PRÁCTICA 2
LIMPIEZA Y ANÁLISIS DE DATOS

PABLO MARTÍN SÁNCHEZ

JULIA CAMARENA PÉREZ

Contenido

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

El dataset consta de los datos de los pasajeros que iban a bordo del Titanic durante su hundimiento. En el dataset consta aquellos pasajeros que sobrevivieron y aquellos que perecieron en el accidente. Con este dataset se podrá responder a la pregunta de si hay algunos grupos de personas que tenían más probabilidades de sobrevivir que otros.

Variable	**Definition**	**Key**
passengerid	Identification number of the passenger	
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
name	Name of the passenger	
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Como vemos en la tabla y cómo podemos observar también en el informe realizado en R, el dataset consta de 5 variables categóricas y 6 numéricas.

2. Integración y selección de los datos de interés a analizar.

Para trabajar con los datos, teniendo en cuenta su relevancia, hemos decidido eliminar las variables passengerid, name, ticket y fare. Hemos considerado que el nombre del pasajero, el identificador que se le ha asignado, el número de su ticket o la tarifa del mismo no nos aportan información relevante a la hora de saber si sobrevivieron o no.

Trabajaremos por ello con el resto de las variables.

3. Limpieza de los datos.

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

Todos los datos están completos, exceptuando la edad de 177 de las observaciones, que están vacías.

Teniendo en cuenta que son muchas observaciones, hemos considerado que no sería óptimo establecer en todas estas la edad media de los pasajeros, ya que podría tratarse de niños o ancianos. Además, sabiendo que tenemos más observaciones de las que podemos extraer respuestas, hemos decidido eliminar estos datos vacíos.

Tras analizar de nuevo los datos, hemos visto que la variable cabin presenta 687 valores vacíos, de un total de 891 observaciones, por lo que se trata de una variable que nos aporta poca información. Es por ello por lo que, aunque a priori no queríamos eliminarla, lo hemos considerado lo más conveniente.

La variable embarked presenta también 2 valores vacíos que hemos decidido eliminar.

3.2 Identificación y tratamiento de valores extremos.

Para identificar los valores extremos, además de ver los máximos y mínimos de las variables mediante la función summary de R, hemos decidido mostrar los boxplot de las mismas, para poder identificarlos de una manera más visual.

Con ello, hemos observado algunos valores extremos, como en el caso de la variable que contiene el número de hermanos que una persona puede tener a bordo del Titanic, siendo el valor extremo 8. Sin embargo, aunque no es muy usual, hemos decidido que es posible que una persona tenga 8 hermanos y por tanto no lo hemos considerado un outlier.

3.3 Otras observaciones

Hemos decidido categorizar la variable pclass, que indica el estatus socioeconómico del pasajero. Inicialmente era una variable cuantitativa (1, 2, 3) y la hemos transformado en categórica (upper, middle, lower).

Del mismo modo hemos categorizado la variable survived, modificando 0 por “no” y 1 por “yes”, que indica si el pasajero sobrevivió.

Hemos modificado el nombre del lugar de embarque (variable embarked). C por Cherbourg, Q por Queenstown y S por Southampton

Hemos creado la variable numérica sex_numeric para poder emplearla más adelante en el estudio de la correlación. Hemos cambiado male por 1 y female por 0.

Del mismo modo, la variable embarked_numeric desde la variable embarked.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Primero hemos decidido discretizar la variable age para trabajar con grupos de edades y ver así cuáles fueron las más afectadas, creando la variable age_segmented.

Así, hemos seleccionado las variables sex, pclass, embarked y age_segmented para analizar cómo han influido en la variable survived o qué relación tienen. Mostraremos su relación de manera visual.

Para estudiar la correlación hemos seleccionado las variables pclass, age, sex_numeric, sibsp, embarked_numeric y parch, en relación con survived.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Asumimos normalidad, ya que el número de muestras es mayor que 30. Además, hemos mostrado las gráficas de Normal Q-Q, que confirman la normalidad en las variables cuantitativas.

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.

Tras aplicar la prueba de correlación se ha observado que la variable que más influye en si los pasajeros sobrevivieron es sex_numeric con 0.54. La siguiente variable más influyente es pclass con 0.36 aunque ninguna de las variables influye en gran cantidad.

En cuanto al contraste de hipótesis, hemos decidido plantear tres preguntas:

La **primera**, ¿sobrevivieron más hombres que mujeres? Tras aplicar el contraste de hipótesis, **aceptamos la hipótesis nula**, ya que el pvalue>0,05 (teniendo en cuenta que hemos escogido un porcentaje de confianza del 95%). Afirmamos, por tanto, que sobrevivieron más mujeres que hombres.

La **segunda**, ¿sobrevivieron más personas menores de 30 años que mayores de 30? Tras aplicar el contraste de hipótesis**, aceptamos de nuevo la hipótesis nula**, afirmando que sobrevivieron más pasajeros mayores de 30 años.

La **tercera**, ¿sobrevivieron más pasajeros de clase alta que de clase media y baja? En este caso, el valor de pvalue es muy inferior a 0,05, por lo que **rechazamos la hipótesis nula**, afirmando que la proporción de los pasajeros que sobrevivieron de clase Medium y Lower no son más que los que sobrevivieron de clase Upper.

Finalmente, se aplica regresión logística viendo cómo influyen las variables age, pclass, age_segmented y sex en la variable survived. Vemos que todas las variables menos segmented_age70-79 son significativas (0.081582>0.05).

Hemos calculado también su odds ratio, que muestra que los pasajeros de clase Medium y Upper sobrevivieron más que los de clase Lower de la variable pclass. También que sobrevivieron más mujeres que hombres. El grupo de edad de personas que más sobrevivieron es el de 0 a 9 años.

5. Representación de los resultados a partir de tablas y gráficas.

Podemos observar los resultados de las tablas y las gráficas en el punto 4 del html Practica2.html accesible a través del siguiente enlace en GitHub:

https://github.com/pmartinsanc/Practica2_Tipologia/tree/main/Codigo%20y%20datasets

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

Tras haber realizado el análisis de los datos hemos llegado a algunas conclusiones. En primer lugar, hemos observado mediante la prueba de correlación y comprobado más tarde mediante la regresión logística que la variable más influyente en si los pasajeros del Titanic sobrevivieron es sex, seguida de la variable pclass.

Además, gracias al contraste de hipótesis y de las preguntas que decidimos plantear, sabemos que proporcionalmente sobrevivieron más mujeres que hombres. También, que la mayoría de los pasajeros que sobrevivieron eran mayores de 30 años y que la proporción de los pasajeros que sobrevivieron de clase Medium y Lower no eran más que los que sobrevivieron de clase Upper.

Mediante la regresión logística hemos comprobado la influencia de las variables a la hora de predecir si los pasajeros sobrevivieron. Las variables age, pclass, age_segmented y sex son significativas, exceptuando del grupo de edad de 70 a 79 años de la variable age_segmented.

Finalmente, mediante su odds ratio hemos visto que los pasajeros de clase Medium y Upper sobrevivieron más que los de clase Lower de la variable pclass. También que sobrevivieron más mujeres que hombres y que el grupo de edad de personas que más sobrevivieron es el de 0 a 9 años.

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.

El código es accesible a través del siguiente enlace:

https://github.com/pmartinsanc/Practica2_Tipologia/tree/main/Codigo%20y%20datasets

Contribuciones	Firma
Investigación previa	Pablo Martín Sánchez, Julia Camarena Pérez
Limpieza de dataset	Pablo Martín Sánchez, Julia Camarena Pérez
Esbozo inicial de los	Pablo Martín Sánchez, Julia Camarena Pérez