

Práctica de Evaluación FAD - Métodos de Análisis de Datos

Isabela Ignacio, Luisa Yáñez, Miguel García

18/12/2021

0. Introducción

La práctica consiste en la elaboración y presentación de un informe de un proyecto de Ciencia de Datos, utilizando las técnicas aprendidas durante el curso, aplicadas a los datos seleccionados.

1. Uso de herramienta de control de versiones

El grupo eligió trabajar en lenguaje R (RStudio version 1.4.1717) y utilizar como herramienta de control de versiones Github. El proyecto “/practica_fd_final” fue creado por Luisa Yáñez (usuario lyanezgu) y compartido con los restantes participantes del grupo Isabela Ignacio (usuario IsaPires1329) y Miguel García (usuario mgarciasanc2021).

2. Conjunto de datos elegido

El conjunto de datos elegido por el grupo se llama “Hospital Charges in America” y incluye información que compara las tarifas de los servicios de hospitalización en diferentes estados de los EEUU para los 100 principales grupos de diagnósticos.

Link del data set: <https://www.kaggle.com/dhirajnirne/hospital-charges-in-america>.

2.1 Paquetes

```
library(readr)
library(ggplot2)
library(GGally)
library(dplyr)
library(tidyr)
library(missForest)
library(VIM)
library(formattable)
library(usmap)
library(cowplot)
```

2.2 Cargar los datos

El conjunto de datos “Hospital Charges in America” contiene 12 columnas y 163065 filas, y lo obtenemos en formato .csv. Inicialmente se han guardado los datos en un data frame llamado “hospital_charges” y se ha realizado un estudio inicial sobre su contenido utilizando la función head y summary.

```
hospital_charges <- read_csv("notebooks/hospital-charges.csv")  
hospital_charges
```

```
## # A tibble: 163,065 x 12  
##   'DRG Definition' 'Provider Id' 'Provider Name' 'Provider Street ~  
##   <chr>           <dbl> <chr>          <chr>  
## 1 039 - EXTRACRANIAL PRO~ 10001 SOUTHEAST ALABAMA M~ 1108 ROSS CLARK C~  
## 2 039 - EXTRACRANIAL PRO~ 10005 MARSHALL MEDICAL CE~ 2505 U S HIGHWAY ~  
## 3 039 - EXTRACRANIAL PRO~ 10006 ELIZA COFFEE MEMORI~ 205 MARENGO STREET~  
## 4 039 - EXTRACRANIAL PRO~ 10011 ST VINCENT'S EAST    50 MEDICAL PARK E~  
## 5 039 - EXTRACRANIAL PRO~ 10016 SHELBY BAPTIST MEDI~ 1000 FIRST STREET~  
## 6 039 - EXTRACRANIAL PRO~ 10023 BAPTIST MEDICAL CEN~ 2105 EAST SOUTH B~  
## 7 039 - EXTRACRANIAL PRO~ 10029 EAST ALABAMA MEDICA~ 2000 PEPPERELL PA~  
## 8 039 - EXTRACRANIAL PRO~ 10033 UNIVERSITY OF ALABA~ 619 SOUTH 19TH ST~  
## 9 039 - EXTRACRANIAL PRO~ 10039 HUNTSVILLE HOSPITAL 101 SIVLEY RD  
## 10 039 - EXTRACRANIAL PRO~ 10040 GADSDEN REGIONAL ME~ 1007 GOODYEAR AVE~  
## # ... with 163,055 more rows, and 8 more variables: Provider City <chr>,  
## #   Provider State <chr>, Provider Zip Code <dbl>,  
## #   Hospital Referral Region Description <chr>, Total Discharges <dbl>,  
## #   Average Covered Charges <chr>, Average Total Payments <chr>,  
## #   Average Medicare Payments <chr>
```

```
head(hospital_charges)
```

```
## # A tibble: 6 x 12  
##   'DRG Definition' 'Provider Id' 'Provider Name' 'Provider Street ~  
##   <chr>           <dbl> <chr>          <chr>  
## 1 039 - EXTRACRANIAL PROC~ 10001 SOUTHEAST ALABAMA M~ 1108 ROSS CLARK C~  
## 2 039 - EXTRACRANIAL PROC~ 10005 MARSHALL MEDICAL CE~ 2505 U S HIGHWAY ~  
## 3 039 - EXTRACRANIAL PROC~ 10006 ELIZA COFFEE MEMORI~ 205 MARENGO STREET~  
## 4 039 - EXTRACRANIAL PROC~ 10011 ST VINCENT'S EAST    50 MEDICAL PARK E~  
## 5 039 - EXTRACRANIAL PROC~ 10016 SHELBY BAPTIST MEDI~ 1000 FIRST STREET~  
## 6 039 - EXTRACRANIAL PROC~ 10023 BAPTIST MEDICAL CEN~ 2105 EAST SOUTH B~  
## # ... with 8 more variables: Provider City <chr>, Provider State <chr>,  
## #   Provider Zip Code <dbl>, Hospital Referral Region Description <chr>,  
## #   Total Discharges <dbl>, Average Covered Charges <chr>,  
## #   Average Total Payments <chr>, Average Medicare Payments <chr>
```

```
summary(hospital_charges)
```

```
## DRG Definition      Provider Id      Provider Name      Provider Street Address  
## Length:163065      Min.    : 10001      Length:163065      Length:163065  
## Class :character   1st Qu.:110092     Class :character   Class :character  
## Mode  :character   Median :250007     Mode  :character   Mode  :character  
##                           Mean   :255570  
##                           3rd Qu.:380075
```

```

##          Max.    :670077
## Provider City      Provider State      Provider Zip Code
## Length:163065      Length:163065      Min.    : 1040
## Class  :character  Class  :character  1st Qu.:27261
## Mode   :character  Mode   :character  Median  :44309
##                               Mean    :47938
##                               3rd Qu.:72901
##                               Max.   :99835
## Hospital Referral Region Description Total Discharges  Average Covered Charges
## Length:163065          Min.    : 11.00  Length:163065
## Class  :character      1st Qu.: 17.00  Class  :character
## Mode   :character      Median : 27.00  Mode   :character
##                               Mean   : 42.78
##                               3rd Qu.: 49.00
##                               Max.  :3383.00
## Average Total Payments Average Medicare Payments
## Length:163065          Length:163065
## Class  :character      Class  :character
## Mode   :character      Mode   :character
## 
## 
## 
```

3. Detección, tratamiento e imputación de datos faltantes

A través de la función summary empezamos comprobando que no hay datos faltantes en el data set. Por ello el grupo ha tenido que añadirlos manualmente para aproximarnos a un caso más real donde lo normal es encontrarlos y tener que lidiar con ellos. Los datos faltantes han sido imputados exclusivamente en las columnas que no van a servir de análisis principal para este estudio, para así intentar que la predicción que hagamos sea lo más precisa posible.

Utilizamos la librería missForest y generamos una semilla para que el resultado sea siempre el mismo.

```

set.seed(101)
hospital_charges <- bind_cols(hospital_charges[c(1, 3, 5, 7,
  8, 9, 11, 12)], missForest::prodNA(hospital_charges[c(-1,
  -3, -5, -7, -8, -9, -11, -12)], noNA = 0.1))

hospital_charges

## # A tibble: 163,065 x 12
##   'DRG Definition'      'Provider Name'      'Provider City' 'Provider Zip C~
##   <chr>                  <chr>                  <chr>                <dbl>
## 1 039 - EXTRACRANIAL PRO~ SOUTHEAST ALABAMA M~ DOTHAN            36301
## 2 039 - EXTRACRANIAL PRO~ MARSHALL MEDICAL CE~ BOAZ              35957
## 3 039 - EXTRACRANIAL PRO~ ELIZA COFFEE MEMORI~ FLORENCE          35631
## 4 039 - EXTRACRANIAL PRO~ ST VINCENT'S EAST     BIRMINGHAM        35235
## 5 039 - EXTRACRANIAL PRO~ SHELBY BAPTIST MEDI~ ALABASTER         35007
## 6 039 - EXTRACRANIAL PRO~ BAPTIST MEDICAL CEN~ MONTGOMERY        36116
## 7 039 - EXTRACRANIAL PRO~ EAST ALABAMA MEDICA~ OPELIKA           36801
## 8 039 - EXTRACRANIAL PRO~ UNIVERSITY OF ALABA~ BIRMINGHAM        35233
## 9 039 - EXTRACRANIAL PRO~ HUNTSVILLE HOSPITAL HUNTSVILLE       35801

```

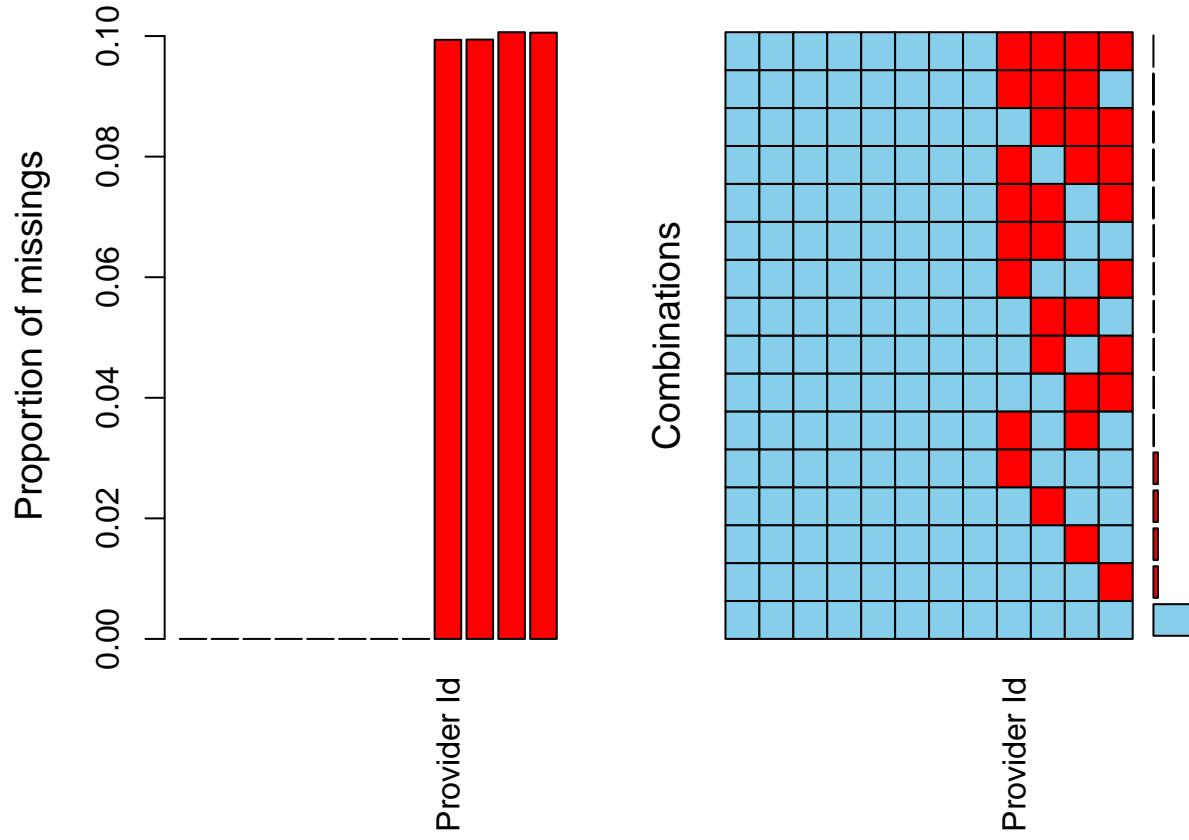
```

## 10 039 - EXTRACRANIAL PRO~ GADSDEN REGIONAL ME~ GADSDEN          35903
## # ... with 163,055 more rows, and 8 more variables:
## #   Hospital Referral Region Description <chr>, Total Discharges <dbl>,
## #   Average Total Payments <chr>, Average Medicare Payments <chr>,
## #   Provider Id <dbl>, Provider Street Address <chr>, Provider State <chr>,
## #   Average Covered Charges <chr>

```

Haciendo uso de la librería VIM, analizamos un poco la estructura que tienen nuestros datos faltantes dentro de nuestra data set para ver y entender como se distribuyen.

```
summary(aggr(hospital_charges))
```



```

##
## Missings per variable:
##                               Variable Count
## DRG Definition           0
## Provider Name             0
## Provider City             0
## Provider Zip Code         0
## Hospital Referral Region Description  0
## Total Discharges          0
## Average Total Payments    0
## Average Medicare Payments 0
## Provider Id 16205          0
## Provider Street Address 16211 0

```

```

##                                     Provider State 16411
##                               Average Covered Charges 16399
##
## Missings in combinations of variables:
##          Combinations   Count    Percent
## 0:0:0:0:0:0:0:0:0:0:0:0 106990 65.61187257
## 0:0:0:0:0:0:0:0:0:0:0:1 11936  7.31978046
## 0:0:0:0:0:0:0:0:0:0:1:0 11914  7.30628890
## 0:0:0:0:0:0:0:0:0:0:1:1 1371   0.84076902
## 0:0:0:0:0:0:0:0:0:1:0:0 11864  7.27562628
## 0:0:0:0:0:0:0:0:0:1:0:1 1327   0.81378591
## 0:0:0:0:0:0:0:0:0:1:1:0 1321   0.81010640
## 0:0:0:0:0:0:0:0:0:1:1:1 137   0.08401558
## 0:0:0:0:0:0:0:0:0:1:0:0 11814  7.24496366
## 0:0:0:0:0:0:0:0:0:1:0:1 1310   0.80336062
## 0:0:0:0:0:0:0:0:0:1:0:1:0 1377   0.84444853
## 0:0:0:0:0:0:0:0:0:1:0:1:1 142   0.08708184
## 0:0:0:0:0:0:0:0:0:1:1:0:0 1257   0.77085825
## 0:0:0:0:0:0:0:0:0:1:1:0:1 156   0.09566737
## 0:0:0:0:0:0:0:0:0:1:1:1:0 129   0.07910956
## 0:0:0:0:0:0:0:0:0:1:1:1:1 20    0.01226505

```

```
# referencia https://rpubs.com/sediaz/na\_aggr
```

4. Partición del conjunto de datos: data set training y data set test

Una vez vistos por encima la estructura general de los datos, y habiendo añadido los datos faltantes que nos hacían falta, pasamos a dividir el conjunto de datos en dos, para diferenciar los que usaremos de entrenamiento de los que usaremos de test (viendo la cantidad de datos de la que disponemos, la distribución elegida ha sido: 20% test y 80% training). Establecemos una semilla que nos guarde de forma permanente la división que hacemos, para que la división de los datos sea siempre la misma.

Guardamos además la partición de datos de test para ser utilizada a futuro para la validación del modelo final, y pasamos a trabajar de aquí en adelante con la partición de training.

```

set.seed(101)
sample <- sample.int(n = nrow(hospital_charges), size = floor(0.8 *
  nrow(hospital_charges)), replace = F)
train <- hospital_charges[sample, ]
test <- hospital_charges[-sample, ]

train

## # A tibble: 130,452 x 12
##   'DRG Definition'      'Provider Name'   'Provider City' 'Provider Zip C-
##   <chr>                  <chr>           <chr>           <dbl>
## 1 064 - INTRACRANIAL HEMOR~ RIVERVIEW HOSPITAL NOBLESVILLE        46060
## 2 439 - DISORDERS OF PANCR~ ST LUKE'S ROOSEVE~ NEW YORK            10025
## 3 853 - INFECTIOUS & PARAS~ ST JOSEPH'S MEDIC~ YONKERS           10701
## 4 329 - MAJOR SMALL & LARG~ UNIVERSITY OF KAN~ KANSAS CITY         66160
## 5 195 - SIMPLE PNEUMONIA &~ GARDEN CITY HOSPI~ GARDEN CITY         48135
## 6 176 - PULMONARY EMBOLISM~ HORIZON MEDICAL C~ DICKSON            37055

```

```

## 7 641 - MISC DISORDERS OF ~ BAYLOR UNIVERSITY~ DALLAS 75246
## 8 638 - DIABETES W CC ST ELIZABETH FLOR~ FLORENCE 41042
## 9 872 - SEPTICEMIA OR SEVE~ ST JOSEPH'S HOSPI~ SYRACUSE 13203
## 10 439 - DISORDERS OF PANCR~ SOUTH POINTE HOSP~ WARRENSVILLE HE 44122
## # ... with 130,442 more rows, and 8 more variables:
## #   Hospital Referral Region Description <chr>, Total Discharges <dbl>,
## #   Average Total Payments <chr>, Average Medicare Payments <chr>,
## #   Provider Id <dbl>, Provider Street Address <chr>, Provider State <chr>,
## #   Average Covered Charges <chr>

test

## # A tibble: 32,613 x 12
##   'DRG Definition'     'Provider Name'    'Provider City' 'Provider Zip C~
##   <chr>                <chr>            <chr>           <dbl>
## 1 039 - EXTRACRANIAL PR~ MARSHALL MEDICAL CEN~ BOAZ      35957
## 2 039 - EXTRACRANIAL PR~ SOUTH BALDWIN REGION~ FOLEY      36535
## 3 039 - EXTRACRANIAL PR~ MOBILE INFIRMARY      MOBILE      36652
## 4 039 - EXTRACRANIAL PR~ TUCSON MEDICAL CENTER TUCSON      85712
## 5 039 - EXTRACRANIAL PR~ CARONDELET ST JOSEPH~ TUCSON      85711
## 6 039 - EXTRACRANIAL PR~ ST JOSEPH'S HOSPITAL~ PHOENIX      85013
## 7 039 - EXTRACRANIAL PR~ BANNER BOSWELL MEDIC~ SUN CITY      85351
## 8 039 - EXTRACRANIAL PR~ SUMMIT HEALTHCARE RE~ SHOW LOW      85901
## 9 039 - EXTRACRANIAL PR~ BANNER HEART HOSPITAL MESA      85206
## 10 039 - EXTRACRANIAL PR~ CONWAY REGIONAL MEDI~ CONWAY     72034
## # ... with 32,603 more rows, and 8 more variables:
## #   Hospital Referral Region Description <chr>, Total Discharges <dbl>,
## #   Average Total Payments <chr>, Average Medicare Payments <chr>,
## #   Provider Id <dbl>, Provider Street Address <chr>, Provider State <chr>,
## #   Average Covered Charges <chr>

```

5. EDA - Análisis exploratorio de datos

5.1 Definición de las variables que componen los datos de estudio

Empezando ya el análisis en profundidad del conjunto de datos que tenemos, vemos que las 12 variables que componen los datos pueden ser descritas como:

- **DRG Definition:** Grupo relativo a un diagnóstico. Los grupos de diagnóstico relacionado (DRG) se utilizan para clasificar la gravedad de la enfermedad en las visitas hospitalarias de pacientes hospitalizados, el riesgo de mortalidad, el pronóstico, la dificultad del tratamiento, la necesidad de intervención y la intensidad de los recursos que necesitan. El sistema DRG fue desarrollado en la Universidad de Yale en la década de 1970 para la clasificación estadística de casos hospitalarios. Realmente la variable DRG es relativa al código y la descripción que identifican el MS-DRG. Los MS-DRG son un sistema de clasificación que agrupa condiciones clínicas similares (diagnósticos) y los procedimientos proporcionados por el hospital durante la estancia. El sistema de Medicare (Sistema de Seguridad Social en EEUU) los utiliza para determinar los reembolsos para hospitales, centros de enfermería especializada y hospicios. Una estadía en el hospital puede variar de un día a 100 días. Los MS-DRG más caros tienen las estadías promedio más largas. El establecimiento del cada DRG se establece según las condiciones clínicas del paciente, necesidad de cantidades similares de recursos para pacientes hospitalizados y sexo y edad del paciente. Para ello se utiliza el sistema de DRG llamado “Medicare Severity DRGs (MS-DRGs)” para reflejar en mejor manera la severidad de la enfermedad del paciente y su consumo

de recursos para su recuperación. Para clasificar el nivel de severidad de un paciente dentro del sistema “MS-DRGs” hay códigos secundarios de diagnóstico:

- MCC: Major Complication/Comorbidity -> El nivel más alto de severidad.
- CC: Complication/Comorbidity -> El siguiente nivel de severidad.
- Non-CC: Non-Complication/Comorbidity -> Este nivel no supone una gran severidad en la enfermedad ni un gran gasto de recursos;

- **Provider ID:** ID o número identificativo de referencia del hospital;
- **Provider Name:** Nombre del hospital;
- **Provider Street Address:** Dirección postal donde se ubica el hospital;
- **Provider City:** Ciudad donde se ubica el hospital;
- **Provider State:** Estado federal de EEUU donde se ubica el hospital;
- **Provider Zip Code:** Código postal donde se ubica el hospital;
- **Hospital Referral Region Description:** Delinación geográfica específica creada por la organización norteamericana “Dartmouth Atlas of Health Care”, para estudiar los mercados vinculados al sector salud en EEUU;
- **Total Discharges:** Número de personas dadas de alta;
- **Average Covered Charges:** Gastos medios del hospital por los servicios cubiertos por la seguridad social para todas las altas del grupo relacionado con el diagnóstico. Por lo tanto cargo promedio según grupo de diagnóstico DRG establecido. Los pacientes que tienen características clínicas similares y costos de tratamiento similares se asignan a un Grupo de Diagnóstico Relacionado (DRG). El DRG está vinculado a un monto de pago fijo basado en el costo promedio del tratamiento de los pacientes del grupo. La asignación de DRG se basa en el diagnóstico del paciente, los procedimientos recibidos, la edad y otra información. Por lo tanto esta variable contiene el cargo promedio por cada DRG proporcionado por el hospital. Sus cargos promedio podrían ser más o menos dependiendo de las necesidades específicas de su paciente y los servicios prestados. Esto es lo que el hospital cobra en la factura final del hospital y es equivalente al “sticker price”. Este es en gran medida un número irrelevante, ya que no importa lo que cobren los diferentes hospitales, a todos se les pagará la misma cantidad de Medicare por cualquier DRG dado. Prácticamente nadie paga el “sticker price” en un hospital. Cuando un paciente ha sido admitido como hospitalizado en un hospital, ese hospital asigna un DRG cuando este paciente es dado de alta, basándolo en la atención que necesitaba durante su estadía en el hospital. Al hospital se le paga una cantidad fija por ese DRG, independientemente de cuánto dinero realmente gaste en su tratamiento. Si un hospital puede tratar a un paciente de forma efectiva por menos dinero del que Medicare paga por su DRG, entonces el hospital gana dinero con esa hospitalización. Si el hospital gasta más dinero cuidando del paciente de lo que Medicare le da para su DRG, entonces el hospital pierde dinero en esa hospitalización;
- **Average Medicare Payments:** Importe medio cubierto por la Seguridad Social de EEUU. Esto es lo que Medicare paga al hospital por ese DRG;
- **Average Total Payments:** Importe medio total a pagar por persona. Esto es lo que realmente se le paga al hospital e incluye lo que paga Medicare más los copagos que paga el paciente más cualquier cosa que pague el seguro secundario (seguro privado).

5.2. Definición de objetivos

El objetivo final del proyecto es llegar a un modelo que permita recomendar cual es el hospital o grupo de hospitales óptimo que debe elegir un paciente enfermo en EEUU, en base a la posible enfermedad que le

van a diagnosticar, su localización geográfica y los costes que su caso clínico puede llegar a tener en base al sistema sanitario estadounidense.

Para esta primera entrega, el objetivo es realizar el tratamiento de datos adecuado y seleccionar las mejores variables que servirán para llegar al modelo de Machine Learning deseado. Se realizará de la misma manera un ajuste, interpretación y diagnosis del modelo de regresión lineal múltiple, en base a las variables que mejor expliquen los datos.

5.3. Transformaciones de variables cuantitativas y procesado de variables cualitativas - Limpieza de datos

5.3.1 Cambiar los nombres de las columnas

Se ha decidido realizar un cambio en el nombre de las variables que aparecen en las columnas de los datos para así seguir un mismo patrón y a al vez evitar tener espacios que nos pueden llegar a dar problemas a futuro.

```
train
```

```
## # A tibble: 130,452 x 12
##   'DRG Definition'      'Provider Name'    'Provider City'  'Provider Zip C-
##   <chr>                  <chr>            <chr>           <dbl>
## 1 064 - INTRACRANIAL HEMOR~ RIVERVIEW HOSPITAL NOBLESVILLE 46060
## 2 439 - DISORDERS OF PANCR~ ST LUKE'S ROOSEVE~ NEW YORK 10025
## 3 853 - INFECTIOUS & PARAS~ ST JOSEPH'S MEDIC~ YONKERS 10701
## 4 329 - MAJOR SMALL & LARG~ UNIVERSITY OF KAN~ KANSAS CITY 66160
## 5 195 - SIMPLE PNEUMONIA &~ GARDEN CITY HOSPI~ GARDEN CITY 48135
## 6 176 - PULMONARY EMBOLISM~ HORIZON MEDICAL C~ DICKSON 37055
## 7 641 - MISC DISORDERS OF ~ BAYLOR UNIVERSITY~ DALLAS 75246
## 8 638 - DIABETES W CC     ST ELIZABETH FLOR~ FLORENCE 41042
## 9 872 - SEPTICEMIA OR SEVE~ ST JOSEPH'S HOSPI~ SYRACUSE 13203
## 10 439 - DISORDERS OF PANCR~ SOUTH POINTE HOSP~ WARRENSVILLE HE 44122
## # ... with 130,442 more rows, and 8 more variables:
## #   Hospital Referral Region Description <chr>, Total Discharges <dbl>,
## #   Average Total Payments <chr>, Average Medicare Payments <chr>,
## #   Provider Id <dbl>, Provider Street Address <chr>, Provider State <chr>,
## #   Average Covered Charges <chr>
```

```
names(train) <- c("drg_def", "prov_name", "prov_city", "prov_zip",
  "referral_reg", "total_discharges", "mean_total_payments",
  "mean_medicare_payments", "prov_id", "prov_address", "prov_state",
  "mean_covered_charges")
```

```
head(train)
```

```
## # A tibble: 6 x 12
##   drg_def      prov_name    prov_city  prov_zip referral_reg total_discharges
##   <chr>        <chr>       <chr>      <dbl> <chr>          <dbl>
## 1 064 - INTRACRA~ RIVERVIEW H~ NOBLESVIL~ 46060 IN - Indian~         11
## 2 439 - DISORDER~ ST LUKE'S R~ NEW YORK    10025 NY - Manhat~        13
## 3 853 - INFECTIO~ ST JOSEPH'S~ YONKERS   10701 NY - White ~        17
```

```

## 4 329 - MAJOR SM~ UNIVERSITY ~ KANSAS CI~      66160 MO - Kansas~        44
## 5 195 - SIMPLE P~ GARDEN CITY~ GARDEN CI~      48135 MI - Dearbo~        52
## 6 176 - PULMONAR~ HORIZON MED~ DICKSON       37055 TN - Nashvi~        14
## # ... with 6 more variables: mean_total_payments <chr>,
## #   mean_medicare_payments <chr>, prov_id <dbl>, prov_address <chr>,
## #   prov_state <chr>, mean_covered_charges <chr>

```

5.3.2 División de la columna drg_def

Realizamos una división de la columna “drg_ref”. Separamos la columna en dos diferenciando entre código de la enfermedad y descripción de la enfermedad. Nos servirá a futuro para simplificar el análisis y visualización de los datos de interés.

```

#test <- data.frame(x = train$drg_def)

train <- train %>%
  separate(data = ., col = drg_def,
           into = c("codigo_enf", "desc_enf"), sep = "-")
train

## # A tibble: 130,452 x 13
##   codigo_enf desc_enf          prov_name    prov_city prov_zip referral_reg
##   <chr>      <chr>          <chr>        <chr>      <dbl> <chr>
## 1 "064 "     " INTRACRANIAL HEM~ RIVERVIEW HOS~ NOBLESVI~  46060 IN - Indian-
## 2 "439 "     " DISORDERS OF PAN~ ST LUKE'S ROO~ NEW YORK   10025 NY - Manhat-
## 3 "853 "     " INFECTIOUS & PAR~ ST JOSEPH'S M~ YONKERS   10701 NY - White ~
## 4 "329 "     " MAJOR SMALL & LA~ UNIVERSITY OF~ KANSAS C~  66160 MO - Kansas~
## 5 "195 "     " SIMPLE PNEUMONIA~ GARDEN CITY H~ GARDEN C~  48135 MI - Dearbo~
## 6 "176 "     " PULMONARY EMBOLI~ HORIZON MEDIC~ DICKSON   37055 TN - Nashvi~
## 7 "641 "     " MISC DISORDERS O~ BAYLOR UNIVER~ DALLAS   75246 TX - Dallas
## 8 "638 "     " DIABETES W CC" ST ELIZABETH ~ FLORENCE  41042 KY - Coving-
## 9 "872 "     " SEPTICEMIA OR SE~ ST JOSEPH'S H~ SYRACUSE  13203 NY - Syracu~
## 10 "439 "    " DISORDERS OF PAN~ SOUTH POINTE ~ WARRENSV~ 44122 OH - Clevel-
## # ... with 130,442 more rows, and 7 more variables: total_discharges <dbl>,
## #   mean_total_payments <chr>, mean_medicare_payments <chr>, prov_id <dbl>,
## #   prov_address <chr>, prov_state <chr>, mean_covered_charges <chr>

```

5.3.3 Cambio de tipo de variable

Se ha decidido eliminar el símbolo de moneda de dólar de las últimas tres columnas, transformando las columnas a tipo numérico.

```

train$mean_covered_charges = as.numeric(gsub("\\\\$", "", train$mean_covered_charges))

train$mean_total_payments = as.numeric(gsub("\\\\$", "", train$mean_total_payments))

train$mean_medicare_payments = as.numeric(gsub("\\\\$", "", train$mean_medicare_payments))

train$prov_zip = as.factor(train$prov_zip)

train$prov_id = as.factor(train$prov_id)

head(train)

```

```

## # A tibble: 6 x 13
##   codigo_enf desc_enf prov_name prov_city prov_zip referral_reg total_discharges
##   <chr>      <chr>      <chr>      <chr>      <chr>                  <dbl>
## 1 "064 "     " INTRA~ RIVERVIE~ NOBLESVI~ 46060    IN - Indian~           11
## 2 "439 "     " DISOR~ ST LUKE'~ NEW YORK  10025    NY - Manhat~          13
## 3 "853 "     " INFEC~ ST JOSEP~ YONKERS  10701    NY - White ~          17
## 4 "329 "     " MAJOR~ UNIVERSI~ KANSAS C~ 66160    MO - Kansas~          44
## 5 "195 "     " SIMPL~ GARDEN C~ GARDEN C~ 48135    MI - Dearbo~          52
## 6 "176 "     " PULMO~ HORIZON ~ DICKSON  37055    TN - Nashvi~          14
## # ... with 6 more variables: mean_total_payments <dbl>,
## #   mean_medicare_payments <dbl>, prov_id <fct>, prov_address <chr>,
## #   prov_state <chr>, mean_covered_charges <dbl>

```

```
str(train)
```

```

## # tibble [130,452 x 13] (S3: tbl_df/tbl/data.frame)
## $ codigo_enf : chr [1:130452] "064 " "439 " "853 " "329 " ...
## $ desc_enf   : chr [1:130452] "INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W MCC" "P...
## $ prov_name  : chr [1:130452] "RIVERVIEW HOSPITAL" "ST LUKE'S ROOSEVELT HOSPITAL" "ST JO...
## $ prov_city  : chr [1:130452] "NOBLESVILLE" "NEW YORK" "YONKERS" "KANSAS CITY" ...
## $ prov_zip   : Factor w/ 3040 levels "1040","1060",...: 1471 203 221 1991 1557 1132 2271 ...
## $ referral_reg: chr [1:130452] "IN - Indianapolis" "NY - Manhattan" "NY - White Plains" "...
## $ total_discharges: num [1:130452] 11 13 17 44 52 14 110 15 70 20 ...
## $ mean_total_payments: num [1:130452] 11086 11853 47649 42984 5235 ...
## $ mean_medicare_payments: num [1:130452] 8772 11076 44184 41458 4358 ...
## $ prov_id    : Factor w/ 3309 levels "10001","10005",...: 1056 NA 1915 NA 1551 2668 NA NA...
## $ prov_address: chr [1:130452] "395 WESTFIELD RD" "1111 AMSTERDAM AVENUE" "127 SOUTH BRO...
## $ prov_state : chr [1:130452] "IN" "NY" "NY" "KS" ...
## $ mean_covered_charges: num [1:130452] NA 40094 43851 NA 9830 ...

```

5.3.4 Creando columna nueva relativa a los copagos que deben realizar los pacientes: mean_total_payments - mean_medicare_payments

Nueva variable representativa del valor de los copagos que debe realizar el paciente o su seguro privado (en caso de contar con uno), para completar, junto a lo que cubre el Estado con el Medicare, el coste total de la intervención hospitalaria.

```

train <- train %>%
  mutate(copagos = mean_total_payments - mean_medicare_payments)
train

```

```

## # A tibble: 130,452 x 14
##   codigo_enf desc_enf      prov_name      prov_city prov_zip referral_reg
##   <chr>      <chr>      <chr>      <chr>      <fct>      <chr>
## 1 "064 "     " INTRACRANIAL HEM~ RIVERVIEW HOS~ NOBLESVI~ 46060    IN - Indian~
## 2 "439 "     " DISORDERS OF PAN~ ST LUKE'S ROO~ NEW YORK  10025    NY - Manhat~
## 3 "853 "     " INFECTIOUS & PAR~ ST JOSEPH'S M~ YONKERS  10701    NY - White ~
## 4 "329 "     " MAJOR SMALL & LA~ UNIVERSITY OF~ KANSAS C~ 66160    MO - Kansas~
## 5 "195 "     " SIMPLE PNEUMONIA~ GARDEN CITY H~ GARDEN C~ 48135    MI - Dearbo~
## 6 "176 "     " PULMONARY EMBOLI~ HORIZON MEDIC~ DICKSON  37055    TN - Nashvi~
## 7 "641 "     " MISC DISORDERS O~ BAYLOR UNIVER~ DALLAS  75246    TX - Dallas
## 8 "638 "     " DIABETES W CC"  ST ELIZABETH ~ FLORENCE 41042    KY - Coving~

```

```

## 9 "872 "      " SEPTICEMIA OR SE~ ST JOSEPH'S H~ SYRACUSE 13203    NY - Syracu~
## 10 "439 "      " DISORDERS OF PAN~ SOUTH POINTE ~ WARRENSV~ 44122    OH - Clevel~
## # ... with 130,442 more rows, and 8 more variables: total_discharges <dbl>,
## #   mean_total_payments <dbl>, mean_medicare_payments <dbl>, prov_id <fct>,
## #   prov_address <chr>, prov_state <chr>, mean_covered_charges <dbl>,
## #   copagos <dbl>

```

5.3.5 Creando columna nueva relativa a la tasa de cobertura de la Seguridad Social estadounidense: mean_medicare_payments/mean_total_payments

Nueva variable representativa de la tasa de cobertura que ofrece el sistema de salud de la Seguridad Social americana según el hospital, el grupo de diagnóstico y la gravedad del paciente. Nos ayudamos para ello de la librería formattable, obteniendo resultados en forma de porcentaje de cobertura sobre el total a pagar al hospital.

```

train <- train %>% mutate(cobertura = percent(mean_medicare_payments/mean_total_payments))
train

```

```

## # A tibble: 130,452 x 15
##   codigo_enf desc_enf           prov_name     prov_city prov_zip referral_reg
##   <chr>       <chr>           <chr>        <chr>      <fct>     <chr>
## 1 "064 "      " INTRACRANIAL HEM~ RIVERVIEW HOS~ NOBLESVI~ 46060    IN - Indian-
## 2 "439 "      " DISORDERS OF PAN~ ST LUKE'S ROO~ NEW YORK 10025    NY - Manhat-
## 3 "853 "      " INFECTIOUS & PAR~ ST JOSEPH'S M~ YONKERS 10701    NY - White ~
## 4 "329 "      " MAJOR SMALL & LA~ UNIVERSITY OF~ KANSAS C~ 66160    MO - Kansas~
## 5 "195 "      " SIMPLE PNEUMONIA~ GARDEN CITY H~ GARDEN C~ 48135    MI - Dearbo~
## 6 "176 "      " PULMONARY EMBOLI~ HORIZON MEDIC~ DICKSON 37055    TN - Nashvi~
## 7 "641 "      " MISC DISORDERS O~ BAYLOR UNIVER~ DALLAS 75246    TX - Dallas
## 8 "638 "      " DIABETES W CC" ST ELIZABETH ~ FLORENCE 41042    KY - Coving~
## 9 "872 "      " SEPTICEMIA OR SE~ ST JOSEPH'S H~ SYRACUSE 13203    NY - Syracu~
## 10 "439 "     " DISORDERS OF PAN~ SOUTH POINTE ~ WARRENSV~ 44122   OH - Clevel~
## # ... with 130,442 more rows, and 9 more variables: total_discharges <dbl>,
## #   mean_total_payments <dbl>, mean_medicare_payments <dbl>, prov_id <fct>,
## #   prov_address <chr>, prov_state <chr>, mean_covered_charges <dbl>,
## #   copagos <dbl>, cobertura <formtbl>

```

5.4 Análisis exhaustivo de los datos críticos para el estudio

5.4.1 Gráfico EEUU: valor de los copagos por Estados

```
# Haciendo la media de lo que cobra el hospital por estado
```

```

region_geog <- train %>%
  group_by(prov_state) %>%
  summarise(mean_total_price = mean(copagos))
region_geog

```

```

## # A tibble: 52 x 2
##   prov_state mean_total_price

```

```

##      <chr>          <dbl>
## 1 AK           1639.
## 2 AL           1151.
## 3 AR           1092.
## 4 AZ           1320.
## 5 CA           1142.
## 6 CO           1337.
## 7 CT           1279.
## 8 DC           1216.
## 9 DE           1384.
## 10 FL          1164.
## # ... with 42 more rows

# Libreria usmap tiene el mapa de EEUU por estado
library(usmap)

statepop #en libreria usmap hay un dataframe que es la poblacion para cada estado (siglas -abbr ) y n

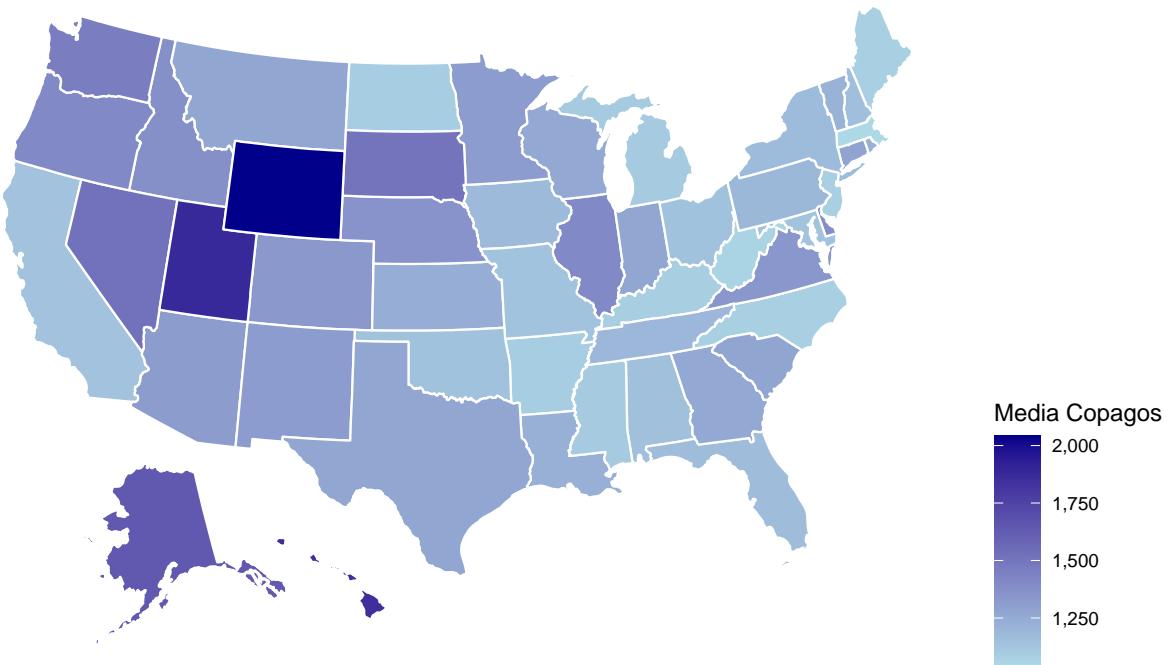
## # A tibble: 51 x 4
##   fips abbr   full      pop_2015
##   <chr> <chr> <chr>      <dbl>
## 1 01   AL    Alabama  4858979
## 2 02   AK    Alaska   738432
## 3 04   AZ    Arizona  6828065
## 4 05   AR    Arkansas 2978204
## 5 06   CA    California 39144818
## 6 08   CO    Colorado  5456574
## 7 09   CT    Connecticut 3590886
## 8 10   DE    Delaware  945934
## 9 11   DC    District of Columbia 672228
## 10 12  FL    Florida   20271272
## # ... with 41 more rows

names(statepop) <- c("fips", "prov_state", "full", "pop_2015") #cambiamos el nombre de la columna abbr

statepop <- statepop %>%
  left_join(region_geog, by = "prov_state") #juntamos region_geog y statepop

plot_usmap(data = statepop, values = "mean_total_price", color = "white") +
  scale_fill_continuous(low = "light blue", high = "dark blue",
                        name = "Media Copagos", label = scales::comma) + theme(legend.position = "right")

```

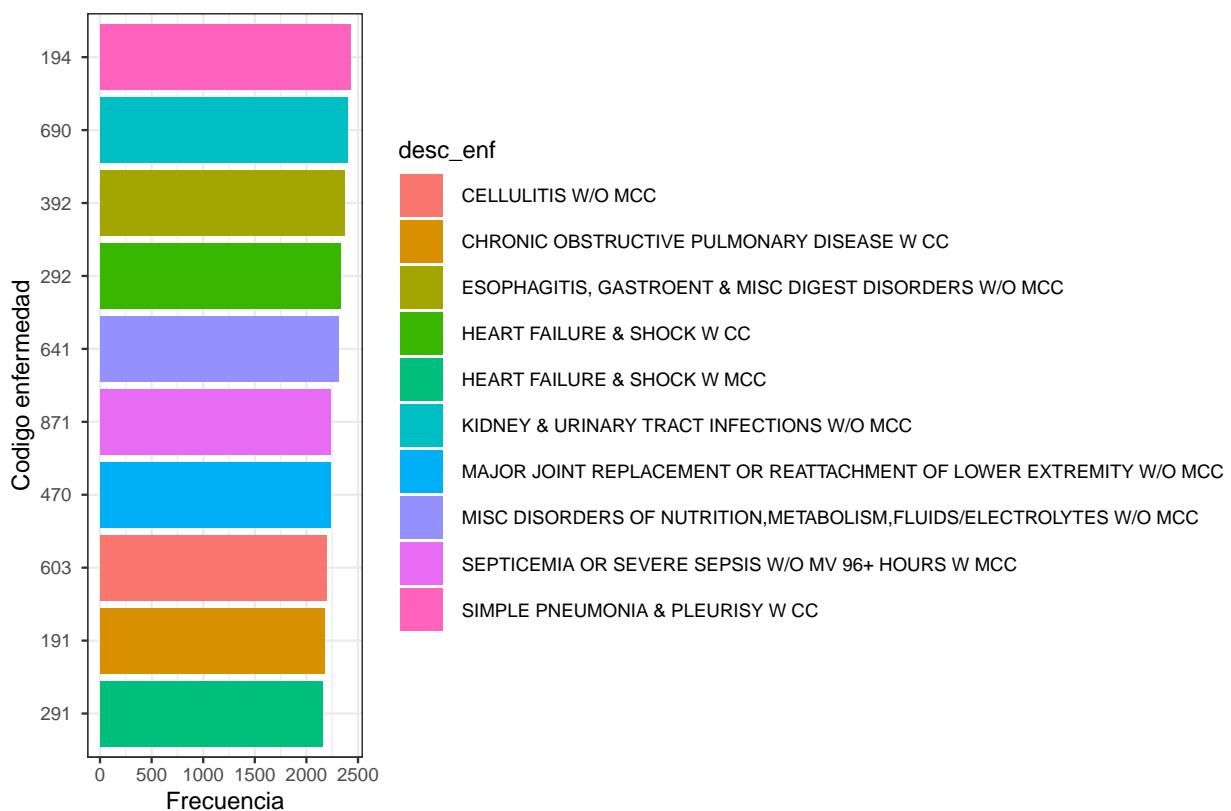


5.4.2 Top 10 enfermedades más comunes detectadas

```
d2 <- train %>%
  count(codigo_enf) %>%
  top_n(10) %>%
  arrange(n, codigo_enf) %>%
  mutate(codigo_enf = factor(codigo_enf, levels = unique(codigo_enf)))

q <- train %>%
  filter(codigo_enf %in% d2$codigo_enf) %>%
  mutate(codigo_enf = factor(codigo_enf, levels = levels(d2$codigo_enf))) %>%
  ggplot(aes(x = codigo_enf, fill=desc_enf)) + geom_bar() + coord_flip() + theme_bw(base_size=9) + xlab(
    ylab("Frecuencia") +
    ggtitle("10 enfermedades mas comunes")
q
```

10 enfermedades más comunes



5.4.3 Top 10 enfermedades más caras

```
d3 <- train %>%
  group_by(codigo_enf) %>% summarise(mean=mean(mean_total_payments)) %>% arrange(desc(mean))

top_10_caras <- head(d3,10)
top_10_caras
```

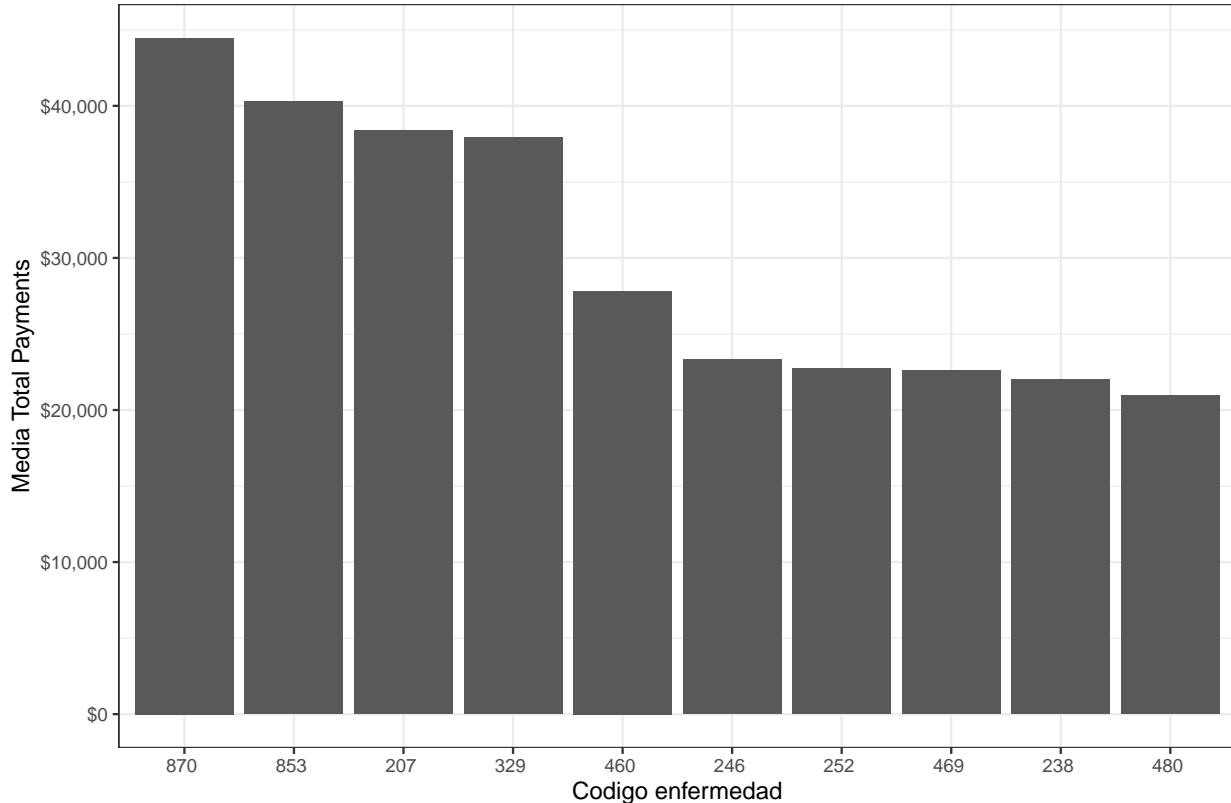
```
## # A tibble: 10 x 2
##   codigo_enf     mean
##   <chr>       <dbl>
## 1 "870 "      44467.
## 2 "853 "      40296.
## 3 "207 "      38372.
## 4 "329 "      37913.
## 5 "460 "      27822.
## 6 "246 "      23337.
## 7 "252 "      22731.
## 8 "469 "      22618.
## 9 "238 "      22009.
## 10 "480 "     20968.
```

```

q <- ggplot(data=top_10_caras, mapping = aes(x = reorder(codigo_enf,-mean),mean)) + geom_bar(stat = "identity")
q

```

10 enfermedades más caras



5.4.4 Gráfico de calor - ratio de cobertura por Estado y por enfermedad

```

test <- top_10_caras %>% inner_join(train)
test

```

```

## # A tibble: 9,048 x 16
##   codigo_enf    mean desc_enf      prov_name    prov_city prov_zip referral_reg
##   <chr>        <dbl> <chr>       <chr>        <chr>      <fct>     <chr>
## 1 "870 "      44467. " SEPTICEMIA~ SOUTHWEST GE~ MIDDLEBU~ 44130 OH - Clevel~
## 2 "870 "      44467. " SEPTICEMIA~ DOWNEY REGIO~ DOWNEY    90241 CA - Los An~
## 3 "870 "      44467. " SEPTICEMIA~ CORONA REGIO~ CORONA    92882 CA - Orange~
## 4 "870 "      44467. " SEPTICEMIA~ ST JOSEPH'S ~ PATERSON 7503  NJ - Paters~
## 5 "870 "      44467. " SEPTICEMIA~ UNIVERSITY O~ SEATTLE    98195 WA - Seattle
## 6 "870 "      44467. " SEPTICEMIA~ KINGSTON HOS~ KINGSTON 12401 NY - Albany
## 7 "870 "      44467. " SEPTICEMIA~ MERCY HEALTH~ MUSKEGON 49444 MI - Muskeg~
## 8 "870 "      44467. " SEPTICEMIA~ ST LUKE'S HO~ MAUMEE    43537 OH - Toledo
## 9 "870 "      44467. " SEPTICEMIA~ UNIVERSITY O~ BIRMINGH~ 35233 AL - Birmin~
## 10 "870 "     44467. " SEPTICEMIA~ BETH ISRAEL ~ BOSTON    2215  MA - Boston
## # ... with 9,038 more rows, and 9 more variables: total_discharges <dbl>,

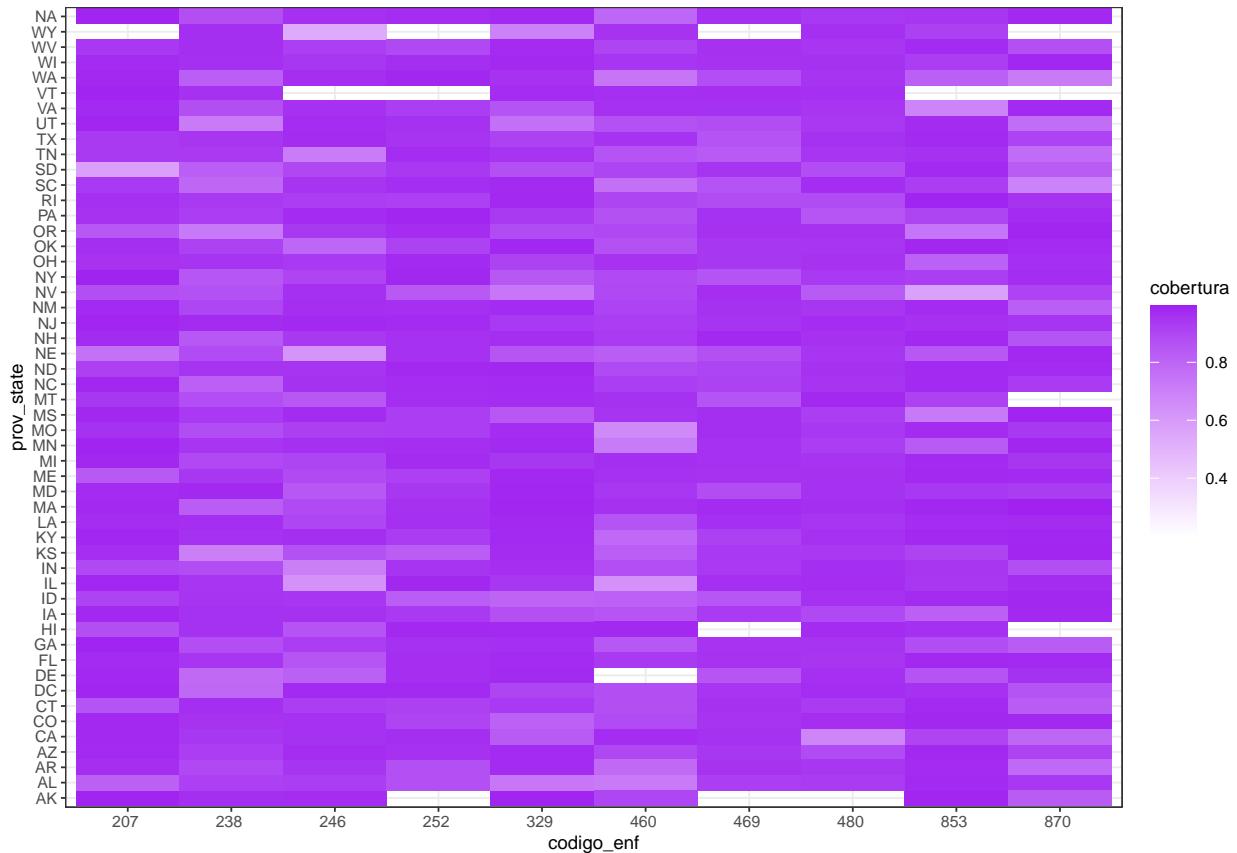
```

```

## #  mean_total_payments <dbl>, mean_medicare_payments <dbl>, prov_id <fct>,
## #  prov_address <chr>, prov_state <chr>, mean_covered_charges <dbl>,
## #  copagos <dbl>, cobertura <formttbl>

ggplot(test, aes(codigo_enf,prov_state, fill=cobertura))+geom_tile() + theme_bw(base_size=7) + scale_fi

```



```

library(PerformanceAnalytics)

## Loading required package: xts

## Warning: package 'xts' was built under R version 4.1.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.2

##
## Attaching package: 'zoo'

```

```

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

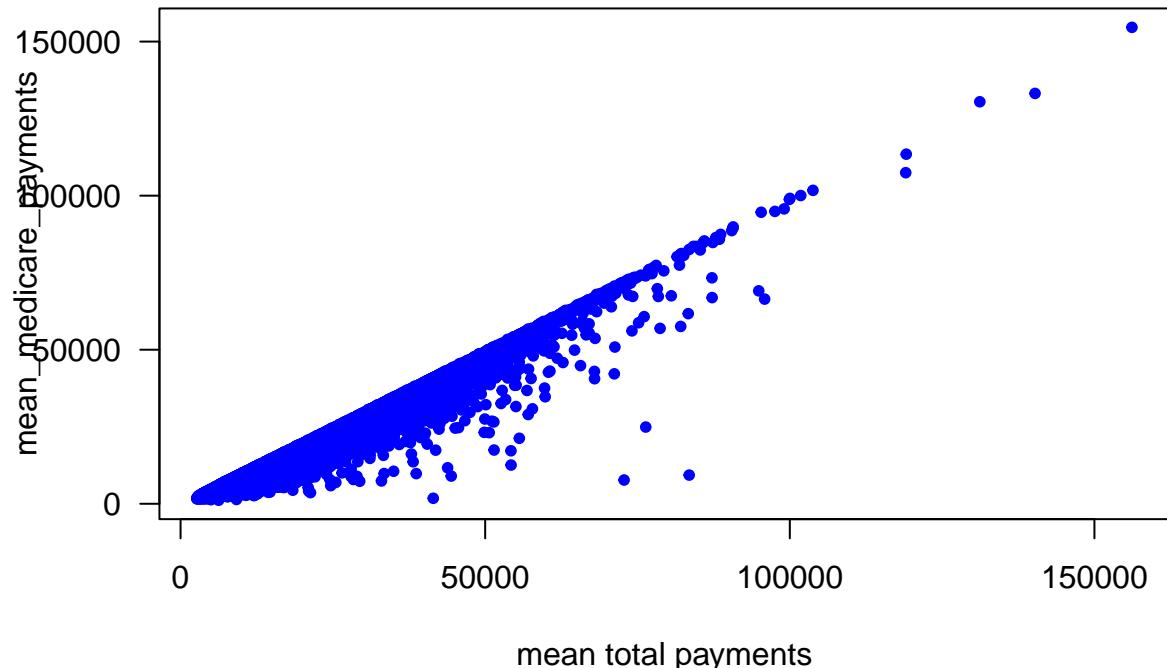
## The following object is masked from 'package:graphics':
##
##     legend

cor(x=train$mean_total_payments, y=train$mean_medicare_payments)

## [1] 0.9893899

with(train, plot(x=mean_total_payments, y=mean_medicare_payments, pch=20, col='blue',
                 xlab='mean total payments', las=1,
                 ylab='mean_medicare_payments'))

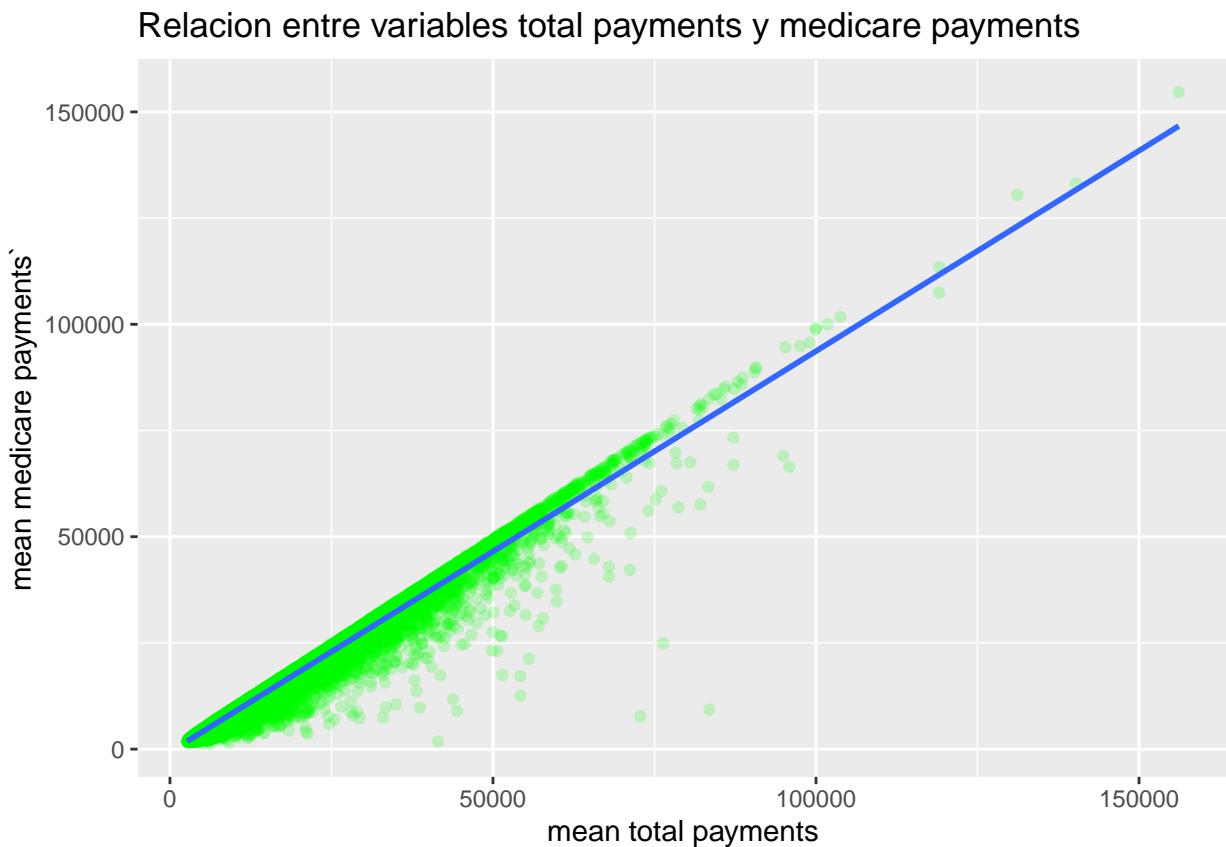
```



```

library(dplyr)
library(ggplot2)
train %>% ggplot(aes(mean_total_payments, mean_medicare_payments)) +
  geom_point(alpha=0.2, colour="green") +
  geom_smooth(formula= 'y ~ x',method = 'lm') +
  labs(title='Relacion entre variables total payments y medicare payments',
       x='mean total payments',
       y='mean medicare payments')

```



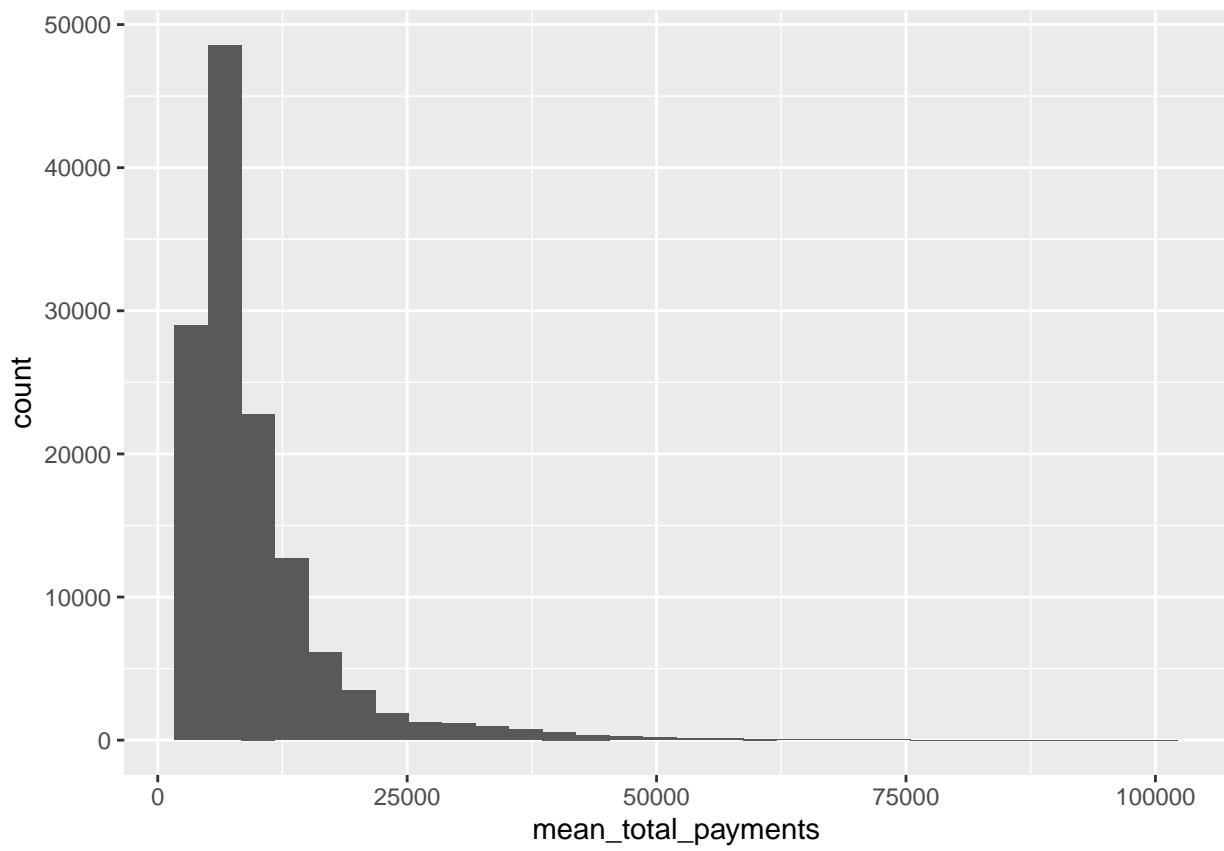
5.4.6 Análisis de la distribución de las variables

```

# ver si el mean total payments sigue una normal
train%>%
  filter(mean_total_payments<100000 ) %>%
  ggplot(aes(x=mean_total_payments))+ geom_histogram()

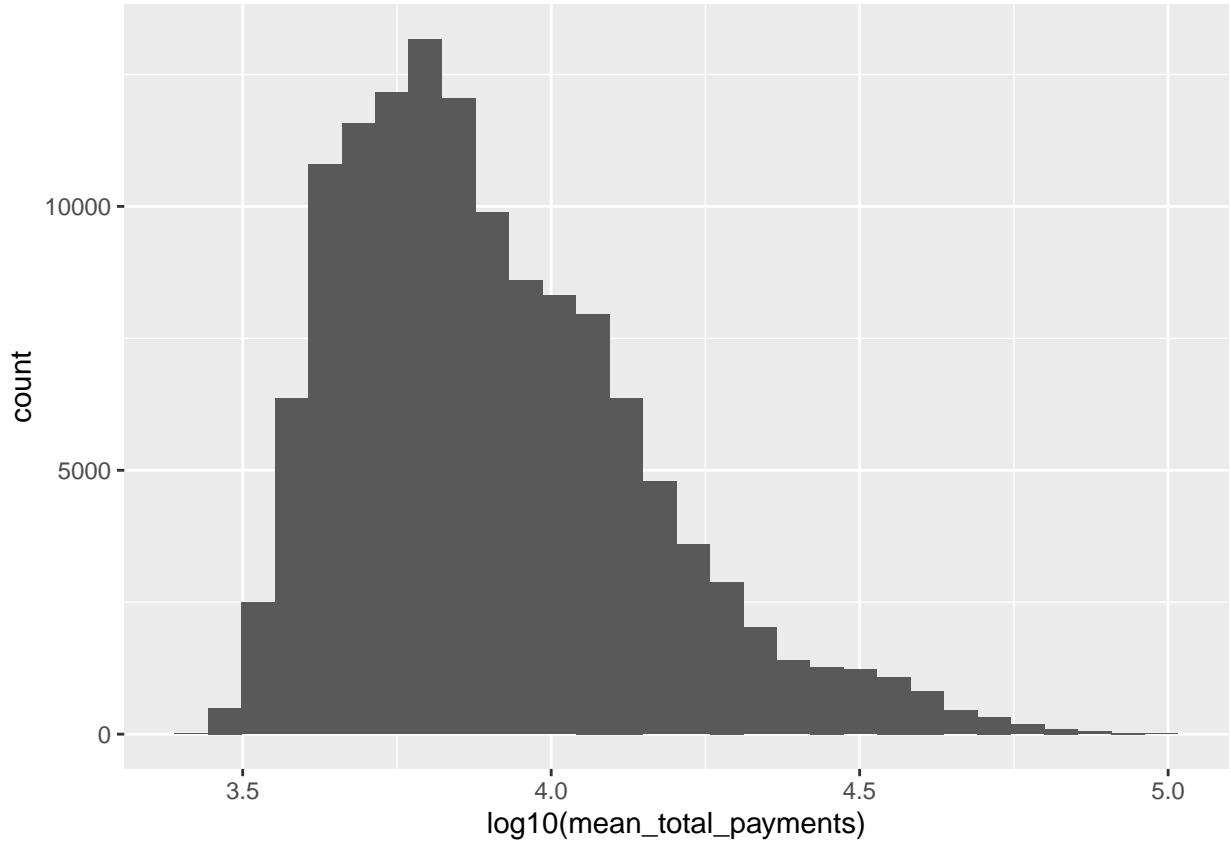
```

‘stat_bin()’ using ‘bins = 30’. Pick better value with ‘binwidth’.



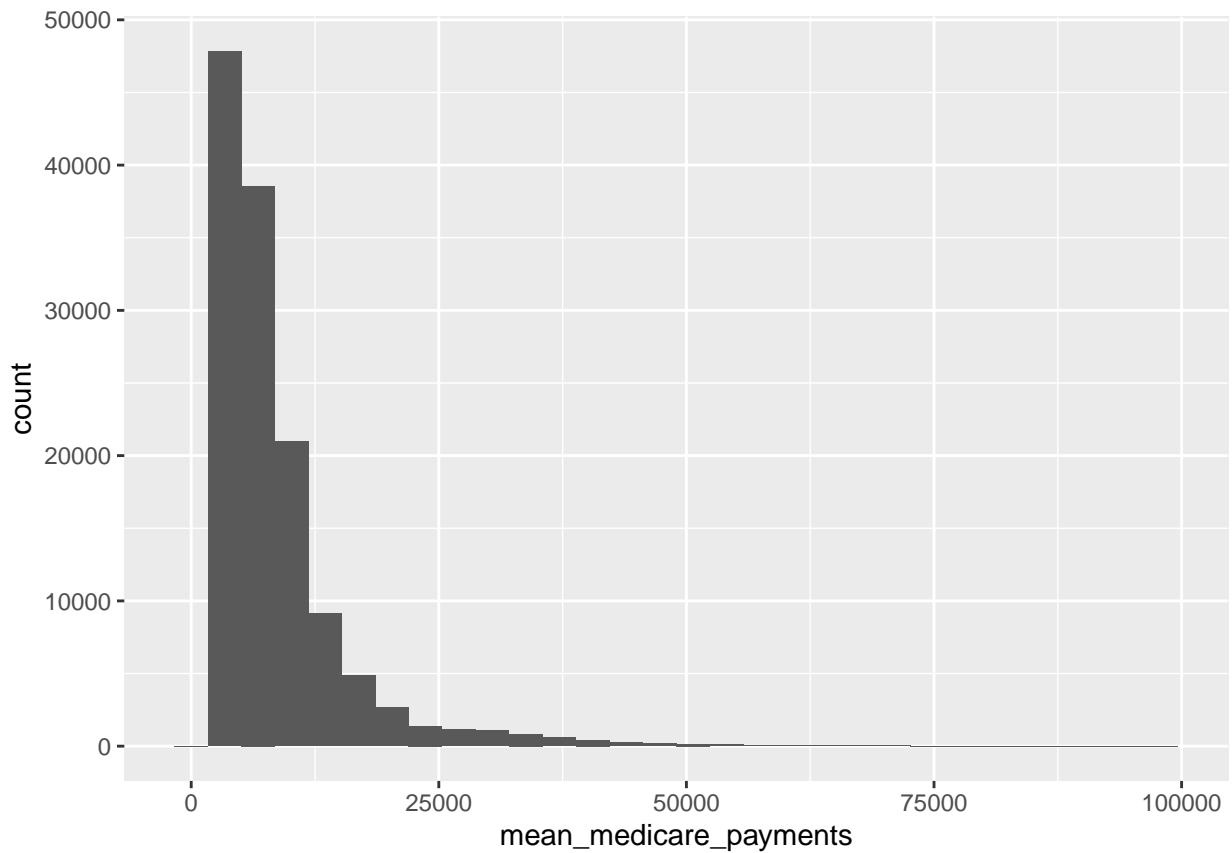
```
# ver si el mean total payments sigue una normal
train%>%
  filter(mean_total_payments<100000 ) %>%
  ggplot(aes(x=log10(mean_total_payments)))+ geom_histogram()

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



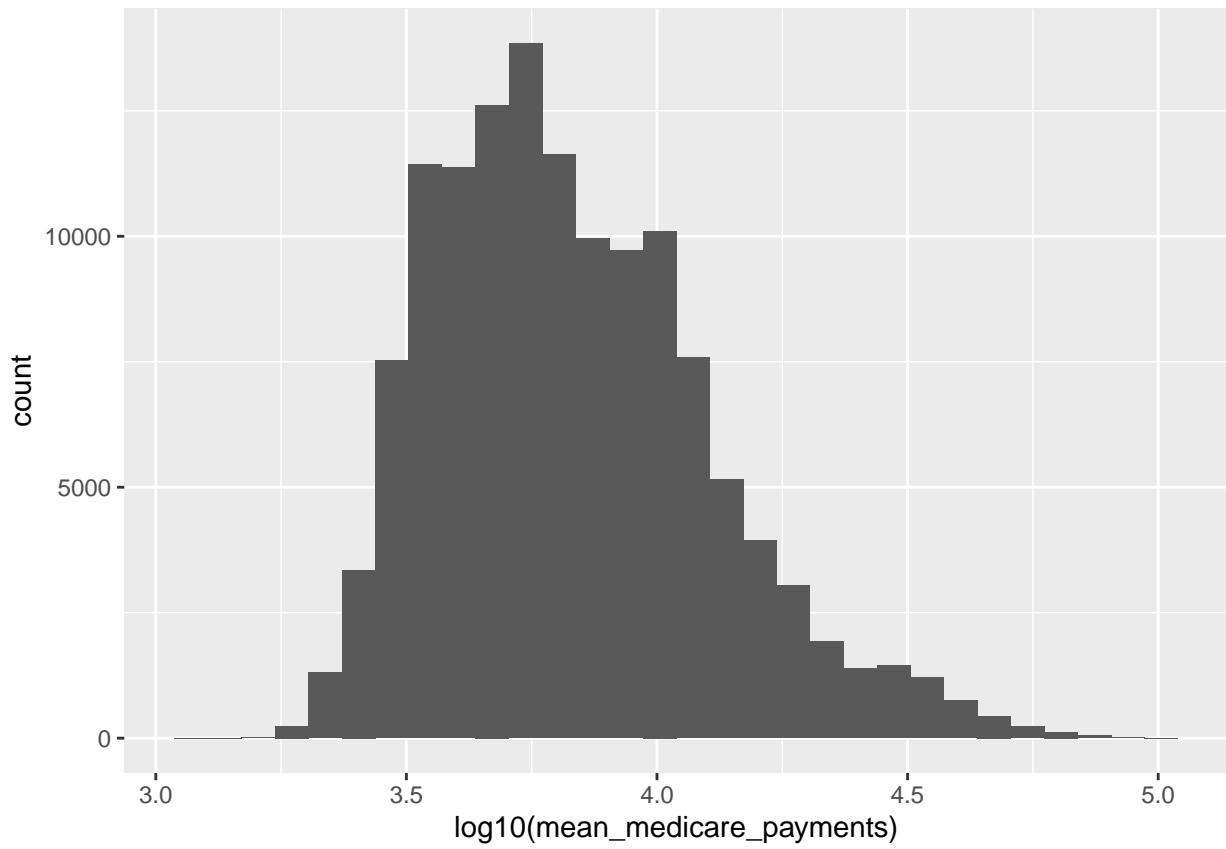
```
# ver si el mean total payments sigue una normal
train%>%
  filter(mean_medicare_payments<100000 ) %>%
  ggplot(aes(x=mean_medicare_payments))+ geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



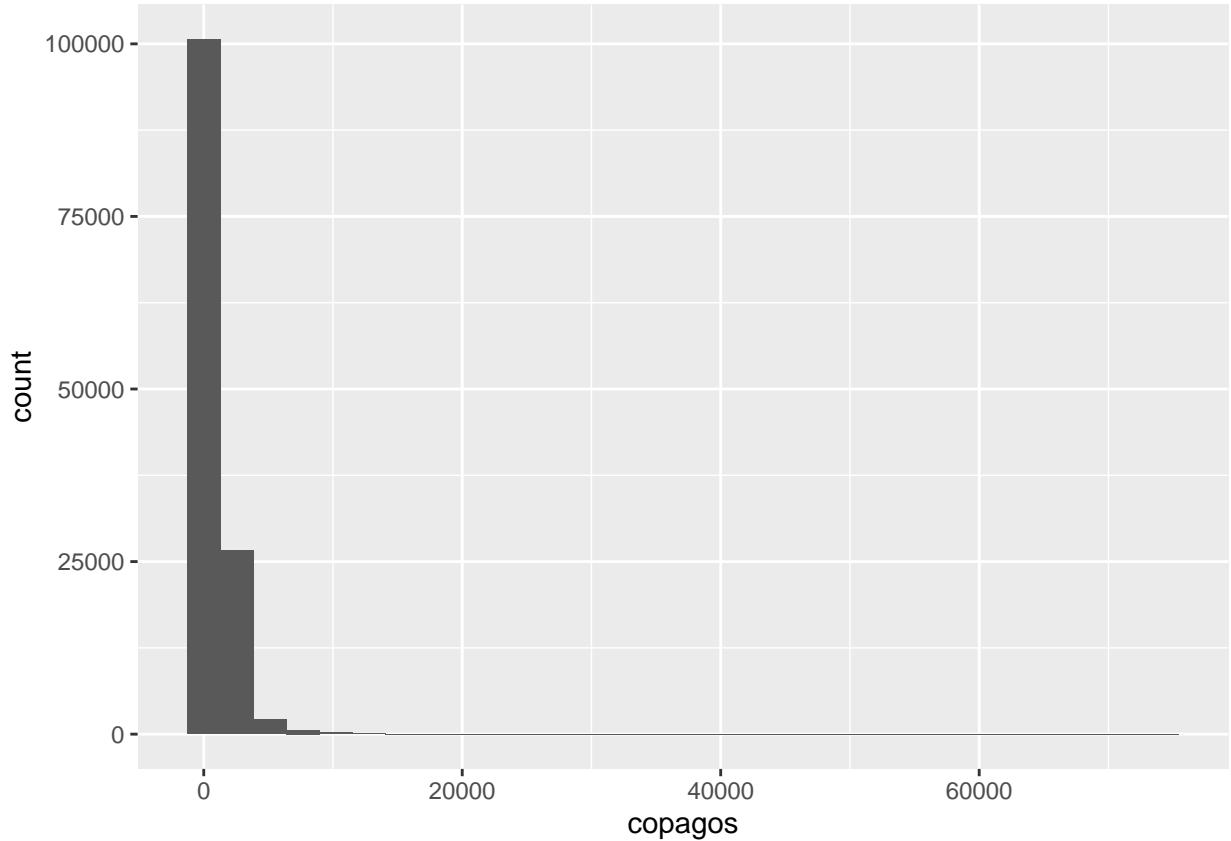
```
# ver si el mean total payments sigue una normal
train%>%
  filter(mean_medicare_payments<100000 ) %>%
  ggplot(aes(x=log10(mean_medicare_payments)))+ geom_histogram()
```

```
## ‘stat_bin()’ using ‘bins = 30’. Pick better value with ‘binwidth’.
```



```
# ver si el mean total payments sigue una normal
train%>%
  filter(mean_medicare_payments<100000 ) %>%
  ggplot(aes(x=copagos))+ geom_histogram()

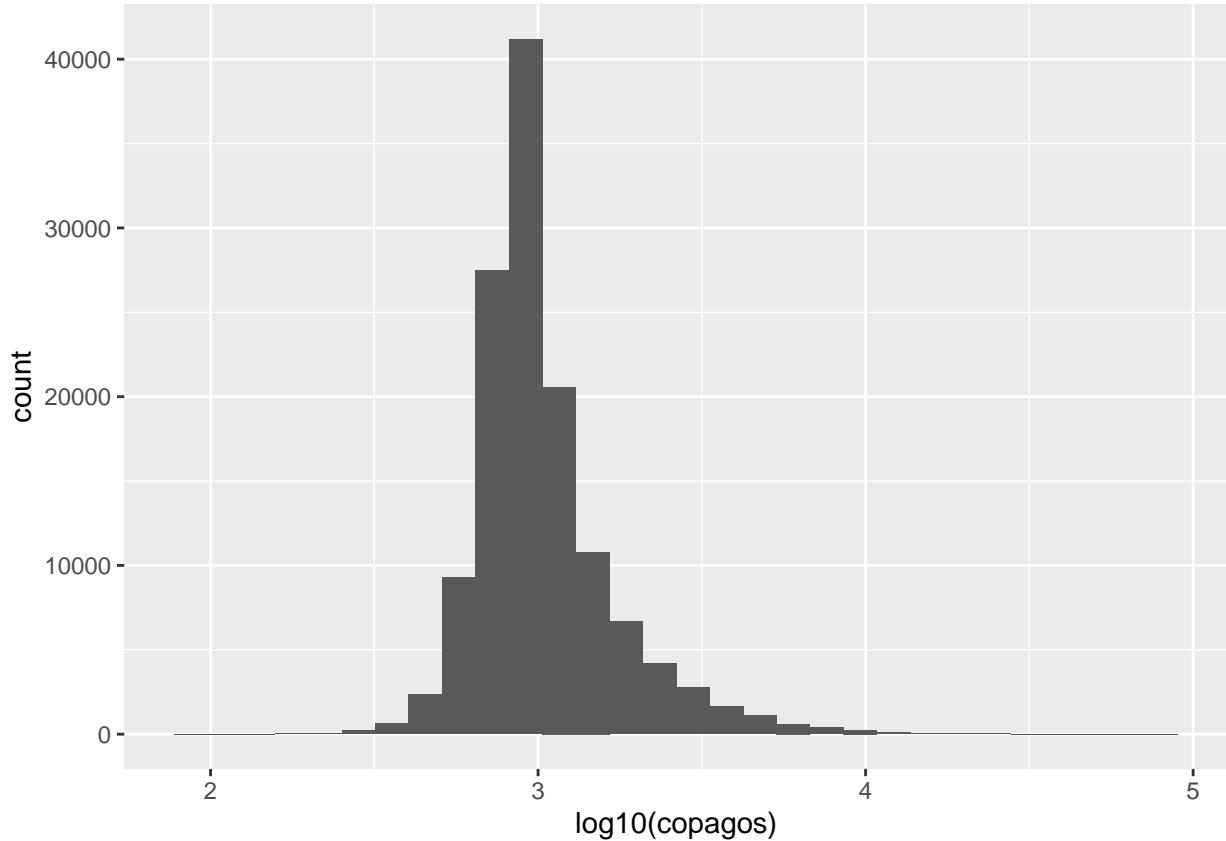
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# ver si el mean total payments sigue una normal
train%>%
  filter(mean_medicare_payments<100000 ) %>%
  ggplot(aes(x=log10(copagos)))+ geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```
#train %>% select(1:14) %>%
#  na.omit() %>%
#  ggpairs(columns = 1:13, ggplot2::aes(colour=group), cardinality_threshold=50000)
```

5.4.7 Boxplot - análisis de la variables de relevancia y de los atípicos observados

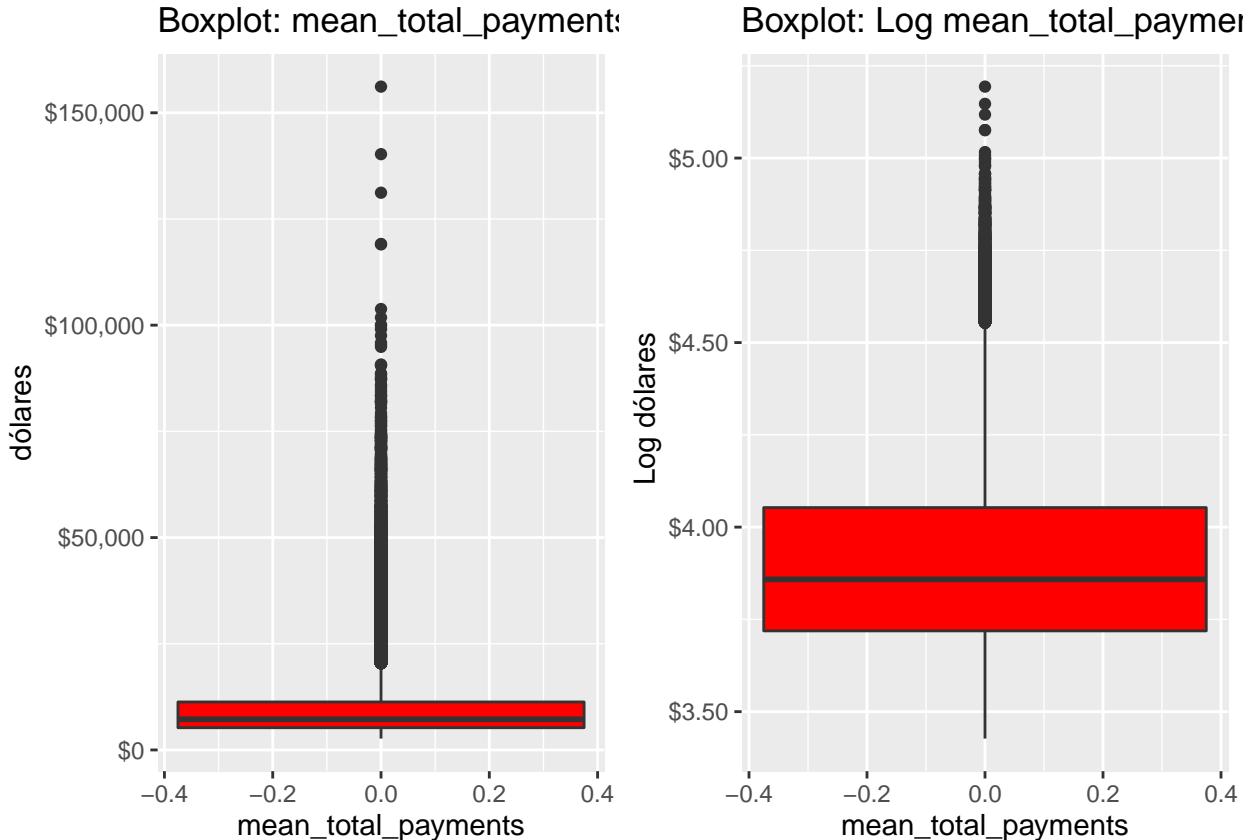
```
train_num <- train %>% select_if(is.numeric)
train_num

## # A tibble: 130,452 x 6
##   total_discharges mean_total_payme~ mean_medicare_pa~ mean_covered_ch~ copagos
##             <dbl>            <dbl>            <dbl>            <dbl>      <dbl>
## 1                 11           11086.          8772.        NA       2314.
## 2                 13           11853.          11076.        40094.     776.
## 3                 17           47649.          44184.        43851.     3465.
## 4                 44           42984.          41458.        NA       1526.
## 5                 52           5235.           4358.         9830.      877.
## 6                 14           6612.           5248.        22260.     1364
## 7                110           5407.           4304.        16005.     1103.
## 8                 15           5160.           4043.        13639.     1117.
## 9                 70           8729.           6518.        16849.     2211.
## 10                20           6411.           5708.        NA       703.
## # ... with 130,442 more rows, and 1 more variable: cobertura <formttbl>
```

```

p1 <- ggplot (train_num, aes(y= train_num$mean_total_payments)) + geom_boxplot(fill = "red") + scale_y_continuous(label = "dólares")
p11 <- ggplot (train_num, aes(y= log10(train_num$mean_total_payments))) + geom_boxplot(fill = "red") + scale_y_continuous(label = "Log dólares")
plot_grid(p1, p11)

```



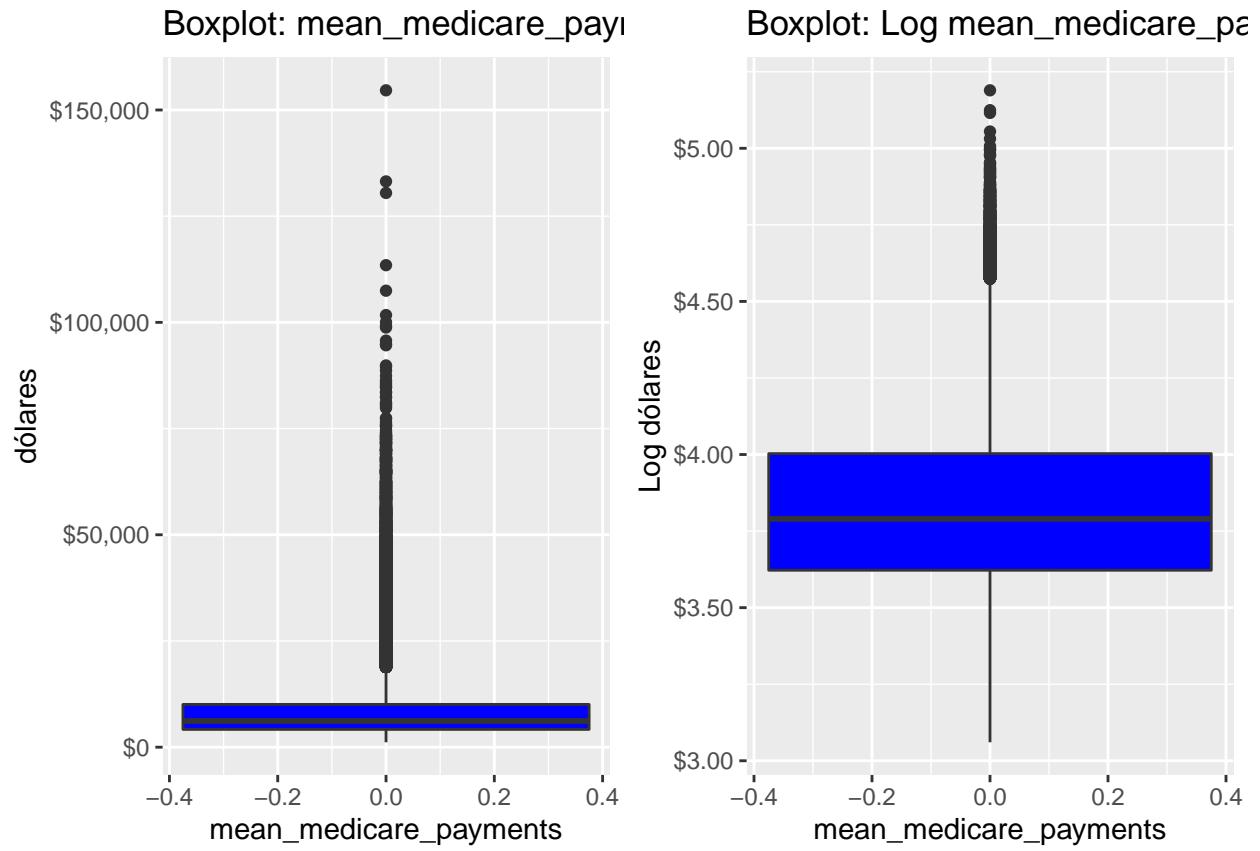
```
mean(train_num$mean_total_payments)
```

```
## [1] 9715.202
```

```

p2 <- ggplot (train_num, aes(y=train_num$mean_medicare_payments)) + geom_boxplot(fill = "blue") + scale_y_continuous(label = "dólares")
p22 <- ggplot (train_num, aes(y= log10(train_num$mean_medicare_payments))) + geom_boxplot(fill = "blue") + scale_y_continuous(label = "Log dólares")
plot_grid(p2, p22)

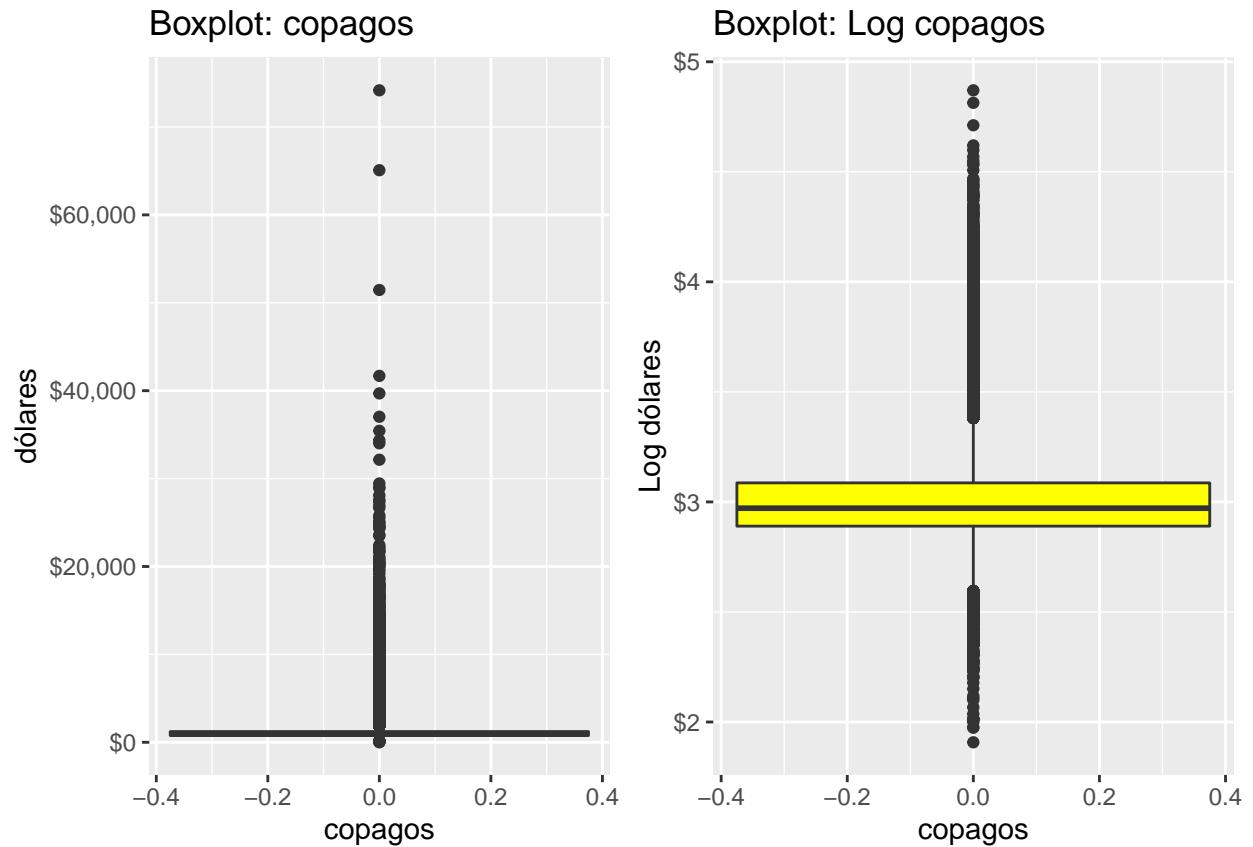
```



```
mean(train_num$mean_medicare_payments)
```

```
## [1] 8501.48
```

```
p3 <- ggplot (train_num, aes(y=train_num$copagos)) + geom_boxplot(fill = "yellow") + scale_y_continuous()
p33 <- ggplot (train_num, aes(y=log10(train_num$copagos))) + geom_boxplot(fill = "yellow") + scale_y_continuous()
plot_grid(p3, p33)
```



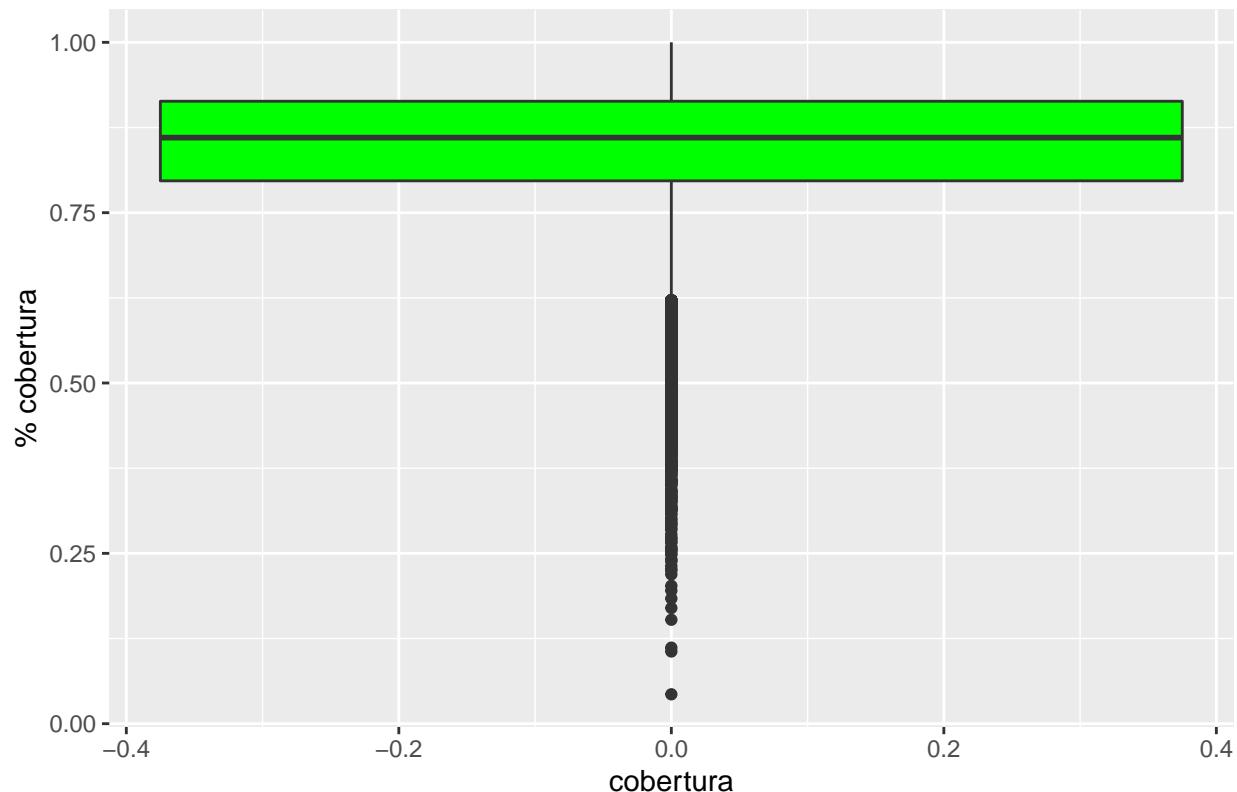
```
mean(train_num$copagos)
```

```
## [1] 1213.723
```

```
p4 <- ggplot (train_num, aes(y=train_num$cobertura)) + geom_boxplot(fill = "green") + scale_y_continuous
```

p4

Boxplot: cobertura



```
mean(train_num$cobertura)
```

```
## [1] 84.66%
```