Tarea 2: Solución

NOMBRE:

3/10/2018

Explicativo

Esta tarea debe ser entregada el miércoles 3 de Octubre . Debe ser realizada en Rmarkdown y debe incluir en el documento el link a tu repositorio personal en GitHub con los archivos fuente para reproducir los resultados. A su vez debe compartir el pdf en EVA.

Los datos que vamos a utilizar en este ejercicio están disponibles en el catálogo de datos abiertos Uruguay https://catalogodatos.gub.uy.

Los datos corresponden a los gastos realizados por actos médicos, **cada fila representa un acto médico**. Los datos y los metadatos se encuentran disponibles:

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

Los pueden leer en R de la siguiente forma:

```
gastolink <- 'https://catalogodatos.gub.uy/dataset/96e636e5-4f78-49a7-8e14-60e90173a0c0/resource/936ac9

gastos <- read.csv(gastolink, header = TRUE, dec = ",",encoding="latin1")
```

Ejercicio 1

Usando las funciones de la librería `dplyr` respondé:

a. ¿Cuál es la prestación con mayor cantidad de actos médicos en Montevideo?

```
library(tidyverse)
gastos %>%
  filter(Prestador_departamento == "MONTEVIDEO") %>%
  group_by(Prestacion) %>%
summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(n = 1)
```

```
## # A tibble: 1 x 2
##   Prestacion                      n
##   <fct>                       <int>
## 1 PCI-Cateterismo izq.adultos  2514
```

b. Creá una variable con los totales de actos médicos por Departamento de residencia (`Departamento_residencia`). ¿Cuál es el departamento de residencia con menor cantidad de actos médicos?

```
library(forcats)

gastos %>%
  group_by(Departamento_residencia) %>%
  summarise(n = n()) %>%
  arrange(n) %>%
  head(n = 1)
```

```
## # A tibble: 1 x 2
##   Departamento_residencia     n
##   <fct>                   <int>
## 1 FLORES                    138
```

c. ¿Qué cantidad de actos médicos son prestados por ASSE o IAMC?

```
gastos %>%
  filter(Prestador_tipo == "ASSE"| Prestador_tipo == "IAMC") %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 22978
```

d. Cada fila representa un acto médico, por lo que puede haber filas que se correspondan con la misma persona. ¿Cómo se puede verificar si esto es así? ¿Cuantos pacientes distintos hay en los datos?

Verificamos que hay 3900 pacientes que están más de una vez.

```
gastos %>%
  filter( duplicated(Paciente) ) %>%
  dim()
```

```
## [1] 3900    9
```

```
gastos %>%
  filter( !duplicated(Paciente) ) %>%
  dim()
```

```
## [1] 19911     9
```

Hay 19911 pacientes distintos en los datos.

e. Crear un **nuevo** conjunto de datos en que cada fila sea un paciente. Agregar dos variables: el gasto total de la persona en actos médicos y la cantidad de visitas. Conservá el resto de las variables originales menos `Prestacion` e `Importe`(Sugerencia usar `summarise_all`).

```
gasto.pp <- gastos %>%
  group_by( Paciente ) %>%
  mutate( importe.total = sum(Importe), visitas = n() ) %>%
  select(-Importe, -Prestacion) %>%
  summarise_all( first )
```

Ejercicio 2

a. Replique el siguiente gráfico (Figura 1) usando `ggplot2` y `forcats` para ordenar el gráfico.

```
library(forcats)

gastos%>%
  group_by(Departamento_residencia) %>%
  summarise(n = n()) %>%
  mutate(prop = n/sum(n, na.rm = TRUE)) %>%
 ggplot(aes(y = fct_reorder(Departamento_residencia,n), x = prop)) +
  geom_point() + labs(x = "Proporción de actos médicos", y =  "Departamento de residencia")
```

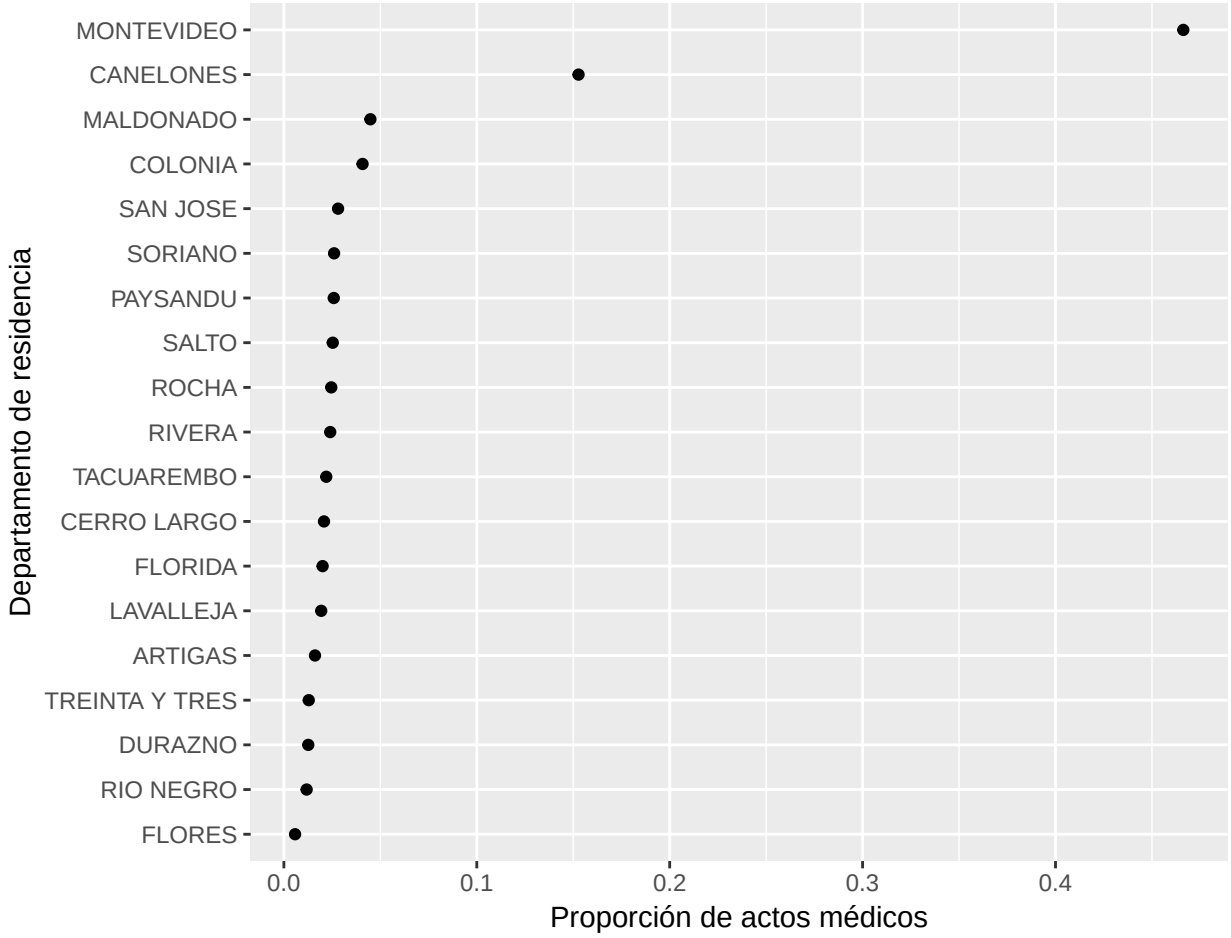


Figura 1: Proporción de actos médicos según departamento de residencia

b. Replique el siguiente gráfico (Figura 2) usando `ggplot2` y `forcats` para ordenar.

```
gastos %>%
  filter(Prestador_departamento == "MONTEVIDEO") %>%
ggplot( aes(x = fct_infreq(Prestacion))) + geom_bar() +
 theme(axis.text.x = element_text(angle = 90, size = 4)) +
  labs(x = "", y = "Total de actos médicos")
```

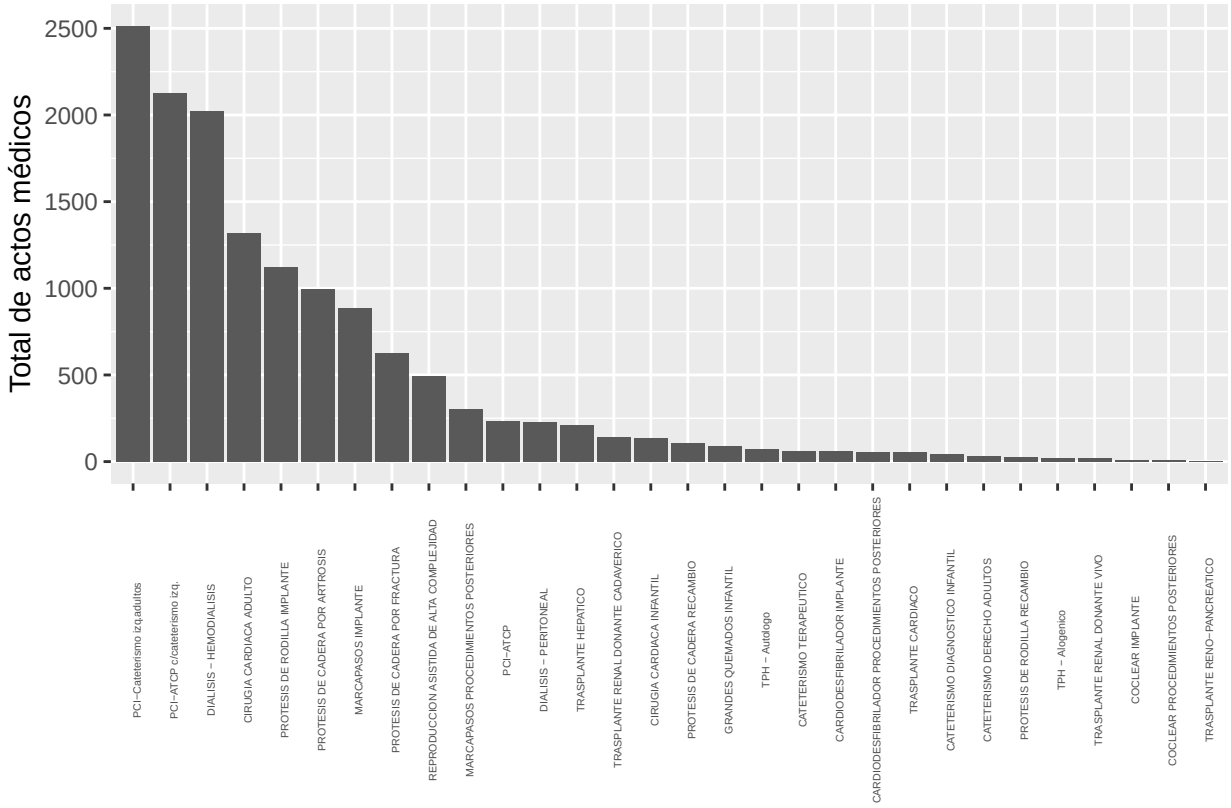


Figura 2: Gráfico de barras de la cantidad de actos médicos en Montevideo por tipo de Prestación

c. Replique el siguiente gráfico (Figura 3) usando `ggplot2` y `forcats` para ordenar.

Notar que el orden de los niveles de `Prestacion` fueron ordenados haciendo:

- una variable auxiliar que vale 1 si `Prestador_tipo` es ASSE y 0 en otro caso
- ordenamos los niveles de prestación según la media de la variable auxiliar
- usamos la función `fct_reorder`

Comente algo interesante que surge de este gráfico.

```
gastos %>%
  mutate( pp = if_else(Prestador_tipo == 'ASSE', 1, 0) ) %>%
      ggplot(aes(x = fct_reorder(Prestacion, pp, .fun = mean) , fill = Prestador_tipo)) +
  geom_bar( position = 'fill') +
  theme(axis.text.y = element_text( hjust = 1,face = 'bold', size = 5),
        legend.position = "bottom") +
  labs(x = 'Tipo de prestación', y = 'Proporción') +
  coord_flip()
```

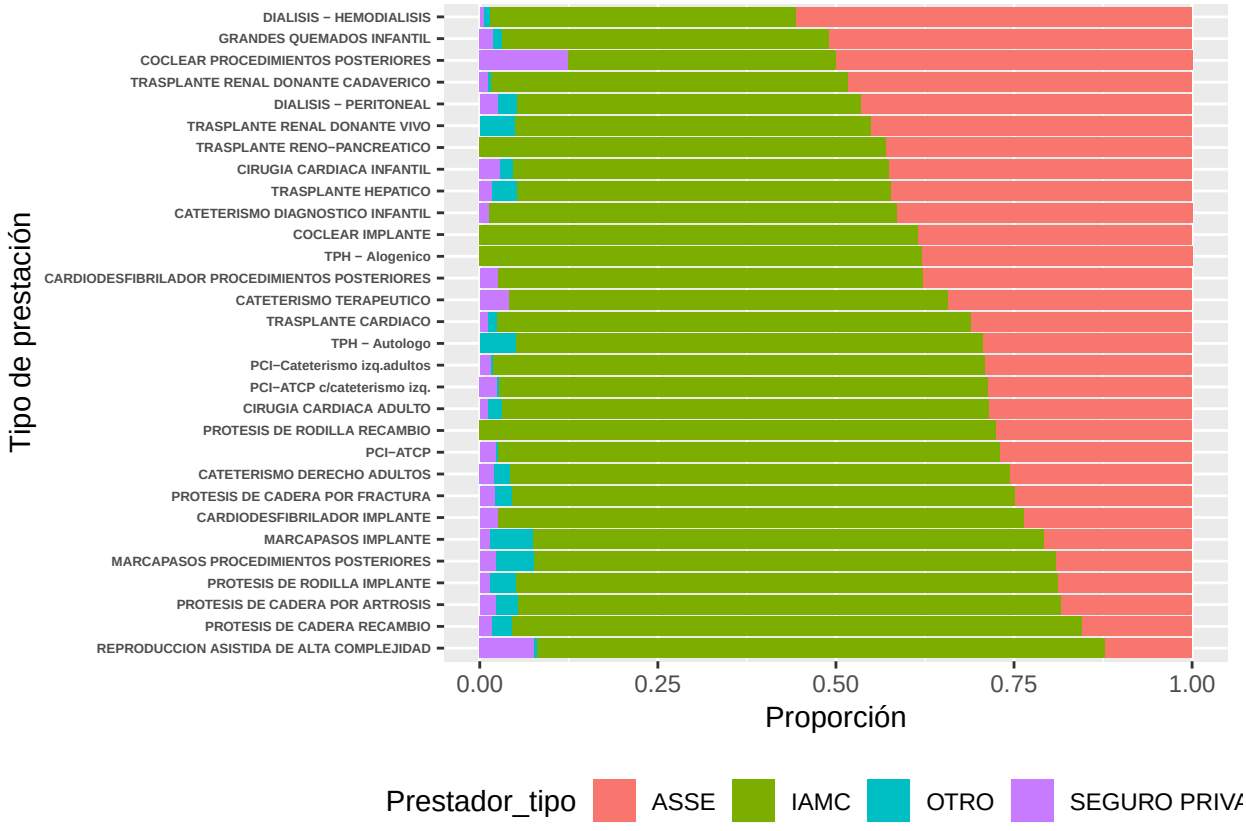


Figura 3: Gráfico de barras apiladas de la cantidad de actos médicos en Montevideo por tipo de Prestacion

Ejercicio 3

a. Usando `ggplot2` elabore una visualización que permita responder la pregunta de ¿Cuáles son las 10 instituciones prestadoras (`Prestador`) que brindaron mayor proporción de actos médicos en Montevideo (`Prestador_deparamento`)?

Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un `caption` (Título) en la figura y algún comentario de interés que describa el gráfico. Puede utilizar `fig.cap` en el chunck de código.

```
gastos %>%
  filter(Prestador_departamento == "MONTEVIDEO") %>%
  group_by(Prestador) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(n = 10) %>%
  mutate(prop = (n/sum(n))) %>%
  ggplot() +
  geom_point(aes(x = fct_reorder(Prestador, prop, .desc = F), y = prop)) + coord_flip() +
  labs(x = "Prestador", y = "Poroporción de actos medicos")
```

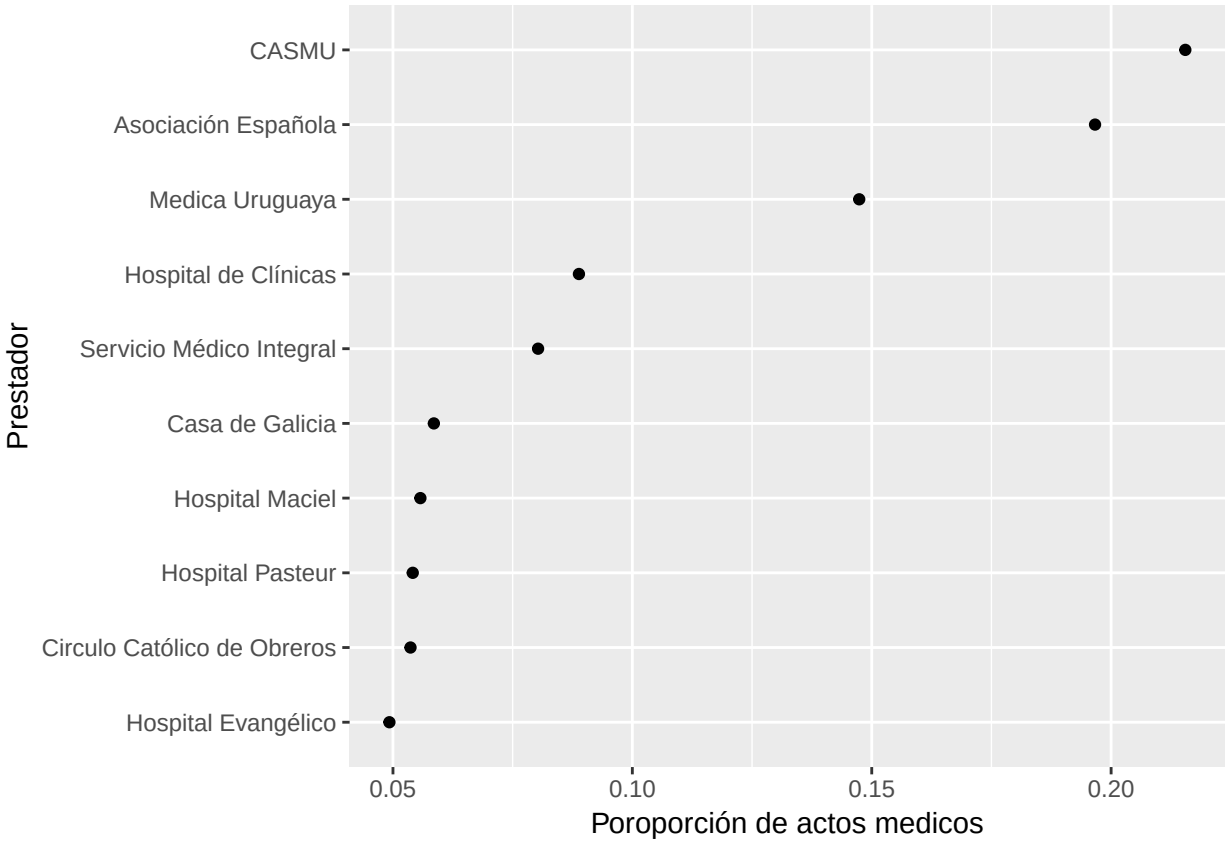


Figura 4: Gráfico de puntos con las 10 instituciones prestadoras más relevantes en Montevideo según proporción de actos médicos

Se puede observar que en Montevideo el CASMU es el prestador más relevante, con una proporción de actos médicos superior al 20 %. En segundo lugar se encuentra la Asociación Española con una particiopación un poco inferior al 20 %.

b. Usando `ggplot2` elabore un gráfico de cajas con el importe del acto médico (en logaritmos) según tipo

de prestador y sexo.

Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un `caption` (Título) en la figura y algún comentario de interés que describa el gráfico. Puede utilizar `fig.cap` en el chunck de código.

```
gastos %>% mutate(lg.imp = log(Importe)) %>%
  ggplot(aes(x = Prestador_tipo, y = lg.imp) ) + geom_boxplot() +
  facet_wrap( ~Sexo) +
  labs(x = "Tipo de Prestador", y = 'Importe (en logs)')
```

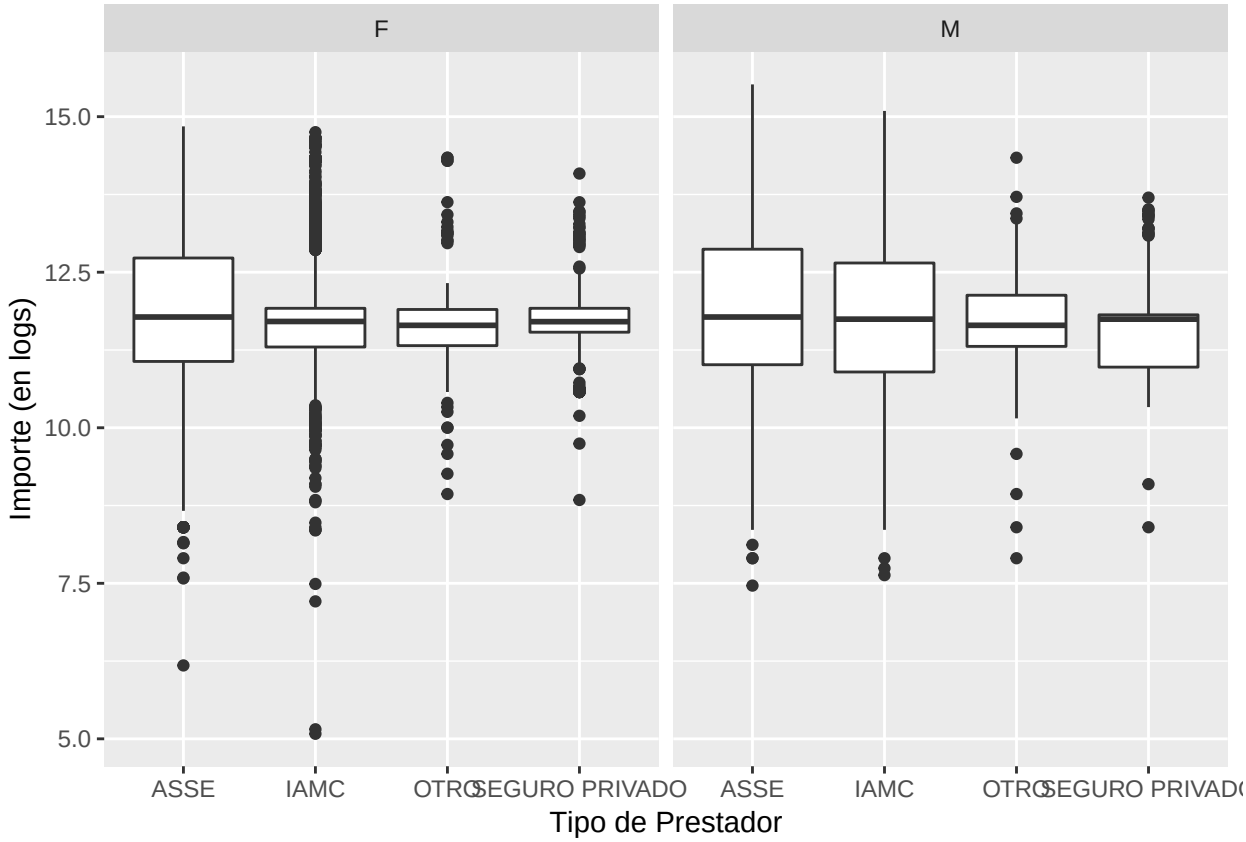


Figura 5: Gráfico de cajas del importe (en log) por tipo de prestador y sexo

Se puede observar que la variabilidad en término de importes es menor para mujeres que para hombres. Por otro lado la mediana de importe (en logs) es similar en todos los casos.

c. Se desea explorar la asociación entre la edad del paciente y el importe de los actos médicos (en logaritmos). Realiza alguna visualización para estudiar dicha asociación, y ver como esta varía según el sexo del paciente y el tipo de prestador.

```
gastos %>%  mutate(lg.imp = log(Importe)) %>%
  ggplot(aes(x = Edad_años, y = lg.imp) ) + geom_point(alpha = 1/5) +
  facet_wrap( ~Sexo) +
  labs(x = "Edad", y = 'Importe (en logs)')
```

Se puede observar que no hay un patrón claro entre las variables analizadas.

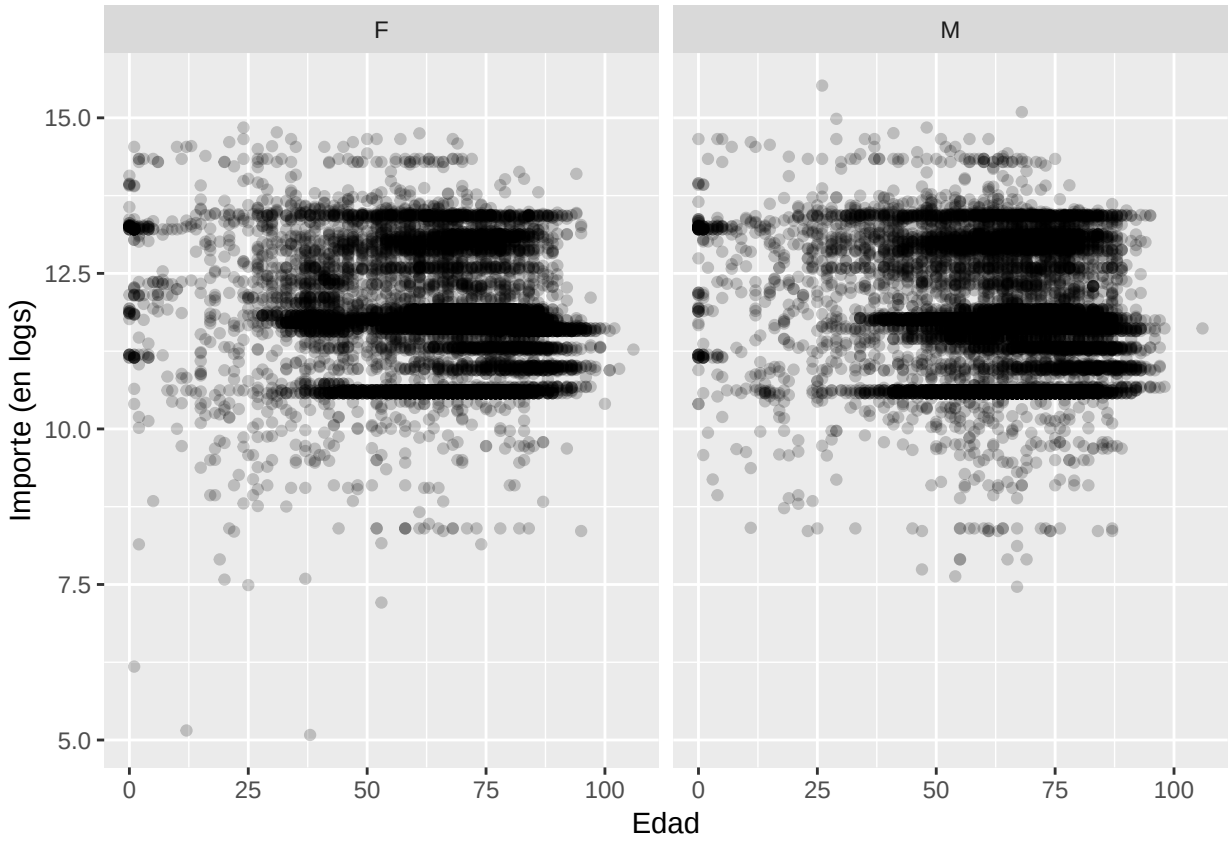


Figura 6: Diagrama de dispersión del importe de los actos médicos según edad y sexo

```
gastos %>%  mutate(lg.imp = log(Importe)) %>%
  ggplot(aes(x = Edad_años, y = lg.imp) ) + geom_point(alpha = 1/5) +
  facet_wrap( ~Prestador_tipo) +
  labs(x = "Edad", y ='Importe (en logs)')
```

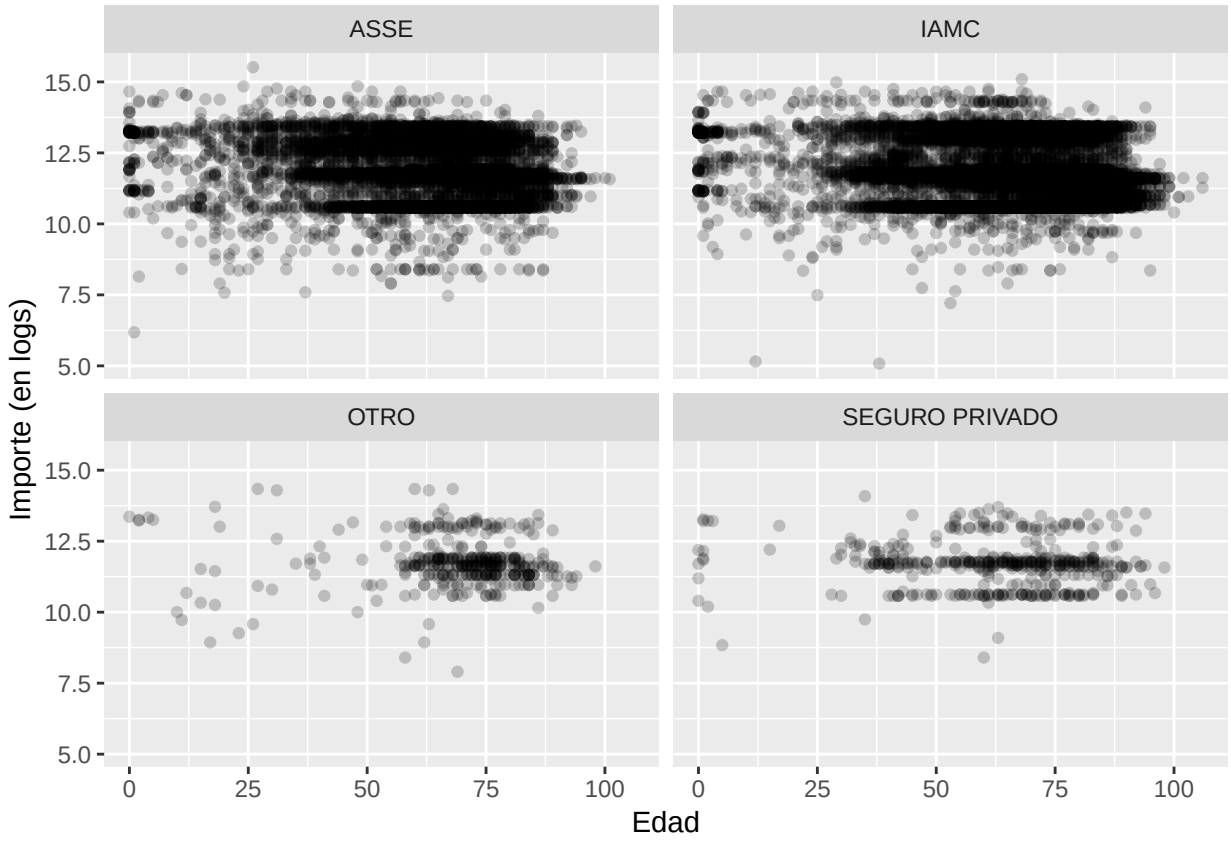
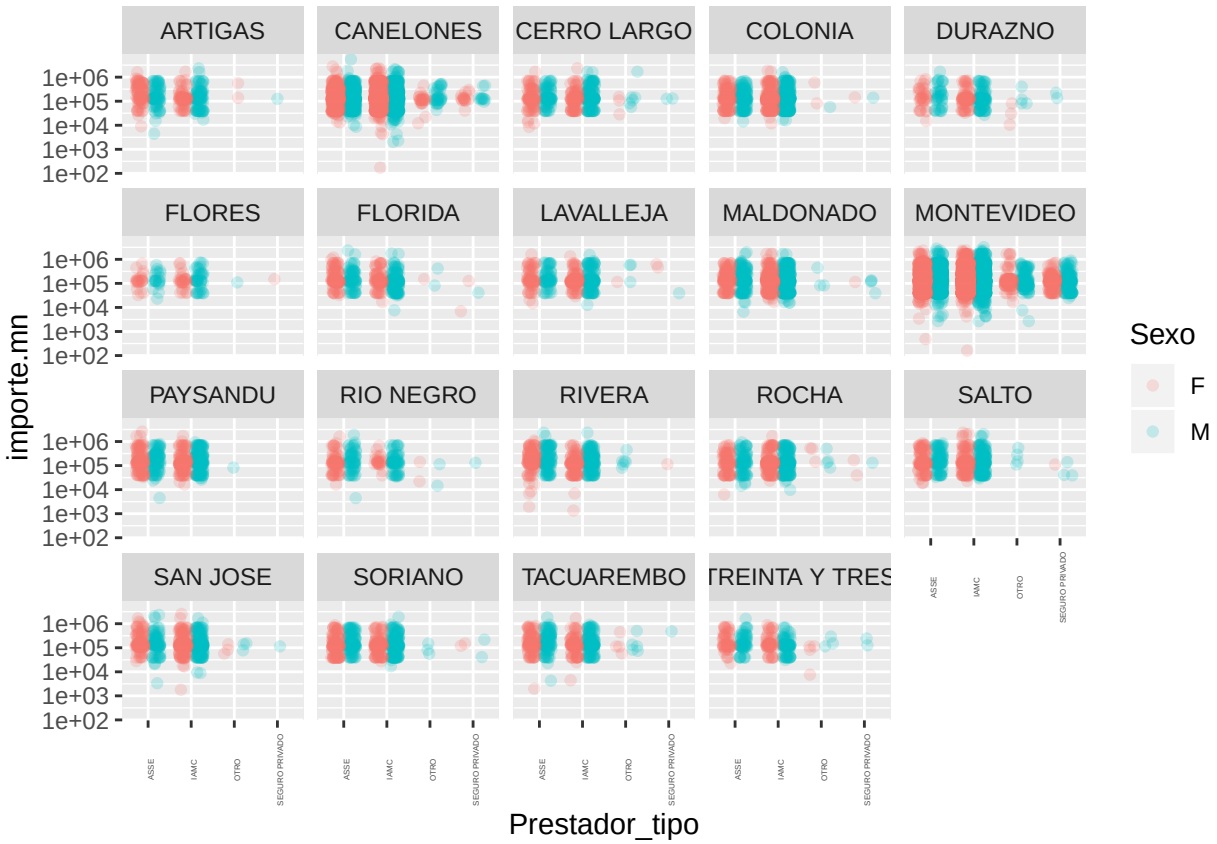


Figura 7: Diagrama de dispersión del importe de los actos médicos según prestador y edad

Se puede observar que en el caso de los seguros privados hay observaciones en 0 años y luego de de los 25 mientras que para ASSE y IAMC la distribucioón es a lo largo de todos las edades. Para ver mejor como se distribuye el importe por edades sería mejor discretizar la variable edades en tramos de edad y hacer gráficos de cajas.

d. Realiza alguna visualización para estudiar el gasto promedio por persona en cada departamento, tipo de prestador y sexo.

```
gasto.pp %>%
  mutate(importe.mn = importe.total / visitas) %>%
  ggplot( aes(x = Prestador_tipo, y = importe.mn, color=Sexo ) ) +
  geom_point(position = position_jitterdodge(), alpha = 1/5) +
  facet_wrap( ~ Departamento_residencia) +
  scale_y_log10() + theme(axis.text.x = element_text(angle = 90, size = 3))
```



e. Realiza alguna visualización para estudiar el peso de las prestaciones en cantidad de actos y en monto relativo. ¿Son las prestaciones más comunes las más caras?

```
gastos %>%
  group_by(Prestacion) %>%
  summarise( gto = sum(Importe), nn = n() ) %>%
  ungroup() %>%
  mutate(gto.tot = sum(gto), prest.tot = sum(nn),
         gto.prop = gto/gto.tot, nn.prop = nn/prest.tot) %>%
  ggplot() +
  geom_point(aes(y = gto.prop, x =  nn.prop)) +
  labs(x = "Proporción de prestaciones", y = "Proporción del gasto")
```

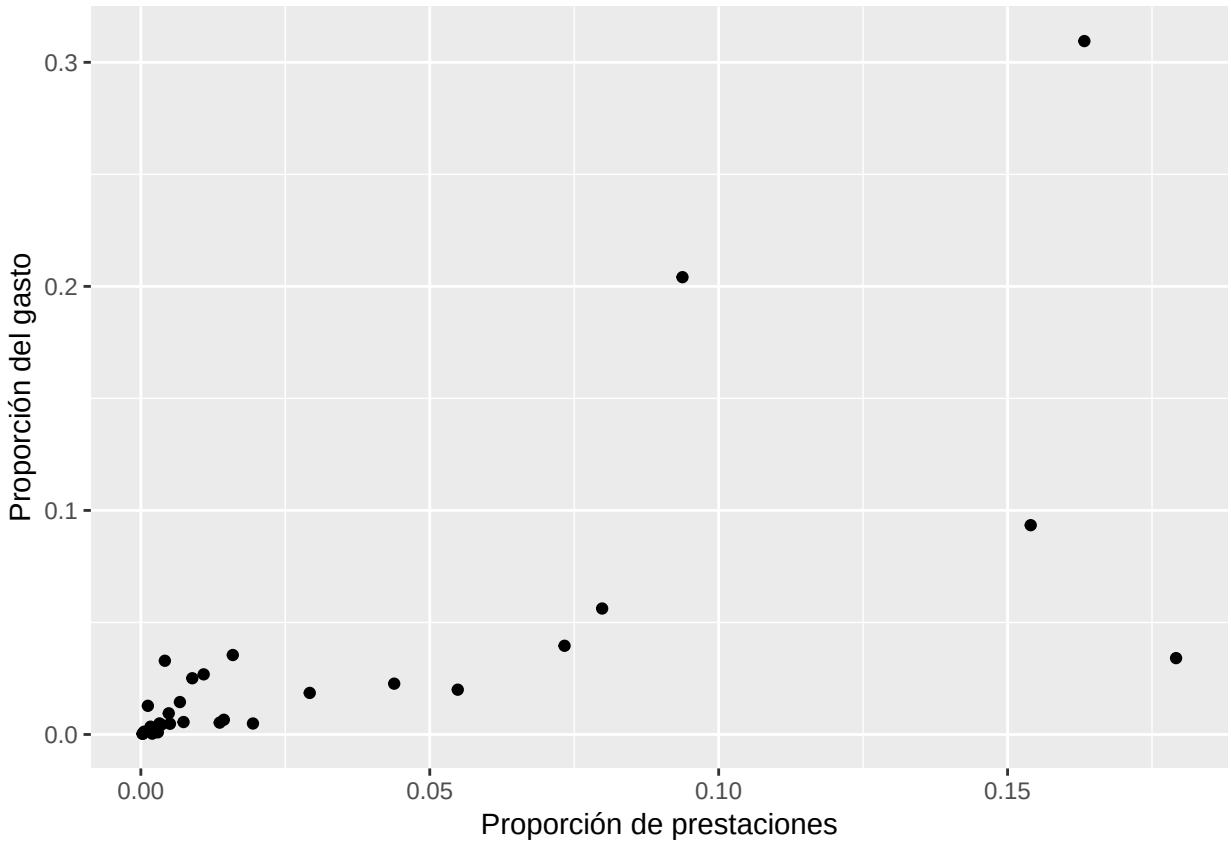


Figura 8: Gráfico de dispersión de proporción de prestaciones y proporción de gastos