FACTORS ASSOCIATED WITH ANXIETY

#Clear R environment  
rm(list=ls())  
#Working directory  
setwd("C:/Users/USER/Desktop/Portfolio")  
#Import dataset  
library(readr)  
Anxiety <- read\_csv("Anxiety.csv")

Rows: 12000 Columns: 20  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (7): Gender, Occupation, Smoking, Family History of Anxiety, Dizziness,...  
dbl (13): ID, Age, Sleep Hours, Physical Activity (hrs/week), Caffeine Intak...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

\*\*Data cleaning

Anxiety$Gender<-as.factor(Anxiety$Gender)  
Anxiety$Occupation<-as.factor(Anxiety$Occupation)  
Anxiety$Smoking<-as.factor(Anxiety$Smoking)  
Anxiety$`Family History of Anxiety`<-as.factor(Anxiety$`Family History of Anxiety`)  
Anxiety$Dizziness<-as.factor(Anxiety$Dizziness)  
Anxiety$Medication<-as.factor(Anxiety$Medication)  
Anxiety$`Recent Major Life Event`<-as.factor(Anxiety$`Recent Major Life Event`)

Explanatory Data Analysis for categorical variables

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ purrr 1.0.2  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(summarytools)

Attaching package: 'summarytools'  
  
The following object is masked from 'package:tibble':  
  
 view

*Univariate Analysis for categorical variables*

*Tables*

freq(Anxiety$Gender,report.nas = FALSE)

Frequencies   
Anxiety$Gender   
Type: Factor   
  
 Freq % % Cum.  
------------ ------- -------- --------  
 Female 5809 48.41 48.41  
 Male 5723 47.69 96.10  
 Other 468 3.90 100.00  
 Total 12000 100.00 100.00

freq(Anxiety$`Family History of Anxiety`,report.nas = FALSE)

Frequencies   
Anxiety$`Family History of Anxiety`   
Type: Factor   
  
 Freq % % Cum.  
----------- ------- -------- --------  
 No 7179 59.82 59.82  
 Yes 4821 40.17 100.00  
 Total 12000 100.00 100.00

freq(Anxiety$Occupation,report.nas = FALSE)

Frequencies   
Anxiety$Occupation   
Type: Factor   
  
 Freq % % Cum.  
---------------- ------- -------- --------  
 Doctor 2004 16.70 16.70  
 Engineer 1953 16.28 32.98  
 Other 1971 16.43 49.40  
 Student 1953 16.28 65.68  
 Teacher 1980 16.50 82.18  
 Unemployed 2139 17.82 100.00  
 Total 12000 100.00 100.00

freq(Anxiety$Smoking,report.nas = FALSE)

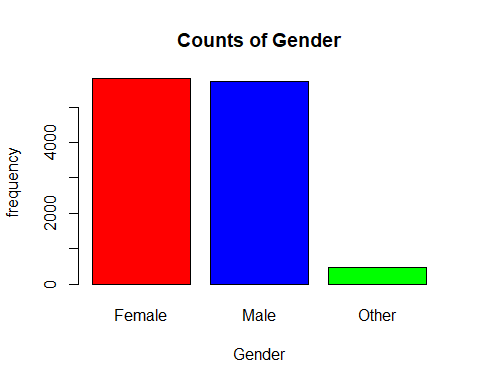
Frequencies   
Anxiety$Smoking   
Type: Factor   
  
 Freq % % Cum.  
----------- ------- -------- --------  
 No 8417 70.14 70.14  
 Yes 3583 29.86 100.00  
 Total 12000 100.00 100.00

freq(Anxiety$Dizziness,report.nas = FALSE)

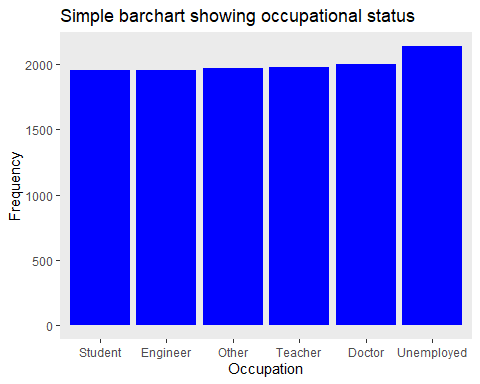
Frequencies   
Anxiety$Dizziness   
Type: Factor   
  
 Freq % % Cum.  
----------- ------- -------- --------  
 No 8406 70.05 70.05  
 Yes 3594 29.95 100.00  
 Total 12000 100.00 100.00

Simple Bar charts

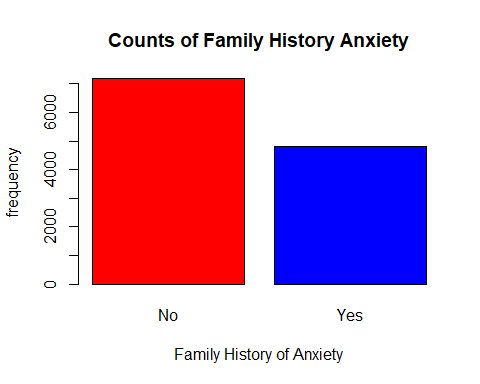
Counts<-table(Anxiety$Gender)  
barplot(Counts,  
 main="Counts of Gender",  
 xlab="Gender",  
 ylab="frequency",  
 legend=rownames(count),  
 col=c("red","blue","green"))



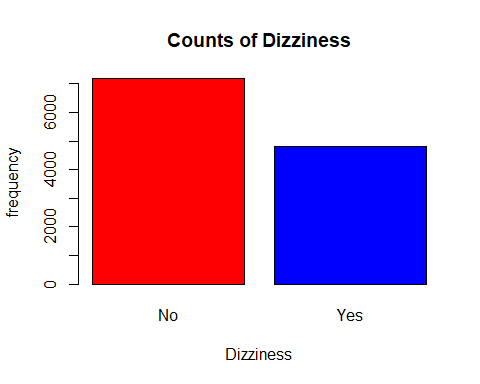
Anxiety%>%  
 mutate(Occupation=fct\_infreq(Occupation))%>%  
 mutate(Occupation=fct\_rev(Occupation))%>%  
 ggplot(aes(Occupation))+  
 geom\_bar(fill="blue")+  
 theme()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank())+  
labs(title = "Simple barchart showing occupational status",  
 x="Occupation",  
 y="Frequency")



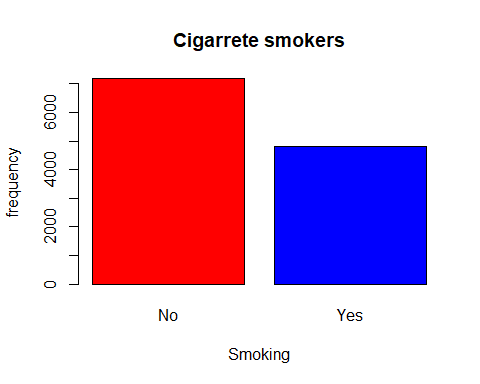
Counts<-table(Anxiety$`Family History of Anxiety`)  
barplot(Counts,  
 main="Counts of Family History Anxiety",  
 xlab="Family History of Anxiety",  
 ylab="frequency",  
 legend=rownames(count),  
 col=c("red","blue"))



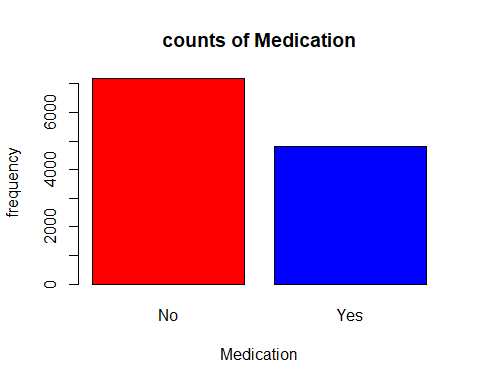
counts<-table(Anxiety$Dizziness)  
barplot(Counts,  
 main="Counts of Dizziness",  
 xlab="Dizziness",  
 ylab="frequency",  
 legend=rownames(count),  
 col=c("red","blue"))



counts<-table(Anxiety$Smoking)  
barplot(Counts,  
 main="Cigarrete smokers",  
 xlab="Smoking",  
 ylab="frequency",  
 legend=rownames(count),  
 col=c("red","blue"))



counts<-table(Anxiety$Medication)  
barplot(Counts,  
 main="counts of Medication",  
 xlab="Medication",  
 ylab="frequency",  
 legend=rownames(count),  
 col=c("red","blue"))



Bivariate Analysis

Association between categorical variables using chi square test of independent .

Outcome variable:Family history of anxiety

ctable(  
 x=Anxiety$Gender,  
 y=Anxiety$`Family History of Anxiety`,  
 chisq = TRUE,  
 headings=FALSE  
)

-------- --------------------------- -------------- -------------- ----------------  
 Family History of Anxiety No Yes Total  
 Gender   
 Female 3498 (60.2%) 2311 (39.8%) 5809 (100.0%)  
 Male 3426 (59.9%) 2297 (40.1%) 5723 (100.0%)  
 Other 255 (54.5%) 213 (45.5%) 468 (100.0%)  
 Total 7179 (59.8%) 4821 (40.2%) 12000 (100.0%)  
-------- --------------------------- -------------- -------------- ----------------  
  
----------------------------  
 Chi.squared df p.value   
------------- ---- ---------  
 5.9228 2 0.0517   
----------------------------

ctable(  
 x=Anxiety$Occupation,  
 y=Anxiety$`Family History of Anxiety`,  
 chisq = TRUE,  
 headings=FALSE  
)

------------ --------------------------- -------------- -------------- ----------------  
 Family History of Anxiety No Yes Total  
 Occupation   
 Doctor 1186 (59.2%) 818 (40.8%) 2004 (100.0%)  
 Engineer 1178 (60.3%) 775 (39.7%) 1953 (100.0%)  
 Other 1184 (60.1%) 787 (39.9%) 1971 (100.0%)  
 Student 1172 (60.0%) 781 (40.0%) 1953 (100.0%)  
 Teacher 1164 (58.8%) 816 (41.2%) 1980 (100.0%)  
 Unemployed 1295 (60.5%) 844 (39.5%) 2139 (100.0%)  
 Total 7179 (59.8%) 4821 (40.2%) 12000 (100.0%)  
------------ --------------------------- -------------- -------------- ----------------  
  
----------------------------  
 Chi.squared df p.value   
------------- ---- ---------  
 1.9637 5 0.8541   
----------------------------

ctable(  
 x=Anxiety$Smoking,  
 y=Anxiety$`Family History of Anxiety`,  
 chisq = TRUE,  
 headings=FALSE  
)

--------- --------------------------- -------------- -------------- ----------------  
 Family History of Anxiety No Yes Total  
 Smoking   
 No 5030 (59.8%) 3387 (40.2%) 8417 (100.0%)  
 Yes 2149 (60.0%) 1434 (40.0%) 3583 (100.0%)  
 Total 7179 (59.8%) 4821 (40.2%) 12000 (100.0%)  
--------- --------------------------- -------------- -------------- ----------------  
  
----------------------------  
 Chi.squared df p.value   
------------- ---- ---------  
 0.0409 1 0.8397   
----------------------------

ctable(  
 x=Anxiety$Dizziness,  
 y=Anxiety$`Family History of Anxiety`,  
 chisq = TRUE,  
 headings=FALSE  
)

----------- --------------------------- -------------- -------------- ----------------  
 Family History of Anxiety No Yes Total  
 Dizziness   
 No 4995 (59.4%) 3411 (40.6%) 8406 (100.0%)  
 Yes 2184 (60.8%) 1410 (39.2%) 3594 (100.0%)  
 Total 7179 (59.8%) 4821 (40.2%) 12000 (100.0%)  
----------- --------------------------- -------------- -------------- ----------------  
  
----------------------------  
 Chi.squared df p.value   
------------- ---- ---------  
 1.8424 1 0.1747   
----------------------------

ctable(  
 x=Anxiety$Medication,  
 y=Anxiety$`Family History of Anxiety`,  
 chisq = TRUE,  
 headings=FALSE  
)

------------ --------------------------- -------------- -------------- ----------------  
 Family History of Anxiety No Yes Total  
 Medication   
 No 5764 (60.0%) 3841 (40.0%) 9605 (100.0%)  
 Yes 1415 (59.1%) 980 (40.9%) 2395 (100.0%)  
 Total 7179 (59.8%) 4821 (40.2%) 12000 (100.0%)  
------------ --------------------------- -------------- -------------- ----------------  
  
----------------------------  
 Chi.squared df p.value   
------------- ---- ---------  
 0.6502 1 0.42   
----------------------------

Explanatory Data Analysis for continuous variables

Univariate Analysis

Sstat<-Anxiety%>%  
 select(Age,`Sleep Hours`,`Physical Activity (hrs/week)`,`Caffeine Intake (mg/day)`,`Alcohol Consumption (drinks/week)`,`Stress Level (1-10)`,`Heart Rate (bpm during attack)`,`Breathing Rate (breaths/min)`,`Sweating Level (1-5)`,`Therapy Sessions (per month)`,`Diet Quality (1-10)`,`Severity of Anxiety Attack (1-10)`)

library(psych)

Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':  
  
 %+%, alpha

Summary Statistics for continuous variables

describe(Sstat)

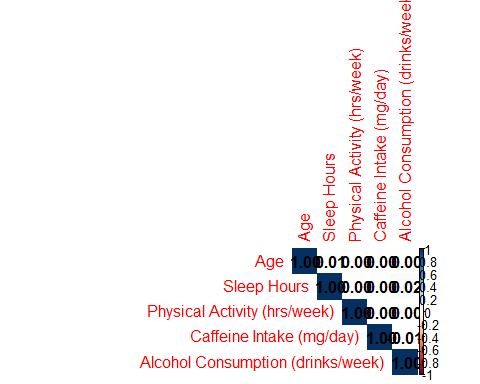
vars n mean sd median trimmed  
Age 1 12000 40.97 13.47 41.0 40.96  
Sleep Hours 2 12000 6.48 2.01 6.5 6.48  
Physical Activity (hrs/week) 3 12000 5.03 2.89 5.0 5.04  
Caffeine Intake (mg/day) 4 12000 246.70 144.49 244.0 246.23  
Alcohol Consumption (drinks/week) 5 12000 9.49 5.77 9.0 9.49  
Stress Level (1-10) 6 12000 5.46 2.90 5.0 5.45  
Heart Rate (bpm during attack) 7 12000 119.40 34.81 119.0 119.39  
Breathing Rate (breaths/min) 8 12000 25.46 8.09 25.0 25.45  
Sweating Level (1-5) 9 12000 2.99 1.41 3.0 2.98  
Therapy Sessions (per month) 10 12000 4.52 2.87 5.0 4.52  
Diet Quality (1-10) 11 12000 5.50 2.87 5.0 5.49  
Severity of Anxiety Attack (1-10) 12 12000 5.51 2.86 6.0 5.51  
 mad min max range skew kurtosis se  
Age 17.79 18 64 46 0.01 -1.18 0.12  
Sleep Hours 2.52 3 10 7 0.02 -1.18 0.02  
Physical Activity (hrs/week) 3.71 0 10 10 -0.01 -1.20 0.03  
Caffeine Intake (mg/day) 185.32 0 499 499 0.02 -1.20 1.32  
Alcohol Consumption (drinks/week) 7.41 0 19 19 0.01 -1.20 0.05  
Stress Level (1-10) 4.45 1 10 9 0.01 -1.24 0.03  
Heart Rate (bpm during attack) 44.48 60 179 119 0.00 -1.22 0.32  
Breathing Rate (breaths/min) 10.38 12 39 27 0.01 -1.21 0.07  
Sweating Level (1-5) 1.48 1 5 4 0.01 -1.30 0.01  
Therapy Sessions (per month) 2.97 0 9 9 0.01 -1.21 0.03  
Diet Quality (1-10) 2.97 1 10 9 0.01 -1.22 0.03  
Severity of Anxiety Attack (1-10) 2.97 1 10 9 -0.01 -1.21 0.03

Relationship between continuous variables

library(corrplot)

corrplot 0.95 loaded

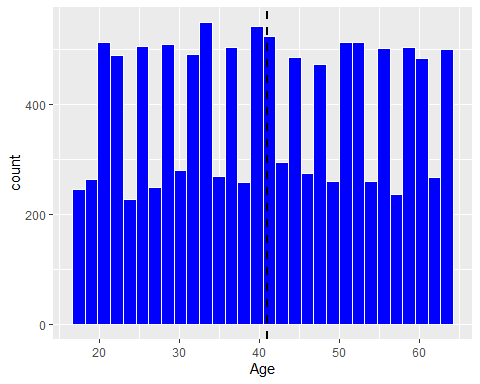
corr <-Anxiety%>%   
 select(Age,`Sleep Hours`,`Physical Activity (hrs/week)`,`Caffeine Intake (mg/day)`,`Alcohol Consumption (drinks/week)`) %>%   
 replace(is.na(.), 0)  
correlation = cor(corr)  
corrplot(correlation, type="upper", method="color", addCoef.col = "black")



Normality test

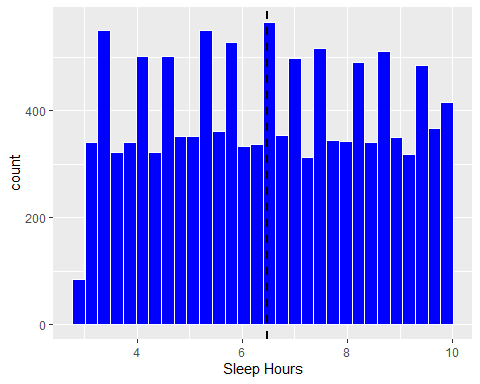
#Age  
ggplot(Anxiety,aes(x=Age))+  
 geom\_histogram(fill="blue",color="white")+  
 geom\_vline(aes(xintercept=mean(Age)),color="black",  
 linewidth=1,linetype="dashed")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



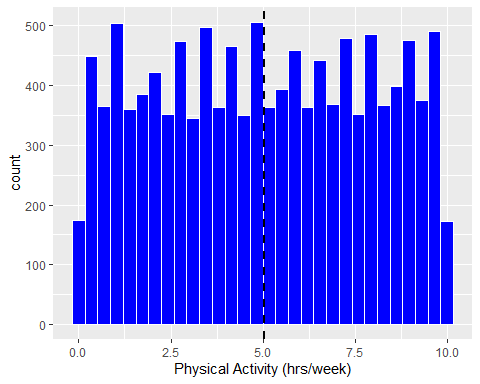
#Sleeping Hours  
ggplot(Anxiety,aes(`Sleep Hours`))+  
 geom\_histogram(fill="blue",color="white")+  
 geom\_vline(aes(xintercept=mean(`Sleep Hours`)),color="black",  
 linewidth=1,linetype="dashed")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



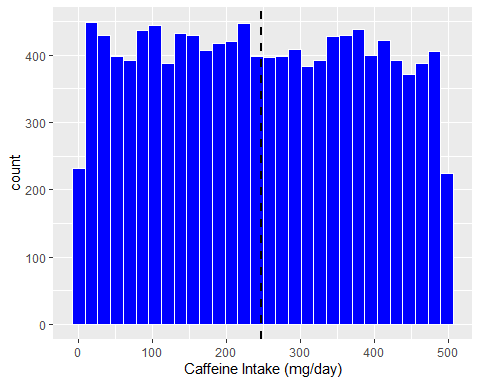
ggplot(Anxiety,aes(`Physical Activity (hrs/week)`))+  
 geom\_histogram(fill="blue",color="white")+  
 geom\_vline(aes(xintercept=mean(`Physical Activity (hrs/week)`)),color="black",  
 linewidth=1,linetype="dashed")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(Anxiety,aes(`Caffeine Intake (mg/day)`))+  
 geom\_histogram(fill="blue",color="white")+  
 geom\_vline(aes(xintercept=mean(`Caffeine Intake (mg/day)`)),color="black",  
 linewidth=1,linetype="dashed")

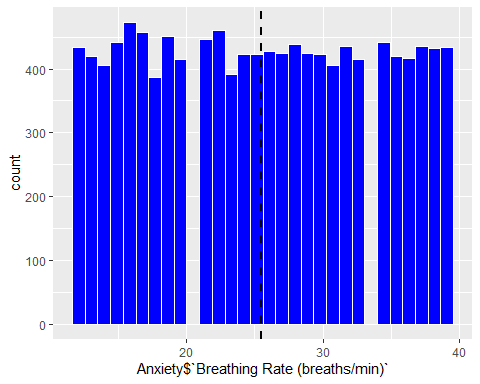
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Histogram of Breathing rate

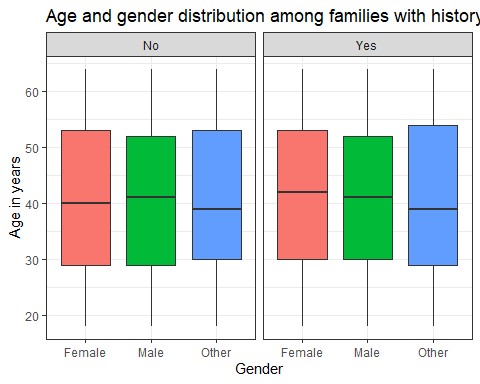
ggplot(Anxiety,aes(Anxiety$`Breathing Rate (breaths/min)`))+  
 geom\_histogram(fill="blue",color="white")+  
 geom\_vline(aes(xintercept=mean(`Breathing Rate (breaths/min)`)),color="black",  
 linewidth=1,linetype="dashed")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



EDA between categorical variables and continuous variables

ggplot(Anxiety,aes(Gender,Age,fill=Gender))+  
 geom\_boxplot(show.legend = FALSE)+  
 facet\_wrap(~`Family History of Anxiety`)+  
 theme\_bw()+  
 labs(title="Age and gender distribution among families with history of anxiety",y="Age in years")



ggplot(Anxiety,aes(Gender,`Sleep Hours`,fill=Gender))+  
 geom\_boxplot(show.legend = FALSE)+  
 facet\_wrap(~`Family History of Anxiety`)+  
 theme\_bw()+  
 labs(title="Sleeping hours and gender distribution among families with history of anxiety",y="Sleeping hours")

