

Tarea individual 2

Lorena Pérez

25/05/2021

```
## Warning: package 'gapminder' was built under R version 4.0.5
```

Entrega

Esta tarea tiene que estar disponible en su repositorio de GitHub con el resto de las actividades y tareas del curso el 26 de Mayo. Asegurate que tanto Federico como yo seamos colaboradoras de tu proyecto privado Tareas_STAT_NT. Recordar seleccionar en en opciones de proyecto, codificación de texto UTF-8. La tarea debe ser realizada en RMarkdown, la tarea es individual por lo que cada uno tiene que escribir su propia versión de la misma. El repositorio debe contener el archivo .Rmd con la solución de la tarea y los archivos que sean necesarios para su reproducibilidad. Vamos a utilizar la biblioteca `gapminder`, por lo que si no la usaste anteriormente tenés que instalarla y luego cargarla. Para obtener la descripción del paquete `library(help = "gapminder")` y para saber sobre la base `?gapminder`.

Recordá que todas las Figuras deben ser autocontenidas, deben tener toda la información necesaria para que se entienda la información que se presenta. Todas las Figuras deben tener leyendas, títulos. El título (caption) debe contener el número de la Figura así como una breve explicación de la información en la misma. Adicionalmente en las Figuras los nombre de los ejes tienen que ser informativos. En el YAML en Tarea_2.Rmd verás `fig_caption: true` para que salgan los `caption` en el chunk de código debes incluir `fig.cap = "Poner el que tipo de gráfico es y algún comentario interesante de lo que ves"`. Luego en el cuerpo del documento podés hacer comentarios extendidos sobre lo que muestra la figura.

Idea básica de regresión lineal

Una regresión lineal es una aproximación utilizada para modelar la relación entre dos variables que llamaremos X e Y. Donde Y es la variable que queremos explicar y X la variable explicativa (regresión simple).

El análisis de regresión ajusta una curva a través de los datos que representa la media de Y dado un valor especificado de X. Si ajustamos una regresión lineal a los datos consideramos “la curva media” como aquella que mejor ajusta a los datos.

Algunas veces ajustamos curvas genéricas promediando puntos cercanos entre si con métodos de suavizado no necesariamente lineales. ¿Cómo incluimos una recta de regresión en nuestro gráfico?

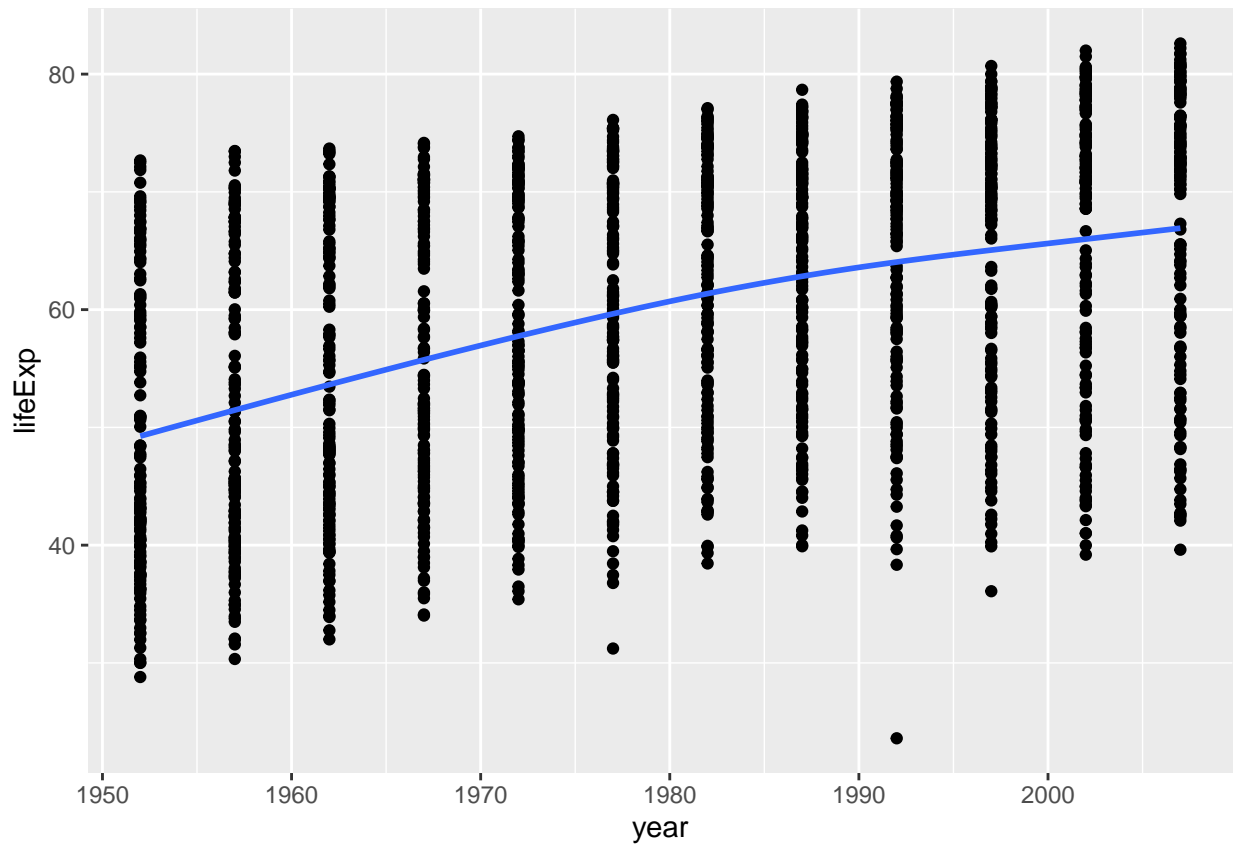
Para agregar una linea de regresión o una curva tienen que agregar una capa a tu gráfico `geom_smooth`. Probablemente dos de los argumentos más útiles de `geom_smooth` son:

- `method = ...`
 - ... “lm” para una linea recta. `lm` “Linear Model”.
 - ... otro para una curva genérica (llamada de suavizado; por defecto, es la parte `smooth` de `geom_smooth`).
 - `se=...` controla si los intervalos de confianza son dibujados o no.

Ejemplo:

```
vc1 <- ggplot(gapminder, aes(year, lifeExp)) +  
  geom_point()  
vc1 + geom_smooth(se = FALSE)
```

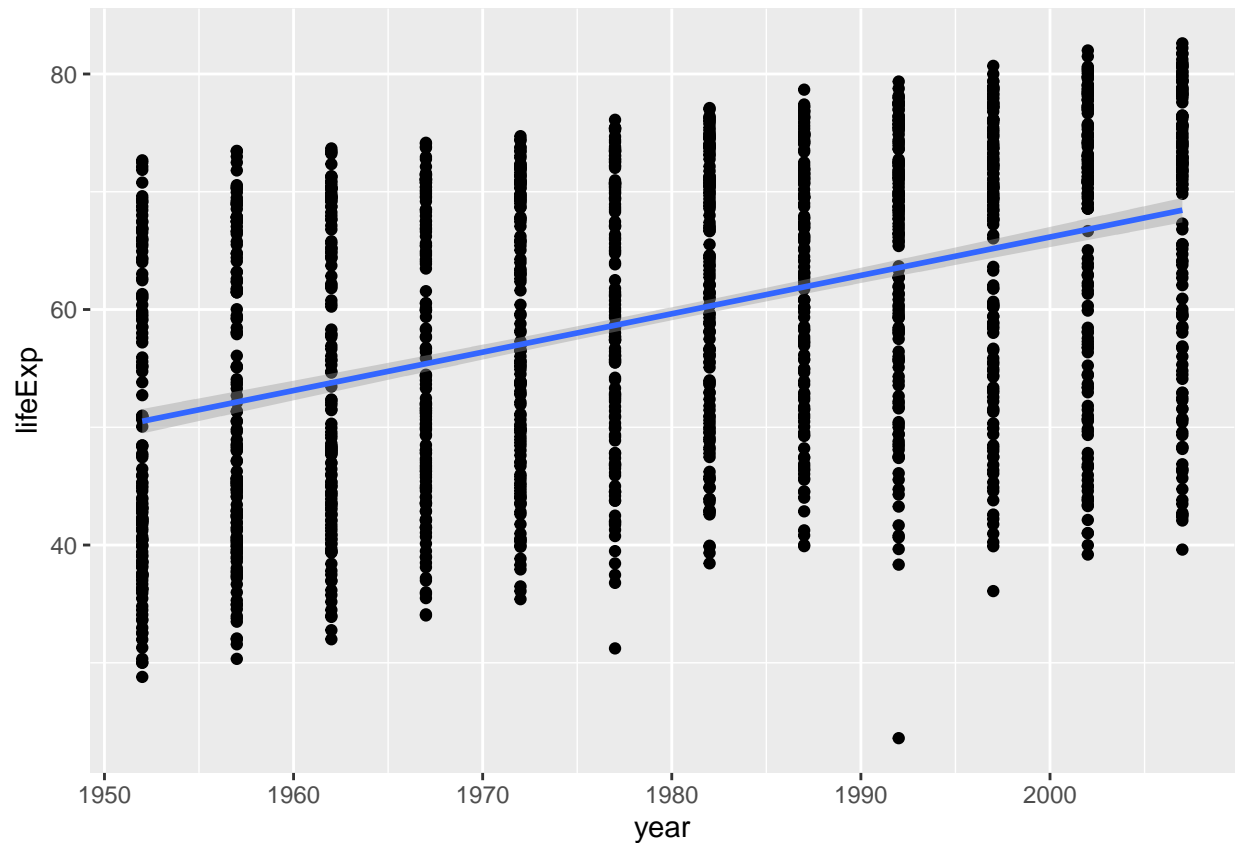
```
## 'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



En este caso `geom_smooth()` está usando `method = 'gam'`

```
vc1 + geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



Ejercicio 1

1. Hacer un gráfico de dispersión que tenga en el eje y `year` y en el eje x `lifeExp`, los puntos deben estar coloreados por la variable `continent`. Para este plot ajustá una recta de regresión para cada continente sin incluir las barras de error. Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un **caption** en la Figura con algún comentario de interés que describa el gráfico. El resto de los comentarios del gráfico se realizan en el texto.

```
install.packages("gapminder")
```

```
## Warning: package 'gapminder' is in use and will not be installed
```

```
library(gapminder)
data<-gapminder
ggplot(data, aes(lifeExp, year, colour = continent)) +
geom_point()+labs(x = "Esperanza de vida al momento de nacer", y = "Año") + geom_smooth(method = "lm", se
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

2. Omitir la capa de `geom_point()` del gráfico anterior. Las líneas aún aparecen aunque los puntos no. ¿Porqué sucede esto?

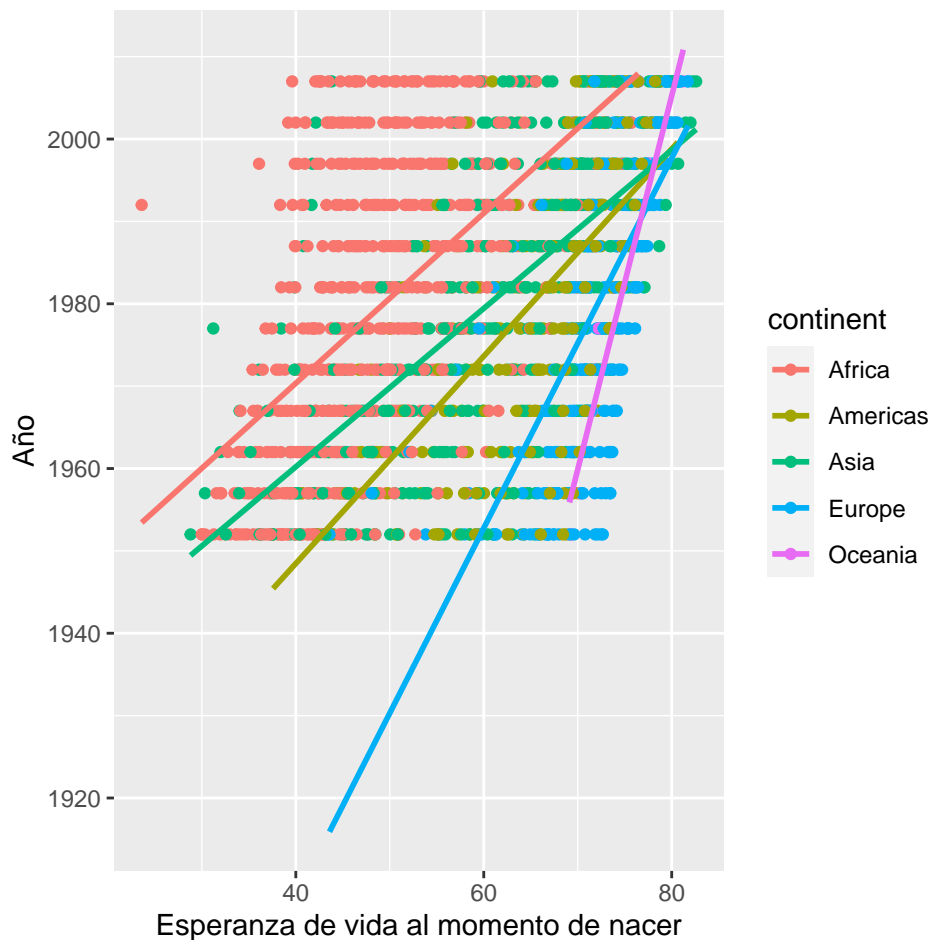


Figure 1: Figura 1: Gráfico de dispersión entre el año y la esperanza de vida al momento de nacer, por continente.

```
ggplot(data, aes(lifeExp, year, colour = continent)) +
  labs(x = "Esperanza de vida al momento de nacer", y = "Año") + geom_smooth(method = "lm", se = FALSE)
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

La figura 2 aún muestra las rectas de regresión dado que el argumento “+geom_smooth(method = “lm”, se = FALSE)” sigue estando presente en el chunk de código. Las líneas se trazan a partir de los puntos, pero la capa que los incluye fue removida del código.

3. El siguiente es un gráfico de dispersión entre `lifeExp` y `gdpPercap` coloreado por la variable `continent`. Usando como elemento estético color (`aes()`) nosotros podemos distinguir los distintos continentes usando diferentes colores de similar manera usando forma (`shape`).

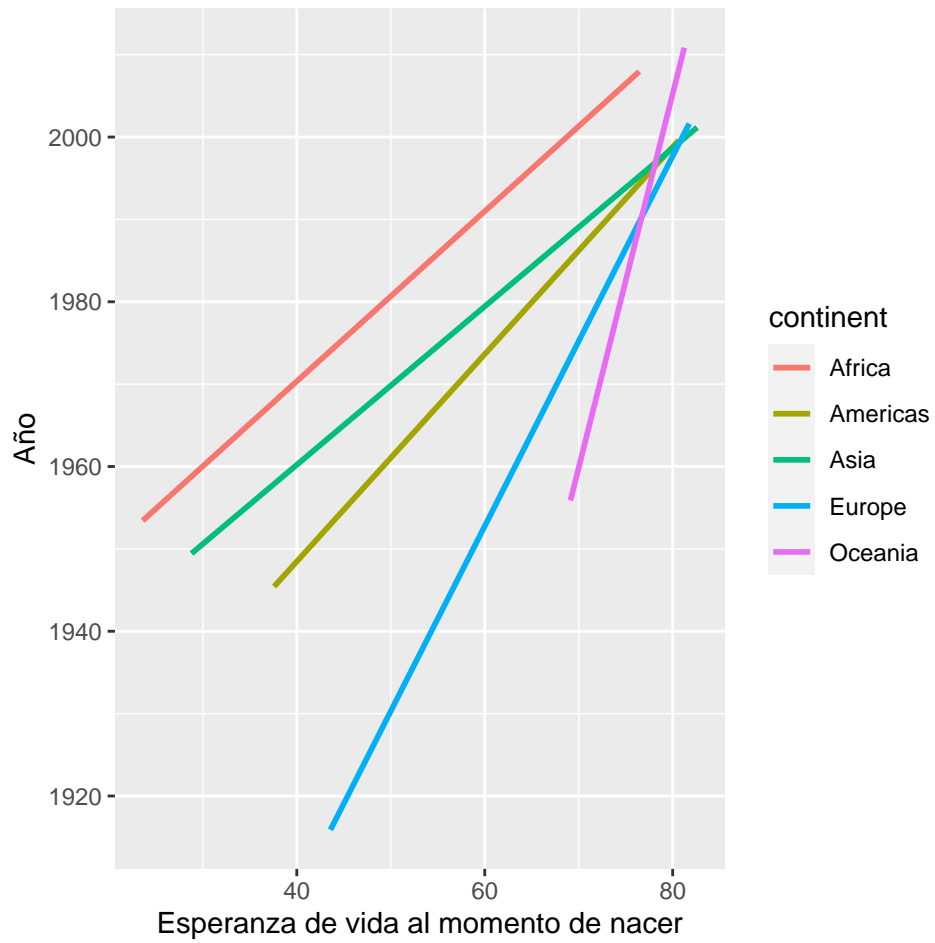
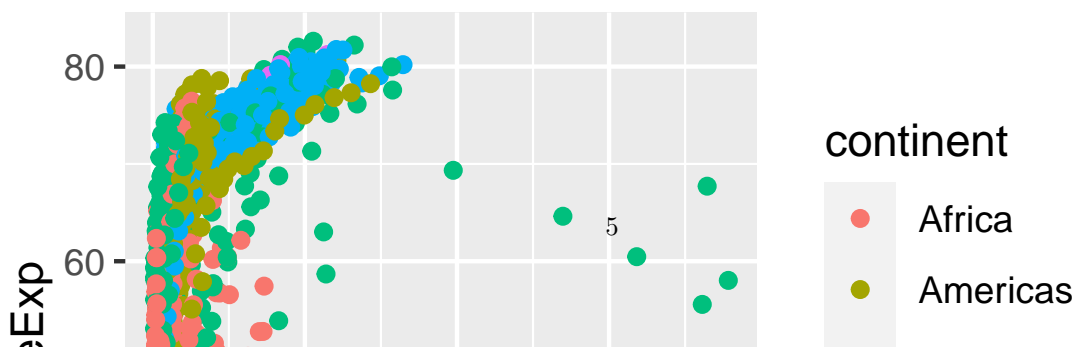


Figure 2: Figura 2: Gráfico de dispersión entre el año y la esperanza de vida al momento de nacer, por continente.



El gráfico anterior está sobrecargado, ¿de qué forma modificarías el gráfico para que sea más clara la comparación para los distintos continentes y por qué? Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Comentá alguna característica interesante que describa lo que aprendes viendo el gráfico.

```
ggplot(data, aes(gdpPercap,lifeExp,color =continent)) +labs(x = "PBI per cápita", y = "Esperanza de vida
```

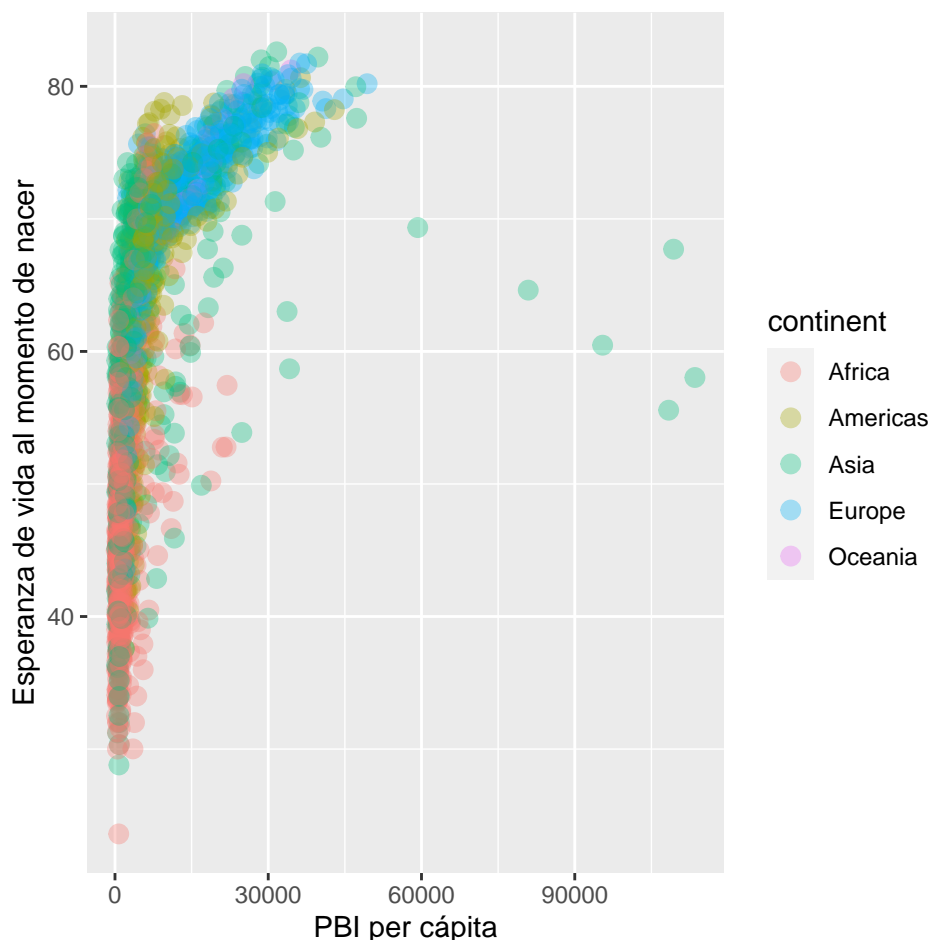


Figure 3: Figura 3: Gráfico de dispersión entre PBI per cápita y la esperanza de vida al momento de nacer, por continente.

A partir del gráfico se observa que continentes como Europa tienen una mayor esperanza de vida a mayor PBI per cápita, mientras que continentes como África tienen menor PBI per cápita, y en general, menor esperanza de vida. En el caso de América, podría plantearse un escenario similar al de Europa. En líneas generales, podría pensarse en una correlación positiva entre PBI per cápita y la esperanza de vida al momento de nacer; a mayor PBI p/c mayor esperanza de vida. Para Asia se observan algunas observaciones que no siguen esta línea (quizás datos atípicos), observaciones con un muy alto PBI p/c pero que aún así no tienen niveles altos de esperanza de vida.

4. Hacer un gráfico de líneas que tenga en el eje **x** **year** y en el eje **y** **gdpPercap** para cada continente en una misma ventana gráfica. En cada continente, el gráfico debe contener una línea para cada país a lo largo del tiempo (serie de tiempo de **gdpPercap**). Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un **caption** en la Figura con algún comentario de interés que describa el gráfico.

```
ggplot(data, aes(year, gdpPerCap)) + geom_line(aes(colour = country)) + theme(legend.position = "none") +
  labs(x = "Año", y = "PBI per cápita")
```

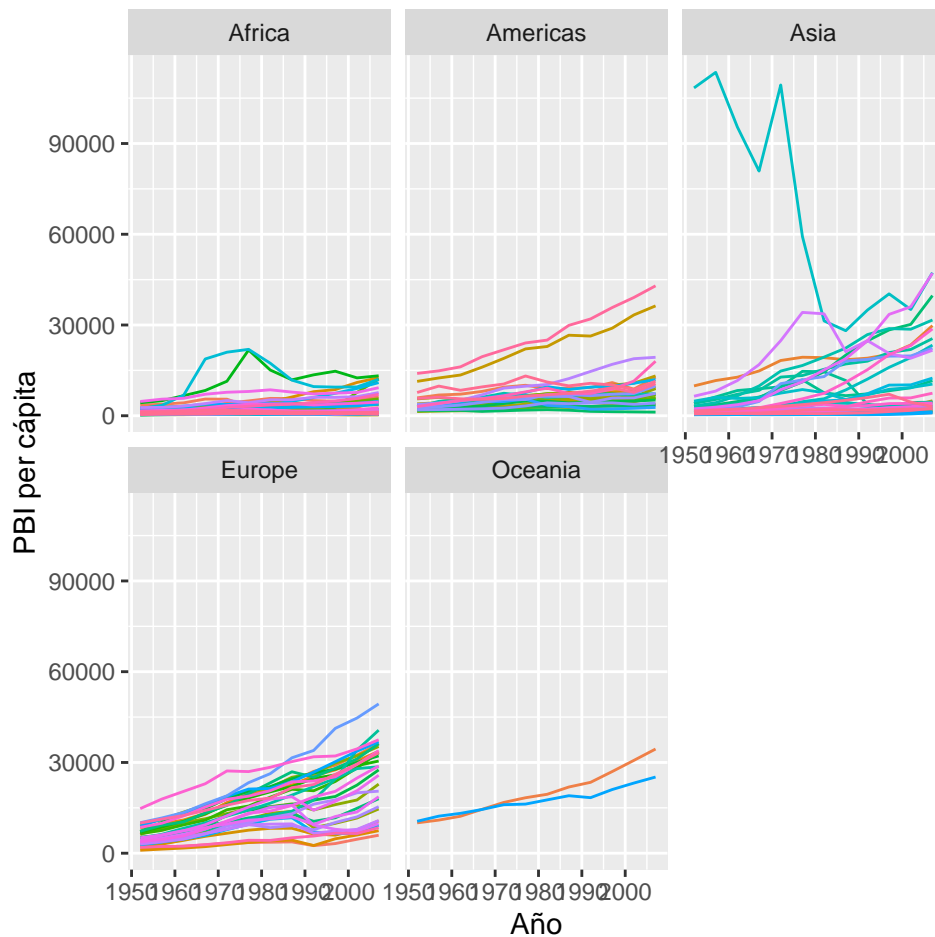


Figure 4: Figura 4: Evolución del PBI per cápita por continente, para cada país.

La representación por países en colores no me parece la más adecuada, dado que es sumamente difícil identificar cuál se corresponde con cuál. Además de que para distintos continentes se usan los mismos colores, aún siendo países distintos. Lo cual puede llevar a interpretaciones erróneas.

5. Usando los datos `gapminder` seleccione una visualización que describa algún aspecto de los datos que no exploramos. Comente algo interesante que se puede aprender de su gráfico.

```
ggplot(data, aes(lifeExp)) + geom_density(aes(fill=continent), alpha=1/3) + labs(x = "Esperanza de vida al
```

En este caso se grafica la distribución para la variable continua “Esperanza de vida”. Se observa que continentes como Europa y Oceanía siguen una distribución en cierta forma similar, con valores mayores por ejemplo que si la comparamos con África, que además parece tener una distribución mucho menos simétrica.

Ejercicio 2

1. Con los datos `mpg` que se encuentran disponible en `ggplot2` hacer un gráfico de barras para la variables `drv` con las siguientes características:

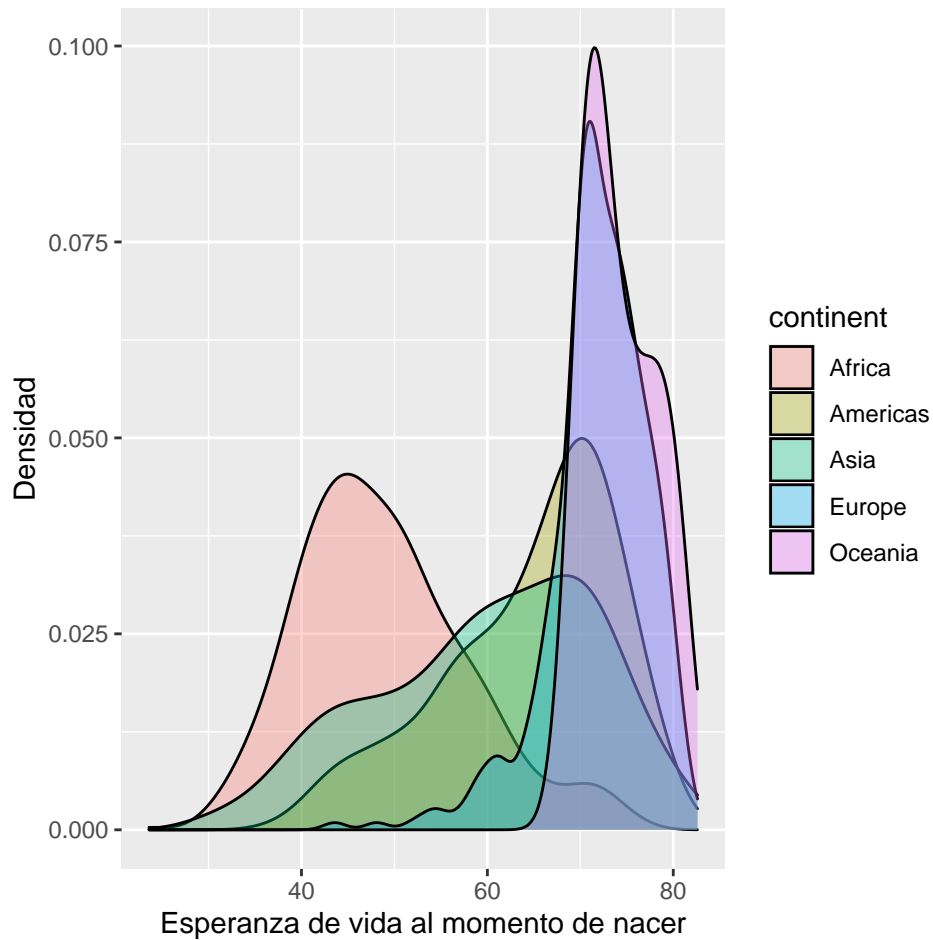


Figure 5: Figura 5: Evolución del PBI per cápita por continente, para cada país.

- Las barras tienen que estar coloreadas por `drv`
- Incluir usando `labs()` el nombre de los ejes y título informativo.
- Usá la paleta de colores `Dark2`, mirá la ayuda de `scale_colour_brewer()`.

```
data2<-mpg
data2$drv <- factor(data2$drv,
                    levels = c("f","r",4),
                    labels = c("Front-wheel drive", "Rear wheel drive", "4 wheel drive"))
ggplot(data2)+ geom_bar(mapping = aes(x = drv, fill = drv)) +labs(title="Type of drive train frequency")
```

2. Usando como base el gráfico anterior:

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

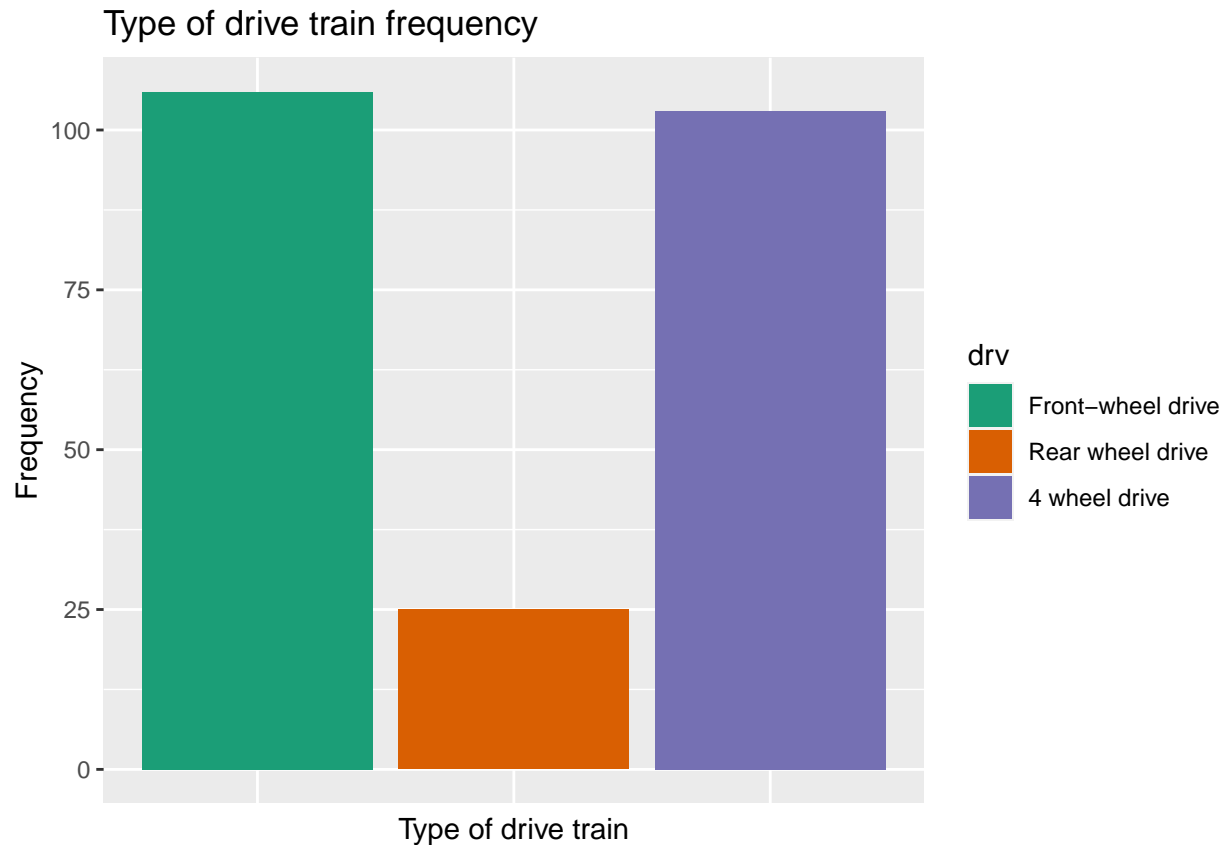


Figure 6: Figura 6: Frecuencia de los tipos de trenes

```
ggplot(data2, aes(x = drv, y = prop.table(stat(count)), fill = drv, label = scales::percent(prop.table(s
  position = position_dodge(.9),
  vjust = -0.5, size = 3) + scale_y_continuous(labels = scales::percent) + labs(x = 'Drv', y = 'Frequency'))
```

Ejercicio 3

Los datos que vamos a utilizar en este ejercicio están disponibles en el catálogo de datos abiertos uruguay <https://catalogodatos.gub.uy>. Los datos que seleccioné son sobre las emisiones de dióxido de carbono (CO2) correspondientes a las actividades de quema de los combustibles en las industrias de la energía y los sectores de consumo. Se incluyen también emisiones de CO2 provenientes de la quema de biomasa y de bunkers internacionales, las cuales se presentan como partidas informativas ya que no se consideran en los totales. En el siguiente link se encontrarán los datos y los meta datos con información que describe la base de datos <https://catalogodatos.gub.uy/dataset/miem-emisiones-de-co2-por-sector>.

Por simplicidad te damos los datos reestructurados (veremos como se hace más adelante en el curso), el archivo se llama datos_emisión.csv, contiene tres columnas AÑO, fuente y emisión.

1. Leer los datos usando el paquete `readr` y la función `read_csv`, guardarlos en un objeto llamado `datos`.

```
library(readr)
```

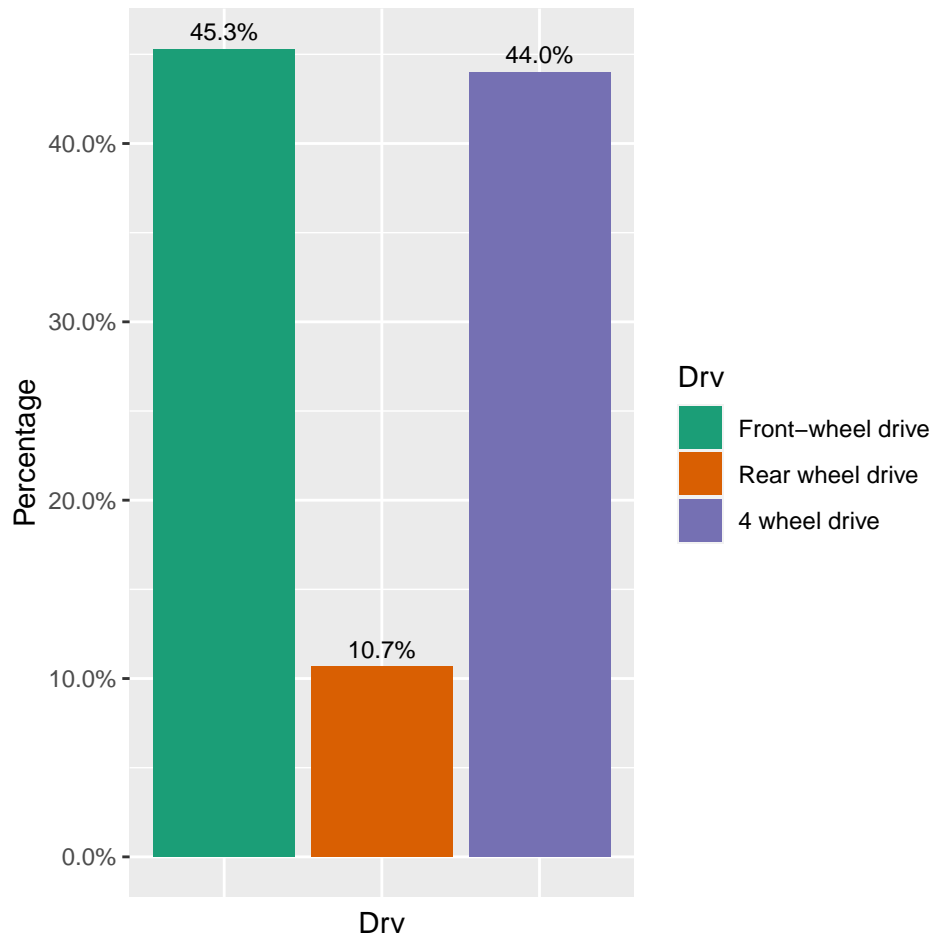


Figure 7: Figura 7: Frecuencia relativa de los tipos de trenes

Warning: package 'readr' was built under R version 4.0.5

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

```
##
## -- Column specification -----
## cols(
##   AÑO = col_double(),
##   fuente = col_character(),
##   emission = col_double()
## )
```

- Usando las funciones de la librería **dplyr** obtenga qué fuentes tienen la emisión máxima. Recuerde que TOTAL debería ser excluido para esta respuesta así como los subtotales.

```
library(dplyr)
```

```
##
## 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
```

```
datos %>% filter(fuente!="TOTAL"&fuente!="S_C"&fuente!="I_E")%>%group_by(fuente)%>%summarise(maximo=max
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
## # A tibble: 1 x 1
##   fuente
##   <chr>
## 1 Q_B
```

3. ¿En qué año se dió la emisión máxima para la fuente que respondió en la pregunta anterior?

```
datos %>% filter(fuente=="Q_B")%>%group_by(AÑO)%>%summarise(maximo=max(emision,na.rm=TRUE) )%>% arrange
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
## # A tibble: 1 x 1
##   AÑO
##   <dbl>
## 1 2017
```

4. Usando las funciones de la librería `dplyr` obtenga las 5 fuentes, sin incluir TOTAL ni subtotales, que tienen un valor medio de emisión a lo largo de todos los años más grandes.

```
top5<-datos %>% filter(fuente!="TOTAL"&fuente!="S_C"&fuente!="I_E")%>%group_by(fuente)%>%summarise(medi
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

5. Usando `ggplot2` realice un gráfico de las emisiones a lo largo de los años para cada fuente. Utilice dos elementos geométricos, puntos y líneas. Seleccione para dibujar solamente las 5 fuentes que a lo largo de los años tienen una emisión media mayor que el resto (respuesta de la pregunta 5). Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluya un `caption` en la figura con algún comentario de interés que describa el gráfico.

```
datos %>% filter(fuente%in%top5)%>%
  ggplot(aes(x=AÑO, y=emision))+
  geom_point(aes(color=fuente))+geom_line(aes(color=fuente))+ theme(aspect.ratio = 1)+labs(x = "Año", y =
```

6. Relpique el siguiente gráfico usando `ggplot2`. Incluya un `caption` en la figura con algún comentario de interés que describa el gráfico.



Figure 8: Figura 8: Evolución de la emisión de CO2 para las 5 fuentes con mayor promedio

```
datos%>%ggplot(aes(x=reorder(fuente,-emision,median,na.rm=TRUE),y=emision))+geom_boxplot()+labs(x="Fuen
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```

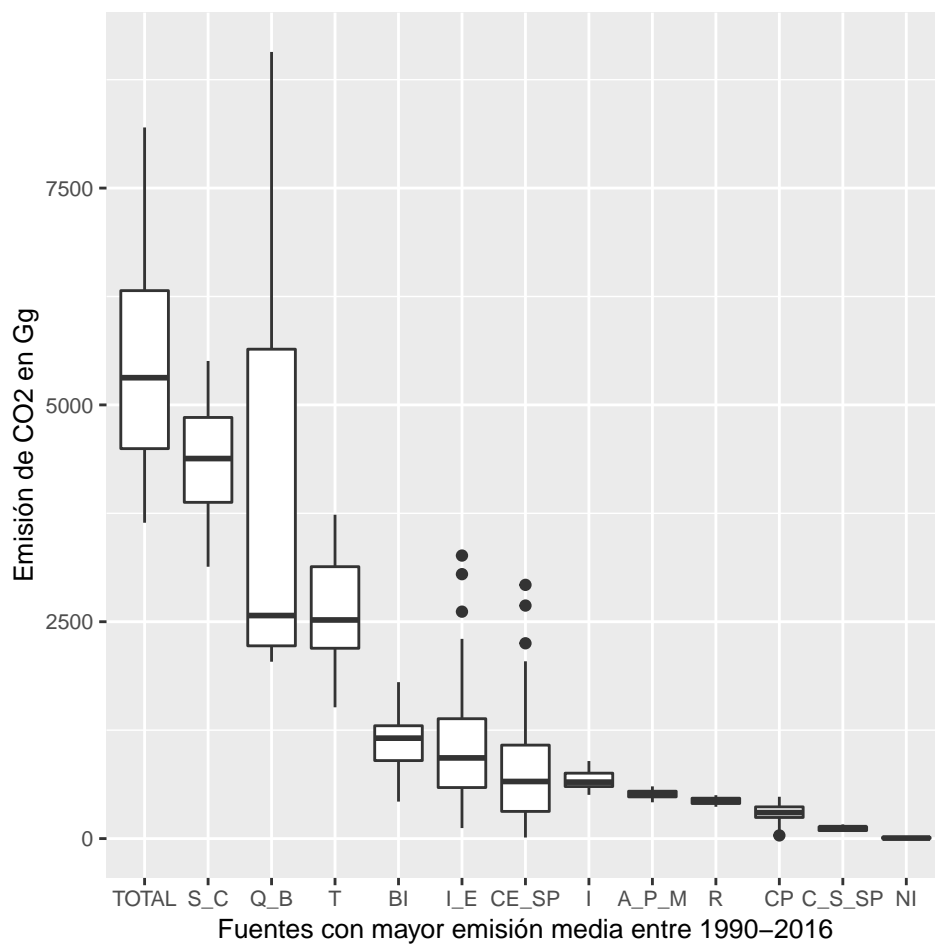
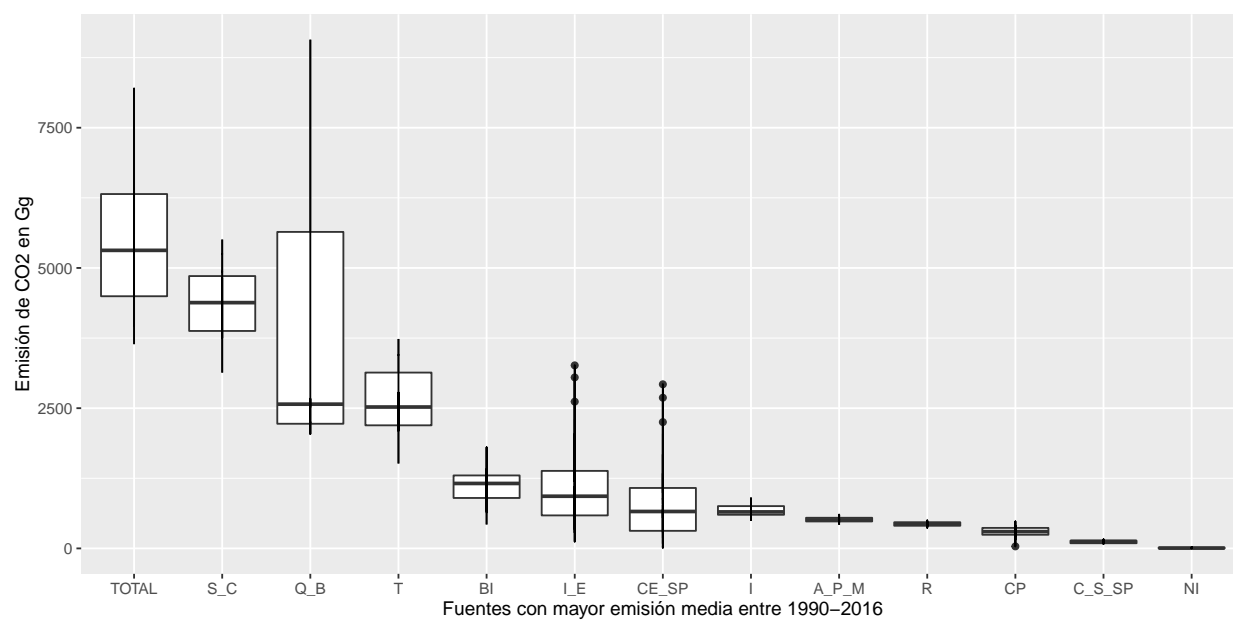


Figure 9: Figura 9: Diagrama de caja para la emisión de CO2 por tipo de fuente



7. Usando la librería `ggplot2` y `ggpmisc` replique el siguiente gráfico de las emisiones totales entre 1990 y 2016. Los puntos rojos indican los máximos locales o picos de emisión de CO₂ en Gg. Use `library(help = ggpmisc)` para ver todas las funciones de la librería `ggpmisc` e identificar cual o cuales necesita para replicar el gráfico. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.

```
library(ggpmisc)
```

```
## Warning: package 'ggpmisc' was built under R version 4.0.5
```

```
##
```

```
## Attaching package: 'ggpmisc'
```

```
## The following object is masked from 'package:ggplot2':
```

```
##
```

```
## annotate
```

```
datos %>% filter(fuente=="TOTAL") %>% group_by(AÑO) %>% ggplot(aes(x=AÑO,y=emision)) + geom_line() + geom_point()
```

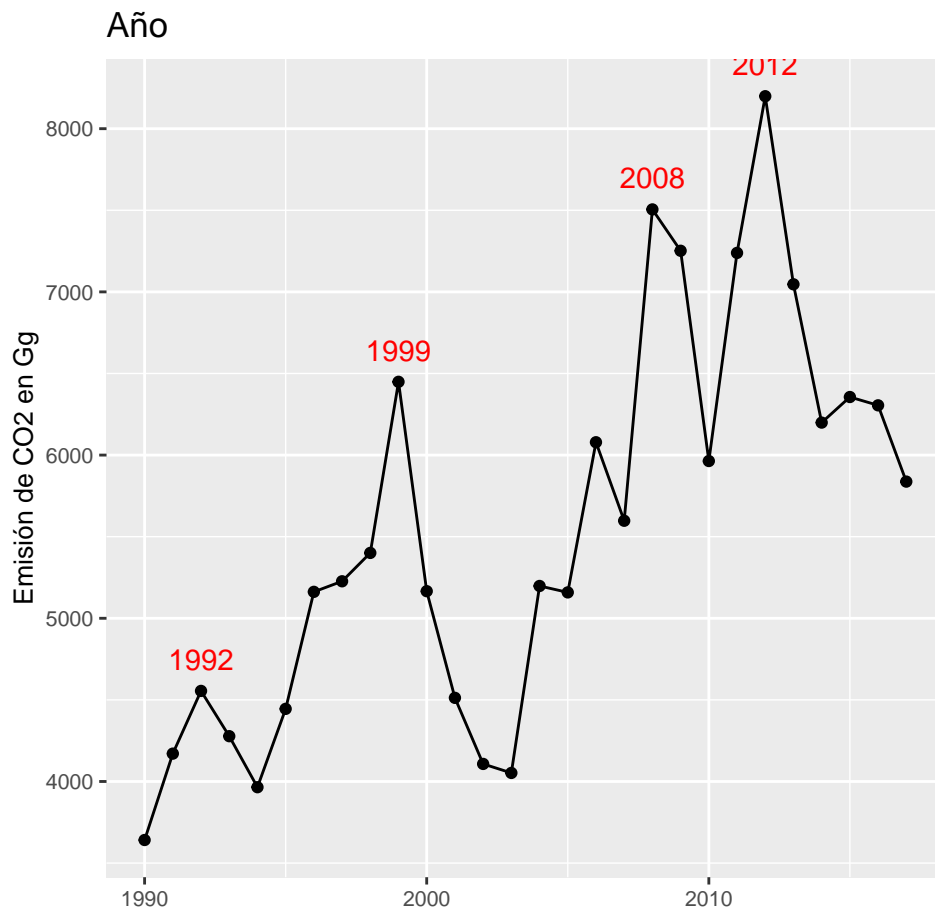
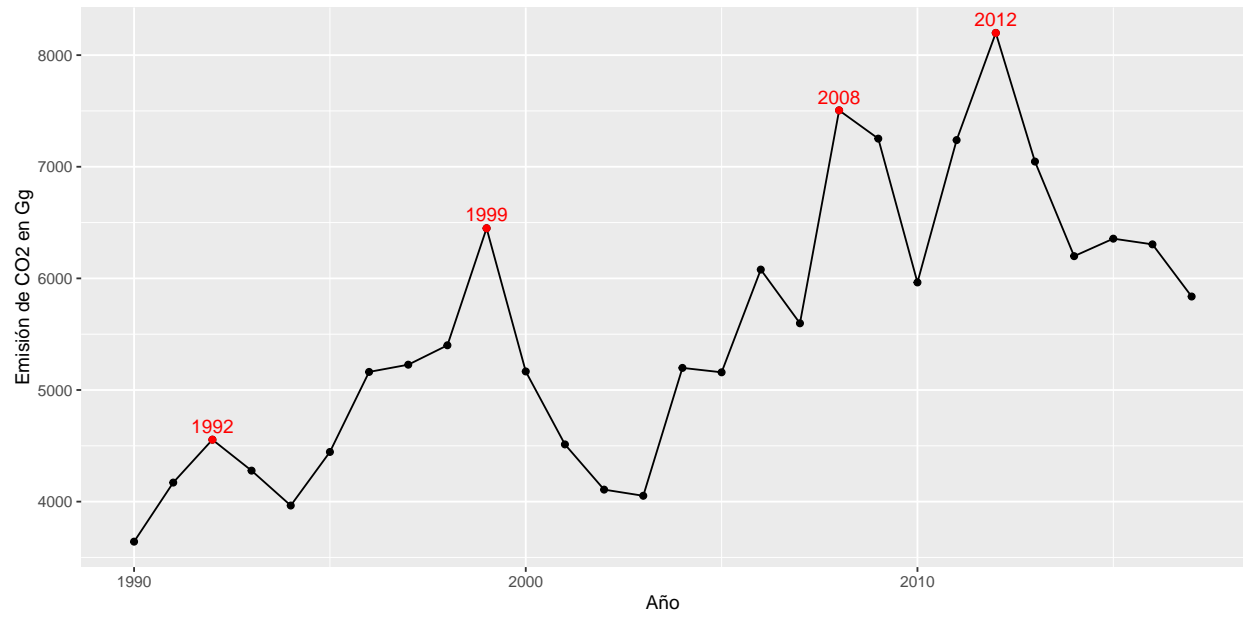


Figure 10: Figura 9: Evolución de la emisión total de CO en Gg



Ejercicio 4

Los datos que vamos a utilizar en este ejercicio son una muestra de datos a nivel nacional sobre abandono escolar en los años 2016.

Table 1: Variables en **muestra.csv**

Variable	Descripción
documento	Cédula de Identidad del alumno
nro_doc_centro_educ	Liceo que concurre el alumno en 2016
nombre_departamento	Nombre del Departamento del centro educativo
grupo_desc	Grupo del alumno en 2016
coberturaT	Cobertura en el primer semestre de 2016
Centro_Grupo	Liceo y grupo del alumno en 2016
cl	Cluster - contexto sociocultural del liceo en 1016
Grado_2016_UE	Grado del alumno en el 2016 según UE
Grado2013	Grado del alumno en 2013 según CRM
Grado2014	Grado del alumno en 2014 según CRM
Grado2015	Grado del alumno en 2015 según CRM
Grado_2016	Grado del alumno en 2016 según CRM
Sexo	Sexo del alumno
Fecha.nacimiento	Fecha de nacimiento del alumno
Grupo_UE_2017	Grupo del alumno en 2017
inasistencias	cantidad de inasistencias en el primer semestre de 2016
asistencias	cantidad de asistencias en el primer semestre de 2016

En el Cuadro 1 se presentan las variables en el conjunto de datos **muestra.csv**.

Este ejercicio tiene como objetivo que realice tres preguntas de interés que le surgen como parte del análisis exploratorio de datos utilizando todo lo aprendido en el curso.

Debe plantear 3 preguntas orientadoras y visualizaciones apropiadas para responderlas. La exploración deberá contener las preguntas a responder sus respuestas con el correspondiente resumen de información o visualización. Incluya en su exploración el análisis de la variabilidad tanto de variables cuantitativas como cualitativas y covariaciones entre las mismas. Recuerde que en las visualizaciones, las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un **caption** en la figura con algún comentario de interés que describa el gráfico y lo que ve en el mismo.

```
library(readr)
data<-read_csv("muestra.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- Column specification -----
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   documento = col_double(),
##   nro_doc_centro_educ = col_double(),
##   coberturaT = col_double(),
##   cl = col_double(),
##   Grado_2016_UE = col_double(),
```



```
## 'Fecha nacimiento' = col_date(format = ""),
## inasistencias = col_double(),
## asistencias = col_double(),
## Abandono = col_double()
## )
## i Use 'spec()' for the full column specifications.
```

```
summary(data)
```

```
##      X1      documento      nro_doc_centro_educ      nombre_departamento
## Min.   : 1      Min.   :1.822e+05      Min.   : 1201070      Length:4023
## 1st Qu.:1006      1st Qu.:5.347e+07      1st Qu.: 1204404      Class :character
## Median :2012      Median :5.407e+07      Median :12101024      Mode  :character
## Mean   :2012      Mean   :5.480e+07      Mean   : 7727218
## 3rd Qu.:3018      3rd Qu.:5.504e+07      3rd Qu.:12121205
## Max.   :4023      Max.   :1.128e+09      Max.   :12191908
## nombre_localidad      grupo_desc      coberturaT      Centro_Grupo
## Length:4023      Length:4023      Min.   : 0.00      Length:4023
## Class :character      Class :character      1st Qu.:47.00      Class :character
## Mode  :character      Mode  :character      Median :53.00      Mode  :character
##                               Mean   :48.75
##                               3rd Qu.:56.00
##                               Max.   :57.00
##      cl      Grado_2016_UE      Grado2013      Grado2014
## Min.   :1.000      Min.   :1      Length:4023      Length:4023
## 1st Qu.:2.000      1st Qu.:1      Class :character      Class :character
## Median :3.000      Median :1      Mode  :character      Mode  :character
## Mean   :3.206      Mean   :1
## 3rd Qu.:5.000      3rd Qu.:1
## Max.   :5.000      Max.   :1
##      Grado2015      Grado2016      Sexo      Fecha nacimiento
## Length:4023      Length:4023      Length:4023      Min.   :1980-06-11
## Class :character      Class :character      Class :character      1st Qu.:2002-10-25
## Mode  :character      Mode  :character      Mode  :character      Median :2003-07-11
##                               Mean   :2003-03-31
##                               3rd Qu.:2003-11-27
##                               Max.   :2005-11-07
##      Grupo_UE_2017      inasistencias      asistencias      Abandono
## Length:4023      Min.   : 0.000      Min.   : 0.00      Min.   :0.00000
## Class :character      1st Qu.: 1.000      1st Qu.:40.00      1st Qu.:0.00000
## Mode  :character      Median : 2.000      Median :48.00      Median :0.00000
##                               Mean   : 4.915      Mean   :43.83      Mean   :0.06488
##                               3rd Qu.: 6.000      3rd Qu.:53.00      3rd Qu.:0.00000
##                               Max.   :47.000      Max.   :57.00      Max.   :1.00000
```

Mirando la descripción de las variables y el resumen de cada una de ellas, podemos preguntarnos: cómo varía la cantidad de inasistencias según el departamento en el que reside el estudiante?

A continuación se presenta un gráfico de barras de las mismas según el lugar de residencia.

```
data$Sexo <- factor(data$Sexo,
                    levels = c("F", "M"),
                    labels = c("Femenino", "Masculino"))
inasistencias_sexo <- data %>% group_by(Sexo) %>% summarise(media = mean(inasistencias, na.rm = TRUE)) %>% arrange(desc(media))
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
ggplot(inasistencias_sexo, aes(x = Sexo, y = media, fill=Sexo)) + geom_bar(stat="identity") + geom_text(aes(
  labs(title="Cantidad promedio de inasistencias en el primer semestre de 2016 por sexo", x = "Sexo", y =
    panel.border = element_blank(), panel.background = element_blank(), axis.title.x = element_text
```

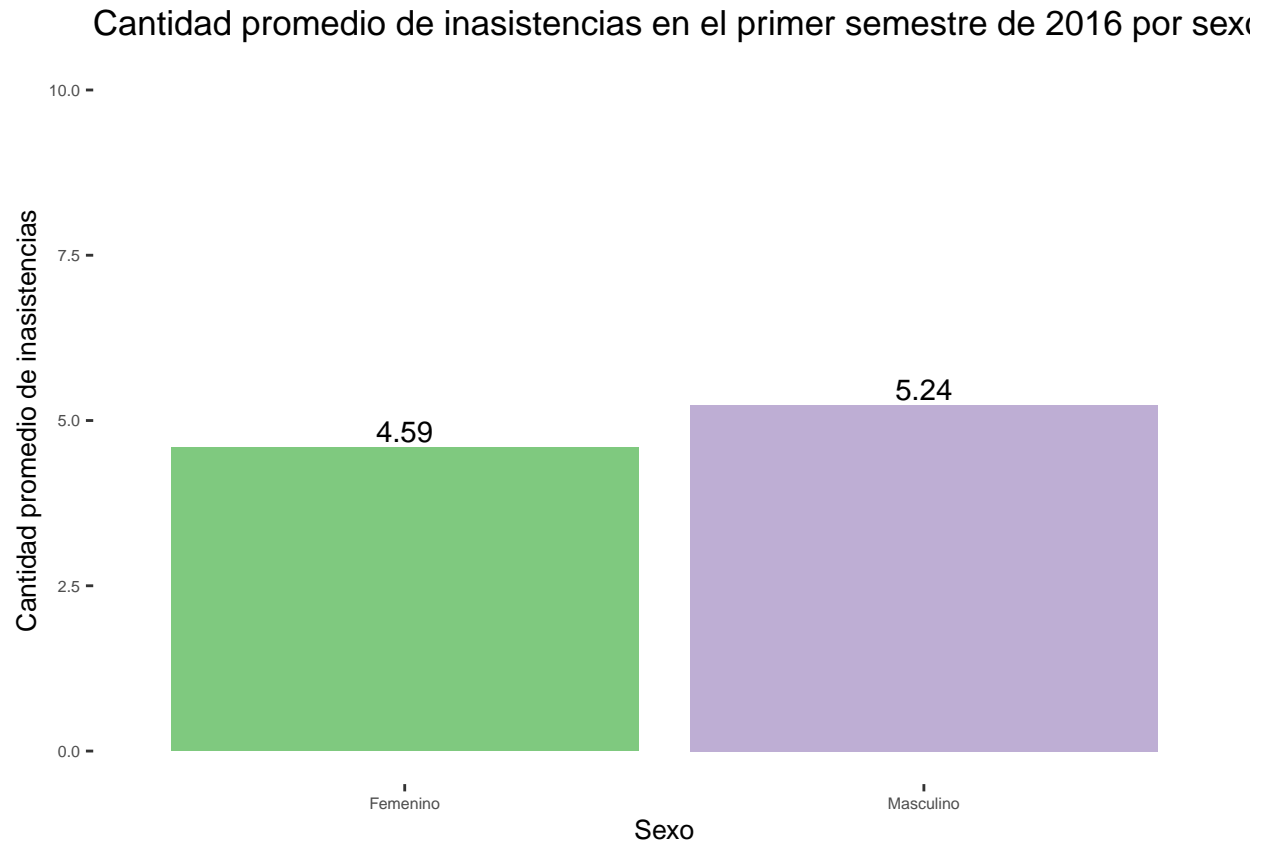


Figure 11: Figura 10: Cantidad promedio de inasistencias por sexo

Se observa que las personas de sexo masculino tienen, en promedio, mayor número de inasistencias que las de sexo femenino. Si bien no se puede concluir nada a partir del gráfico, uno podría preguntarse a qué está relacionado esto. Podría compararse la edad de los encuestados según el sexo, como también el lugar de residencia o el contexto socioeconómico.

```
masinasistencias<-data%>%group_by(nombre_departamento)%>%summarise(media=mean(inasistencias,na.rm=TRUE)
  labs(title="Cantidad promedio de inasistencias en el primer semestre de 2016 por lugar de residencia"
    panel.border = element_blank(), panel.background = element_blank(), axis.title.x = element_text
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

La figura 11 muestra que Montevideo es el departamento con mayor cantidad promedio de inasistencias, seguido por Río Negro y Salto. A partir de esto, cabe preguntarse si incide el hecho de que el estudiante resida en una zona metropolitana en la cantidad de inasistencias? De ser así uno tendería a pensar que Canelones podría estar entre los departamentos con más cantidad de inasistencias promedio, sin embargo se encuentra más abajo en la lista.

```

inasistencias_cl<-data %>% group_by(cl)%>%summarise(media=mean(inasistencias,na.rm=TRUE) )%>% arrange(m

## 'summarise()' ungrouping output (override with '.groups' argument)

ggplot(inasistencias_cl,aes(x = cl, y = media, fill=as.factor(cl))) +geom_bar(stat="identity")+geom_text
  labs(title="Cantidad promedio de inasistencias en el primer semestre de 2016 por nivel sociocultural"
        panel.background = element_blank(), axis.title.x = element_text(size=10,hjust = 0.5),axis.title

```

Cantidad promedio de inasistencias en el primer semestre de 2016 por nive

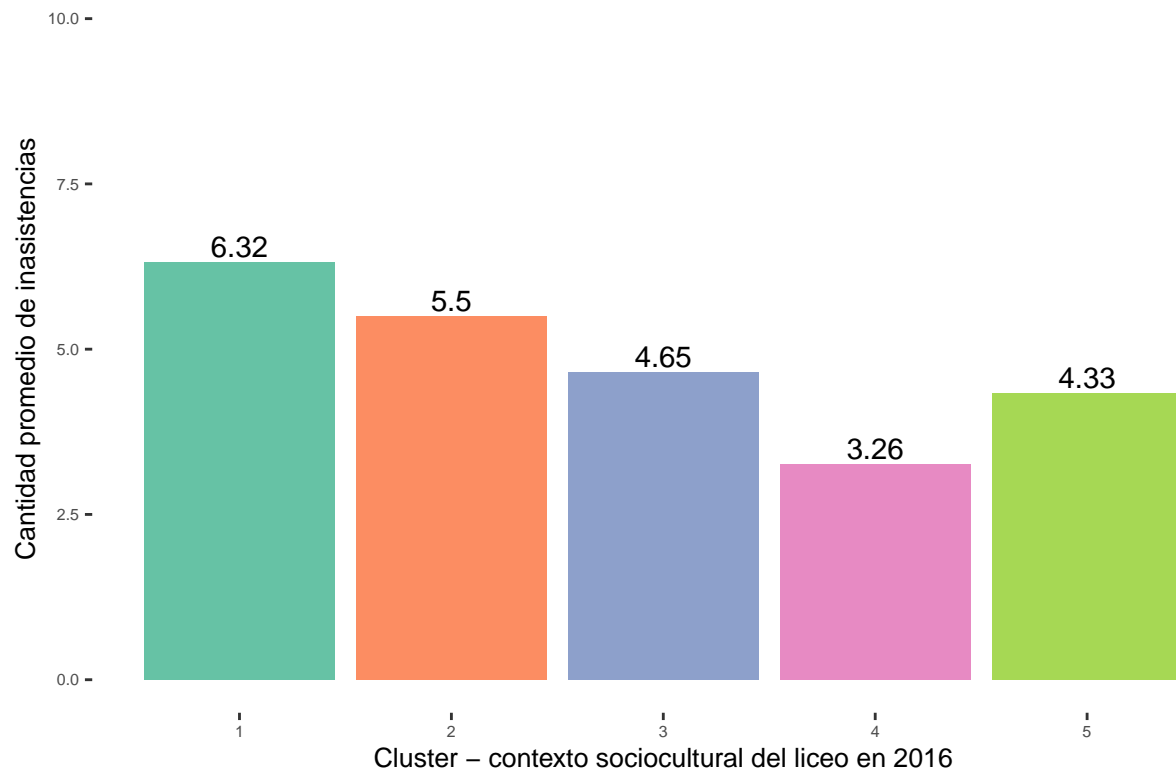


Figure 12: Figura 12: Cantidad promedio de inasistencias por contexto sociocultural del liceo en 2016.

No me parece muy correcto interpretar el gráfico ya que no hay una descripción de la variable cl que permita identificar cada uno de los contextos socioculturales de los liceos. Sin embargo, podríamos preguntarnos si existe relación alguna entre la cantidad de inasistencias de los estudiantes y el nivel del liceo. Por ejemplo, varios estudios y encuestas indican que el nivel de deserción de los estudiantes de secundaria se exagera en contextos más carenciados, por lo tanto podría ocurrir lo mismo con el nivel de inasistencia.