Práctica 2. Tipología y ciclo de vida del dato.

Alberto Mariscal y Elena Naranjo

29/5/2022

Contents

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 1
2. Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir. 2
3. Limpieza de los datos 5
4. Análisis de los datos 10
5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica. 15
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 15
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python . . . 15

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

EL dataset elegido se puede encontrar en la plataforma Kaggle, conocida por los concursos que organiza en temas relacionados con Data Science y los datasets que ofrece (https://www.kaggle.com/competitions/actuarial-loss-estimation/data). Se trata de un conjunto de datos que incluye 90000 casos realistas de compensaciones por seguros laborales en casos de accidentes en el trabajo. Para cada registro se tiene información demográfica y del empleado, así como una descripción del accidente.

El dataset incluye 15 variables que tienen los siguientes nombres y descripciones:

- ClaimNumber: es un identificador único de la poliza que servirá para identificarla y distinguirla de las demás
- DateTimeOfAccidente: fecha y hora del accidente
- DateReported: Fecha en la que se reporta el accidente
- Age: Edad del trabajador
- Gender: Sexo del trabajador
- MaritalStatus: estado civil del trabajador que podrá estar casado (M), soltero (S) o no tener información al respecto (U)
- DependentChildren: número de niños dependientes del trabajador
- DependentsOther: número de dependientes del trabajador sin contabilizar los niños
- WeeklyWage: salario semanal del trabajador
- PartTimeFullTime: tipo de contrato laboral que podrá ser a tiempo parcial (P) o a jornada completa (F)
- HoursWorkedPerWeek: número total de horas trabajadas a la semana
- DaysWorkedPerWeek: Número total de días trabajados a la semana
- ClaimDescription: campo libre con comentarios descriptivos sobre el registro

- InitialIncurredClaimCost: coste inicial estimado por la aseguradora
- UltimateIncurredClaimCost: pagos totales de la aseguradora al trabajador

El dataset se ha elegido porque nos parece que está muy completo, permite realizar análisis interesantes y trata un tema de gran relevancia, interés y actualidad. Anualmente, las compañías aseguradoras deben hacer frente a grandes costes por las polizas que deben abonar, muchos de ellos debido a fraudes o desvíos significativos sobre lo inicialmente presupuestado. Estamos seguros que las grandes compañías, al igual que los bancos hacen con sus productos por ejemplo, cuentan con grandes departamentos de data analysis y data science que tratan de desarrollar modelos para ajustar lo mejor posible sus polizas y será lo que en esta práctica intentaremos, en la medida de lo posible, replicar.

El objetivo de la práctica es el de crear un modelo para establecer el precio a pagar en función del accidente, así como analizar si existen diferencias entre diferentes grupos como, por ejemplo, hombres y mujeres.

```
claim <- read.csv("dataset_original.csv")
```

2. Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

En un principio no consideramos que sea necesario filtrar los datos ni hacer selecciones de los mismos ya que Kaggle nos los proporciona en dos archivos diferentes divididos en train y test. Por otro lado, no hay categorías o grupos diferentes como para separarlo, por lo que consideramos que no tendría ningún sentido hacerlo.

En primer lugar, lo que nos interesa será cambiar el nombre de las columnas:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
names(claim) <-  c('Id', 'Ocurrencia', 'Apertura', 'Edad', 'Sexo', 'Estado', 'Dependientes', 'OtrosDepe

head(claim)
```

```
##          Id           Ocurrencia             Apertura Edad Sexo Estado
## 1 WC8285054 2002-04-09T07:00:00Z 2002-07-05T00:00:00Z   48    M      M
## 2 WC6982224 1999-01-07T11:00:00Z 1999-01-20T00:00:00Z   43    F      M
## 3 WC5481426 1996-03-25T00:00:00Z 1996-04-14T00:00:00Z   30    M      U
## 4 WC9775968 2005-06-22T13:00:00Z 2005-07-22T00:00:00Z   41    M      S
## 5 WC2634037 1990-08-29T08:00:00Z 1990-09-27T00:00:00Z   36    M      M
## 6 WC6828422 1999-06-21T11:00:00Z 1999-09-09T00:00:00Z   50    M      M
##   Dependientes OtrosDepend Salario Jornada HorasSemana DiasSemana
## 1            0           0  500.00       F        38.0          5
## 2            0           0  509.34       F        37.5          5
## 3            0           0  709.10       F        38.0          5
## 4            0           0  555.46       F        38.0          5
## 5            0           0  377.10       F        38.0          5
## 6            0           0  200.00       F        38.0          5
```

```
##                                                                 Descripcion
## 1                      LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2             STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
## 3                                   CUT ON SHARP EDGE CUT LEFT THUMB
## 4                                DIGGING LOWER BACK LOWER BACK STRAIN
## 5 REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
## 6                                 STRUCK HEAD ON HEAD LACERATED HEAD
##   CosteInicio CosteFinal
## 1        1500  4748.2034
## 2        5500  6326.2858
## 3        1700  2293.9491
## 4       15000 17786.4872
## 5        2800  4014.0029
## 6         500   598.7623
```

En segundo lugar, para facilitar los cálculos temporales convertiremos las variables **Ocurrencia y Apertura** a formato fecha para más facilidad en su posterior tratamiento. De la misma manera, creamos una nueva variable llamada **Tiempo** que será el tiempo transcurrido entre la fecha del accidente y la apertura del incidente.

```
claim$Ocurrencia <- as.Date(claim$Ocurrencia)
claim$Apertura <- as.Date(claim$Apertura)

claim$tiempo <- as.integer(claim$Apertura - claim$Ocurrencia, units = "days")
head(claim)
```

```
##          Id Ocurrencia   Apertura Edad Sexo Estado Dependientes OtrosDepend
## 1 WC8285054 2002-04-09 2002-07-05   48    M      M            0           0
## 2 WC6982224 1999-01-07 1999-01-20   43    F      M            0           0
## 3 WC5481426 1996-03-25 1996-04-14   30    M      U            0           0
## 4 WC9775968 2005-06-22 2005-07-22   41    M      S            0           0
## 5 WC2634037 1990-08-29 1990-09-27   36    M      M            0           0
## 6 WC6828422 1999-06-21 1999-09-09   50    M      M            0           0
##   Salario Jornada HorasSemana DiasSemana
## 1  500.00       F        38.0          5
## 2  509.34       F        37.5          5
## 3  709.10       F        38.0          5
## 4  555.46       F        38.0          5
## 5  377.10       F        38.0          5
## 6  200.00       F        38.0          5
##                                                                 Descripcion
## 1                      LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2             STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
## 3                                   CUT ON SHARP EDGE CUT LEFT THUMB
## 4                                DIGGING LOWER BACK LOWER BACK STRAIN
## 5 REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
## 6                                 STRUCK HEAD ON HEAD LACERATED HEAD
##   CosteInicio CosteFinal tiempo
## 1        1500  4748.2034     87
## 2        5500  6326.2858     13
## 3        1700  2293.9491     20
## 4       15000 17786.4872     30
## 5        2800  4014.0029     29
## 6         500   598.7623     80
```

Vamos a crear también una variable categórica llamada **Clasificación** relativa al tiempo calculado anterior-

mente.

```
claim$Clasificacion <- cut(as.double(claim$tiempo), breaks = c(0, 15, 30, 89, Inf), labels = c("Muy rapi
head(claim)
```

```
##          Id Ocurrencia   Apertura Edad Sexo Estado Dependientes OtrosDepend
## 1 WC8285054 2002-04-09 2002-07-05   48    M      M            0           0
## 2 WC6982224 1999-01-07 1999-01-20   43    F      M            0           0
## 3 WC5481426 1996-03-25 1996-04-14   30    M      U            0           0
## 4 WC9775968 2005-06-22 2005-07-22   41    M      S            0           0
## 5 WC2634037 1990-08-29 1990-09-27   36    M      M            0           0
## 6 WC6828422 1999-06-21 1999-09-09   50    M      M            0           0
##   Salario Jornada HorasSemana DiasSemana
## 1  500.00       F        38.0          5
## 2  509.34       F        37.5          5
## 3  709.10       F        38.0          5
## 4  555.46       F        38.0          5
## 5  377.10       F        38.0          5
## 6  200.00       F        38.0          5
##                                                          Descripcion
## 1                     LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2             STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
## 3                                  CUT ON SHARP EDGE CUT LEFT THUMB
## 4                               DIGGING LOWER BACK LOWER BACK STRAIN
## 5 REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
## 6                                STRUCK HEAD ON HEAD LACERATED HEAD
##   CosteInicio CosteFinal tiempo Clasificacion
## 1        1500  4748.2034     87         Lento
## 2        5500  6326.2858     13    Muy rapido
## 3        1700  2293.9491     20        Rapido
## 4       15000 17786.4872     30        Rapido
## 5        2800  4014.0029     29        Rapido
## 6         500   598.7623     80         Lento
```

Además de estas variables, crearemos una columna llamada **RiesgoSM** que contenga información sobre el riesgo de enfermedades relacionadas con la salud mental, tema de total actualidad, en el trabajador. Para ello, haremos uso de la columna que contiene una descripción en texto de lo sucedido.

```
claim$RiesgoSM <- as.factor(case_when(grepl("STRESS|ANXIETY|HARASSMENT|DEPRESSION", claim$Descripcion) =
table(claim$RiesgoSM)
```

```
##
##     0     1
## 53736   264
```

Por último, exploramos los datos para tener una idea inicial tras los cambios y creaciones de nuevas columnas:

```
summary(claim)
```

```
##       Id              Ocurrencia           Apertura              Edad
##  Length:54000       Min.   :1988-01-01   Min.   :1988-01-08   Min.   :13.00
##  Class :character   1st Qu.:1992-06-30   1st Qu.:1992-08-04   1st Qu.:23.00
##  Mode  :character   Median :1997-01-07   Median :1997-02-16   Median :32.00
##                     Mean   :1997-01-02   Mean   :1997-02-11   Mean   :33.84
##                     3rd Qu.:2001-07-09   3rd Qu.:2001-08-25   3rd Qu.:43.00
##                     Max.   :2005-12-31   Max.   :2006-09-23   Max.   :81.00
##      Sexo               Estado           Dependientes     OtrosDepend
```

```
##  Length:54000       Length:54000       Min.   :0.0000   Min.   :0.000000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.000000
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.000000
##                                        Mean   :0.1192   Mean   :0.009944
##                                        3rd Qu.:0.0000   3rd Qu.:0.000000
##                                        Max.   :9.0000   Max.   :5.000000
##     Salario          Jornada           HorasSemana       DiasSemana
##  Min.   :   1.0   Length:54000       Min.   :  0.00   Min.   :1.000
##  1st Qu.: 200.0   Class :character   1st Qu.: 38.00   1st Qu.:5.000
##  Median : 392.2   Mode  :character   Median : 38.00   Median :5.000
##  Mean   : 416.4                      Mean   : 37.74   Mean   :4.906
##  3rd Qu.: 500.0                      3rd Qu.: 40.00   3rd Qu.:5.000
##  Max.   :7497.0                      Max.   :640.00   Max.   :7.000
##  Descripcion         CosteInicio        CosteFinal          tiempo
##  Length:54000       Min.   :      1   Min.   :    122   Min.   :   0.00
##  Class :character   1st Qu.:    700   1st Qu.:    926   1st Qu.:  14.00
##  Mode  :character   Median :   2000   Median :   3371   Median :  22.00
##                     Mean   :   7841   Mean   :  11003   Mean   :  39.31
##                     3rd Qu.:   9500   3rd Qu.:   8197   3rd Qu.:  41.00
##                     Max.   :2000000   Max.   :4027136   Max.   :1095.00
##     Clasificacion   RiesgoSM
##  Muy rapido:16949   0:53736
##  Rapido    :17867   1:  264
##  Lento     :14785
##  Muy lento : 4394
##  NA's      :    5
##
```

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos

Para poder gestionar los valores que faltan deberemos ver qué registros tienen valores desconocidos como "nan" pero también aquellos que están introducidos como "u" de unknown, por ejemplo. Este trabajo deberá hacerse manualmente para cada variable ya que es específico para este dataset y deberemos comprender qué información contiene cada columna para entender cómo tratar los datos.

Comenzamos con la variable **Sexo** para ver cuántos tipos diferentes de registros tiene el dataset:

```
# Variable Sexo
table(claim$Sexo)
```

```
##
##     F     M     U
## 12338 41660     2
```

Como podemos ver, tenemos 3 categorías en la variable Sexo, 12338 catalogados como femenino, 41660 como masculino y 2 como desconocidos. Estos últimos serán aquellos valores desconocidos y haremos que R los trate como tal asignándoles el valor nan de manera que no se incluyan en las estadísticas al hacer cálculos.

```
claim$Sexo <- na_if(claim$Sexo, "U")
```

Continuamos con la varible **Estado**:

```
# Variable Sexo
table(claim$Estado)
```

```
##
```

```
##           M     S     U
##    29 22516 26161  5294
```

Vemos que hay registros que tienen un string vacío ("“) o también desconocido (”u"). Procederemos de la misma manera, asignándoles el valor nan para que R ya distinga que son valores que faltan.

```
claim$Estado <- na_if(claim$Estado, "U")
claim$Estado <- na_if(claim$Estado, "")

table(claim$Estado)
```

```
##
##     M     S
## 22516 26161
```

Dado que casi un 10% de los datos de Estado son nulos, vamos a realizar un filtro, **seleccionando solo con aquellos registros que no tengan datos nulos** de cara a un mejor modelo en el futuro, llamaremos a este nuevo conjunto claimNET:

```
# Comprobamos valores NA
summary(claim)
```

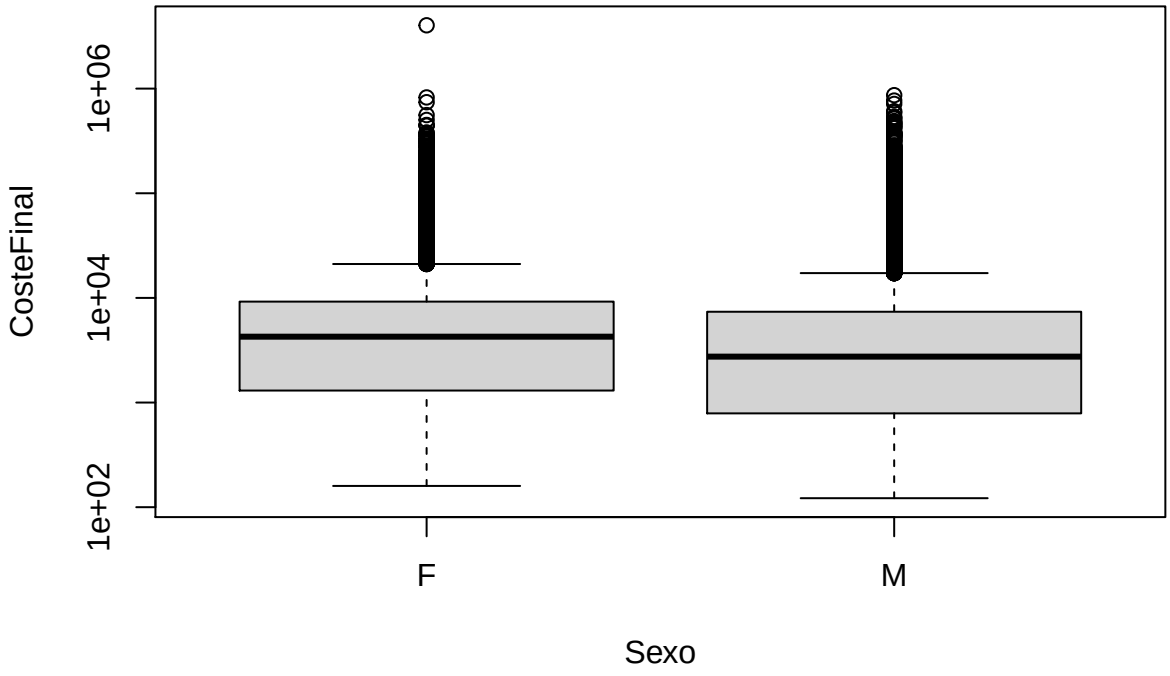
```
##       Id              Ocurrencia            Apertura               Edad
##  Length:54000       Min.   :1988-01-01   Min.   :1988-01-08   Min.   :13.00
##  Class :character   1st Qu.:1992-06-30   1st Qu.:1992-08-04   1st Qu.:23.00
##  Mode  :character   Median :1997-01-07   Median :1997-02-16   Median :32.00
##                     Mean   :1997-01-02   Mean   :1997-02-11   Mean   :33.84
##                     3rd Qu.:2001-07-09   3rd Qu.:2001-08-25   3rd Qu.:43.00
##                     Max.   :2005-12-31   Max.   :2006-09-23   Max.   :81.00
##      Sexo              Estado           Dependientes     OtrosDepend
##  Length:54000       Length:54000       Min.   :0.0000   Min.   :0.000000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.000000
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.000000
##                                        Mean   :0.1192   Mean   :0.009944
##                                        3rd Qu.:0.0000   3rd Qu.:0.000000
##                                        Max.   :9.0000   Max.   :5.000000
##     Salario         Jornada           HorasSemana       DiasSemana
##  Min.   :   1.0   Length:54000       Min.   :  0.00   Min.   :1.000
##  1st Qu.: 200.0   Class :character   1st Qu.: 38.00   1st Qu.:5.000
##  Median : 392.2   Mode  :character   Median : 38.00   Median :5.000
##  Mean   : 416.4                      Mean   : 37.74   Mean   :4.906
##  3rd Qu.: 500.0                      3rd Qu.: 40.00   3rd Qu.:5.000
##  Max.   :7497.0                      Max.   :640.00   Max.   :7.000
##  Descripcion         CosteInicio        CosteFinal          tiempo
##  Length:54000       Min.   :      1   Min.   :    122   Min.   :   0.00
##  Class :character   1st Qu.:    700   1st Qu.:    926   1st Qu.:  14.00
##  Mode  :character   Median :   2000   Median :   3371   Median :  22.00
##                     Mean   :   7841   Mean   :  11003   Mean   :  39.31
##                     3rd Qu.:   9500   3rd Qu.:   8197   3rd Qu.:  41.00
##                     Max.   :2000000   Max.   :4027136   Max.   :1095.00
##     Clasificacion   RiesgoSM
##  Muy rapido:16949   0:53736
##  Rapido    :17867   1:  264
##  Lento     :14785
##  Muy lento : 4394
##  NA's      :    5
##
```

```
claimNET <- drop_na(claim)
```

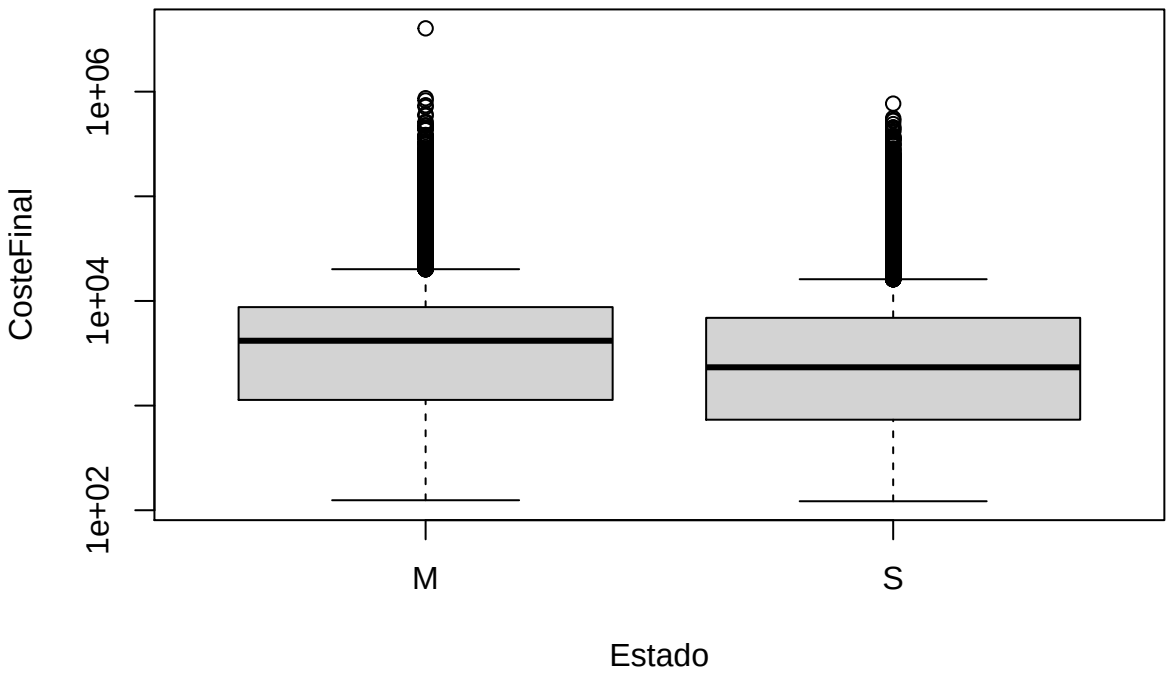
3.2 Identifica y gestiona los valores extremos

Para identificar los valores extremos, procederemos a mostrarr las diferentes variables en boxplots frente a la variable objetivo a poder predecir CosteFinal, de manera que se observe la relación entre ellas y la distribución de las mismas en las distintas categorías.

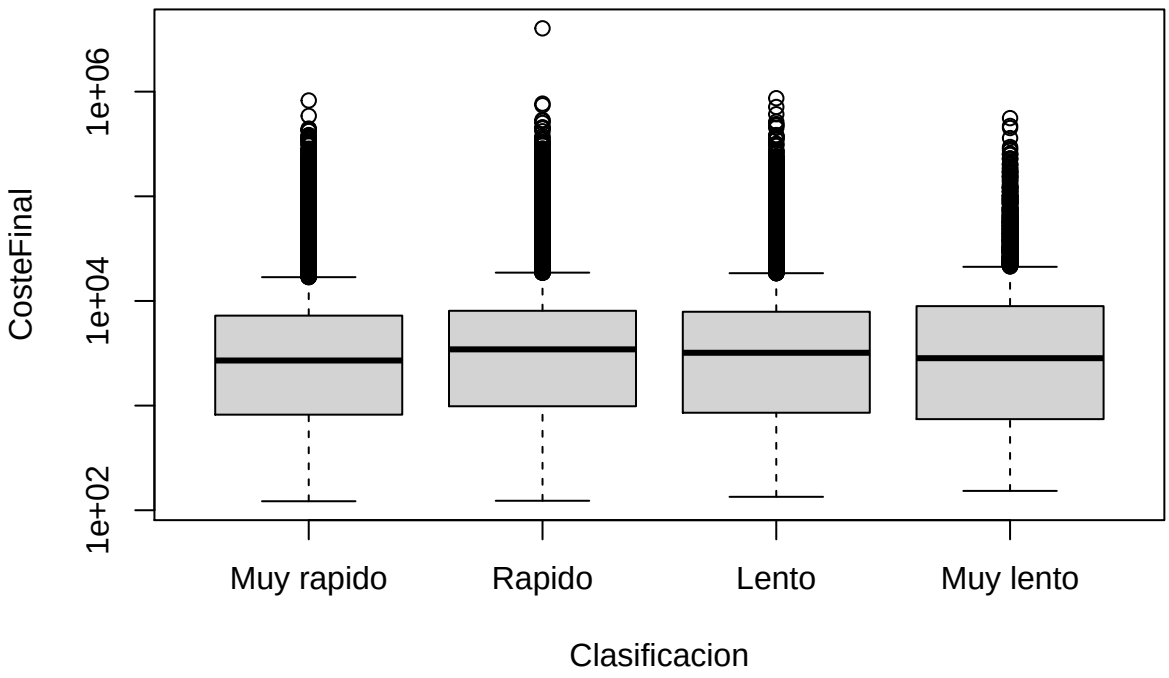
```
# Segun sexo
boxplot(formula = CosteFinal ~ Sexo, data=claimNET, log='y')
```



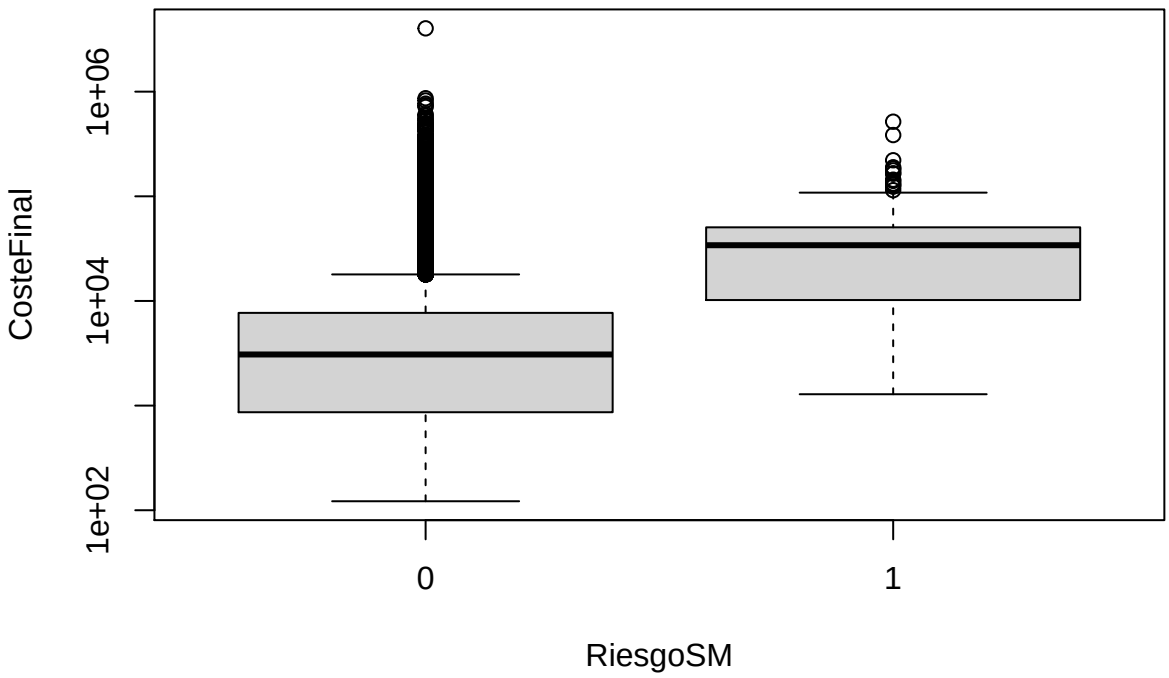

```
# Segun Estado
boxplot(formula = CosteFinal ~ Estado, data=claimNET, log='y')
```




```
#Segun Clasificacion
boxplot(formula = CosteFinal ~ Clasificacion, data=claimNET, log='y')
```




```
# Segun RiesgoSM
boxplot(formula = CosteFinal ~ RiesgoSM, data=claimNET, log='y')
```



Tal y como podemos observar en las representaciones anteriores, los datos están relativamente agrupados para todas las variables y parece que el dataset, en general, está relativamente limpio. No se aprecian en un primer momento registros que se encuentren muy alejados de los demás o de los cuartiles principales, estando también un gran número de registros agrupados en los extremos, sobre todo superior, de los costes. Por tanto, consideramos que a pesar de haber valores extremos, no se trata de valores anómalos y deberán ser tenidos en consideración a la hora de analizar los datos, por lo que no se procederá a eliminarlos.

4. Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar/comparar (p.e. si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?

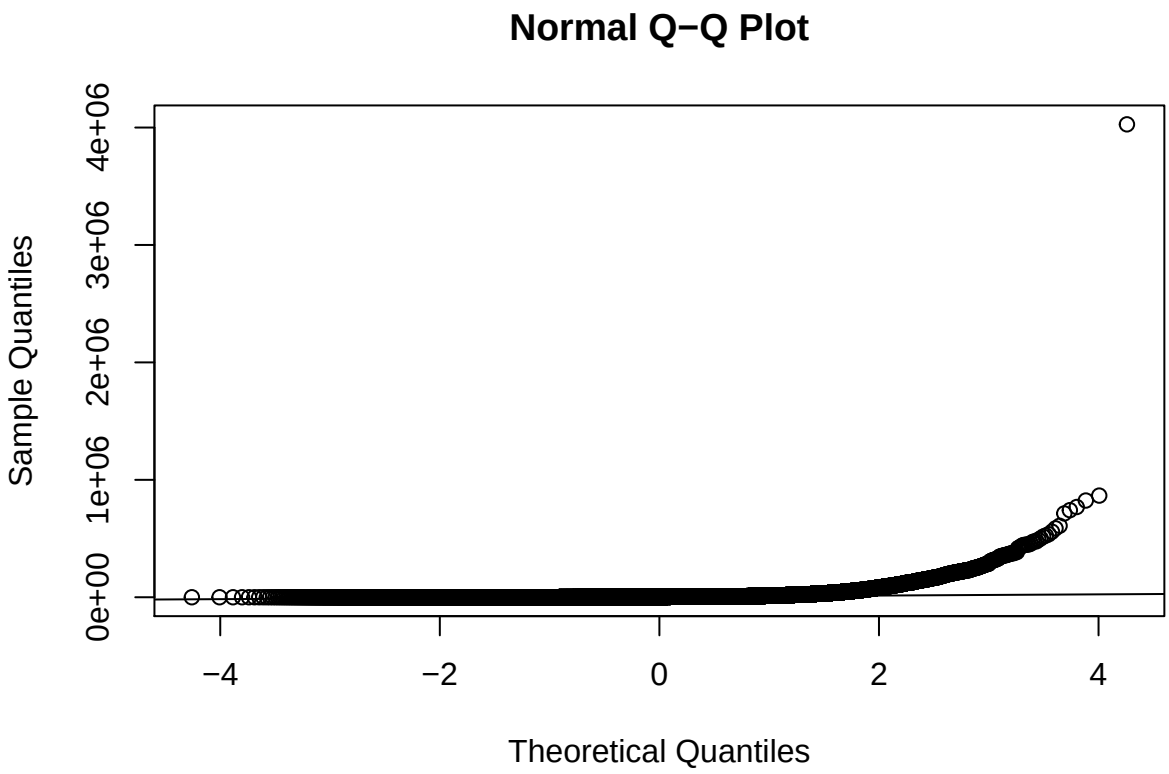
Se diferenciarán dos grupos de acuerdo a su **Sexo**, hombres y mujeres. La comprobación que se realizará es si la indemnización de las mujeres supera la de los hombres. Diferenciamos los grupos de la siguiente forma:

```
mujeres <- claimNET$CosteFinal[claimNET$Sexo=="F"]
hombres <- claimNET$CosteFinal[claimNET$Sexo=="M"]
```

4.2 Comprobación de la normalidad y la homogeneidad de la varianza

En este caso, estamos interesados en conocer si la variable **CosteFinal** sigue una distribución normal. Para la comprobación de la normalidad haremos uso de la representación qqnorm y qqline que nos darán una idea visualmente de cómo se distribuyen los datos. Afortunadamente, estas funciones son fácilmente implementables en R:

```
qqnorm(claimNET$CosteFinal)
qqline(claimNET$CosteFinal)
```



Como se puede observar, los datos no parecen distribuirse de manera normal, ya que de hacerlo la representación sería muy parecida a una recta inclinada.

Podemos realizar una segunda comprobación con el contraste de normalidad de Lilliefors, donde tendremos las siguientes hipótesis:

- H0 = Los datos no proceden de una distribución normal
- H1 = Los datos no proceden de una distribución normal

```
lillie.test(claimNET$CosteFinal)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  claimNET$CosteFinal
## D = 0.37797, p-value < 2.2e-16
```

Como se puede observar, el resultado es $p - value < 0.05$ por lo que podremos rechazar la hipótesis nula de normalidad de datos al igual que sucede en el análisis visual. Por tanto, podemos decir que la variable **CosteFinal no sigue una distribución normal**.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

El primer análisis que vamos a realizar es si podemos aceptar que la indemnización a las mujeres supera en más de 100€ a la de los hombres Plantearemos el siguiente contraste de hipótesis:

$H_0 : \mu_{Mujeres} - \mu_{Hombres} \leq 1000$

$H_1 : \mu_{Mujeres} - \mu_{Hombres} > 1000$

Donde μ se refiere a las respectivas medias poblacionales.

Para poder elegir el test a aplicar haremos en primer lugar un análisis de las varianzas para comprobar si ambos grupos presentan diferencias significativas:

```
# Realizamos el test de varianzas
var.test(mujeres, hombres)
```

```
##
##  F test to compare two variances
##
## data:  mujeres and hombres
## F = 3.2487, num df = 11255, denom df = 37414, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.153713 3.347389
## sample estimates:
## ratio of variances
##           3.248715
```

Al arrojar el test de varianzas un $p - value < 0.05$ rechazaremos la hipótesis de igualdad de varianzas y utilizaremos un test sobre la media de dos muestras independientes con una varianza desconocida y diferente entre ellas:

```
# Calculamos el estadístico de contraste, valor crítico y p-value

mujeres.n <- length(mujeres)
hombres.n <- length(hombres)

mujeres.mean <- mean(mujeres)
hombres.mean <- mean(hombres)

mujeres.sd <- sd(mujeres)
hombres.sd <- sd(hombres)

d <- (mujeres.sd^2/mujeres.n+hombres.sd^2/hombres.n)^2/(mujeres.sd^4/(mujeres.n^2*(mujeres.n-1))+hombres

d
```

```
## [1] 13401.45
```

```
s <- sqrt(mujeres.sd^2/mujeres.n + hombres.sd^2/hombres.n)
obs.value <- (mujeres.mean-hombres.mean-1000) / s
obs.value
```

```
## [1] 3.419415
```

```
pvalue <- pt(obs.value, df= d, lower.tail = FALSE)
pvalue
```

```
## [1] 0.0003147147
```

```
critic.value <- qt(0.05, df=d, lower.tail = FALSE)
critic.value
```

```
## [1] 1.644967
```

El pvalor obtenido ha sido 0.0001648202, por tanto $p - value = 0.0001648202 < 0.05$, lo que indica que se

rechaza la hipótesis nula. Por otra parte, el valor observado es 3.591609, que no se encuentra dentro de la zona de aceptación, dado que $3.591609 \notin [-\infty, 1.64495]$, de forma que se confirma el rechazo de la hipótesis nula y concluimos que en promedio, la indemnización de las mujeres es superior en 1000 euros a la de los hombres.

Una vez comprobada la diferencia entre hombres y mujeres, como segundo análisis, vamos a intentar generar un modelo de regresión lineal que pueda predecir el coste que va a tener un determinado accidente:

```
# Crear un modelo de regresión lineal múltiple: Edad, Sexo, Estado, Dependientes, OtrosDepend, salario,

# Convertimos CosteInicio a logaritmico
claimNET$CosteInicio_log = log(claimNET$CosteInicio)
claimNET$CosteFinal_log <- log(claimNET$CosteFinal)
model_3 <- lm(CosteFinal_log ~ Edad + Sexo + Estado + Dependientes + OtrosDepend + Salario + Jornada + H

summary(model_3)
```

```
##
## Call:
## lm(formula = CosteFinal_log ~ Edad + Sexo + Estado + Dependientes +
##     OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
##     Clasificacion + RiesgoSM + CosteInicio_log, data = claimNET)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2970 -0.4735 -0.1408  0.3238  7.3712
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.378e+00  4.636e-02  29.713  < 2e-16 ***
## Edad                    4.461e-03  3.456e-04  12.906  < 2e-16 ***
## SexoM                  -1.190e-01  8.451e-03 -14.075  < 2e-16 ***
## EstadoS                -4.378e-02  8.372e-03  -5.229 1.71e-07 ***
## Dependientes            4.039e-02  6.620e-03   6.100 1.07e-09 ***
## OtrosDepend             7.298e-02  3.097e-02   2.356  0.01846 *
## Salario                 5.763e-04  1.527e-05  37.746  < 2e-16 ***
## JornadaP                4.138e-02  1.561e-02   2.651  0.00802 **
## HorasSemana             4.431e-04  2.962e-04   1.496  0.13467
## DiasSemana             -4.656e-02  8.342e-03  -5.581 2.40e-08 ***
## ClasificacionRapido     2.537e-02  8.578e-03   2.957  0.00311 **
## ClasificacionLento      1.533e-02  8.984e-03   1.706  0.08801 .
## ClasificacionMuy lento  9.149e-03  1.338e-02   0.684  0.49412
## RiesgoSM1               2.672e-01  5.158e-02   5.181 2.22e-07 ***
## CosteInicio_log         8.415e-01  2.440e-03 344.831  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7527 on 48656 degrees of freedom
## Multiple R-squared:  0.7552, Adjusted R-squared:  0.7551
## F-statistic: 1.072e+04 on 14 and 48656 DF,  p-value: < 2.2e-16
```

Para hacernos una idea de la calidad del modelo, observaremos el valor de R-squared, que presenta un valor de 0.75, lo que indica que el modelo es claramente mejorable pero no del todo desacertado.

Finalmente, como tercer análisis, vamos a realizar un **análisis ANOVA** que contraste si existen diferencias en la variable **CosteFinal** de acuerdo a la **Clasificación** creada en relación al tiempo entre apertura de

incidencia y pago creada al principio.

El factor Clasificacion tiene 4 niveles: 1 Muy lento, 2 Lento, 3 Rápido y 4 Muy rápido. Las hipótesis son:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 : \mu_i \neq \mu_j$ para algún i, j

donde $\mu_1, \mu_2, \mu_3, \mu_4$ denotan, la media poblacional de CosteFinal para las distintas clasificaciones Muy lento, Lento, Rápido y Muy rápido.

Creamos el modelo ANOVA, En primer lugar, con las hipótesis anteriores, vemos si podemos asumir la igualdad de medias entre los cuatro grupos:

```
#aov
Model.5.2.aov <- aov(log(CosteFinal) ~ Clasificacion, claimNET)
kk <- summary(Model.5.2.aov)
kk
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## Clasificacion     3    198   66.11   28.62 <2e-16 ***
## Residuals     48667 112407    2.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#lm
Model.5.2.lm <- lm(log(CosteFinal) ~ Clasificacion, claimNET)
anova(Model.5.2.lm)
```

```
## Analysis of Variance Table
##
## Response: log(CosteFinal)
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Clasificacion     3    198  66.106  28.621 < 2.2e-16 ***
## Residuals     48667 112407   2.310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor (Pr(> F)) es prácticamente 0, menor a 0.05 lo que indica que, el factor analizado es significativo. En conclusión, rechazamos la hipótesis nula de igualdad de medias entre los cuatro grupos del factor.

Estimemos el efecto de cada uno de los niveles:

```
mu <- mean(claimNET$CosteFinal); mu
```

```
## [1] 10456.68
```

```
alpha1 <- mean(claimNET$CosteFinal[claimNET$Clasificacion=="Muy lento"])-mu
alpha2 <- mean(claimNET$CosteFinal[claimNET$Clasificacion=="Lento"])-mu
alpha3 <- mean(claimNET$CosteFinal[claimNET$Clasificacion=="Rapido"])-mu
alpha4 <- mean(claimNET$CosteFinal[claimNET$Clasificacion=="Muy rapido"])-mu
alpha1; alpha2; alpha3; alpha4
```

```
## [1] 1299.148
```

```
## [1] 235.6857
```

```
## [1] 588.3672
```

```
## [1] -1219.519
```

El efecto de Clasificacion es negativo para los clasificados como Muy rápido mientras que es positivo para los clasificados como Rápido, Lento y Muy lento.

5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

Desarrollado a lo largo de la práctica.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Las conclusiones finales son que este dataset permite la estimación del precio a pagar por las aseguradoras en función de sus variables, aunque el modelo conseguido debería ser mejorado para poder emplearlo realmente, dado su R-squared.

Por otra parte, hemos concluido que existen diferencias entre la indemnización de mujeres y hombres.

Finalmente, hemos visto que el tiempo que se tarda desde que se abre la incidencia hasta que resuelve influye en el coste final a pagar.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python

Adjuntado en el github.

Finalmente, extraemos el conjunto de datos final:

```
write.csv(claimNET, "dataset_final.csv")
```