Tarea 2

Luis Gagñevin

5/21/2021

Antes de empezar llamamos las librerias que precisaremos en el trabajo

```
library(gapminder)
library(tidyverse)
library(ggplot2)
library(readr)
library(ggpmisc)
```

Las librerias que utilizaremos seran:

- Gapminder (Datos)
- tidyverse (Manipulacion de los Datos y Matrices)
- ggplot2 (Se utilizara para la creacion de graficos)
- readr (Datos)
- ggpmisc (Complemento para ggplot2)

Todas las figuras y graficos estan autocontenidas y tienen la informacion necesaria para ser entendidas.

Ejercicio 1

1.

Hacer un grafico de dispersion que tenga el eje y: year y en el eje x: lifeExp, los puntos deben estar coloreados por la variable continent. Para este plot ajusta una recta de regresion para cada continente sin incluit las barras de error. Las etiquetas de los ejes deben ser claras y describir las variab les involucradas. Incluir un caption en la Figura con algun comentario de interes que describa el grafico. El resto de los comentarios del grafico se realizaran en el texto

```
gd<- ggplot(gapminder, aes(lifeExp, year, color=continent))
gd+geom_point()+geom_smooth(se=FALSE)+
  labs(x="Esperanza de Vida",
       y="Año",
       colour="Continente",
       title="Esperanza de vida por año segun cada continente")
```

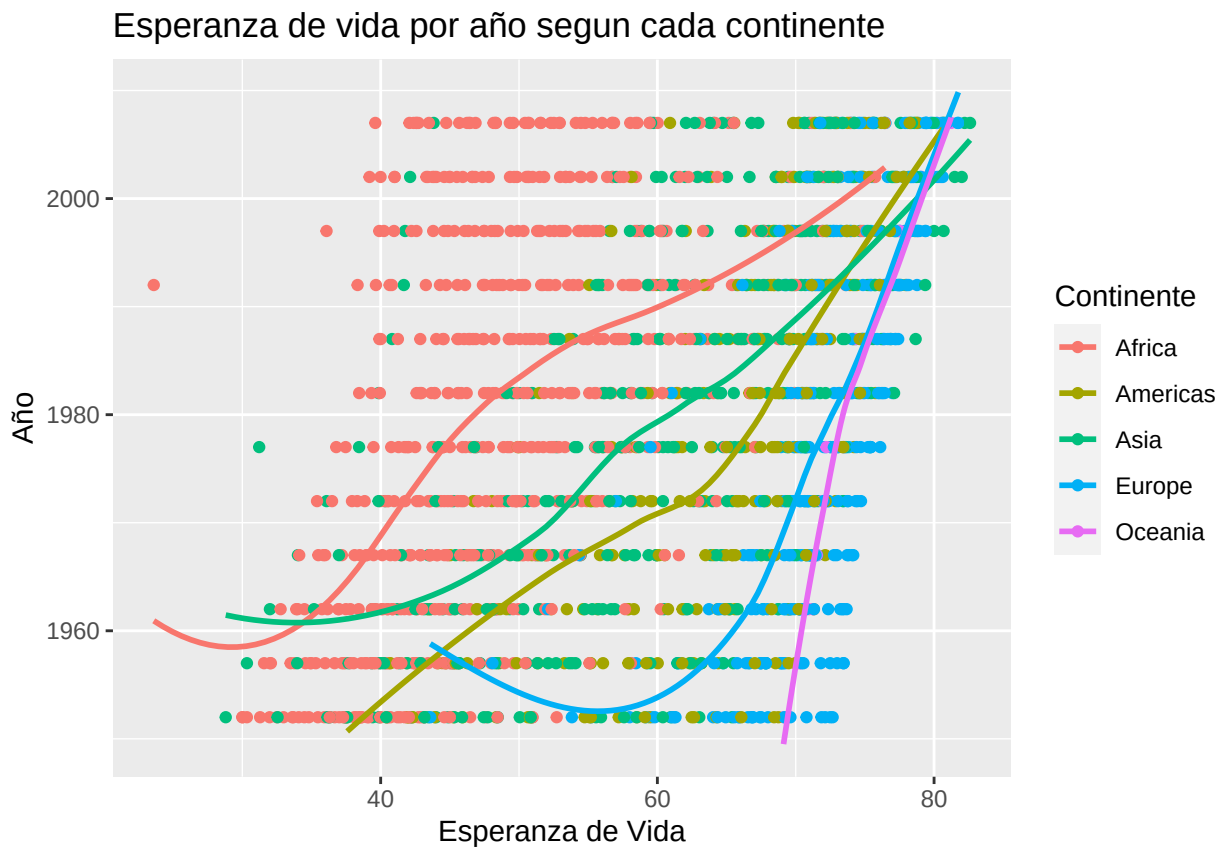


Figure 1: Diagrama de dispersion con plot de regresion lineal

Aqui tenemos un grafico que contiene las variables de esperanza de vida y año y continente, podemos ver como la esperanza de vida en general aumenta en todo continente, en algunos continentes la esperanza aumenta menos que en otros, por ejemplo en Oceania, pasa de los 69 años hasta los 80 mientras que en paises como Africa toma un recorrido desde los 20 hasta los 75 aprox.

2.

Omitir la capa de geom_point(), Las lineas aun aparecen pero los puntos no. ¿Porque sucede esto?

```
gd+geom_smooth(se=FALSE)
```

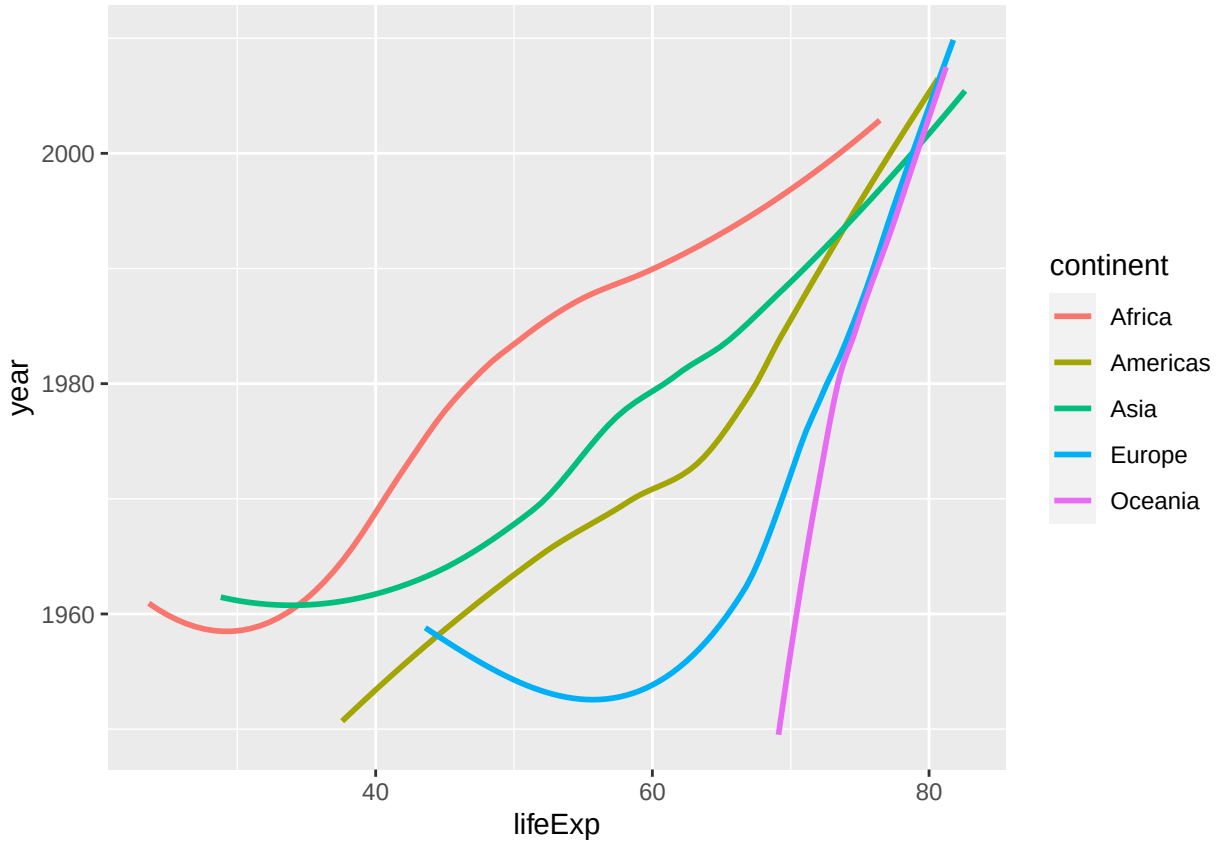


Figure 2: Diagrama de regresion lineal

La respuesta a esto es sencilla, estamos sobreponiendo dos ploteos los cuales comparten variables para poder ver la dispersion en el grafico anterior, sin embargo, las lineas que se forman son independientes a la existnecia de los puntos en el ploteo ya que son parte de geom_smooth() y no de geom_point().

3.

El siguiente es un grafico de dispersion entre lifeExp y gdpPercap coloreado por continent. Usando como elemento estetico color (aes()) nosotros podemos distinguir los distintos continentes usando diferentes colores de similar manera usando forma (shape) Dicho grafico esta sobrecargado. ¿Como lo modificarias para que sea mas clara la comparacion para los distintos continentes y porque? Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Comenta alguna caracteristica interesante que describa lo que aprendes viendo el grafico

```
gd<- ggplot(gapminder, aes(gdpPercap, lifeExp))
gd+geom_point(color='skyblue')+facet_grid(.~continent, scales = "free_x")
```

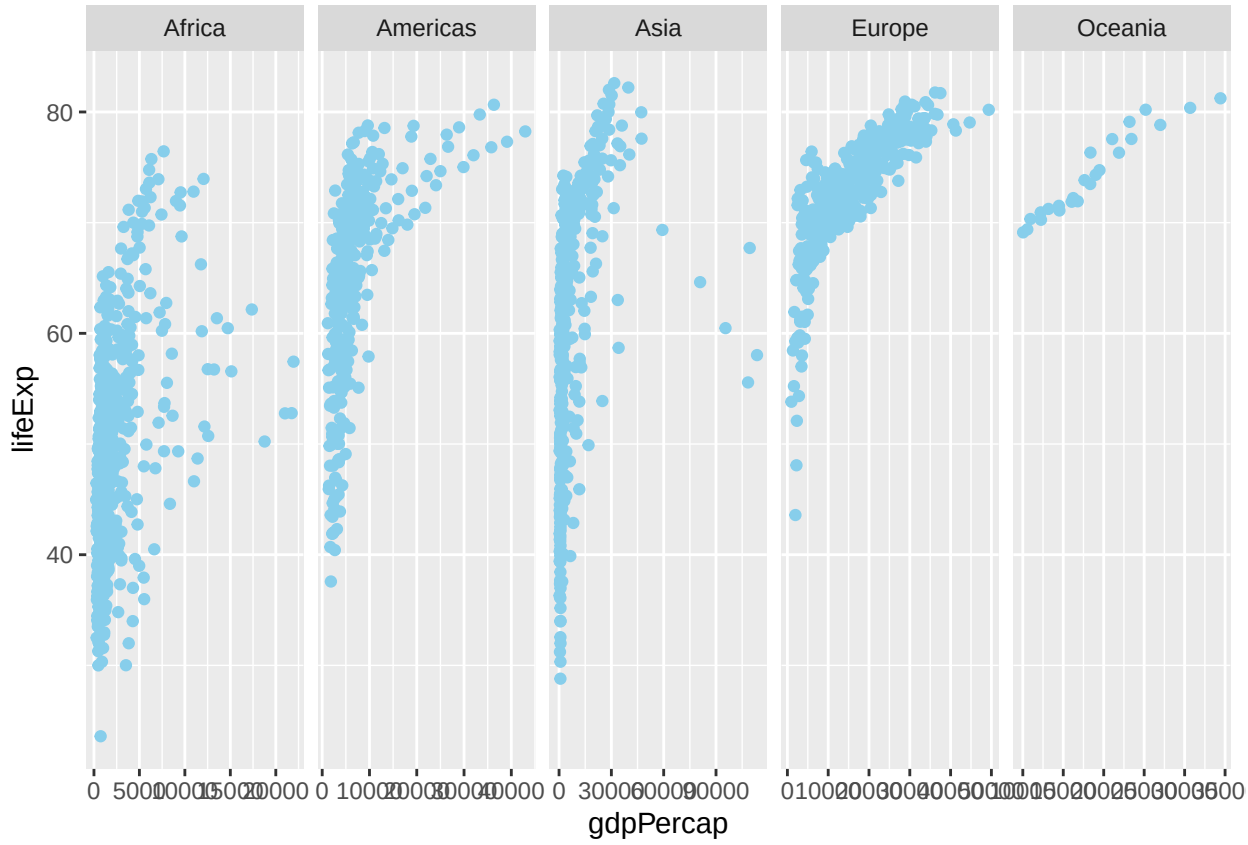


Figure 3: Comparativa entre esperanza de vida y PBI per capita

Es interesante ver como el continente de oceania tiene la mayor esperanza de vida y no varia mucho segun el GDP per capita, como sucede en otros continentes donde la variacion es notablemente importante segun el nivel de vida.

Para hacer el grafico mas lindo a la vista, los separe segun el continente usando un facet_grid(), deja mejor visto y abierto a una comparacion mas linda y a la vista que un facet_wrap() y que sin utilizarlo. Tambien elimine los valores sueltos que serian los de un GDP mayor a 40000 ya que los casos mayores son muy aislados y no son relevantes para una comparacion eficaz entre los continentes.

Comentario: Bien, te modifiqué que en vez de fijar los límites a mano, liberes la escala del eje x. Otra opción es que uses log. No elimines valores atípicos, analizalos. Y si queres plantea un gráfico adicional sin esos casos, pero eso de eliminarlos en un análisis exploratorio esta mal.

4.

Hacer un grafico de lineas que tenga en el eje x year y en el eje y gdpPercap para cada continente en una misma ventana grafica. En cada continente, el grafico debe contener una linea para cada pais a lo largo del tiempo (Serie de tiempo de gdpPercap). Las etiquetas de los ejes deben ser claras y describ ir las variables involucradas. Incluir un caption en la Figura con algun comentario de interes que describa el grafico

```
gd<- ggplot(gapminder, aes(year, gdpPercap, group = country))
gd+ geom_line(color="darkred", alpha = 0.4)+facet_grid(gapminder$continent)+ylim(0,40000)+
  labs(y="PBI per capita",
       x="Año",
       title = "Relacion PBI per capita por año para cada continente",
       caption=)
```

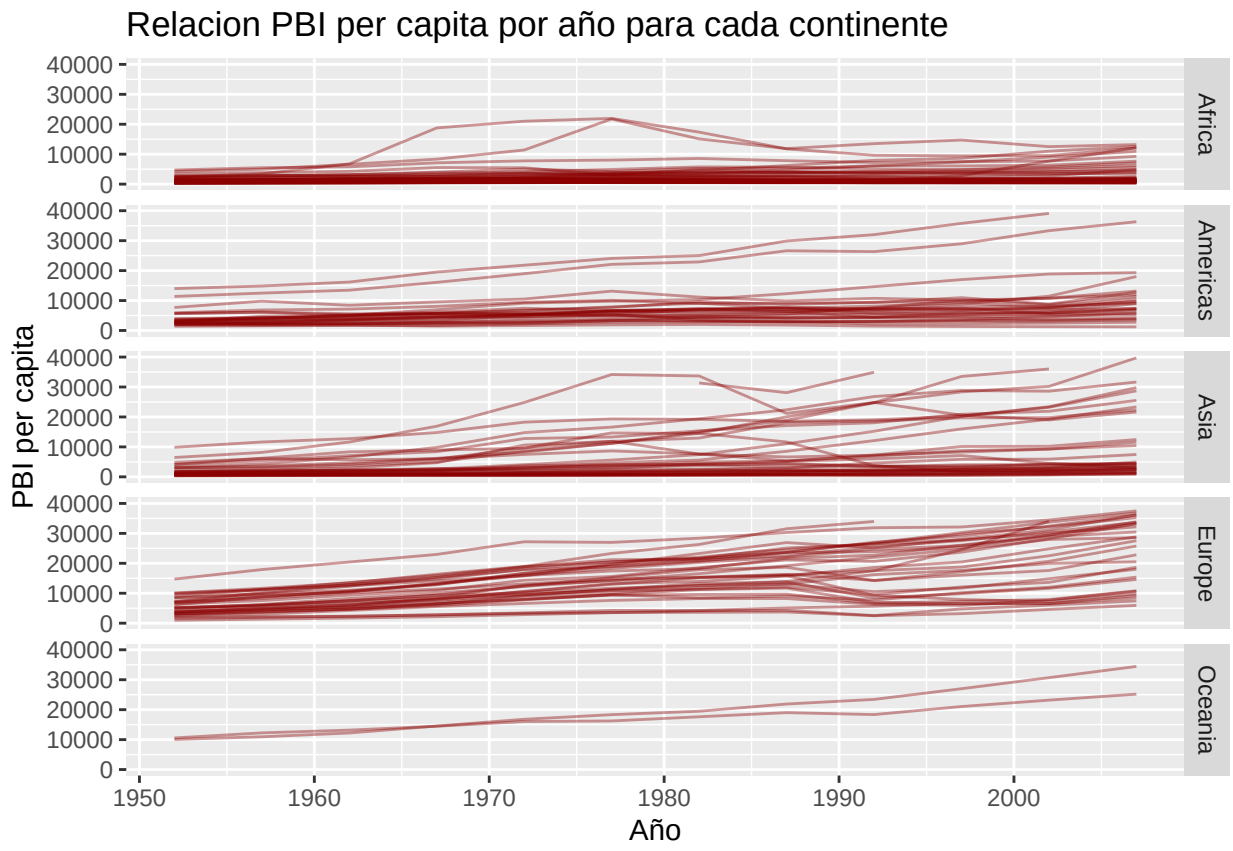


Figure 4: Grafico de lineas entre el año y el GDP per capita de gapminder

Comentario: Esta mal, falta agrupar por país y darle transparencia.

5.

Usando los datos de gapminder seleccione una visualizacion que describa algun aspecto de los datos que no exploramos. Comente algo interesante que se pueda aprender de su grafico

```
dg<- gapminder %>% filter(continent=="Oceania" | continent=="Asia")
ggplot(dg, aes(year, pop))+geom_col()+
  facet_wrap(dg$continent, scales="free")+
  scale_y_continuous(labels = scales::label_number_si())+
  labs(x="Año", y="Poblacion")
```

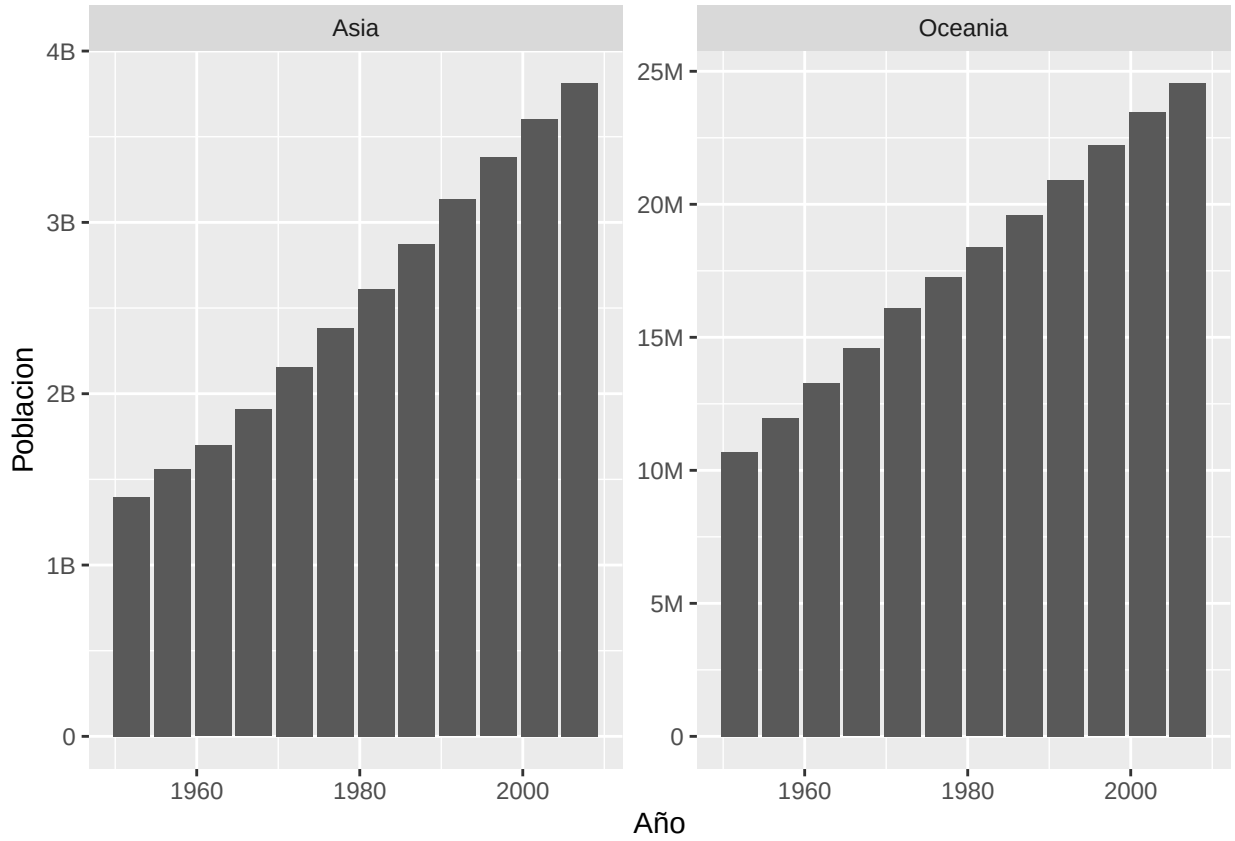


Figure 5: CRecimiento poblacional entre Asia y Oceania

Decidi investigar primero que continente tenia el mayor crecimiento y cual el menor. Vi que el continente con mayor crecimiento es ASIA y el menor es Oceania. Oceania tiene una poblacion 160 veces menor que la de ASIA.

Ejercicio 2

1.

Con los datos mpg que se encuentran en ggplot2, hacer un grafico de barras para la variables drv con las siguientes caracteristicas:

- Las barras tienen que estar coloreadas por drv
- Incuir usando labs() el nombre de los ejes y titulo informativo
- Usa la paleta Dark2

```
data(mpg)
ggplot(mpg, aes(x=drv, fill=drv))+geom_bar()+scale_fill_brewer(palette = "Dark2")+
  labs(x="Traccion", y="Cantidad de vehiculos", title="Traccion vehicular")
```

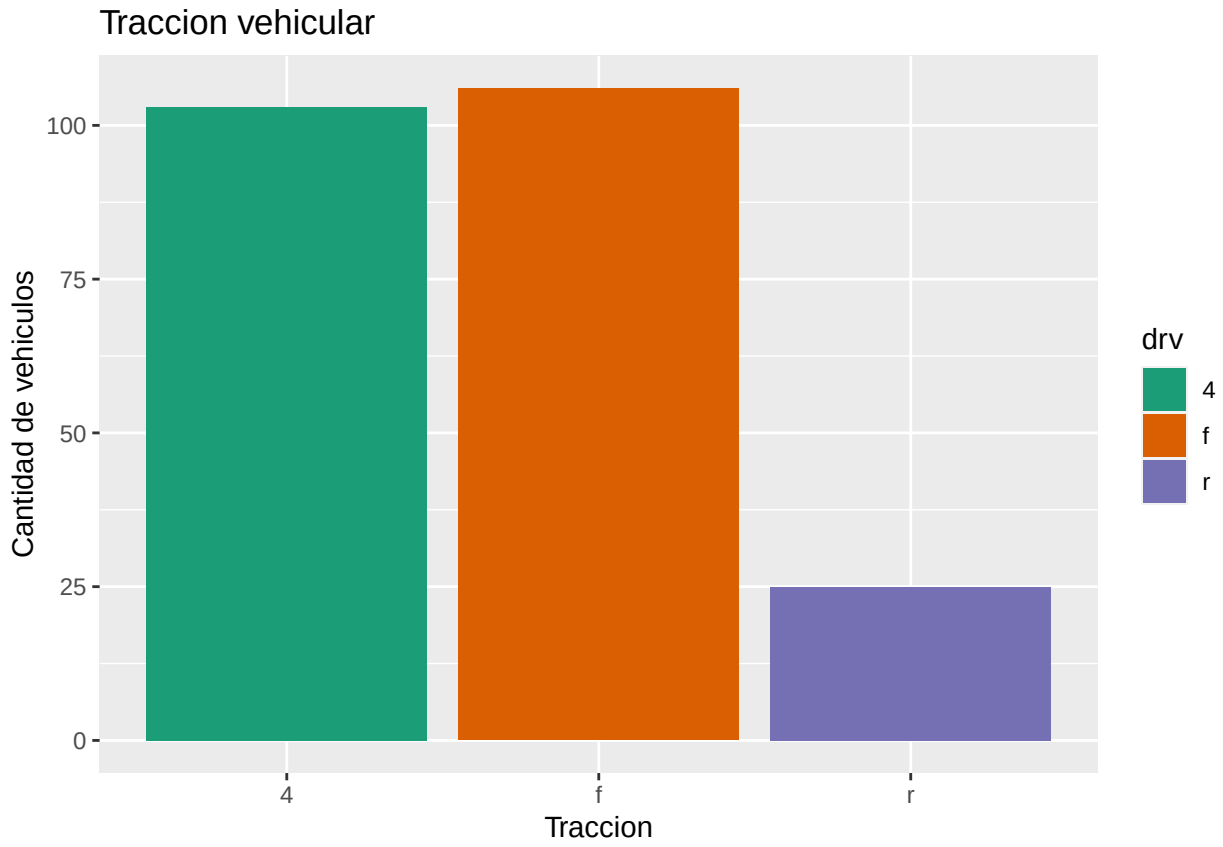


Figure 6: Comparativa de tracciones vehiculares en el mercado

Comentario: Ordenar de forma descendente y agregar descripción a las etiquetas.

2.

Usando como base el grafico anterior:

- Incluir en el eje y porcentaje en vez de conteos
- Usando scale_y_continous() cambiar la escala del eje y a porcentajes
- Usando geom_text() incluir texto con porcentajes arriba de cada barra

Primero debemos de generar con tidyverse un grupo de valores con el porcentaje de veces que aparece en la columna drv cada tipo de traccion (4 f r).
Luego pasamos al grafico donde utilizamos un geom_bar que utilizara la x segun los valores porcentuales de Y y colocara sobre estas columnas el % aproximado a 1 cifra de cada una.

En la parte estetica, utilizamos un scale_fill_brewer y no un color ya que estamos utilizando un fill= en la parte grafica.
Luego renombramos las columnas, titulos y ejes y borramos la leyenda

```
data(mpg)
mpgp<- mpg %>% group_by(drv) %>% summarize(veces = n()) %>% mutate(prct=veces/sum(veces))

ggplot(mpgp, aes(drv,prct, fill=drv))+
  geom_bar(stat='identity')+
  geom_text(aes(label=scales::percent(prct),vjust= -.3))+
  scale_y_continuous(labels=scales::percent)+
  scale_fill_brewer(palette = 'Dark2')+
  scale_x_discrete(labels=c("4 ruedas", "Frontal", "Trasera"))+
  labs(x="Traccion", y="Cantidad de vehiculos", title="Traccion vehicular")+
  theme(legend.position="none")
```

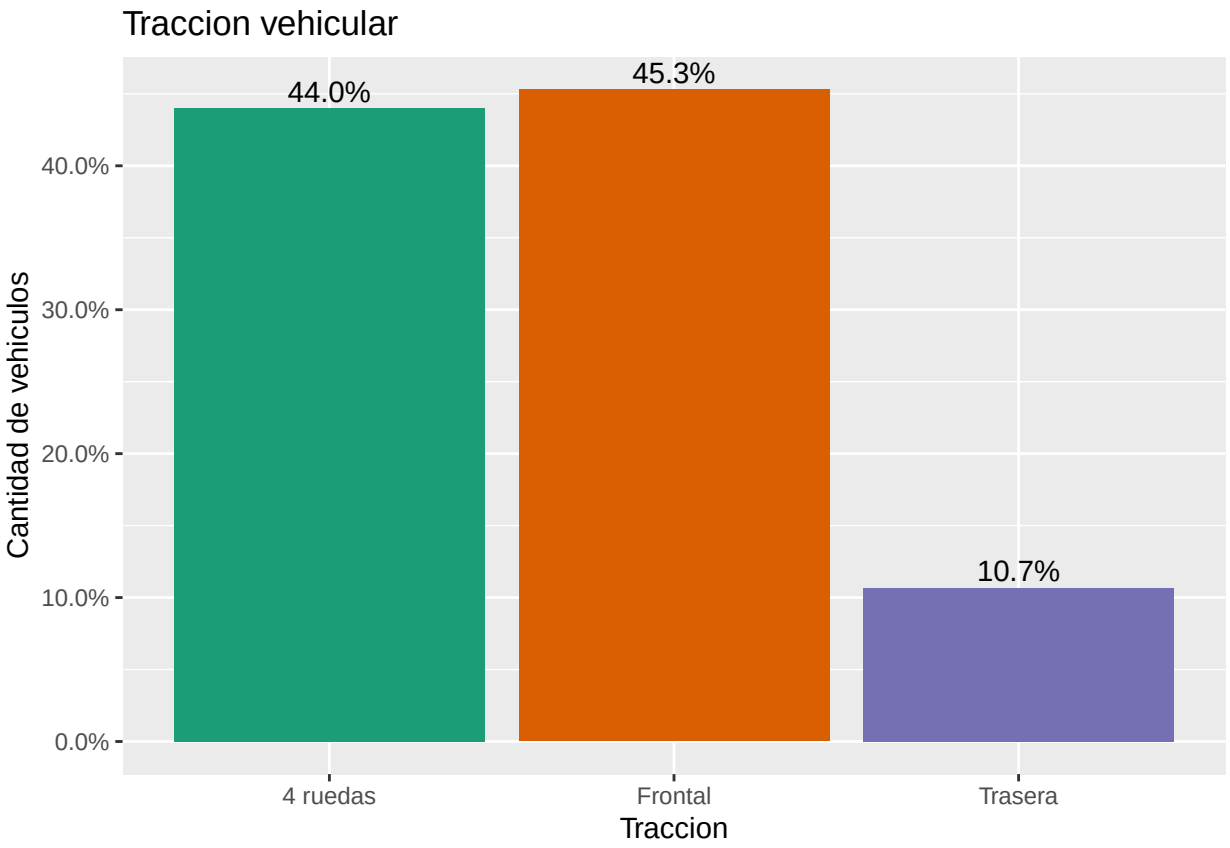


Figure 7: Traccion predominante del mercado

Ejercicio 3.

1.

Leer los datos usando el paquede readr y la funcion read_csv(). Guardarlos en un objeto llamado datos.

```
datos <- read_csv('dato_emision.csv')
```

2.

Usando las funciones de la libreria dplyr, obtenga que fuentes tienen la emision maxima. Recuerde que total deberia ser excluido para esta respuesta asi como los subtotales

```
datosnew<- datos %>% filter(fuente != 'TOTAL' & fuente != 'I_E' & fuente != 'S_C' & !is.na(emision)) %>%
datosnew[1:3,]
```

```
## # A tibble: 3 x 3
##     AÑO fuente emision
##   <dbl> <chr>    <dbl>
## 1  2017 Q_B      9070.
## 2  2016 Q_B      8831.
## 3  2015 Q_B      8497.
```

3.

En que año se dio la emision maxima para la fuente que responde la pregunta anterior.

Con la funcion siguiente podemos ver el año en que se dio la emision maxima.

```
datosnew[which.max(datosnew$emision), 1]
```

La emision mas alta fue en: 2017

4.

Usando las funciones de la librerya dplyr obtenga las 5 fuentes, sin total ni subtotales que tienen un valor medio de emision a lo largo de todos los años mas grandes.

Para hacer eso, agrupamos y creamos el "ValorMedio" el cual tendra el mean. Luego los ordenamos de manera descendente y tomatmos los primeros 5 valores y seleccionamos solo la columna para saber quienes son los que mayor emision producen en promedio.

```
datos2 <- datosnew %>% group_by(fuente) %>% summarise(ValorMedio = mean(emision, na.rm=TRUE)) %>% arran
datos2[1:5, 1]
```

Las 5 mayores emisiones son: c("Q_B", "T", "BI", "CE_SP", "I")

5.

Usando ggplot2 realice un grafico de las emisiones a lo largo de los años para cada fuente. Utilice dos elementos geometricos, puntos y lineas. Selecciones para dibujar solamente las 5 fuentes que a lo largo de los años tienen una emision media mayor que el resto. Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un caption en la figura con algun comentario de interes que describa el grafico.

Renombre las variables para que sea entendible para el publico.
Para dar una mayor claridad coloree los graficos de forma que contrasten y se noten los puntos donde hubo un dato.

```
datosmax<- filter(datosnew, fuente %in% as.matrix(datos2[1:5,1]))
datosmax$fuente <- factor(datosmax$fuente, labels=c("Bunkers Internacionales","Centrales Electricas S.Pu
                                                    'Industrial','Quema de Biomasa',"Transporte"))


a<- ggplot(datosmax, aes(AÑO, emision))
a+geom_line(color="darkgreen")+geom_point(color="darkred")+facet_wrap(~datosmax$fuente, scales = 'free')
  labs(y="Emision de CO2", x= "Año",
       title = "Evolucion de Emision de Co2 segun sector")
```

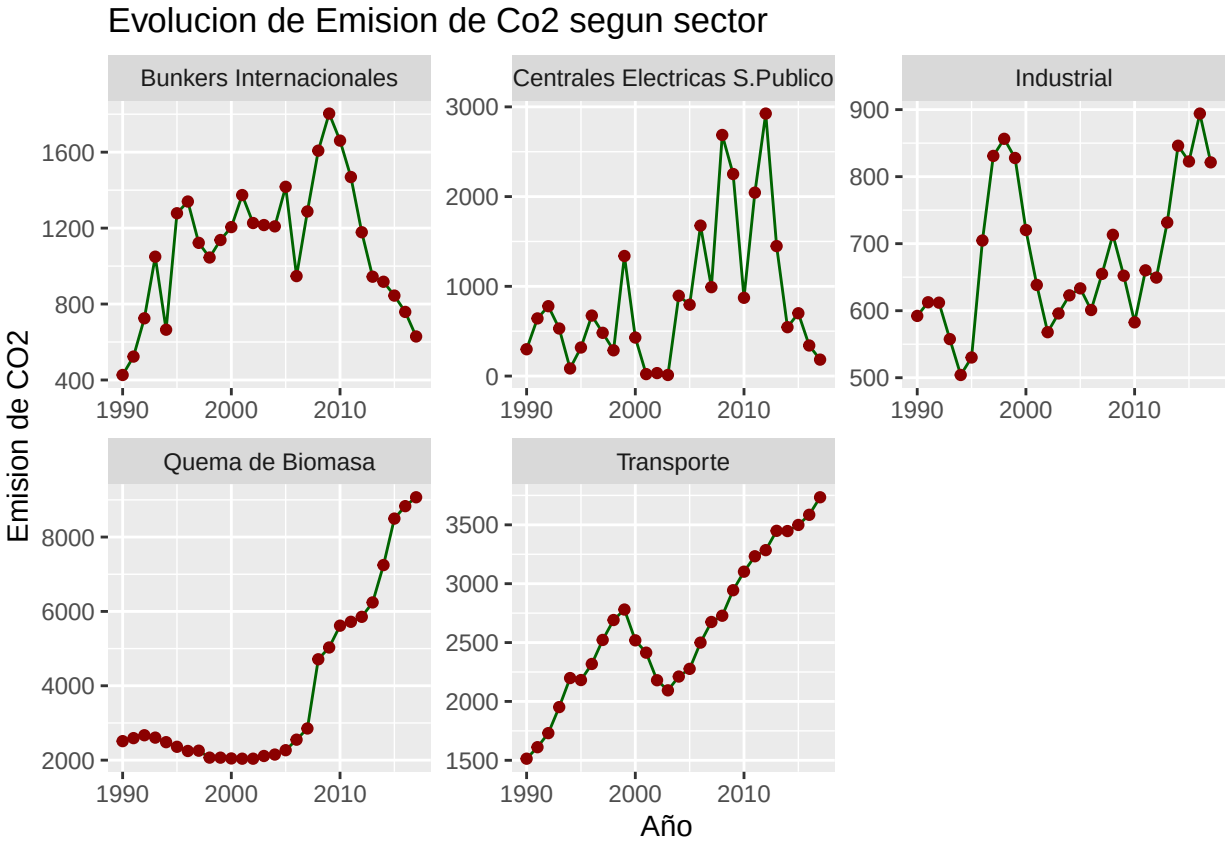


Figure 8: Emision de Co2 desde 1990-2017, datos tomados por el MIEM

6.

Replique un grafico usando ggplot2.

```
ggplot(datos, aes(reorder(fuente, -emision, mean, na.rm=TRUE),emision))+geom_boxplot()+geom_line()+
labs(y="Emision de CO2 en Gg", x="Fuentes con mayor emision media entre 1990-2016")
```

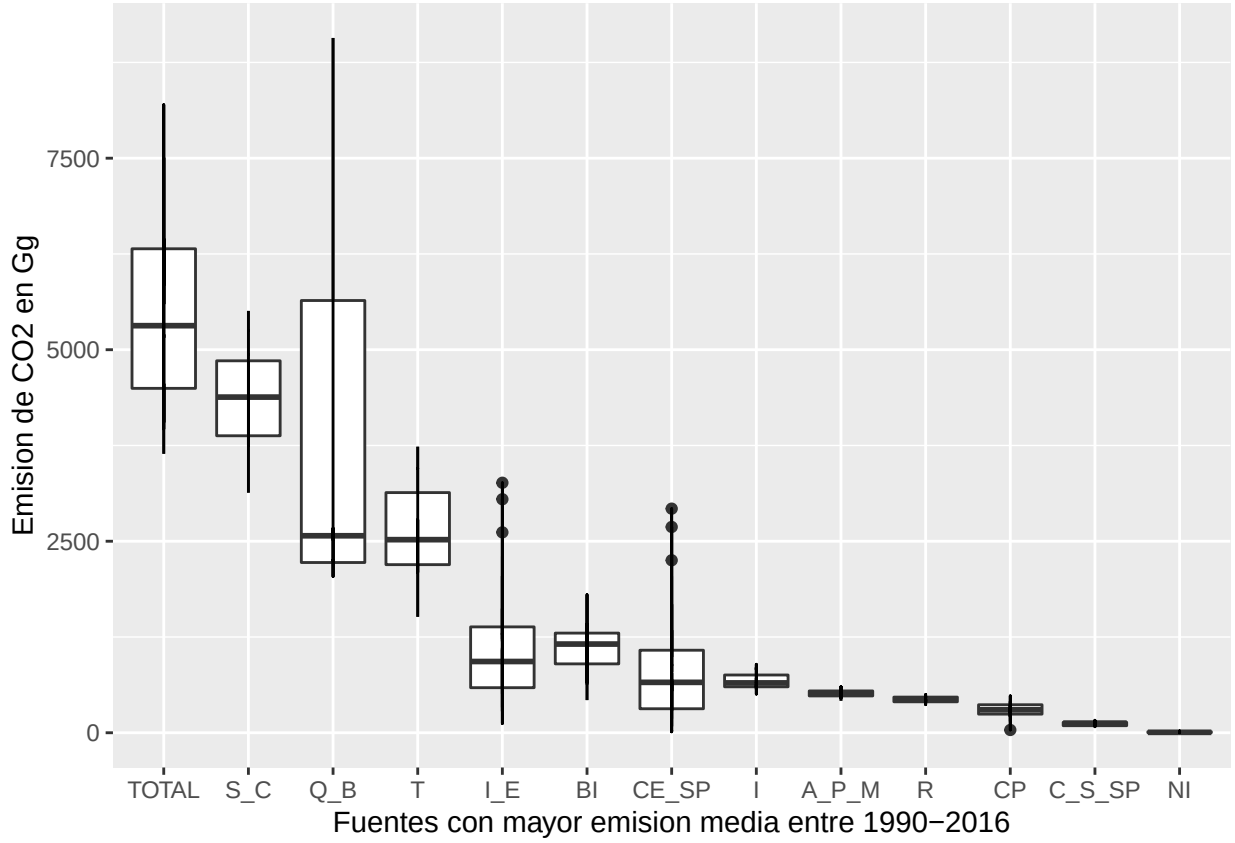


Figure 9: Grafico de cajas y bigotes con segmento de linea que muestra el rango

7.

Usando la libreria de ggplot2 y ggpmisc replique el siguiente grafico de las emisiones totales entre 1990 y 2016. Los puntos rojos indican maximos locales o picos de emiision de CO2 en Gg.

```
a<- datos %>% filter(fuente=='TOTAL')
ggplot(a, aes(x=AÑO, y=emision))+geom_line()+geom_point()+
  stat_peaks(color='red')+stat_peaks(geom='text', vjust=-.5, color='red')+
  labs(y="Emision de CO2 en Gg", x="Año",
       title='Picos de emision entre 1990-2017')
```

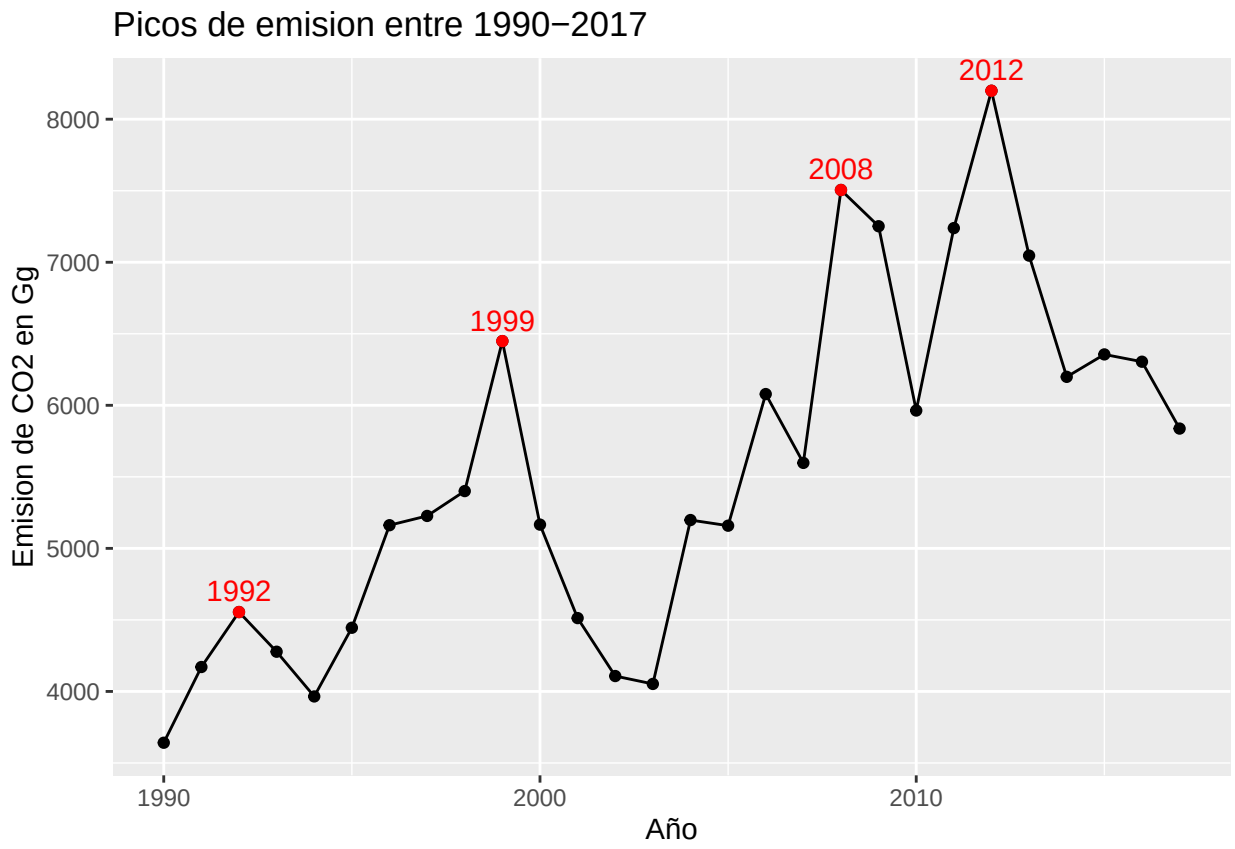


Figure 10: El grafico esta hecho por un geom_line, geom_point y utilizando stat_peaks para ver los picos

Ejercicio 4.

Estudiar una muestra de datos a nivel nacional sobre abandono en los años 2016.
Realizar tres preguntas de interes que surjan como parte del analisis exploratorio de datos.

```
datos <- read_csv('muestra.csv')
```

Para mi tres preguntas importantes son:

- ¿Cuál es el sexo que abandona más? (Tomando en cuenta que ABandono=0 No, Abandono=1 Si,)
- ¿En qué departamento hay mayor abandono promedio?
- ¿En qué Departamentos el Contexto sociocultural de los liceos es más bajo?

Pregunta 1:

```
sexoab<- datos %>% select(Sexo, Abandono)
sexoab<- sexoab %>% group_by(Sexo) %>% summarise(Abandono=sum(Abandono))


ggplot(sexoab, aes(Sexo, Abandono, fill=Sexo))+geom_bar(stat='identity', width=0.5)+
  geom_text(label=sexoab$Abandono, vjust=-.5, color='DarkGreen')+
  theme(aspect.ratio = 2/1)+scale_fill_brewer(palette = "Dark2")+
  labs(x="Sexo", y="Cantidad de alumnos que abandonaron",
       title="Comparacion de abandono entre sexos")
```

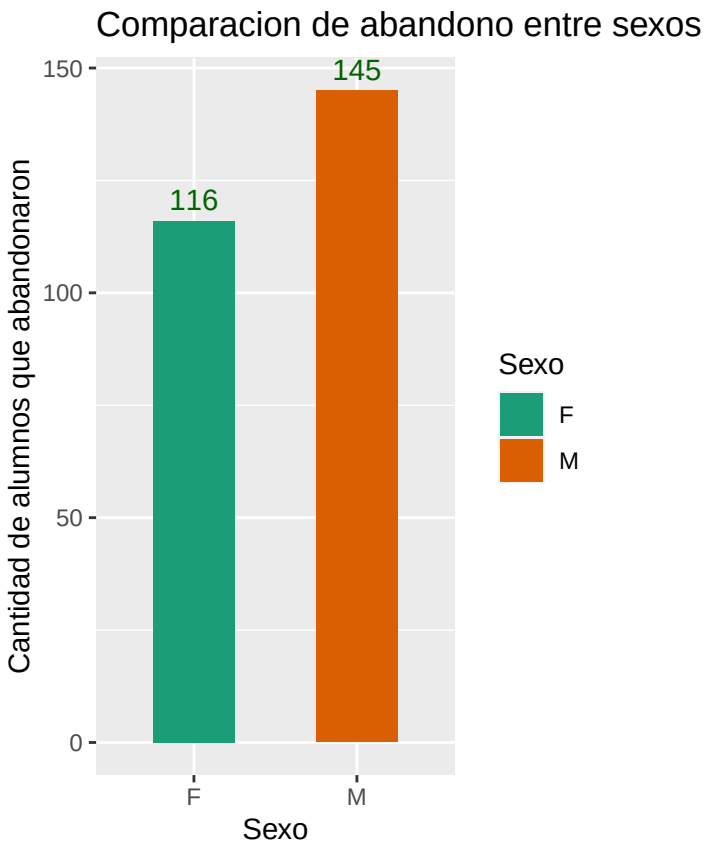


Figure 11: Grafico de barras que muestra que sexo es mas propenso al abandono de clase

Podemos ver que es mas propenso a abandonar el sexo masculino que el fenemnino

Pregunta 2:

```
dep<- datos %>% select(nombre_departamento, Abandono)

NoAbandono <- dep %>% group_by(nombre_departamento)%>% count(a=Abandono==0) %>%filter(a==TRUE) %>% sele

dep<- dep %>% group_by(nombre_departamento) %>% summarise(Abandono=sum(Abandono))%>% mutate(NoAbandono)

dep2<- dep %>% summarise('indice'=Abandono/(Abandono+NoAbandono))

dep<- dep %>% select(nombre_departamento) %>% mutate(dep2)



ggplot(dep, aes(x=reorder(nombre_departamento,-indice), indice, fill=indice))+geom_bar(stat='identity')+
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.25))+
  labs(x="", y="Porcentaje de abandono", title= 'Indice de Abandono escolar por departamento')
```

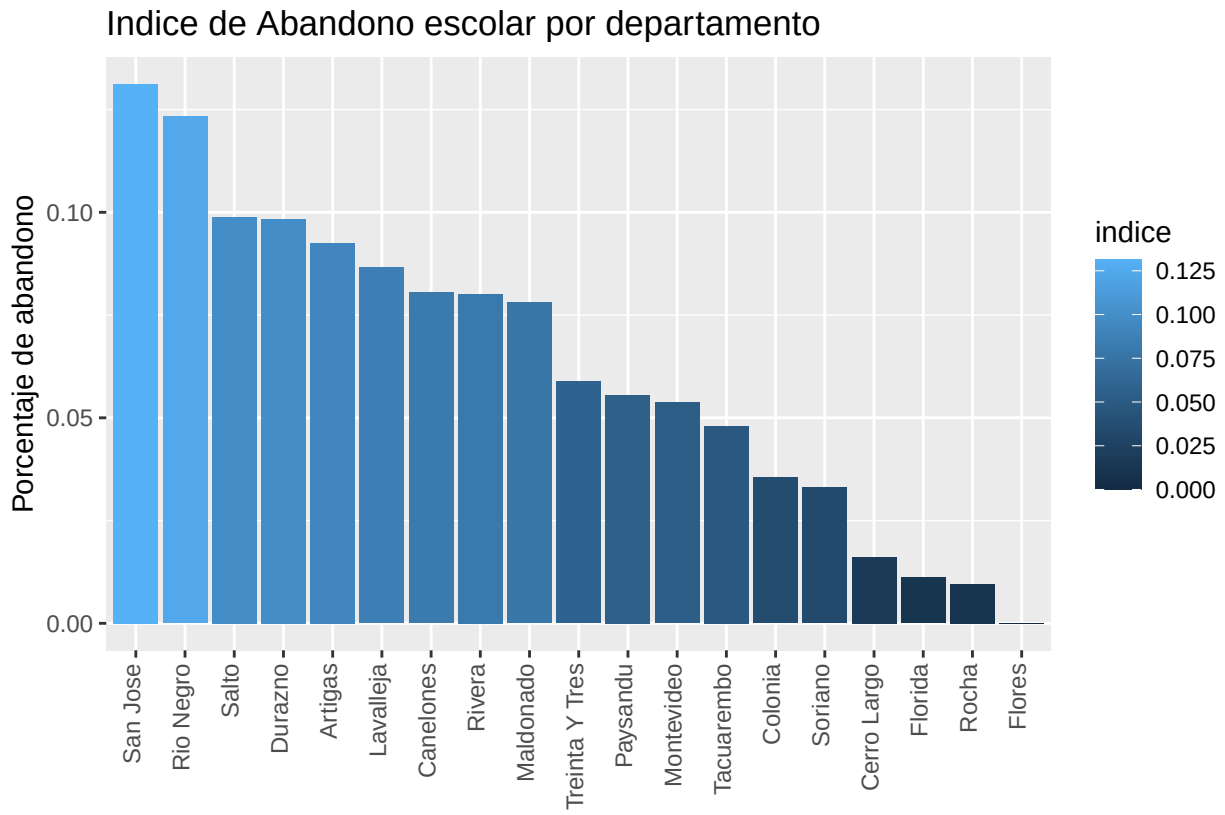


Figure 12: Indice de abandono escolar por cada departamento

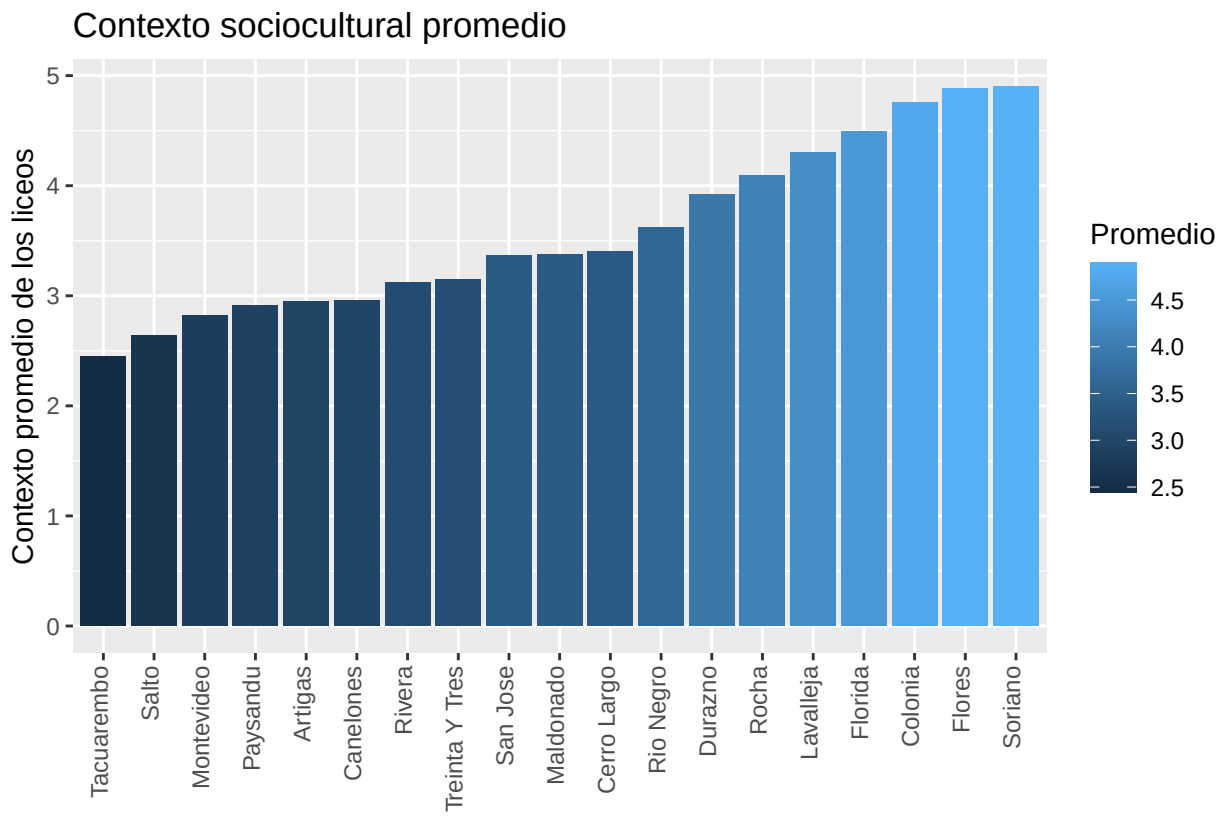
Podemos ver que el departamento con mayor indice de abandono escolar es San José, y el que menor indice tiene es Flores.

Comentario: Muy bien! El índice sería el porcentaje de abandonos.

Pregunta 3.

```
contx<- datos %>% select(cl, nombre_departamento)%>% group_by(nombre_departamento) %>%
  summarise(Promedio =mean(cl))

ggplot(contx, aes(x=reorder(nombre_departamento, Promedio),y=Promedio, fill=Promedio))+geom_bar(stat='i
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.25))+
  labs(x="", y="Contexto promedio de los liceos",
       title= 'Contexto sociocultural promedio')
```



Podemos ver comparando este gráfico y el anterior, que las zonas con mejor contexto tienen un menor indice de abandono.

Comentario: Esta buena la idea pero cl no es una variable continua para calcular su promedio.

{Muy buen trabajo! Hay un par de errores respecto a fijar el method = "lm", agrupar cuando se va a gráficar lineas y otros comentarios respecto a ordenar las barras y fijar etiquetas explicativas. Se puede ahondar más en las preguntas y análisis planteados, por ejemplo plantear hipótesis que surjen de estos análisis. Importante: no sacar datos que pueden ser atípicos del análisis exploratorio, de hecho allí es donde surgen preguntas respecto a esas observaciones.}