# Ejemplo de ajuste de distribución

Vásquez Guerra Carlos Fernando

El archivo `payments.csv` contiene información de clientes que hicieron pagos por una deuda del mismo tipo, es de interés para la compañía modelar los pagos que se están realizando.

La base proporcionada tiene 6 variables en total las siguientes se enlistan a continuación:

- `id` : Identificador de la persona que realiza el pago.
- `Gender`: Genero del cliente (`male` o `female`).
- `Age`: Edad del cliente.
- `Payment.Method` : Métodod de pago que uso el usuario (`credit card`, `cheque` o `cash`).
- `Churn` : Identifica si el cliente se ha retirado de la empresa (`churn`) o sigue con ella (`loyal`).
- `Payment` : Cantidad monetaria que realizo el usuario.

| id | Gender | Age | Payment.Method | Churn | Payment |
|---|---|---|---|---|---|
| 1 | male | 64 | credit card | loyal | 91.74816 |
| 2 | male | 35 | cheque | churn | 79.20001 |
| 3 | female | 25 | credit card | loyal | 90.15703 |
| 4 | female | 39 | credit card | NA | 69.85222 |
| 5 | male | 39 | credit card | loyal | 79.12670 |
| 6 | female | 28 | cheque | churn | 74.03494 |

Tratando a los datos de una forma más exhaustiva, se tienen las siguientes características:

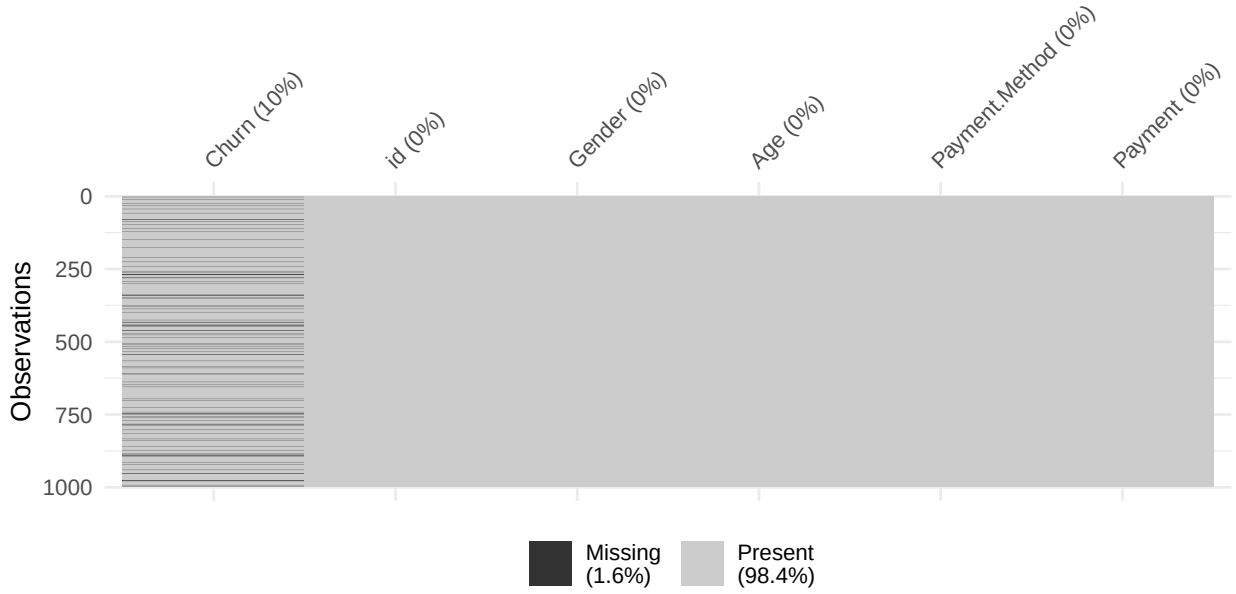
| | Unicos | %Unicos | Valores Perdidos | Moda | Tipo |
|---|---|---|---|---|---|
| Gender | 2 | 0.2008032 | 0 | male | character |
| Payment.Method | 3 | 0.3012048 | 0 | credit card | character |
| Churn | 3 | 0.3012048 | 96 | loyal | character |

| | Unicos | %Unicos | Mínimo | Máximo | Media | Desviación estandar | Tipo |
|---|---|---|---|---|---|---|---|
| id | 996 | 100.000000 | 1.00000 | 996.0000 | 498.50000 | 287.66474 | double |
| Age | 74 | 7.429719 | 17.00000 | 91.0000 | 45.61647 | 18.77675 | double |
| Payment | 996 | 100.000000 | 48.50817 | 111.4355 | 84.64396 | 9.49986 | double |

Hay que destacar algunos puntos importantes:

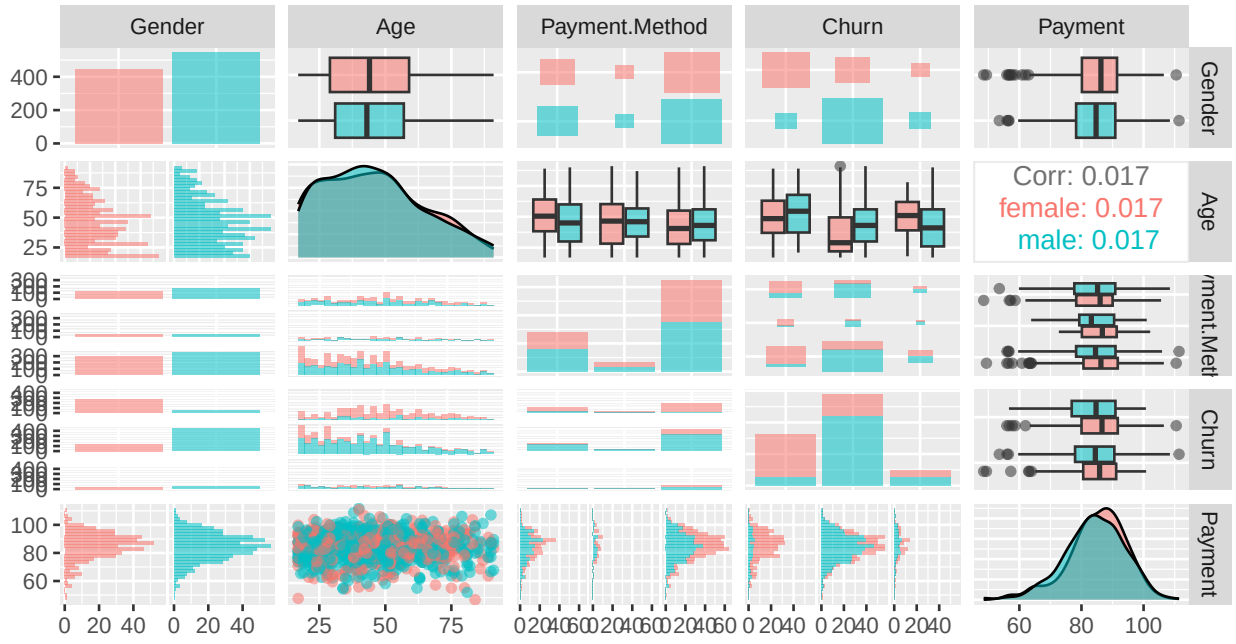
- No importa que existan valores nulos en la variable `Churn` ya que no es la variable de interés, al menos para las instrucciones dadas.
- De manera rápida identificamos que la mayoría de clientes son hombres, tiene 42 años, al menos 17 y máximo 91 años, utilizan tarjeta de crédito, todos los usuarios tienen han realizado pagos diferentes y, en promedio, los usuarios deben $84.64396 \pm 9.49986$ unidades monetarias.

De manera general, esta la información que se tienen de los datos nulos sobre las variables.




No es de nuestro interés, al menos por el momento, arreglar los problemas con los valores perdidos, por que eso es todo lo que se mencionará sobre estos; además, no afectan a la variable que es de nuestro interés en este momento.

De manera rápida, para ver algunas relaciones y comportamientos de las variables dos a dos; se tiene la siguiente gráfica donde las gráficas de color rojo representan a los clientes de genero femenino y las de color azul para los hombres.

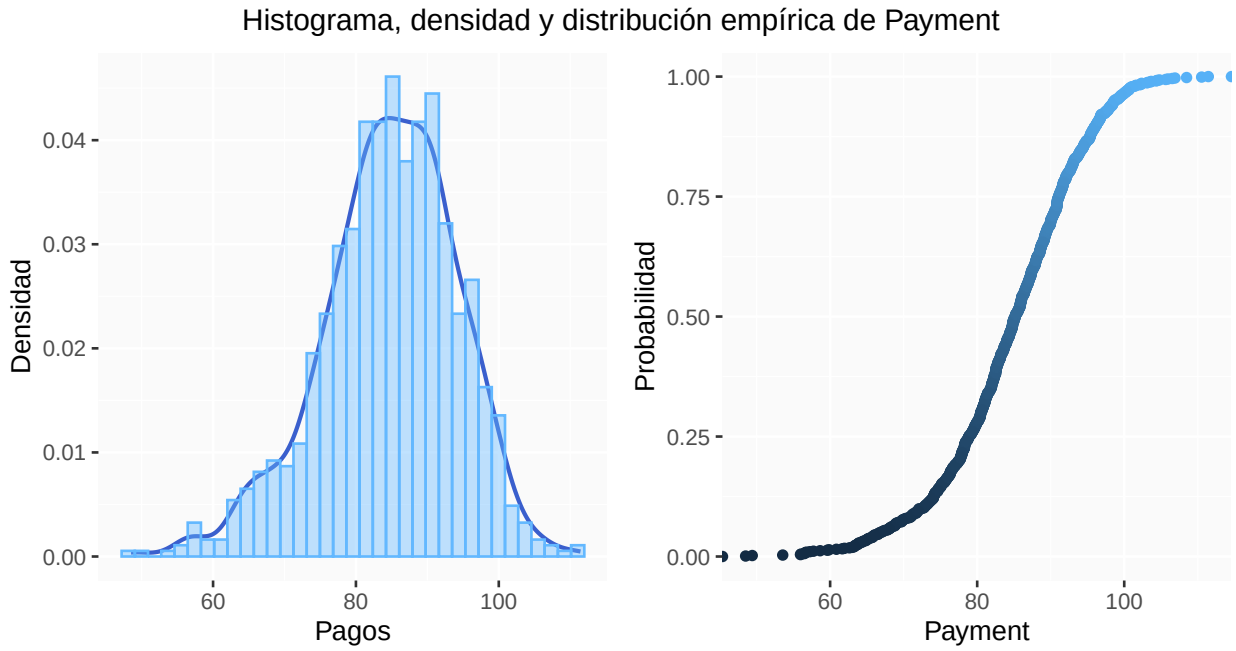


Con esto, podemos ver la siguiente información de manera más clara:

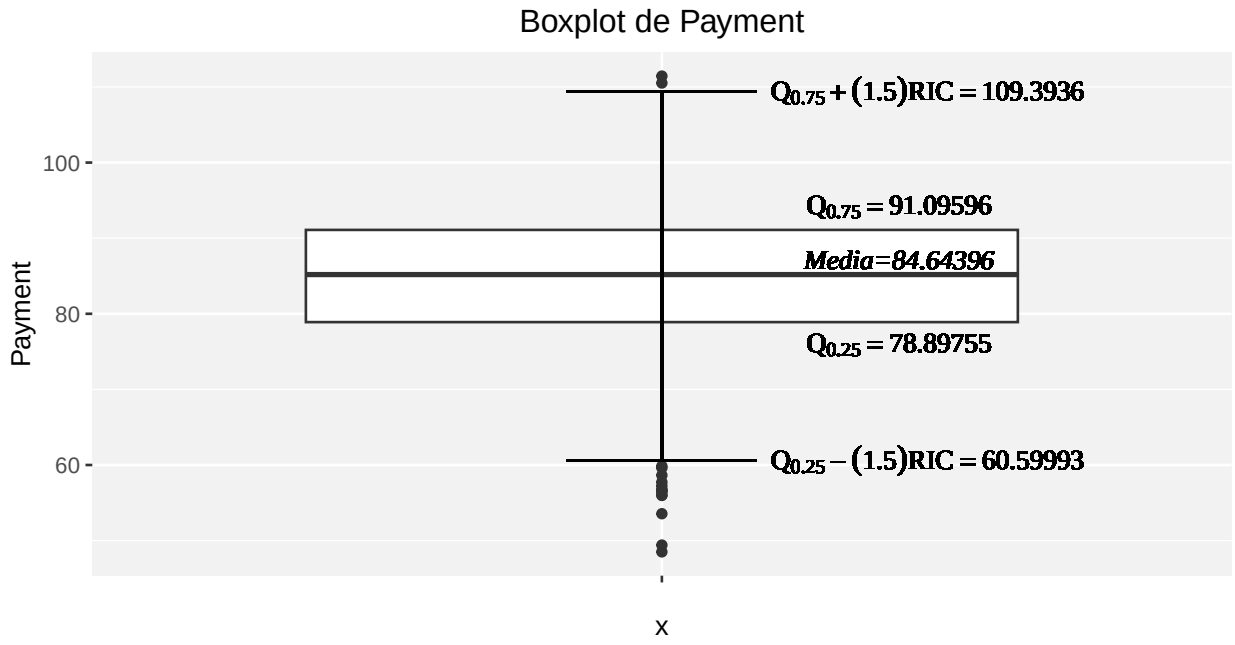
- La cantidad de hombres y mujeres solo varia por cien clientes.
- No hay una tendencia significativa que indique que las mujeres o los hombres pagan deudas más grandes que el otro género y lo mismo sucede con el método de pago y la edad.

Es de principal interés analizar el comportamiento de los pagos, por lo que se tiene una ampliación de la densidad de estos a continuación y más gráficas que serán de ayuda

Primero, veremos un histograma, densidad y distribución empirica de los pagos




Ahora, un boxplot




Ahora, nos interesa buscar de forma **empírica** qué variable aleatoria está detrás de los pagos de los clientes y así proponer una distribución.

Antes que nada, veamos algunas características que debe tener la variable aleatoria que propongamos:

- Valores continuos mayores a cero, ya que son precios y ningún precio en esta base de datos es negativo.
- Se tienen las siguientes estadísticas (aunque algunas ya se pudieron apreciar en las gráficas anteriores):

1. ***Media*** = 84.64396
2. ***Mediana*** = 85.1851347
3. ***Moda*** = 82.5
4. ***Varianza*** = 90.2473479
5. ***Desviación estandar*** = 9.4998604
6. ***Rango intercuantil*** = [78.8975482, 91.0959555]
7. ***Coeficiente de variación*** = 11%
8. ***Coeficiente de asimetría (Skewness)*** = -0.4685155
9. ***Curtosis*** = 3.4315899

Con los datos y las gráficas anteriores podemos resumir lo siguiente:

- Al ser los valores mayores a cero y continuos, descartamos cualquier variable discreta (como *poisson*, etc) y cualquiera que tenga un rango con valores negativos (como la *normal* y la *t de Student*), además de que se tienen valores mayores a uno por lo que se descartan otras distribuciones como la *beta*
- La densidad no tiene colas pesadas, en todo caso, la cola donde tiene la mayor cantidad de outliers es la cola izquierda (notable de ver en cualquiera de los tres gráficos anteriores); por lo que descartamos distribuciones con colas muy pesadas (como la *log-Normal*, *Pareto* y *Burr*)
- Al tener una curtosis positiva, inidica la presencia, como se puede ver en la gráfica de la densidad, de un pico y al tener un coeficiente de variación del 11%; los valores no varían demasiado de la media, lo cual, generalmente no cumple la distribución *exponencial*, por lo que proponemos una función gamma o weibull para ajustar estos datos.

Para la distribución gamma, se utilizaron los estimadores por momentos y para la función weibull, los estimadores máximo verosimil. Ambos con la función `fitdistrplus::fitdist()`

- Gamma: forma = 79.4682621; rate = 0.9388534.
- Weibull: forma = 10.281482; escala = 88.7567643.

A manera de resumen, se tiene lo siguiente:

| Distribución | Prueba | Estadístico | P-Value |
|---|---|---|---|
| Gamma | Anderson-Darling | 5.8184513 | 0.0011751 |
| Gamma | Kolmogorov-Smirnoff | 0.0497377 | 0.0144837 |
| Weibull | Anderson-Darling | 1.0908218 | 0.3129271 |
| Weibull | Kolmogorov-Smirnoff | 0.0246616 | 0.5798048 |

Por lo que elegimos la distribución Weibull como la distribución que mejor se ajusta a los datos; ya que, con esta distribución, no existe evidencia estadística suficiente para reachazar la hipótesis nula, es decir, no se recahza que no siguen una distribución weibull. Siendo más específicos sobre la distribución, se propone el siguiente modelo:

$$Weibull(x; \alpha = 10.28148, \beta = 88.75676) = \frac{\alpha}{\beta}(\frac{1}{\beta}x)^{(\alpha-1)}e^{-(\frac{1}{\beta}x)^{\alpha}} = \frac{10.28148}{88.75676}(\frac{1}{88.75676}x)^{(10.28148-1)}e^{-(\frac{1}{88.75676}x)^{10.28148}}$$

Finalmente, como algo extra; fijando una semilla ,`set.seed(9)`, y realizando simulaciones sobre la weibull propuesta, se tiene la siguiente comparación de densidades