TIC

ACTIVIDAD #2

Tipo actividad: análisis exploratorio con la base de datos del titanic

Aplicando lo aprendido, análisis exploratorio con la base de datos del titanic (6 horas).

Para realizar el análisis exploratorio con un enfoque más formal, inicar respondiendo a las preguntas:

- ¿Cada columna es una variable?
- ¿Cada fila es una observación?

Esto permite saber si los datos están ordenados. Existen casos en los que no necesariamente son así las cosas, requiriendo transformaciones para poder iniciar el análisis exploratorio. En este caso los datos están ordenados y se cuenta con 12 columnas y 891 filas. Posteriormente se puede determinar qué datos hay en cada columna, que deberían ser similares a la tabla 1:

	0
Survived	Número
Pclass	Número
Name	Texto
Sex	Texto
Age	Número
SibSp	Número
Parch	Número
Ticket	Texto
Fare	Número
Cabin	Texto
Embarked	Texto

Para los datos numéricos, se puede realizar análisis univariados no gráficos. Los estudiantes deben calcular, utilizando las funciones de hojas de datos, las siguientes variables: conteo, media, desviación estándar, valor mínimo y valor máximo de las siguientes columnas:

Survived, Pclass, Age, SibSp, Parch, Fare

Discutir con los estudiantes ¿Qué significan esos datos?
Si se observa la media de la columna survived (debería dar 0.383838) es posible entender que solo el 38% de las personas a bordo sobrevivieron. Quiere decir que la mayoría (61.61%) no lo hizo. Estas son algunas conclusiones que se pueden orientar con los estudiantes a partir del análisis univariado, no gráfico.
Para complementar, se puede realizar el análisis univariado gráfico por medio de conteos. Por ejemplo, para la columna survived (sólo contiene ceros o unos) se puede realizar el conteo de ceros y el conteo de unos. Crear una tabla manualmente similar a la de la figura 1 y graficarla con la hoja de cálculo (figura 2) empleando un gráfico de columnas.

No sobrevientes	549
Sobrevivientes	342

Figura 1: conteo de la columna survived

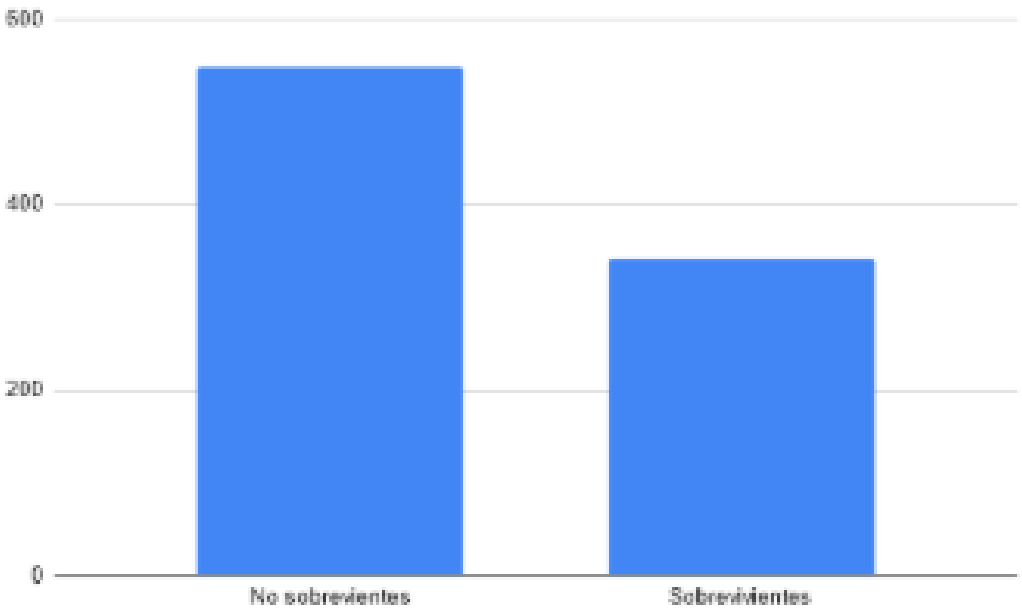
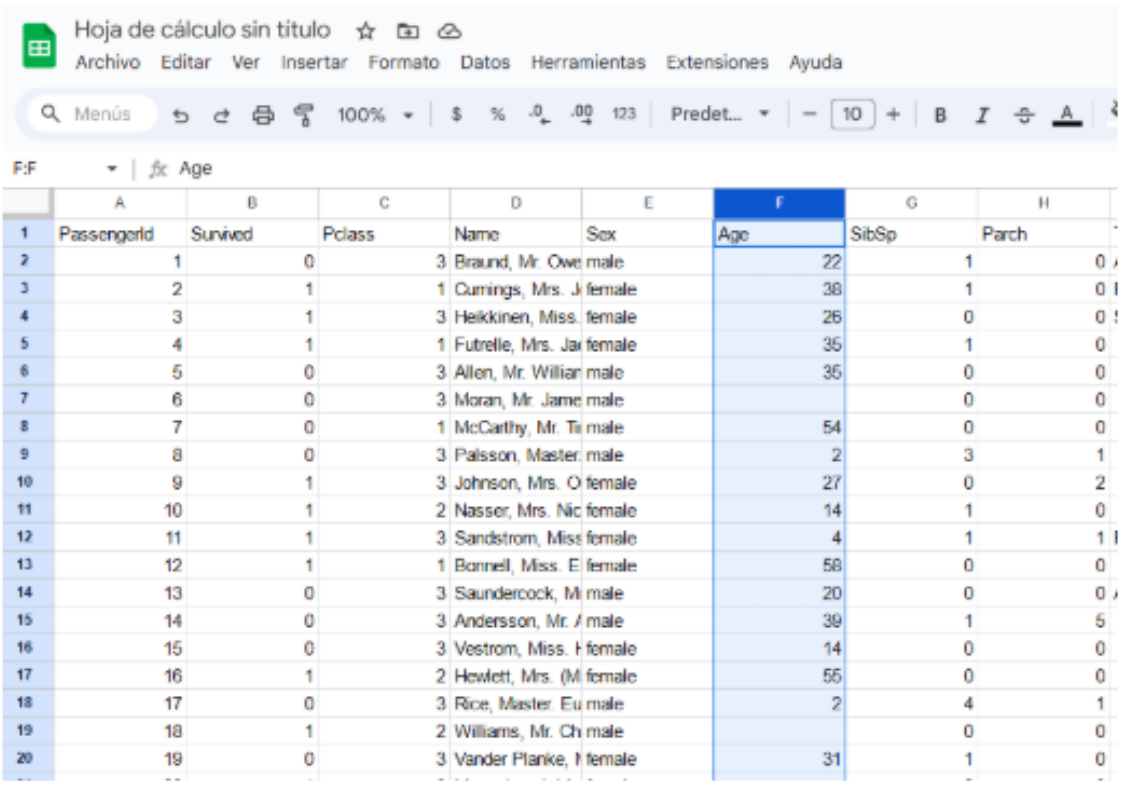


Figura 2: visualización del conteo con un diagrama de barras

Con el análisis gráfico es posible entender mejor la proporción de sobrevivientes en contraste con la de no sobrevivientes. Orientar a los estudiantes a realizar los conteos de las demás variables numéricas. Este análisis de barras también se puede efectuar en algunas columnas en donde la variable es una categoría, por ejemplo, en la columna Sex que determina el género de cada persona. Determinando la proporción de hombres y mujeres a bordo del barco, y entendiendo mejor la historia que narran los datos.

Para variables numéricas (no categóricas), como el caso de la edad o del costo del tiquete (fare) se pueden hacer histogramas. En la figura 3 se aprecia este proceso, seleccionado la columna, y en el menú insertar gráficos



	A	B	C	D	E	F	G	H
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
2	1	0	3	Braund, Mr. Owe	male	22	1	0
3	2	1	1	Cumings, Mrs. J	female	38	1	0
4	3	1	3	Heikkinen, Miss.	female	26	0	0
5	4	1	1	Futrelle, Mrs. Ja	female	35	1	0
6	5	0	3	Allen, Mr. Williar	male	35	0	0
7	6	0	3	Moran, Mr. Jame	male		0	0
8	7	0	1	McCarthy, Mr. Ti	male	54	0	0
9	8	0	3	Palsson, Master.	male	2	3	1
10	9	1	3	Johnson, Mrs. O	female	27	0	2
11	10	1	2	Nasser, Mrs. Nic	female	14	1	0
12	11	1	3	Sandstrom, Miss	female	4	1	1
13	12	1	1	Bonnell, Miss. E	female	58	0	0
14	13	0	3	Saundercock, M	male	20	0	0
15	14	0	3	Andersson, Mr. A	male	39	1	5
16	15	0	3	Vestrom, Miss. H	female	14	0	0
17	16	1	2	Hewlett, Mrs. (M	female	55	0	0
18	17	0	3	Rice, Master. Eu	male	2	4	1
19	18	1	2	Williams, Mr. Ch	male		0	0
20	19	0	3	Vander Planke, M	female	31	1	0

Figura 2: visualización del conteo con un diagrama de barras

Una vez se cree el gráfico, en la configuración (panel derecho) seleccionar gráfico de histograma (como se muestra en la figura 4).

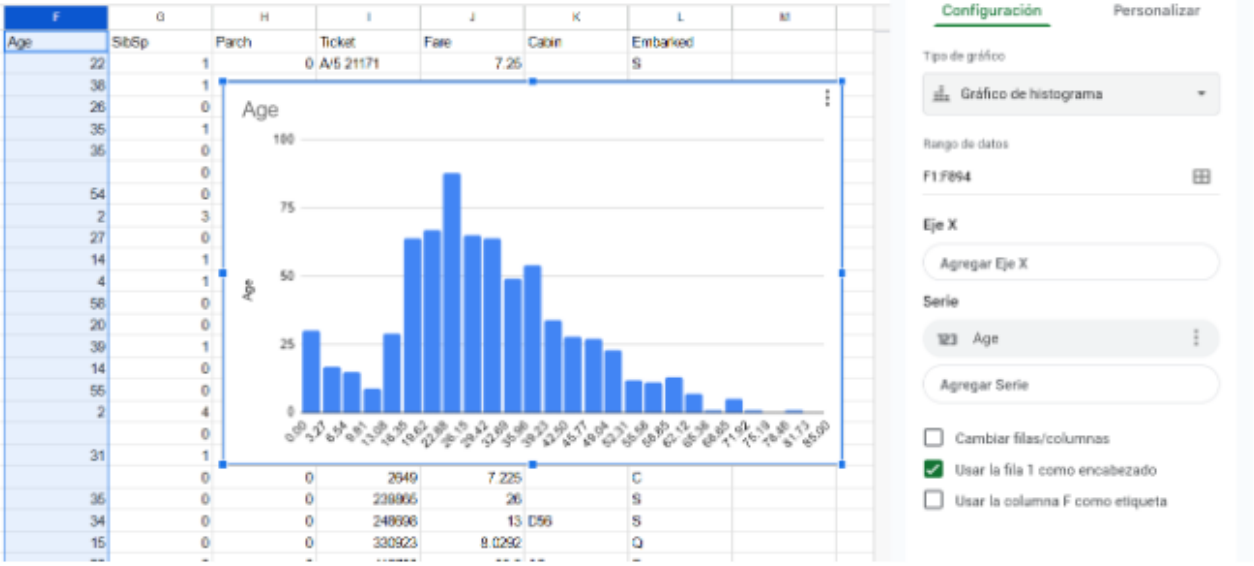


Figura 2: visualización del conteo con un diagrama de barras

Se propone realizar el histograma de la columna fare y Passenger ID, preguntar a los estudiantes si tiene sentido el histograma de Passenger ID y analizar qué sucede si se elimina esta columna. ¿Es algo positivo o negativo?.

Para el caso del análisis multivariado, con los estudiantes se deben analizar las variables de supervivencia y género (dos variables). Para esto se debe filtrar el género y calcular la media de supervivencia, se sugiere para facilidad utilizar la herramienta filtro en la columna sex, y copiar los datos filtrados a una nueva hoja.

Al calcular la media de la columna de supervivencia, solo para hombres se aprecia que solo el 18.82 % sobrevivieron.
En el caso de las mujeres sobrevivieron el 73.73%
Orientar a los estudiantes a analizar otras variables filtradas por la supervivencia (análisis multivariados) para determinar, por ejemplo, la supervivencia según el costo del tiquete, el puerto de embarque y la clase de pasajero.
Finalmente, validar los datos de tres variables, es decir, determinar con los estudiantes la supervivencia entre hombres y mujeres dependiendo de la clase de tiquete. Con esto se tienen los valores promedio de supervivencia dadas diferentes circunstancias (análisis multivariado).
Para el análisis gráfico multivariante, se propone contar columnas y hacer gráficos de barras, como el de las figuras 5 y 6 que muestran la supervivencia entre hombres y mujeres. Emplear otros tipos de gráficos que ayuden a visualizar mejor, como el diagrama de torta (figura 6).

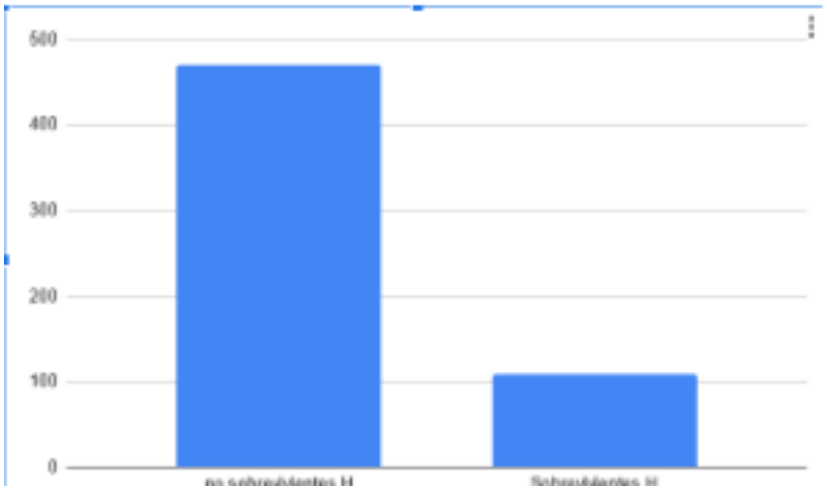


Figura 5: Supervivencia en hombres.

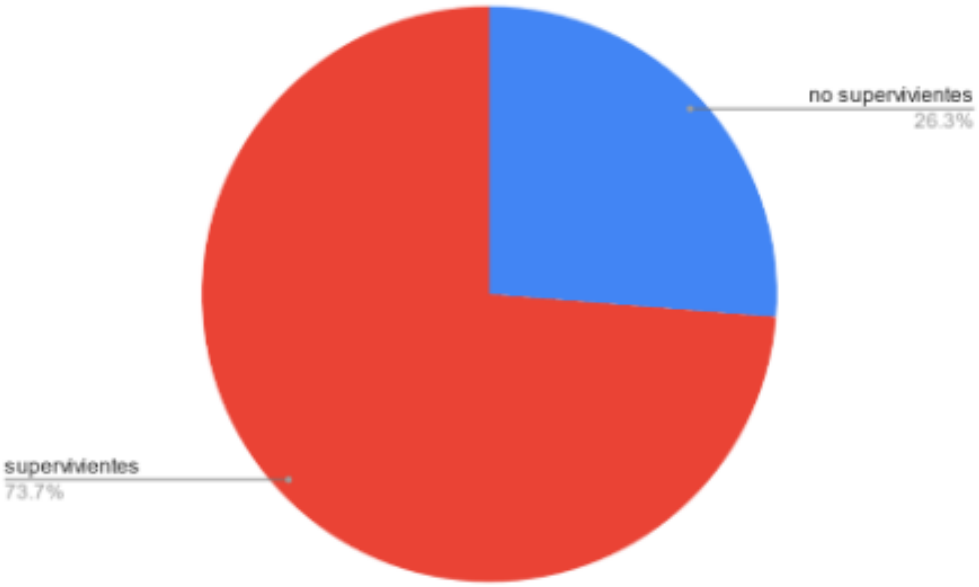


Figura 6: Supervivencia en mujeres

Con estos gráficos se aprecian contrastes y se explora la historia que la información recolectada cuenta. Como ejercicio, proponer a los estudiantes realizar gráficos de supervivencia filtrando por género y por tiquete, o por género y por rangos de edad.

Orientar a los estudiantes a intentar responder las siguientes preguntas con los datos:

1. ¿Cuántos sobrevivientes hay según el tipo de tiquete?
2. ¿Qué tasa de supervivencia tenían las personas según el tipo de tiquete?
3. ¿Cambia la tasa de supervivencia según el tiquete si se tiene en cuenta el género?
4. ¿Cambia la tasa de supervivencia según el tiquete, el género y la cabina?
5. ¿Cómo se puede relacionar la variable sibsp? ¿afecta en los datos?
6. ¿Qué supervivencia hay por edad?
7. ¿Qué supervivencia hay por rango de edad y género?
8. ¿Qué supervivencia hay por edad, género y cabina?
9. ¿Cómo relacionar la columna Parch?
10. ¿Qué relación tiene la edad con el costo del tiquete?
11. ¿Qué relación tiene el lugar de embarque con el costo del tiquete?