Universidad de Costa Rica

Facultad de Ciencias

Escuela de Matemática

Departamento de Matemáticas y Ciencias Actuariales

Herramientas de Ciencias de Datos I

Análisis de los sobrevivientes al hundimiento del Titanic

Alessandro Umaña Vega C37963

Joshua Cervantes Artavia

Grupo 2

II Ciclo

Viernes 14 de noviembre, 2025

Este documento presenta un análisis exploratorio del dataset Titanic disponible en R con la librería "Titanic". El objetivo es aplicar técnicas de manipulación, limpieza e imputación de datos para facilitar su estudio, además de realizar análisis descriptivo de las estadísticas y visualizaciones para comprender factores de importancia que pudieron influir en la supervivencia de los pasajeros.

Primero se explicará un poco acerca del dataset, que se logró importar por la librería readr. Este trabajo se concentró en las variables numéricas y de tipo factor, ya que las variables de tipo carácter indican el nombre del pasajero, el número de tiquete y el lugar donde abarcaron, los cuales no son de importancia para el trabajo. Sin embargo, las variables numéricas y de factor consta de clase del pasajero, edad, sexo, número de hermanos, padres o hijos abordo, si sobrevivió o no, las cuales se usarán para estudiarlas. Al examinar la base de datos se observan valores faltantes en la variable de edad, los cuales se trabajarán más adelante. También se cambió las variables sexo, clase del pasajero y sobrevivencia a factor.

Después se trabajaron los valores faltantes en la variable edad, por lo que se puede observar en la Tabla 1 hay una cantidad de 177 valores faltantes el cual es casi un 20% de las entradas totales, por lo que se considera importante imputarlas y no eliminarlas. Para imputarlas se decide hacerlo por la media de edad por la clase de pasajeros para mantener la estructura de la base, además de que los pasajeros de cada clase van a tener un nivel socioeconómico parecido y junto con este un rango de edad que varía, pues es menos probable que una persona de 18 años tenga el poder adquisitivo para poder adquirir un tiquete de primera clase que una persona de 30 años. Entonces se calcula la mediana de la edad por cada clase y después con una función "for" e "if" se recorre fila por fila y si la edad es un NA el pasajero pertenece a una clase específica y se reemplaza con la mediana específica, evitando alterar otros datos y mantener la información de la base de datos. Además de hacer uso de la librería "dplyr" para usar la función "select" y corregir la base de datos.

Tabla 1: Valores faltantes en edad por sexo

	False	True
Mujer	261	53
Hombre	453	124

Por otro lado, también se compararon algunas variables haciendo uso de la librería de "ggplot2" y "tidyverse", usando sus diferentes funciones para lograr crear gráficos de barras, de puntos con dispersión, de densidad e histogramas. De primero se tiene un gráfico comparando la supervivencia por sexo y como se observa en el Grafico 1 las mujeres tuvieron una tasa mucho más alta de supervivencia que los hombres de alrededor de un 50% de diferencia entre las tasas de supervivencia, lo cual hace sentido, ya que se opta en este tipo de emergencias primero salvar mujeres y niños y después los hombres.

También se puede analizar la supervivencia por clase del pasajero, para determinar si el nivel socioeconómico afectaba la supervivencia en este accidente. En el Gráfico 2 se observa que efectivamente hay una mayor tasa de supervivencia entre más alta fuera la clase

del pasajero. Lo cual puede indicar que aquellos que tenían una clase más alta tenían algún tipo de prioridad en el rescate, así como una mejor ubicación dentro del barco.

Gráfico 1: Supervivencia por sexo

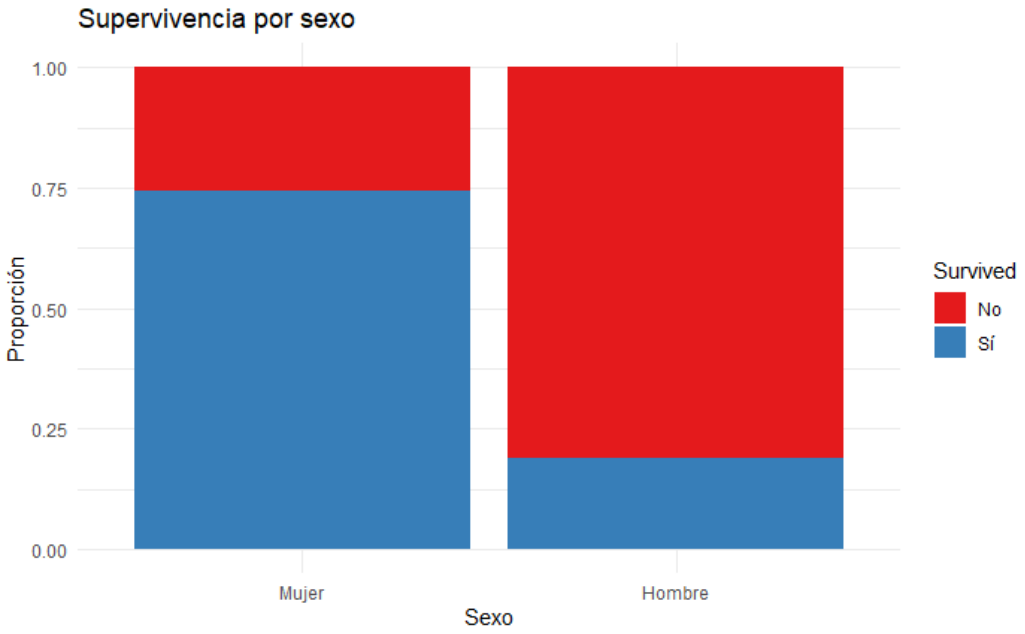
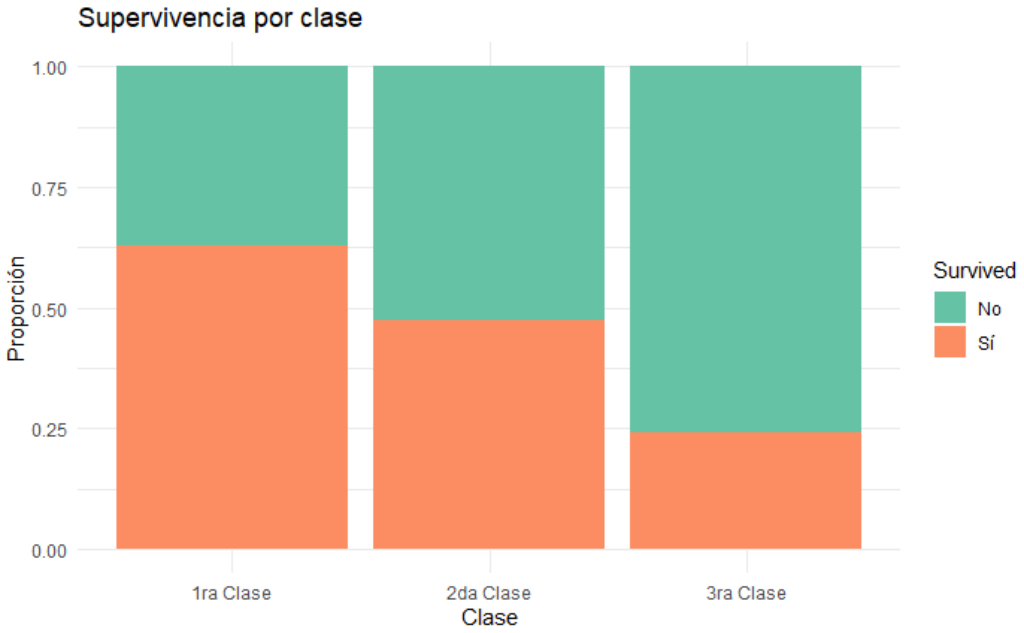
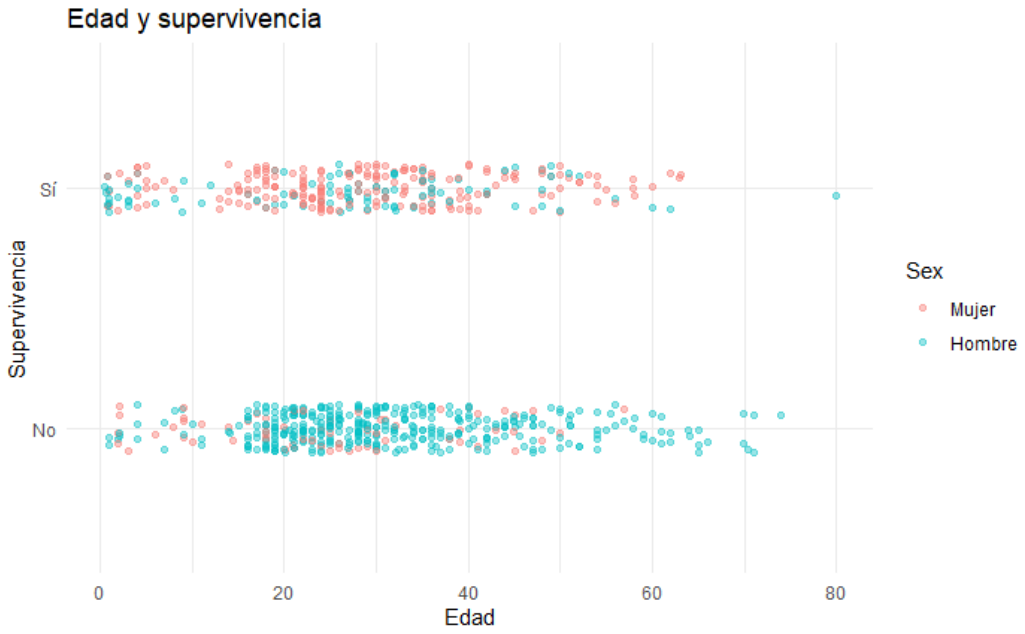


Gráfico 2: Supervivencia por clase



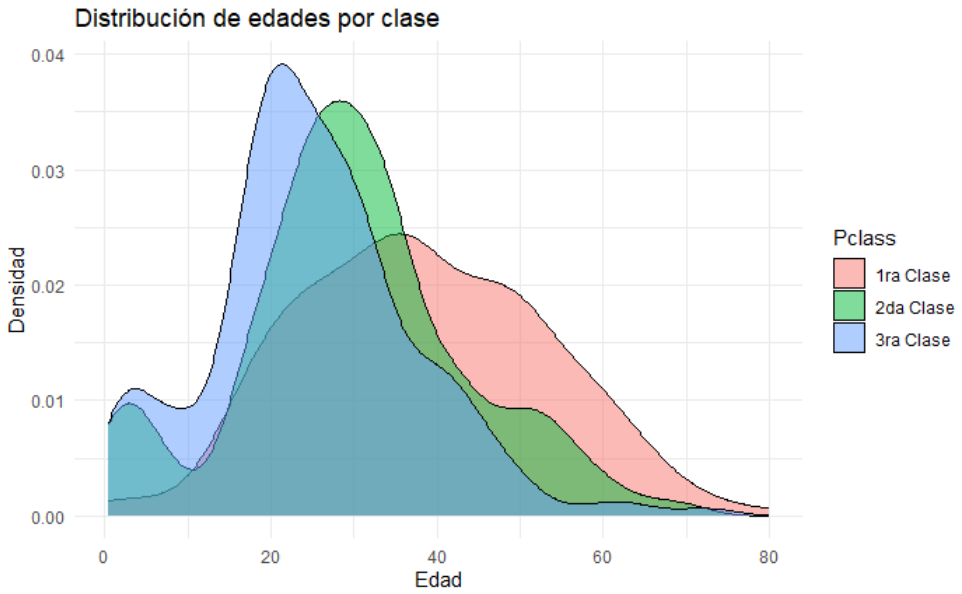
Ahora se analizará la relación entre la supervivencia, edad y sexo, dondeen el Gráfico 3 se puede concluir que los niños muestran una ligera mayor probabilidad de supervivencia, sin embargo, la relación no es tan marcada como lo es con la clase y el sexo. Esto sugiere que la edad no fue un factor tan determinante a la hora de evacuar a los pasajeros como lo pudo ser la clase o el sexo.

Gráfico 3: Supervivencia por edad



Por último, se verá la distribución de la edad por clases, donde se observa en el Gráfico 4 que en la tercera clase hay mayor concentración alrededor de los 20 años y existe un pico en tempranas edades, para la segunda clase hay mayor densidad en los 25-30 años y también varias personas entre los 40-60 años están dentro, pero en menor magnitud que la primera clase y para la tercera clase tiene una distribución muy dispersa, pero se concentra en los 30-40 años. Por lo que se puede intuir que entre mayor edad tenían mayor poder adquisitivo para poder en primera clase y las personas jóvenes se concentraban en la tercera clase.

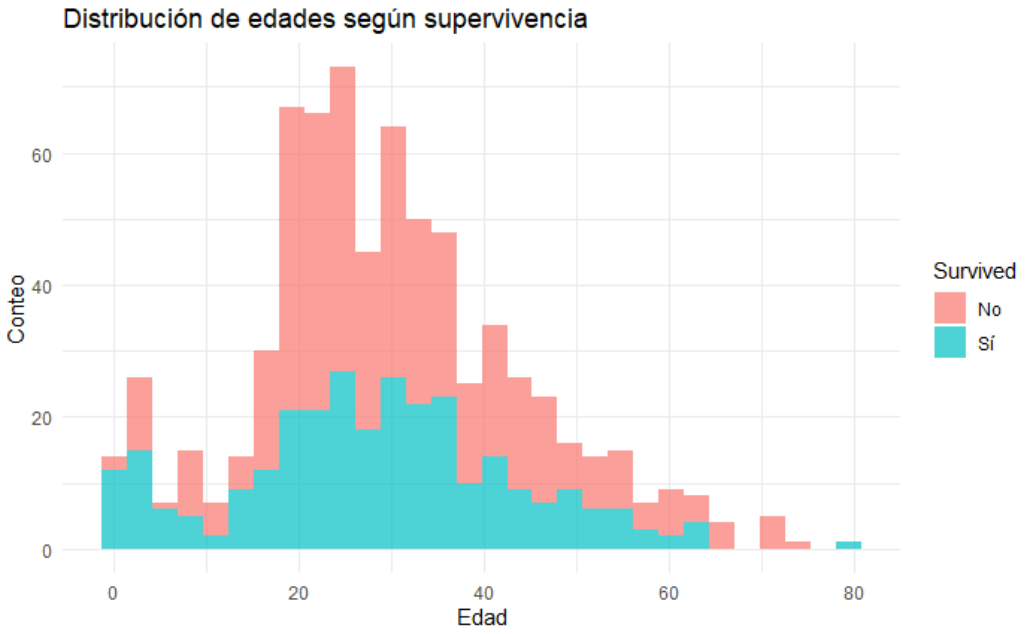
Gráfico 4: Distribución de edades por clase



Por último, se analiza la edad con la supervivencia, donde se observa en el Gráfico 5que hubo la mayor cantidad de supervivencia la tuvieron las personas alrededor de los 20 a los 35 años y las personas menores y mayores a este rango tuvieron menos supervivientes, sin embargo, la mayor cantidad de no sobrevivientes también está en el mismo rango, por lo que se puede asumir que su alta tasa de supervivencia es debido a la alta cantidad de personas que estaban en este rango de edad. También se puede observar que

en magnitud la cantidad de niños y perosnas mayores de edad sobrevivientes es mayor a la cantidad que no sobrevivieron.

Gráfico 5: Distribución de edades según supervivencia



También se hizo una correlación entre las variables, para identificar qué factores influyeron en la supervivencia, además de interpretar otros patrones. Con la función "cor" se puede lograr esto, donde solo se tiene ingresar las variables numéricas de la base de datos trabajada, en caso de que existan valores faltantes se puede usar "use = "complete.obs"" para ignorarlas. En la Tabla 2 se puedeobservar varias relaciones entre las variables.

Se empieza analizando la variable de supervivencia donde se observa que presenta correlaciones débiles o moderadas con casi todas las variables, en excepción con la clase y la tarifa. Donde con la clase tiene un –0.3384 que indica que las personas de clases altas tuvieron un mayor índice de supervivencia, esto se puede explicar por su cercanía a los botes salvavidas y condición de alojamientos más favorables. De la misma manera con la tarifa (0.2573) indica que las personas que pagaron una tarifa más alta tuvieron mayores posibilidades de sobrevivir. Lo cual hace sentido por la gran relación entre la tarifa que se paga con la clase en la que están.

Respecto a la edad se observa que la relación con la supervivencia es muy débil (–0.0472) lo que indica que, aunque los niños tuvieran la prioridad había en menos magnitud que las personas adultas. Sin embargo, la edad si presenta mayor correlación con otras variables, como con la clase (–0.4085) donde en promedio las personas de tercera clase eran más jóvenes mientras los de primera clase incluían una mayor cantidad de personas adultas mayores. Además de una correlación negativa con la cantidad de hermanos (-0.2435) y la cantidad de papas o hijos (-0.1711), lo que indica que las personas con más acompañantes suelen ser personas jóvenes, lo cual hace sentido por la presencia de familias completas.

También, las variables de cantidad de hermanos y de padres o hijos poseen una correlación fuerte (0.4148), lo cual es coherente pues como se indicó anteriormente existían familias viajando juntos. Además, ambas tienen una relación positiva con la tarifa (0.1596 para hermanos y 0.2162 para padres o hijos), sugiriendo que los grupos numerosos pagaban tarifas más altas.

Otra relación importante es la de clase y tarifa (-0.5495), lo cual confirma que entre más alta la tarifa, más alta la clase del pasajero. Esto también confirma que haber pagado una mayor tarifa supone mayor supervivencia, pues pagar más significaba estar ubicado en una zona más privilegiada.

Tabla 2: Matriz de correlaciones

	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.00000000	-0.33848104	-0.04725461	-0.03532250	0.08162941	0.2573065
Pclass	-0.33848104	1.00000000	-0.40848699	0.08308136	0.01844267	-0.5494996
Age	-0.04725461	-0.40848699	1.00000000	-0.24352572	-0.17109485	0.1237837
SibSp	-0.03532250	0.08308136	-0.24352572	1.00000000	0.41483770	0.1596510
Parch	0.08162941	0.01844267	-0.17109485	0.41483770	1.00000000	0.2162249
Fare	0.2573065	-0.5494996	0.1237837	0.1596510	0.2162249	1.00000000

Entonces se puede concluir que la matriz de correlación evidencia que la clase social, la tarifa pagada y la composición familiar influyen en variables de edad y de supervivencia. La supervivencia se vio marcada más por factores socioeconómicos y que no dependían tanto de la edad, sino de estructuras sociales dentro del barco, tales como la ubicación de las habitaciones, acceso a los botes y normas de evacuación.

El análisis del conjunto de datos del Titanic revela que la supervivencia estuvo fuertemente influenciada por factores socioeconómicos, especialmente la clase del pasajero y la tarifa pagada. Los resultados muestran que los pasajeros de primera clase tuvieron una mayor tasa de supervivencia, mientras los que viajaban en tercera clase una tasa mucho menor. Esta relación se confirma por mediante la correlación negativa entre la clase y la supervivencia y la correlación positiva entre la tarifa y la supervivencia. En términos familiares, se observa que aquellos pasajeros que viajaban en familia reflejan dinámicos a

bordo, pues las familias más grandes pagaron tarifas más altas. También se demostró que la supervivencia y la edad no fue un factor tan importante, pues, aunque los niños tuvieran buena tasa de supervivencia la mayor parte de los pasajeros eran personas jóvenes. Además, la distribución de edades presenta que las personas jóvenes se concentraban en la terca clase, mientras las personas de primera clase eran más mayores. En conclusión, el acceso a botes, la ubicación dentro del barco y el estatus socioeconómico fueron factores determinantes de la supervivencia.

Por último, se recomienda usar imputación segmentada si existe alguna relación clara con otras variables. También, el uso de gráficos ayuda a entender mejor la relación que se tienen entre sí, así como un análisis de correlación. En términos de presentación complementar el análisis numérico con visualizaciones más detalladas como curvas de densidad segmentadas, así como profundizar en modelos predictivos más complejos para cuantificar el impacto combinado de diferentes variables. Además de las limitaciones de la base de datos, ya que no incluye variables que podrían ser importantes, como cuantas familias había, de cuantas personas eran cada familia, que tan riesgoso era la habitación en la que estaban. Estas recomendaciones permitirían un mejor análisis