Depression Draft 1

Christy Hui

2024-11-30

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1

## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'purrr' was built under R version 4.4.1

## Warning: package 'stringr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

## -- 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 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```r
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.4.2

##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.1

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
```

```
##
##     lift
```

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.4.2
```

```r
depression = read.csv("final_depression_dataset_1.csv")
```

```r
# find the dimension of depression
dim(depression)
```

```
## [1] 2556   19
```

```r
# find if there exist duplicates
sum(duplicated(depression))
```

```
## [1] 0
```

```r
# find number of NAs for each column
sapply(depression, function(x) {sum(is.na(x))})
```

```
##                                  Name                                Gender 
##                                     0                                     0 
##                                   Age                                  City 
##                                     0                                     0 
##       Working.Professional.or.Student                            Profession 
##                                     0                                     0 
##                     Academic.Pressure                         Work.Pressure 
##                                  2054                                   502 
##                                  CGPA                    Study.Satisfaction 
##                                  2054                                  2054 
##                      Job.Satisfaction                        Sleep.Duration 
##                                   502                                     0 
##                        Dietary.Habits                                Degree 
##                                     0                                     0 
## Have.you.ever.had.suicidal.thoughts..                      Work.Study.Hours 
##                                     0                                     0 
##                      Financial.Stress      Family.History.of.Mental.Illness 
##                                     0                                     0 
##                            Depression 
##                                     0
```

```r
# delete columns with NAs
depression = depression[, -c(7:11)]
sapply(depression, function(x) {sum(is.na(x))})
```

```
##                                  Name                                Gender 
##                                     0                                     0 
##                                   Age                                  City 
##                                     0                                     0 
##       Working.Professional.or.Student                            Profession 
##                                     0                                     0 
##                        Sleep.Duration                        Dietary.Habits 
##                                     0                                     0 
##                                Degree Have.you.ever.had.suicidal.thoughts.. 
##                                     0                                     0 
##                      Work.Study.Hours                      Financial.Stress 
##                                      0                                     0
##      Family.History.of.Mental.Illness                            Depression
##                                      0                                     0
```

```r
# due to a large amount of varied answers for "City" and "Profession," we delete the variables
# we also delete name because we don't care about that variable
unique(depression$City)
```

```
##  [1] "Ghaziabad"     "Kalyan"        "Bhopal"        "Thane"        
##  [5] "Indore"        "Pune"          "Bangalore"     "Hyderabad"    
##  [9] "Srinagar"      "Nashik"        "Kolkata"       "Ahmedabad"    
## [13] "Varanasi"      "Chennai"       "Jaipur"        "Surat"        
## [17] "Vasai-Virar"   "Rajkot"        "Patna"         "Mumbai"       
## [21] "Vadodara"      "Lucknow"       "Faridabad"     "Meerut"       
## [25] "Kanpur"        "Visakhapatnam" "Ludhiana"      "Nagpur"       
## [29] "Delhi"         "Agra"
```

```r
unique(depression$Profession)
```

```
##  [1] "Teacher"                 "Financial Analyst"       "UX/UI Designer"         
##  [4] "Civil Engineer"          "Accountant"              "Lawyer"                 
##  [7] "Content Writer"          ""                        "Pilot"                  
## [10] "Customer Support"        "Judge"                   "Architect"              
## [13] "HR Manager"              "Digital Marketer"        "Sales Executive"        
## [16] "Business Analyst"        "Mechanical Engineer"     "Consultant"             
## [19] "Data Scientist"          "Pharmacist"              "Software Engineer"      
## [22] "Travel Consultant"       "Manager"                 "Entrepreneur"           
## [25] "Doctor"                  "Researcher"              "Plumber"                
## [28] "Finanancial Analyst"     "Marketing Manager"       "Educational Consultant" 
## [31] "Chemist"                 "Research Analyst"        "Chef"                   
## [34] "Electrician"             "Graphic Designer"        "Investment Banker"
```

```r
depression = subset(depression, select = -c(Name, City, Profession))
```

```r
# degree has many varied answers as well; however, they can be recoded into three main categories: high
unique(depression$Degree)
```

```
##  [1] "MA"       "B.Com"    "M.Com"    "MD"       "BE"       "MCA"     
##  [7] "BA"       "LLM"      "BCA"      "Class 12" "B.Ed"     "M.Tech"  
## [13] "LLB"      "B.Arch"   "ME"       "MBA"      "M.Pharm"  "MBBS"    
## [19] "PhD"      "BSc"      "MSc"      "MHM"      "BBA"      "BHM"     
## [25] "B.Tech"   "M.Ed"     "B.Pharm"
```

```r
depression$Degree = case_when(depression$Degree == "Class 12" ~ "High School Equivalent",
                              grepl("^[BL]", depression$Degree) ~ "Bachelors Degree",
                              grepl("^[MP]", depression$Degree) ~ "Post-Graduate Degree")

table(depression$Degree)
```

```
## 
##       Bachelors Degree High School Equivalent   Post-Graduate Degree 
##                   1193                    275                   1088
```

```r
# find type of each variable so we can change each type
sapply(depression, function(x) {class(x)})
```

```
##                                 Gender                                   Age 
##                            "character"                             "integer"
##       Working.Professional.or.Student                         Sleep.Duration
##                           "character"                            "character"
##                        Dietary.Habits                                 Degree
##                           "character"                            "character"
## Have.you.ever.had.suicidal.thoughts..                       Work.Study.Hours
##                           "character"                              "integer"
##                      Financial.Stress       Family.History.of.Mental.Illness
##                             "integer"                            "character"
##                            Depression
##                           "character"
```

```
# change each categorical into a factor, changing the base/ordering them if needed
depression$Gender = as.factor(depression$Gender)
depression$Working.Professional.or.Student = as.factor(depression$Working.Professional.or.Student)
depression$Sleep.Duration = factor(depression$Sleep.Duration, levels = c("Less than 5 hours", "5-6 hour
depression$Dietary.Habits = factor(depression$Dietary.Habits, levels = c("Unhealthy", "Moderate", "Heal
depression$Degree = factor(depression$Degree, levels = c("High School Equivalent", "Bachelors Degree",
depression$Have.you.ever.had.suicidal.thoughts.. = as.factor(depression$Have.you.ever.had.suicidal.thoug
depression$Financial.Stress = factor(depression$Financial.Stress, levels = c(1, 2, 3, 4, 5))
depression$Family.History.of.Mental.Illness = as.factor(depression$Family.History.of.Mental.Illness)
depression$Depression = as.factor(depression$Depression)
```

```
depressionFactored = select(depression, where(is.factor))
sapply(depressionFactored, table)
```

```
## $Gender
##
## Female   Male
##   1223   1333
##
## $Working.Professional.or.Student
##
##              Student Working Professional
##                  502                 2054
##
## $Sleep.Duration
##
## Less than 5 hours         5-6 hours         7-8 hours More than 8 hours
##               648               628               658               622
##              TRUE
##                 0
##
## $Dietary.Habits
##
## Unhealthy  Moderate   Healthy
##       882       832       842
##
## $Degree
##
## High School Equivalent       Bachelors Degree   Post-Graduate Degree
##                    275                   1193                   1088
##
## $Have.you.ever.had.suicidal.thoughts..
##
##   No  Yes
## 1307 1249
##
## $Financial.Stress
##
##   1   2   3   4   5
## 517 549 488 501 501
##
## $Family.History.of.Mental.Illness
##
##   No  Yes
## 1311 1245
##
## $Depression
##
##   No  Yes
## 2101  455
```

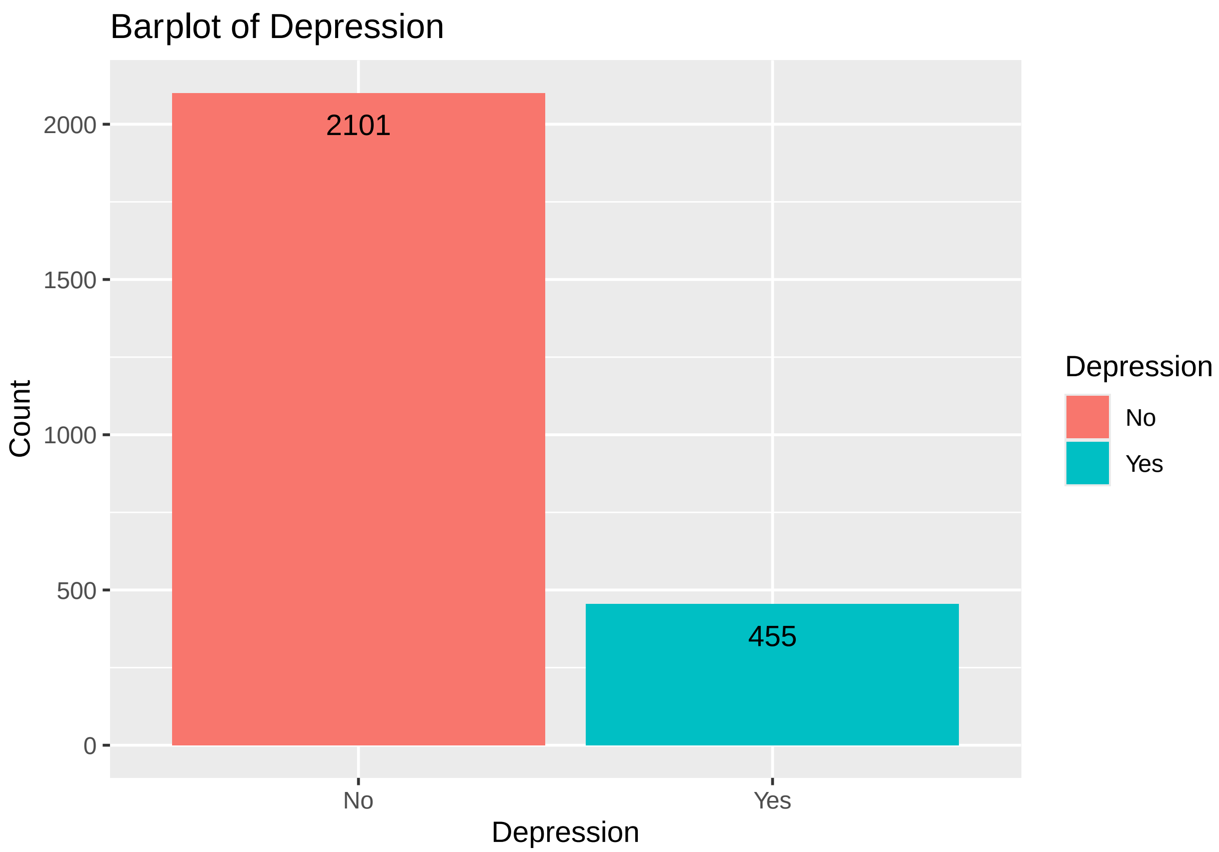
IF YOU WANT TO CHANGE THE COLOR, PLEASE USE THESE TWO LINKS:

https://sape.inf.usi.ch/quick-reference/ggplot2/colour

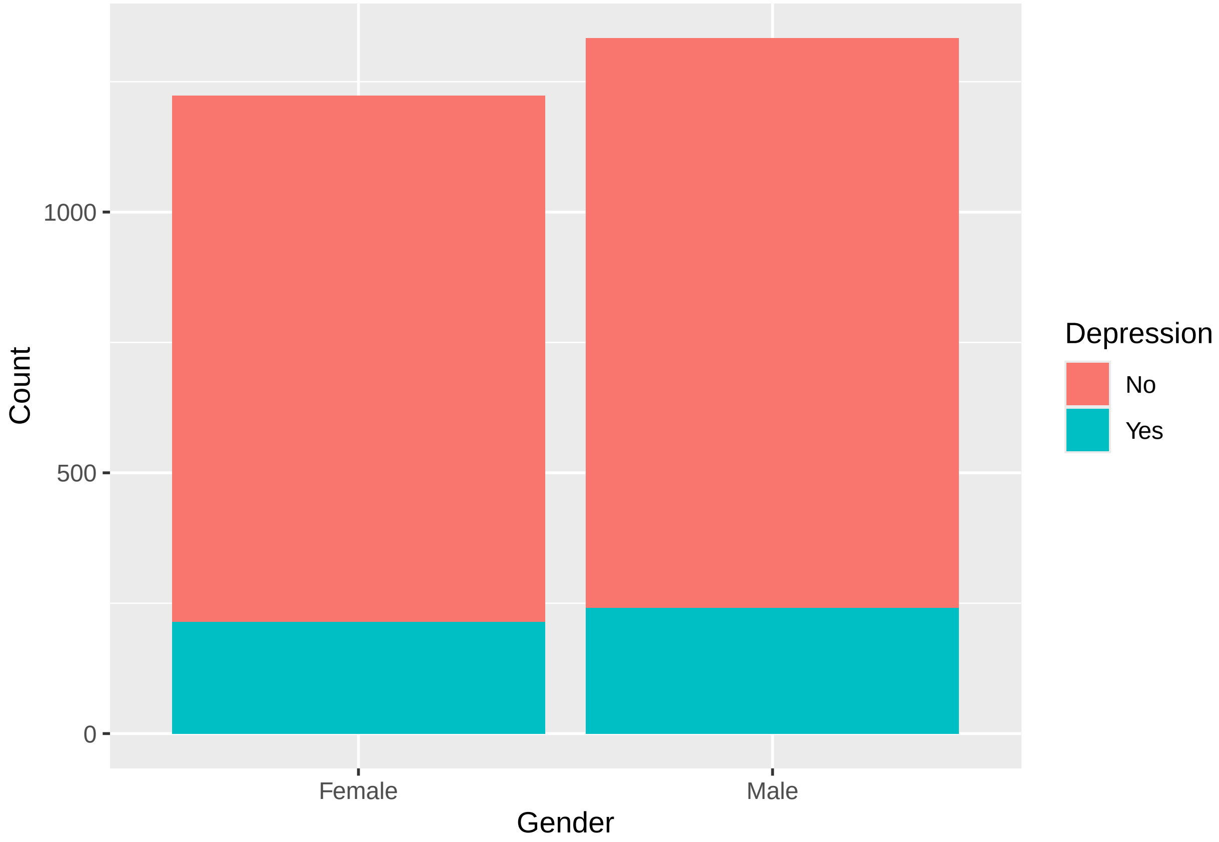
https://www.rapidtables.com/web/color/RGB_Color.html

```
# plot depression count
ggplot(depression, aes(x = Depression)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Depression") +
  ylab("Count") +
  ggtitle("Barplot of Depression") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



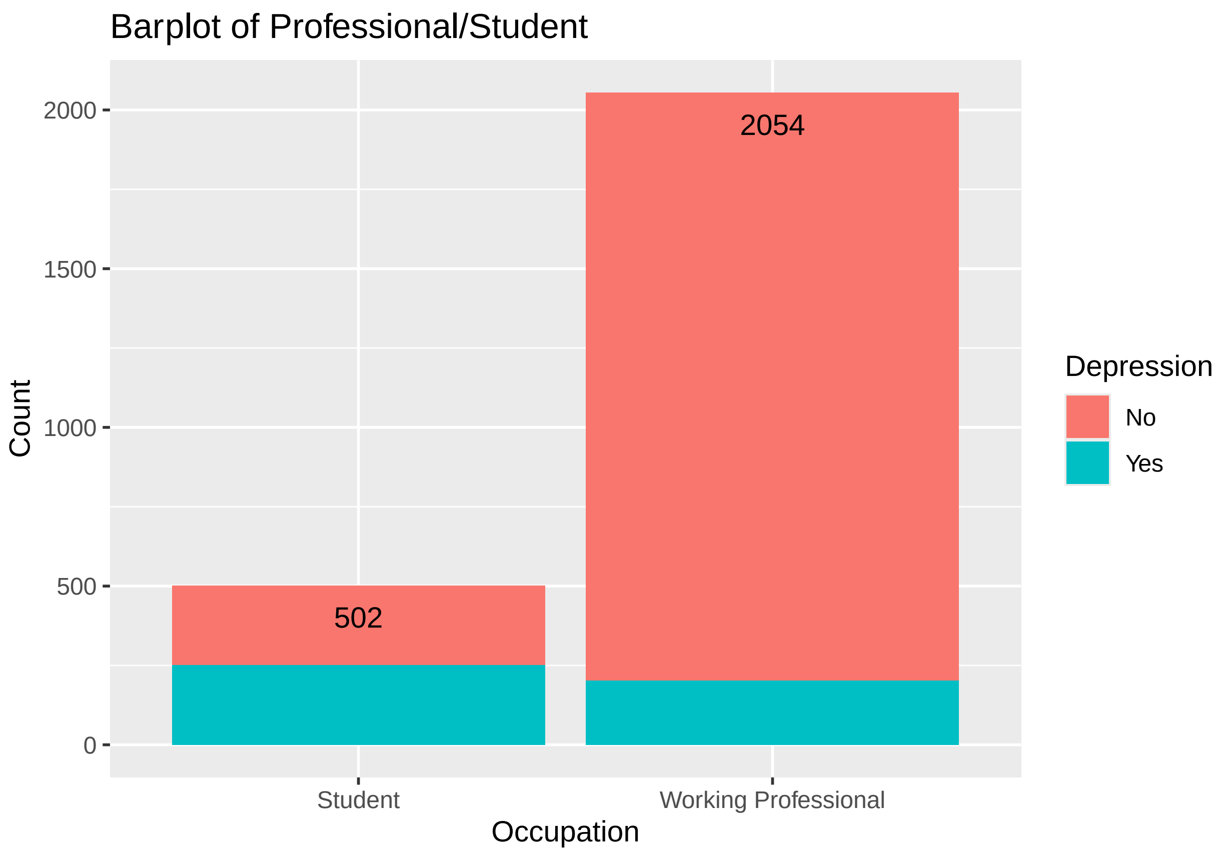
```
# plot gender
ggplot(depression, aes(x = Gender)) +
  geom_bar(aes(fill = Depression)) +
  ylab("Count")
```



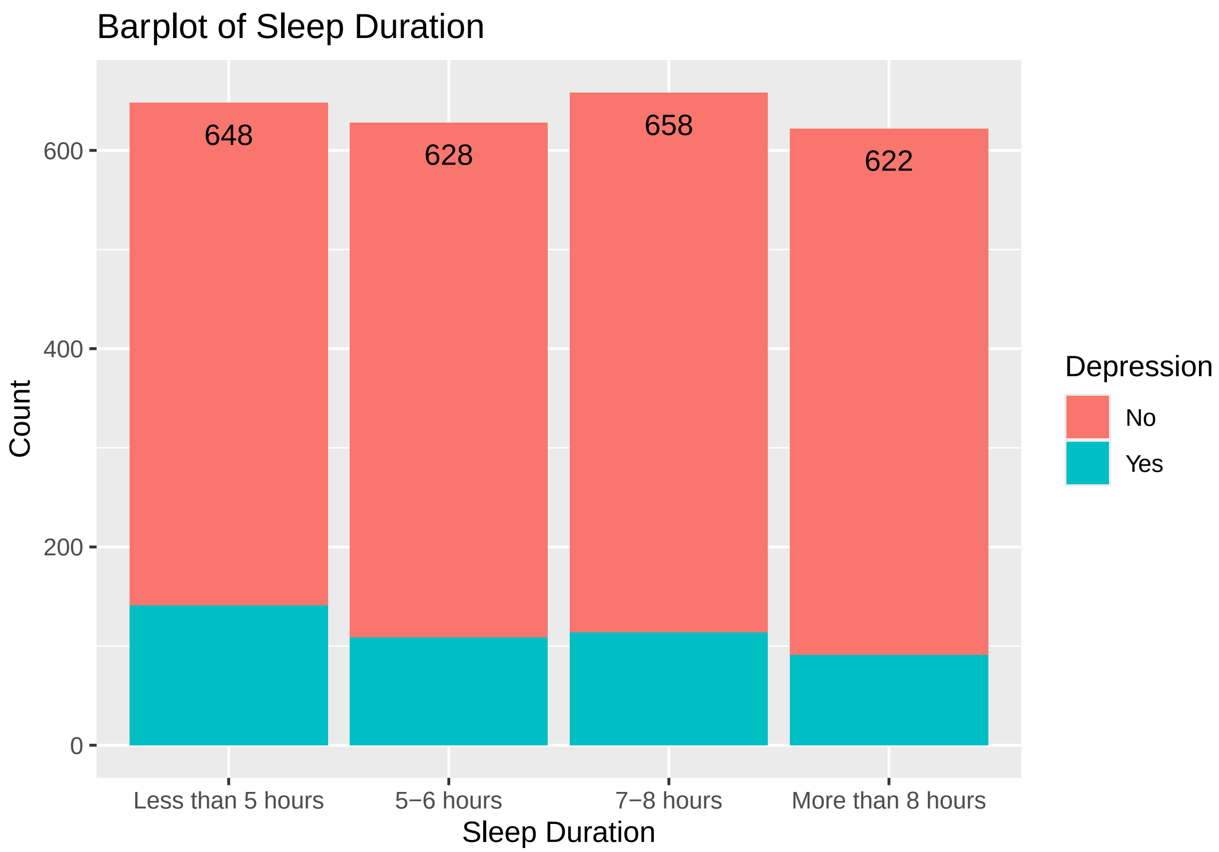
```
  ggtitle("Barplot of Gender") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```

```
## NULL
```

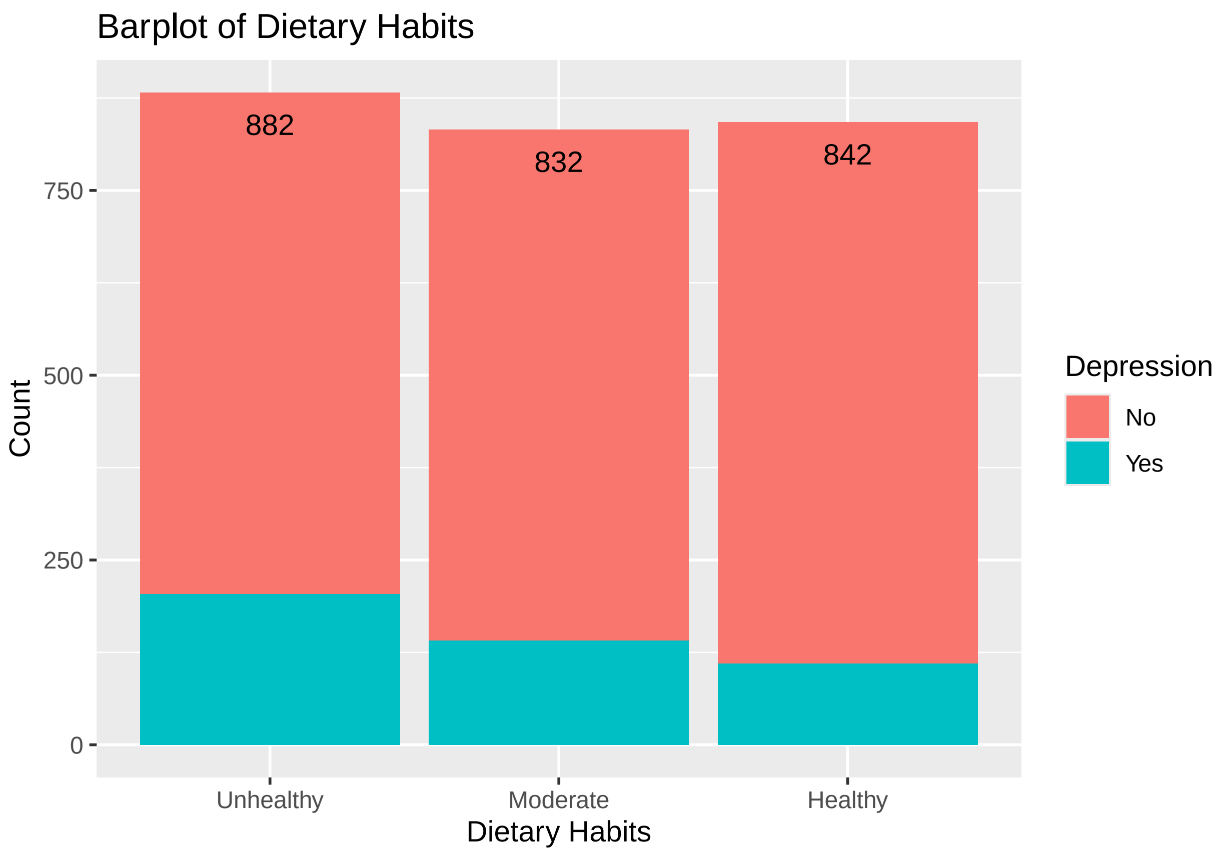
```
# plot whether or not person is a working professional or student
ggplot(depression, aes(x = Working.Professional.or.Student)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Occupation") +
  ylab("Count") +
  ggtitle("Barplot of Professional/Student") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



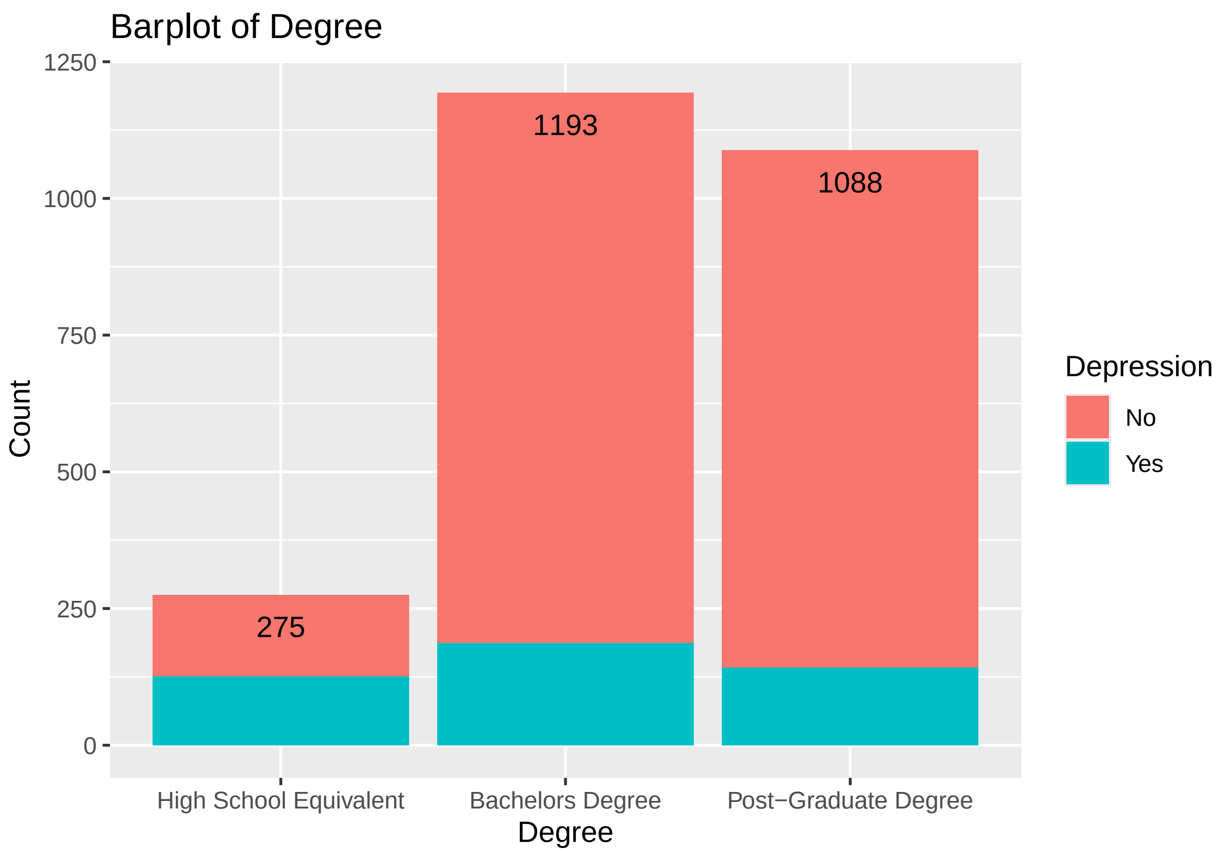
```
# plot sleep duration habits
ggplot(depression, aes(x = Sleep.Duration)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Sleep Duration") +
  ylab("Count") +
  ggtitle("Barplot of Sleep Duration") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



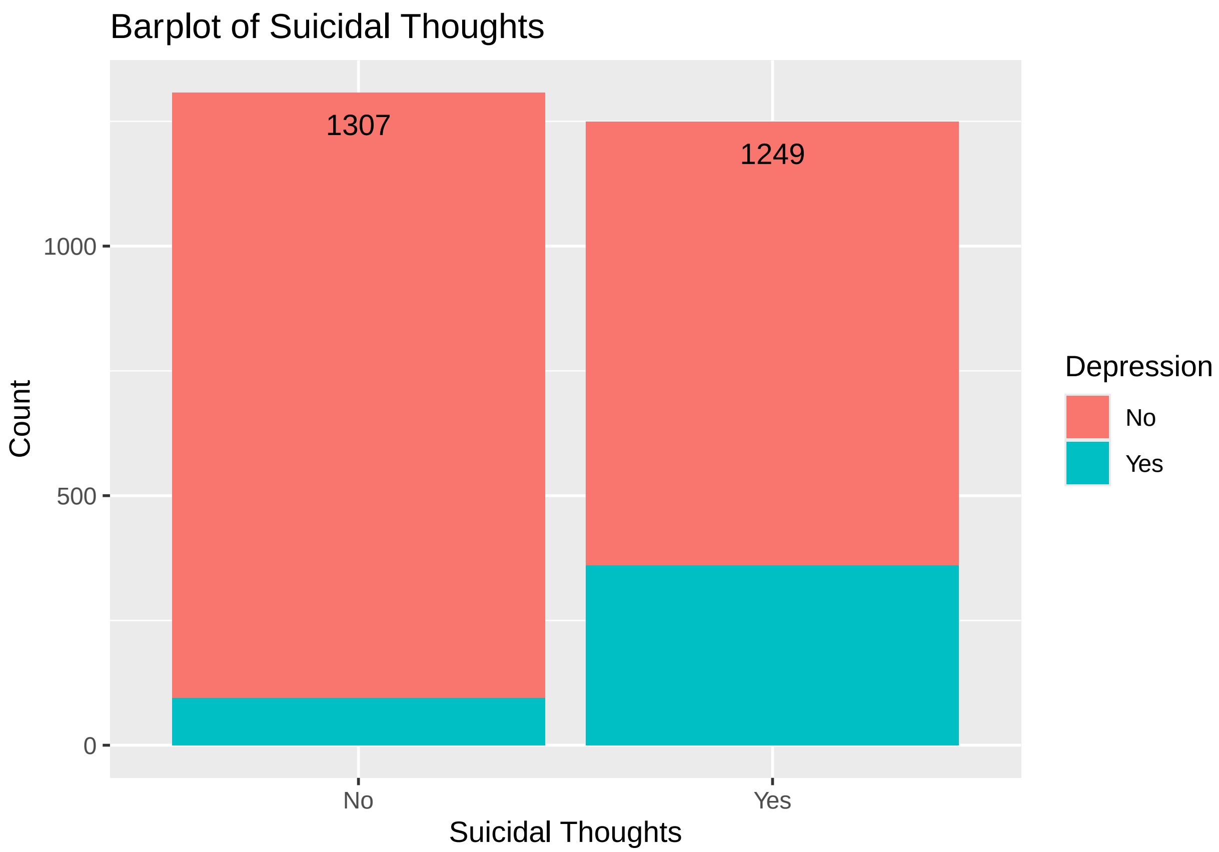
```
# plot dietary habits
ggplot(depression, aes(x = Dietary.Habits)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Dietary Habits") +
  ylab("Count") +
  ggtitle("Barplot of Dietary Habits") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



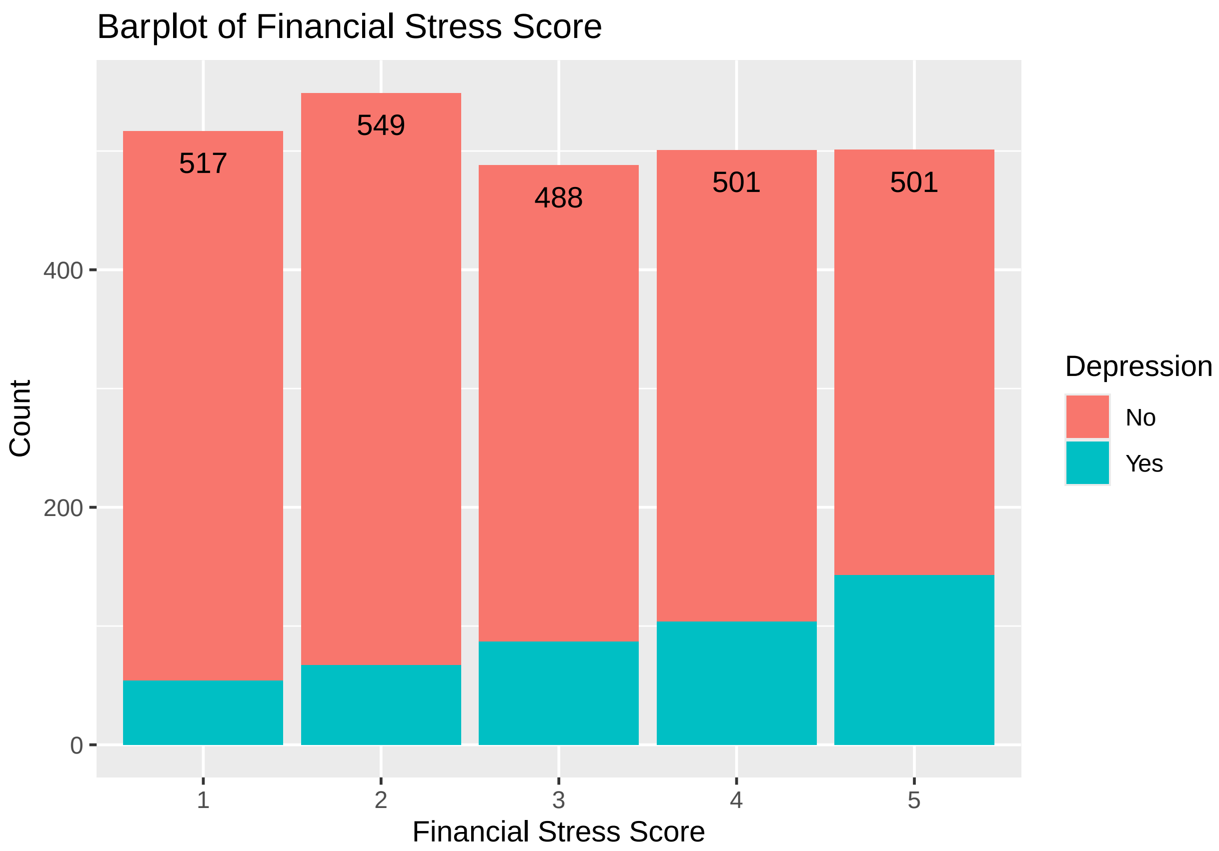
```r
# plot degree count
ggplot(depression, aes(x = Degree)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Degree") +
  ylab("Count") +
  ggtitle("Barplot of Degree") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



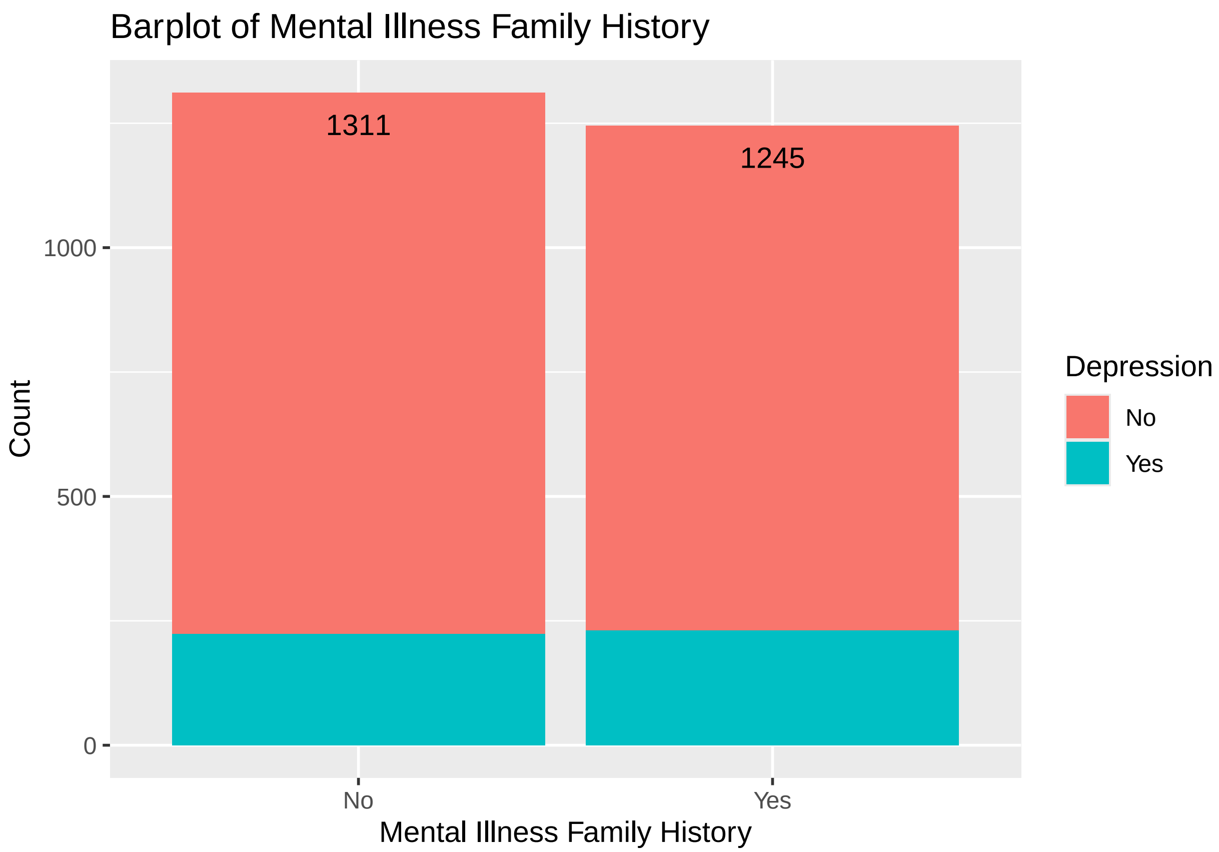
```r
# plot degree count
ggplot(depression, aes(x = Have.you.ever.had.suicidal.thoughts..)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Suicidal Thoughts") +
  ylab("Count") +
  ggtitle("Barplot of Suicidal Thoughts") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



```r
# plot financial stress count
ggplot(depression, aes(x = Financial.Stress)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Financial Stress Score") +
  ylab("Count") +
  ggtitle("Barplot of Financial Stress Score") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```



```r
# plot family history of mental illness count
ggplot(depression, aes(x = Family.History.of.Mental.Illness)) +
  geom_bar(aes(fill = Depression)) +
  xlab("Mental Illness Family History") +
  ylab("Count") +
  ggtitle("Barplot of Mental Illness Family History") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 2)
```

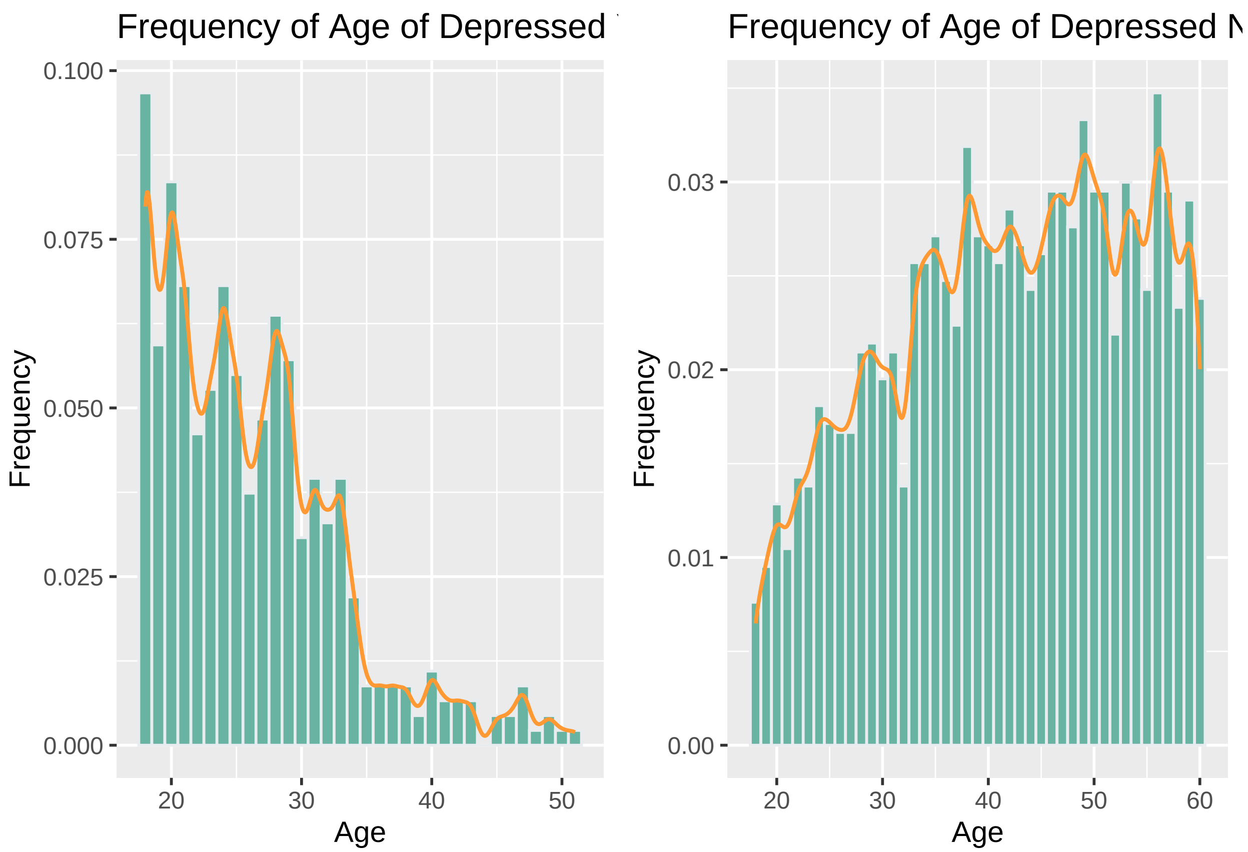


```
depressionYes = depression[depression$Depression == "Yes", ]
depressionNo = depression[depression$Depression == "No", ]
```

```
p1 = ggplot(depressionYes, aes(x = Age, y = after_stat(density))) +
  geom_histogram(binwidth = 1, fill="#69B3A2", color = "#E9ECEF") +
  geom_density(color = "#FF9933", linewidth = 0.7, adjust = 0.3) +
  ggtitle("Frequency of Age of Depressed Yes") +
  ylab("Frequency")

p2 = ggplot(depressionNo, aes(x = Age, y = after_stat(density))) +
  geom_histogram(binwidth = 1, fill="#69B3A2", color = "#E9ECEF") +
  geom_density(color = "#FF9933", linewidth = 0.7, adjust = 0.3) +
  ggtitle("Frequency of Age of Depressed No") +
  ylab("Frequency")
```

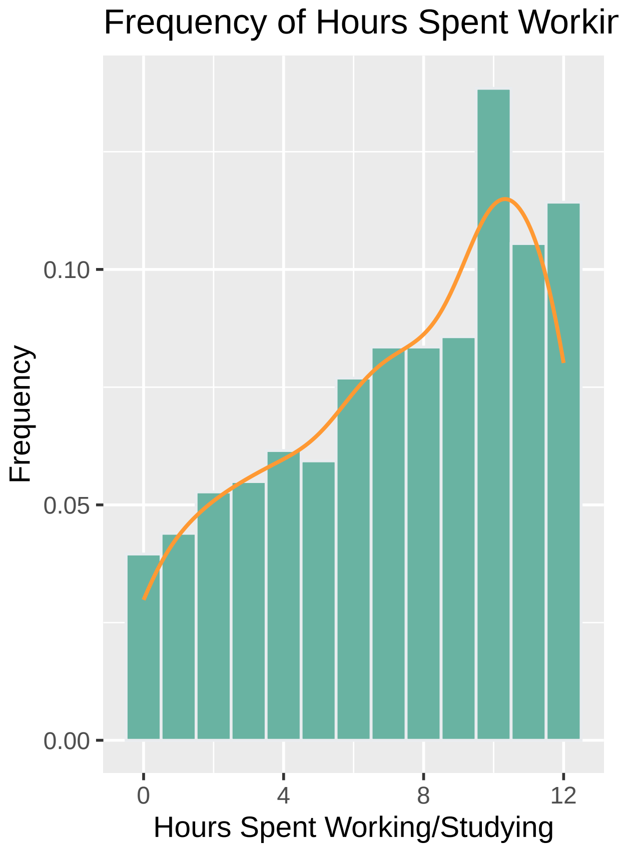
```
plot_grid(p1, p2)
```

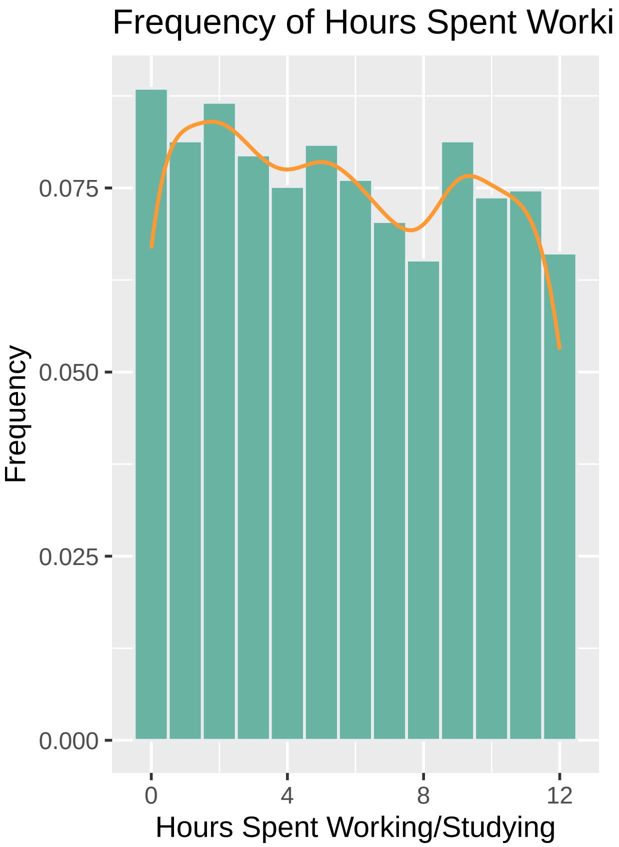


```
p3 = ggplot(depressionYes, aes(x = Work.Study.Hours, y = after_stat(density))) +
  geom_histogram(binwidth = 1, fill="#69B3A2", color = "#E9ECEF") +
  geom_density(color = "#FF9933", linewidth = 0.7, adjust = 1) +
  ggtitle("Frequency of Hours Spent Working/Studying of Depressed Yes") +
  xlab("Hours Spent Working/Studying") +
  ylab("Frequency")

p4 = ggplot(depressionNo, aes(x = Work.Study.Hours, y = after_stat(density))) +
  geom_histogram(binwidth = 1, fill="#69B3A2", color = "#E9ECEF") +
  geom_density(color = "#FF9933", linewidth = 0.7, adjust = 1) +
  ggtitle("Frequency of Hours Spent Working/Studying of Depressed No") +
  xlab("Hours Spent Working/Studying") +
  ylab("Frequency")
```

```
plot_grid(p3, p4)
```



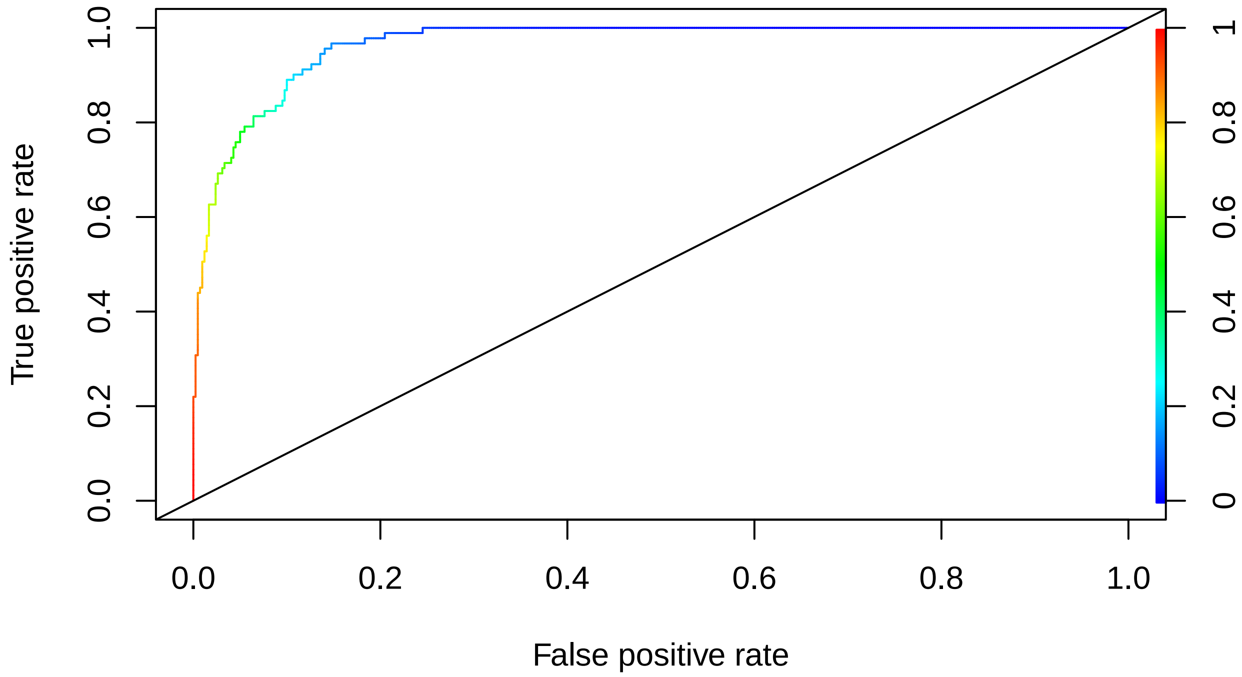


```r
# create train and test set
set.seed(213)
index = createDataPartition(depression$Depression, p = 0.80, list = FALSE, times = 1)
depression_train = depression[index,]
depression_test = depression[-index,]
```

```r
# create model with all predictors (no interaction effects)
depression_glm = glm(Depression ~ ., data = depression_train, family = "binomial")
summary(depression_glm)
```

```
##
## Call:
## glm(formula = Depression ~ ., family = "binomial", data = depression_train)
##
## Coefficients:
##                                                        Estimate Std. Error z value
## (Intercept)                                             2.14025    0.46513   4.601
## GenderMale                                              0.06564    0.17679   0.371
## Age                                                    -0.18223    0.01365 -13.349
## Working.Professional.or.StudentWorking Professional -1.56469    0.20159  -7.762
## Sleep.Duration5-6 hours                                -0.42603    0.23919  -1.781
## Sleep.Duration7-8 hours                                -0.67252    0.24313  -2.766
## Sleep.DurationMore than 8 hours                        -1.09576    0.25715  -4.261
## Dietary.HabitsModerate                                 -0.54505    0.21060  -2.588
## Dietary.HabitsHealthy                                  -1.21926    0.22512  -5.416
## DegreeBachelors Degree                                 -0.45768    0.27516  -1.663
## DegreePost-Graduate Degree                             -0.64609    0.28661  -2.254
## Have.you.ever.had.suicidal.thoughts..Yes             3.03646    0.22686  13.385
## Work.Study.Hours                                     0.19778    0.02528   7.823
## Financial.Stress2                                    0.66448    0.30611   2.171
## Financial.Stress3                                    1.24175    0.29339   4.232
## Financial.Stress4                                    1.56984    0.30520   5.144
## Financial.Stress5                                    2.56514    0.29828   8.600
## Family.History.of.Mental.IllnessYes                  0.60204    0.18084   3.329
##                                                    Pr(>|z|)    
## (Intercept)                                        4.20e-06 ***
## GenderMale                                         0.710441    
## Age                                                 < 2e-16 ***
## Working.Professional.or.StudentWorking Professional 8.37e-15 ***
## Sleep.Duration5-6 hours                            0.074882 .  
## Sleep.Duration7-8 hours                            0.005673 ** 
## Sleep.DurationMore than 8 hours                    2.03e-05 ***
## Dietary.HabitsModerate                             0.009652 ** 
## Dietary.HabitsHealthy                              6.10e-08 ***
## DegreeBachelors Degree                             0.096246 .  
## DegreePost-Graduate Degree                         0.024180 *  
## Have.you.ever.had.suicidal.thoughts..Yes            < 2e-16 ***
## Work.Study.Hours                                   5.15e-15 ***
## Financial.Stress2                                  0.029953 *  
## Financial.Stress3                                  2.31e-05 ***
## Financial.Stress4                                  2.69e-07 ***
## Financial.Stress5                                   < 2e-16 ***
## Family.History.of.Mental.IllnessYes                0.000871 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1915.51  on 2044  degrees of freedom
## Residual deviance:  853.32  on 2027  degrees of freedom
## AIC: 889.32
## 
## Number of Fisher Scoring iterations: 7
```

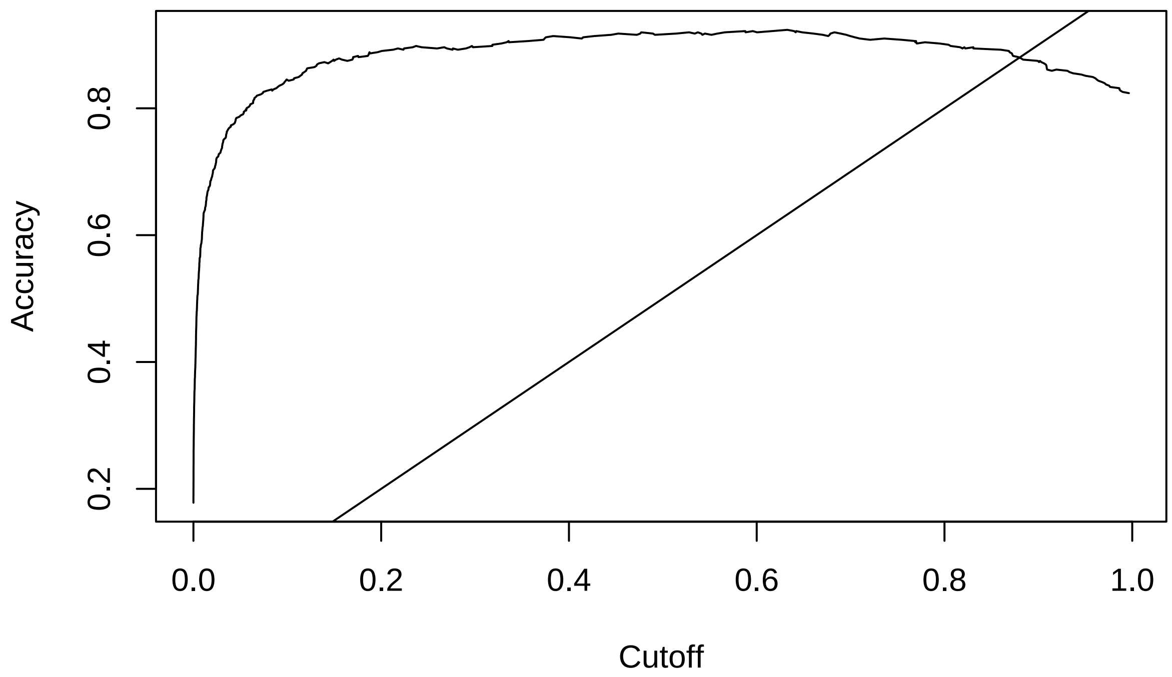
```
# draw a roc curve for true positive rate and true negative rate to find the optimal cutoff
glm_predictions = predict(depression_glm, newdata = depression_test, type = "response")
prob_predictions = prediction(glm_predictions, depression_test$Depression)
roc_curve = performance(prob_predictions, "tpr", "fpr")
plot(roc_curve, colorize = TRUE)
abline(0, 1)
```



```
# auc value
unlist(slot(performance(prob_predictions, "auc"), "y.values"))
```

```
## [1] 0.9652538
```

```
acc = performance(prob_predictions, "acc")
plot(acc)
abline(0, 1)
```



```
glm_predictions2 = predict(depression_glm, newdata = depression_test)
glm_predictions2 = ifelse(glm_predictions2 > 0.30, "Yes", "No")
glm_predictions2 = as.factor(glm_predictions2)
confusionMatrix(glm_predictions2, depression_test$Depression)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  406  26
##        Yes  14  65
##
##                Accuracy : 0.9217
##                  95% CI : (0.8949, 0.9435)
##     No Information Rate : 0.8219
##     P-Value [Acc > NIR] : 7.318e-11
##
##                   Kappa : 0.718
##
##  Mcnemar's Test P-Value : 0.08199
##
##             Sensitivity : 0.9667
##             Specificity : 0.7143
##          Pos Pred Value : 0.9398
##          Neg Pred Value : 0.8228
##              Prevalence : 0.8219
##          Detection Rate : 0.7945
##    Detection Prevalence : 0.8454
##       Balanced Accuracy : 0.8405
##
##        'Positive' Class : No
##
```

```r
train_control = trainControl(method = "repeatedcv", number = 10, repeats = 3, classProbs = TRUE)
depression_cvglm = train(Depression ~ .,
                         data = depression_train,
                         method = "glm",
                         family = binomial,
                         trControl = train_control)
```

```r
depression_cvglm$results
```

```
##   parameter  Accuracy     Kappa AccuracySD    KappaSD
## 1      none 0.8961698 0.6253827 0.02050593 0.07911171
```

```r
cvglm_predictions = predict(depression_cvglm, depression_test)
confusionMatrix(cvglm_predictions, depression_test$Depression)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  400  22
##        Yes  20  69
##
##                Accuracy : 0.9178
##                  95% CI : (0.8905, 0.9401)
##     No Information Rate : 0.8219
##     P-Value [Acc > NIR] : 4.57e-10
##
##                   Kappa : 0.7168
##
##  Mcnemar's Test P-Value : 0.8774
##
##             Sensitivity : 0.9524
##             Specificity : 0.7582
##          Pos Pred Value : 0.9479
##          Neg Pred Value : 0.7753
##              Prevalence : 0.8219
##          Detection Rate : 0.7828
##    Detection Prevalence : 0.8258
##       Balanced Accuracy : 0.8553
##
##        'Positive' Class : No
##
```

```r
varImp(depression_cvglm)
```

```
## glm variable importance
##
##                                                       Overall
## Have.you.ever.had.suicidal.thoughts..Yes              100.000
## Age                                                      99.726
## Financial.Stress5                                        63.229
## Work.Study.Hours                                         57.263
## `Working.Professional.or.StudentWorking Professional`   56.791
## Dietary.HabitsHealthy                                    38.764
## Financial.Stress4                                        36.672
## `Sleep.DurationMore than 8 hours`                        29.890
## Financial.Stress3                                        29.670
## Family.History.of.Mental.IllnessYes                      22.728
## `Sleep.Duration7-8 hours`                                18.402
## Dietary.HabitsModerate                                   17.034
## `DegreePost-Graduate Degree`                             14.469
## Financial.Stress2                                        13.827
## `Sleep.Duration5-6 hours`                                10.834
## `DegreeBachelors Degree`                                  9.929
## GenderMale                                                0.000
```