08/06/2018

**TIPOLOGIA DE DATOS – PRACTICA 2**

*Nombre de los participantes del grupo: María del Carmen Terrats Reviriego*

1.-Descripcion del dataset. .Por qué es importante y que pregunta/problema pretende responder?

El fichero elegido ha sido el del titanic obtenido de la pagina web: Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El fichero a usar será el fichero train.csv que contiene un subconjunto de registros de los pasajeros del barco Titanic que se hundió.

El fichero contiene estos campos:

PassengerId: numero correlativo que identifica el registro

Survived: 0 No sobrevivió 1 Sobrevivió

Pclass: 1 Primera Clase, 2 Segunda clase, 3 Tercera clase

Name: Nombre del pasajero

Sex: male/female

Age: Años del pasajero

SibSp: Numero de hermanos o conyuge

Parch: Numero de padres/hijos

Ticket: Número de ticket

Fare: Importe

Cabin: camarote donde se alojaba

Embarked: Puerto de embarque: C = Cherbourg, Q = Queenstown, S = Southampton

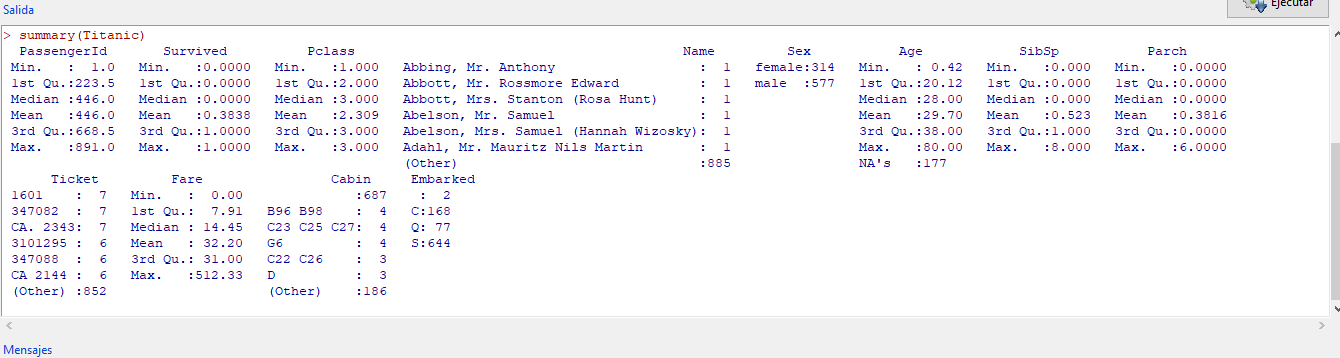
Que indica si cada pasajero sobrevivió al desastre o no lo hizo.

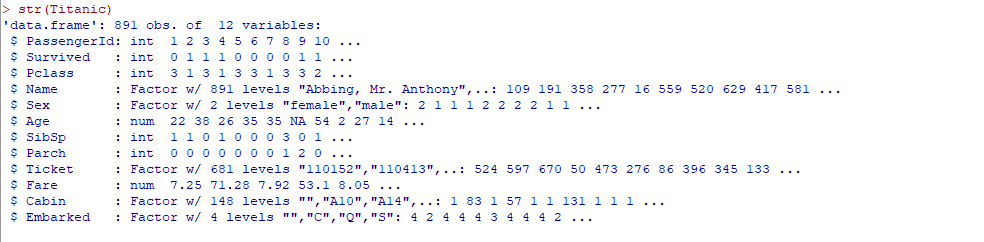
Voy a tratar de resolver la pregunta de cómo eran las personas que sobrevivieron y como eran las personas que no sobrevivieron, si hay características comunes dentro de estos dos grupos.

El fichero contiene 891 registros y las 12 columnas antes indicadas.

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

Primero hago una inspección del fichero:





Donde veo los tipos de cada campo y veo que hay campos como nombre o passenger id que no aportan nada al tener tantas variantes como registros.

El puerto donde embarcaron, pese a ser un campo factorizado tampoco creo que influyese en la supervivencia o no.

Tampoco creo que interese para resolver la pregunta los campos fare, sibsp o parch. Por tanto todos estos campos los elimino.

3. Limpieza de los datos.

3.1. .Los datos contienen ceros o elementos vacios? .Como gestionarias

cada uno de estos casos?

Con los campos que nos han quedado ahora vemos si tienen nulos. En ese caso vemos que el campo Age tiene muchos nulos por tanto no sabemos la edad de esas personas.

> sapply(Titanic, function(x)(sum(is.na(x))))

Survived Pclass Sex Age Cabin

0 0 0 177 0

Y busco si hay 0 pero veo que solo hay 0 en el campo survived, lo que es normal al ser unos de los posibles valores. El resto de campos no tiene ese valor.

> summary(Titanic)

Survived Pclass Sex Age Cabin

Min. :0.0000 Min. :1.000 female:314 Min. : 0.42 :687

1st Qu.:0.0000 1st Qu.:2.000 male :577 1st Qu.:20.12 B96 B98 : 4

Median :0.0000 Median :3.000 Median :28.00 C23 C25 C27: 4

Mean :0.3838 Mean :2.309 Mean :29.70 G6 : 4

3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:38.00 C22 C26 : 3

Max. :1.0000 Max. :3.000 Max. :80.00 D : 3

NA's :177 (Other) :186

En el caso del campo Age al ser un valor numérico decido cambiarlo por su valor de mediana de dicha columna que antes ya he visto al hacer el summary que era 28.

Ahora voy a identificar por el campo Age si el pasajero es un niño(<16 años), adulto (entre 16 y 60), y senior (>60 años).

El campo cabin aunque podría ser interesante para saber en que cubierta estaba el pasajero alojado, está casi siempre vacio y solo lo tienen informado los pasajeros de primera clase así que decido quitarlo.

Factorizo todos los campos.

3.2. Identificación y tratamiento de valores extremos.

Los campos seleccionados para hacer el análisis no tienen valores extemos según hemos visto al hacer sumarry y al hacer la factorización.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar

(Planificación de los análisis a aplicar).

Queremos saber si el sobrevivir o no tuvo que ver con la edad o con la clase en la que viajaban o con el sexo.

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

> with(Titanic, tapply(Survived, Sex, var, na.rm=TRUE))

female male

0.1920291 0.1534879

> var.test(Survived ~ Sex, alternative='two.sided', conf.level=.95, data=Titanic)

F test to compare two variances

data: Survived by Sex

F = 1.2511, num df = 313, denom df = 576, p-value = 0.02218

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

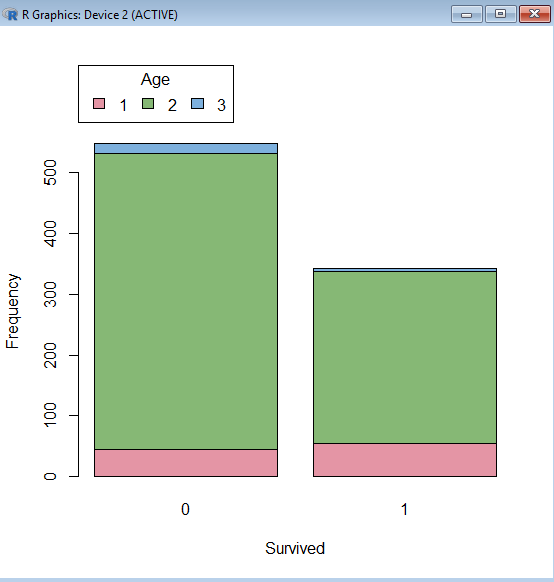
1.032458 1.524708

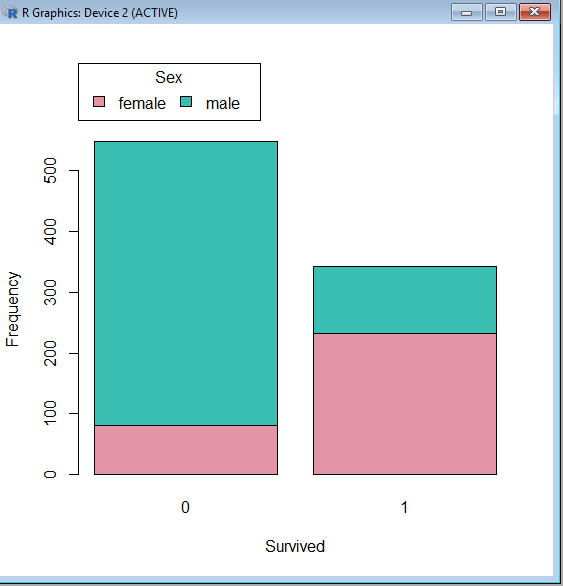
sample estimates:

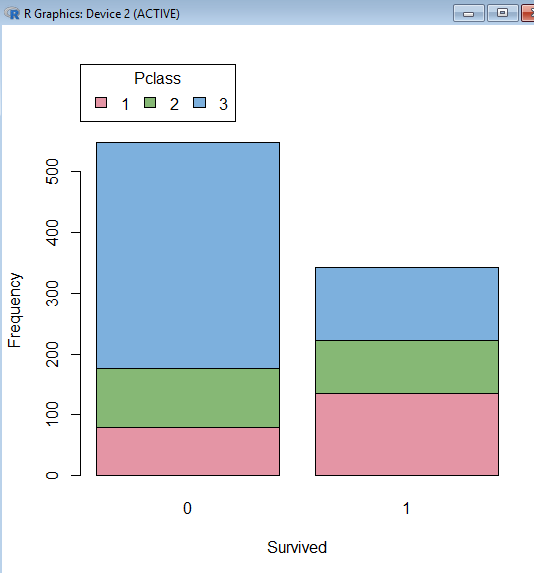
ratio of variances

1.251103

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







6. Resolución del problema. A partir de los resultados obtenidos, ¿cuales son las

conclusiones?. Los resultados permiten responder al problema?

Según las gráficas podemos ver que tuvieron más posibilidades de vivir las mujeres. Luego vemos que la edad no influyó mucho para sobrevivir o no, pero que si influyó el estar en primera clase. Los pasajeros de primera clase tuvieron más probabilidades para sobrevivir.

7. Código: Hay que adjuntar el codigo, preferiblemente en R, con el que se ha

realizado la limpieza, analisis y representacion de los datos.

Adjunto el código en R en el fichero llamado Instrucciones\_Titanic.R