Primer reto

Oliver Rodriguez

8/3/2021

Estas son algunas librerias que se utilizaran:

```
# para trabajar on regex:
library(stringr)
library(tidyverse)
library(kableExtra)
```

Aqui se realizo todo el procedimiento de organizar los datos:

```
setwd('C:/Users/Oliver/Documents/9/TAE')
txt <- readLines('hilo tae2.txt', encoding = "UTF-8")

#así soluciono  el problema de los que escribieron separado por \n
ali <- paste(cbind(txt[seq(5,10)])  , collapse = " | ")
nico <- paste(cbind(txt[seq(64,69)])  , collapse = " | ")

txt_final <- c(txt[-c(seq(4,11), seq(64,70))],ali, "",nico)

#lo que me interesa tiene tamaÃ±o mayor aprox a 50
vec <- c()
for (i in txt_final) {
  if (nchar(i) >50 ) {
    vec <- c(vec, i)
  }
}

#Para normalizar primero paso a minusculas
vec <- str_to_lower(vec)

# Les quito las tildes
library(stringi)
vec <- stri_trans_general(vec,'Latin-ASCII')

# Hay unos errores usando como separaciÃ³n 1 y I en vez de |
length(str_which(vec, ' \\| ')) # vec  y la condiciÃ³n no son del mismo tamaÃ±o entonces
                                # voy a encontrar los raritos
arreglos <- vec[str_which(vec, ' \\| ', T)] # los encontrÃ©

# los siguientes son los arreglos correspondientes
arreglos[1] <- str_replace_all(arreglos[1],',',' |')
arreglos[2] <- str_replace_all(arreglos[2], ' l ', ' | ')
```

```r
arreglos[2] <- str_replace_all(arreglos[2], '[:alpha:]{1}[:digit:]{2}[:alpha:]{1}', '| 21 |')
arreglos[3] <- str_replace_all(arreglos[3], ',', ' | ')
arreglos[3] <- str_replace_all(arreglos[3], 'hola \\|  soy ', '') # me doy cuenta de que ya esta

# aqui concateno los buenos con los arreglos
vec3 <- c(vec[str_which(vec, ' \\| ')], arreglos[c(1,2)])

# elimino puntaucion y simbolos
vec3 <- str_replace_all(vec3,'[:punct:]','')
# aqui los separo por |
vec4 <- str_split(vec3, '\\|') # me resulto mas facil separarlo sin tener en cuenta los espacios

# decubro que el 61 tiene problemas de "" y debo borrar para que
# se extraigan en el mism orden que lo necesit y hago lo siguiente:
vec4 <- c(vec4[-61],list(vec4[[61]][c(2,3,4,5,6,7)]))

# aqui selecciono los que tienen 6 elementos que es lo que nos piden
# y saco los que no cumplen para revisar
vec_raros <- c()
vec5 <- c()
for (i in vec4) {
  if (length(i) == 6 ){
    vec5 <- c(vec5,list(i))
  }else{vec_raros <- c(vec_raros,list(i))}
}


# decido excluirlos por el momennto para luego de tener los demas datos proceder como indico
falto_trabajo <- vec_raros[c(1,2,5,8,11)] # se le asignara un 'no'
falto_semestre <- vec_raros[c(3,4,7,10)] #  se le asignra promedio
falto_mucho <- vec_raros[c(6)]  # semestre promedio no labora y tendra pasatiempo leer
separar_ultima_entrada <- vec_raros[c(9)]



# creo los vectores y voy borrando espacios en blanco incio y final
nombres <- trimws(sapply(vec5, '[', 1))
edad <- trimws(sapply(vec5, '[', 2))
programa0 <- trimws(sapply(vec5, '[', 3))
semestre0 <- trimws(sapply(vec5, '[', 4))
trabaja <- trimws(sapply(vec5, '[', 5))
pasatiempo <- trimws(sapply(vec5, '[', 6))

# VOy a añadir los raritos con lo dicho:
ft <- trimws(sapply(falto_trabajo, '[', c(1:5)))
fs <- trimws(sapply(falto_semestre, '[', c(1:5)))
fm <- trimws(sapply(falto_mucho, '[', c(1:3)))
sue <- trimws(sapply(separar_ultima_entrada, '[', c(1:5)))



nombres <- c(nombres,ft[1,],fs[1,],fm[1,],sue[1,])
edad <- c(edad,ft[2,],fs[2,],fm[2,],sue[2,])
```

```r
programa0 <- c(programa0,ft[3,],fs[3,],fm[3,],sue[3,])
# semestre luego lo anadimos cuando podamos sacar el promedio
trabaja <- c(trabaja,rep('no',5),fs[4,],'no','desarrollo series')
pasatiempo <- c(pasatiempo,ft[5,],fs[5,],'leer','deporte')

#para la edad quito las letras y espacios en blanco y lo convierto a numerico:
edad <- as.numeric(str_replace_all(edad,'[:alpha:]|[:space:]',''))
# normalizo más los de sistemas llevandolos solo 'sitemas':
programa <- c()
for (i in programa0) {
  if (str_detect(i,'sistemas') == TRUE){
    programa <- c(programa,'sistemas')
  }else{
    programa <- c(programa,i)
  }

}
# quito las palabras ingenieria y 'de '
programa <- str_replace_all(programa, 'ingenieria |ingenieria de |ing ','')
programa <- str_replace_all(programa, 'de ','')

#Semestre es un desastre. vamos despacio, primero elimimo la palabra semestre.
semestre <- c()
for (i in semestre0) {
  if (str_detect(i,'semestre') == TRUE){
    semestre <- c(semestre,str_replace_all(i,'semestre',''))
  }else{    semestre <- c(semestre,i)
  }
}

#Luego identifico los patrones indeceados y los modifico correctamente
semestre <- str_replace_all(semestre,'octavo','8')
semestre <- str_replace_all(semestre,'xviii','16')
semestre <- str_replace_all(semestre,'septimo','7')
semestre <- str_replace_all(semestre,'ultimo','10')
semestre <- str_replace_all(semestre,'ix','9')
semestre <- str_replace_all(semestre,'3er ','3')
semestre[46] <- '12'
semestre <- as.numeric(semestre) #buala

#añado los que faltaban del semestre:
round(mean(semestre),0) #la media es 8 para los que no pusieron el semestre
semestre <- as.numeric(c(semestre,ft[4,],rep(8,4),8,sue[4,]))
# para trabajo reemplazo algunas cadenas que representan que no trabaja:
trabaja <- str_replace_all(trabaja,'no [:alpha:]*.|na|ninguna|desempleado','no')
trabaja[48] <- 'no'# no me quiso reemplazar, por eso lo hago mas manual
# un 'no' para los campos vacios
trabaja1 <- c()
for (i in trabaja) {
  if (nchar(i) == 0 ) {
    trabaja1 <- c(trabaja1, 'no')
  }else{trabaja1 <- c(trabaja1, i)}
}
```

```r
#Creamos el DF:
df <- data.frame(nombres, edad, programa, semestre, trabaja1, pasatiempo )
ifelse(test = df$trabaja1 == 'no', 'no','si')

# Para Buscar el sexo extraemos los primeros nombres:
name_1 <- sapply(str_split(df[,1], ' '), '[', 1)

# trabajo manual de separar hombre y mujer si repetir
n_h <- c("juan","santiago","sebastian","jean","daniel","julian","carlos","alejandro","arley","david","is
  "mateo","jhonier",  "eider" ,"esteban", "miguel","diego" ,"james","hans","jelssin","edwar","nicolas",
  "esteban","oliver","edhy" )
n_m <- c("carolina", "laura","cristina", "jennifer","stephany","catherine","zuleima","estefania","julia
  ,"sara"  ,    "yuliza","salome", "ana","vanessa", "isabela")

# este loop arrojara los generos
sexo <- c()
for (i in name_1){
  if (i %in% n_h == T){
    sexo <- c(sexo,'H')
  }
  if (i %in% n_m == T) {
    sexo <- c(sexo,'M')
  }
}

# anado la columma al DF:
df['sexo'] <- sexo
# un pequeño errorcito
df[6,6] <- 'leer libros'
df['trabaja'] <- ifelse(df$trabaja1 == 'no', 'no', 'si')

write.csv(df,'C:/Users/Oliver/Documents/9/TAE/presentacion_TAE.csv',sep = ',',row.names = F)
```

Finalmente la base se ve asi:

```r
head(df,10) %>%
  kbl()
```

nombres	edad	programa	semestre	trabaja1	pasatiempo
juan esteban cendales sora	21	sistemas	8	trascender global	piano y prog
santiago ramirez zapata	22	control	9	no	bailar y hac
juan jose hurtado alvarez	26	sistemas	7	no	jugar futbol
juan pablo ortega medina	21	sistemas	8	ddb latino puerto rico inc trascender global	viajar
sebastian rendon giraldo	20	sistemas	8	freelancer	jugar videoj
jean paul yepes	22	sistemas	10	no	leer libros
daniel alexander naranjo rios	21	sistemas	9	no	videojuegos
julian alejandro usuga ortiz	21	estadistica	4	no	deporte y m
sebastian lopez restrepo	24	sistemas	9	trascender global	jugar videoj
carlos daniel montoya	22	sistemas	10	experimentality	explorar tec

Completitud

En la siguiente tabla podemos observar el resumen de las varibles numericas, que son el semestre y la edad, ademas se le anade las desviaciones estandares.

```
# se quiere presentar estadistico de resumen
resum <- summary(df[,c(2,4)])
# para mostrar la desviacion estandar se hace los siguiente
desv1 <- as.character(round(sd(df[,2]),3)) # se calcula y redondea y luego se conviente a caracter
desv2 <- as.character(round(sd(df[,4]),3))
# luego lo añadimos a los resumenes y finalmente se imprime
resum <- rbind(resum,c(paste('sd     :',desv1,sep = ''),paste('sd     :',desv2,sep = '')))
resum%>%
  kbl()
```

	edad	semestre
	Min. :18.00	Min. : 3.000
	1st Qu.:21.00	1st Qu.: 8.000
	Median :21.00	Median : 8.000
	Mean :22.16	Mean : 7.971
	3rd Qu.:23.00	3rd Qu.: 9.000
	Max. :33.00	Max. :16.000
	sd :2.305	sd :2.086

Aqui tenemos las medidas de escala. Podemos ver que el RIQ es pequeño para ambos casos y esto da una idea de que en este porcentaje, la mayoria de la edad de los estudiantes difiere en 2 años y que la mayoria de los estudiantes difieren en 1 semestre:

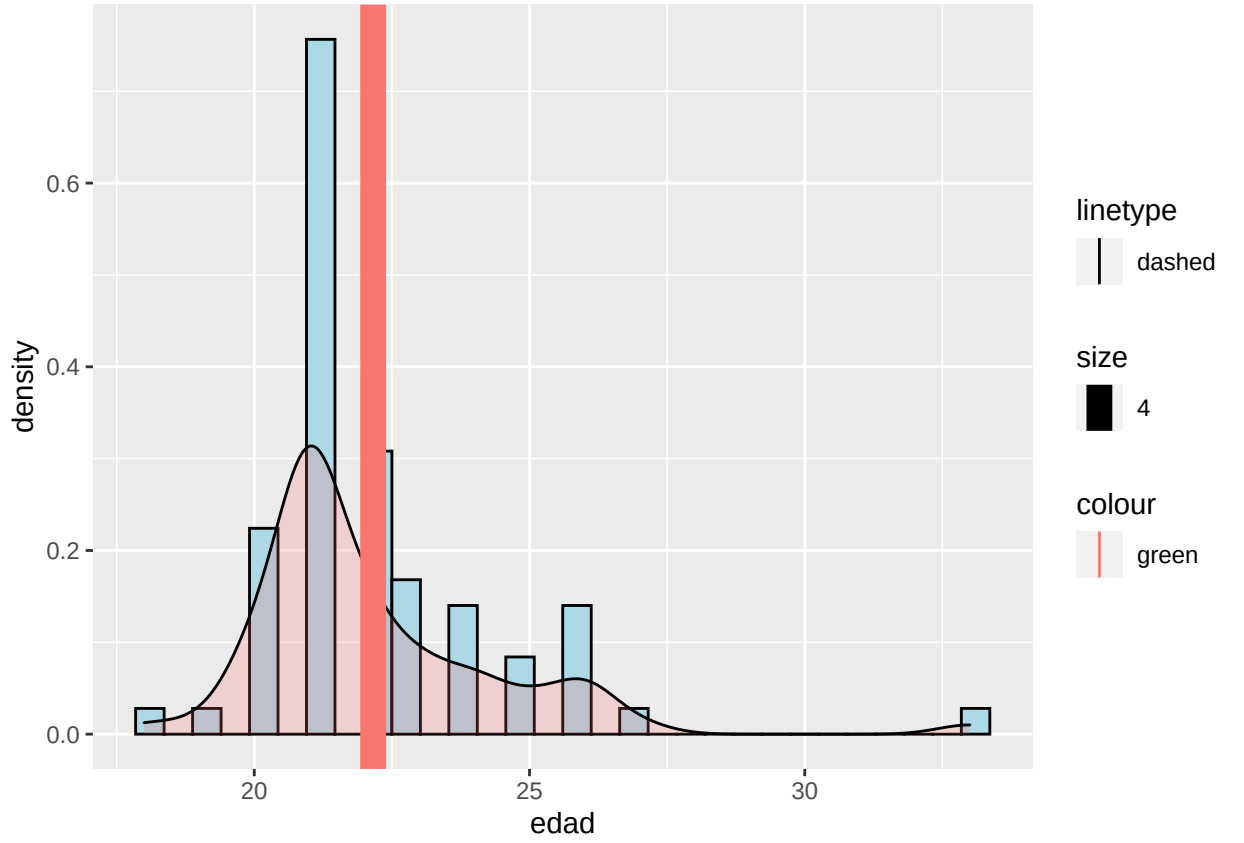
```
# se crean dos funciones para rango intecuatil y rango:
intercuart <- function(x) quantile(x)[4]-quantile(x)[2]
rango <- function(x) max(x)-min(x)
# DF de las escalas
escalas <- data.frame(rango=c(rango(df[,2]),rango(df[,4])), 'rango intercuartil' = c(intercuart(df[,2])

kbl(escalas)
```

	rango	rango.intercuartil
edad	15	2
semestre	13	1

Algunos histogramas para la variables numéricas:

En el histograma de la edad podemos ver sus frecuencias relativas y su grafica de desnsidad, vemos que la edad más comun es 21 años, tambien se observa que el promedio es aproximadamente 22, tambien tienen algunos datos atípicos como 18 y mayor 30. Tambien se muestra la tabla de sus frecuencias absolutas:

```
library(ggplot2)

ggplot(df, aes(x=edad)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightblue", bins=30)+
  geom_density(alpha=.2, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(edad), color="green", linetype="dashed", size=4))
```

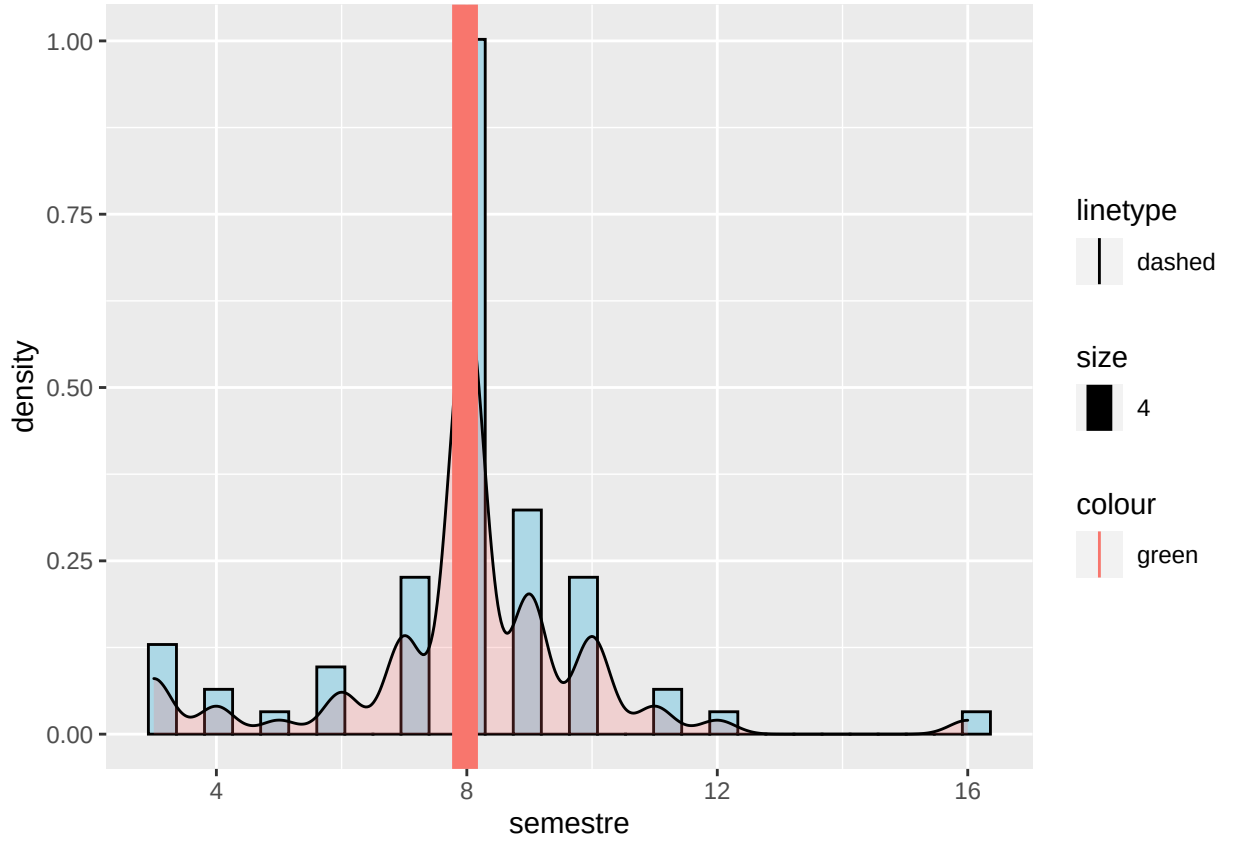



```
kbl(table(df[,2]))
```

Var1	Freq
18	1
19	1
20	8
21	27
22	11
23	6
24	5
25	3
26	5
27	1
33	1

El histograma de los semestres en frecuencia relativa con su grafica de densidad nos muestra que el semetre más frecuente es el 8, y que la media es muy cercana a el octavo semestre. Tambien se muestran las frecuencias absolutas en la tabla:

```
ggplot(df, aes(x=semestre)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightblue", bins=30)+
  geom_density(alpha=.2, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(semestre), color="green", linetype="dashed", size=4))
```




```
kbl(table(df[,2]))
```

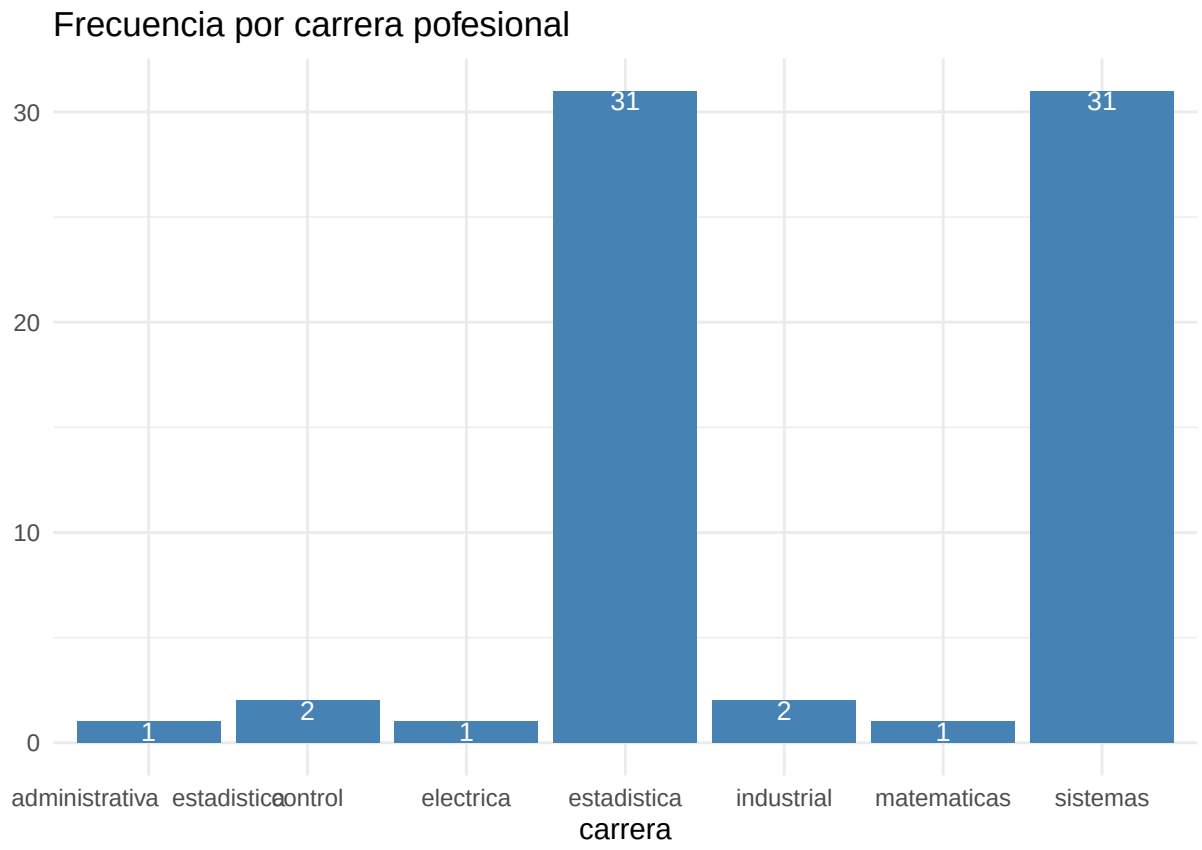
Var1	Freq
18	1
19	1
20	8
21	27
22	11
23	6
24	5
25	3
26	5
27	1
33	1

Variables categoricas:

primero obsevamos la cantidad de estudiantes por carrera, en sus frecuencias absulas y una tabla con sus frecuencias relativas, donde ingenieria de sistemas y estadistica son las mas frecuentes, siendo estadistica por un estudiante de doble titulacion la carrera mas frecuente en este curso:

```
ggplot(data= data.frame(table(df$programa)), aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue")+
```

```
  geom_text(aes(label=Freq), vjust=1, color="white", size=3.5)+
  theme_minimal()+
  labs(title="Frecuencia por carrera pofesional ",x="carrera", y = "")
```

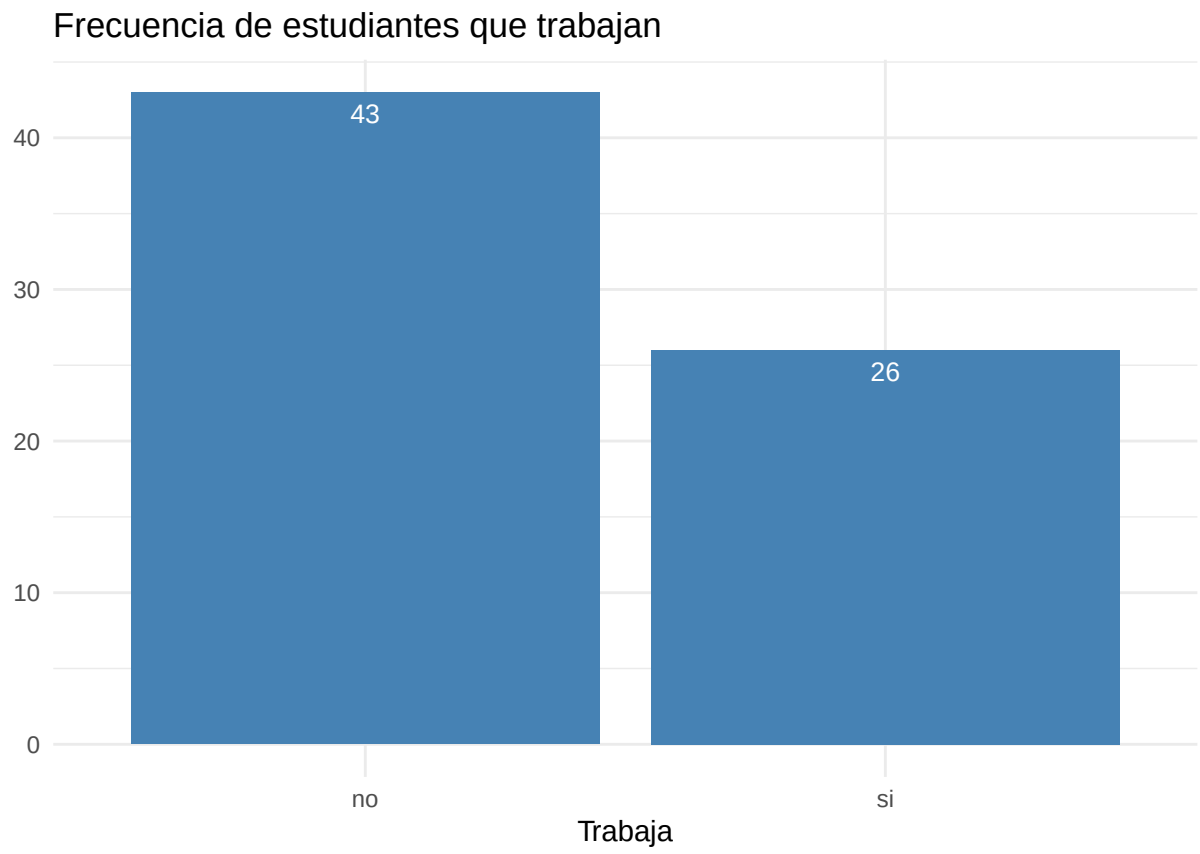



```
kbl(table(df$programa)/length(df$programa)*100)
```

Var1	Freq
administrativa estadistica	1.449275
control	2.898551
electrica	1.449275
estadistica	44.927536
industrial	2.898551
matematicas	1.449275
sistemas	44.927536

Veamos la cantidad de estudiantes que trabajan y los que no en sus frecuencias absulas y una tabla con sus frecuencias relativas, donde observamos que aproximadamente 62% no trabajan:

```
ggplot(data= data.frame(table(ifelse(df$trabaja1 == 'no', 'no','si'))), aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=Freq), vjust=1.6, color="white", size=3.5)+
  theme_minimal()+
  labs(title="Frecuencia de estudiantes que trabajan ",x="Trabaja", y = "")
```

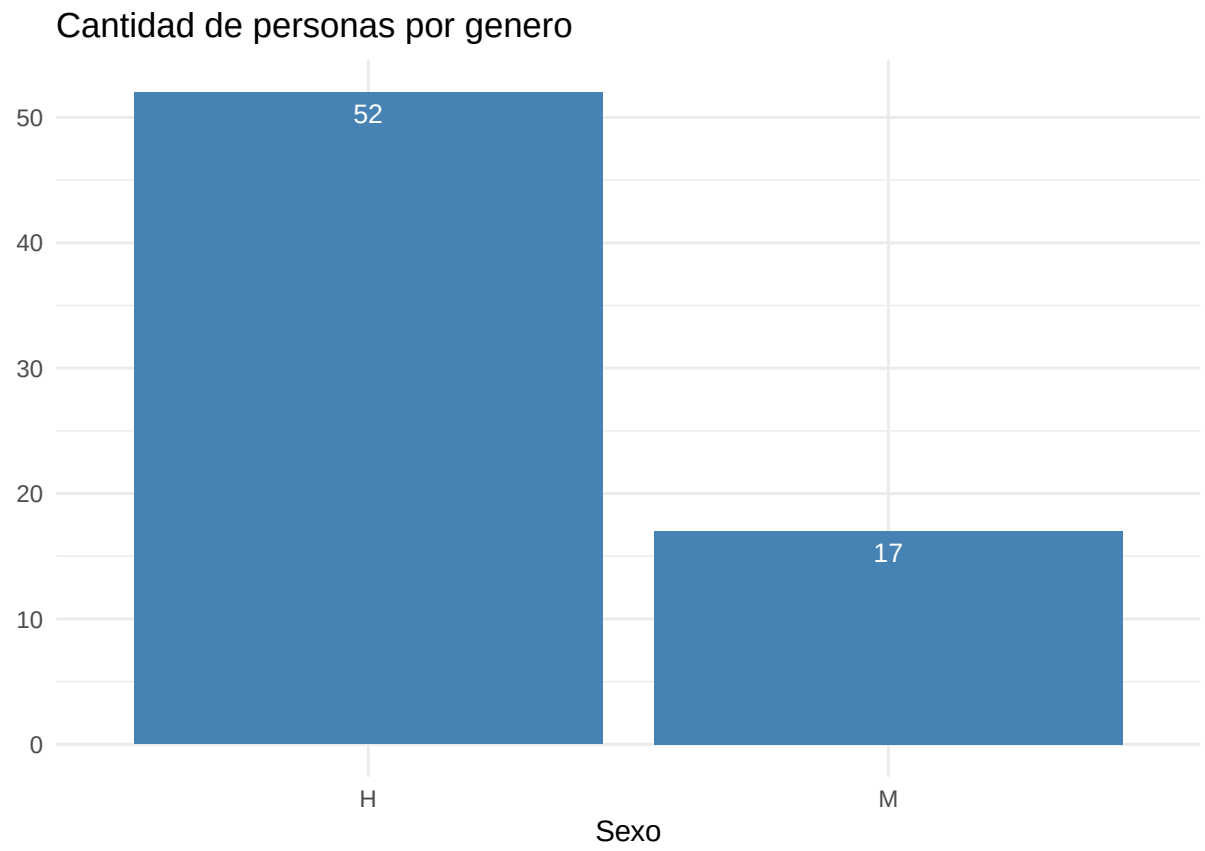



```
kbl(table(ifelse(df$trabaja1 == 'no', 'no','si'))/length(df$trabaja1)*100)
```

Var1	Freq
no	62.31884
si	37.68116

Ahora observamos la cantidad de estudiantes discriminados por sexo en sus frecuencias aobsulas y una tabla con sus frecuencias relativas, aproximadamente el 75% de los estudiantes son hombres:

```
ggplot(data= data.frame(table(df$sexo)), aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=Freq), vjust=1.6, color="white", size=3.5)+
  theme_minimal()+
  labs(title="Cantidad de personas por genero ",x="Sexo", y = "")
```




```
kbl(table(df$sexo)/length(df$sexo)*100)
```

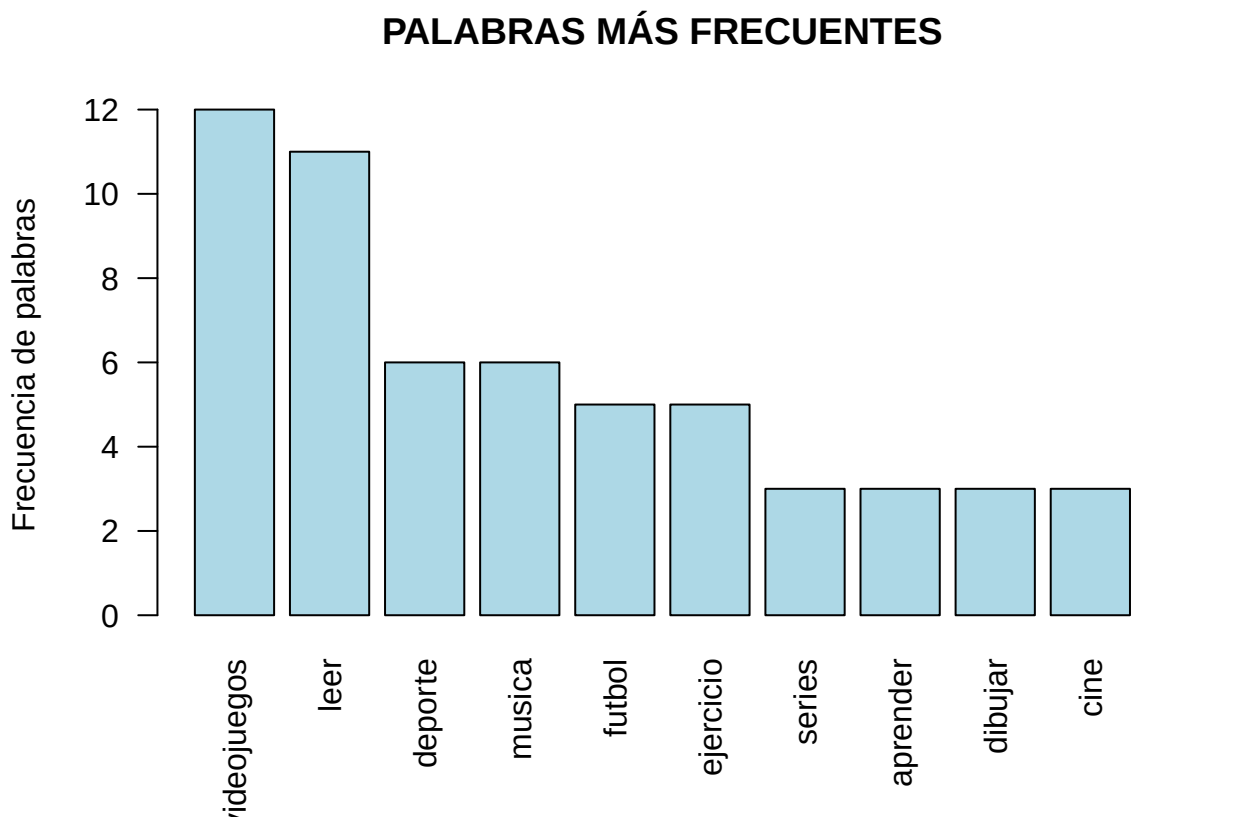
Var1	Freq
H	75.36232
M	24.63768

```
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("wordcloud2")
```

```
corpus <- Corpus(VectorSource(df$pasatiempo))
d  <- tm_map(corpus, stripWhitespace)
d <- tm_map(d, removePunctuation)
d <- tm_map(d, removeNumbers)
d <- tm_map(d, removeWords, stopwords("spanish"))
d <- tm_map(d, removeWords, c('hacer', 'ver', 'jugar'))
# crea matriz de términos
tdm <- TermDocumentMatrix(d)
frecuentes<-findFreqTerms(tdm, lowfreq=1)
#Sumarización
m <- as.matrix(tdm) #lo vuelve una matriz
v <- sort(rowSums(m),decreasing=TRUE) #lo ordena y suma
dff <- data.frame(word = names(v),freq=v) # lo nombra y le da formato de data.frame
```

En esta grafica se puede obserar los pasatiempos más comunes:

```
### TRAZAR FRECUENCIA DE PALABRAS
barplot(dff[1:10,]$freq, las = 2, names.arg = dff[1:10,]$word,
col ="lightblue", main ="PALABRAS MÁS FRECUENTES", ylab = "Frecuencia de palabras")
```



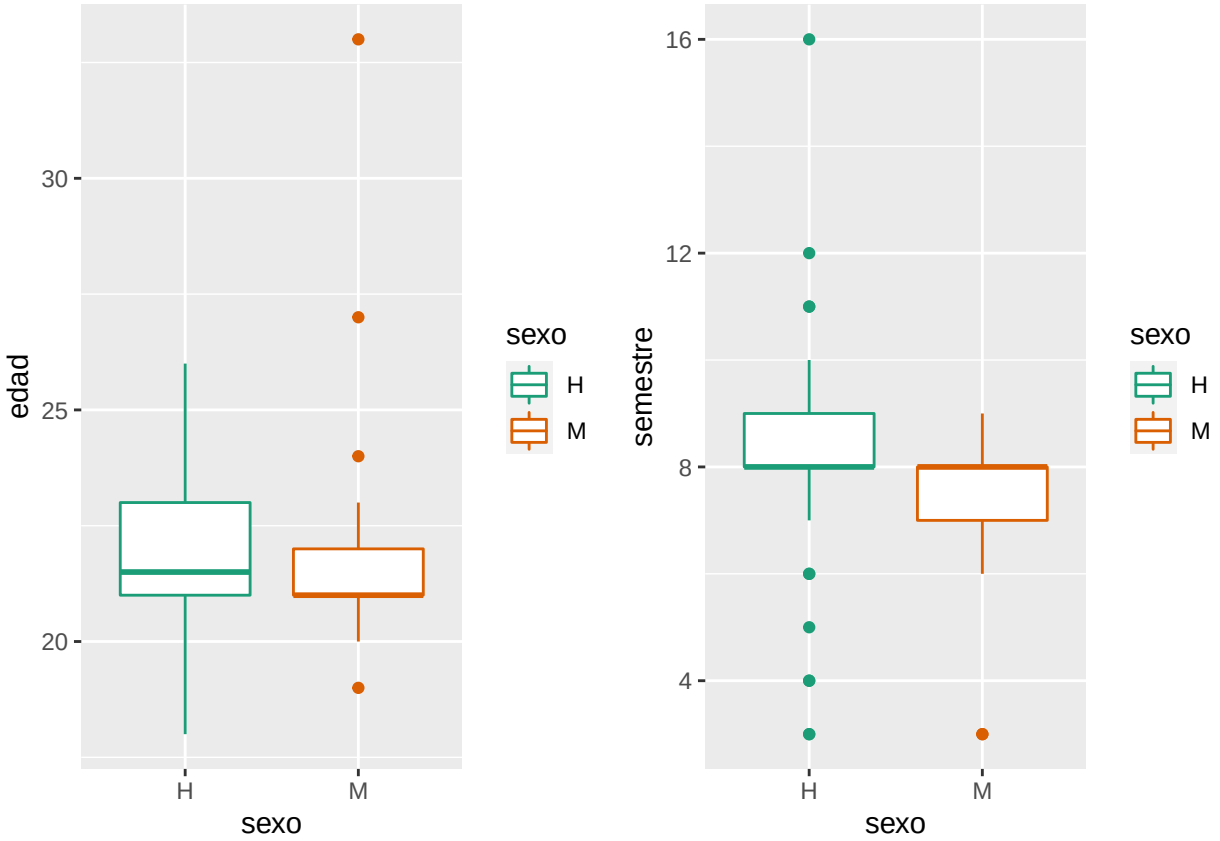
Este grafico es interactivo y le permite observar la cantidad de veces que aparecen las palabras, siento muy comun videojuegos sobre todo, luego el leer, deporte etc.

```
#wordcloud2(dff, size=0.6)
```

Relaciones multivariadas:

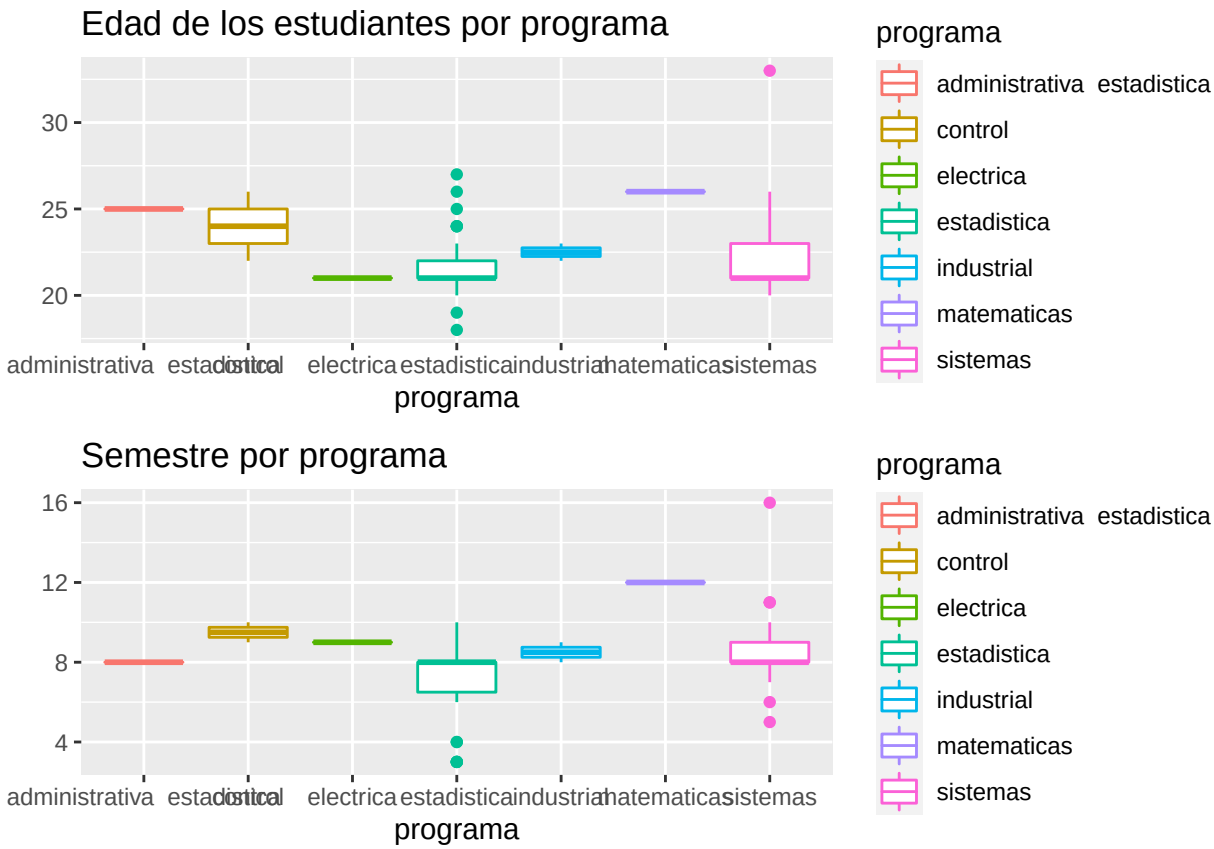
Se observa en la siguentes graficas que la edad de las mujeres es mas variable que la de los hombres, mientras que los hombres muestran mayor variabilidad en el semestre cursado.

```
library(gridExtra)

g1 <- ggplot(df, aes(x=sexo, y=edad, color=sexo)) +
  geom_boxplot()+scale_color_brewer(palette="Dark2")
g2 <- ggplot(df, aes(x=sexo, y=semestre, color=sexo)) +
  geom_boxplot()+scale_color_brewer(palette="Dark2")
grid.arrange(g1,g2,ncol= 2)
```



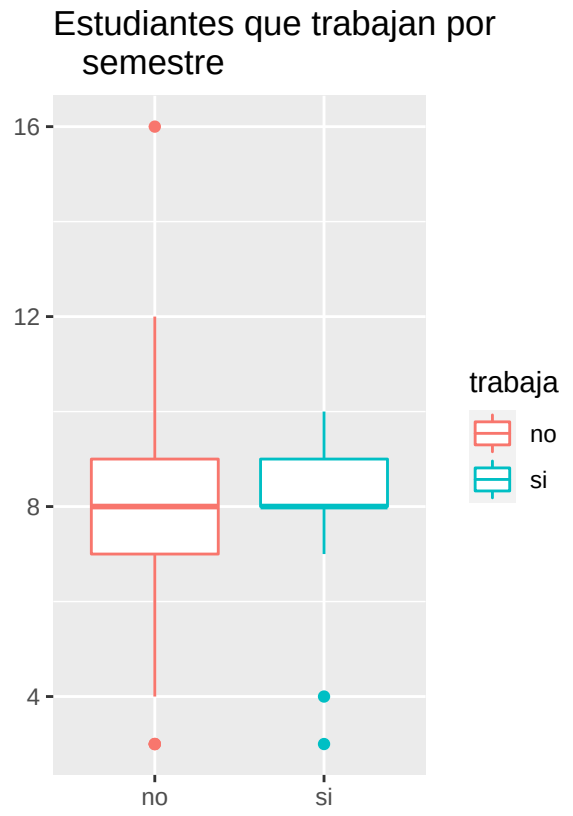
Se tienen pocos datos de alagunos cursos pero podemos observar que estadsticamentes se observan diferencias significativas entre ingenieria de control y estadistica en cuanto a la edad, podemos ver que la persona de mayor edad es de ingenieria de sistemas y que los estudiantes de estadistica tienen mayor variabilidad y son tambien los más jovenes y son los de menor semestre.

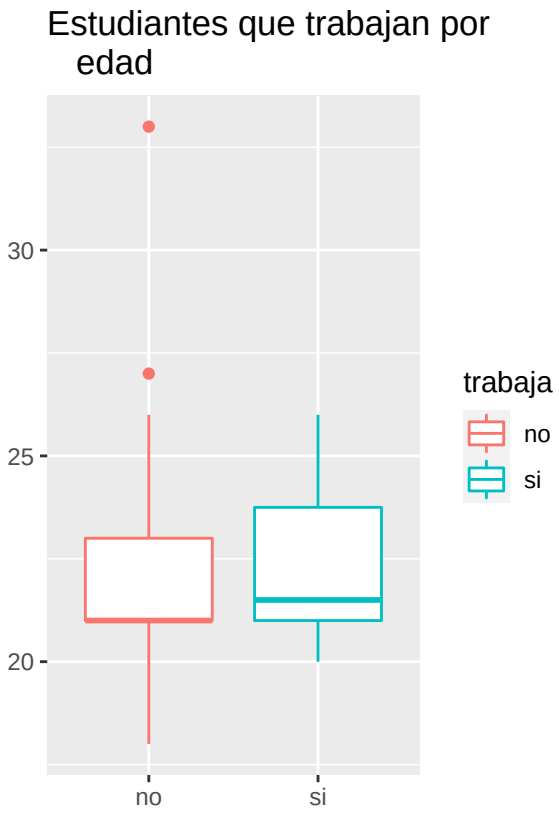
```
g3 <- ggplot(df, aes(x=programa, y=edad, color=programa)) +
  geom_boxplot()+
  labs(title="Edad de los estudiantes por programa ", y = "")
g4 <- ggplot(df, aes(x=programa, y=semestre, color=programa)) +
  geom_boxplot()+
  labs(title="Semestre por programa" , y = "")
grid.arrange(g3,g4,nrow=2)
```



En las siguientes graficas se observa que la pesonas que no trabajan tienen mas dispercion de en cuanto a sus semestres y tambien son mas dispersos en cuanto a su edad.

```
g5 <- ggplot(df, aes(x=trabaja, y=semestre, color=trabaja)) +
  geom_boxplot()+
  labs(title="Estudiantes que trabajan por \n \t semestre ", x= '', y = "")

g6 <- ggplot(df, aes(x=trabaja, y=edad, color=trabaja)) +
  geom_boxplot()+
  labs(title="Estudiantes que trabajan por \n \t edad ", x= '', y = "")

grid.arrange(g5,g6, ncol=2)
```





Para observar un grado de relación lineal entre las variables edad y semestre se calcula su correlacion y se obtien una realcion aprox. del 31%. Tambien se plantea un modelo que involucre estas variables y se aprecia que el semestre es una variable que es significativa para explicar la edad y que el semetre puede explicar aproximadamente el 10% de la variabilidad de las edades, aunque es poco, podria postularse posteriormente un mejor modelo.

```
kbl(cor(df$edad,df$semestre))
```

x
0.3190492

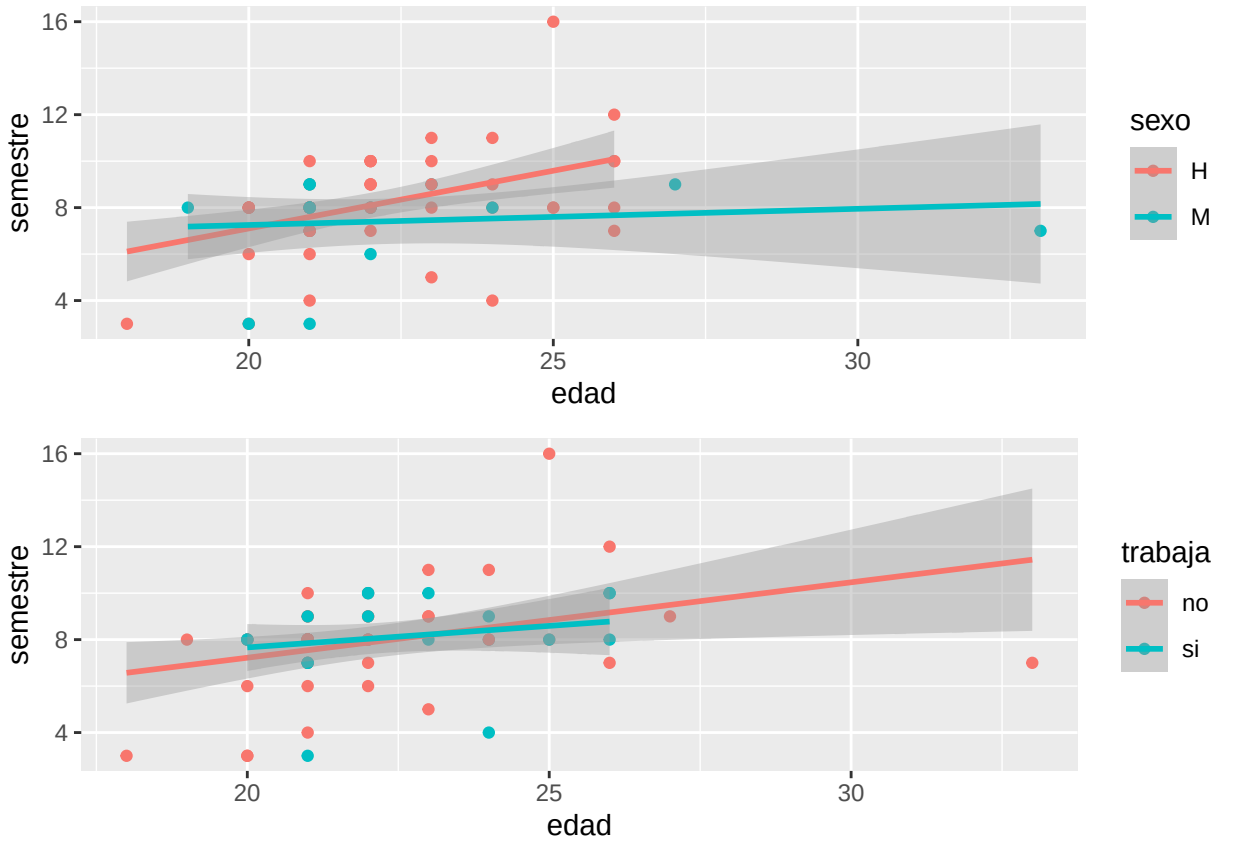
```
mod <- lm(df$edad ~ df$semestre,)

summary(mod)
```

```
##
## Call:
## lm(formula = df$edad ~ df$semestre)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1696 -1.1696 -0.8171  0.5354 11.1829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

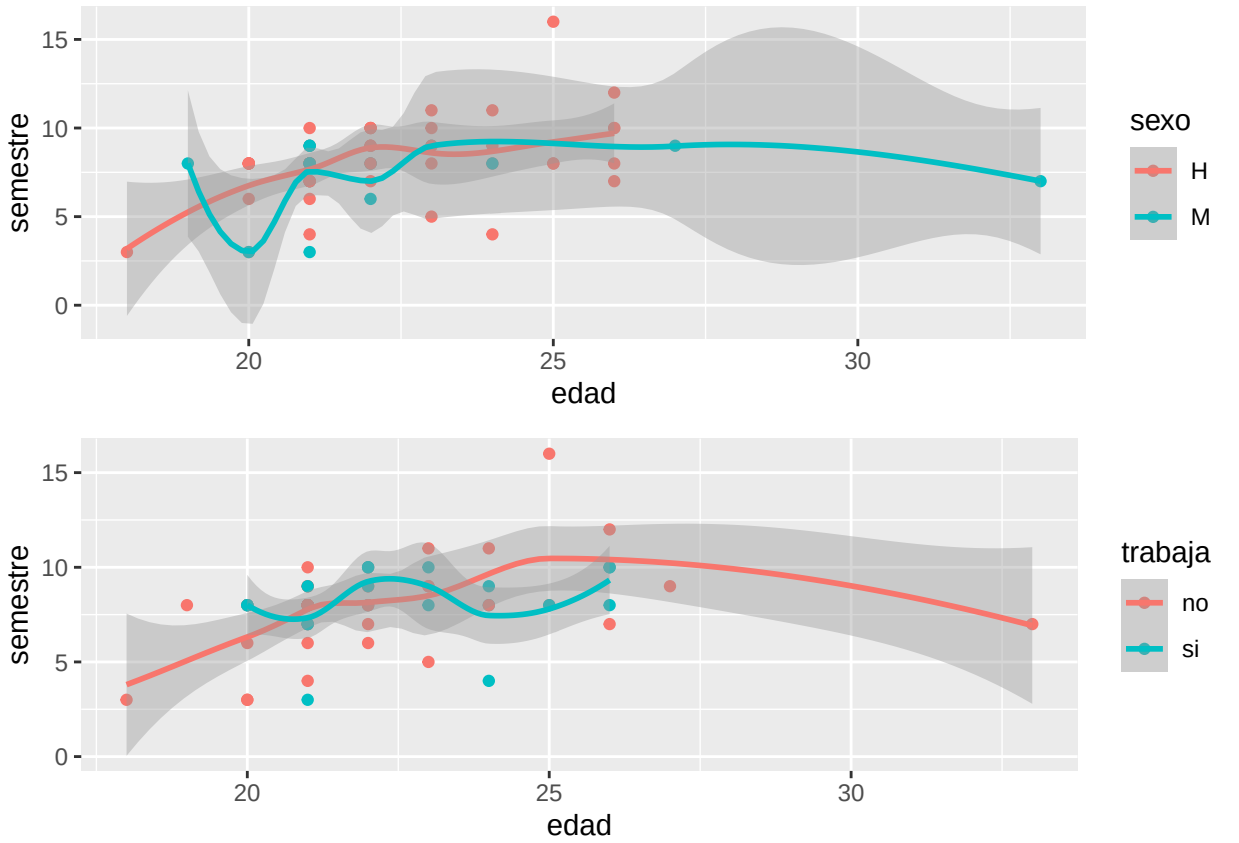
```
## (Intercept)  19.3497     1.0535  18.366  < 2e-16 ***
## df$semestre   0.3525     0.1279   2.756  0.00754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.201 on 67 degrees of freedom
## Multiple R-squared:  0.1018, Adjusted R-squared:  0.08839
## F-statistic: 7.593 on 1 and 67 DF,  p-value: 0.007539
```

Acontinuación se observa graficas con regresion lineal y Loess discriminados por sexo y por trabajo, graficamente se puede pensar que el modelo loess tiene una curva más suave que capta mejor el comportamiento de los datos, para aclarar esto analiticamente seria conveniente compararlos con medidas numericas.

```
# Add the regression line
g7 <- ggplot(df, aes(x=edad, y=semestre,color=sexo)) +
  geom_point()+
  geom_smooth(method=lm)

g8 <- ggplot(df, aes(x=edad, y=semestre,color=trabaja)) +
  geom_point()+
  geom_smooth(method=lm)

g9 <- ggplot(df, aes(x=edad, y=semestre,color=sexo)) +
  geom_point()+
  geom_smooth()

g10<- ggplot(df, aes(x=edad, y=semestre,color=trabaja)) +
  geom_point()+
  geom_smooth()

grid.arrange(g7,g8,nrow=2)
```

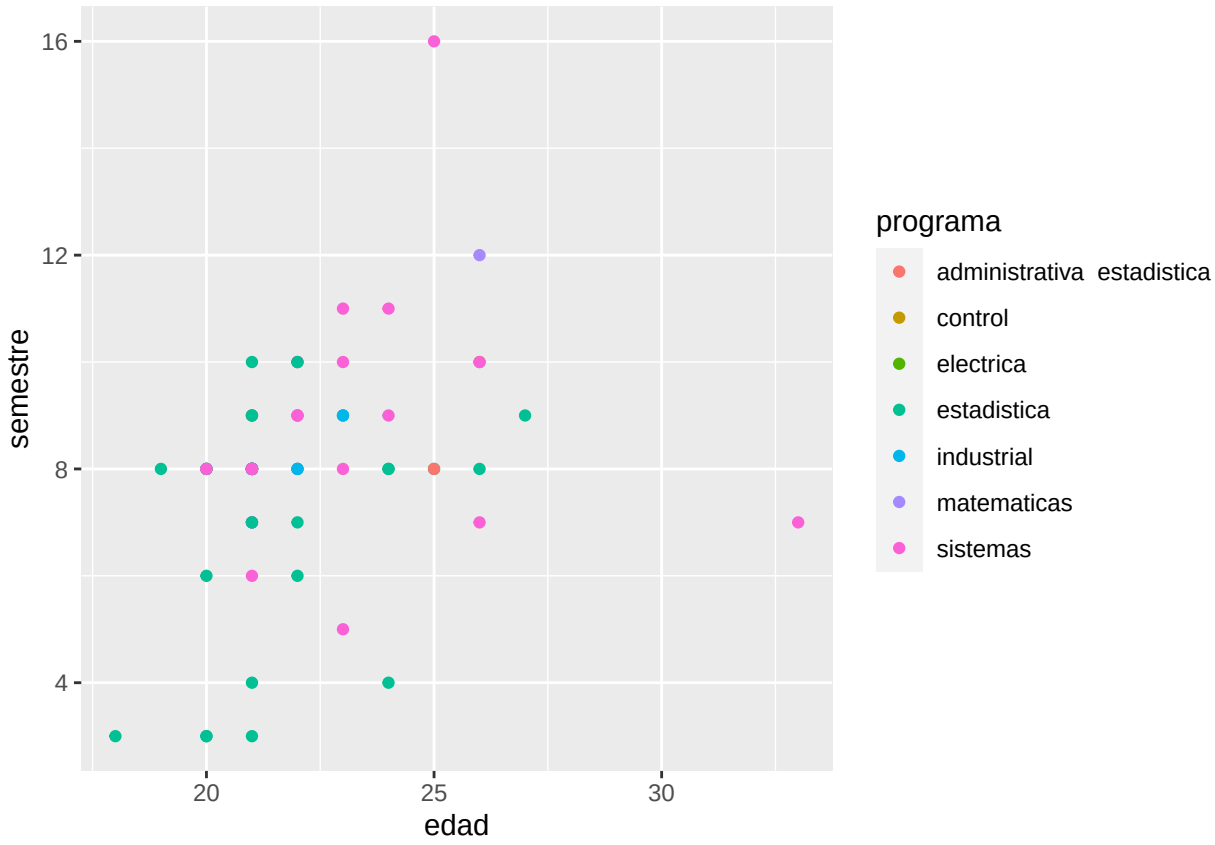



```
grid.arrange(g9, g10, nrow =2)
```



En esta grafica se observa que las observaciones atipicas grandes hacen parte del programa de ingenieria de sistemas y que las atipicas pequenas son de estadisticas.

```
ggplot(df, aes(x=edad, y=semestre,color=programa)) +
  geom_point()
```



Formulacion e hipotesis:

Si consideramos que nuestros datos son prospectivos podemos hacernos las siguientes hipotesis:

hipotesis 1: la poblacion de estudiantes de TAE son mayormente hombres?

hipotesis 2: la mediana de los esetudiantes son de 8avo semestre?

hipotesis 3: videojuegos y leer son preferidos igualmente?

hipotesis 4: La edad promedio de los estudiantes esta entre 22 y 23 años?

hipotesis 5: Los estudiantes que no trabajan son mas que los que si?