Mathematical Statistics. A statistical analysis of NHANES variables. Part *1.*

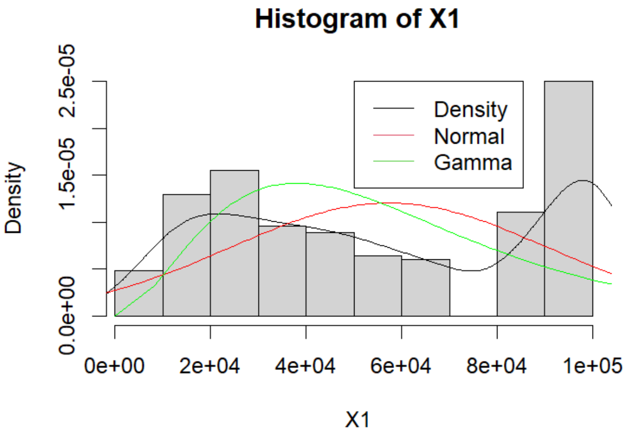
1. Perform a descriptive analysis of the variable X1. Estimate the density function of the variable using the sample and compare it with a Normal and a Gamma distribution, with the same mean and variance as the sample. Perform a descriptive analysis of X2. Perform a descriptive analysis of the variables F1 and F2.

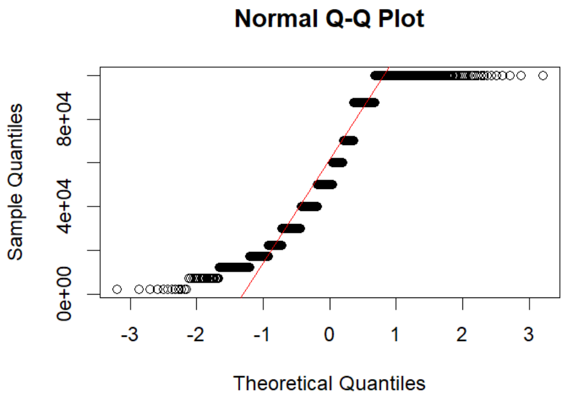
We have selected the following variables:

- **X1**: Average income of each person
- **X2**: Number of days that a person drinks alcohol a year-
- **F1**: "YES" or "NO" if the person consumes marihuana regularly
- **F2**: Education.

SUMMARY OF X1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2500	30000	50000	57364	93750	100000	65

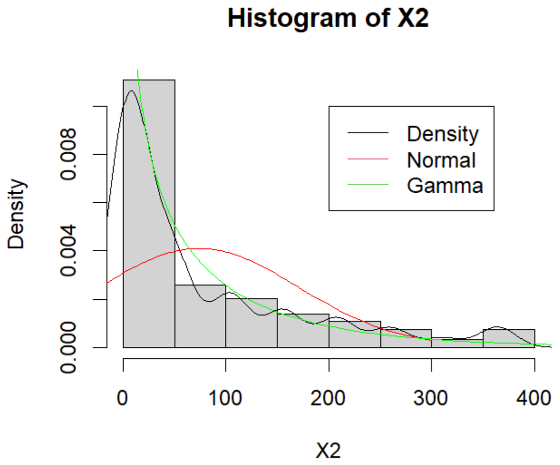


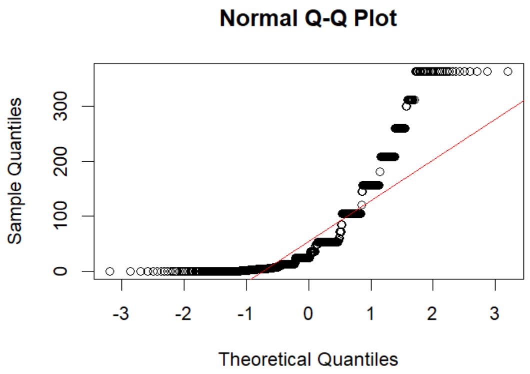


We see that it does not follow a normal nor a gamma distribution. However, he data may be biased, since nobody has an income in the range 70000-80000 dollars per year.

SUMMARY OF X2 :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	4.00	24.00	74.07	104.00	364.00	65



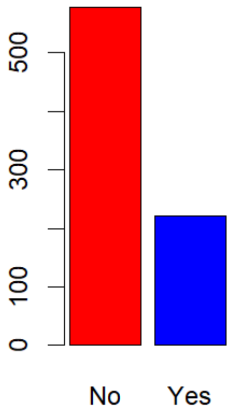


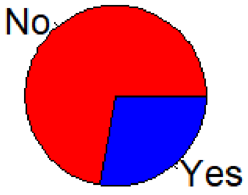
We can see that its distribution is very asymmetrical. Thus it does not follow a normal distribution, but a gamma one.

FACTOR F1

There are 578 people that do not consume marihuana regularly, and 222 that does. That is **72.25%** of people who does **not consume** it and **27.75% who does**.

We see it in a barplot.

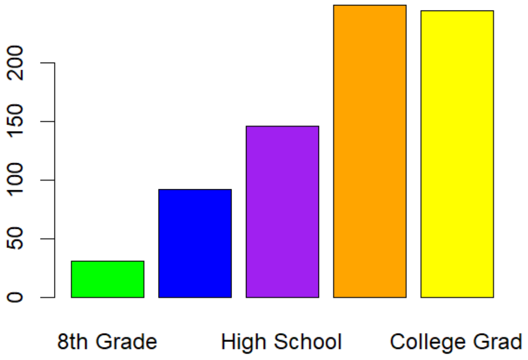


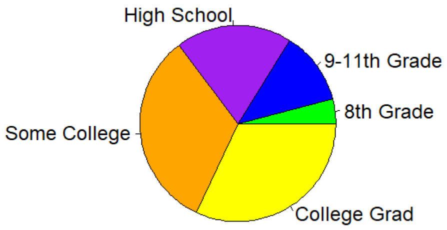


FACTOR F2.

We see the number of people in each group and its percentage.

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Frequency	**31**	**92**	**146**	**249**	**244**
Percentages	**3.875**	**11.500**	**18.250**	**31.125**	**30.500**



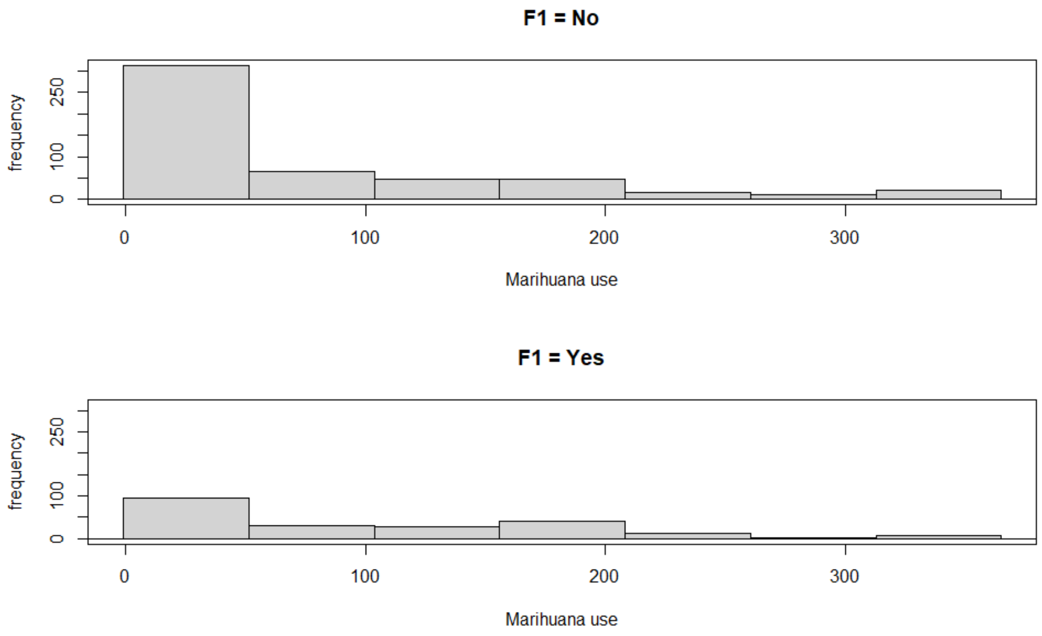


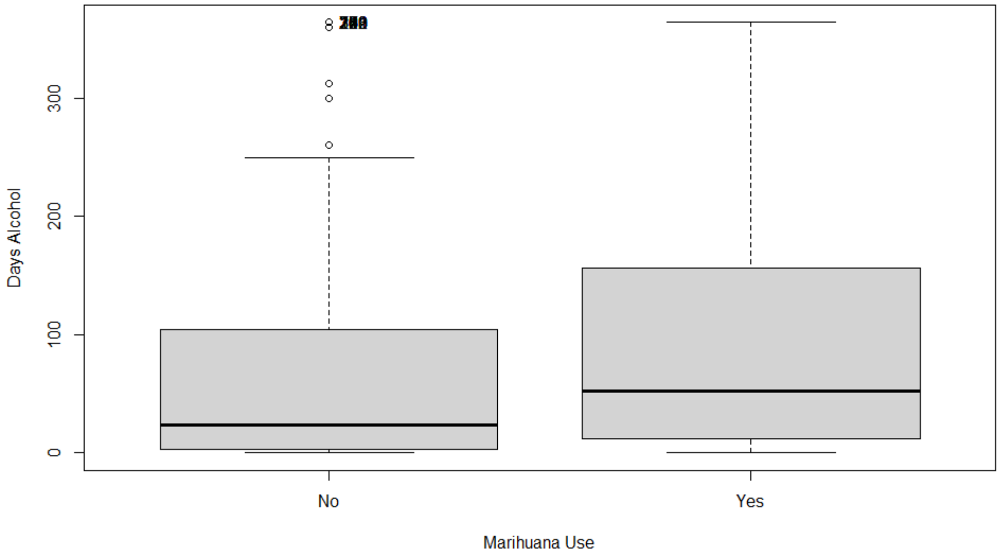
2. Perform a descriptive analysis to study whether there is a dependency between the quantitative variable X1 and the factor F1.

We think there might be some kind of relationship between the alcohol and the marihuana use. To analyse its dependency, we use first a contingency table to study the number of day that a person drinks alcohol in a year conditioned to the regular use of marihuana.

	mean	sd	IQR	0%	25%	50%	75%	100%	n	NA
DOES NOT CONSUME	66.91	96.36	101	0	3	24	104	364	518	60
CONSUMES	91.17	97.19	144	0	12	52	156	364	217	5

We see that, in average, people who consume marihuana drink more alcohol. This may indicate that the variables are related. We represent it in an histogram and a boxplot.



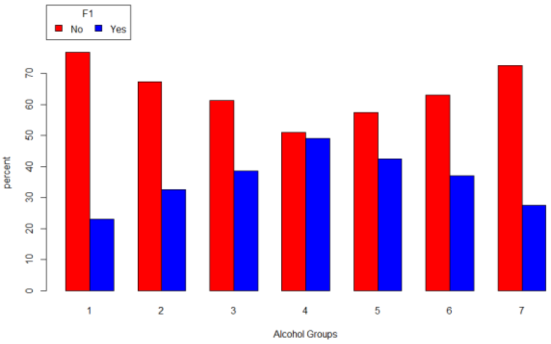


We see a slightly difference in the shape of the histograms, specially in the first category. In the boxplot we can appreciate the difference between the medians. However, it is not enough to see a clear dependence.

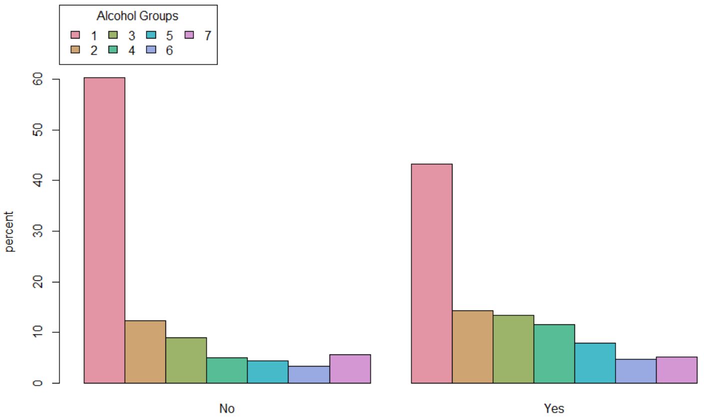
To check it we group the values of the variable “Alcohol” in a qualitative variable “Alcohol Groups” with 7 categories indicating the number of days a year that a person drinks alcohol: 0-50, 51-100, …, 301-365.

We get the following contingency tables that we can represent with a bar graphic.

PERCENTS	**1**	**2**	**3**	**4**	**5**	**6**	**7**
NO	76.9	67.4	61.3	51	57.5	63	72.5
YES	23.1	32.6	38.7	49	42.5	37	27.5
TOTAL	100	100	100	100	100	100	100
COUNT	407	95	75	51	40	27	40



PERCENTAGES	**DOES NOT CONSUME**	**CONSUMES**
1	60.4	43.3
2	12.4	14.3
3	8.9	13.4
4	5.0	11.5
5	4.4	7.8
6	3.3	4.6
7	5.6	5.1
Total	100	100
Count	518	217



In the first graphic there is a clear correlation between the variables. For the first four groups (number of days of alcohol use from 0 to 200) the more days a person drinks alcohol, the more likely they are to consume marihuana. For groups 5,6 and 7 (number of days of alcohol use from 200 to 365) the more days a person drinks alcohol, the less likely they are to consume marihuana.

However, if we pick another sample, we would get another bar graphic and we would see that, in reality, the percentage of marihuana users in groups 5,6 and 7 is higher than the values we get in this research. In reality this dependence is almost linear; the more alcohol a person consumes, the more likely it is to consume marihuana.

Looking at the second graph we deduce that if a person consumes marihuana they drink, in average, more alcohol.

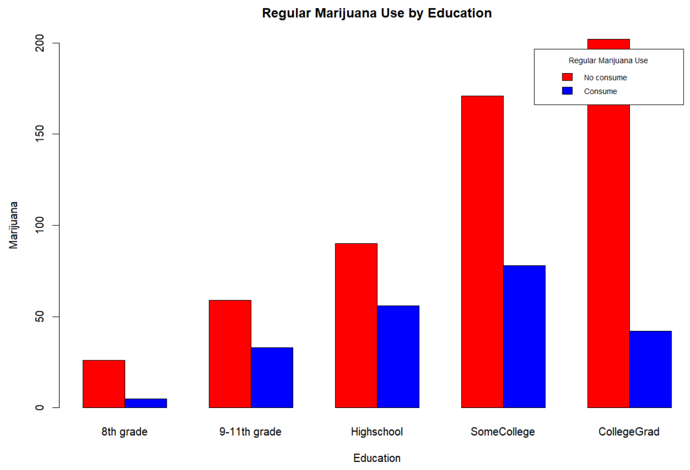
The chi-squared test gives a value of **0.0003562,** which confirms our belief that the two variables are related

3. Perform a descriptive analysis to study the dependence between factors F1 and F2.

Now we will look at how education level may impact the likelihood of marijuana consumption.

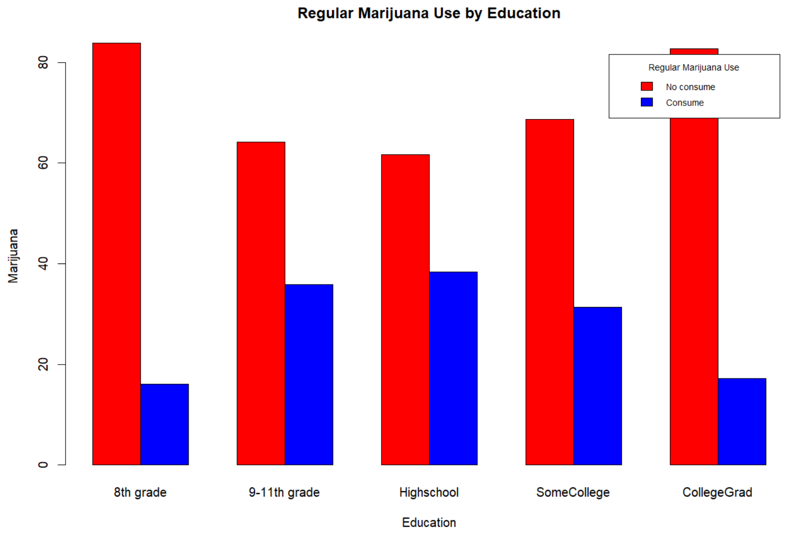
	8th grade	9-11th grade	Highschool	SomeCollege	CollegeGrad
No	26	59	90	171	202
Yes	5	33	56	78	42
TOTAL	31	92	146	249	244

At first glance, we observe that the sample grows according to the level of education, reaching its peak in Some College.



Despite the fact that there are fewer people in some areas, we observe that they more or less maintain a similar relationship. But this will be better observed percentagewise.

	8th grade	9-11th grade	Highschool	SomeCollege	CollegeGrad
No	83.9	64.1	61.6	68.6	82.8
Yes	16.1	35.9	38.4	31.3	17.2
TOTAL	100	100	100	100	100



Upon analyzing the graphs and contingency tables that show the relationship between marijuana consumption and education level, a clear trend can be observed: as education level increases or decreases, marijuana consumption decreases. This suggests a correlation between both variables, reaching its peak at high school.

However, we need to check if the relationship obtained is strong or not, and therefore it is important to analyze the correlation coefficient or **p-value, which is 1.16e-05** (<0.05)and that indicates that they are **dependent.**

Thus, after analyzing the available data, it appears that education level is a significant factor that influences marijuana consumption, among other factors that may play an important role. Nonetheless, it is important to continue studying to gain a deeper understanding of the complex factors that can impact drug use.

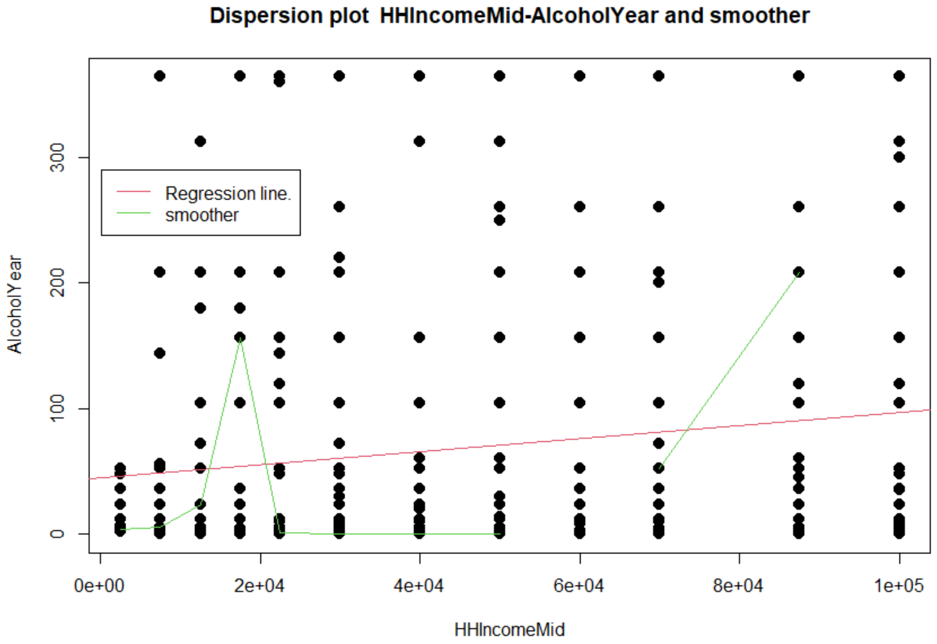
4. Perform a descriptive analysis to study the dependence between the quantitative variables X1 and X2. Analyze this dependence in the whole sample and also as a function of the factor F1.

We want to study if the number of days per year that people drink alcohol beverages depends on the annual income for the household and if this relationship changes according to their marijuana consume. To achieve this goal, we will calculate the correlation coefficient and a dispersion plot with a regression line between the two variables.

First of all, we are going to analyse this dependence in the whole sample. The regression line we obtain is

$$X_2 = (5.157 * 10^{-4})X_1 + 45.42$$

Plotting,

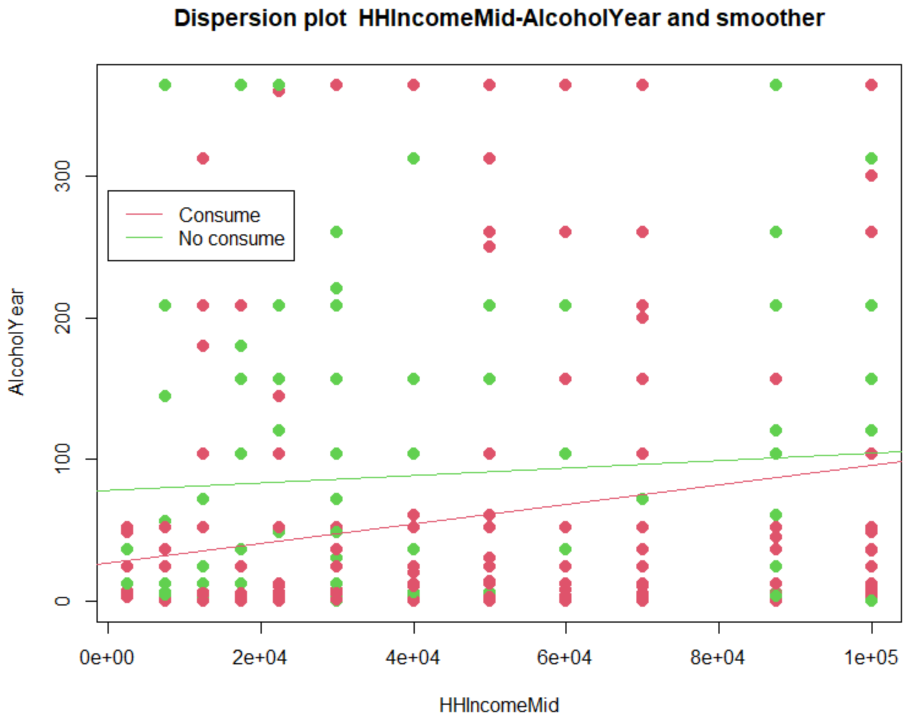


Note that the regression line isn't a good choice to model our sample. (We get a standard error of 97.3 days per year).

The value of the correlation coefficient is 0.17 which is NOT big enough to confirm that the number of days drinking alcohol per year is dependent to the annual income.

Repeating this process for each group of the variable F_1, we get

Regular consume of marijuana	Regression line between X_1 and X_2
No consume	$X_2 = (6.872 * 10^{-4})X_1 + 27.02$
Consume	$X_2 = (2.624 * 10^{-4})X_1 + 77.84$



Regular consume of marijuana	Correlation coefficient between X_1 and X_2
No consume	0.229
Consume	0.09

As we can see, there is a remarkable difference between both groups. X_1 and X_2 are dependent on every group, but this dependence is more accentuated for the ones not consuming marijuana.

APPENDIX

CODE:

```
knitr::opts_chunk$set(echo = TRUE)
library(NHANES)
library(RcmdrMisc)
#group H(j=8)
set.seed(800)
n=800
which(!is.na(NHANES$RegularMarij))
X1.<-NULL
X2.<-NULL
F1.<-NULL
F2.<-NULL
X1<-NULL
X2<-NULL
F1<-NULL
F2<-NULL
for (i in 1:length(which(!is.na(NHANES$RegularMarij))))
{
  X1.[i]<-NHANES$HHIncomeMid[which(!is.na(NHANES$RegularMarij))[i]]
  X2.[i]<-NHANES$AlcoholYear[which(!is.na(NHANES$RegularMarij))[i]]
  F1.[i]<-NHANES$RegularMarij[which(!is.na(NHANES$RegularMarij))[i]]
  F2.[i]<-NHANES$Education[which(!is.na(NHANES$RegularMarij))[i]]
}
random<-sample(c(1:length(which(!is.na(NHANES$RegularMarij)))), size=800)
for (i in 1:800)
{
  X1[i]<-X1.[random[i]]
  X2[i]<-X2.[random[i]]
  F1[i]<-F1.[random[i]]
  F2[i]<-F2.[random[i]]
}
#Con el warning no hay problema si se ejecuta desde el principio. Comprobado y se pq.
F1<-cut(F1,breaks=2,labels=c("No", "Yes"))
F2<-cut(F2,breaks=5,labels=c("8th Grade", "9-11th Grade", "High School", "Some College","College Grad"))


#EXERCISE 1
#descriptive analysis of X1
summary(X1)
sigma<-sd(X1,na.rm=TRUE)
mu<-mean(X1,na.rm = TRUE)
mu
sigma
qqnorm(X1)
qqline(X1, col='red')
hist(X1, probability=TRUE, breaks=10, xlim=c(2000,100000))
dxE<-density(X1,na.rm=TRUE)
lines(dxE, col="black")
xn<-qnorm(seq(0,1,length.out=100),mu,sigma)
lines(xn,dnorm(xn,mu,sigma), col='red')
xg<-qgamma(seq(0,1,length.out=100),shape=(mu^2)/(sigma^2),scale=(sigma^2)/mu)
lines(xg,dgamma(xg,shape=(mu^2)/(sigma^2),scale=(sigma^2)/mu), col='purple')
#descriptive analysis of X2
summary(X2)
sigma2<-sd(X2,na.rm=TRUE)
mu2<-mean(X2,na.rm=TRUE)
```

```
mu2
qqnorm(X2)
qqline(X2, col='red')
hist(X2, probability=TRUE, breaks=11, xlim=c(0,400))
dxE2<-density(X2,na.rm=TRUE)
lines(dxE2, col="black")
xn2<-qnorm(seq(0,1,length.out=100),mu2,sigma2)
lines(xn2,dnorm(xn2,mu2,sigma2), col='red')
xg2<qgamma(seq(0,1,length.out=100),shape=(mu2^2)/(sigma2^2),scale=(sigma2^2)/mu2)
lines(xg2,dgamma(xg2,shape=(mu2^2)/(sigma2^2),scale=(sigma2^2)/mu2), col='purple')
#descriptive analysis F1
table(as.factor(F1))#valores absolutos
table(as.factor(F1))/length(F1)*100#porcentajes
par(mfrow = c(1, 2))
barplot(table(F1), col = c("green", "blue"))
pie(table(F1), col = c("green", "blue"))
#descriptive analysis F2
table(as.factor(F2))#valores absolutos
table(as.factor(F2))/length(F1)*100#porcentajes
par(mfrow = c(1, 1))
barplot(table(F2), col = c("green", "blue","purple","orange","yellow"))
pie(table(F2),col = c("green", "blue","purple","orange","yellow"))
#PREGUNTA 2
numSummary(X2, group=F1)
par(mfrow=c(1,1))
Hist(X2, groups=F1, breaks= c(-1,50, 100,150,200,250, 300, 365), probability=FALSE)Boxplot(X2~F1)
```

```
groupsX2<- cut(X2, breaks = c(-1,50, 100,150,200,250, 300, 365), labels = c(1, 2, 3, 4, 5,6,7))
groupsX2
X2
summary(groupsX2)
barplot(table(groupsX2))
Tabla<-xtabs(~F1 + groupsX2)
Tabla
T1<-totPercents((Tabla))
T1
T2<-colPercents((Tabla))
T2
T3<-colPercents(t(Tabla))
T3
Barplot(groupsX2, by=F1, style="parallel", col = c("red", "blue"),scale="percent")
#Si se cambia la semilla se ve mejor la correlacion.
Barplot(F1, by=groupsX2, style="parallel", scale="percent", conditional="TRUE")

resultado <- chisq.test(table(groupsX2, F1))
resultado

#PREGUNTA 3

F11 <- factor(F1, labels = c("Consume","No consume"))
F22 <- factor(F2, labels=c("8th grade","9-11th grade", "Highschool", "SomeCollege", "CollegeGrad"))

#Table+Plot F1 and F2
tabla <- xtabs(~F11 + F22, data = NHANES)
print(tabla)
```

```
barplot(tabla, beside = TRUE, legend.text = rownames(tabla), args.legend = list(title = "Regular Marijuana Use", cex = 0.7),
        xlab = "Education", ylab = "Marijuana", col = c("red", "blue"), main = "Regular Marijuana Use by Education")

#Percents
tabla_prop <- prop.table(tabla) * 100
tabla_prop_round <- round(tabla_prop, digits = 1)
tabla_prop_total_round <- addmargins(tabla_prop_round)
print(tabla_prop_total_round)

colpercents <- prop.table(tabla, margin = 2) * 100
colpercents_round <- round(colpercents, digits = 1)
barplot(colpercents, beside = TRUE, legend.text = rownames(tabla), args.legend = list(title = "Regular Marijuana Use", cex = 0.7), xlab = "Education", ylab = "Marijuana", col = c("red", "blue"), main = "Regular Marijuana Use by Education")
resultado <- chisq.test(table(F1, F2))
resultado

#Exercise 4
#Dependence Between X1 and X2 in the whole sample
#Numerical
lm1<-lm(X2 ~ X1)
summary(lm1)
df <- data.frame(X1, X2)
# Eliminar las filas que contienen valores faltantes en ambos vectores
df_complete <- df[complete.cases(df), ]
# Calcular la matriz de correlación entre x e y
correlacion <- cor(df_complete$X1, df_complete$X2)
# Imprimir la matriz de correlación
print(correlacion)
```

```
#Plots
par(mfrow = c(1, 1))
plot(X1, X2, xlab="HHIncomeMid", ylab="AlcoholYear",
    main = "Dispersion plot  HHIncomeMid-AlcoholYear and smoother",  pch=16, cex=1.4)
abline(a=lm1$coefficients[1], b=lm1$coefficients[2], col="2")
lines(lowess(X1, X2, f = 0.2, delta=10), col = "3")
legend(62, 290, inset = 0.1, c("Regression line.","smoother"), lty = 1, col = 2:3)

#Dependence Between X1 and X2 as function of F1
#Numerical
library(dplyr)
df <- data.frame(F11, X1, X2)
# Eliminar las filas que contienen valores faltantes en ambos vectores
df_complete <- df[complete.cases(df), ]

lmG<-df %>% group_by(F11) %>% do(model = lm(X2 ~ X1, data = .))
for (i in 1:2)
{
  print(F11[i])
  print(summary(lmG[[2]][[i]]))
  i <- i+1
}
CorrG<-df_complete %>% group_by(F11) %>% summarize(cor=cor(X1,X2))
print(CorrG)

#Plots
colorF11 = c(2,3)[F11]
plot(X1, X2, xlab="HHIncomeMid", ylab="AlcoholYear",
     main = "Dispersion plot  HHIncomeMid-AlcoholYear and smoother",  pch=16,
cex=1.4, col=colorF11)

for (i in 1:2)

{

  abline(a=lmG[[2]][[i]]$coefficients[1], b=lmG[[2]][[i]]$coefficients[2], col=i+1)

  i <- i+1

}

legend(62, 290, inset = 0.1, c("Consume","No consume"), lty = 1, col =2:3)
```