Tarea 2

Carolina Rodriguez

4/5/2018

Entrega

La tarea debe ser realizada en RMarkdown en un repositorio de GitHub llamado "Tarea 2". La tarea es individual por lo que cada uno tiene que escribir su propia versión de la misma. El repositorio debe contener unicamente el archivo .Rmd con la solución de la tarea.

Ejercicio 1

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/emisiones-de-co2-por-sector.

Debe leer con cuidado la información en los meta datos para responder correctamente.

Los datos fueron reestructurados para simplificar la exploración, de la siguente manera:

```
library(tidyverse)
dato_emision <- gather(dat, key = fuente, value = emisión,
    -AÑO)
```

Con estos datos responda las sigueintes preguntas:

_ 1._ Usando las funciones de la librería `dplyr` obtenga qué fuentes tienen la emisión máxima. Recuerde que TOTAL debería ser excluído para esta respuesta. ‘

```
filter(dato_emision, fuente != "TOTAL" & emision ==
    max(emision, na.rm = TRUE))
```

```
##    AÑO fuente emision
## 1 2016    Q_B  8831.1
```

```
filter(dato_emision, !fuente %in% c("TOTAL", "S_C",
    "I_E")) %>% group_by(fuente)
```

```
## # A tibble: 270 x 3
## # Groups:   fuente [10]
##      AÑO fuente emision
##    <int> <chr>    <dbl>
##  1  1990 CE_SP    299.
##  2  1991 CE_SP    642.
##  3  1992 CE_SP    779.
##  4  1993 CE_SP    530.
##  5  1994 CE_SP     84.7
##  6  1995 CE_SP    318.
##  7  1996 CE_SP    674.
##  8  1997 CE_SP    482.
```

```
##  9  1998 CE_SP    287.
## 10  1999 CE_SP   1338.
## # ... with 260 more rows
```

La fuente con máxima emisión es Q_B, la cual refiere a quema de biomasa.

_ 2._ ¿En qué año se dió la emisión máxima para la fuente que respondió en la pregunta anterior?

```
filter(dato_emision, fuente != "TOTAL" & emision ==
    max(emision, na.rm = TRUE))
```

```
##    AÑO fuente emision
## 1 2016    Q_B  8831.1
```

Se da en el año 2016.

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

```
dato_emision %>% filter(fuente != "TOTAL" & fuente !=
    "S_C" & fuente != "I_E") %>% group_by(fuente) %>%
    summarise(media_fuente = mean(emision, na.rm = TRUE)) %>%
    arrange(desc(media_fuente)) %>% top_n(5)
```

```
## Selecting by media_fuente
```

```
## # A tibble: 5 x 2
##   fuente media_fuente
##   <chr>         <dbl>
## 1 Q_B           3691.
## 2 T             2579.
## 3 BI            1125.
## 4 CE_SP          893.
## 5 I              674.
```

4. 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 3). 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.

```
datos <- filter(dato_emision, !is.na(emision)) %>%
    filter(fuente == c("Q_B", "T", "BI", "CE_SP", "I"))
ggplot(datos, aes(x = AÑO, y = emision)) + geom_point() +
    geom_line() + facet_wrap(~fuente) + labs(x = "AÑOS",
    y = "Emisiones de CO2")
```

5. Relplique el siguiente gráfico usando ggplot2.

```
datos %>% ggplot(aes(x = fct_reorder(fuente, -emision),
    emision)) + geom_boxplot() + labs(x = "Fuentes con mayor emisión media entre 1990-2016",
    y = "Emisión de CO2 en Gg")
```

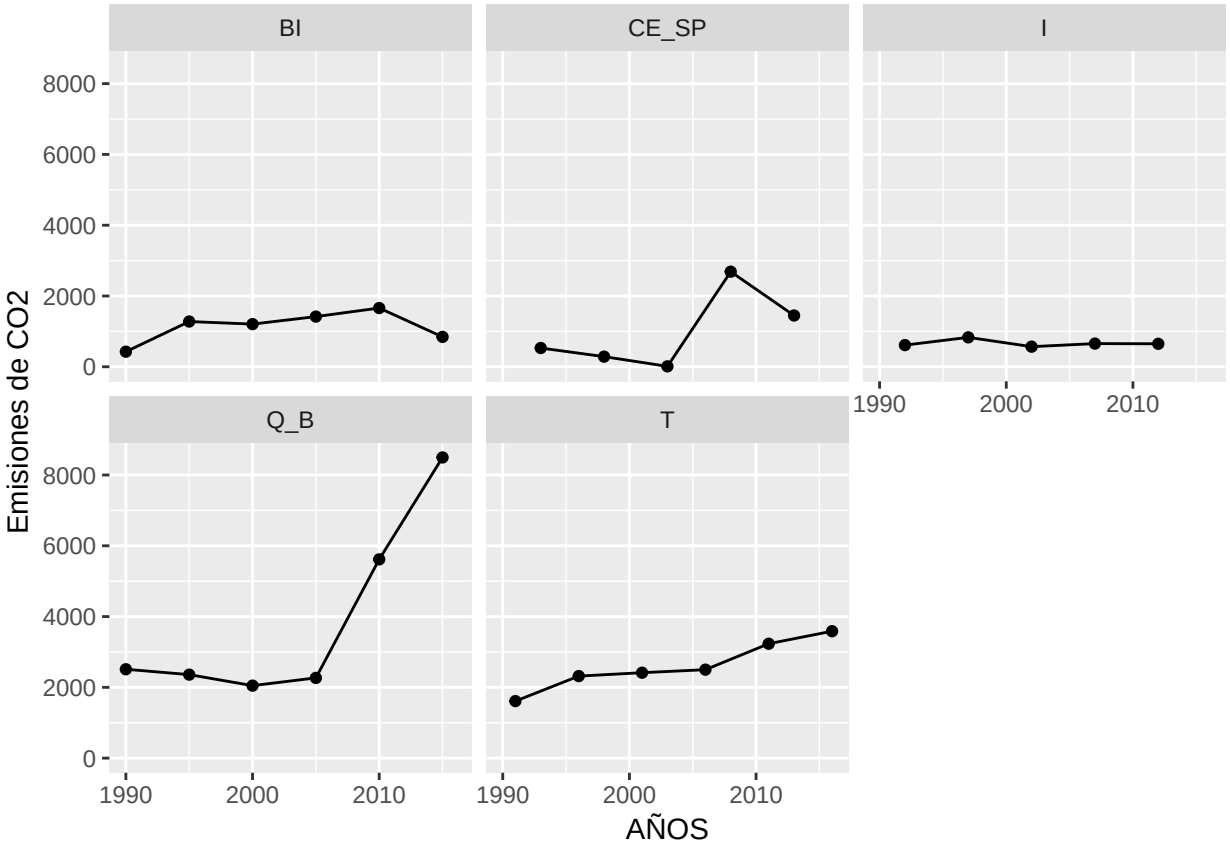
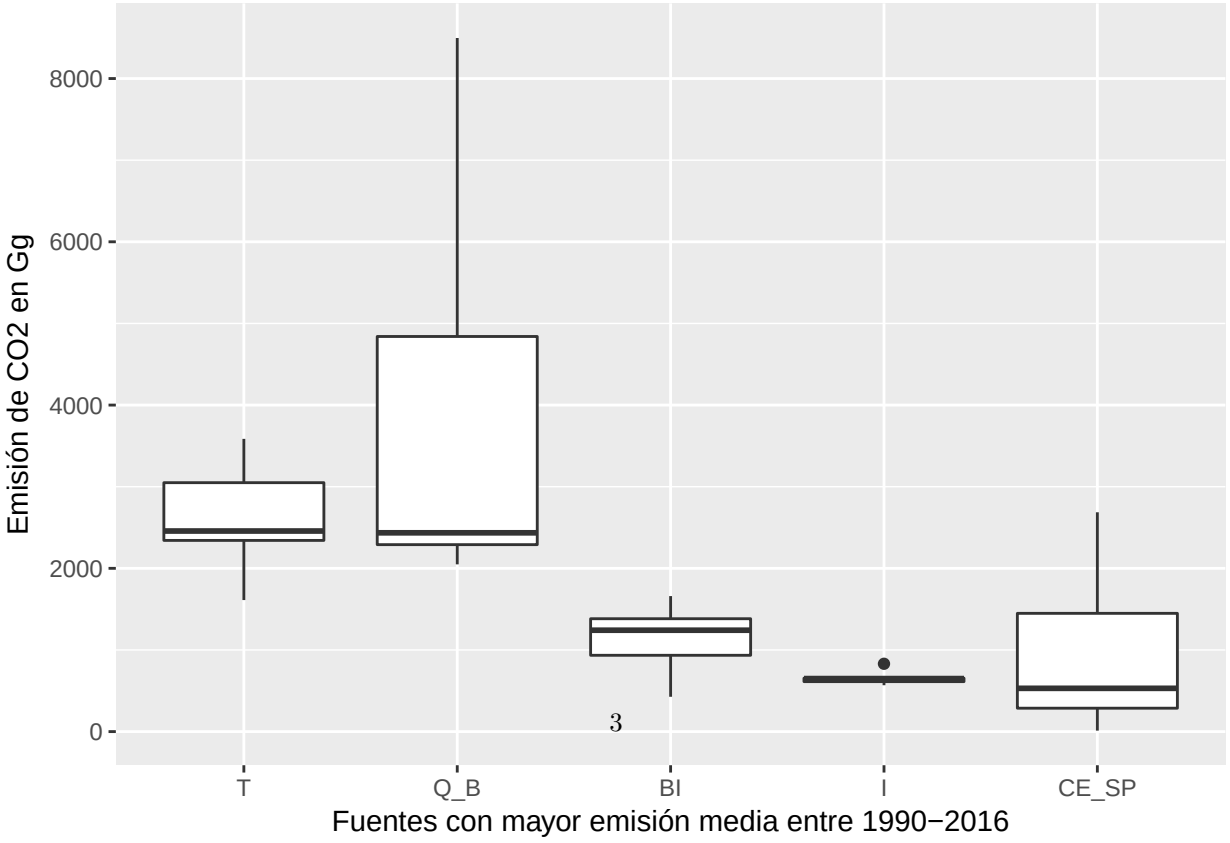


Figure 1: Las emisiones de CO2 de quema de biomasa crecen exponencialmente a partir del año 2005, también se observa un crecimiento para el transporte pero es mucho más leve. La industria se mantiene en valores bajos y constantes para el periodo de análisis, mientras que el comportamiento de las emisiones para las centrales electricas servicio público tienen un comportamiento bastante varaible.



6. 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.

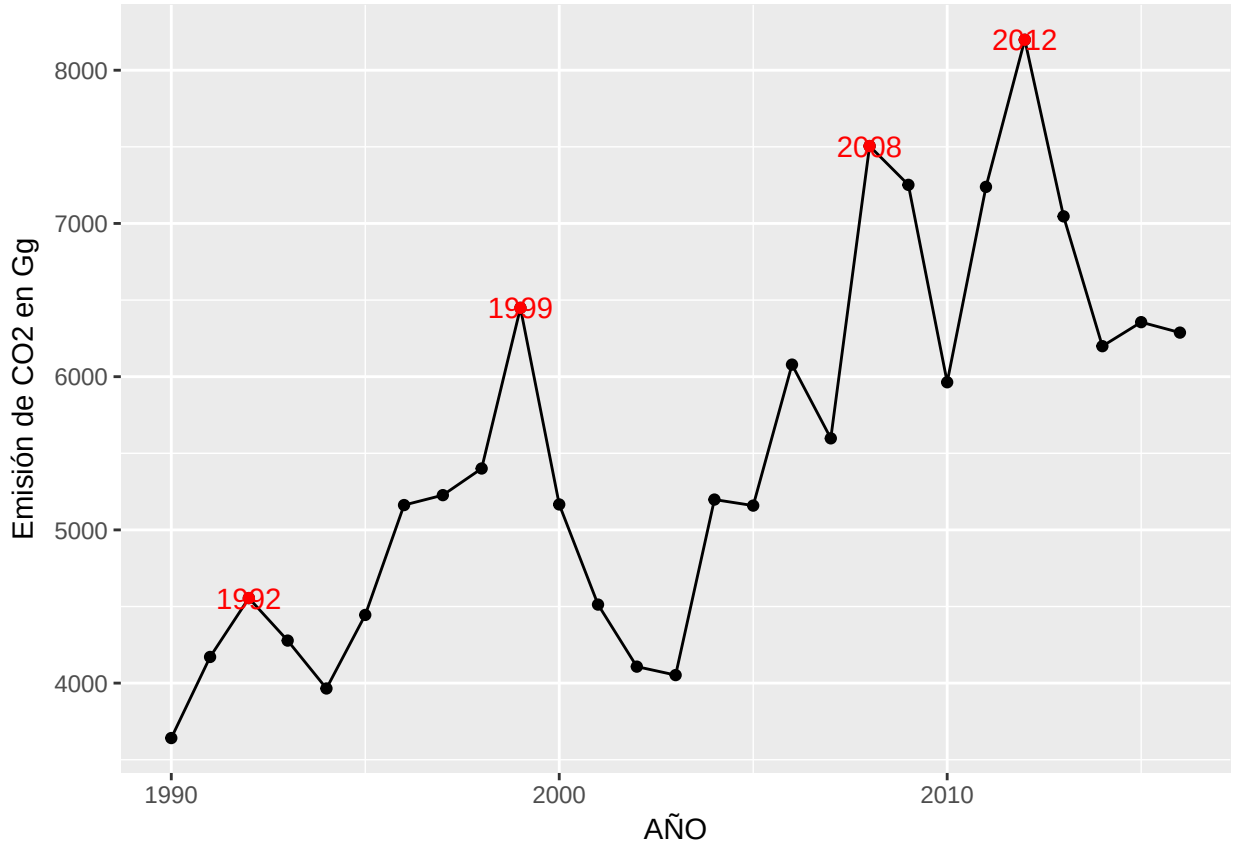
```
library(ggpmisc)
```

```
## Warning: package 'ggpmisc' was built under R version 3.4.4
```

```
## For news about 'ggpmisc', please, see http://www.r4photobiology.info/
```

```
## For on-line documentation see http://docs.r4photobiology.info/ggpmisc/
```

```
library(tidyverse)
dato_emision %>% filter(fuente == "TOTAL") %>% group_by(AÑO) %>%
    ggplot(aes(AÑO, emision)) + geom_point() + geom_line() +
    labs(x = "AÑO", y = "Emisión de CO2 en Gg") + stat_peaks(colour = "red") +
    stat_peaks(geom = "text", colour = "red")
```



Ejercicio 2

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 los gastos realizados por actos médicos Los datos y los metadatos se encuentran disponibles en:

https://catalogodatos.gub.uy/dataset/gasto_am_2016_fondo-nacional-de-recursos/resource/936ac9e6-b0f6-424a-9b53-ee408a

Este ejercicio tiene como objetivo que realice un análisis exploratorio de datos utilizando todo lo aprendido en el curso. Debe contener al menos 5 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.

1

¿Qué tipo de variables contiene nuestra base de datos? en caso de tener variables categoricas ver los niveles de las mismas.

```
names(gastos)  # nombres de las variables
```

```
## [1] "Prestacion"              "Paciente"
## [3] "Edad_aÃ.os"              "Sexo"
## [5] "Departamento_residencia" "Prestador"
## [7] "Prestador_tipo"          "Prestador_departamento"
## [9] "Importe"
```

```
str(gastos)  # estructura del conjunto de datos
```

```
## 'data.frame':    23811 obs. of  9 variables:
##  $ Prestacion             : Factor w/ 30 levels "CARDIODESFIBRILADOR IMPLANTE",..: 1 1 1 1 1 1 1 1 1
##  $ Paciente               : int  48572 56004 67813 78014 80694 82954 84293 97516 98726 111481 ...
##  $ Edad_aÃ.os             : int  34 77 60 66 54 69 77 79 75 73 ...
##  $ Sexo                   : Factor w/ 2 levels "F","M": 2 2 2 2 1 2 2 2 2 2 ...
##  $ Departamento_residencia: Factor w/ 19 levels "ARTIGAS","CANELONES",..: 14 2 10 10 10 9 7 10 4 2 .
##  $ Prestador              : Factor w/ 117 levels "AMECOM IAMPP",..: 62 75 92 112 111 1 61 92 14 19 .
##  $ Prestador_tipo         : Factor w/ 4 levels "ASSE","IAMC",..: 1 2 1 2 4 2 1 1 2 2 ...
##  $ Prestador_departamento : Factor w/ 20 levels "ARTIGAS","CANELONES",..: 8 2 10 10 10 9 7 10 4 10 .
##  $ Importe                : num  424476 405983 401947 404193 420440 ...
```

```
dim(gastos)
```

```
## [1] 23811     9
```

```
levels(gastos$Prestacion)
```

```
##  [1] "CARDIODESFIBRILADOR IMPLANTE"
##  [2] "CARDIODESFIBRILADOR PROCEDIMIENTOS POSTERIORES"
##  [3] "CATETERISMO DERECHO ADULTOS"
##  [4] "CATETERISMO DIAGNOSTICO INFANTIL"
##  [5] "CATETERISMO TERAPEUTICO"
##  [6] "CIRUGIA CARDIACA ADULTO"
##  [7] "CIRUGIA CARDIACA INFANTIL"
##  [8] "COCLEAR IMPLANTE"
##  [9] "COCLEAR PROCEDIMIENTOS POSTERIORES"
## [10] "DIALISIS - HEMODIALISIS"
## [11] "DIALISIS - PERITONEAL"
## [12] "GRANDES QUEMADOS INFANTIL"
## [13] "MARCAPASOS IMPLANTE"
## [14] "MARCAPASOS PROCEDIMIENTOS POSTERIORES"
## [15] "PCI-ATCP"
## [16] "PCI-ATCP c/cateterismo izq."
```

```
## [17] "PCI-Cateterismo izq.adultos"
## [18] "PROTESIS DE CADERA POR ARTROSIS"
## [19] "PROTESIS DE CADERA POR FRACTURA"
## [20] "PROTESIS DE CADERA RECAMBIO"
## [21] "PROTESIS DE RODILLA IMPLANTE"
## [22] "PROTESIS DE RODILLA RECAMBIO"
## [23] "REPRODUCCION ASISTIDA DE ALTA COMPLEJIDAD"
## [24] "TPH - Alogenico"
## [25] "TPH - Autologo"
## [26] "TRASPLANTE CARDIACO"
## [27] "TRASPLANTE HEPATICO"
## [28] "TRASPLANTE RENAL DONANTE CADAVERICO"
## [29] "TRASPLANTE RENAL DONANTE VIVO"
## [30] "TRASPLANTE RENO-PANCREATICO"
```

```
levels(gastos$Sexo)
```

```
## [1] "F" "M"
```

```
levels(gastos$Departamento_residencia)
```

```
##  [1] "ARTIGAS"        "CANELONES"      "CERRO LARGO"    "COLONIA"
##  [5] "DURAZNO"        "FLORES"         "FLORIDA"        "LAVALLEJA"
##  [9] "MALDONADO"      "MONTEVIDEO"     "PAYSANDU"       "RIO NEGRO"
## [13] "RIVERA"         "ROCHA"          "SALTO"          "SAN JOSE"
## [17] "SORIANO"        "TACUAREMBO"     "TREINTA Y TRES"
```

```
levels(gastos$Prestador)
```

```
##   [1] "AMECOM IAMPP"
##   [2] "AMEDRIN IAMPP"
##   [3] "ANCAP"
##   [4] "Asociaci?n Espa?ola"
##   [5] "Asociaci?n M?dica de San Jos? IAMPP"
##   [6] "ASSE - Sin especificar"
##   [7] "Banco Hipotecario"
##   [8] "Blue Cross FONASA"
##   [9] "CAAMEPA IAMPP"
##  [10] "CAMCEL IAMPP"
##  [11] "CAMDEL IAMPP"
##  [12] "CAMEC IAMPP"
##  [13] "CAMEDUR IAMPP"
##  [14] "CAMOC IAMPP"
##  [15] "CAMS IAMPP"
##  [16] "CAMY"
##  [17] "Casa de Galicia"
##  [18] "CASMER IAMPP"
##  [19] "CASMU"
##  [20] "Centro auxiliar Ciudad de la Costa"
##  [21] "Centro auxiliar de Aigu?"
##  [22] "Centro auxiliar de Bella Uni?n"
##  [23] "Centro auxiliar de Cardona-Florencio Sanchez"
##  [24] "Centro auxiliar de Carmelo"
##  [25] "Centro auxiliar de Castillos"
##  [26] "Centro auxiliar de Chuy"
##  [27] "Centro auxiliar de Ciudad del Plata"
```

```
##  [28] "Centro auxiliar de Dolores"
##  [29] "Centro auxiliar de Ecilda Paullier"
##  [30] "Centro auxiliar de Guich?n"
##  [31] "Centro auxiliar de Jos? Batlle y Ordo?ez"
##  [32] "Centro auxiliar de Juan Lacaze"
##  [33] "Centro auxiliar de La Paloma"
##  [34] "Centro auxiliar de Las Piedras"
##  [35] "Centro auxiliar de Lascano"
##  [36] "Centro auxiliar de Libertad"
##  [37] "Centro auxiliar de Nueva Helvecia"
##  [38] "Centro auxiliar de Nueva Palmira"
##  [39] "Centro auxiliar de Pan de Azucar"
##  [40] "Centro auxiliar de Pando"
##  [41] "Centro auxiliar de Paso de los Toros"
##  [42] "Centro auxiliar de Rinc?n de la Bolsa"
##  [43] "Centro auxiliar de Rio Branco"
##  [44] "Centro auxiliar de Rosario"
##  [45] "Centro auxiliar de San Gregorio de Polanco"
##  [46] "Centro auxiliar de San Ram?n"
##  [47] "Centro auxiliar de Santa Luc?a"
##  [48] "Centro auxiliar de Sarand? del Y?"
##  [49] "Centro auxiliar de Sarand? Grande"
##  [50] "Centro auxiliar de Soca"
##  [51] "Centro auxiliar de Tala"
##  [52] "Centro auxiliar de Young"
##  [53] "Centro de Salud del Cerro"
##  [54] "Centro de Salud Dr. Enrique Claveaux"
##  [55] "Centro Departamental de Artigas"
##  [56] "Centro Departamental de Canelones"
##  [57] "Centro Departamental de Cerro Largo"
##  [58] "Centro Departamental de Colonia"
##  [59] "Centro Departamental de Durazno"
##  [60] "Centro Departamental de Flores"
##  [61] "Centro Departamental de Florida"
##  [62] "Centro Departamental de Lavalleja"
##  [63] "Centro Departamental de Maldonado"
##  [64] "Centro Departamental de Paysand?"
##  [65] "Centro Departamental de Rio Negro"
##  [66] "Centro Departamental de Rivera"
##  [67] "Centro Departamental de Rocha"
##  [68] "Centro Departamental de Salto"
##  [69] "Centro Departamental de San Jos?"
##  [70] "Centro Departamental de Soriano"
##  [71] "Centro Departamental de Tacuaremb?"
##  [72] "Centro Departamental de Treinta y Tres"
##  [73] "Circulo Cat?lico de Obreros"
##  [74] "Clinica Castillo"
##  [75] "COMECA IAMPP"
##  [76] "COMEF IAMPP"
##  [77] "COMEFLO IAMPP"
##  [78] "COMEPA"
##  [79] "COMERI"
##  [80] "COMERO IAMPP"
##  [81] "COMETT"
```

```
##  [82] "COMTA IAMPP"
##  [83] "COPAMHI"
##  [84] "COSEM"
##  [85] "CRAME"
##  [86] "CRAMI IAMPP"
##  [87] "CUDAM"
##  [88] "Direcci?n Nacional de Sanidad de las Fuerzas Armadas"
##  [89] "GREMCA"
##  [90] "GREMEDA IAMPP"
##  [91] "Hospital Brit?nico"
##  [92] "Hospital de Cl?nicas"
##  [93] "Hospital Espa?ol"
##  [94] "Hospital Etchepare"
##  [95] "Hospital Evang?lico"
##  [96] "Hospital Maciel"
##  [97] "Hospital Pasteur"
##  [98] "Hospital Pereira Rossell"
##  [99] "Hospital Pi?eyro Del Campo"
## [100] "Hospital Policial"
## [101] "Hospital Saint Bois"
## [102] "Hospital Vilardeb?"
## [103] "IAC Treinta y Tres"
## [104] "IMPASA"
## [105] "Instituto Nacional de Ordopedia y Traumatolog?a"
## [106] "Instituto Nacional del C?ncer"
## [107] "Medica Uruguaya"
## [108] "Medicina Personalizada"
## [109] "Otra instituci?n p?blica"
## [110] "Red de Atenci?n Primaria RAP"
## [111] "Seguro Americano"
## [112] "Servicio M?dico Integral"
## [113] "SISMED"
## [114] "Sociedad M?dica Quirurgica del Salto"
## [115] "SUMMUM FONASA"
## [116] "UMER Cardona"
## [117] "Universal"
```

```
levels(gastos$Prestador_tipo)
```

```
## [1] "ASSE"           "IAMC"           "OTRO"           "SEGURO PRIVADO"
```

```
levels(gastos$Prestador_departamento)
```

```
##  [1] "ARTIGAS"        "CANELONES"      "CERRO LARGO"    "COLONIA"
##  [5] "DURAZNO"        "FLORES"         "FLORIDA"        "LAVALLEJA"
##  [9] "MALDONADO"      "MONTEVIDEO"     "PASANDU"        "RIO NEGRO"
## [13] "RIVERA"         "ROCHA"          "SALTO"          "SAN JOSE"
## [17] "SIN DATO"       "SORIANO"        "TACUAREMBO"     "TREINTA Y TRES"
```

2

¿Cuánto se gastó por tipo de prestación y por sexo?

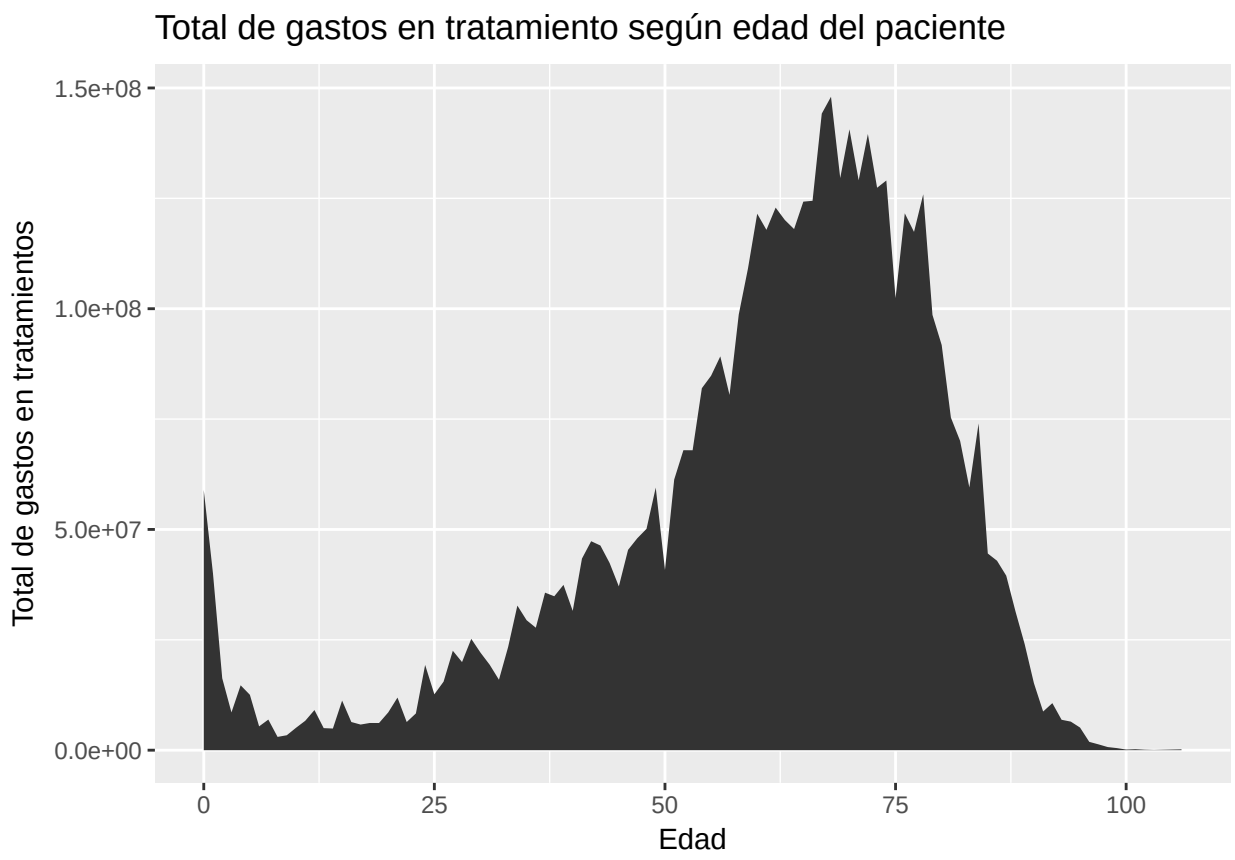
```r
group_by(gastos, Prestacion) %>% summarise(gastra = sum(Importe)) %>%
    arrange(desc(gastra))
```

```
## # A tibble: 30 x 2
##    Prestacion                              gastra
##    <fct>                                    <dbl>
##  1 DIALISIS - HEMODIALISIS             1551191180.
##  2 CIRUGIA CARDIACA ADULTO             1023054368.
##  3 PCI-ATCP c/cateterismo izq.          468258951.
##  4 PROTESIS DE RODILLA IMPLANTE         281821808.
##  5 PROTESIS DE CADERA POR ARTROSIS      198460342.
##  6 DIALISIS - PERITONEAL                177760873.
##  7 PCI-Cateterismo izq.adultos          170851003.
##  8 TPH - Autologo                       164927350.
##  9 TRASPLANTE RENAL DONANTE CADAVERICO  134516984.
## 10 CIRUGIA CARDIACA INFANTIL            125776369.
## # ... with 20 more rows
```

3

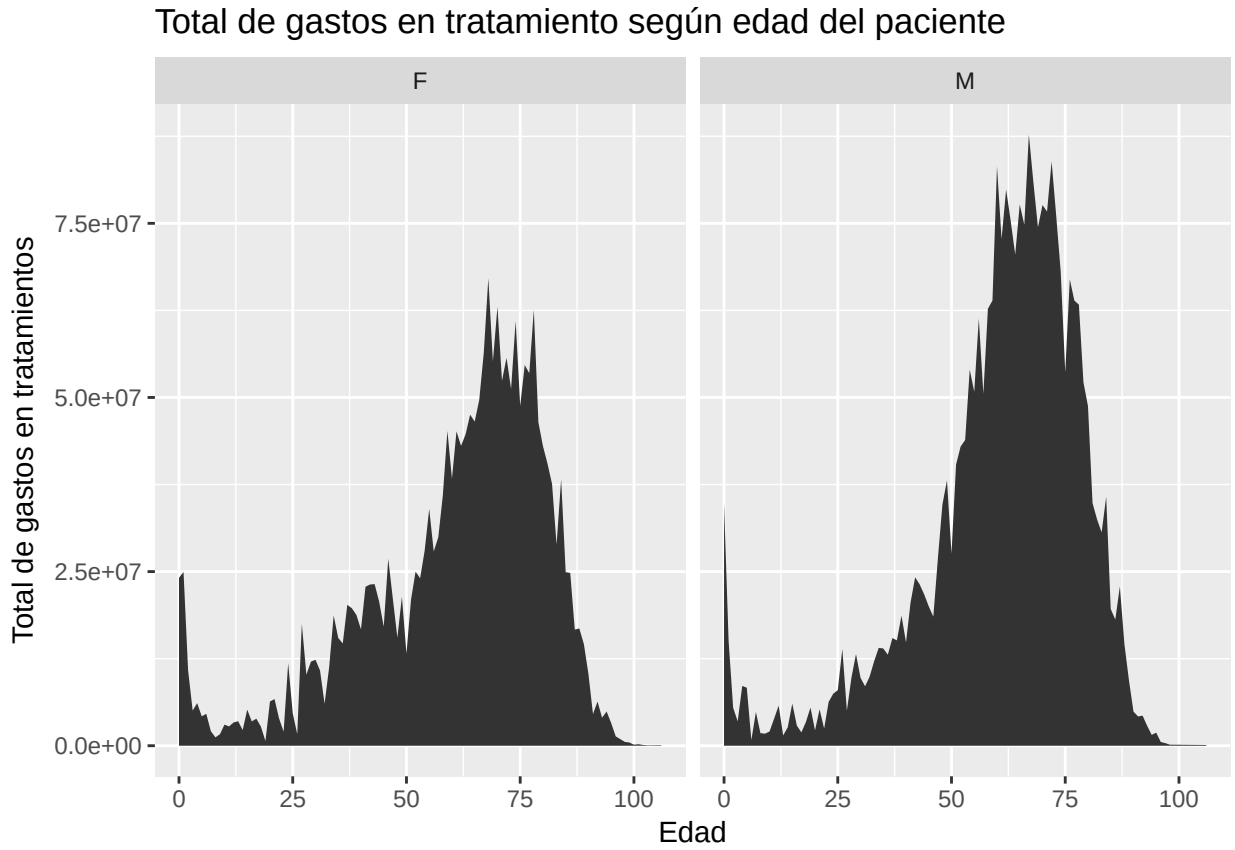
¿Cuánto fue el gasto en tratamientos por edad?

```r
group_by(gastos, Edad_aÃ.os) %>% summarise(suma = sum(Importe)) %>%
    ggplot(aes(Edad_aÃ.os, suma)) + geom_area() + labs(x = "Edad",
    y = "Total de gastos en tratamientos", title = "Total de gastos en tratamiento según edad del pacien
```



Esta gráfica resulta interesante para visualizar donde se encuentran los grupos de edad que más gastan en tratamietos médicos, veamos que pasa si lo faceteamos de acuerdo a al sexo.

```
group_by(gastos, Edad_aÃ.os, Sexo) %>% summarise(suma = sum(Importe)) %>%
    ggplot(aes(Edad_aÃ.os, suma)) + geom_area() + labs(x = "Edad",
    y = "Total de gastos en tratamientos", title = "Total de gastos en tratamiento según edad del pacie
    facet_wrap(~Sexo)
```



De acuerdo a los gáficos observamos que luego de los 50 años se gasta más en tratamientos para hombres que en mujeres, entre los 25 y los 45 el gasto por tratamiento en mujeres es mayor que en los hombres, esto se puede deber a la edad en que las mujeres quedan emabarazadas.

Qué tratamientos son los que predominan por edad?

```
filter(gastos, Edad_aÃ.os > 55 & Edad_aÃ.os < 75) %>%
    group_by(Prestacion, Sexo) %>% summarise(cantidad = n()) %>%
    ggplot(aes(Prestacion, cantidad, fill = Sexo)) +
    geom_col(position = "dodge") + theme(axis.text.x = element_text(angle = 90)) +
    coord_flip() + labs(title = "Cantidad de tratamientos por Prestacion para personas entre 55 y 70 añ
```

¿Cuántos tratamientos se realizaron por tipo de prestación y como se distribuyen entre sexo femenino y masculino?

```
library(ggplot2)
library(tidyverse)
ggplot(gastos, aes(x = fct_infreq(Prestacion), fill = Sexo)) +
    geom_bar() + theme(axis.text.x = element_text(angle = 90)) +
    labs(x = "tipo de tratamiento", y = "Numero de tratamientos",
        title = "Cantidad de tratamientos por tipo de prestación y sexo") +
```

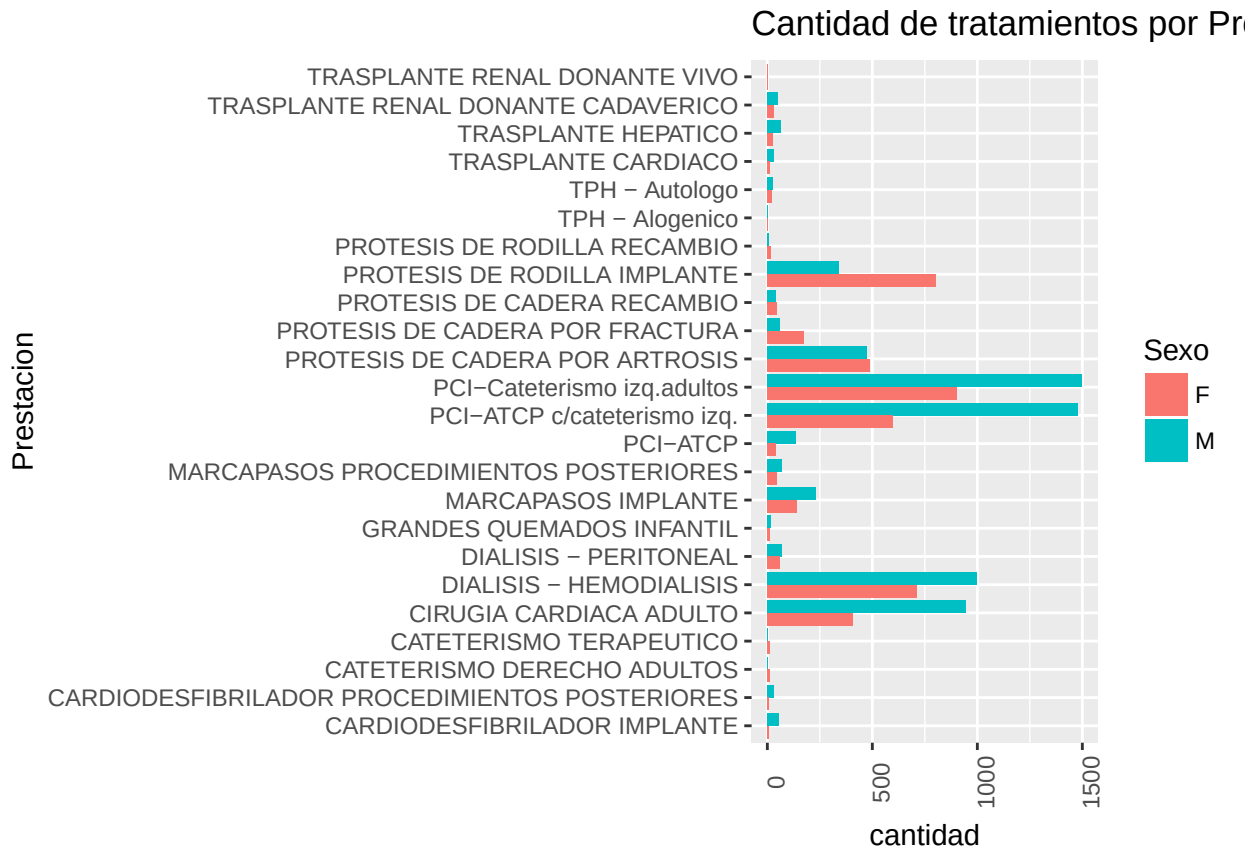


Figure 2: Para este grupo de personas, la cantidad de tratamientos en personas de sexo masculino supera ampliamente al sexo femenino en todos los tipos de tratamiento excepto:implante de protesis de rodilla,protesis de cadera por fractura,protesis de cadera recambio y protesis de cadera por artrosis. El tratamiento que más se realizan los hombres en este grupo de edad son PCI-Cateterismo izq.adultos y PCIc/cateterismo izq.

```
    coord_flip()
```

Veamos el mismo gráfico en terminos del porcentaje:

```
ggplot(gastos, aes(x = fct_infreq(Prestacion), fill = Sexo)) +
    geom_bar(position = "fill") + theme(axis.text.x = element_text(angle = 90)) +
    labs(x = "tipo de tratamiento", y = "Numero de tratamientos",
        title = "Cantidad de tratamientos por tipo de prestación y sexo") +
    coord_flip()
```

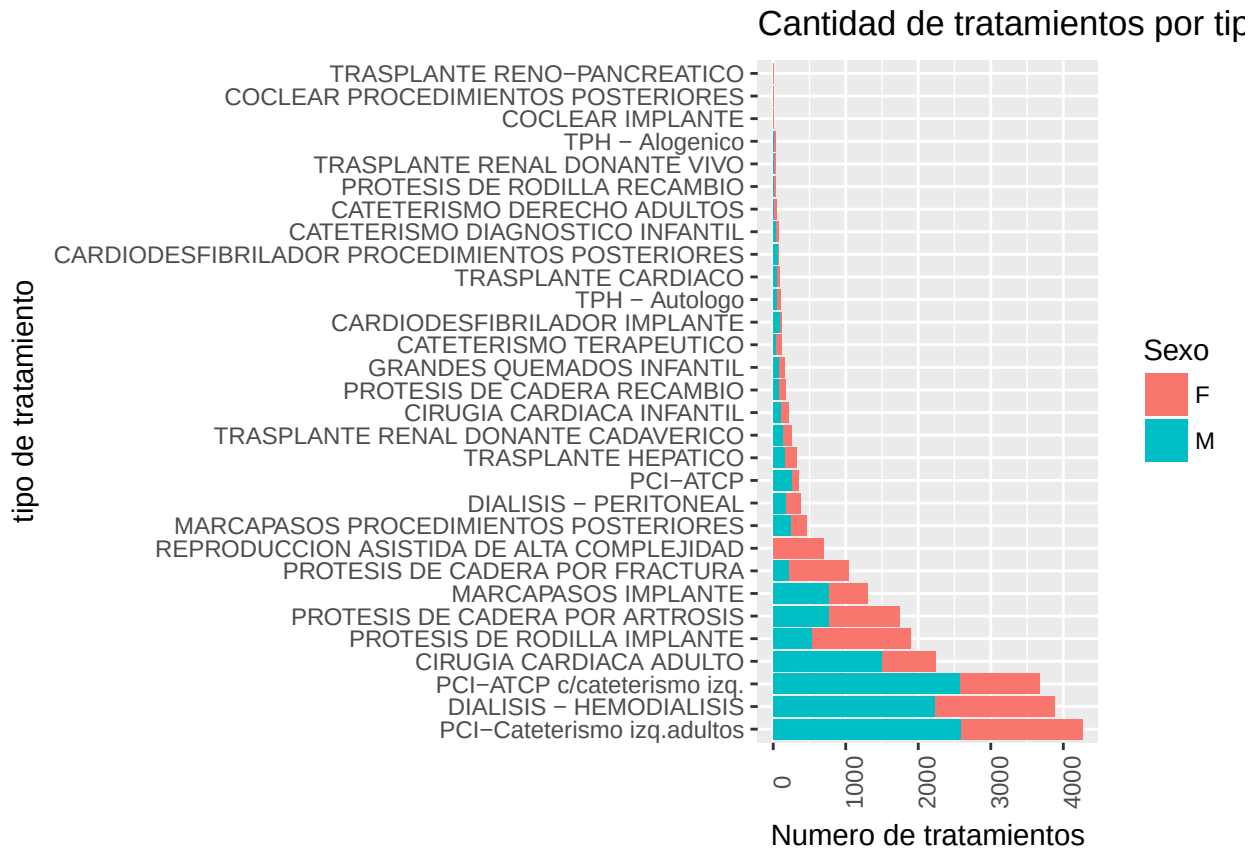
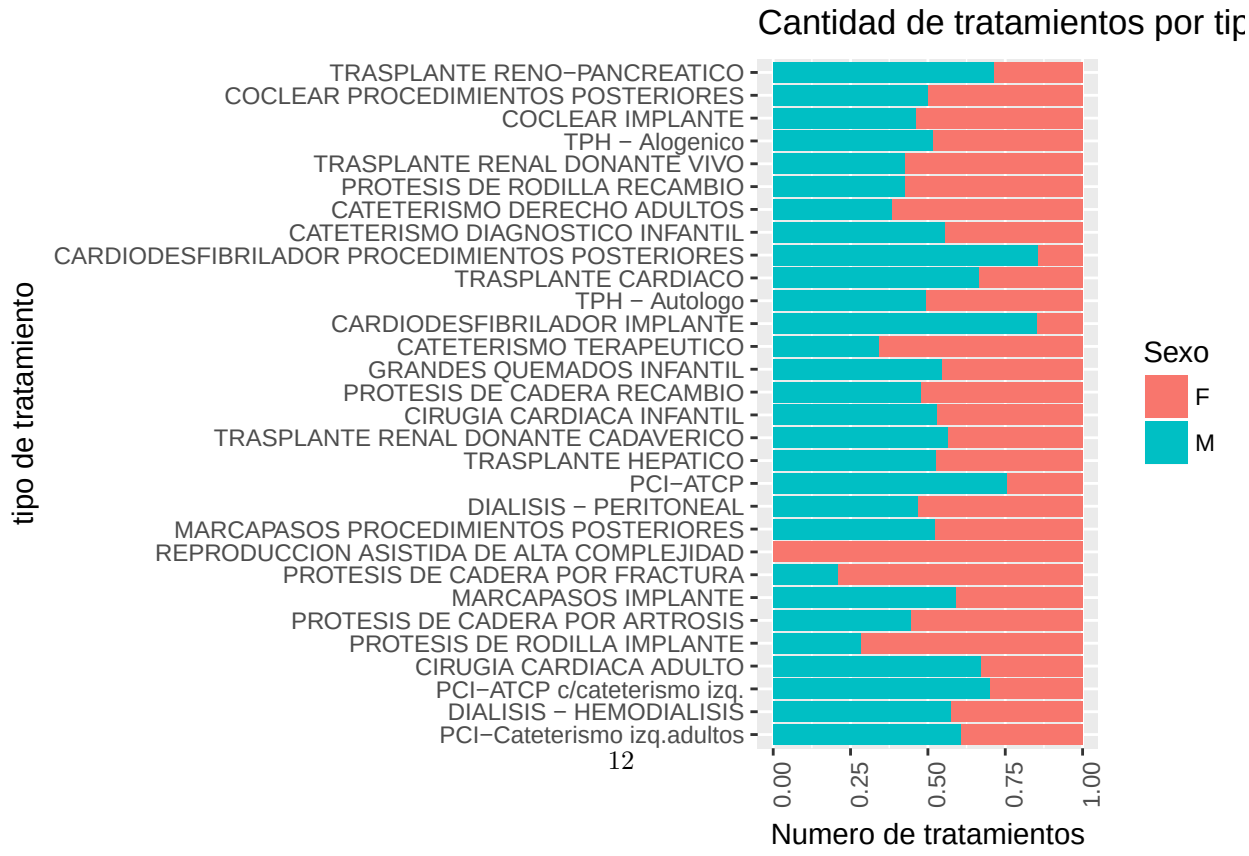


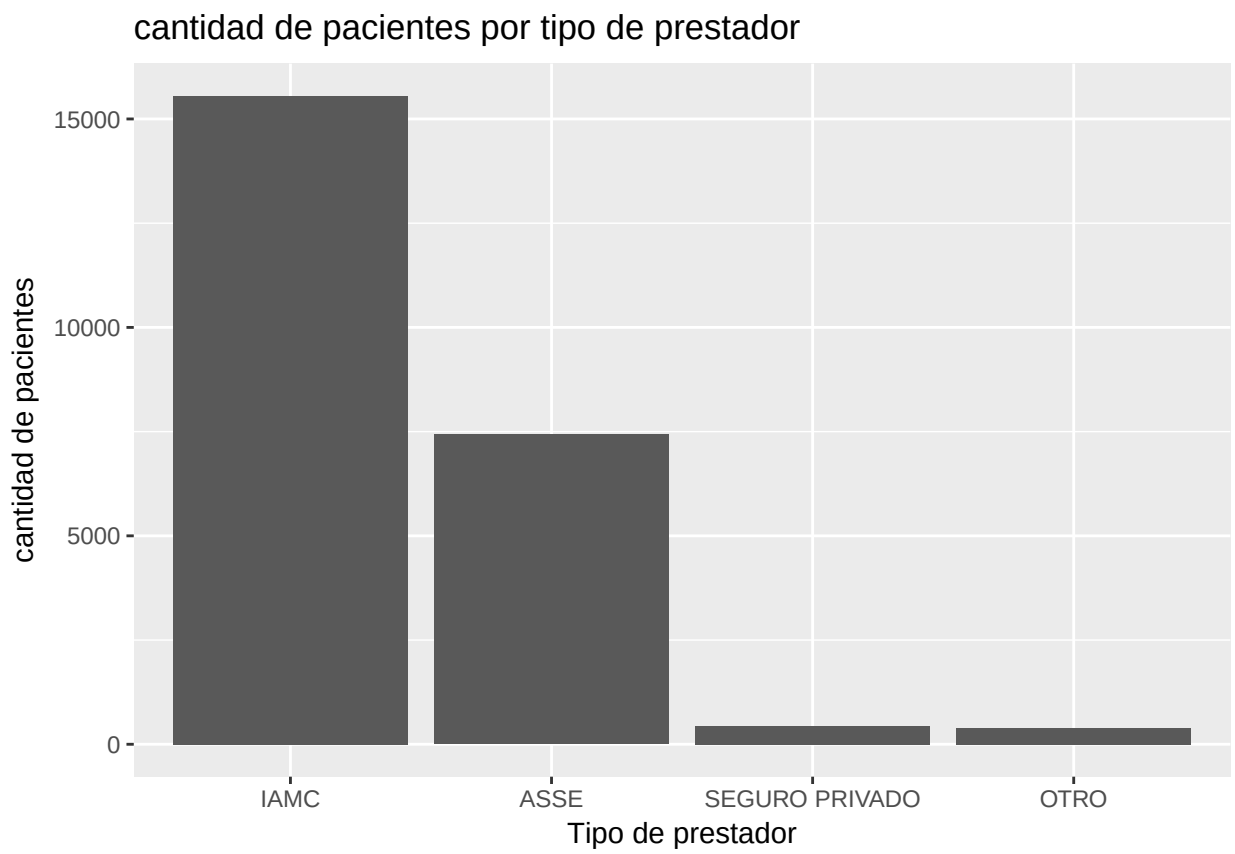
Figure 3: Los tres tratamientos mas realizados son:PCI-cateterismo adulto izq, Dialisis-hemodialisis,PCI-ATPC c/cateterismo izq. Para la mayoría de las prestaciones, los hombres se realizan más tratamientos que las mujeres.



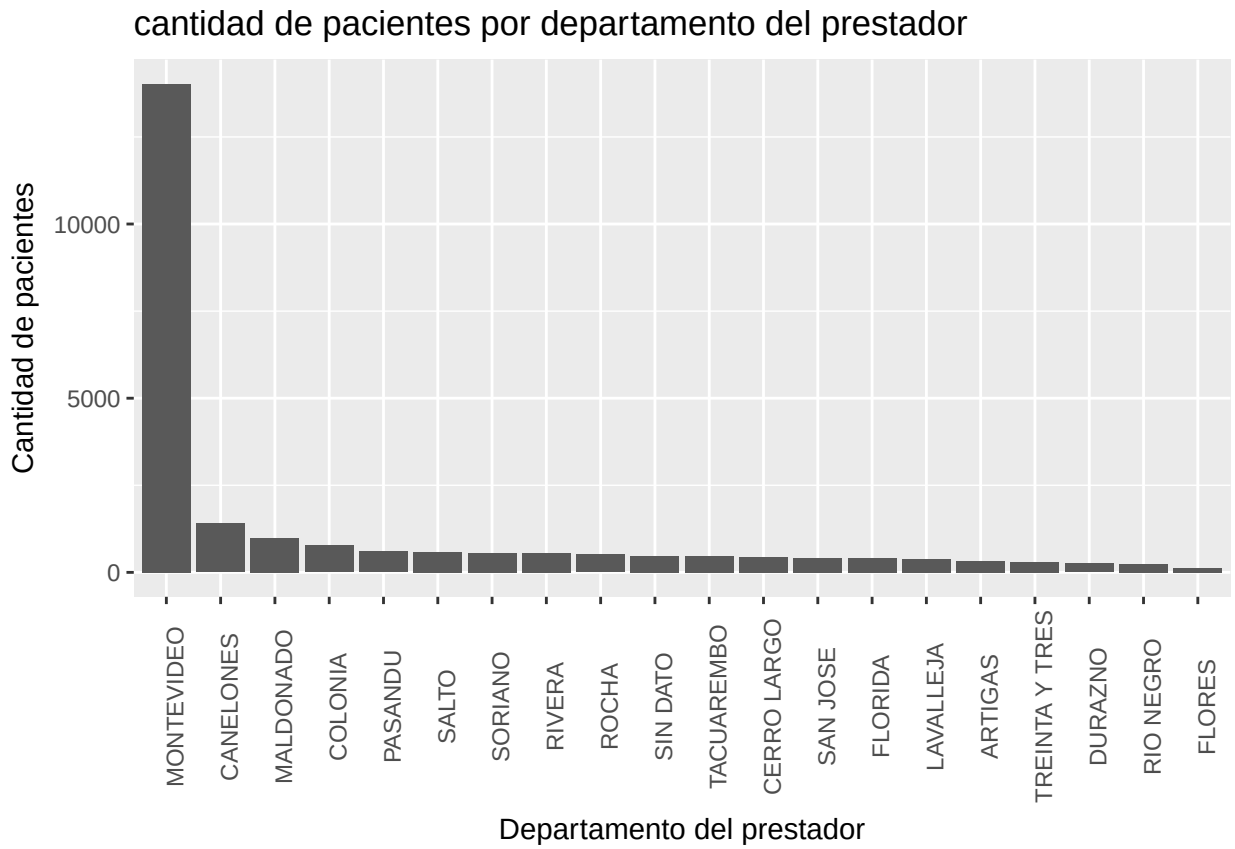
Veamos ahora algunos gráficos que nos permiten visualizar las variaciones de los datos.

Para el caso de variables cualitativas utilizamos graficos de barra.

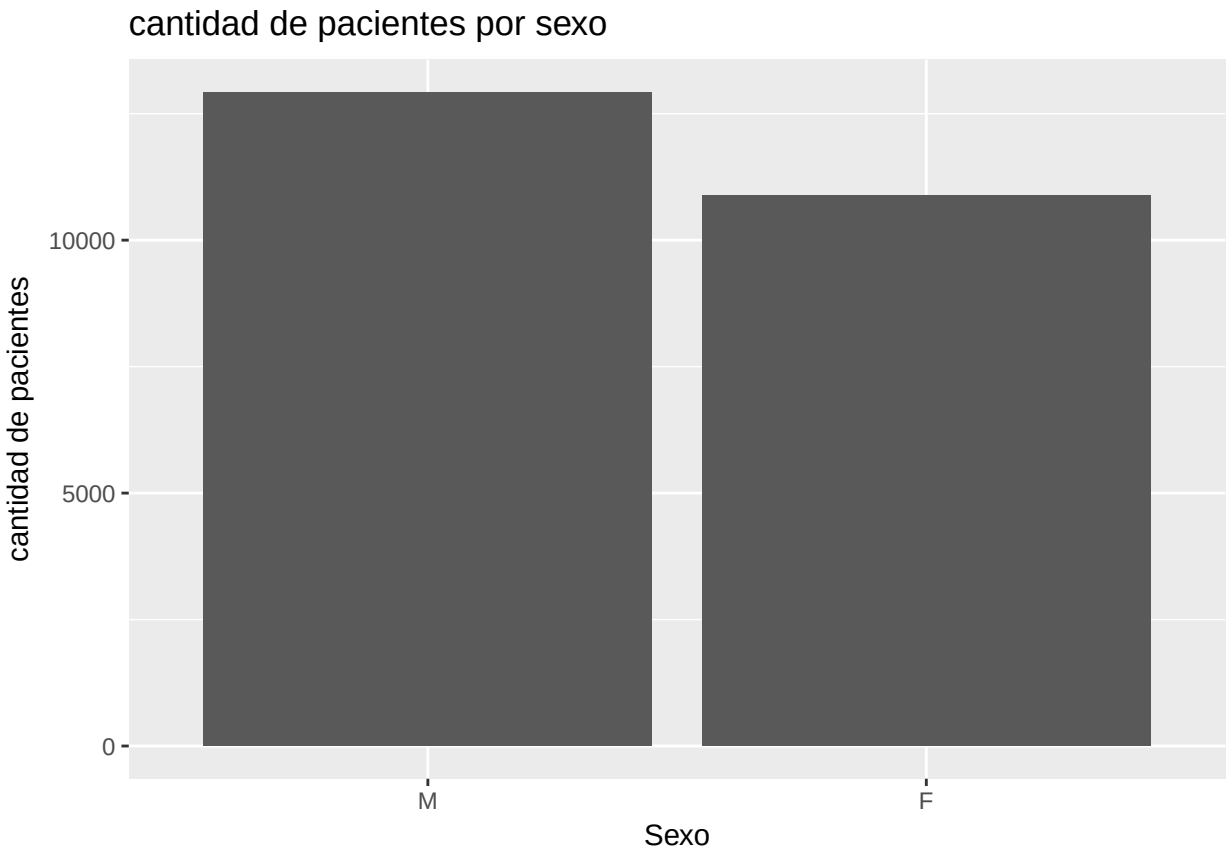
```
ggplot(gastos, aes(x = fct_infreq(Prestador_tipo))) +
    geom_bar() + labs(y = "cantidad de pacientes",
    x = "Tipo de prestador", title = "cantidad de pacientes por tipo de prestador")
```




```
ggplot(gastos, aes(x = fct_infreq(Prestador_departamento))) +
    geom_bar() + labs(x = "Departamento del prestador",
    y = "Cantidad de pacientes", title = "cantidad de pacientes por departamento del prestador") +
    theme(axis.text.x = element_text(angle = 90))
```

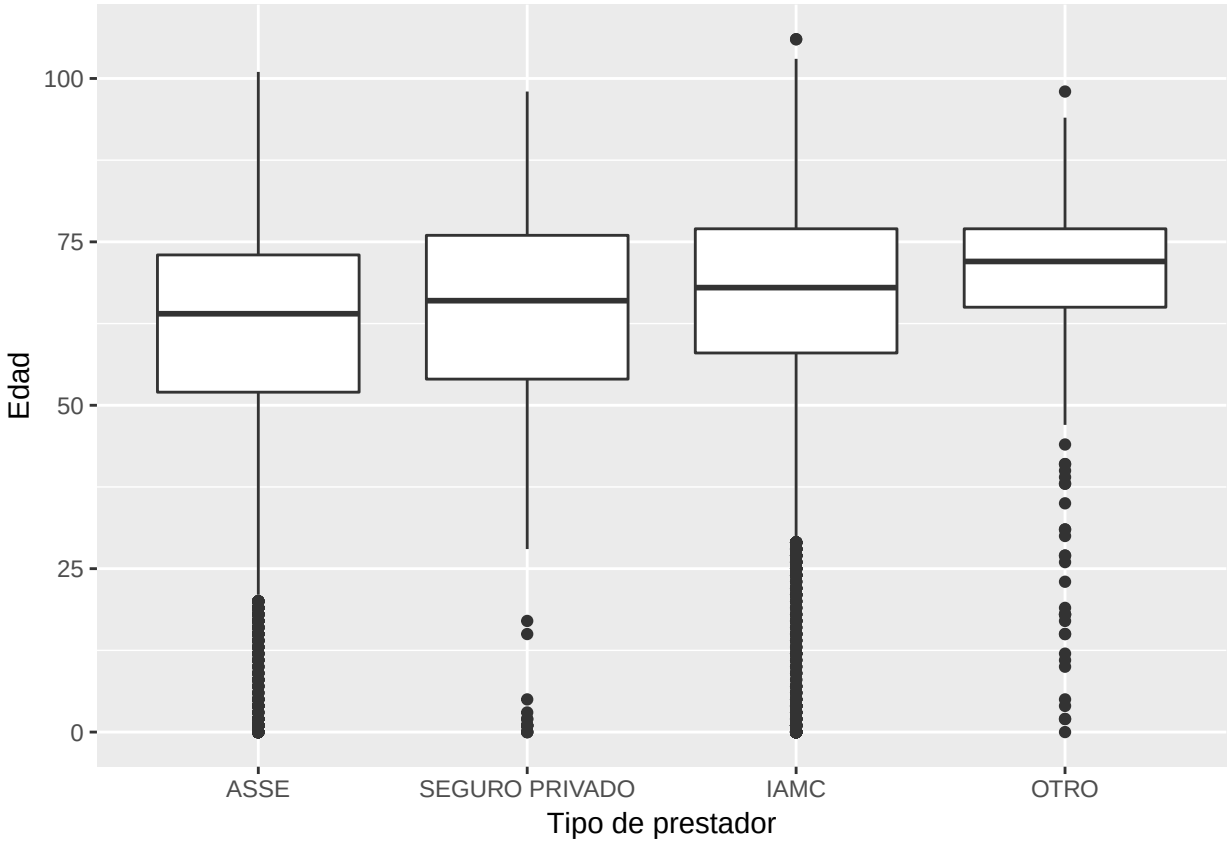



```
ggplot(gastos, aes(x = fct_infreq(Sexo))) + geom_bar() +
    labs(x = "Sexo", y = "cantidad de pacientes", title = "cantidad de pacientes por sexo")
```



Para variables cuantitativas usamos boxplot si queremos mirar la variabilidad.

```
ggplot(gastos, aes(x = fct_reorder(Prestador_tipo,
    Edad_aÃ.os), Edad_aÃ.os)) + geom_boxplot() + labs(x = "Tipo de prestador",
    y = "Edad")
```


100
75
50
25
0
Edad
ASSE
SEGURO PRIVADO
IAMC
OTRO
Tipo de prestador