Data Science Analysis

2023-09-05

library(readr)  
shamiri\_imputed\_dataset <- read\_csv("C:/Users/USER/Downloads/shamiri\_imputed\_dataset.csv")

## Rows: 658 Columns: 36  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): ParticipantID, Tribe, Gender, School, School\_Resources  
## dbl (31): PHQ1, PHQ2, PHQ3, PHQ4, PHQ5, PHQ6, PHQ7, PHQ8, Av\_PHQ, GAD1, GAD2...  
##   
## 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.

View(shamiri\_imputed\_dataset)  
knitr::opts\_chunk$set(echo = TRUE)

## Data Pre-processing

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

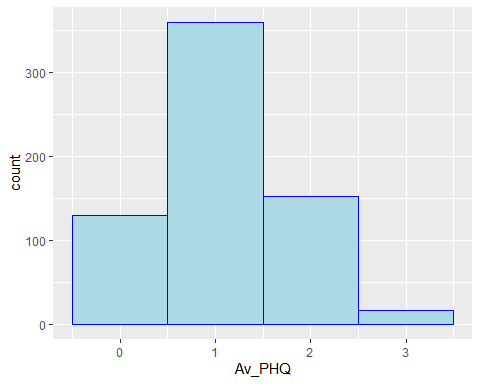
shamiri\_imputed\_dataset %>% select(c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-11,-12,-13,-14,-15,-16,-17,-19,-20,-21,-22,-23,-24,-25,-26,-27,-28,-29,-30))->shamiri\_dataset  
shamiri\_dataset$Gender<-factor(shamiri\_dataset$Gender,labels = c("Female","Male"))  
View(shamiri\_dataset)

## Data Visualization

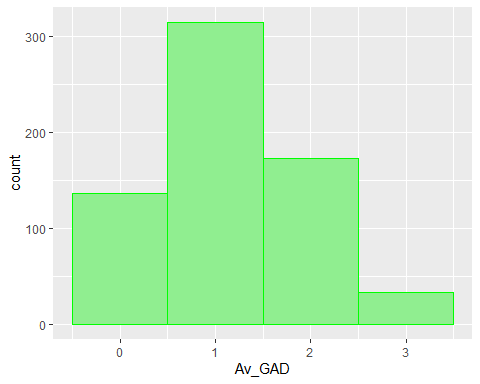
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

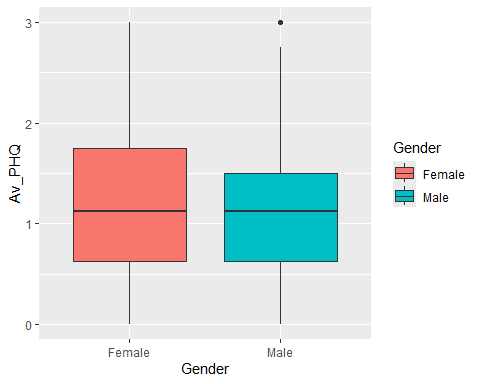
attach(shamiri\_dataset)  
ggplot(data = shamiri\_dataset, aes(x=Av\_PHQ))+geom\_histogram(bins=4, fill="lightblue", col="blue")



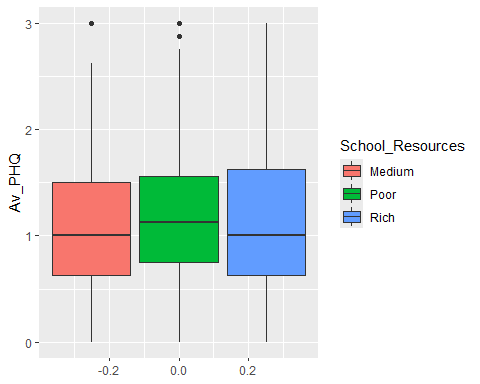
ggplot(data = shamiri\_dataset, aes(x=Av\_GAD))+geom\_histogram(bins = 4, fill="lightgreen", col="green")



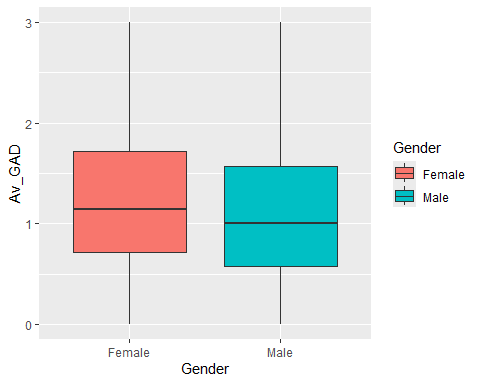
ggplot(data = shamiri\_dataset, aes(y=Av\_PHQ, x=Gender, fill=Gender))+geom\_boxplot()



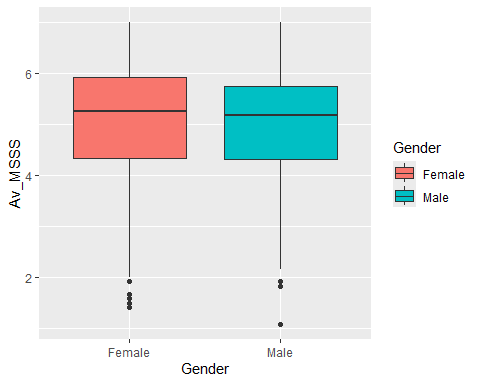
ggplot(data = shamiri\_dataset, aes(y=Av\_PHQ, X=School\_Resources, fill=School\_Resources))+geom\_boxplot()



ggplot(data = shamiri\_dataset, aes(y=Av\_GAD, x=Gender, fill=Gender))+geom\_boxplot()

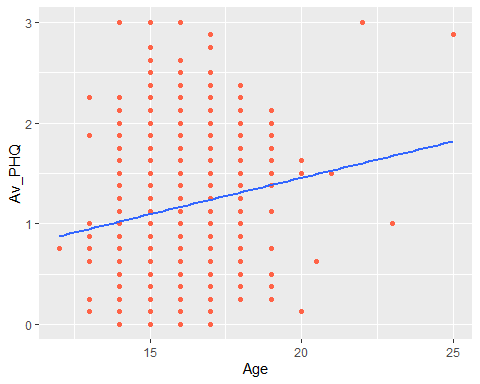


ggplot(data = shamiri\_dataset, aes(y=Av\_MSSS, x=Gender, fill=Gender))+geom\_boxplot()



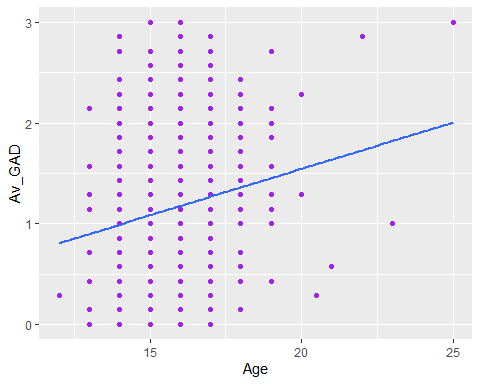
ggplot(data = shamiri\_dataset, aes(y=Av\_PHQ, x=Age))+geom\_point(col="tomato")+geom\_smooth(method = "lm", se=F)

## `geom\_smooth()` using formula = 'y ~ x'



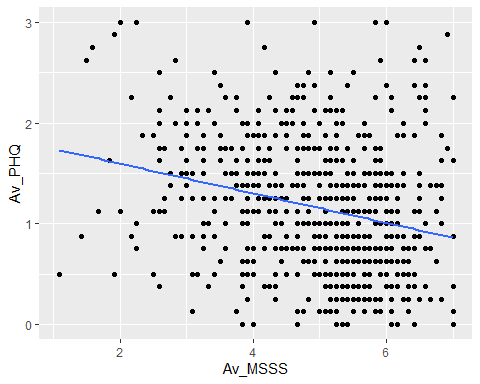
ggplot(data = shamiri\_dataset, aes(x=Age, y=Av\_GAD))+geom\_point(col="purple")+geom\_smooth(method = "lm", se=F)

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(data = shamiri\_dataset, aes(x=Av\_MSSS, y=Av\_PHQ))+geom\_point(col="black")+geom\_smooth(method = "lm", se=F)

## `geom\_smooth()` using formula = 'y ~ x'



## Splitting Data

library(caTools)

## Warning: package 'caTools' was built under R version 4.2.3

sample.split(shamiri\_dataset$Av\_PHQ, SplitRatio = 0.65)->split\_index  
train<-subset(shamiri\_dataset,split\_index==T)  
test<-subset(shamiri\_dataset,split\_index==F)  
nrow(train)

## [1] 427

nrow(test)

## [1] 231

## Model Building

lm(Av\_PHQ~.,data = train)->mod1  
predict(mod1,test)->result

## Warning in predict.lm(mod1, test): prediction from a rank-deficient fit may be  
## misleading

cbind(actual=test$Av\_PHQ,predicted=result)->compare\_result  
as.data.frame(compare\_result)->compare\_result  
compare\_result$actual-compare\_result$predicted->error  
cbind(compare\_result,error)->compare\_result  
sqrt(mean(compare\_result$error^2))->rmse1  
print(rmse1)

## [1] 0.4473839

summary(mod1)

##   
## Call:  
## lm(formula = Av\_PHQ ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.16617 -0.28194 -0.01652 0.25031 1.63026   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.581949 0.305246 1.906 0.057273 .   
## Av\_GAD 0.633955 0.031495 20.129 < 2e-16 \*\*\*  
## Av\_MSSS -0.072950 0.019609 -3.720 0.000226 \*\*\*  
## TribeMinority -0.042467 0.052815 -0.804 0.421809   
## GenderMale -0.005642 0.059207 -0.095 0.924127   
## SchoolAHS 0.026558 0.089813 0.296 0.767606   
## SchoolElite -0.043732 0.089165 -0.490 0.624065   
## SchoolOlympic -0.030687 0.068786 -0.446 0.655743   
## SchoolStarays -0.095849 0.098517 -0.973 0.331159   
## Age 0.016320 0.018513 0.882 0.378529   
## School\_ResourcesPoor NA NA NA NA   
## School\_ResourcesRich NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4479 on 417 degrees of freedom  
## Multiple R-squared: 0.5392, Adjusted R-squared: 0.5293   
## F-statistic: 54.22 on 9 and 417 DF, p-value: < 2.2e-16

lm(Av\_PHQ~.-Tribe-Gender-School\_Resources-Age-School,data=train)->mod2  
predict(mod2,test)->result2  
cbind.data.frame(actual=test$Av\_PHQ,predicted=result2)->compare\_result2  
compare\_result2$actual-compare\_result2$predicted->error  
cbind(compare\_result2,error)->compare\_result2  
sqrt(mean(compare\_result2$error^2))->rmse2  
print(rmse2)

## [1] 0.4513124

summary(mod2)

##   
## Call:  
## lm(formula = Av\_PHQ ~ . - Tribe - Gender - School\_Resources -   
## Age - School, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.13631 -0.28006 -0.03105 0.24507 1.72198   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.79230 0.11019 7.190 2.95e-12 \*\*\*  
## Av\_GAD 0.62941 0.03033 20.753 < 2e-16 \*\*\*  
## Av\_MSSS -0.07291 0.01922 -3.794 0.00017 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4463 on 424 degrees of freedom  
## Multiple R-squared: 0.5348, Adjusted R-squared: 0.5326   
## F-statistic: 243.7 on 2 and 424 DF, p-value: < 2.2e-16