Tarea individual 2 - Martín Olivera - CI. 4.845.488-3

Entregar el Viernes 26 de Abril

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 travéz 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?

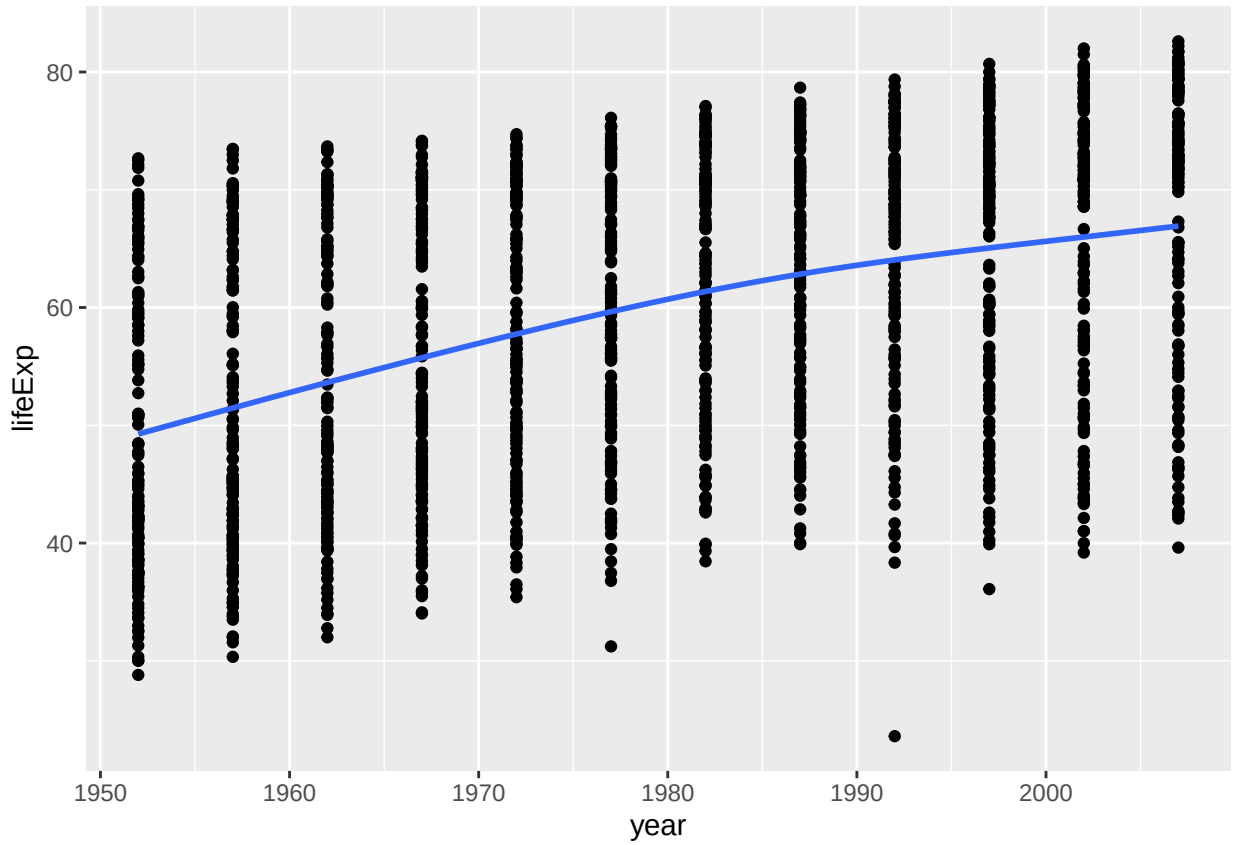
Para agregar una linea de regresión o una curva tinenes que agregar una capa a tu gráfico `geom_smoth`. Probablemente dos de los argumentos más útiles de `geom_smoth` son:

- `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)
```

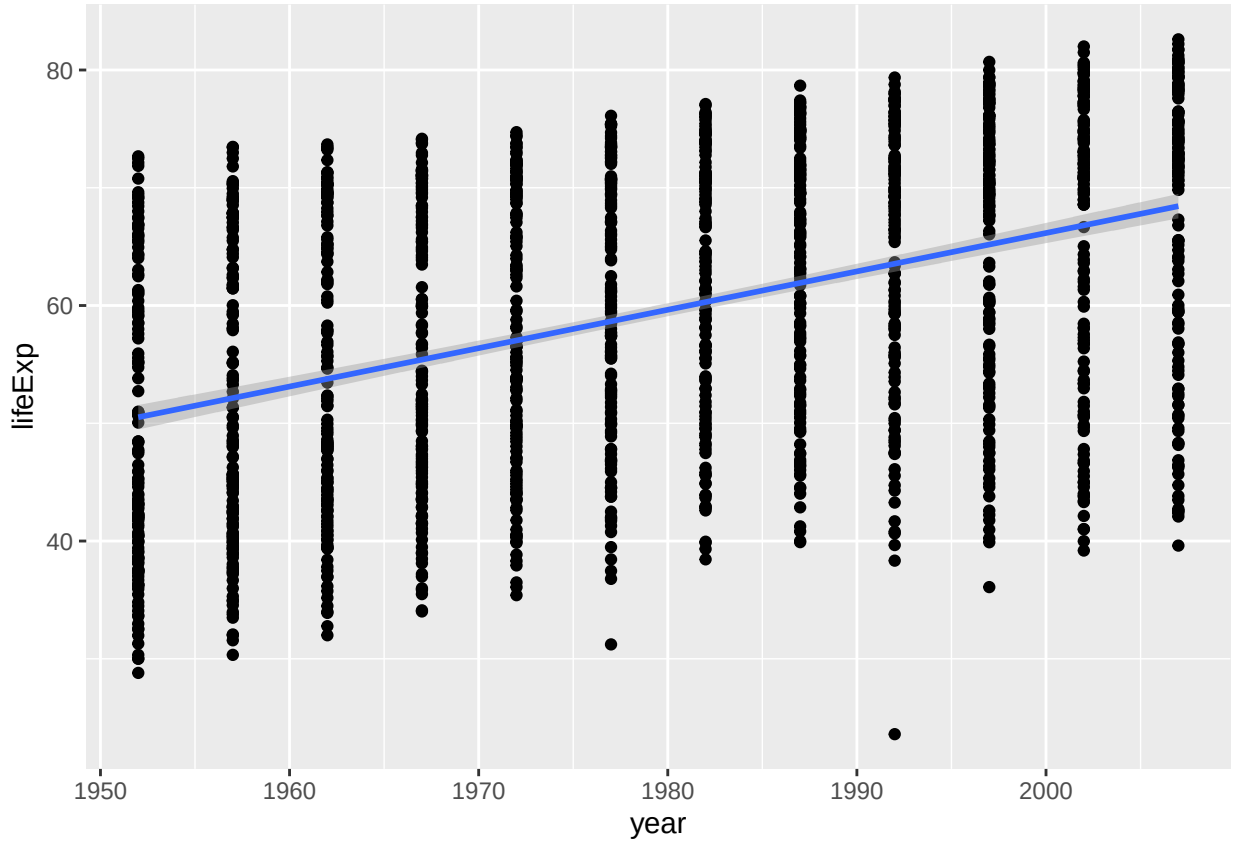
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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 graáfico se realizan en el texto.

```
ggplot(gapminder, aes(lifeExp, year, colour = continent)) +
    geom_point() + theme(aspect.ratio = 1) + geom_smooth(method = "lm",
    se = FALSE)
```

```
labs(x = "Life expectancy", y = "Year")
```

```
## $x
## [1] "Life expectancy"
##
## $y
## [1] "Year"
##
## attr(,"class")
## [1] "labels"
```

La Figura 1 muestra evidencia a favor de que la esperanza de vida en países africanos es inferior a la que se observa en europa y sensiblemente mejor a la observada en Asia y las Américas. Parece existir una fuerte relación entre los niveles de desarrollo de los continentes y sus niveles de esperanza de vida.

2. Omitir la capa de `geom_point()` del gráfico anterior. Las lineas aún aparecen aunque los puntos no.

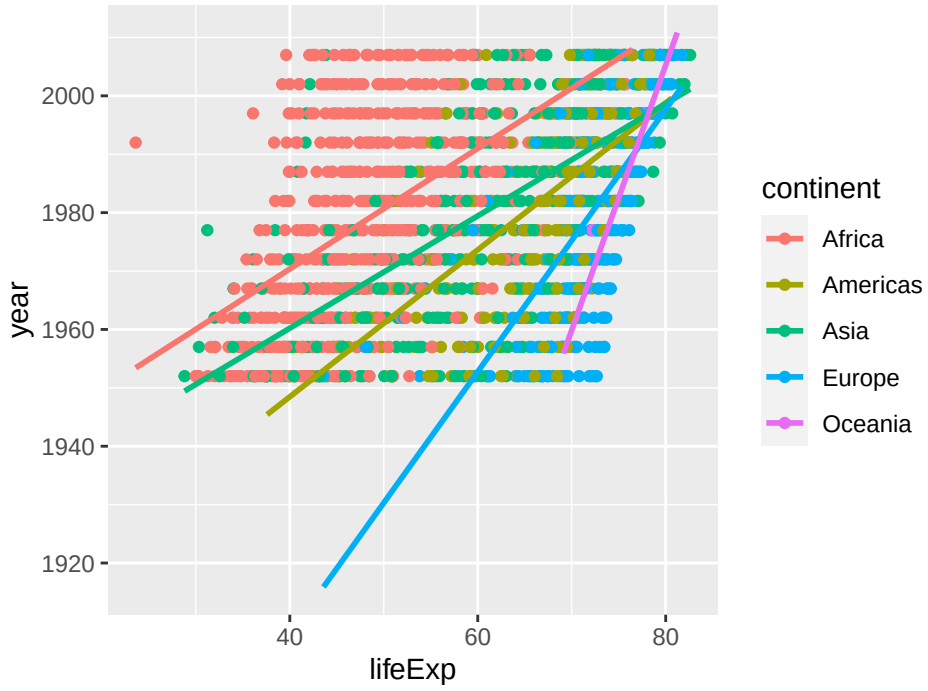


Figure 1: Dispersión de la esperanza de vida por año, según continente.

¿Porqué sucede esto?

```
ggplot(gapminder, aes(lifeExp, year, colour = continent)) +
    theme(aspect.ratio = 1) + geom_smooth(method = "lm",
    se = FALSE)
```

```
labs(x = "Life expectancy", y = "Year")
```

```
## $x
## [1] "Life expectancy"
##
## $y
## [1] "Year"
##
## attr(,"class")
## [1] "labels"
```

Al quitar el geom_point omito que se imprima la nube de puntos asociada a la relación entre la variable year y lifeExp, sin embargo, la relación aún existe y bajo la capa geom_smooth se grafica solo la recta de regresión asociada a la relación entre las variables, nuevamente, discriminando por continentes.

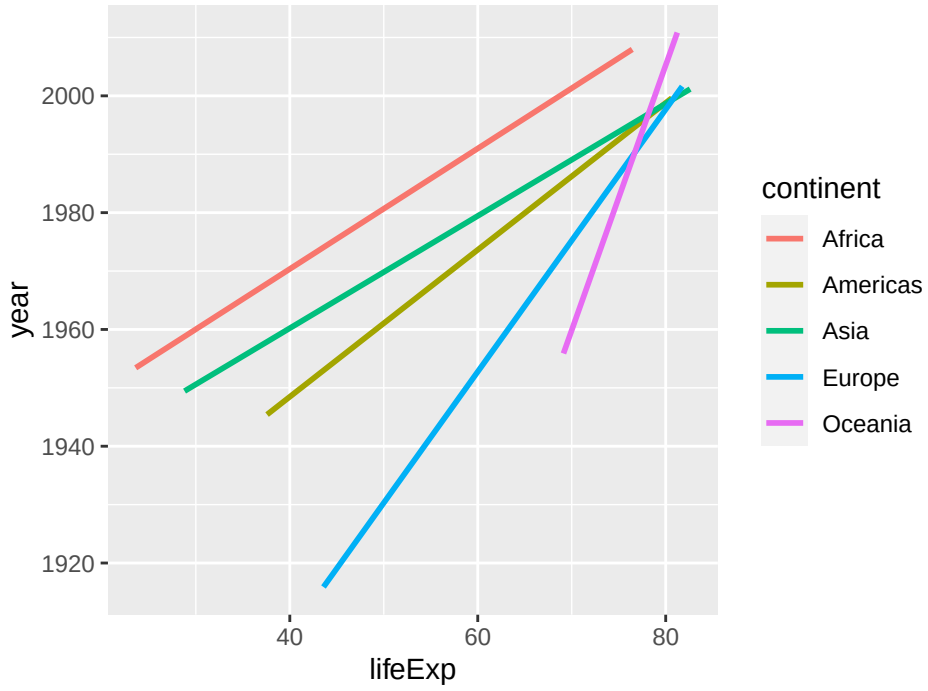
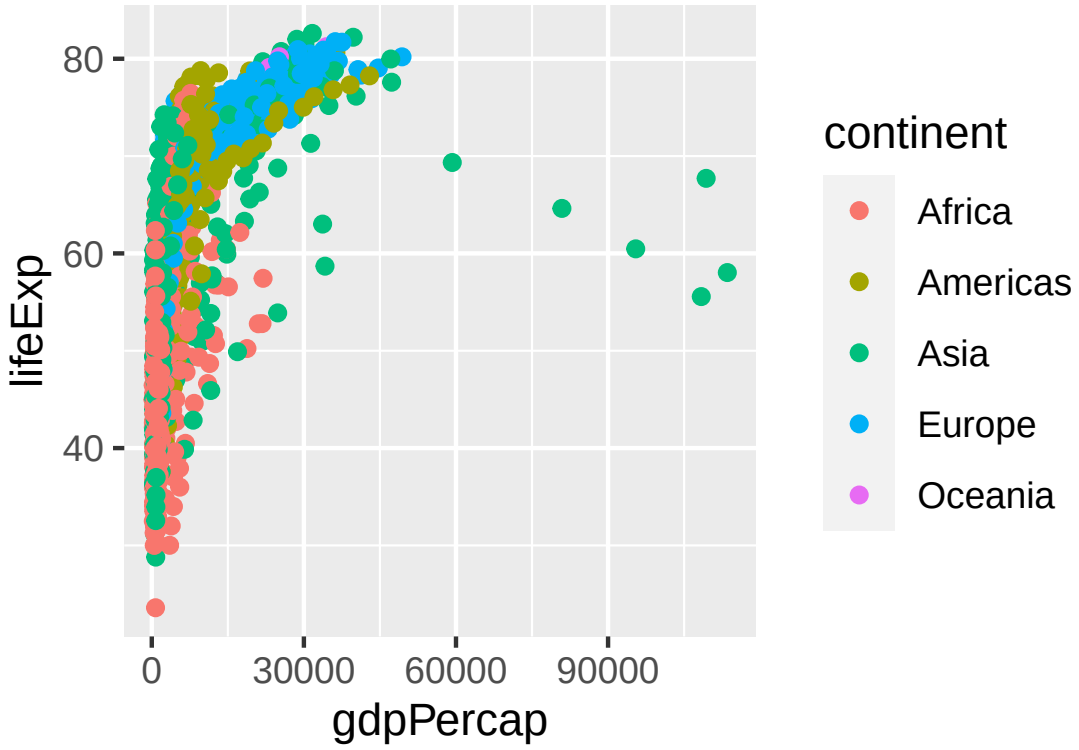


Figure 2: Esperanza de vida por año, según continente.

Comentario: Correcto! En estos ejercicios me equívoque en una pregunta del foro respecto si los ejes están bien especificados. Matemáticamente esta bien, pero se analizar al revez (y el análisis que hiciste es correcto), es decir en el eje y se gráfica la esperanza de vida y en el eje x los años, en el entendido que la variable a explicar es la esperanza de vida.

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`).


80
60
40
lifeExp
0
30000
60000
90000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania

El gráfico anterior está sobrecargado, ¿de que forma modificarías el gráfico para que sea más clara la comparación para los distintos continentes y porqué? 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(gapminder, aes(gdpPercap, lifeExp, colour = continent)) +
    geom_point(alpha = 1/3) + theme(aspect.ratio = 1)
```

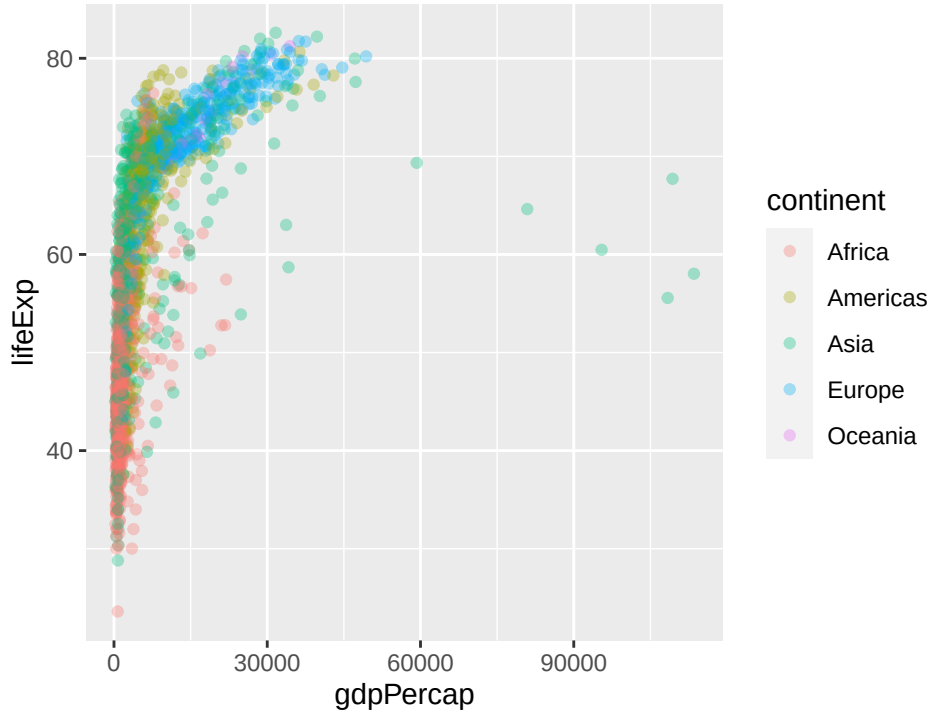


Figure 3: Diagrama de dispersión entre el producto interno bruto por habitante y la esperanza de vida para todo el período, por continente.

```
labs(x = "Gross Domestic Product, per capita", y = "Life expectancy")
```

```
## $x
## [1] "Gross Domestic Product, per capita"
##
## $y
## [1] "Life expectancy"
##
## attr(,"class")
## [1] "labels"
```

Al graficar las dos variables para todo el período, cada país registra un punto en el plano para cada año, sobrecargandolo, entonces con la opción alpha = 1/3 en geom_point, es posible ilustrar cada punto como

1/3 de la "saturación" del color para poder visualizar de mejor forma la la masa de puntos que se ilustra en la figura, denotando mayor cantidad de puntos cuanto más nítido sea el color. Respecto a lo que se puede visualizar de la Figura 3: se describe la relación existente en la esperanza de vida, medida en años y el producto interno bruto del país, por habitante, discriminando por continente. A partir del gráfico se puede observar que aquellos países pertenecientes a Europa con un ingreso per cápita mayor, registran mayores niveles de esperanza de vida. En oposición, África, presenta bajos niveles de ingreso per cápita y puede asociarse a una baja esperanza de vida. En niveles intermedios se encuentra el resto de los continentes. Pero a priori, se podría decir que a mayor ingreso de los países mayor sería la esperanza de vida, lo cual se relaciona a una mayor disponibilidad de recursos para destinar a la salud y el buen vivir y a nivel macroeconómico a un mayor bienestar general de la economía.

Comentario: Muy bien el análisis y gráfico pero FALTA hacer un facet para para que no haya sobreploteo y se puede modificar la escala del eje x (log) o dejarla libre, además recordar que aquí se grafican todos los años en conjunto. Y esto va más allá del gráfico y podrían ser hipótesis a plantear (ojo van bien, y tiene mucho sentido pero esto es un análisis exploratorio): "Pero a priori, se podría decir que a mayor ingreso de los países mayor sería la esperanza de vida (OK, aunque acá un componente temporal que estamos obviando), lo cual se relaciona (podría!) a una mayor disponibilidad de recursos para destinar a la salud y el buen vivir y a nivel macroeconómico a un mayor bienestar general de la economía" (Primero habría que definir que es bienestar, segundo mayor pib es mayor bienestar? Bajo que indicador?)

4. Hacer un gráfico de lineas 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 linea 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(gapminder, aes(year, gdpPercap)) + geom_line(aes(group = country)) +
    facet_wrap(~continent) + labs(x = "Years", y = "Gross Domestic Product, per capita")
```

La Figura anterior da cuenta de la dinámica del producto interno bruto por habitante para el período 1952-2007 para cada país, discriminando por continente. En primer lugar se observa que en el caso de Europa, los niveles de producto son superiores en promedio a los observados en el resto de los continentes. En oposición, los niveles de actividad de los países africanos parecen ubicarse en niveles reducidos. Un comentario distintivo del caso de las Américas y Asia es la variabilidad que existe entre los distintos países del continente en términos de sus niveles de actividad, dando cuenta de una mayor desigualdad en términos de ingreso entre los países de cada continente; cuestión que no parece observarse en los otros continentes.

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(gapminder, aes(pop, lifeExp)) + geom_point(alpha = 1/3) +
    theme(aspect.ratio = 1) + facet_wrap(~continent) +
    labs(x = "Population", y = "Life expectancy")
```

De la Figura 5 se puede extraer algun comentario adicional no trabajado hasta ahora: la relación entre la esperanza de vida y la población de los países. Se puede distinguir en primer lugar que en los casos de las Américas y Asia, existiría alguna relación de asociación positiva entre la esperanza de vida y el nivel poblacional. Esto implicaría que la población no solo aumentaría sino también que se observaría un proceso de envejecimiento poblacional. A priori esto sería contraproducente en términos de lo que se conoce como "Capacidad de Carga" de los países (Verhulst, 1838). Esto implicaría la necesidad de mayores recursos para combatir el potencial agotamiento del bono demográfico y evitar perjuicios en el bienestar social a traves de reducciones del ingreso disponible de la población debido a la mayor población y en sintonía a una mayor cantidad de población dependiente (mayores de edad, asociados a mayores niveles de esperanza de vida). Esta observación podría generaizarse en Europa, aunque de forma no tan clara. En el caso africano, aunque la relación no parece del todo clara, se podría especular a partir de la Firgura 5 que a mayor nivel poblacional, la esperanza de vida no aumentaría. Esto puede verse explicado por altas tasas de natalidad y de mortalidad,

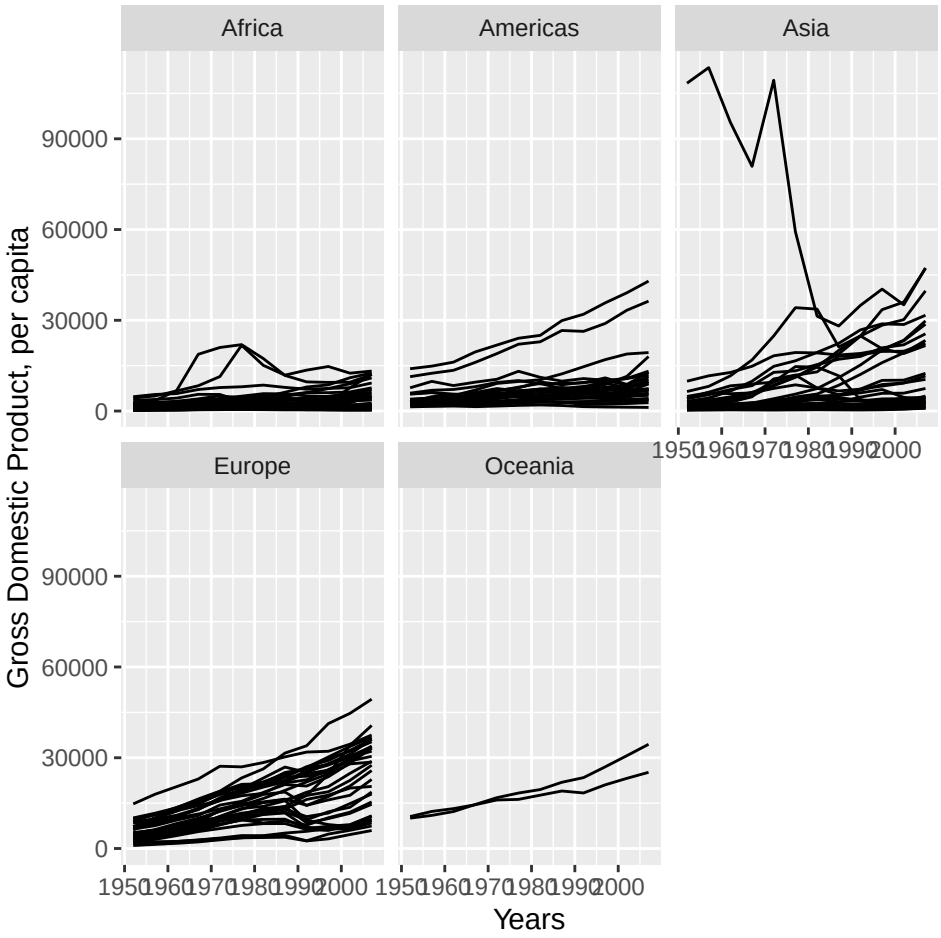


Figure 4: Producto interno bruto por habitante (1952-2007) para cada pais, segun continente.

generando escacez de recursos para la población perjudicando los niveles de vida.

Comentario: Excelente nuevamente en especial el final donde planteas posibles hipótesis, muy bien Martín! A mejorar la escala x podría quedar en log o liberarse por continente porque dado los comentarios realizados y el objetivo buscado se quiere analizar por continente entonces quedaría mejor poder hacer "un zoom" en cada continente puesto que los valores son muy diferentes en el caso de Asia y no permiten ver que sucede con el resto. Por otro lado, acá nuevamente tenes todos los años en el mismo gráfico. Podrías analizar esto para ciertos años particulares o darle color (continuo) a los años (o por lusto o décadas). No sabemos a partir de este gráfico si los países tienen o no tienen bono demográfico.

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:

- 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()`.

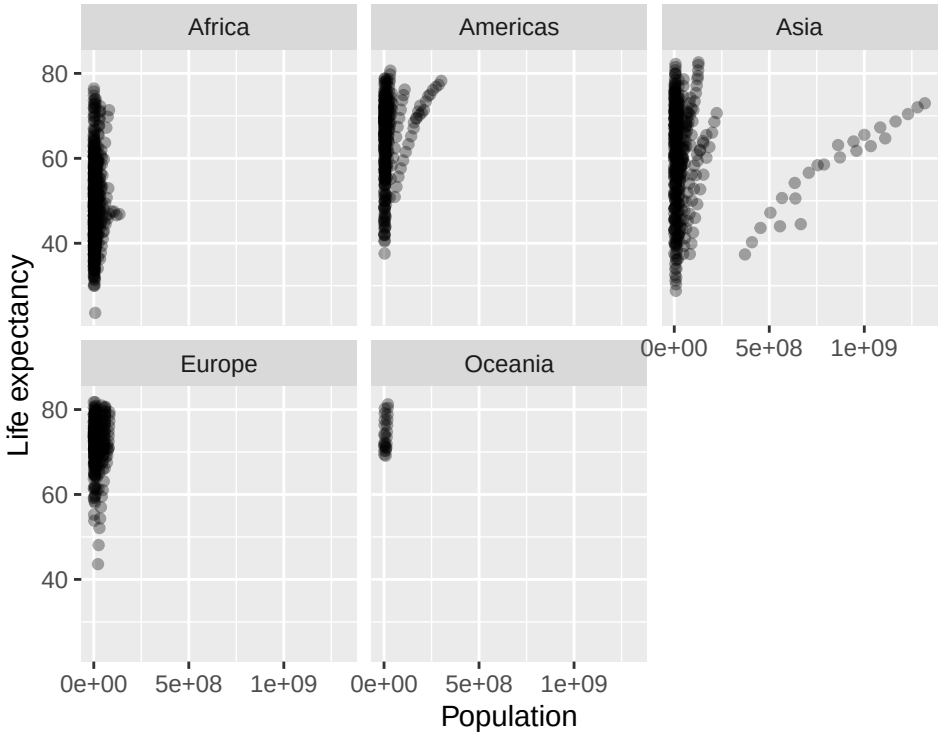


Figure 5: Diagrama de dispersión entre la esperanza de vida y el nivel poblacional de los países, según continente (1952-2007).

```
ggplot(mpg, aes(x = drv, fill = drv)) + geom_bar() +
    scale_color_brewer(palette = "Dark2") + labs(x = "Tipo de tren",
    y = "Cantidad", title = "Distribución según tipo de tren")
```

Se puede observar de la Figura 6 que la mayor parte de los trenes son del tipo “front wheel drive” y “4tw.”

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

```
ggplot(mpg, aes(x = drv, fill = drv)) + geom_bar() +
    scale_color_brewer(palette = "Dark2") + labs(x = "Tipo de tren",
    y = "Cantidad", title = "Distribución según tipo de tren")
```

```
ggplot(mpg, aes(x = drv, y = prop.table(stat(count)),
    fill = drv, label = scales::percent(prop.table(stat(count))))) +
    geom_bar(position = "dodge") + geom_text(stat = "count",
```

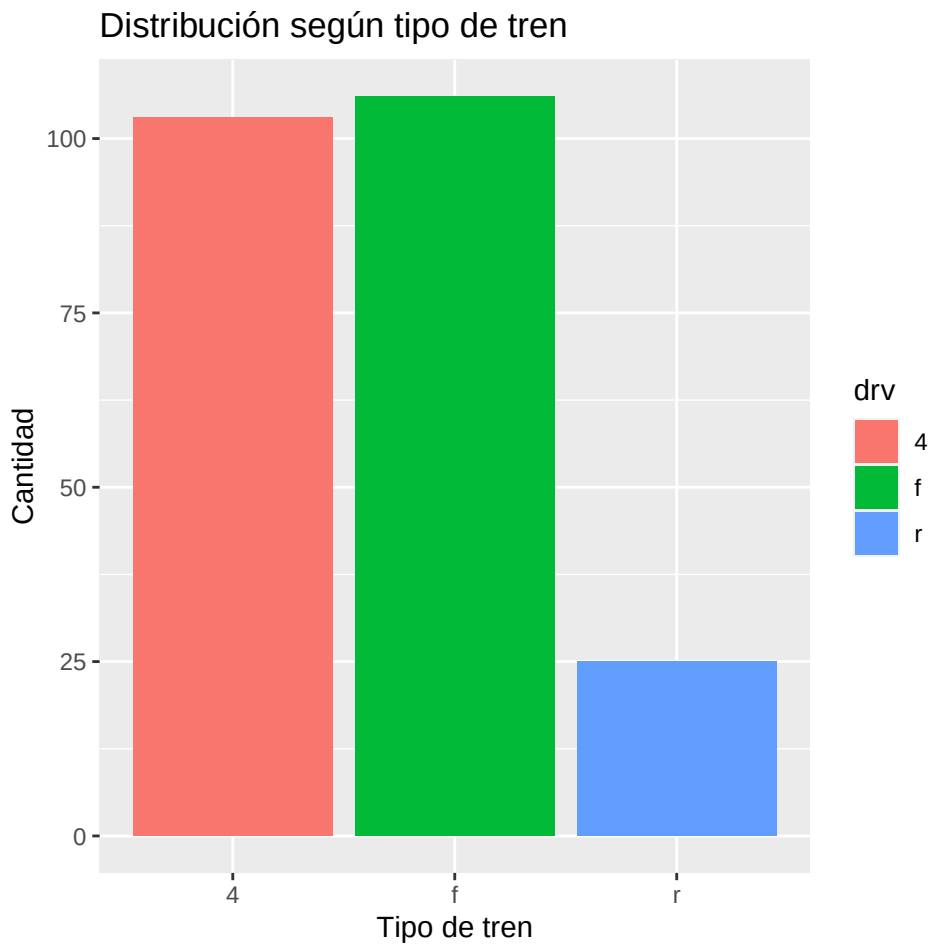


Figure 6: Figura 6: Distribución según tipo de tren.

```
    position = position_dodge(0.9), vjust = -0.5, size = 3) +
    scale_y_continuous(labels = scales::percent) +
    theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
    scale_fill_brewer(palette = "Dark2") + labs(x = "Tipo de tren",
    y = "Porcentaje", title = "Distribución según tipo de tren")
```

En particular, sumado al comentario de la Figrua 6, la Figura 8 muestra que solo el 10% de los trenes son del tipo “rear wheel drive”, luego los otros dos tipos de trenes se distribuyen el restante 90% de forma similar

Comentario: Muy bien. Ordenar las barras de forma descendente y darle nombre explicativo a las etiquetas.

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 encuentrán 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.

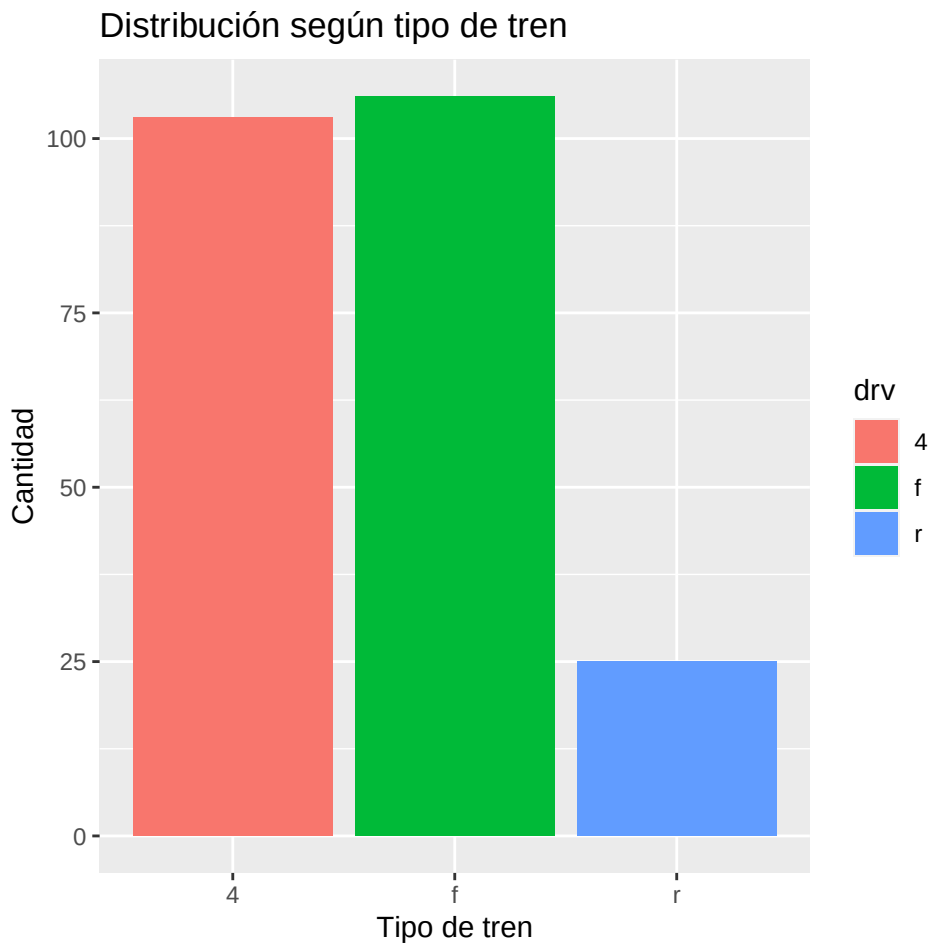


Figure 7: Figura 7: Distribución según tipo de tren.

Por simplicidad te damos los datos restructurados (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)
datos <- read.csv("dato_emision.csv", header = TRUE,
    sep = ",", dec = ".")
```

2. Usando las funciones de la librería `dplyr` obtenga qué fuentes tienen la emisión máxima. Recuerde que TOTAL debería ser excluído 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':
##
```

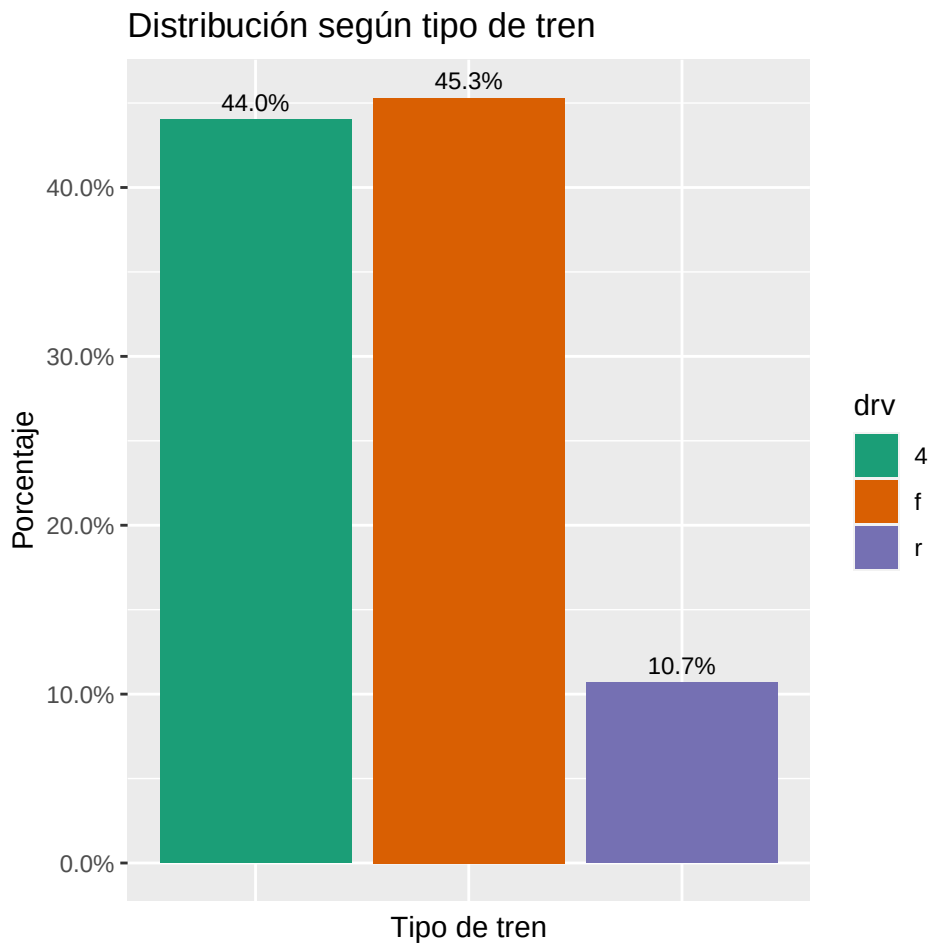


Figure 8: Figura 8: Distribución según tipo de tren.

```
##     intersect, setdiff, setequal, union
```

```
datos %>%
    filter(fuente != "TOTAL" & fuente != "S_C" & fuente !=
        "I_E") %>%
    group_by(fuente) %>%
    summarise(maximo_emision = max(emision, na.rm = TRUE)) %>%
    arrange(maximo_emision, desc = FALSE) %>%
    tail(1) %>%
    select(fuente)
```

```
## # 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?

```
colnames(datos) = c("año", "fuente", "emision")
datos %>%
    filter(fuente == "Q_B") %>%
    group_by(año) %>%
```

```
    summarise(maximo_emision = max(emision, na.rm = TRUE)) %>%
    arrange(maximo_emision, desc = FALSE) %>%
    tail(1) %>%
    select(año)
```

```
## # A tibble: 1 x 1
##     año
##   <int>
## 1  2017
```

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

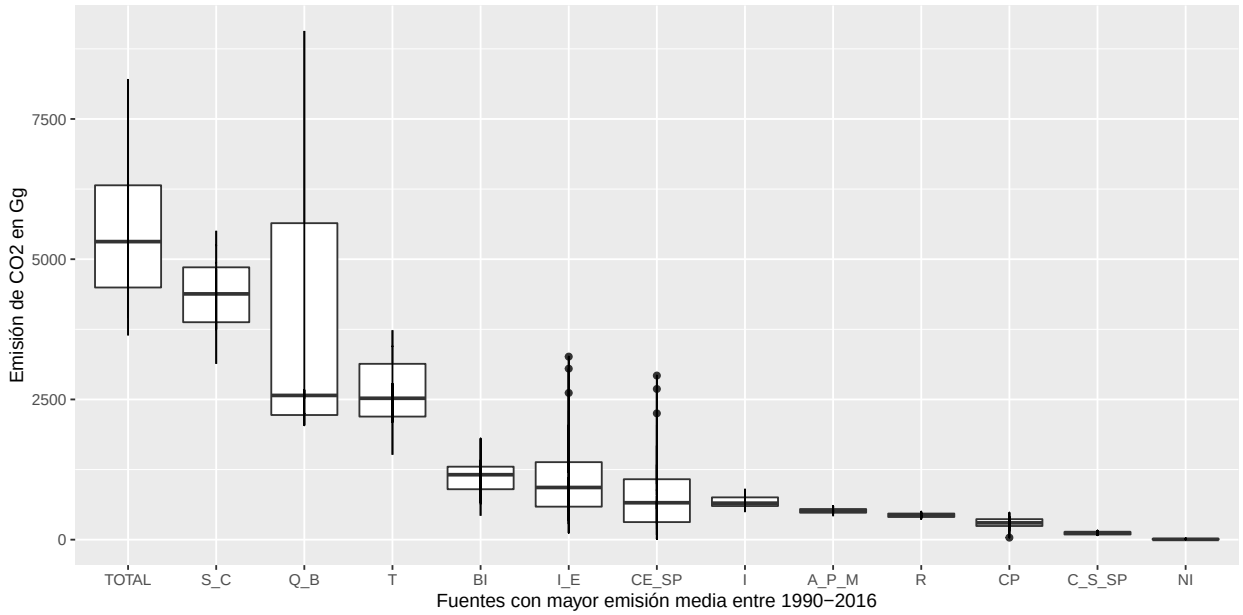
```
datos %>%
    filter(fuente != "TOTAL" & fuente != "S_C" & fuente !=
        "I_E") %>%
    group_by(fuente) %>%
    summarise(media_emision = mean(emision, na.rm = TRUE)) %>%
    arrange(media_emision) %>%
    tail(5) %>%
    select(fuente)
```

```
## # A tibble: 5 x 1
##   fuente
##   <chr>
## 1 I
## 2 CE_SP
## 3 BI
## 4 T
## 5 Q_B
```

5. Usando ggplot2 realice un gráfico de las emisiones a lo largo de los años para cada fuente. Utilice dos elementos geométricos, puntos y lineas. Selecciones 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. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.

Comentario: Los ejercicios anteriores, excelente. Este ejercicio no se realiza.

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




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

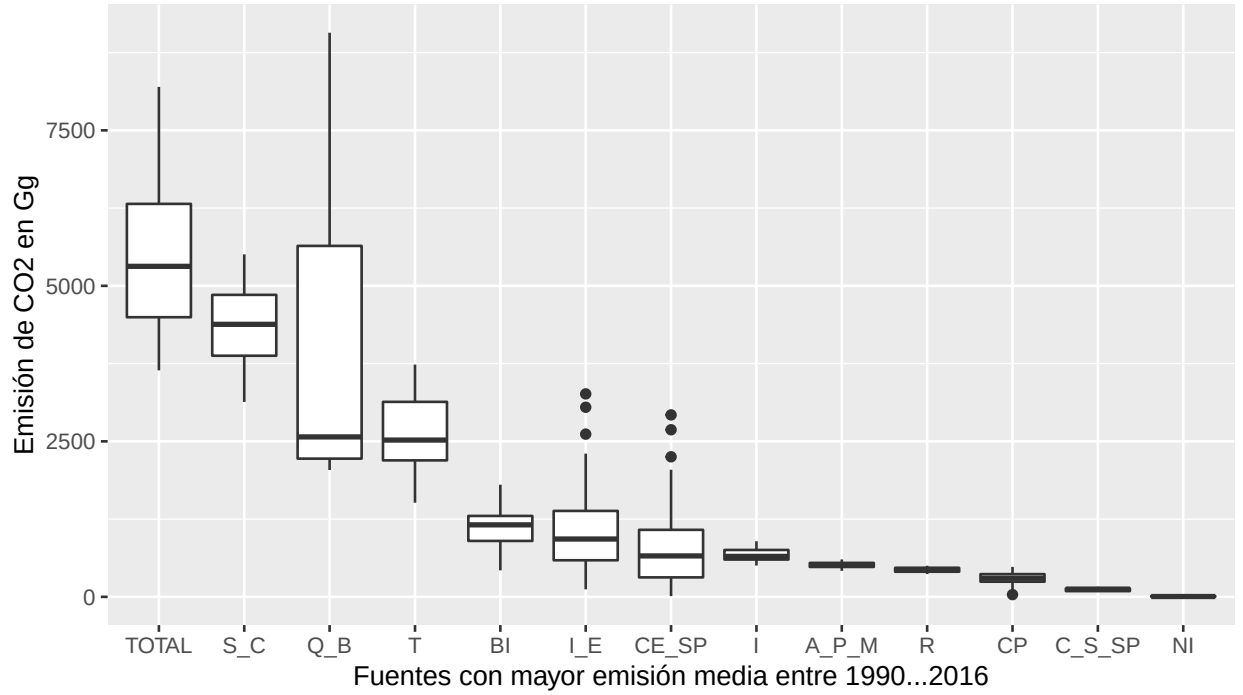
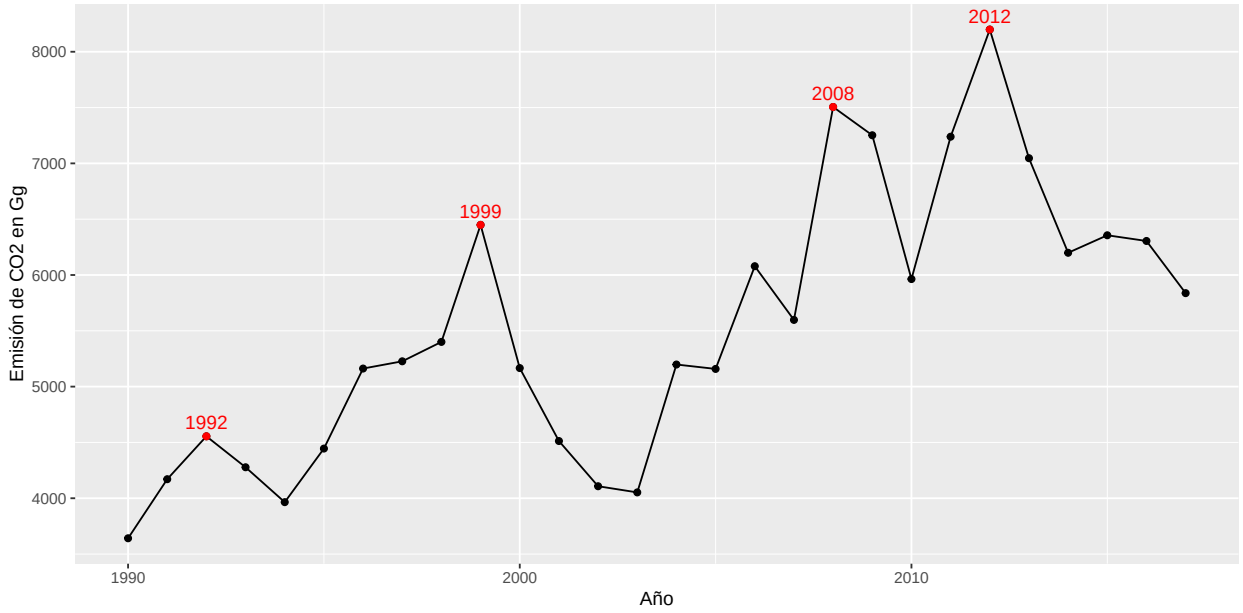


Figure 9: Figura 9: Distribución según tipo de tren.

La Figura 9 muestra la distribucion de los datos en términos de las emisiones de CO2 discriminando según las funetes de emisiones. Se observan a partir del gráfico de caja, los rangos intercuartílicos, las medias y

medianas de las emisiones de las distintas fuentes, datos atípicos, entre otros. En la Figura se observan las distintas fuentes de emisiones ordenadas según su mediana muestral, donde Q_B es la que muestra mayor variabilidad en los datos considerando el rango intercuartílico, mientras que N_I parece ser la que tiene los datos más concentrados.

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)
datos %>%
    filter(fuente == "TOTAL") %>%
    group_by(año) %>%
    ggplot(aes(x = año, y = emision)) + labs(x = "Año",
    y = "Emisión de CO2 en Gg") + geom_line() + geom_point() +
    stat_peaks(colour = "red") + stat_peaks(colour = "red",
    geom = "text", vjust = -1)
```

Se observa que en tendencia las emisiones de CO2 son crecientes a lo largo del tiempo denotando los años "records". Por años records se consideran aquellos años donde las emisiones son superiores al período siguiente. Por ejemplo, suelen utilizarse las sucesiones de records en general, en el estudio de la teoría de la probabilidad. 2012, resulta el año con mayor nivel de emisiones de CO2 mientras que se observa una brusca caida luego de 1999. Esto puede deberse al tratado que llevó al protocolo de Kyoto en 1997 con el objetivo de mitigar las emisiones de gases de efecto invernadero.

Comentario: Muy buen análisis. El único punto es que no se marcan los que son mayores al valor siguiente, notar que eso no se cumple. Si se va marcando el valor máximo (local) a medida que pasa el tiempo, cuando hay un nuevo máximo queda en rojo.

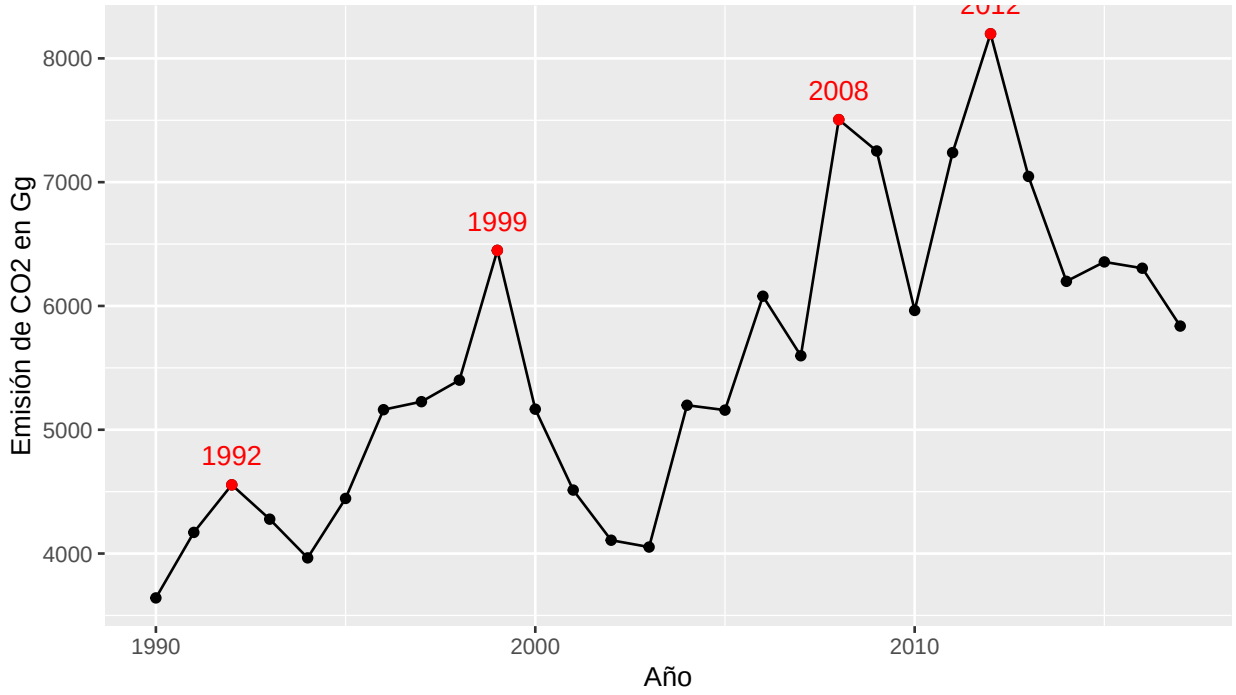


Figure 10: Figura 10: Evolución temporal de las emisiones de dióxido de carbono.

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.

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.

```
datos <- read.csv("muestra.csv", header = TRUE, sep = ",",
    dec = ".")
```

A partir de los datos vinculados al abandono escolar, es de interés entender que factores determinan la decisión de abandono de sus estudios, por parte de los estudiantes. En primera instancia, realizar un análisis exploratorio de datos, parece ser una buena estrategia inicial para este análisis. En particular, se puede hacer énfasis en describir a estos estudiantes que abandonan su estudios. Por ejemplo, distinguirlos según su sexo o el departamente de procedencia y otras dos variables que a priori parecen relevantes, como lo son la situación sociodemográfica de la intitución y la cantidad de inasistencias de estos estudiantes. Es así que se pueden formular las siguientes preguntas guías: \ 1) ¿Cómo se distribuyen los estudiantes que abandonan según departamento distinguiendo entre hombres y mujeres? \ 2) ¿Es posible observar alguna relación de asociación entre la cantidad de inasistencias y la condición de abandono de los estudiantes, distinguiéndolos por sexo? ¿Cómo es dicha relación? \ 3) ¿Se puede observar alguna relación entre la proporción de estudiantes

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

que abandonan sus estudios con la situación sociodemográfica del centro de estudios? \

```
datos %>%
    filter(Abandono == 1) %>%
    ggplot(aes(x = Abandono, y = prop.table(stat(count)),
        fill = Sexo)) + geom_bar() + scale_y_continuous(labels = scales::percent) +
    theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
    scale_fill_brewer(palette = "Dark2") + facet_wrap(~nombre_departamento) +
    labs(x = "Abandono", y = "Proporción de abandono por sexo",
        title = "Distribución de abandonos por sexo, según departamento")
```

En primer lugar se observa que Montevideo y Canelones son aquellos departamentos de Uruguay con mayor abandono de estudiantes, aunque a nivel general de todo el país parece existir un bajo nivel de abandono. Cuando se discrimina por sexo, se puede vislumbrar que la situación de abandono está más presenta en hombres que en mujeres, en el general de los departamenteos.

```
ggplot(datos, aes(x = inasistencias, y = Abandono,
    colour = Sexo)) + geom_point() + theme(aspect.ratio = 1) +
    geom_smooth(method = "lm", se = TRUE) + labs(x = "Abandono",
    y = "Inasistencias", title = "Inasistencias, según condición de abandono")
```

Aunque a priori, observando la nube de puntos asociada a la relación existente en la condición de abandono y la cantidad de inasistencias, no parece existir una relación clara. Sin embargo, esto puede deberse a que la variable de abandono solo toma dos valores, 0 o 1 y la variable inasistencias contiene valores discretos, por tanto la cantidad de observaciones puede no permitir la mejor visualización. Sin embargo, cuando se introducen las rectas de regresión asociada a los puntos discriminando por sexo, sí parece más clara la relación que existe. A partir de los datos parece existir una mayor propensión al abandono ante un escenario de mayores inasistencias. Aunque por sexo, no se encuentran diferencias importantes entre hombres y mujeres el impacto de las inasistencias en el abandono en los varones parecería ser algo superior al de las mujeres. Sin embargo, cuando se observan las rectas de regresión junto con las bandas de confianza para cada sexo, se observa que ambas bandas se superponen, es decir, la diferencia encontrada entre hombres y mujeres en términos de esta pregunta no es significativa, permitiendo encontrar escenarios donde el impacto de las

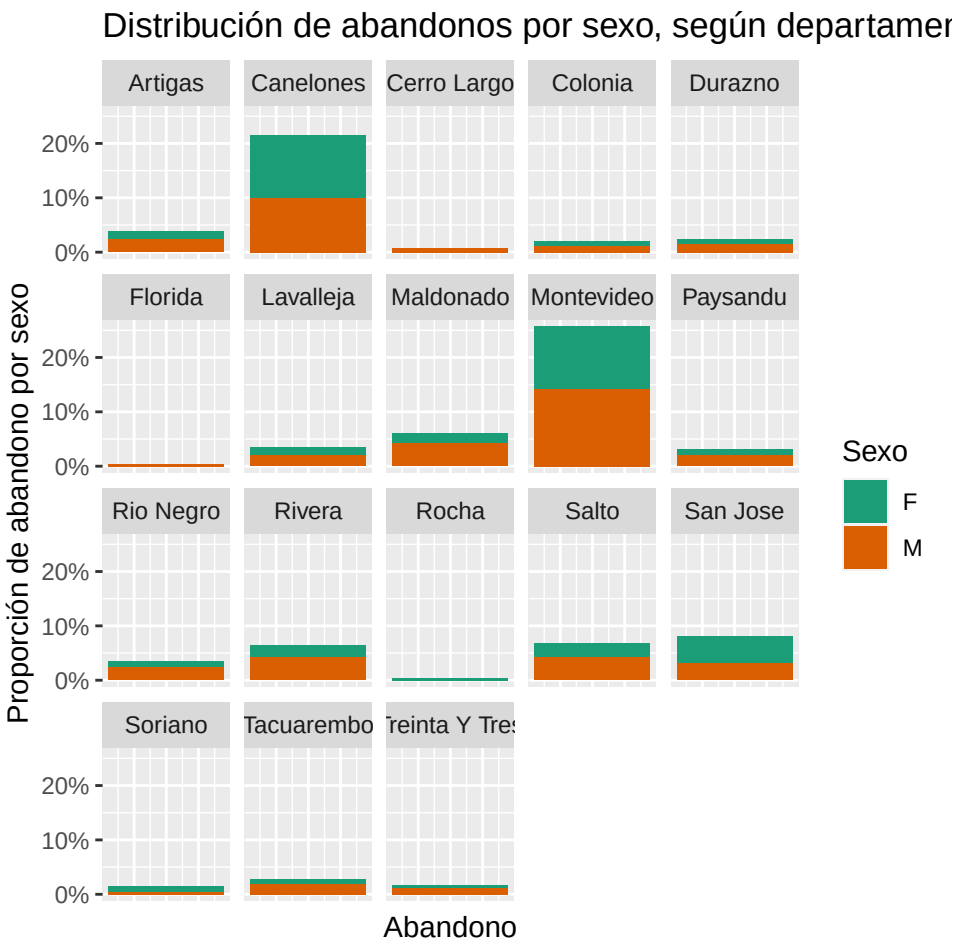


Figure 11: Distribución de abandonos por sexo, según departamento.

inasitencias sera igual o incluso se revierta, entre los sexos. \

En suma, si se logra visualizar un impacto de las inasistencias en la propensión a abandonar de los estudiantes, aunque esta no se diferencia según sexo.

```
# as.factor(datos$cl)


ggplot(datos, aes(x = Abandono)) + geom_bar() + scale_fill_brewer(palette = "Dark2") +
    facet_wrap(~cl) + labs(x = "Abandono", y = "Proporción de abandono por cluster",
    title = "Distribución de abandonos por situación sociodemográfica de la institución, según departame
```

La visualización anterior permite reconocer si la condición de abandono a un centro educativo se relaiona de alguna manera con la condición sociodemográfica del mismo. No se dispone de información sobre qué representa cada nivel de condición sociodemográfica (cada cluster/grupo), pero a priori y dada la escasa cantidad de estudiantes que abandonan no parecerían existir vínculos que llamen la atención entre entre la condición de abandono y la condición sociodemográfica de la institución. En lo que sigue se propone una alternativa sólo condierando a los estudiantes que abandonan.

```
# as.factor(datos$cl)
```

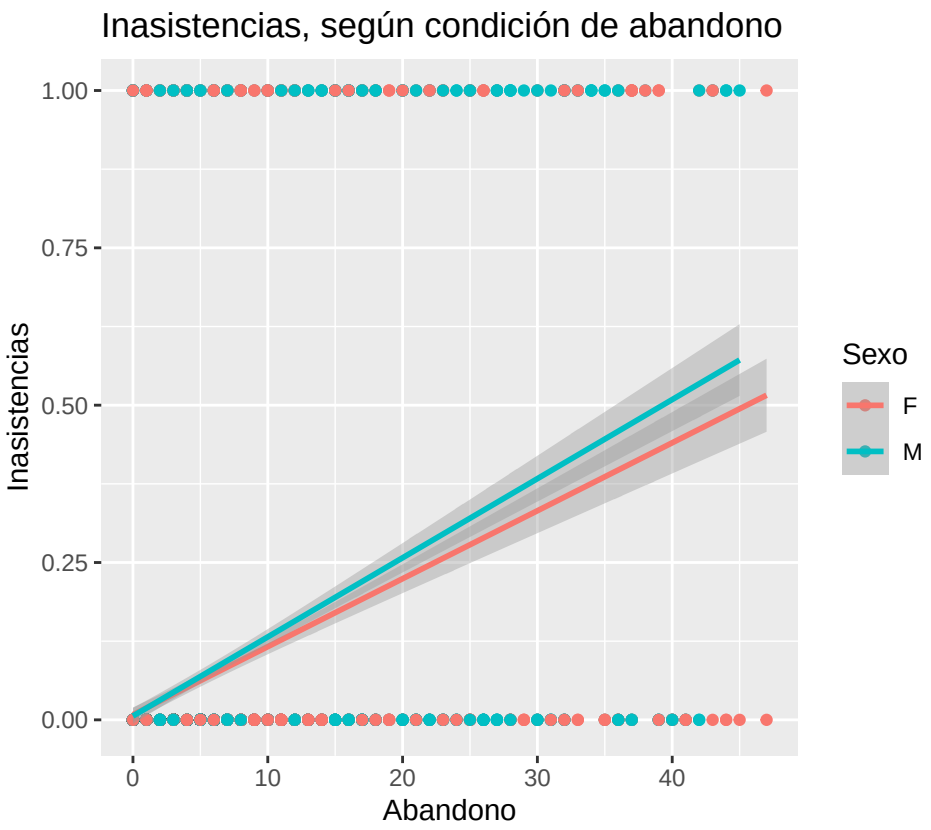


Figure 12: Distribución de abandonos por sexo, según departamento.

```
datos %>%
    filter(Abandono == 1) %>%
    ggplot(aes(x = Abandono)) + geom_bar() + scale_fill_brewer(palette = "Dark2") +
    facet_wrap(~cl) + labs(x = "Abandono", y = "Proporción de abandono por cluster",
    title = "Distribución de abandonos por situación sociodemográfica de la institución, según departam
```

De lo anterior se observa que los clusters 1, 3 y 5 son aquellos con mayor abandono estudiantil. Sin embargo, al no disponer de mayor información sobre las características de cada cluster, no se podría interpretar en demasía la existencia de alguna relación. Suponiendo que las condiciones 1 y 5 son extremos en términos de situación sociodemográfica, no parecería existir mayores diferencias entre ellos en cuanto al abandono y por lo tanto, se podría especular en que la situación sociodemográfica de la institución no afectaría sustantivamente la condición de abandono de un estudiante.

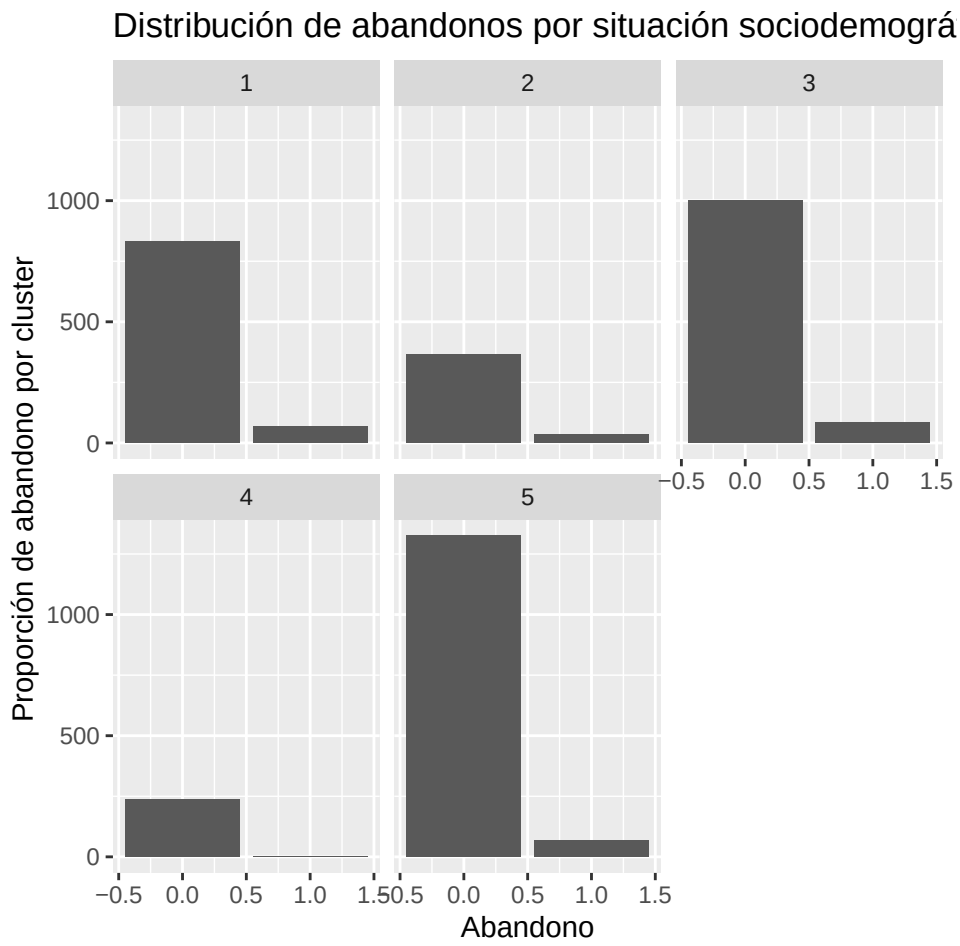


Figure 13: Distribución de abandonos por sexo, según departamento.

Comentario: Muy buen trabajo, excelente nivel de análisis, de gráficas y planteamiento de hipótesis sobre distintas visualizaciones. Previamente realicé algunos comentarios a considerar en algunas interpretaciones y graficos.

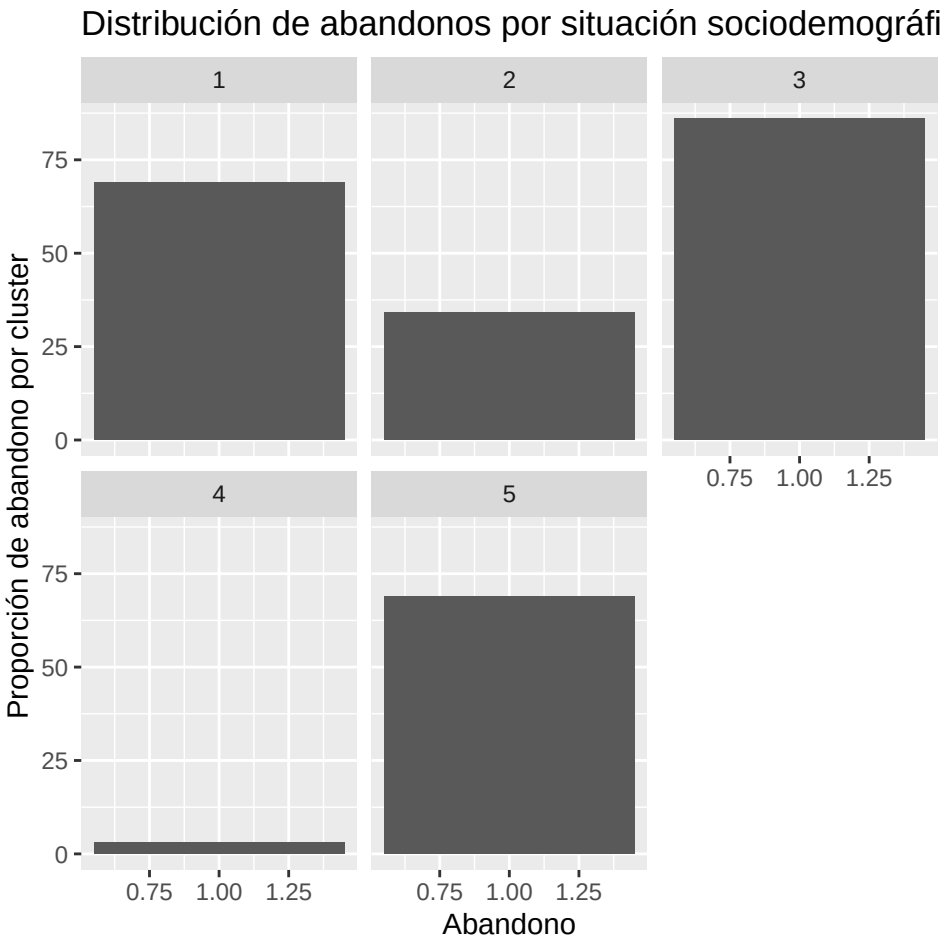


Figure 14: Distribución de abandonos por sexo, según departamento.