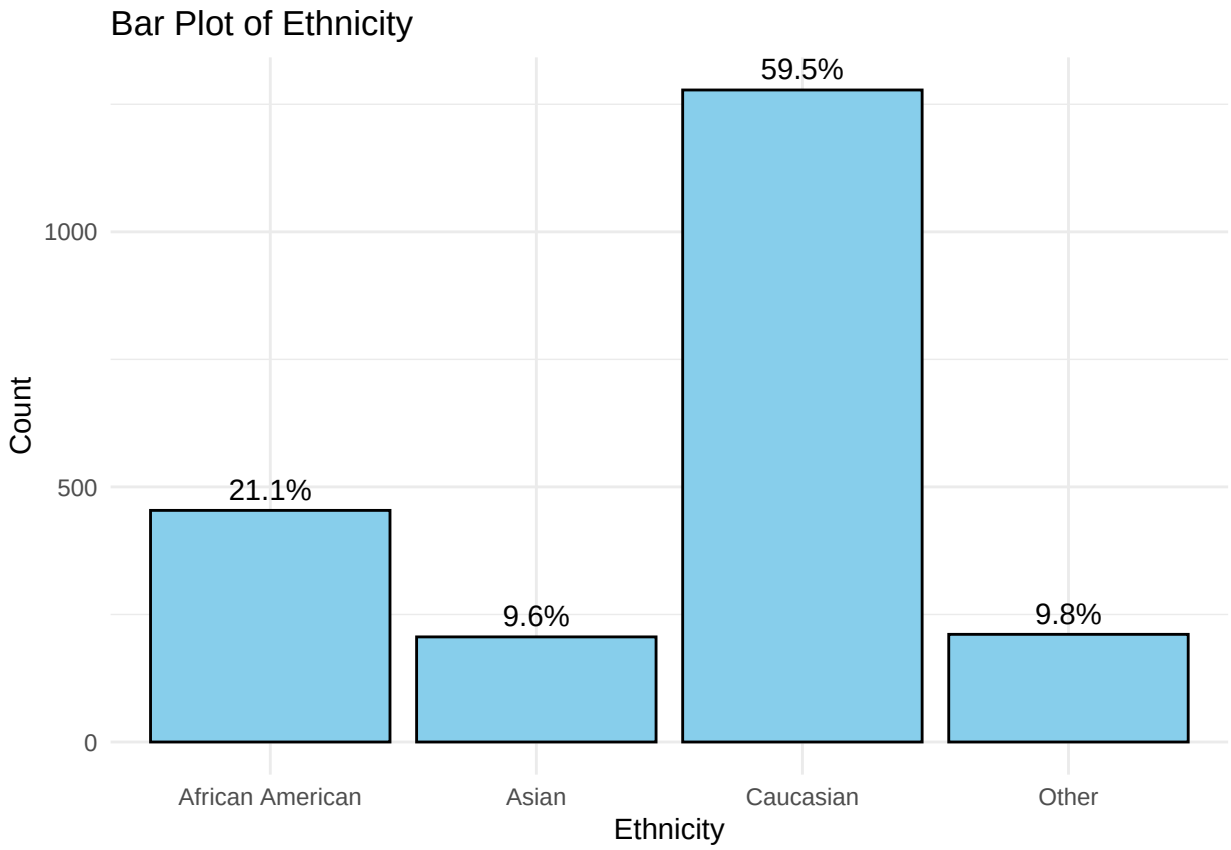
```
cat("\nHistograms of Categorical Variables:\n")
```

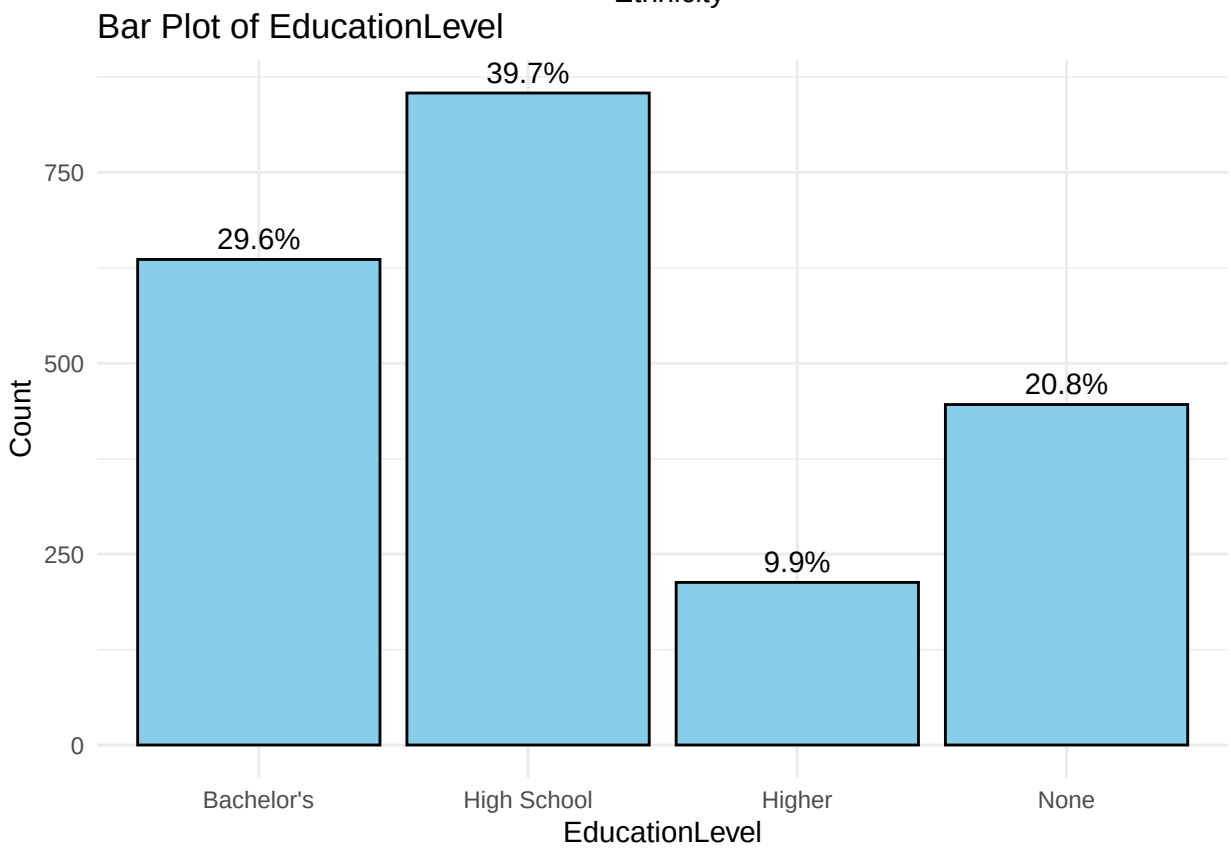
```
##
## Histograms of Categorical Variables:
```

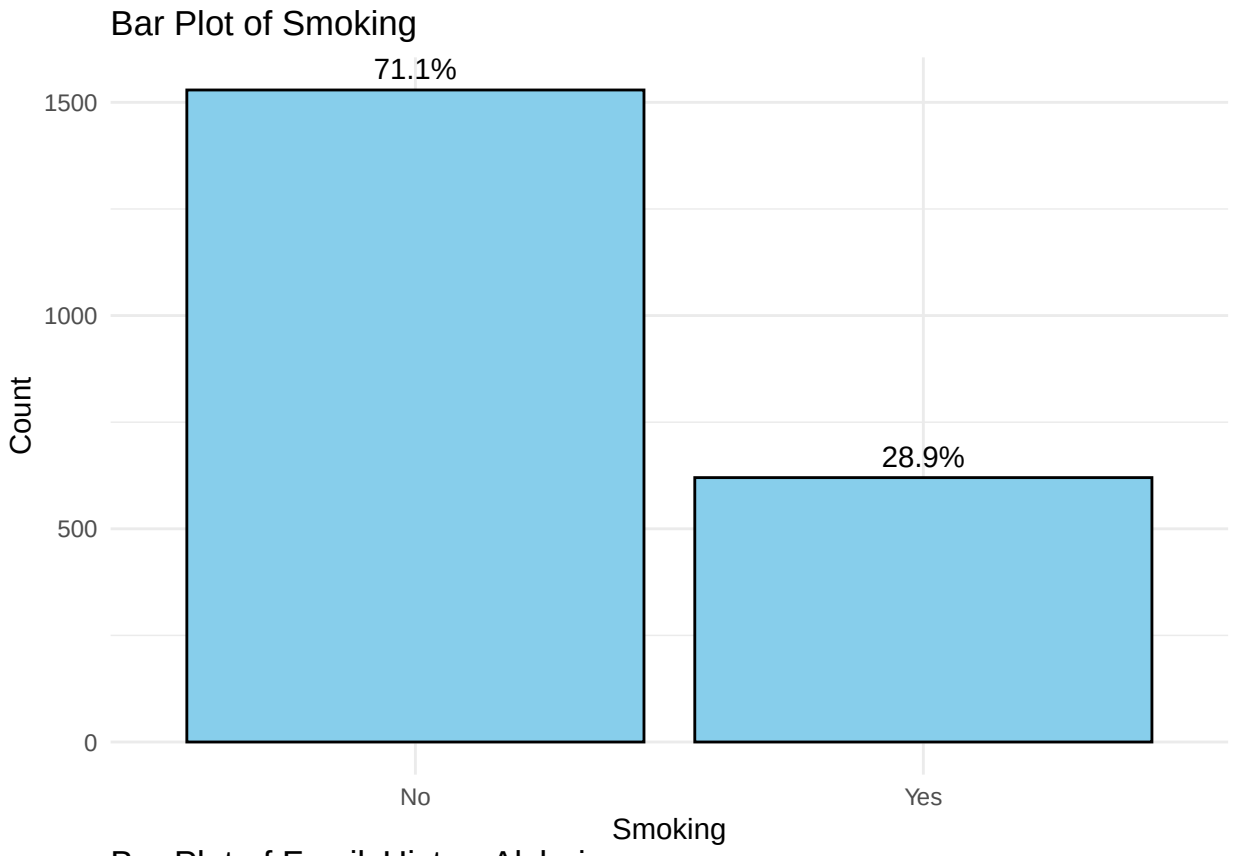
```
categorical_variables <- data[sapply(data, is.factor)]
categorical_variables$Gender <- ifelse(categorical_variables$Gender==0,'Male','Female')
categorical_variables <- categorical_variables %>%
  mutate(Ethnicity = case_when(
    Ethnicity == 0 ~ 'Caucasian',
    Ethnicity == 1 ~ 'African American',
    Ethnicity == 2 ~ 'Asian',
    TRUE ~ 'Other'),
    EducationLevel = case_when(
      EducationLevel == 0 ~ 'None',
      EducationLevel == 1 ~ 'High School',
      EducationLevel == 2 ~ "Bachelor's",
      TRUE ~ 'Higher')
  )

categorical_variables[,4:18] <- ifelse(categorical_variables[4:18]==0,'No','Yes')
```

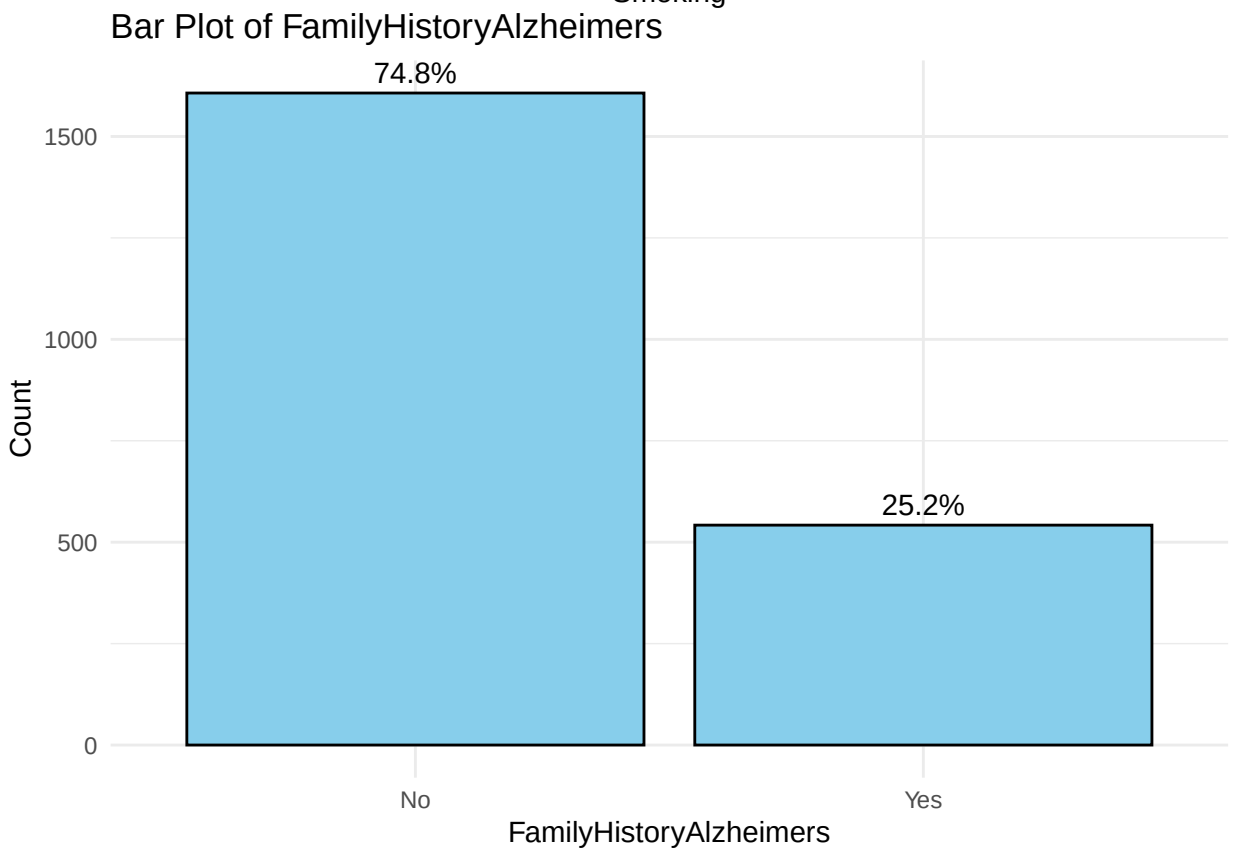
```r
for (c in 1:ncol(categorical_variables)) {
  counts <- categorical_variables %>%
    count(!!sym(colnames(categorical_variables)[c])) %>%
    mutate(Percentage = n / sum(n) * 100)
  p <- ggplot(counts, aes_string(x=colnames(categorical_variables)[c], y = "n")) +
    geom_bar(stat = "identity", fill = "skyblue", color = "black") +
    geom_text(aes(label = paste0(round(Percentage,1),"%")), vjust = -0.5) +
    labs(title = paste("Bar Plot of", colnames(categorical_variables)[c]),
         x = colnames(categorical_variables)[c],
         y = "Count") + theme_minimal()
  print(p)
}
```

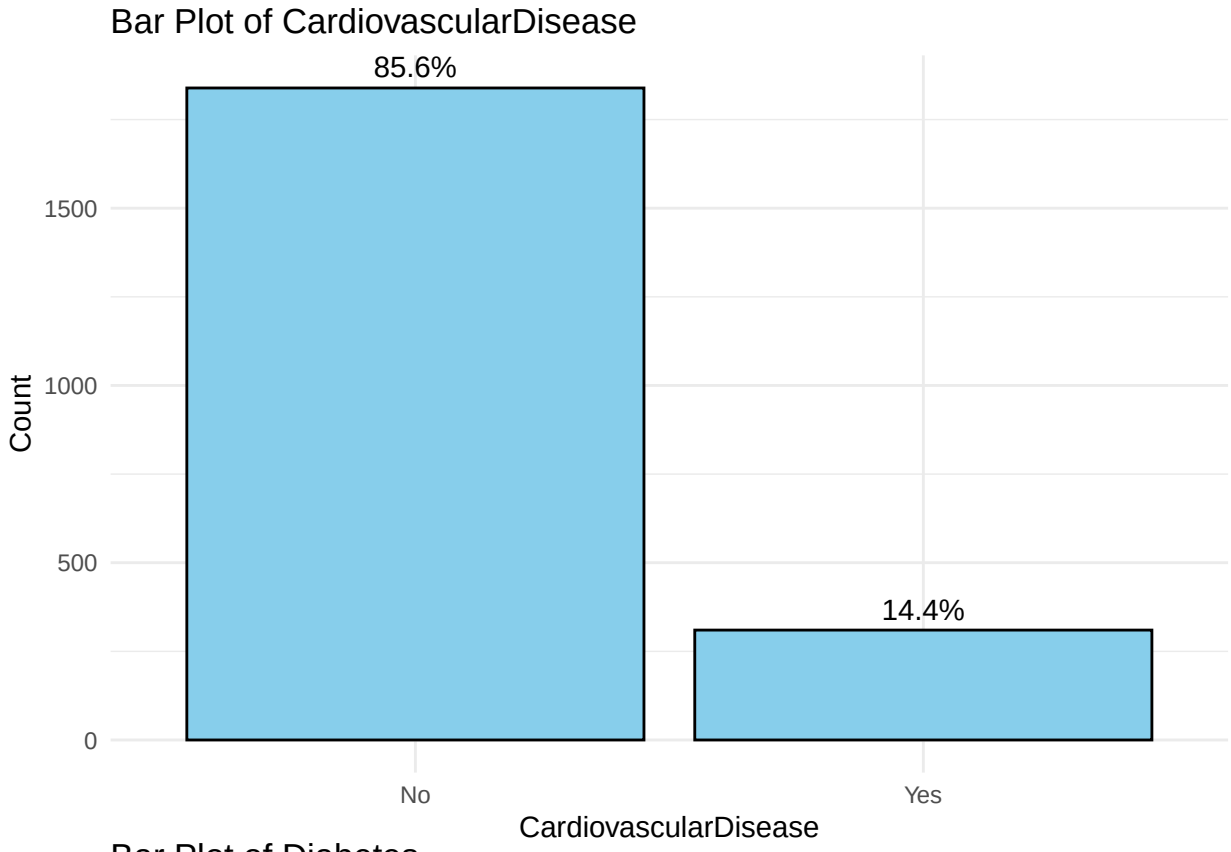


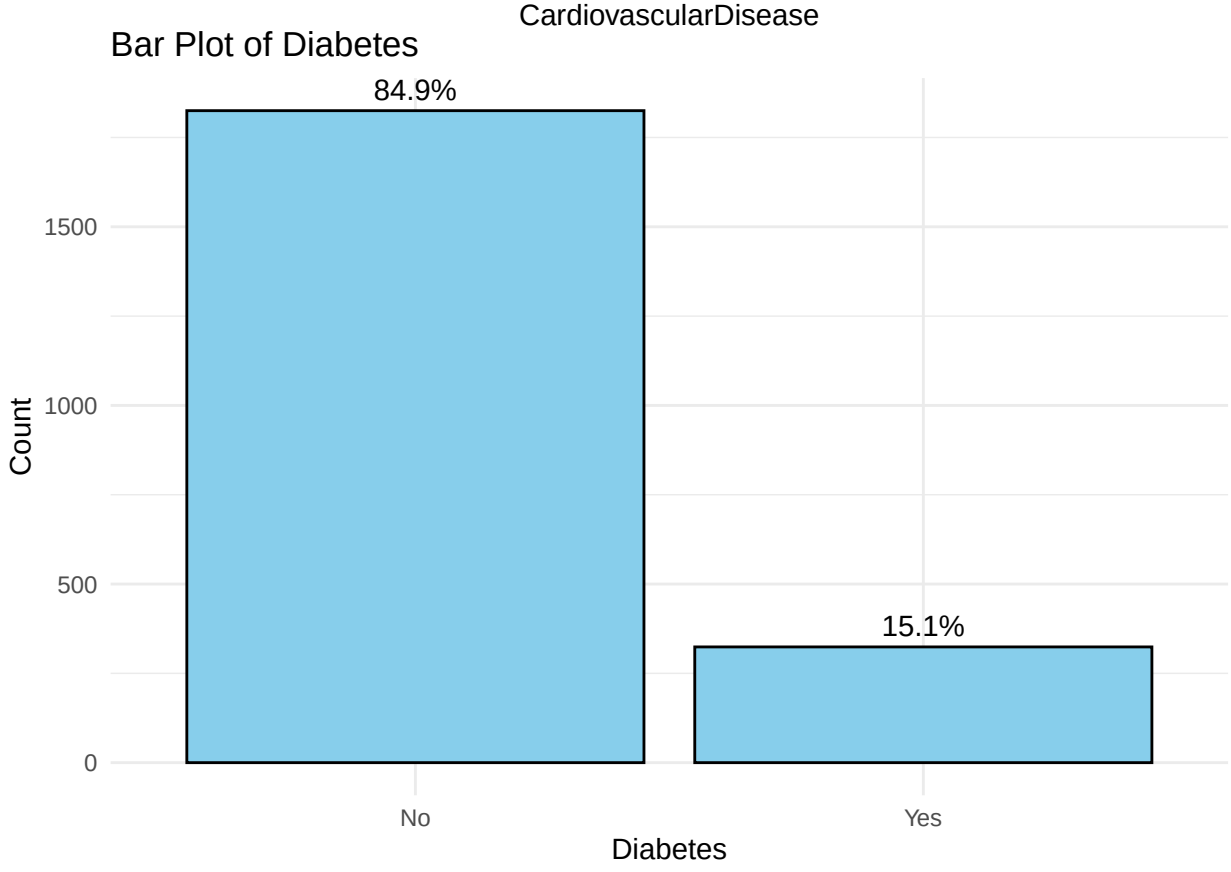

Bar Plot of Ethnicity
59.5%
1000
Count
500
21.1%
9.6%
9.8%
0
African American
Asian
Caucasian
Other
Ethnicity

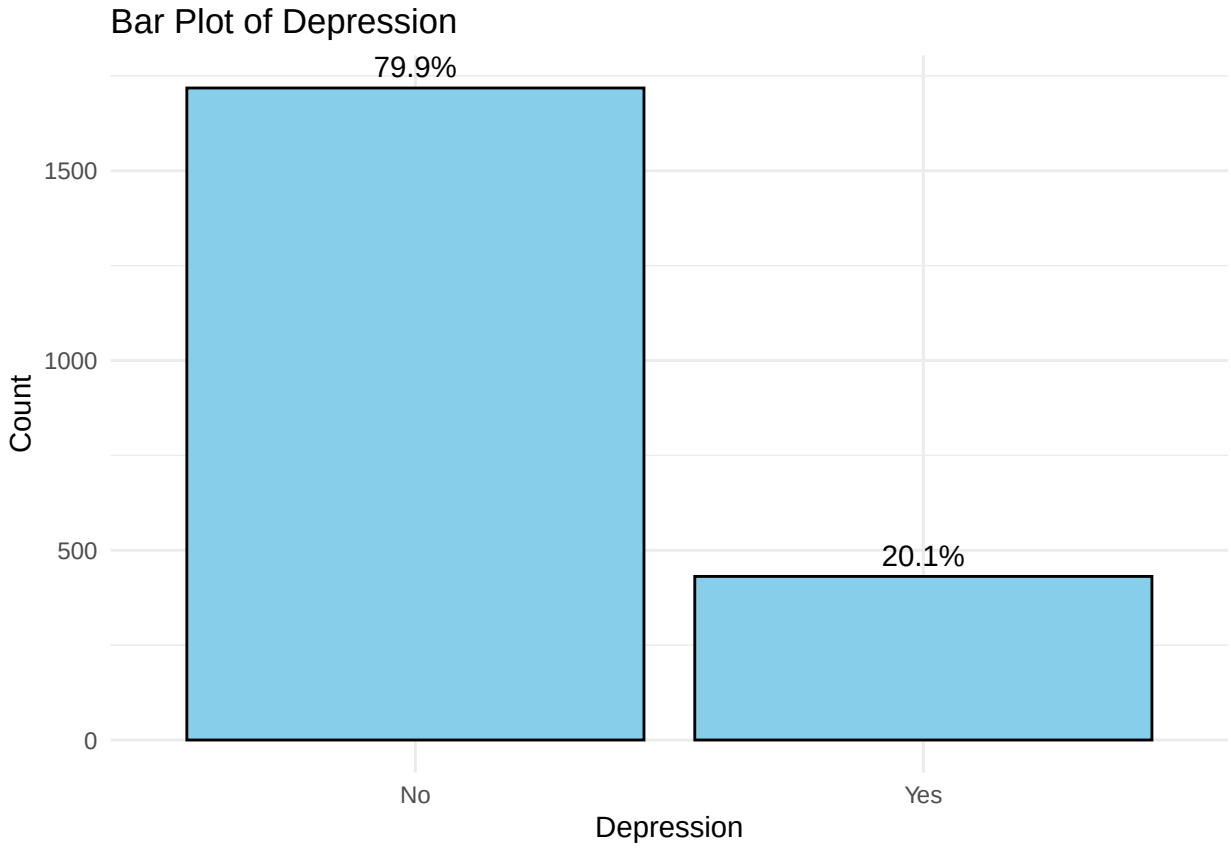

Bar Plot of EducationLevel
39.7%
750
29.6%
Count
500
20.8%
250
9.9%
0
Bachelor's
High School
Higher
None
EducationLevel

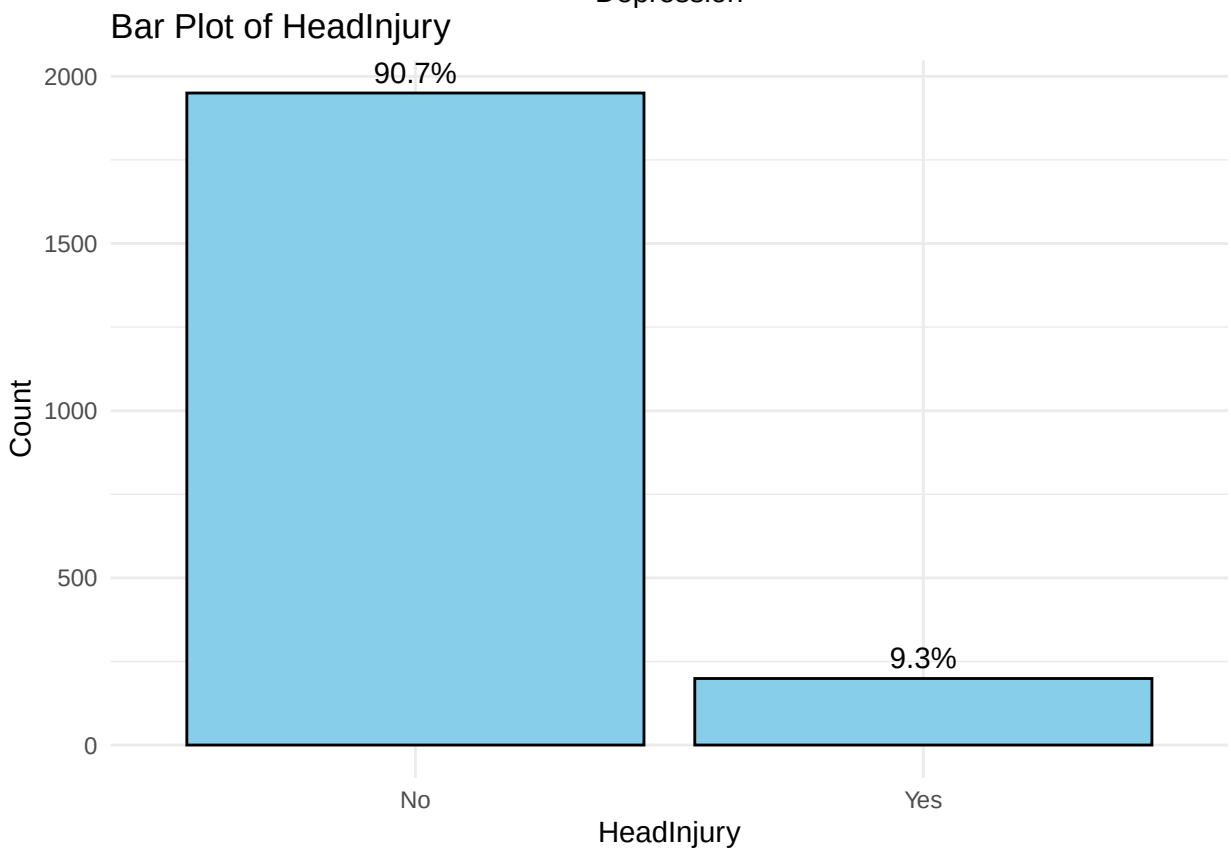

Bar Plot of Smoking
71.1%
28.9%
1500
1000
500
0
Count
No
Yes
Smoking

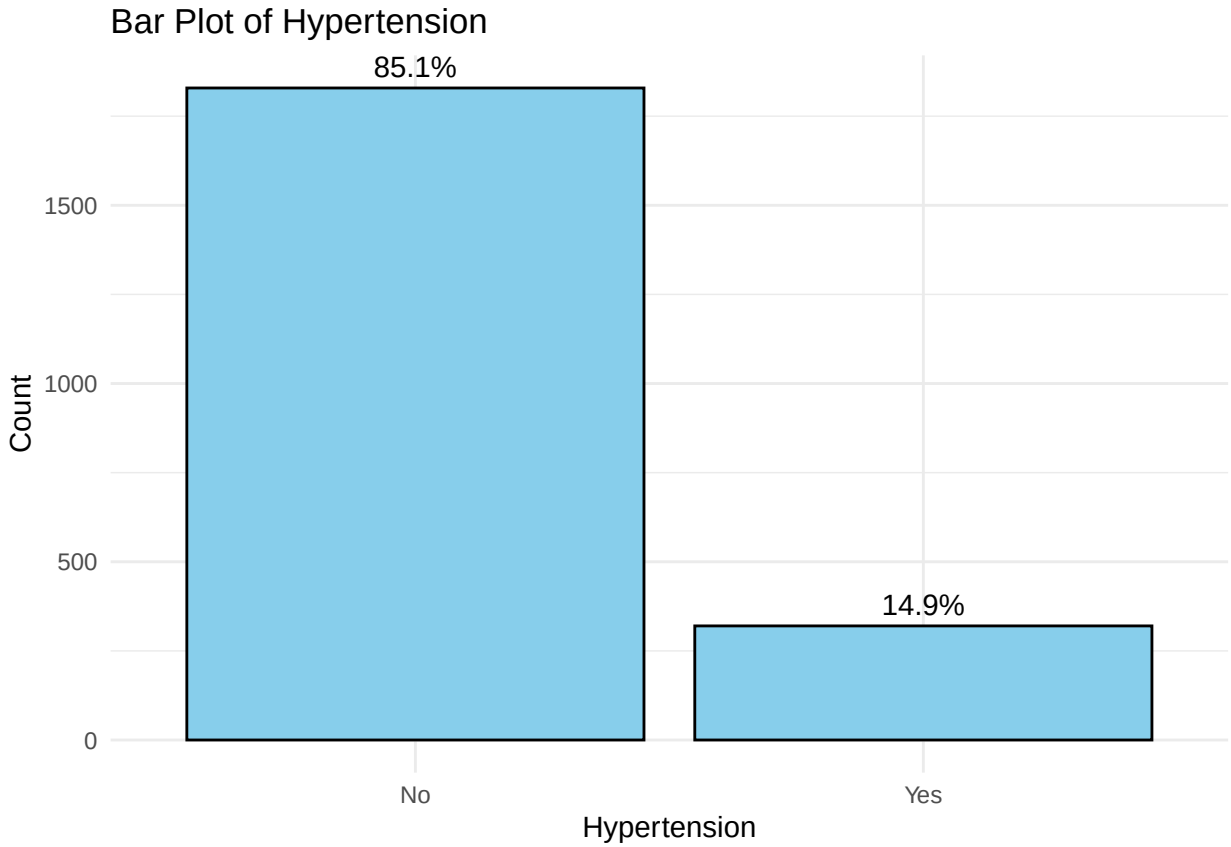

Bar Plot of FamilyHistoryAlzheimers
74.8%
25.2%
1500
1000
500
0
Count
No
Yes
FamilyHistoryAlzheimers

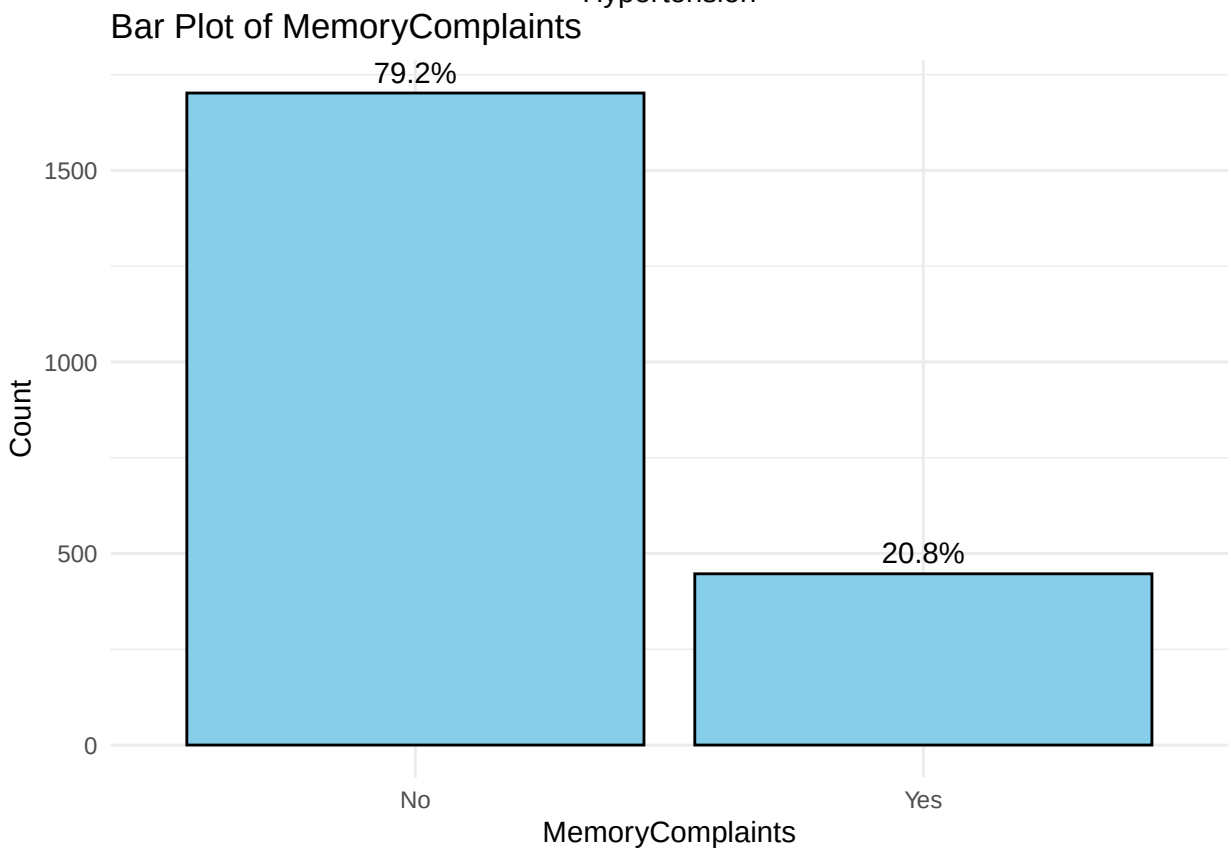

Bar Plot of CardiovascularDisease
85.6%
14.4%
1500
1000
500
0
Count
No
Yes
CardiovascularDisease

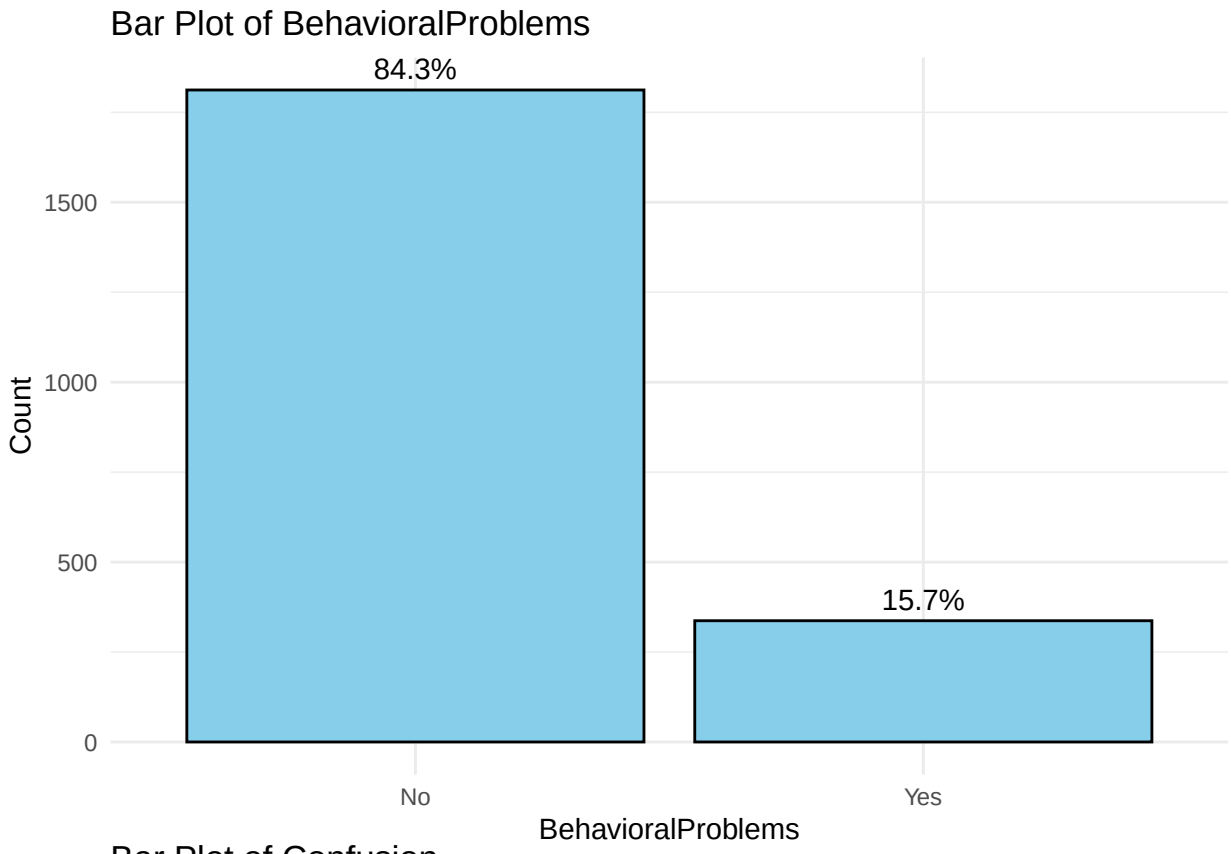

Bar Plot of Diabetes
84.9%
15.1%
1500
1000
500
0
Count
No
Yes
Diabetes

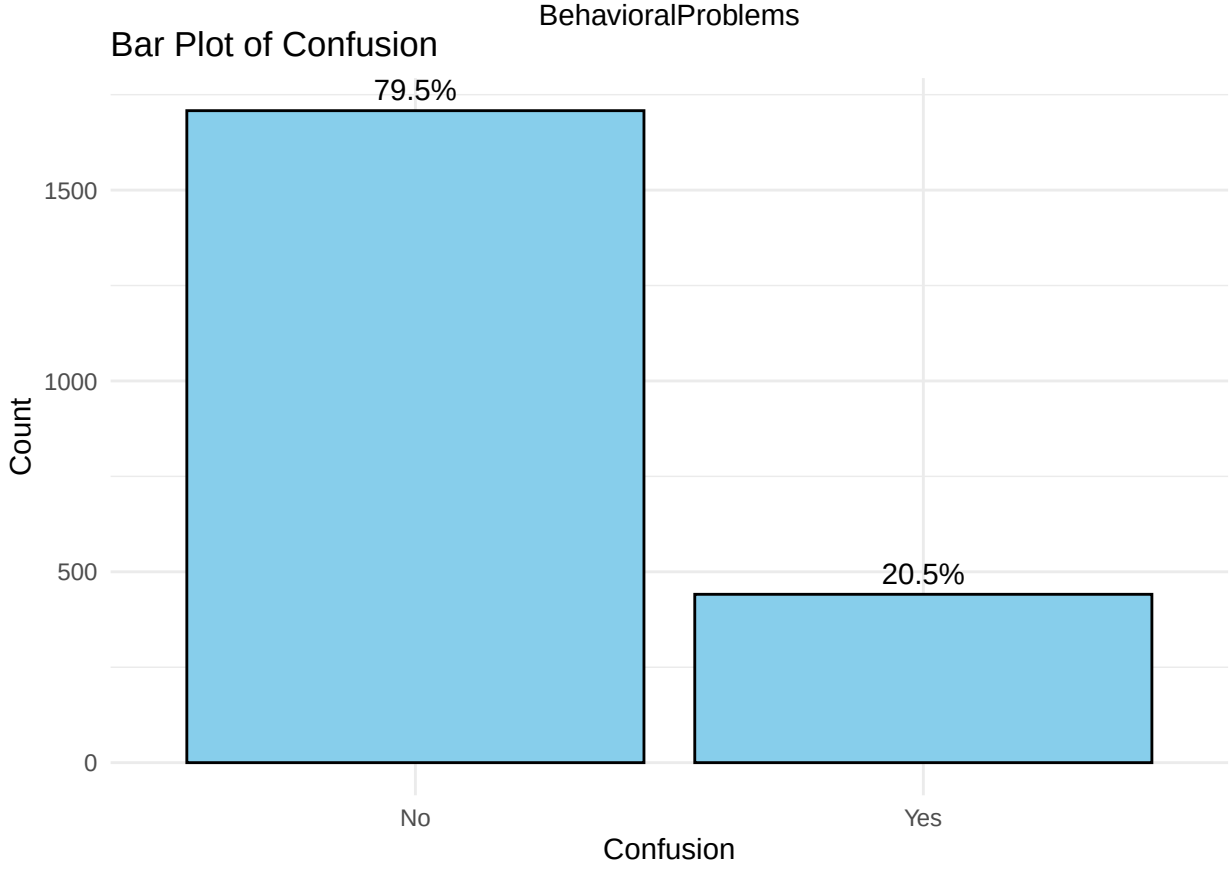

Bar Plot of Depression
79.9%
20.1%
1500
1000
500
0
Count
No
Yes
Depression

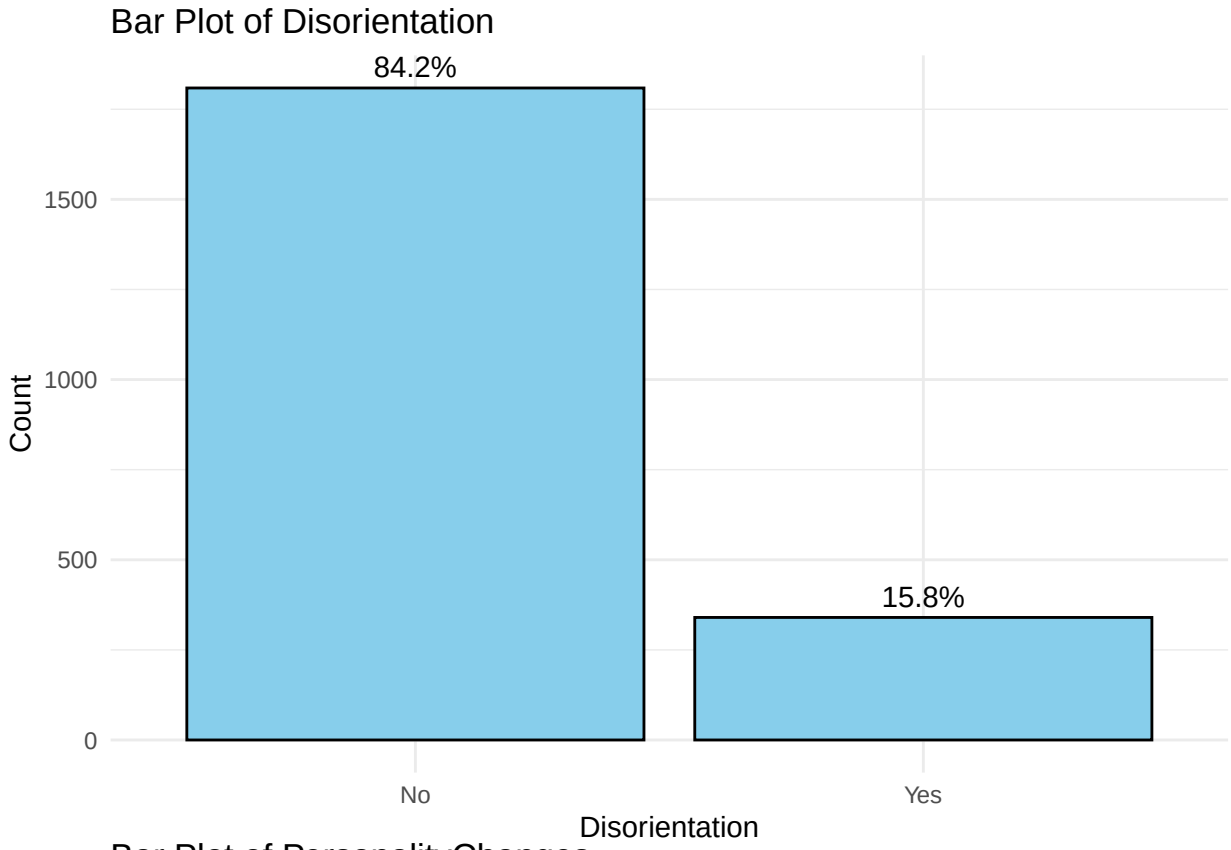

Bar Plot of HeadInjury
90.7%
9.3%
2000
1500
1000
500
0
Count
No
Yes
HeadInjury

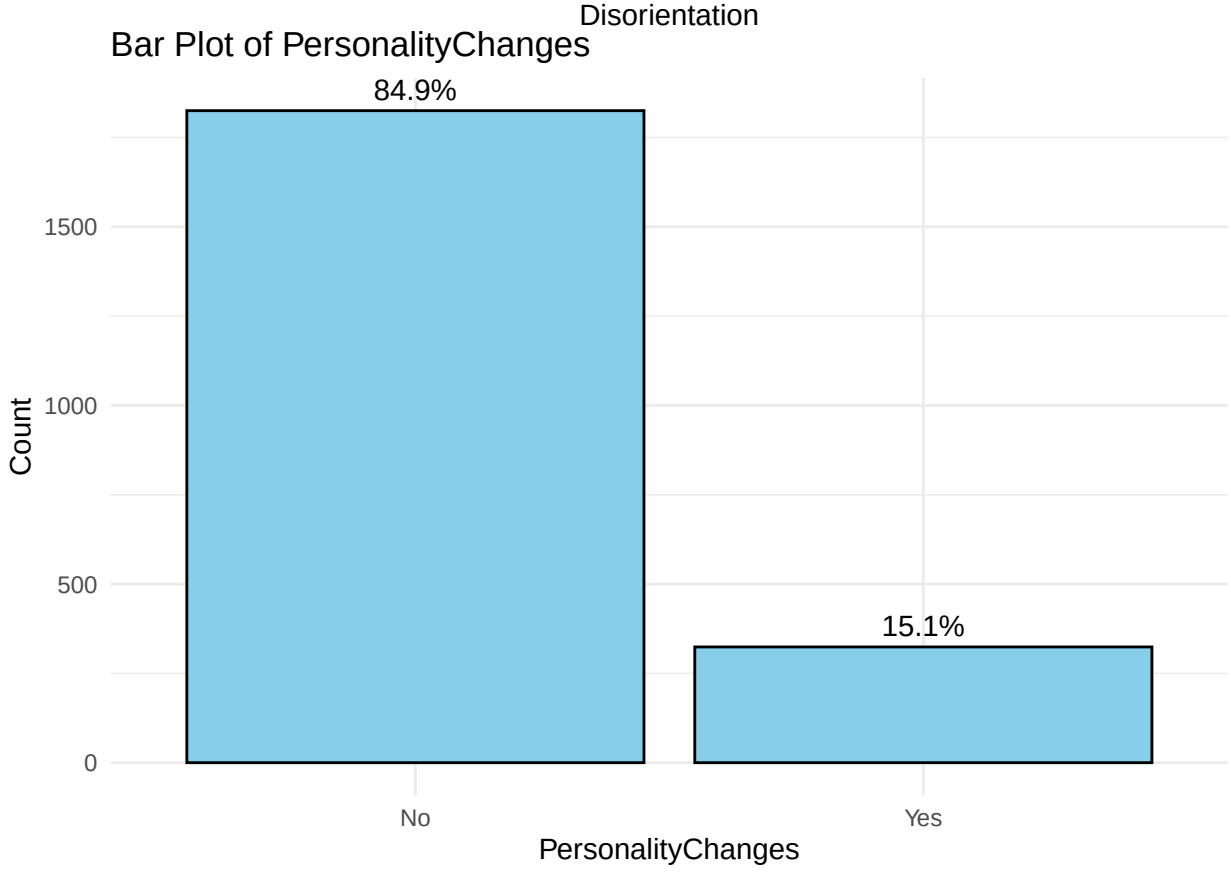

Bar Plot of Hypertension
85.1%
14.9%
1500
1000
500
0
Count
No
Yes
Hypertension

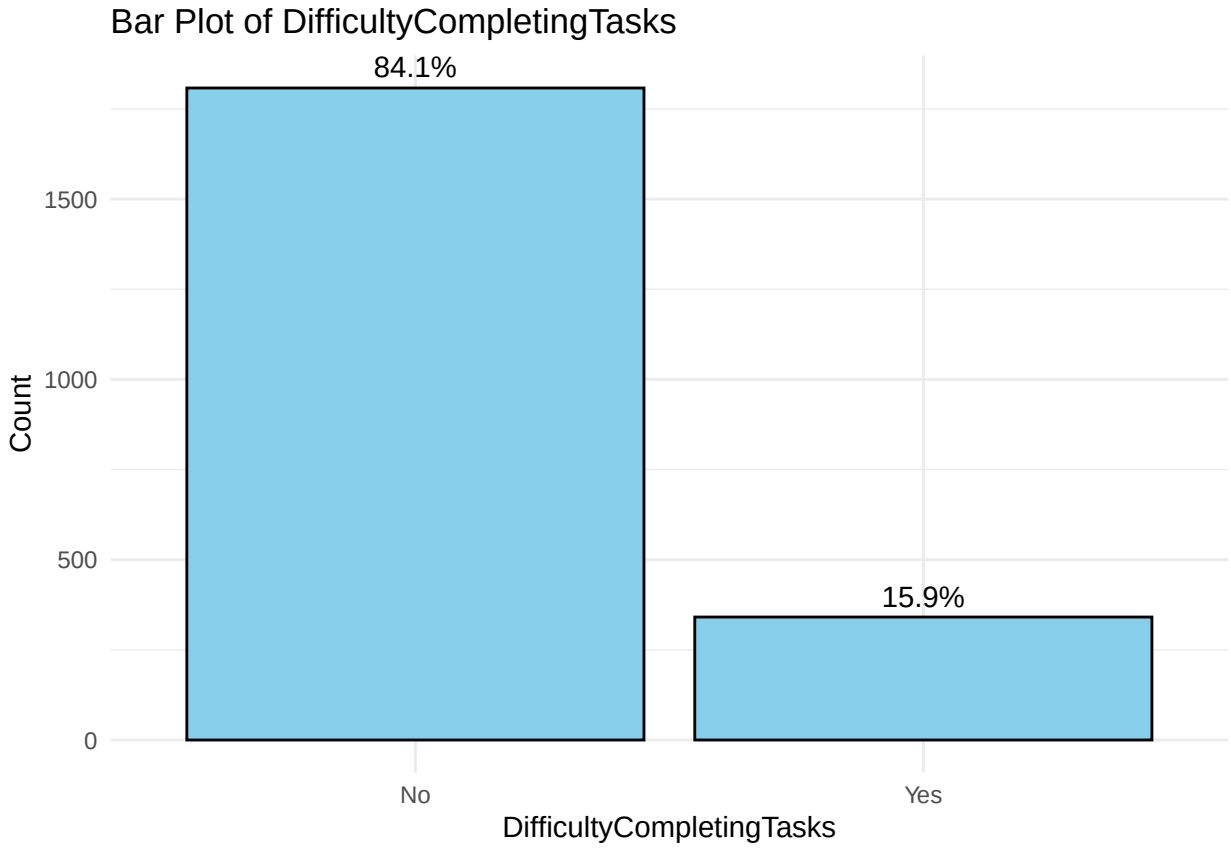

Bar Plot of MemoryComplaints
79.2%
20.8%
1500
1000
500
0
Count
No
Yes
MemoryComplaints

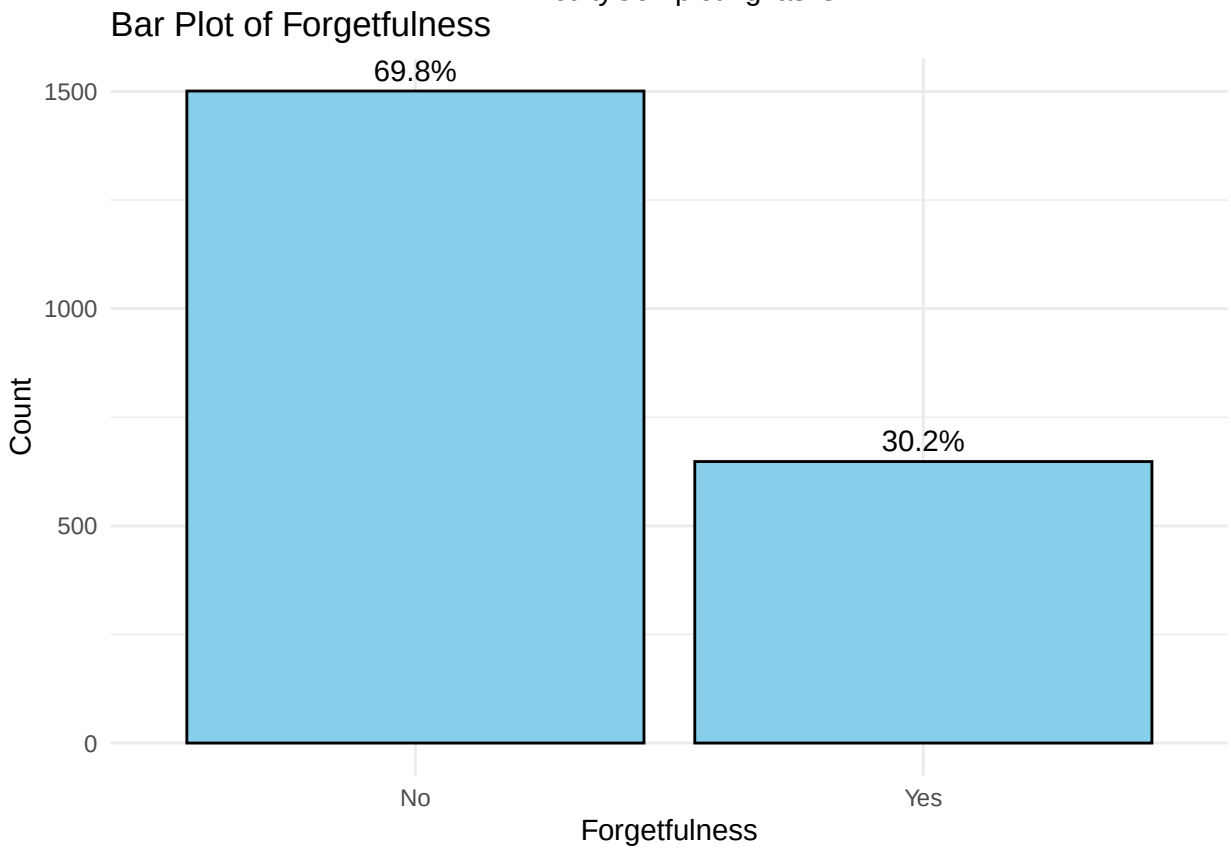

Bar Plot of BehavioralProblems
84.3%
15.7%
1500
1000
500
0
Count
No
Yes
BehavioralProblems

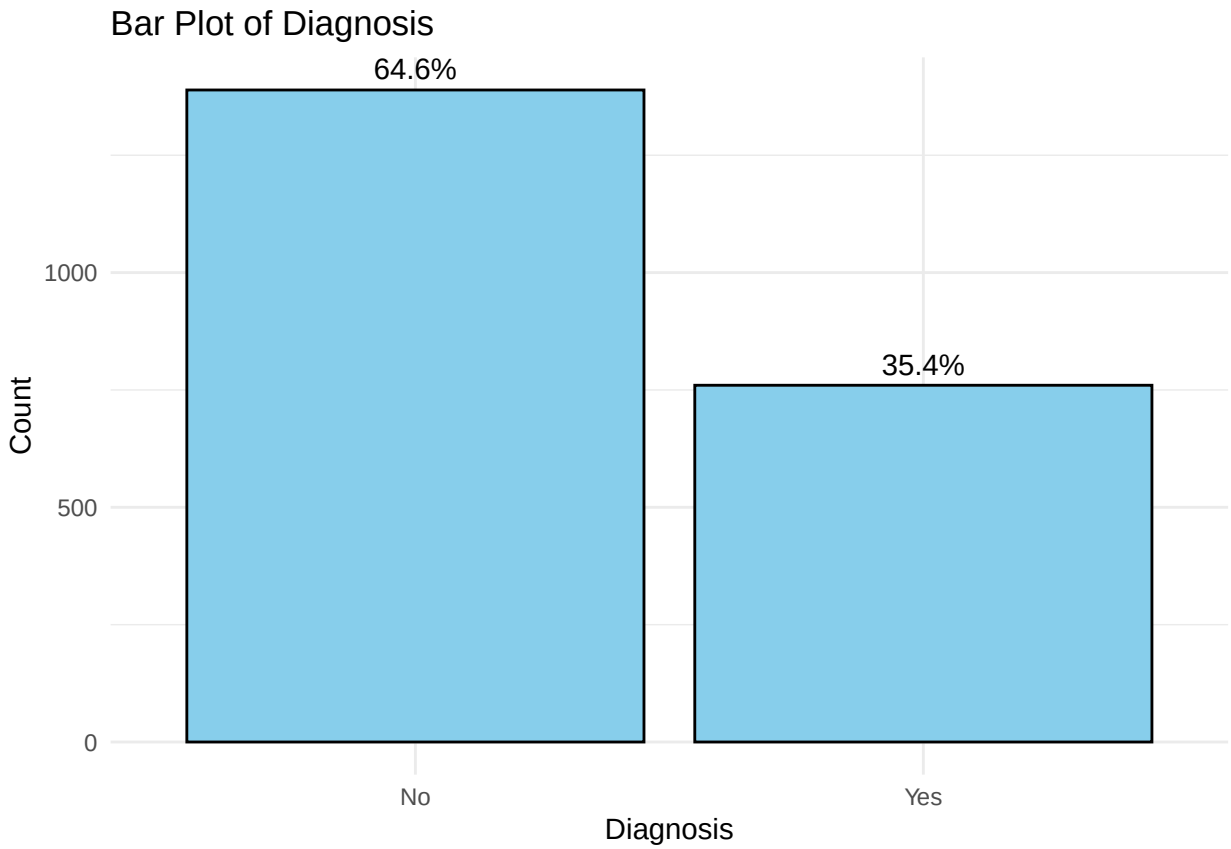

Bar Plot of Confusion
79.5%
20.5%
1500
1000
500
0
Count
No
Yes
Confusion


Bar Plot of Disorientation
84.2%
15.8%
1500
1000
500
0
Count
No
Yes
Disorientation


Bar Plot of PersonalityChanges
84.9%
15.1%
1500
1000
500
0
Count
No
Yes
PersonalityChanges


Bar Plot of DifficultyCompletingTasks
84.1%
15.9%
1500
1000
500
0
Count
No
Yes
DifficultyCompletingTasks


Bar Plot of Forgetfulness
69.8%
30.2%
1500
1000
500
0
Count
No
Yes
Forgetfulness



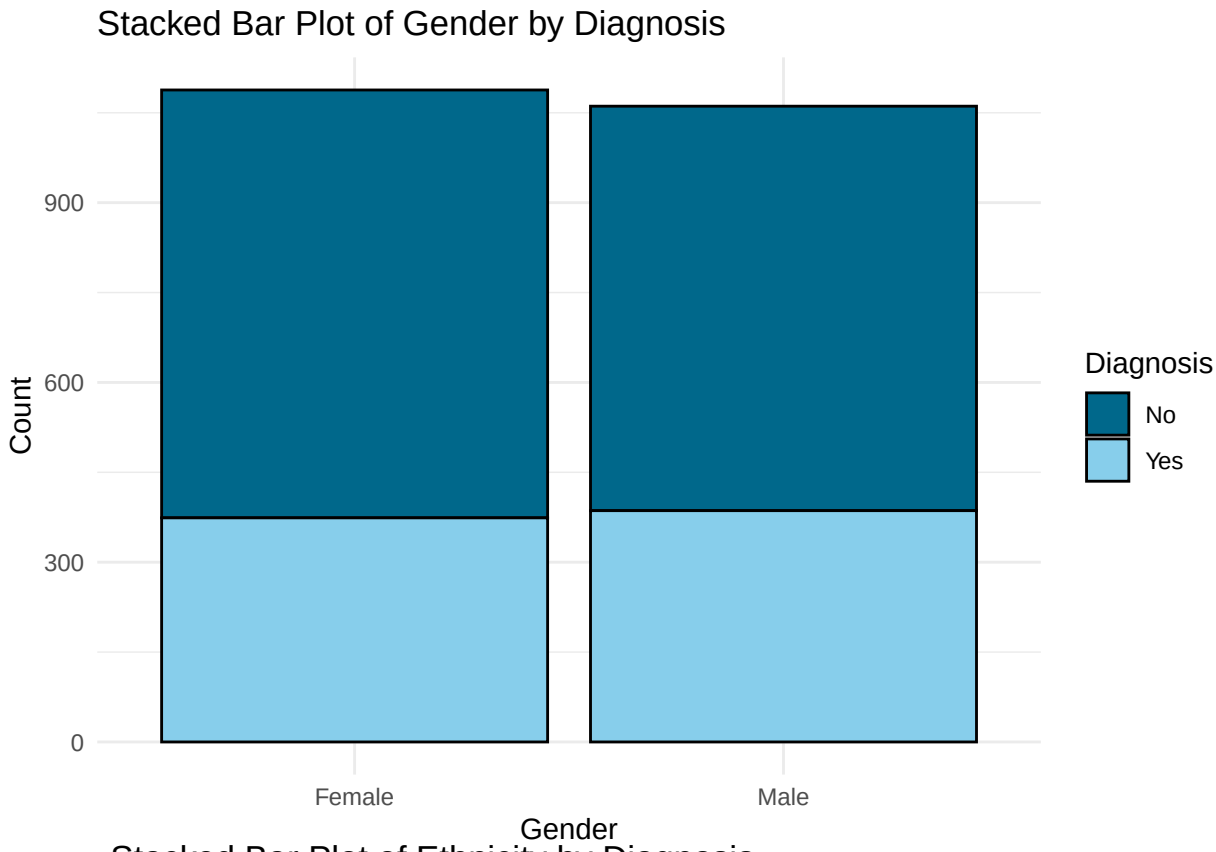
Frequency tables of categorical data are generated.

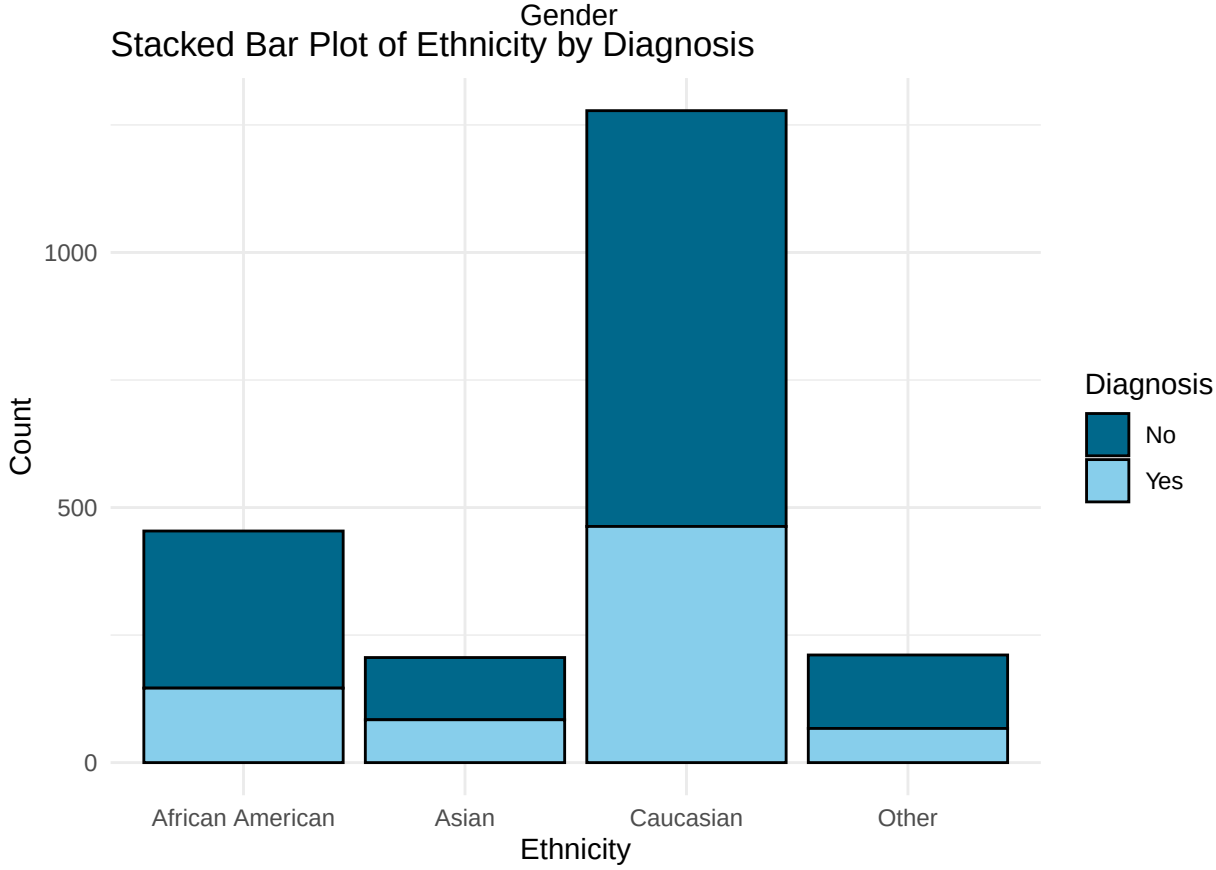
6-2. Distribution of ctegorical variables, based on diagnosis status

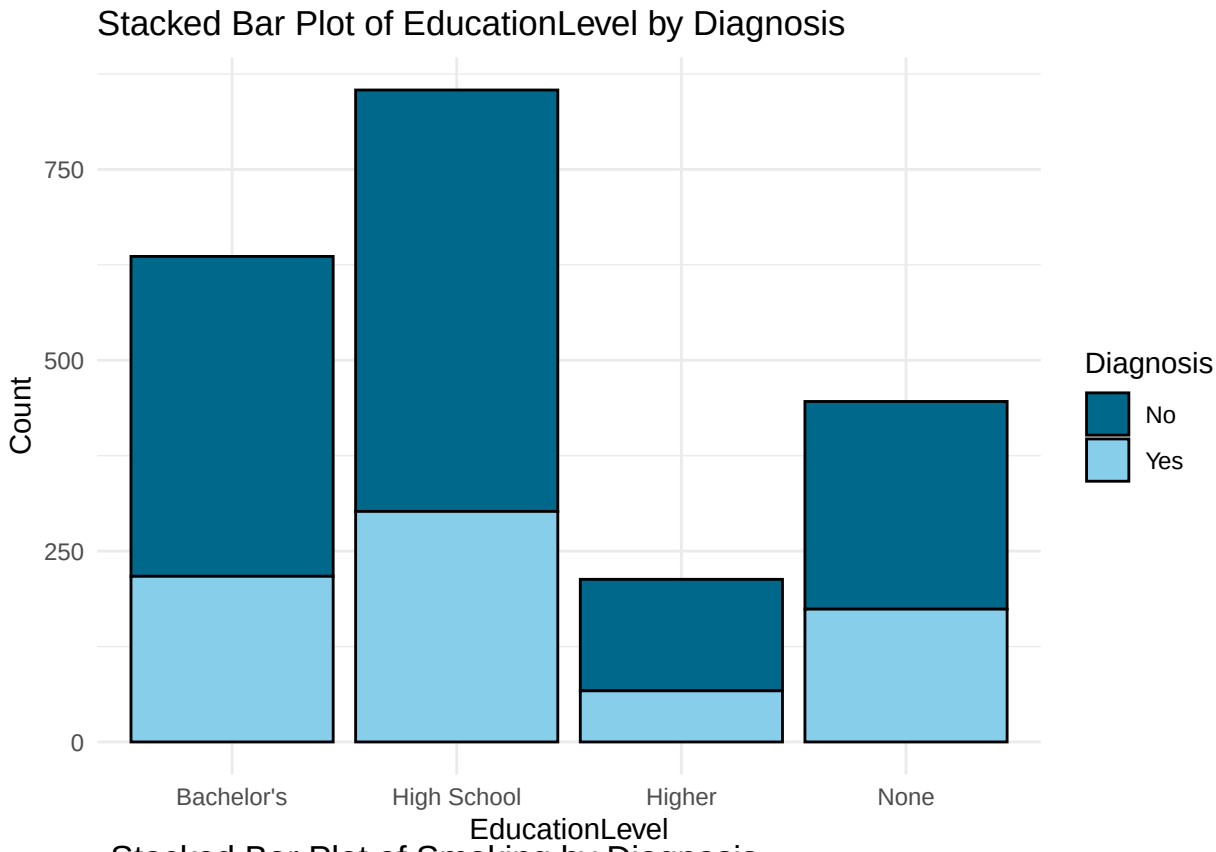
```
cat("\nHistograms of Categorical Variables based on Diagnosis:\n")
```

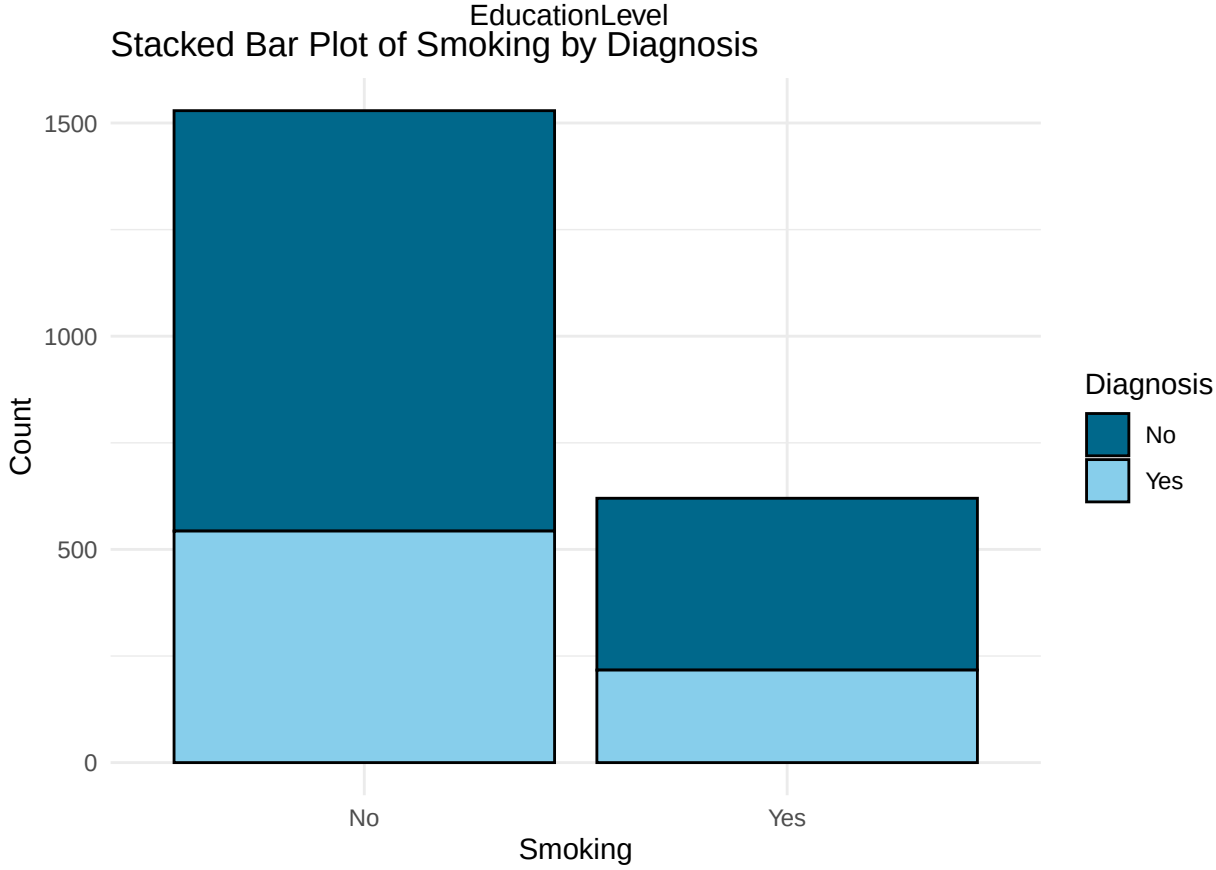
```
##
## Histograms of Categorical Variables based on Diagnosis:
```

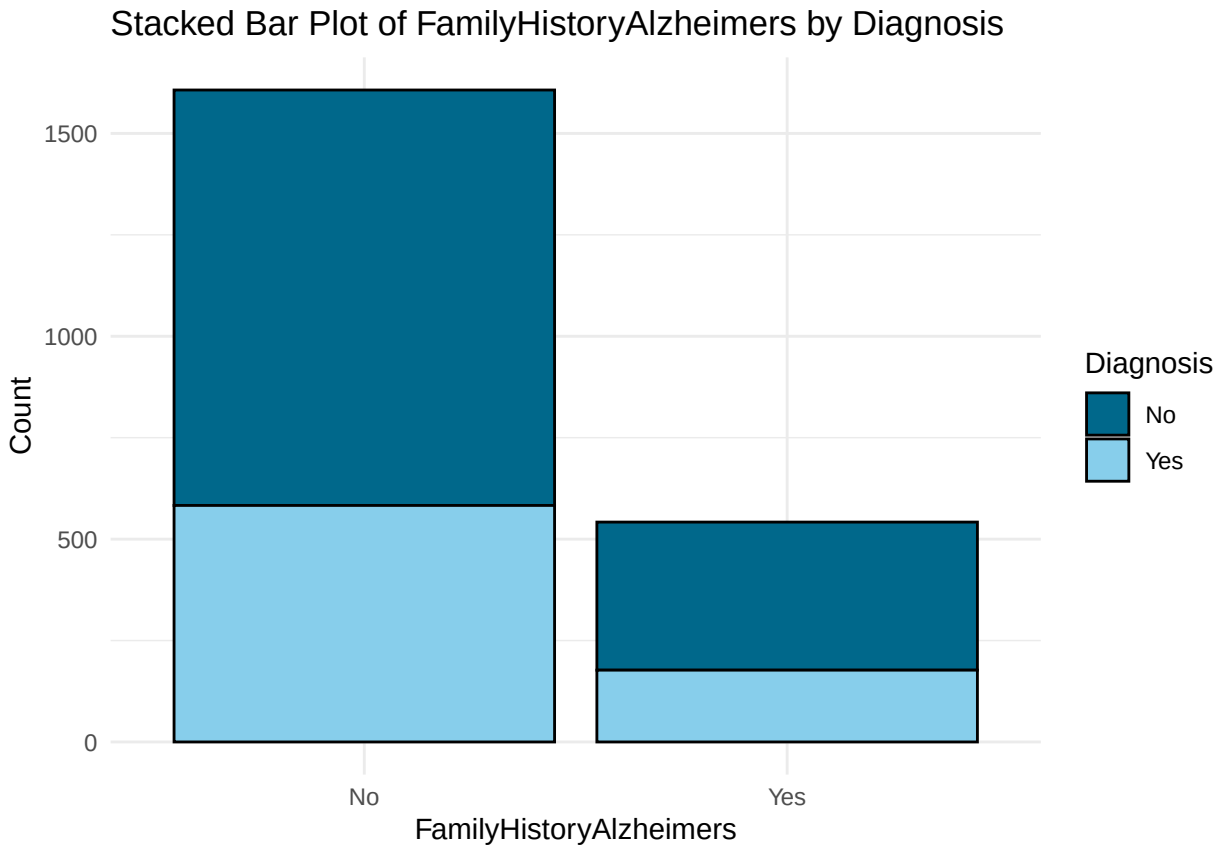
```
for (c in 1:ncol(categorical_variables)) {
  if (colnames(categorical_variables)[c]=='Diagnosis') next
  counts <- categorical_variables %>%
    count(!!sym(colnames(categorical_variables)[c]), Diagnosis)
  p <- ggplot(counts, aes_string(x = colnames(categorical_variables)[c], y = "n", fill = "Diagnosis")) +
    geom_bar(stat = "identity", position = "stack", color = "black") +
    labs(title = paste("Stacked Bar Plot of", colnames(categorical_variables)[c], "by Diagnosis"),
         x = colnames(categorical_variables)[c],
         y = "Count",
         fill = "Diagnosis") +
    theme_minimal() +
    scale_fill_manual(values = c("Yes" = "skyblue", "No" = "deepskyblue4"))
  print(p)
}
```

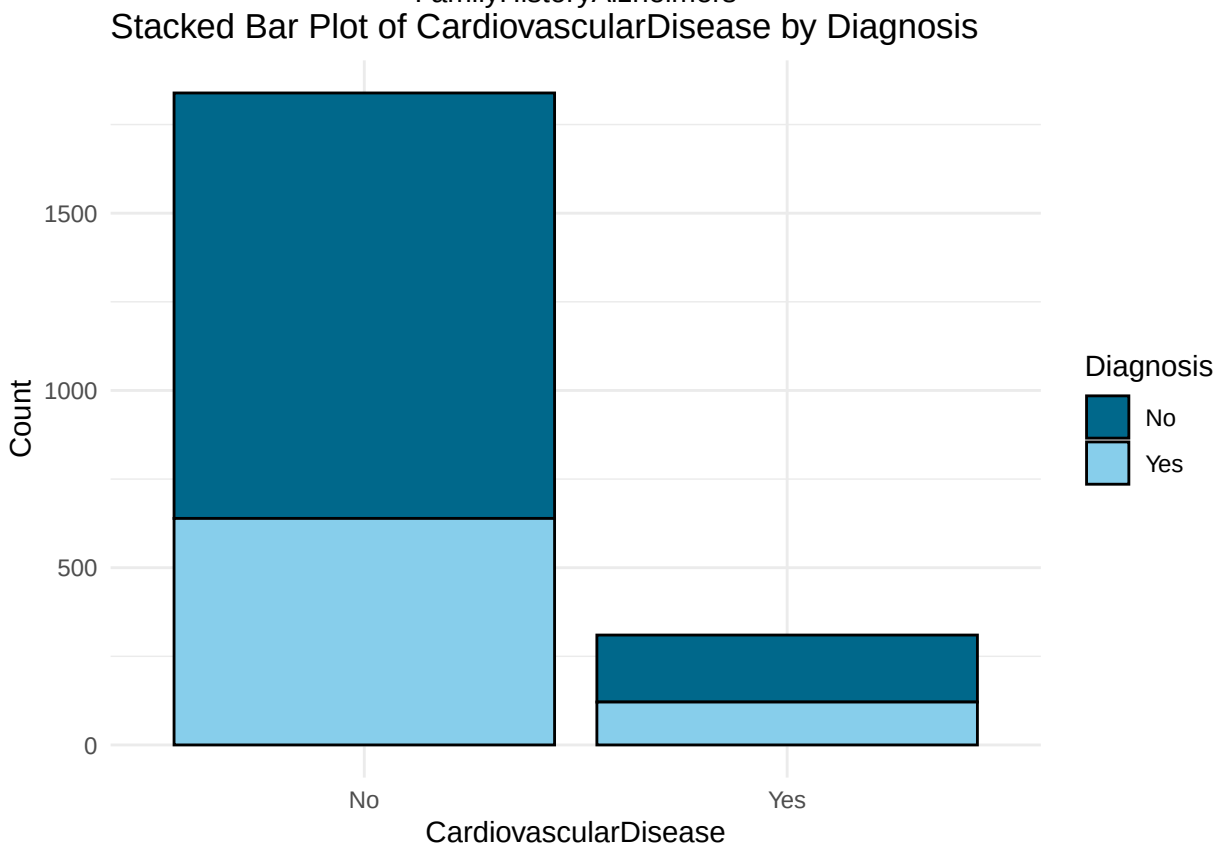

Stacked Bar Plot of Gender by Diagnosis
Count
900
600
300
0
Female
Male
Gender
Diagnosis
No
Yes

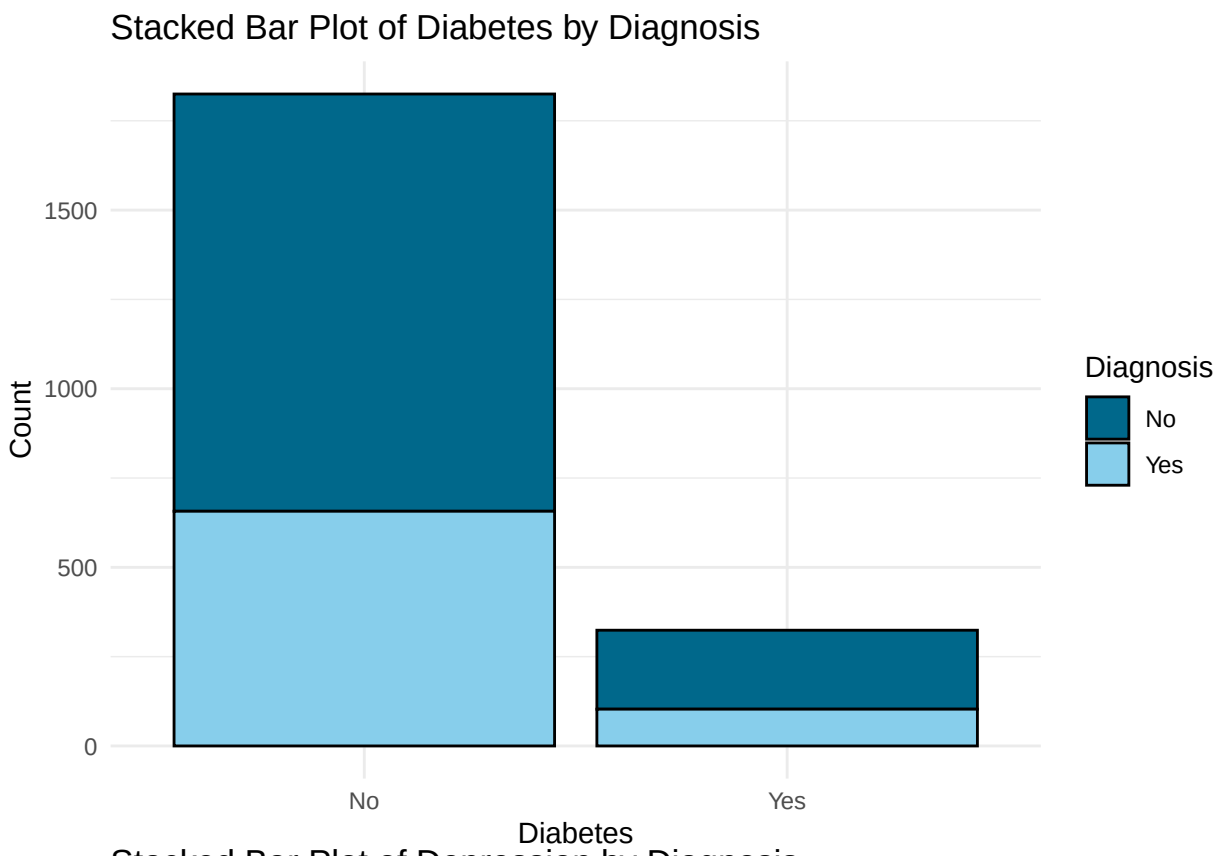

Stacked Bar Plot of Ethnicity by Diagnosis
Count
1000
500
0
African American
Asian
Caucasian
Other
Ethnicity
Diagnosis
No
Yes

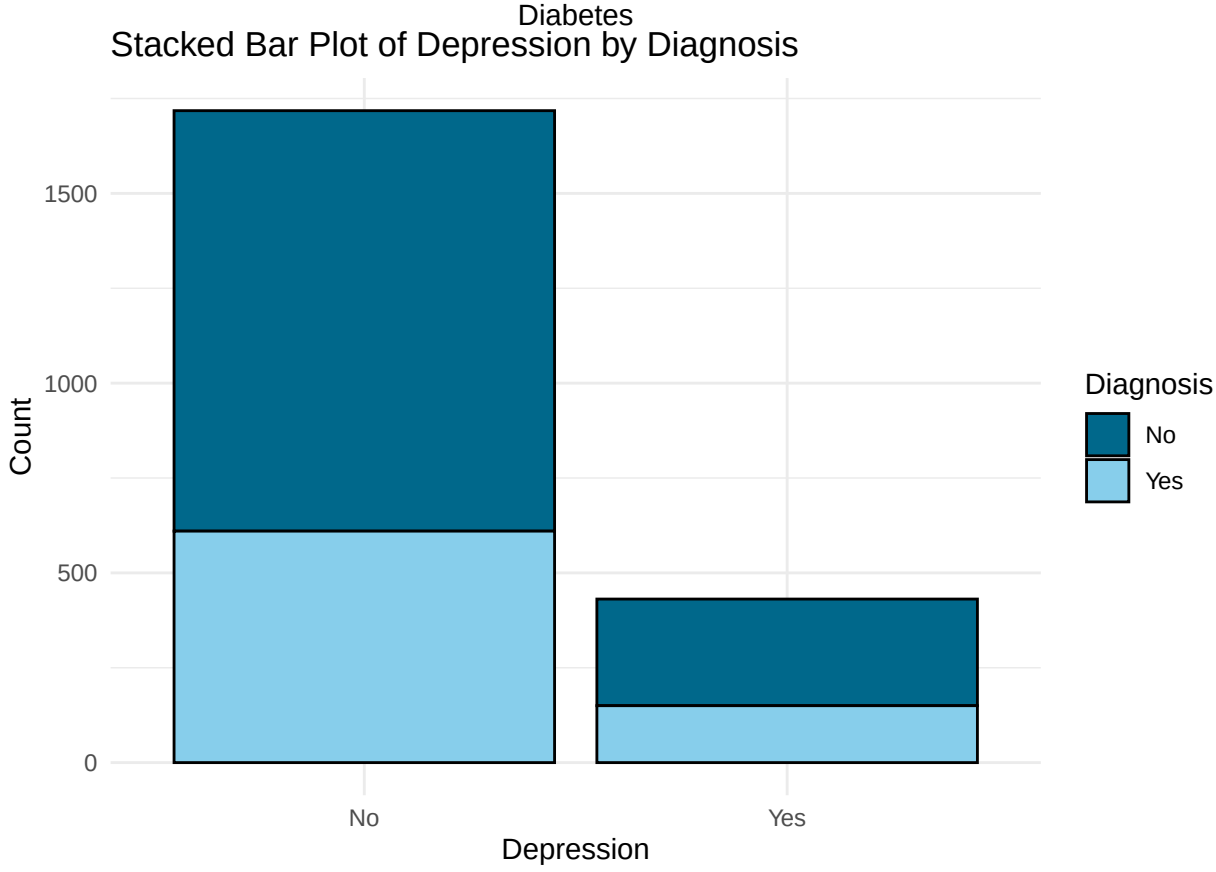

Stacked Bar Plot of EducationLevel by Diagnosis
Count
750
500
250
0
Bachelor's
High School
Higher
None
EducationLevel
Diagnosis
No
Yes

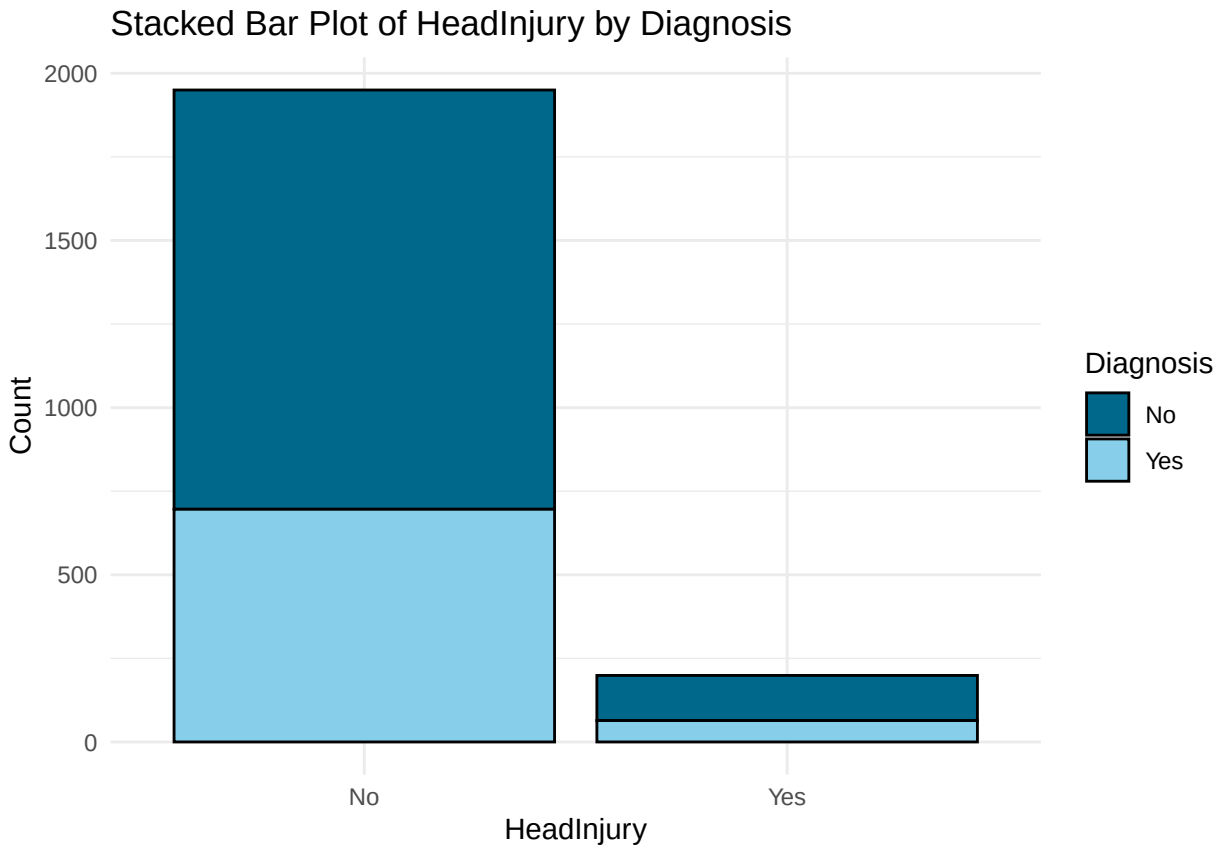

Stacked Bar Plot of Smoking by Diagnosis
Count
1500
1000
500
0
No
Yes
Smoking
Diagnosis
No
Yes

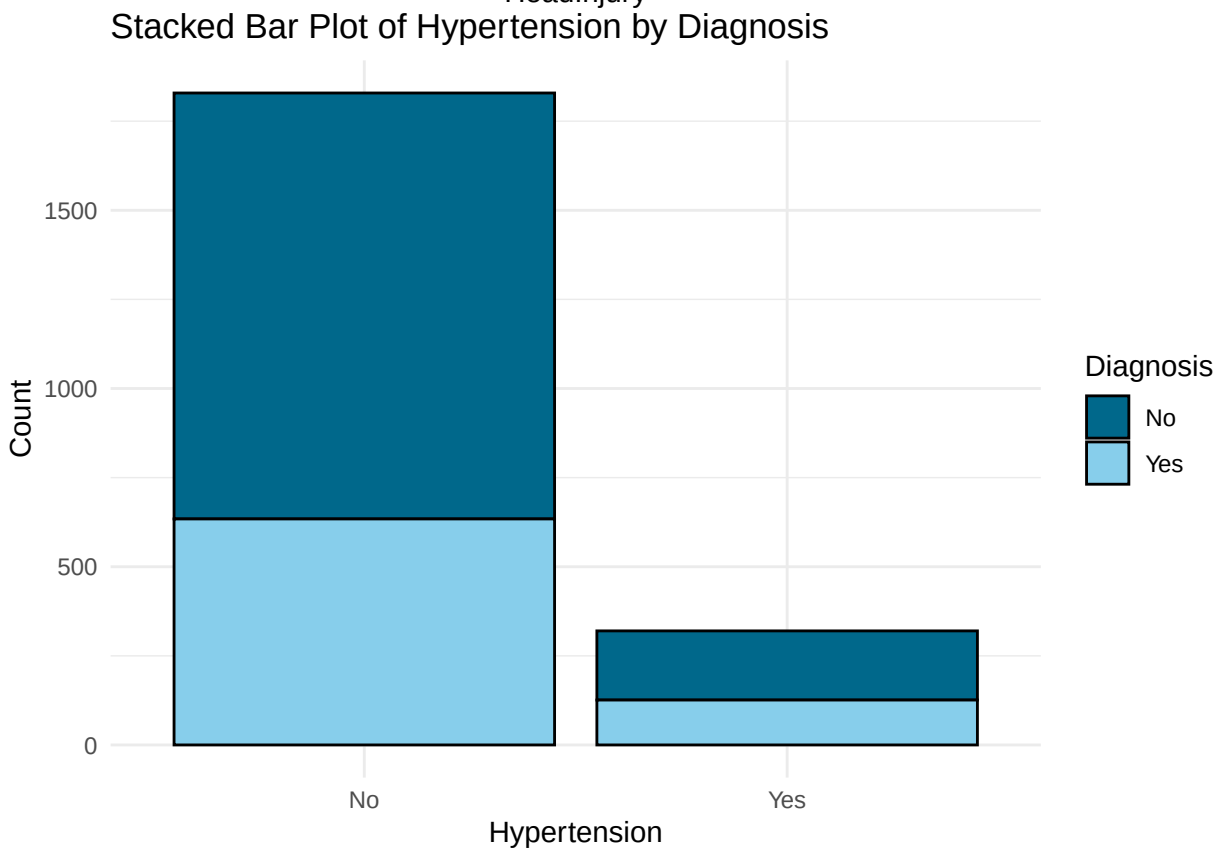

Stacked Bar Plot of FamilyHistoryAlzheimers by Diagnosis
1500
1000
500
0
Count
No
Yes
FamilyHistoryAlzheimers
Diagnosis
No
Yes

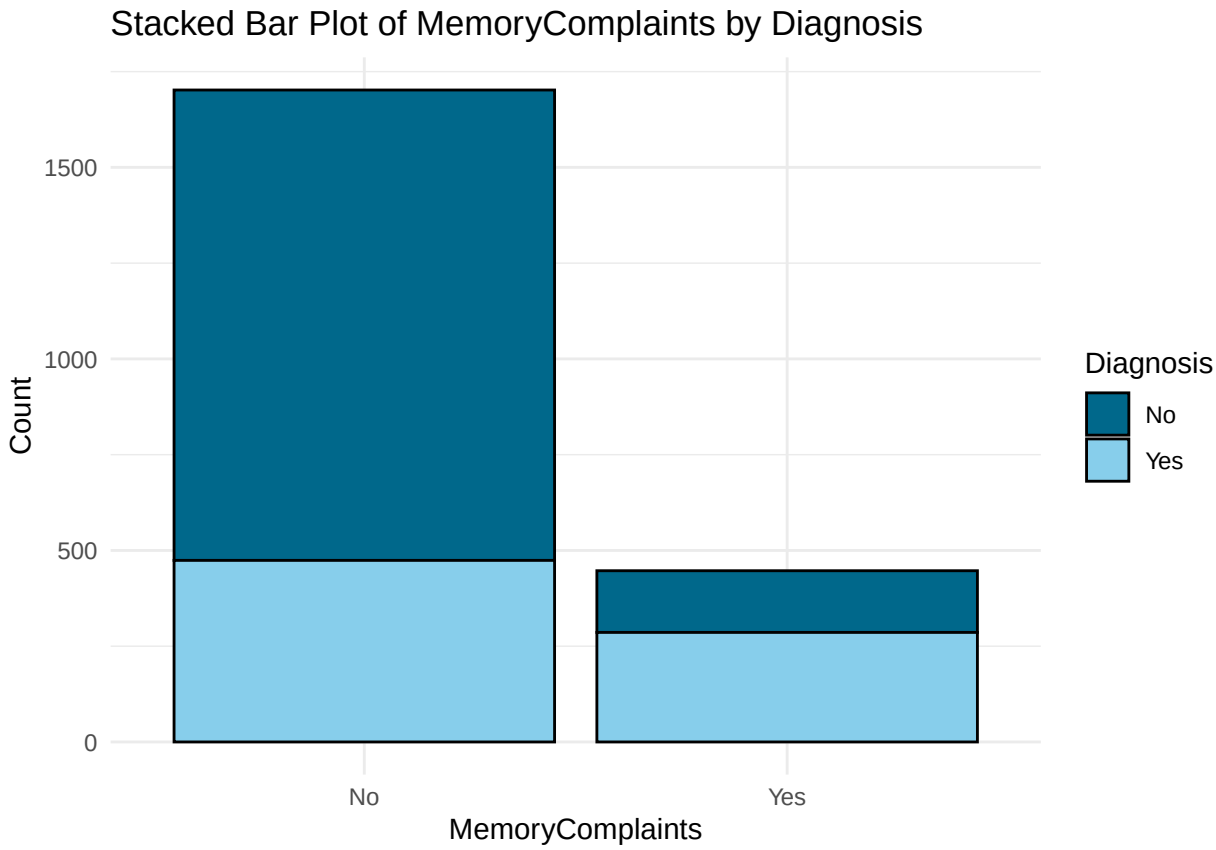

Stacked Bar Plot of CardiovascularDisease by Diagnosis
1500
1000
500
0
Count
No
Yes
CardiovascularDisease
Diagnosis
No
Yes

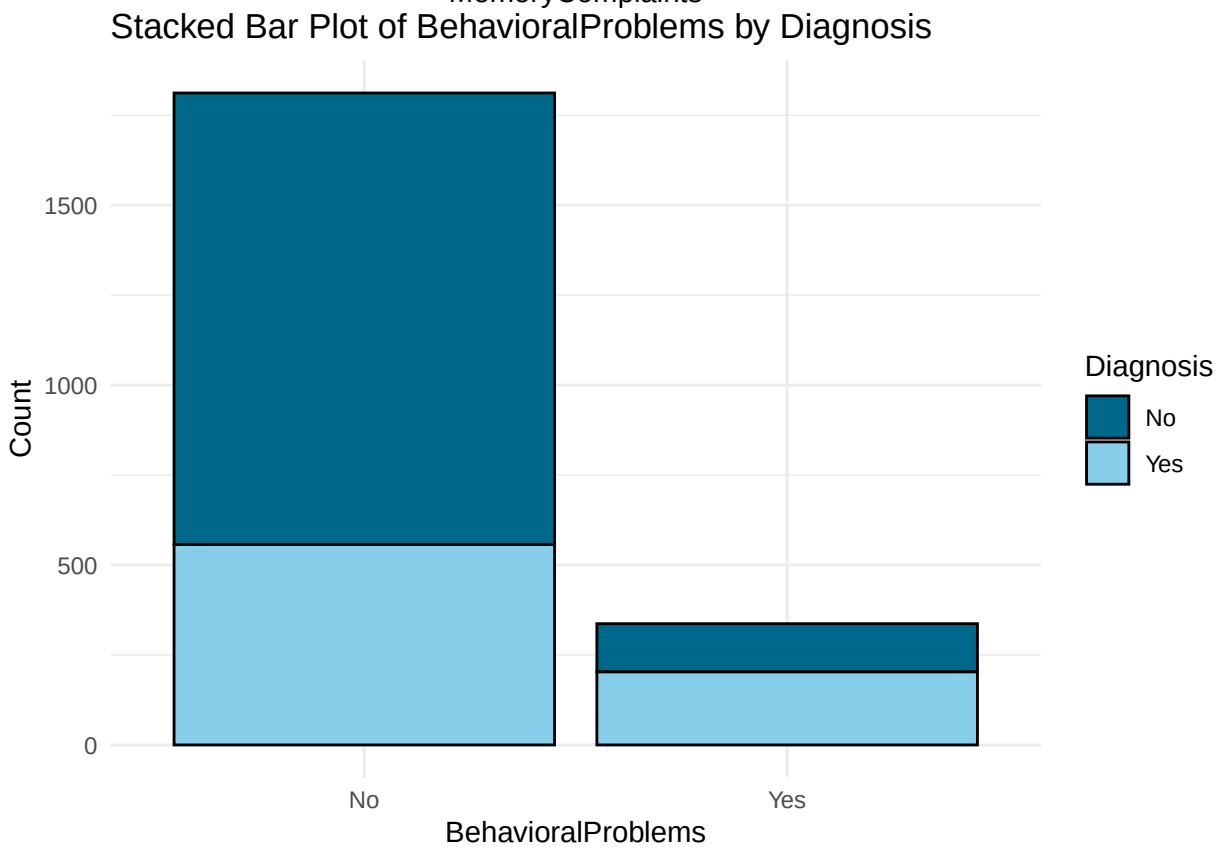

Stacked Bar Plot of Diabetes by Diagnosis
Count
1500
1000
500
0
No
Yes
Diabetes
Diagnosis
No
Yes

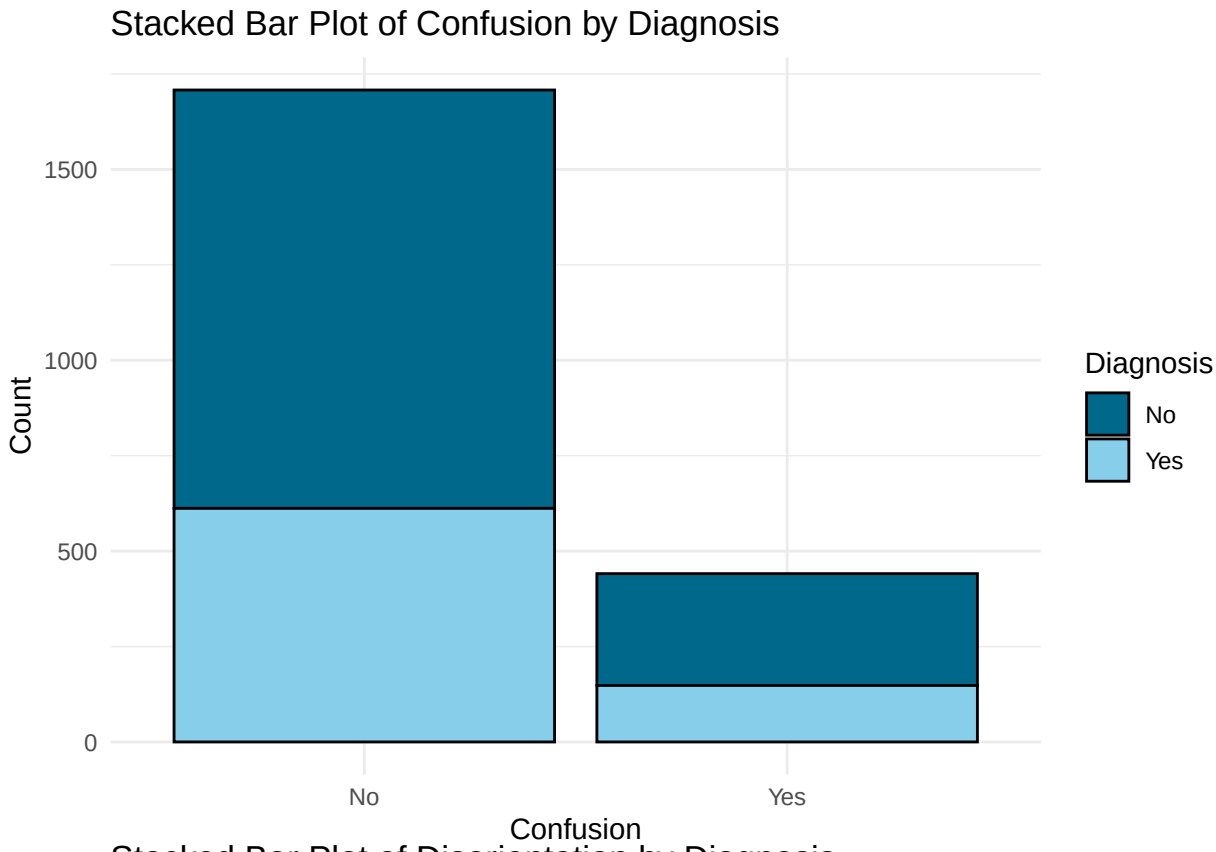

Stacked Bar Plot of Depression by Diagnosis
Count
1500
1000
500
0
No
Yes
Depression
Diagnosis
No
Yes

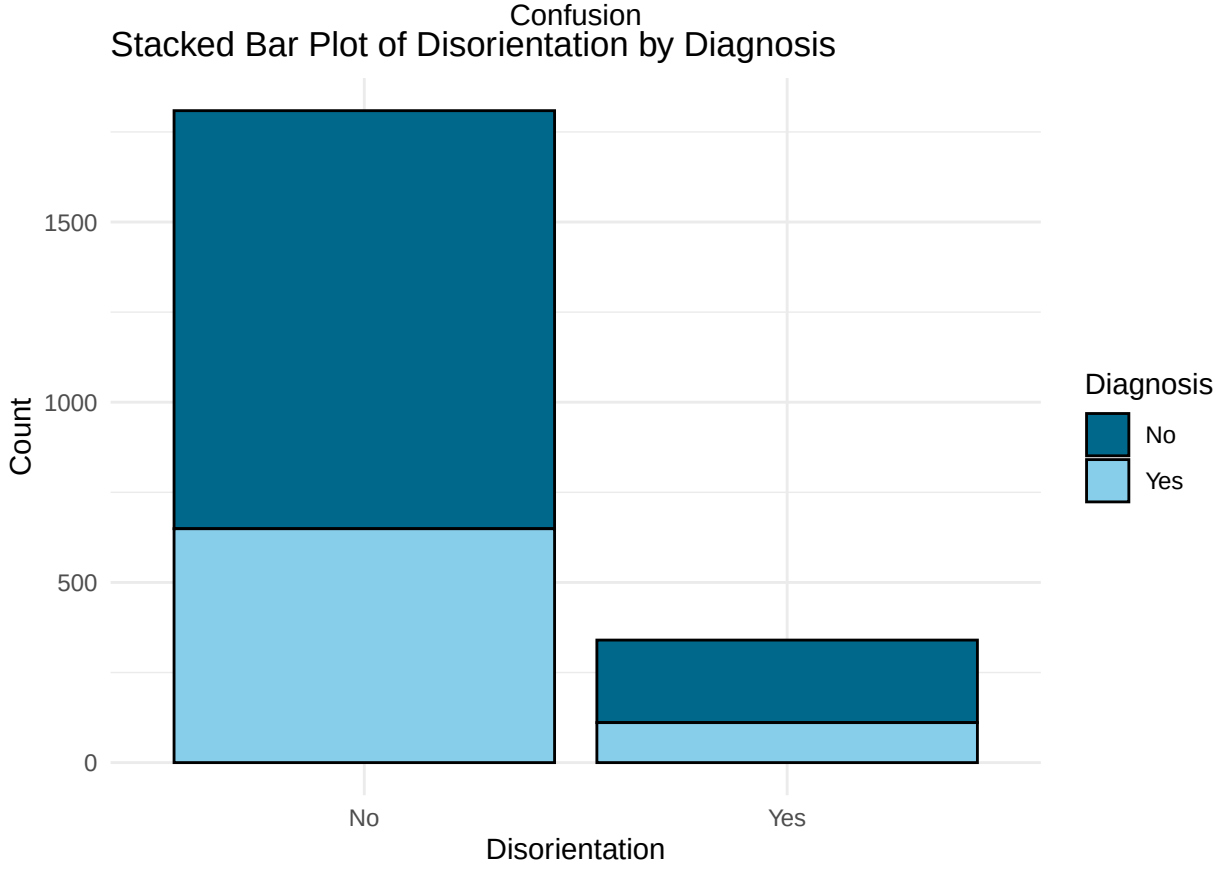

Stacked Bar Plot of HeadInjury by Diagnosis
2000
1500
1000
500
0
Count
No
Yes
HeadInjury
Diagnosis
No
Yes

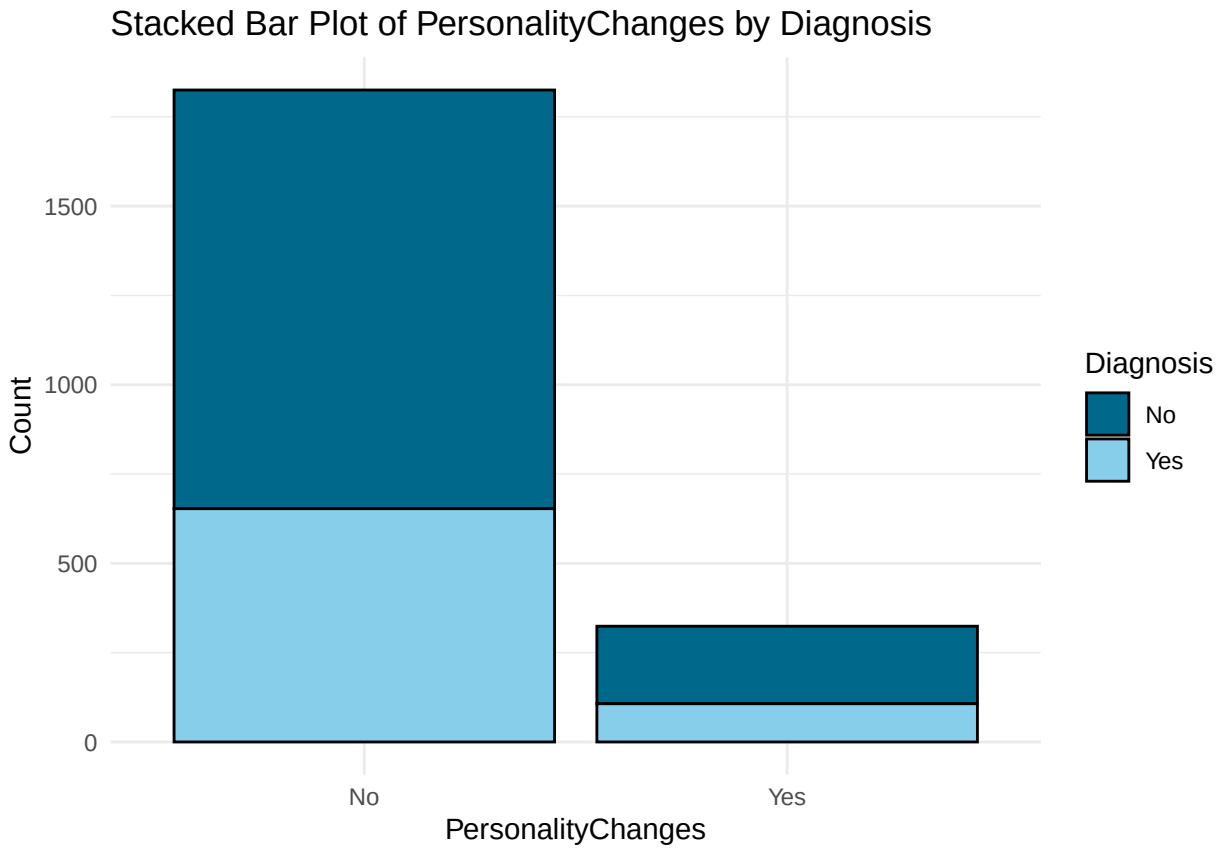

Stacked Bar Plot of Hypertension by Diagnosis
1500
1000
500
0
Count
No
Yes
Hypertension
Diagnosis
No
Yes

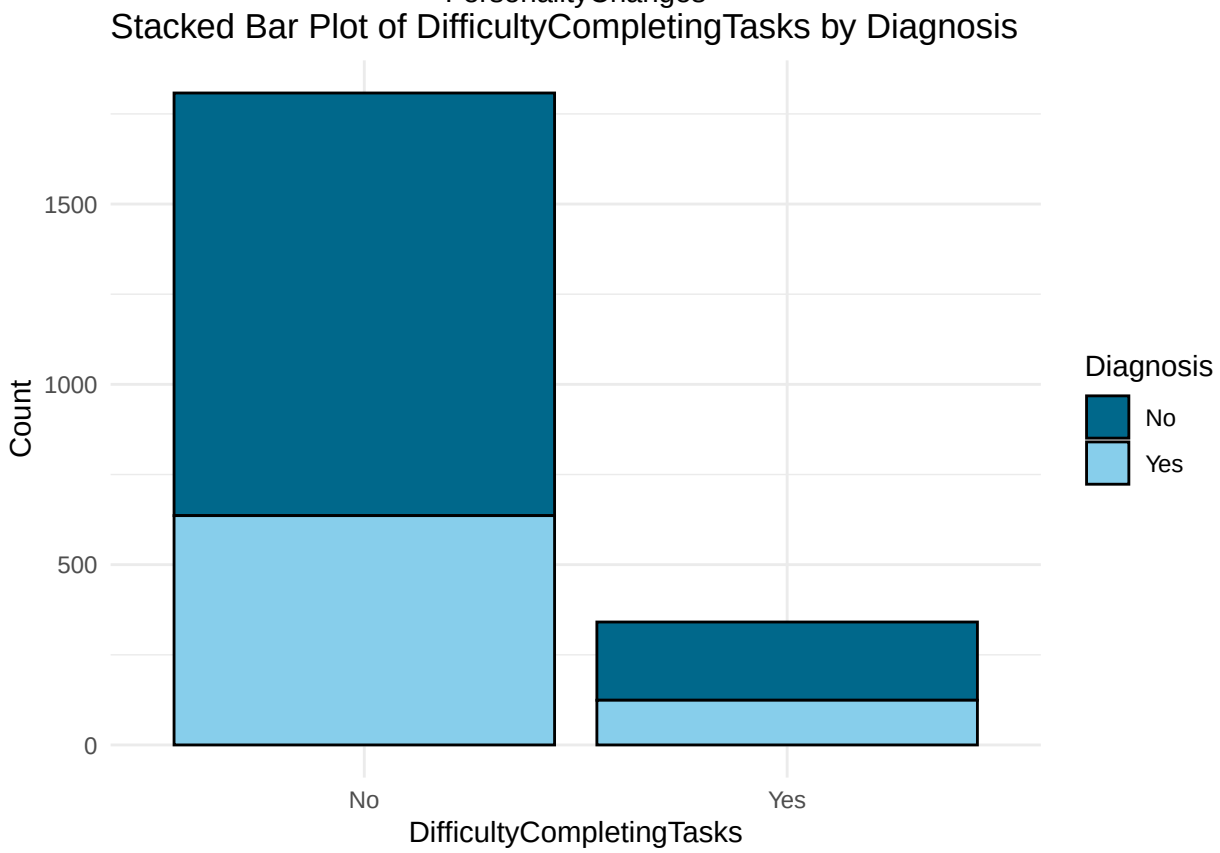

Stacked Bar Plot of MemoryComplaints by Diagnosis
Count
1500
1000
500
0
No
Yes
MemoryComplaints
Diagnosis
No
Yes

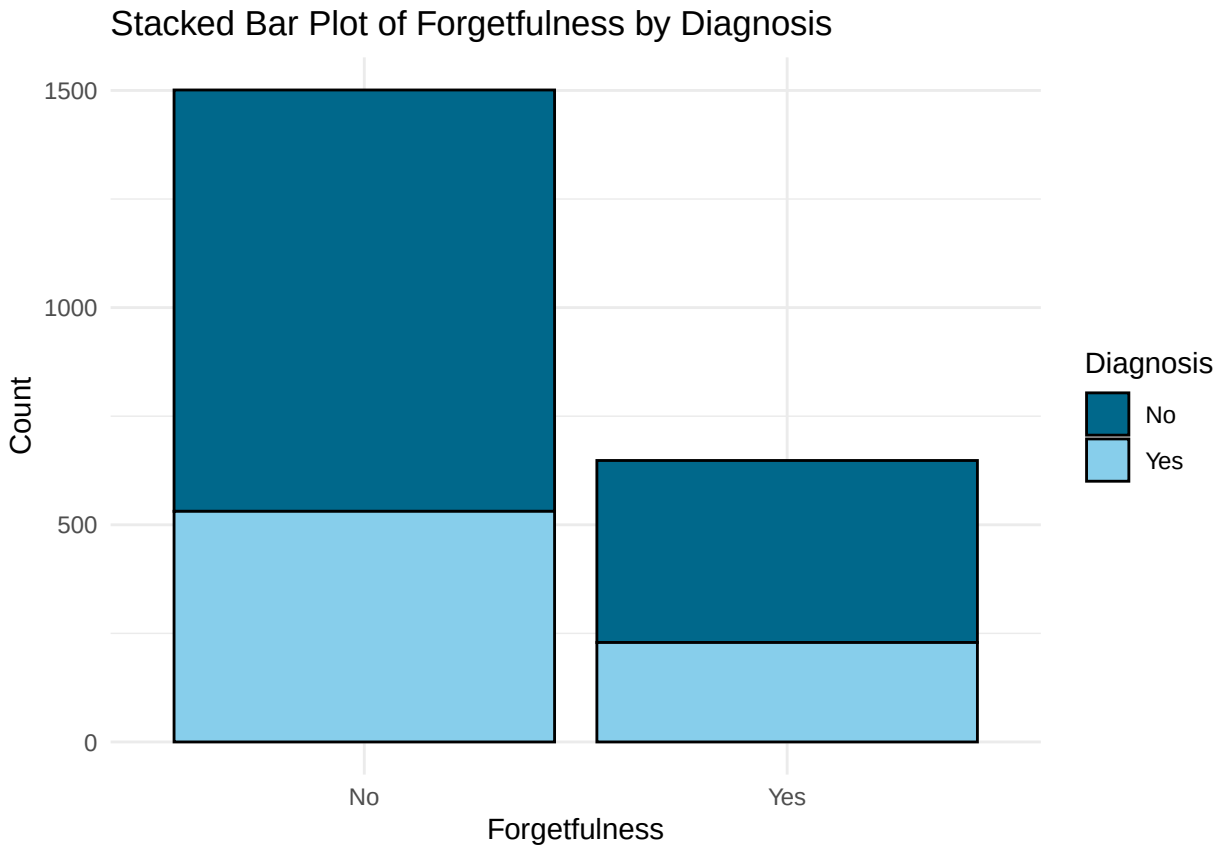

Stacked Bar Plot of BehavioralProblems by Diagnosis
Count
1500
1000
500
0
No
Yes
BehavioralProblems
Diagnosis
No
Yes


Stacked Bar Plot of Confusion by Diagnosis
Count
1500
1000
500
0
No
Yes
Confusion
Diagnosis
No
Yes


Stacked Bar Plot of Disorientation by Diagnosis
Count
1500
1000
500
0
No
Yes
Disorientation
Diagnosis
No
Yes


Stacked Bar Plot of PersonalityChanges by Diagnosis
Count
1500
1000
500
0
No
Yes
PersonalityChanges
Diagnosis
No
Yes


Stacked Bar Plot of DifficultyCompletingTasks by Diagnosis
Count
1500
1000
500
0
No
Yes
DifficultyCompletingTasks
Diagnosis
No
Yes



For better understanding, frequency tables of categorical data are re-generated based on diagnosis status of Alzheimer's disease.