Regresión Logística Múltiple

Link al dataset: https://www.kaggle.com/c/titanic/data?select=train.csv

Las variables para usar son:

|  |  |  |
| --- | --- | --- |
| **Nombres variables** | **Definición** | **Valores** |
| sobrevivio | Sobrevivio al accidente | 0 = No, 1 = Si |
| clase | Clase del ticket | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sexo | Sexo |  |
| edad | Edad en años |  |
| her\_esp | Cantidad de hermanos/conyuges en el Titanic |  |
| padre\_hijo | Cantidad de padres/hijos en el Titanic |  |
| ticket | Numero de ticket |  |
| precio\_ticket | Precio del ticket |  |
| nro\_cabina | Número de cabina |  |
| puerto\_embarcacion | Puerto de embarcación | C = Cherbourg, Q = Queenstown, S = Southampton |

# Exploración de datos

Empezamos viendo las medidas representativas de cada variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | sobrevivio | clase | edad | her\_esp | padre\_hijo | precio\_ticket |
| Min | 0 | 1 | 0.42 | 0 | 0 | 0 |
| 1st Qu. | 0 | 2 | 20.12 | 0 | 0 | 7.91 |
| Median | 0 | 3 | 28.00 | 0 | 0 | 14.45 |
| Mean | 0.3838 | 2.309 | 29.70 | 0.523 | 0.3816 | 32.20 |
| 3rd Qu. | 1 | 3 | 38.00 | 1 | 0.0000 | 31.00 |
| Max. | 1 | 3 | 80.00 | 8.000 | 6.0000 | 512.33 |
|  |  |  | NA's :177 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| sexo | ticket | nro\_cabina | puerto\_embarcacion |
| Length:891 | Length:891 | Length:891 | Length:891 |
| Class :character | Class :character | Class :character | Class :character |
| Mode :character | Mode :character | Mode :character | Mode :character |

Lo que se observa es:

* sobrevivio no lo trata como un factor.
* clase no lo trata como un factor.
* sexo no lo trata como un factor.
* nro\_cabina no lo trata como un factor.
* Ticket no lo trata como un factor.
* Puerto\_embarcacion no lo trata como un factor.
* En edad hay 117 NAs

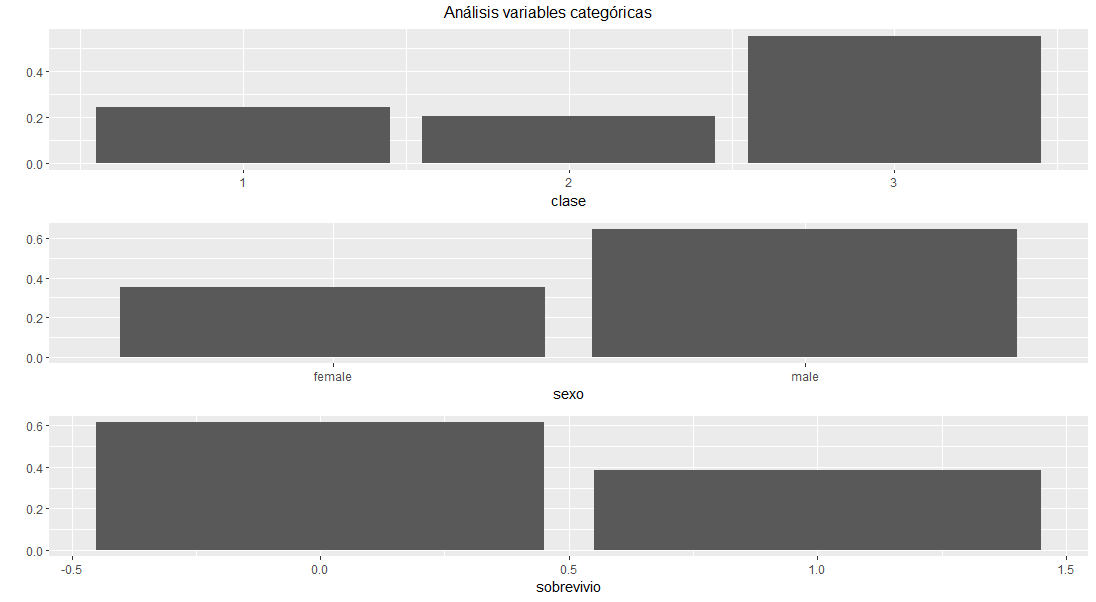
Para las variables que no se tratan como factores las convertimos a factores. Luego, para los valores de edad con NA lo que hacemos es aproximarlo con la media de los valores restantes. No usamos la mediana porque no se presentan muchos outliers entonces no es necesario usar la mediana. Por último, para los valores sin nivel de puerto\_embarcacion decidimos no modificarlos ya que no parece haber inconsistencia en los datos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | edad | her\_esp | padre\_hijo | precio\_ticket |
| Min | 0.42 | 0 | 0 | 0 |
| 1st Qu. | 22 | 0 | 0 | 7.91 |
| Median | 29.7 | 0 | 0 | 14.45 |
| Mean | 29.7 | 0.523 | 0.3816 | 32.2 |
| 3rd Qu. | 35 | 1 | 0 | 31 |
| Max. | 80 | 8 | 6 | 512.33 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sobrevivio | clase | sexo | puerto\_embarcacion | nro\_cabina | ticket |
| 0: 549 | 1: 216 | female: 314 | : 2 | : 687 | 1601: 7 |
| 1: 342 | 2: 184 | male : 577 | C: 168 | B96 B98: 4 | 347082: 7 |
|  | 3: 491 |  | Q: 77 | C23 C25 C27: 4 | CA. 2343: 7 |
|  |  |  | S: 644 | G6: 4 | 3101295: 6 |
|  |  |  |  | C22 C26: 3 | 347088: 6 |
|  |  |  |  | D: 3 | CA 2144: 6 |
|  |  |  |  | (Other): 186 | (Other): 852 |

* edad presenta valores con decimales, esto puede ser porque son bebes menores a un año o porque se está aproximando su edad, en cuyo caso se presenta en la forma de xx.5. Además, se reemplazaron los valores NA con la media de los valores restantes, y por esto es recurrente el valor 29.7.
* puerto\_embarcacion presenta dos valores sin categoría.
* nro\_cabina presenta 687 valores sin categoría.

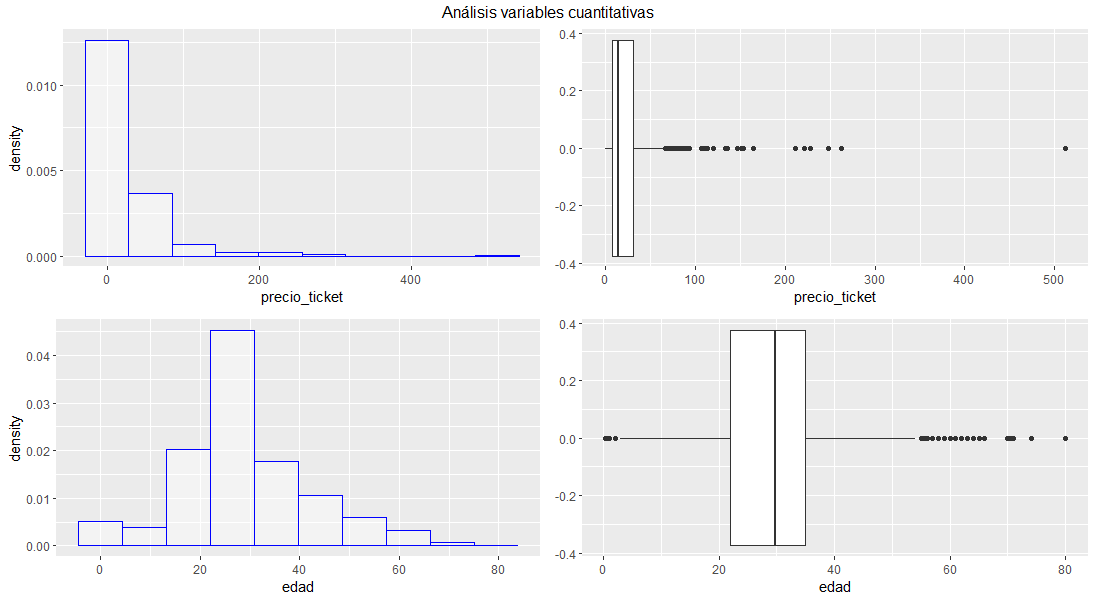
Ahora vamos a analizar las variables **clase,** **sexo**, **sobrevivio**, **edad** y **precio\_ticket**.



Con respecto a **clase**, la mayoría de los pasajeros pertenecían a la tercera clase, siendo estos más del 50%. Luego, le siguen los pasajeros de primera clase y por último los de segunda clase.

Con respecto a **sexo**, la mayoría de los pasajeros eran masculino, con mas del 60% del total.

Con respecto a **sobrevivio**, la mayoría de las personas no sobrevivieron, con aproximadamente el 60% del total.



**precio\_ticket** presenta una gran cantidad de muestras en los primeros grupos y estas disminuyen a medida que aumenta el precio. Además, en el boxplot se puede ver la existencia de varios outliers, siendo el más significativo la observación que se encuentra por encima de los $500.

**edad** presenta una distribución que pareciera ser normal, teniendo gran cantidad de muestras cerca de las edades que se encuentran próximas a los 30 años. En el boxplot se puede ver como hay presencia de outliers, siendo el más alejado un pasajero de 80 años.