Caso de Uso: Introducción y premisas

En respuesta a la práctica que nos ocupa, he decidido utilizar uno de los datasets que he encontrado en el sitio web especializado en Machine Learning www.kaggle.com.

El dataset en concreto hace referencia a un caso de uso de análisis relacionado con la supervivencia del Titanic, y fue publicado hará unos dos años en el contexto de unan competición de Machine Learning. A continuación copio link al dataset mencionado:

https://www.kaggle.com/c/titanic/data

Los motivos por los que me he decidido por este caso de uso y dataset y no otro, son variados, pero expongo a continuación los principales:

1. Considero que el principal propósito de la práctica es precisamente 'practicar' con escenarios de uso relacioandos con la limpieza de datos y manipulación de variables, y considero el dataset ideal para ello ya que no es ni demasiado complejo ni demasiado sencillo.
2. He preferido 'huir' de datasets asociados al uso del lenguaje escrito para identificar patrones de comportamiento u opinión, similar al que cree en respuesta a la práctica 1, dado que considero dichos datasets algo más 'limitados' a la hora de desarrollar técnicas estadísticas, visualizaciones o práctias de limpieza de datos.
3. El dataset dispone de variables categóricas, que pueden ser útiles en el propósito de realizar visualizaciones, y cuantitativas (discretas y continuas), útiles para realizar análisis estadísticos.

En la misma linea de sencillez, se trata de un dataset eminentemente simple en el cual poder realizar técnicas básicas de machine learning (o estadística descriptiva básica) sin necesidad de pretender realizar modelos más complejos, lo cual en mi opinión, no es objetivo de la presente práctica.

La práctica ha sido realizada únicamente por el alumno que escribe estas líneas: Javier Lombana Domínguez.

Link a Github

El programa R generado así como el dataset y un resumen del trabajo realizado pueden realizarse en mi sitio Github en la ubicación que copio a continuación:

https://github.com/jlombanado/M2.8517-PRACTICA2

Ficheros

En el enlace a Github copiado en la sección anterior, se encuentran los principales ficheros asociados a la práctica, a saber:

- Documento pdf que incluye el 'paso a paso'
- Documento wiki Readme.md donde se incluye un resumen del documento anterior
- Documento archivo R con el código realizado
- Fichero csv utilizado
- Imágenes utilizadas para mostrar visualizaciones y que se incluyen en el wiki de la página

1. El dataset:

Como se puede intuir por mi respuesta en el epílogo, considero la selección del dataset clave en el contexto de la respuesta al problema del que se trate. Como se puede apreciar, he seleccionado el dataset en cuestión a raiz de unas premisas clave que he tenido en consideración a la hora de abordar el problema planteado en la práctica, pero en el caso de que estuviéramos tratando de responder a otras preguntas, o el enfoque fuera diferente, el dataset podría bien haber sido diferente. Por poner sólo algunos ejemplos, de algunos otros datasets que podrían haber sido incluidos en la práctica, enumeraré algunos tipos con los que me he encontrado en la propia web especializada kaggle o en otras fuentes de internet:

- Datasets eminentemente cuantitativos apropiados para realizar regresiones

- Datasets asociados a opiniones de usuarios o clientes en forma de 'lenguaje natural'.
- Datasets apropiados para realizar clasificaciones de mayor o menor complejidad (ya sean binarias, como las del caso de análisis de supervivencia del titanic, o multi-clasificaciones).
- Datasets apropiados para escenarios 'no supervisados' y por tanto, no clasificados a priori.

Los tipos de datasets enumerados arriba son sólo algunos ejemplos de los tipos de conjuntos de datos con los que nos podemos encontrar. Todos ellos pueden ser sujeto de tareas de limpieza de datos, análisis y visualizaciones, aunque las técnicas varían ostensiblemente dependiendo de la tipología de datos de que se trate. En la misma línea, las tareas de limpieza son ostensiblemente diferentes si tratamos con variables cuantitativas o categóricas, pero ese es un punto al que volveré más adelante en el apartado correspondiente de la práctica.

A continuación me centraré principalmente en el análisis del dataset escogido, y haré mención a la metodología utilizada en el análisis del dataset, incluida su limpieza, exploración y visualización.

En lo que respecta al dataset, el propio sitio web kaggle define las siguientes variables y claves como parte del dataset escogido (copio literalmente respetando el idioma inglés original):

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

Como podemos apreciar, el dataset se compone de un total de 10 variables de diferentes tipos, destacando una variable clasificatoria que podremos utilizar para identificar qué pacientes sobrevivieron tras el naufragio del Titanic.

En lo que respuesta al ejercicio desarrollado en el lenguaje de programación estadístico R, he decidido realizar el análisis desglosado en diferentes pasos metodológicos lógicos que se dividen en 4 partes fundamentales, de las cuales copio a continuación las cabeceras. El detalle del contenido de las mismas puede apreciarse en el propio sitio git donde he colgado el ejercicio práctico. De la misma forma, profundizaré más en detalle en cada una de las partes en mis respuesas a las preguntas 2 (limpieza de los datos), 3 (análisis de los datos) y 4 (visualizaciones)

PARTE 1: LECTURA DEL DATASET Y CARGA DE LIBRERIAS

1. Librerias graficas
2. Carga del Dataset

PARTE 2: LIMPIEZA DE LOS DATOS : DATA CLEANING

1. Visualizacion de variables (distribucion, dimensiones y valores)
2. Tratamiento de valores nulos y vacios
3. Visualizacion de variables (distribucion, dimensiones y valores)
4. Cleaning de las variables (descripciones de las columnas)
5. Modificacion de la descripcion de las etiquetas de clases
6. Factorizacion de columnas para facilitar su uso

PARTE 3: ANALISIS ESTADISTICO DE LOS DATOS : DATA EXPLORATION

1. Distribuciones de valor por variable
2. Desviaciones (Standard Deviation)
3. Asimetria (Skewness)
4. Correlaciones
5. Probabilidades

PARTE 4: PREGUNTAS QUE PODEMOS RESPONDER DEL DATASET : DATA INTERROGATION

1. PREGUNTA 1: CUALES SON LAS DIMENSIONES DEL DATASET
2. PREGUNTA 2: QUIEN COMPRO LOS TICKETS MAS CAROS?
3. PREGUNTA 3: ¿CUANTOS HOMBRES Y MUJERES SOBREVIVIERON?
4. PREGUNTA 4: CUAL ES LA DISTRIBUCION DE PASAJEROS POR CLASE?
5. PREGUNTA 5: ¿CUANTAS MUJERES HABIA EN EL PASAJE?
6. PREGUNTA 6: ¿CUAL ES LA CORRELACION ENTRE EDAD Y TARIFAS?
7. PREGUNTA 7: ¿CUAL ES LA DISTRIBUCION DE LAS TARIFAS PAGADAS?

PARTE 5:

1. Conclusiones

2. Limpieza de los datos.

En primer lugar realizamos la carga de las librerias que vamos a utilizar en nuestro analisis y a continuación leemos los datasets. Como en este ejemplo en particular los datasets que he encontrado en kaggle venian ya divididos entre training y test set (80\20 aproximadamente), he decidido recombinar de nuevo ambos conjuntos de datos, ya que nuestro principal objetivo en la practica no es generar un modelo sino explorar y analizar los datos.

A continuacion copio un extracto de los primeros registros. Como se puede observar a simple vista, hay numerosos campos vacíos en la columna 'Cabin'. La variable 'Age' también parece tener algunos datos nulos.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500		S
14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.0000		S
17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250		Q

Los siguientes pasos irán orientados a conocer mejor los datos de que disponemos. Los comandos ejecutados a continuación nos indican las dimensiones del dataset (891 filas y 12 columnas) y los valores diferentes dispuestos en nuestras variables categóricas. EL siguiente pantallazo muestra los resultados obtenidos:

```
> dim(titanic_data)
[1] 891  12
>
> unique(titanic_data$Pclass)
[1] 3 1 2
> unique(titanic_data$Sex)
[1] "male"   "female"
> unique(titanic_data$SibSp)
[1] 1 0 3 4 2 5 8
> unique(titanic_data$Parch)
[1] 0 1 2 5 3 4 6
> unique(titanic_data$Embarked)
[1] "S" "C" "Q" ""
```

Como podemos ver, lo más destacable son las variables Parch y SibSp, que ya aparecen discretizadas en conjuntos de datos. Dichas variables nos proporcionan datos sobre la relación de consanguineidad de los pasajeros, tal y como se nos indica en Kaggle (copio fragmento íntegro):

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

A continuación realizaremos el tratamiento de los valores nulos. El siguiente comando nos informa de cuál es el volumen de los mismos en nuestro dataset:

```
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket        Fare       Cabin
          0           0           0           0           0         177           0           0           0           0           0
   Embarked
          0
```

Como podemos apreciar, el caso más preocupantes es la variable 'Age' con 177 valores nulos, que se tratarán a continuación. Copio fragmento de código donde se puede comprobar cómo se ha realizado la labor de sustitución de los valores nulos, por los de la media agrupada asociada a la dupla 'survived'+'clase'. El resultado de dicho análisis nos permitirá obtener un valor medio que utilizaremos para rellenar el campo edad en los valores vacíos:

```
# 2. Tratamiento de valores nulos y vacios

sapply(titanic_data, function(x) sum(is.na(x)))
sum(is.na(titanic_data$titanic_data))
sum(is.na(titanic_data$Age))

summarise(median_age = median(titanic_data$Ag, na.rm=TRUE))

titanic_data_means <- titanic_data %>% group_by(Survived,Pclass) %>% summarise(median_age = median(Age, na.rm=TRUE))

# calculamos los valores medios por grupo de supervivencia y clase
nonsurvived1class = titanic_data_means %>% filter(Survived == 0 & Pclass == 1 ) %>% select(median_age)
nonsurvived2class = titanic_data_means %>% filter(Survived == 0 & Pclass == 2 ) %>% select(median_age)
nonsurvived3class = titanic_data_means %>% filter(Survived == 0 & Pclass == 3 ) %>% select(median_age)
survived1class = titanic_data_means %>% filter(Survived == 1 & Pclass == 1 ) %>% select(median_age)
survived2class = titanic_data_means %>% filter(Survived == 1 & Pclass == 2 ) %>% select(median_age)
survived3class = titanic_data_means %>% filter(Survived == 1 & Pclass == 3 ) %>% select(median_age)

# y despues asignamos la media de edad de cada grupo a a edad que corresponde del dataset
titanic_data$Age[titanic_data$Survived == 0 & titanic_data$Pclass == 1 & is.na(titanic_data$Age)] = nonsurvived1class[,2]
titanic_data$Age[titanic_data$Survived == 0 & titanic_data$Pclass == 2 & is.na(titanic_data$Age)] = nonsurvived2class[,2]
titanic_data$Age[titanic_data$Survived == 0 & titanic_data$Pclass == 3 & is.na(titanic_data$Age)] = nonsurvived3class[,2]
titanic_data$Age[titanic_data$Survived == 1 & titanic_data$Pclass == 1 & is.na(titanic_data$Age)] = survived1class[,2]
titanic_data$Age[titanic_data$Survived == 1 & titanic_data$Pclass == 2 & is.na(titanic_data$Age)] = nonsurvived2class[,2]
titanic_data$Age[titanic_data$Survived == 1 & titanic_data$Pclass == 3 & is.na(titanic_data$Age)] = nonsurvived3class[,2]

# finalmente convertimos la columna edad de tipo 'lista' a numerica, para facilitar su posterior manejo
titanic_data$edad <- sapply(titanic_data$edad, as.numeric)
```

Los siguientes pasos, 3, 4 y 5, son bastante sencillos y básicamente se encargan de 'limpiar' y preparar los datos para su posterior análisis. El paso 3 modifica las descripciones de las columnas para hacerlas más entendibles. El paso 4, modifica algunos de los contenidos de los atributos con el mismo objetivo. Finalmente, se factorizan algunos de los atributos para facilitar su uso y manejo en los pasos posteriores:

```
# 3. Cleaning de las variables (modificamos las descripciones de las columnas para adaptarlas a español)
colnames(titanic_data) = c('pass_id', 'sobrevive', 'clase', 'nombre',
                           'sexo', 'edad', 'num_hermanos', 'num_padres',
                           'billete', 'tarifa', 'cabina', 'embarcado')

# eliminamos la variable cabina del dataset
titanic_data =  subset(titanic_data , select=-c(cabina))

# 4. Modificacion de la descripcion de las etiquetas de clases
titanic_data$clase[titanic_data$clase == 1] = '1a clase'
titanic_data$clase[titanic_data$clase == 2] = '2a clase'
titanic_data$clase[titanic_data$clase == 3] = '3a clase'

# 4.1. Verificamos cambio
unique(titanic_data$clase)

# 5. Factorizacion de columnas para facilitar su uso
titanic_data$clase = as.factor(titanic_data$clase)
titanic_data$edad = as.factor(titanic_data$edad)
titanic_data$sexo = as.factor(titanic_data$sexo)
titanic_data$sobrevive = as.factor(titanic_data$sobrevive)
titanic_data$tarifa = as.factor(titanic_data$tarifa)
```

El resultado es el dataset del cual copio a continuación las primeras 20 filas a modo de ejemplo:

pass_id	sobrevive	clase	nombre	sexo	edad	num_hermanos	num_padres	billete	tarifa	embarcado
1	0	3a clase	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	S
2	1	1a clase	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C
3	1	3a clase	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	S
4	1	1a clase	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	S
5	0	3a clase	Allen, Mr. William Henry	male	35	0	0	373450	8.05	S
6	0	3a clase	Moran, Mr. James	male	25	0	0	330877	8.4583	Q
7	0	1a clase	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	S
8	0	3a clase	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	S
9	1	3a clase	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	S
10	1	2a clase	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	C
11	1	3a clase	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	S
12	1	1a clase	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	S
13	0	3a clase	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	S
14	0	3a clase	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	S
15	0	3a clase	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	S
16	1	2a clase	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	S
17	0	3a clase	Rice, Master. Eugene	male	2	4	1	382652	29.125	Q
18	1	2a clase	Williams, Mr. Charles Eugene	male	30.5	0	0	244373	13	S
19	0	3a clase	Vander Planke, Mrs. Julius (Emelia Maria Vandemoorte...	female	31	1	0	345763	18	S

Referencias:

http://www.markhneedham.com/blog/2015/02/22/rdplyr-extracting-data-frame-column-value-for-filtering-with-in/
https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e

3. Análisis de los datos, normalizaciones y pruebas estadísticas realizadas.

EL comando 'summary' nos proporciona información adicional del dataset, lo cual puede sernos útil a la hora de analizar las variables cuantitativas, ya que podremos identificar de una sola vez las principales medidas estadísticas. Este es el resultado obtenido previamente a haber realizado la limpieza y factorización del dataset:

```
> summary(titanic_data)
  PassengerId      Survived          Pclass          Name              Sex                 Age            SibSp           Parch       
 Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309        Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000  
 1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000  
 Median : 655   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00   Median :0.0000   Median :0.000  
 Mean   : 655   Mean   :0.3838   Mean   :2.295                                         Mean   :29.88   Mean   :0.4989   Mean   :0.385  
 3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000  
 Max.   :1309   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00   Max.   :8.0000   Max.   :9.000  
                NA's   :418                                                            NA's   :263                                     
    Ticket               Fare            Cabin             Embarked        
 Length:1309        Min.   :  0.000   Length:1309        Length:1309        
 Class :character   1st Qu.:  7.896   Class :character   Class :character   
 Mode  :character   Median : 14.454   Mode  :character   Mode  :character   
                    Mean   : 33.295                                         
                    3rd Qu.: 31.275                                         
                    Max.   :512.329                                         
                    NA's   :1                                               
```

Como ejemplo podemos mencionar el caso de la variable 'Age', cuya medición nos está informando de las edades mínimas, máximas y medias entre los pasajeros. Así sabemos que la persona más joven era un bebé de 0.17 meses y la de mayor edad 80 años. La media de los pasajeros era 29.88 antes de realizar la tarea de limpieza y eliminación de valores nulos al sustituirlos por la media.

Los resultados obtenidos, nos permiten tener una idea más clara de las características del conjunto de datos que manejamos, obeniendo resultados muy significativos y que pueden ayudarnos a entender mejor los datos. En las líneas siguientes mostraré la sección del programa R elaborado donde se ha realizado el análisis estadístico, para después pasar a las conclusiones obtenidas tras su ejecución:

```
################## PARTE 3: ANALISIS ESTADISTICO DE LOS DATOS : DATA EXPLORATION  ####################
library(mlbench)
library(e1071)

# En primer lugar identificamos la tipologia de las clases que componen el dataset
sapply(titanic_data, class)

# 1. Desviaciones y medias de atributos cuantitativos (Standard Deviation): edad y tarifa
sapply(titanic_data[,c(6,10)], mean)
sapply(titanic_data[,c(6,10)], sd)

# 2. Asimetria (Skewness) de las variables cuantitativas
skew <- apply(titanic_data[,c(6,10)] , 2, skewness)
print(skew)

# 3. Correlaciones entre las variables
correlations <- cor(titanic_data[,c(2,6)] )
# mostramos la matriz de correlacion
print(correlations)
correlations <- cor(titanic_data[,c(2,10)] )
# mostramos la matriz de correlacion
print(correlations)

# 4. Distribuciones de valor por variable y probabilidades (porcentuales)
# paquete
y <- titanic_data$clase
cbind(freq=table(y), percentage=prop.table(table(y))*100)
y <- titanic_data$sobrevive
cbind(freq=table(y), percentage=prop.table(table(y))*100)
y <- titanic_data$sexo
cbind(freq=table(y), percentage=prop.table(table(y))*100)
################################ FIN PARTE 3 ##################################################################
```

Conviene aclarar en primer lugar el enfoque del análisis realizado y su razonamiento. Como se puede apreciar, la tipología de atributos contenidos en nuestro dataset se compone de numerosas variables numéricas, pero dos de ellas podrian ser consideradas como las más relevantes desde el punto de vista estadístico al ser continuas: edad y tarifa. Otras variables cuantitativas, como número de hijos o de hermanos, son también numéricas y relevantes, pero su rango es más limitado y por tanto no es de esperar que un análisis estadístico nos devuelva diferencias significativas. Enfocaré mi análisis por tanto y por dichos motivos, en edad y tarifa, aunque un estudio más exhaustivo y profundo podría requerir realizar un análisis semejante con las demás variables. Copio a continuación los pasos realizados y las conclusiones:

1. **Desviaciones y medias de los atributos cuantitativos: edad y tarifa:**
   La desviación estándar junto con la media son útiles para saber si los datos tienen una distribución gaussiana y para poder, por ejemplo, tomar la decisión de los valores 'outlier' a eliminar en el supuesto escenario de que existan elementos excesivamente alejados de la desviación estándar y de la mayor parte de los datos.

Como se puede apreciar, la media en el caso de edad ha bajado ligeramente tras la eliminación de los valores nulos, y se mantiene en el caso de tarifa (que no tenía nulos).

En el caso de la desviación, se puede apreciar que es mucho mayor en el caso de tarifa, lo que a fin de cuentas va a provocar que para dicho atributo las lecturas sean mucho más escoradas y asimétricas, dado que los datos se encuentran también más expandidas como veremos en las siguientes partes del análisis.

```
> sapply(titanic_data[,c(6,10)], mean)
    edad   tarifa
29.35429 32.20421
> sapply(titanic_data[,c(6,10)], sd)
    edad   tarifa
13.30808 49.69343
```

2. **Asimetría de los atributos cuantitativos: edad y tarifa**

Si una distribución tiene un aspecto casi gaussiano pero se desplaza hacia la izquierda o hacia la derecha, es útil conocer la asimetría o 'skewness' de los datos. Tener una idea del mismo es mucho más sencillo utilizando gráficos, como un histograma o gráfico de densidad, como se hará en la siguiente pregunta, pero su análisis cuantitativo sí que ofrece una referencia que se podrá utilizar en posteriores análisis.

En el ejemplo estudiado, se puede apreciar de nuevo que el valor de asimetría de tarifas es mucho mayor que el de edad, lo que induce a pensar que, como comentaba en el punto anterior, dicho atributo está mucho más sesgado y es más asimétrico que el otro:

```
> skew <- apply(titanic_data[,c(6,10)] , 2, skewness)
> print(skew)
     edad    tarifa
0.5044234 4.7712097
```

3. **Correlaciones entre las principales variables**

Es también enormemente útil desde el punto de vista estadístico, analizar las interacciones entre atributos calculando las correlaciones entre ellos.

Las funciones utilizadas en este ejercicio calculan las correlaciones entre dos atributos y construyen una matriz simétrica de correlaciones entre ambas. Como se puede apreciar en dicha tabla, las desviaciones de cero muestran una correlación más o menos positiva o negativa dependiendo de la distancia entre dichos valores. En los ejemplos analizados, se muestra la correlación entre los siguientes pares de atributos y los resultados hablan por sí solos:

```
           sobrevive       edad
sobrevive  1.00000000 -0.06202983
edad      -0.06202983  1.00000000

           sobrevive    tarifa
sobrevive 1.0000000 0.2573065
tarifa    0.2573065 1.0000000

             edad    tarifa
edad   1.0000000 0.1216818
tarifa 0.1216818 1.0000000
```

Como se puede apreciar, existen motivos para interpretar que la relación entre la supervivencia y la tarifa pagada es superior a la de las otras correlaciones, por lo que podría ser interesante analizar esto en más profundidad en la siguiente sección mediante visualizaciones.

4. **Distribuciones y probabilidades**

Un último paso estadístico interesante de cara a conocer los volúmenes y porcentajes de datos que manejamos, es el análisis de distribución y frecuencia de los datos. Como se puede apreciar, los resultados son bastante significativos. Algunas de las principales conclusiones a las que podemos llegar son:

- Los pasajeros de 3a clase eran un 55% del total del pasaje, con lo que más de la mitad correspondía a dicha categoría.
- El 61% de los pasajeros fallecieron
- El 65% eran hombres y el 35% mujeres

Por lo que atendiendo únicamente a los números, podríamos aventurarnos a decir que tenías muchas más proabilidades de no sobrevivir en caso de que fueras hombre y viajaras en tercera clase

```
> y <- titanic_data$clase
> cbind(freq=table(y), percentage=prop.table(table(y))*100)
         freq percentage
1a clase  216   24.24242
2a clase  184   20.65095
3a clase  491   55.10662
> y <- titanic_data$sobrevive
> cbind(freq=table(y), percentage=prop.table(table(y))*100)
  freq percentage
0  549   61.61616
1  342   38.38384
> y <- titanic_data$sexo
> cbind(freq=table(y), percentage=prop.table(table(y))*100)
       freq percentage
female  314    35.2413
male    577    64.7587
```

Referencias:

https://www.datascience.com/blog/learn-data-science-intro-to-data-visualization-in-matplotlib
https://becominghuman.ai/how-to-deal-with-skewed-dataset-in-machine-learning-afd2928011cc

4. Tablas y gráficas.

Dado que la parte que he dedicado a visualización y preguntas es bastante extensa, la he dividido en dos partes, aunque ambas están intrínsecamente relacionadas ya que, frecuentemente para responder preguntas es necesario realizar visualizaciones específicas. En este contexto, subdividiré esta sección en dos partes igualmente, y repasaré a continuación las partes fundamentales del código ejecutado y sus resultados.

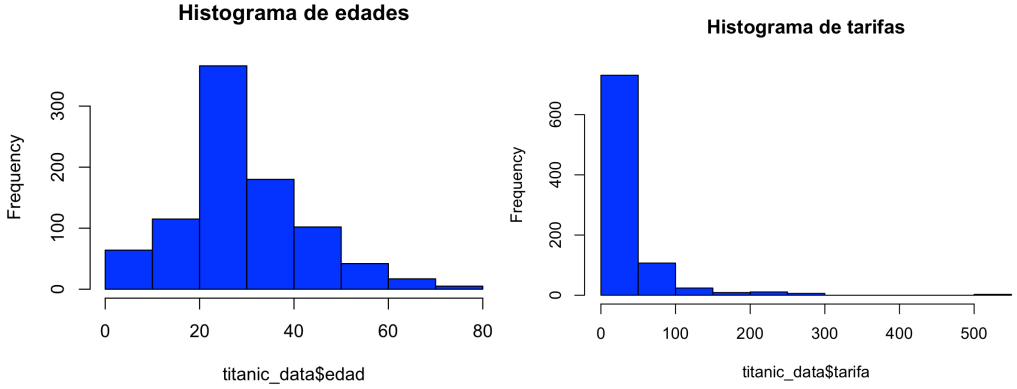
4.1. Visualizaciones

Copio a continuación algunas (no todas) de las visualizaciones que he realizado con el objetivo de entender mejor el conjunto de datos que manejaba. Para consultar otras visualizaciones similares, en este u otros contextos, emplazo a consultar el sitio web Github identificado al comienzo de esta práctica

En el primer paso realizado, he discretizado las variables edad y tarifa para asociarlas a grupos, más fácilmente tratables a nivel visual que un valor numérico. Los resultados obtenidos tras la ejecución del código que muestro a continuación, han sido utilizados en varias de las visualizaciones que incluyo como ejemplo en las siguientes secciones:

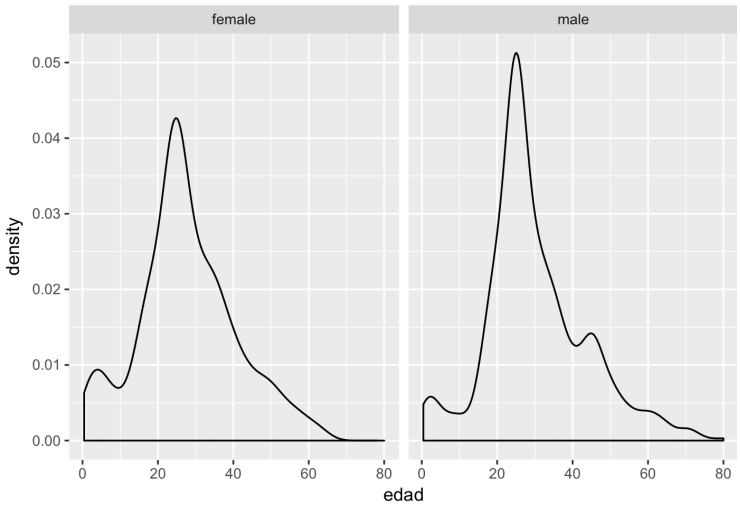
```
summary(titanic_data$edad)
int <- seq(0,80,by=10)
edad.bins <- cut(titanic_data$edad,breaks=int,right=FALSE)
plot(edad,edad.bins)
table(edad,edad.bins)
aggregate(edad, by=list(edad.bins), FUN=mean, na.rm=TRUE)

# Tarifa
summary(titanic_data$tarifa)
int <- seq(10,512,by=25)
tarifa.bins <- cut(titanic_data$tarifa,breaks=int,right=FALSE)
plot(titanic_data$tarifa , tarifa.bins)
table(titanic_data$tarifa , tarifa.bins)
aggregate(titanic_data$tarifa , by=list(tarifa.bins), FUN=mean, na.rm=TRUE)
```

Histogramas: se realizan los siguientes histogramas de edades y tarifas para visualizar la frecuencia de cada uno de los atributos en relación al conjunto de datos:



Conclusiones de los diagramas de densidad : La mayor frecuencia de pasajeros se corresponde a la franja de edad que va de los 20 a los 30 años que, como veremos después, también encaja con la franja de edad con mayor índice de no-supervivencia. En cuanto a la tarifa, la frecuencia es tremendamente superior en el rango que va de 0 a 50. Este factor será visto también más adelante con otra herramienta de visualización que nos permitirá relacionar la tarifa con la supervivencia.

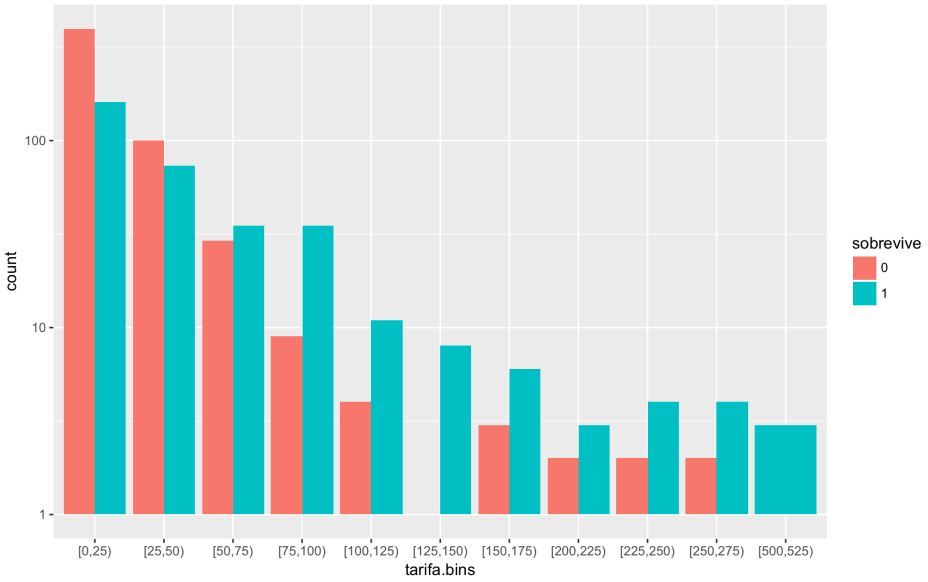
Diagramas de densidad:



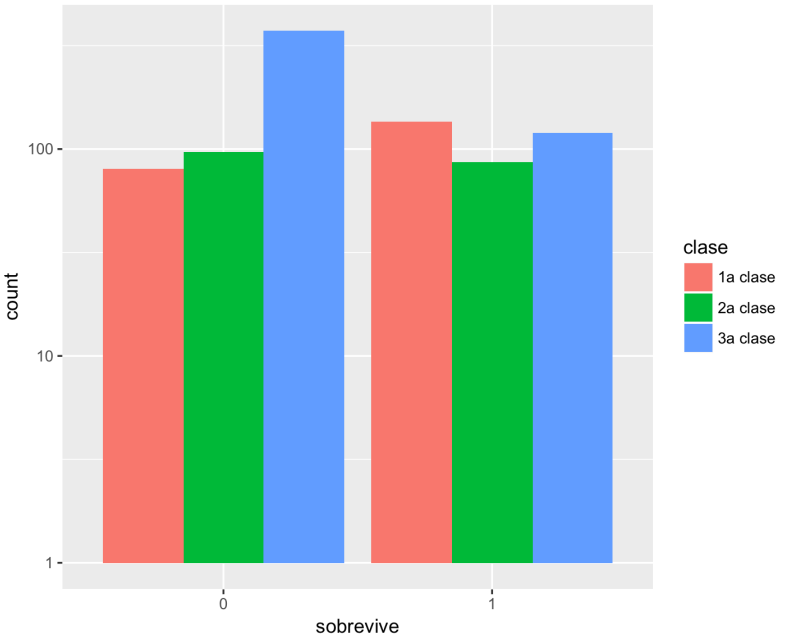
Conclusiones de los diagramas de densidad : los diagramas muestran la densidad del volumen de los datos en relación a la variable edad en relación a la variable sexo, por lo que podemos ver que existe una mayor densidad para los hombres en la franja de edad que va de 20 a 40 años, franja que en el caso de las mujeres es más suave a nivel de densidad.

Diagramas de barras:
Se realizan diferentes tipologías de visualización orientadas a investigar las correlaciones y frecuencias de algunos de los atributos cuando aparecen relacionados en una misma gráfica. Algunas de las variables investigadas son: Supervivencia por tarifas, Supervivencia por clase y Supervivencia por edad:

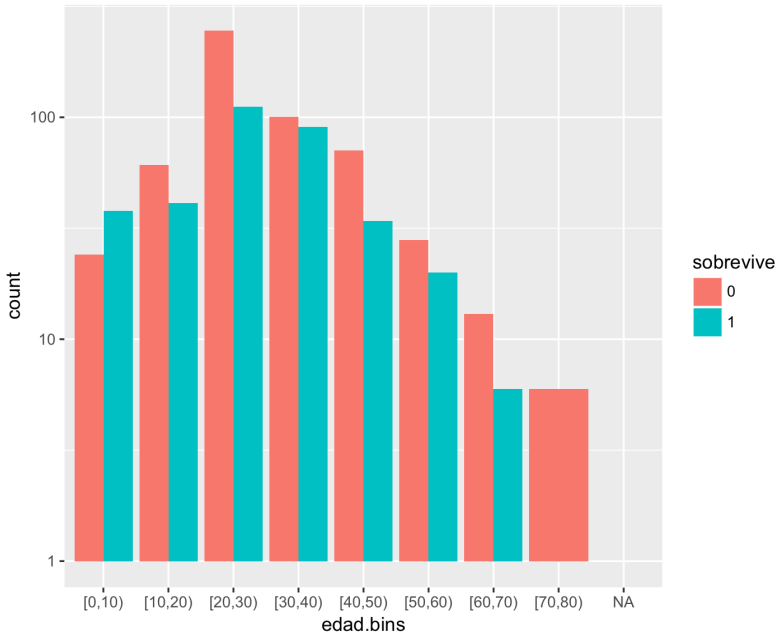
Supervivencia por tarifa



Supervivencia por Clase



Supervivencia por edad



Conclusiones de las visualizaciones de diagramas de barras: esta visualización nuevamente refuerza varios hechos que ya intuíamos por análisis previos, pero también nos permite profundizar en algún factor nuevo:

1. El ratio de supervivencia de los pasajeros de tercera clase es sensiblemente inferior al de las demás clases (esto es, teniendo en cuenta los valores absolutos por clase y no comparando unas clases con otras, menos pasajeros de tercera clase tendían a sobrevivir tras el desastre).
2. Sin tener en cuenta las clases, las edades con menor ratio de supervivencia se encontraban en el rango que va de 20 a 30 años de edad, seguido de cerca por el de 30 a 40.
3. Los ratios de supervivencia disminuyen concluyentemente en las tarifas bajas, especialmente por debajo de 50$, seguramente asociados a tercera clase.

4.2. Preguntas

En mi opinión, es frecuente que en el proceso de exploración y análisis, así como en la fase de visualización, surjan preguntas concretas, y es en ese punto exacto donde nace la sub-fase de interrogación de los datos. Dicha fase se retroalimenta con la anterior (visualización) y frecuentemente enlaza con nuevas y más complejas visualizaciones. Se incluyen a continuación algunas preguntas de ejemplo con las que pretendo dar a entender cómo, en mi opinión, deberíamos orientar el proceso de interrogación y preguntas en un caso práctico como el que nos ocupa:

Pregunta 1: ¿Quién compró los tickets más caros?

Para responder a esta pregunta, se realiza un ranking de los datos por tarifa y ordenado ascendentemente

```
View(titanic_data[order(-titanic_data$tarifa),])
```

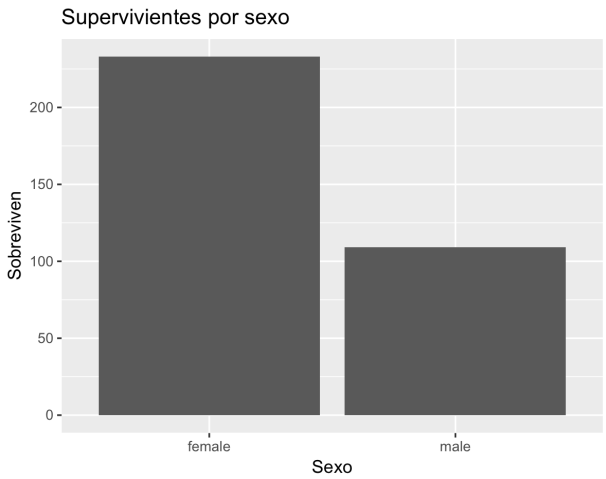
A continuación se puede apreciar un listado ordenado de pasajeros, comenzando por aquellos que pagaron más por su billete. Como se puede apreciar a simple vista, el porcentaje de supervivientes entre las tarifas altas es bastante elevado:

pass_id	sobrevive	clase	nombre	sexo	edad	num_hermanos	num_padres	billete	tarifa	embarcado
259	1	1a clase	Ward, Miss. Anna	female	35.00	0	0	PC 17755	512.3292	C
680	1	1a clase	Cardeza, Mr. Thomas Drake Martinez	male	36.00	0	1	PC 17755	512.3292	C
738	1	1a clase	Lesurer, Mr. Gustave J	male	35.00	0	0	PC 17755	512.3292	C
28	0	1a clase	Fortune, Mr. Charles Alexander	male	19.00	3	2	19950	263.0000	S
89	1	1a clase	Fortune, Miss. Mabel Helen	female	23.00	3	2	19950	263.0000	S
342	1	1a clase	Fortune, Miss. Alice Elizabeth	female	24.00	3	2	19950	263.0000	S
439	0	1a clase	Fortune, Mr. Mark	male	64.00	1	4	19950	263.0000	S
312	1	1a clase	Ryerson, Miss. Emily Borie	female	18.00	2	2	PC 17608	262.3750	C
743	1	1a clase	Ryerson, Miss. Susan Parker "Suzette"	female	21.00	2	2	PC 17608	262.3750	C
119	0	1a clase	Baxter, Mr. Quigg Edmond	male	24.00	0	1	PC 17558	247.5208	C

Pregunta 2: ¿Cuántos hombres y mujeres sobrevivieron?.

El siguiente comando de R nos permite realizar un diagrama de barras desglosado por sexo (x) y supervivencia (y). Como vemos el ratio de supervivientes es muy superior en el caso de mujeres:

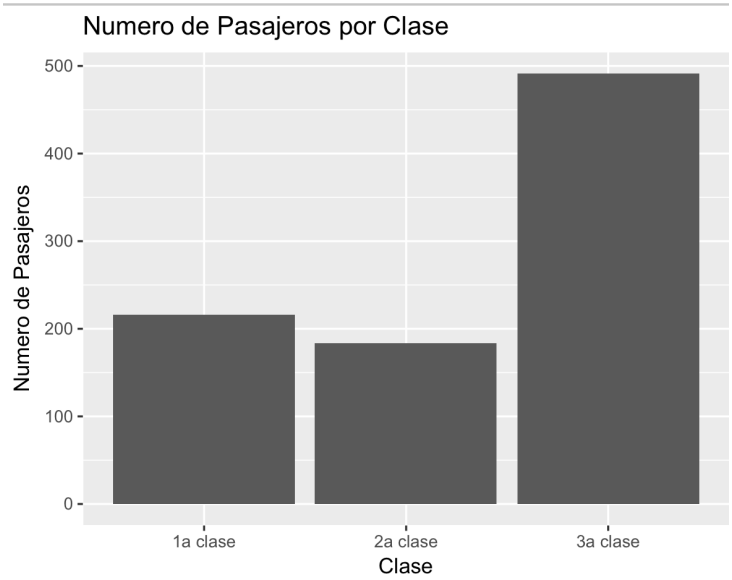
```
# 2.1. Realizamos un summary de los datos
survivors_sex = aggregate(titanic_data$sobrevive, by = list(sexo = titanic_data$sex),FUN = sum)
# 2.2. Y a continuacion un plot del barchart
qplot(sexo, data = survivors_sex, geom='bar', weight = x) +
  ggtitle("Supervivientes por sexo") +
  xlab("Sexo") +
  ylab("Sobreviven")
```



Pregunta 3:¿Cuál es la distribución de los pasajeros por clase?

Volviendo a la pregunta que ya respondimos en la parte anterior (distribución de pasajeros), pero que ahora podemos visualizar utilizando el siguiente código, que nos permitirá construir un diagrama de barras para visualizar la relación de las diferentes categorías con respecto al total de pasajeros:

```
# 3.1. Realizamos un summary de los datos
number_passengers_class = aggregate(titanic_data$clase, by = list(passenger_class = titanic_data$clase),
                                    FUN = length)
# 3.2. Y a continuacion un plot del barchart
qplot(passenger_class, data = number_passengers_class, geom='bar', weight = x) +
  ggtitle("Numero de Pasajeros por Clase") +
  xlab("Clase") +
  ylab("Numero de Pasajeros")
```



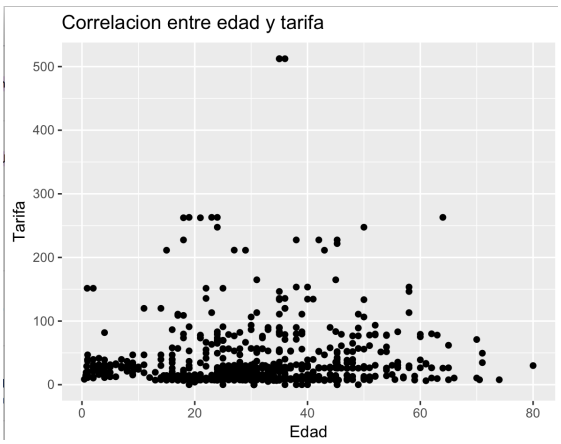
Pregunta 4:¿Cuántas personas tenían el título Mrs en el pasaje?
El siguiente comando nos permite distinguir un total de 129 (entre más de 800 pasajeros) personas con título Mrs en el pasaje, lo que equivale a menos del 10% del total:

```
> nrow(
+   titanic_data[grepl('Mrs', titanic_data$nombre),]
+ )
[1] 129
>
```

Pregunta 5: ¿Cuál es la correlación entre tarifas y edad?
El siguiente código y el gráfico siguiente, nos permiten visualizar de una sola vez la correlación entre las variables tarifas y edad.

```
# 5.1. Correlacion entre tarifas y edad
cor(titanic_data$edad, titanic_data$tarifa, use = 'complete.obs')
# 5.2. Gplot para mostrar graficamente la Correlacion entre tarifas y edad
qplot(edad, tarifa, data = titanic_data) +
  ggtitle("Correlacion entre edad y tarifa") +
  xlab("Edad") +
  ylab("Tarifa")
```

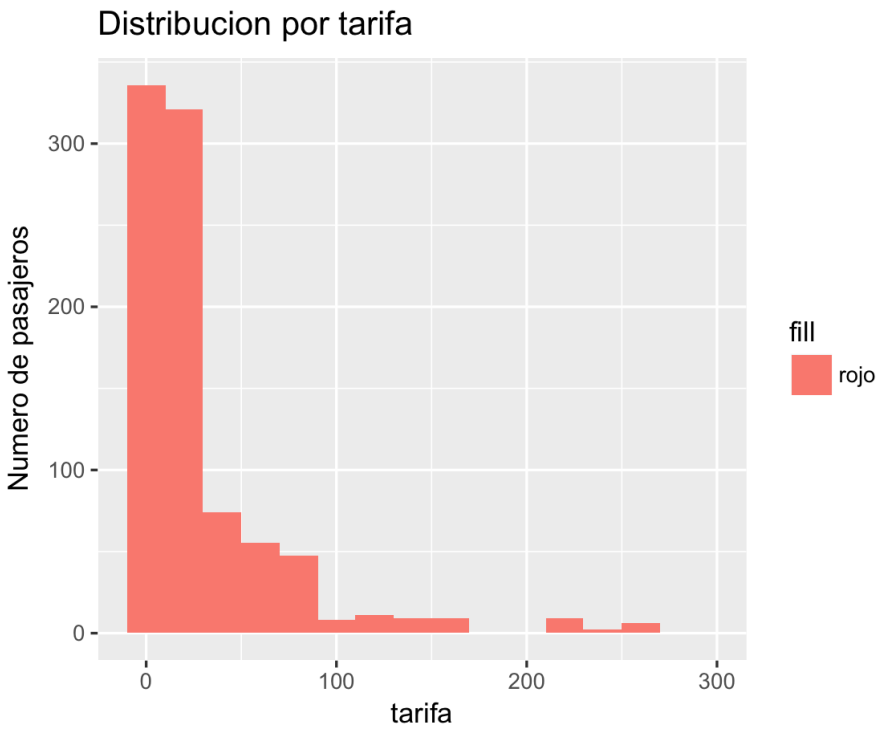
Como se puede apreciar, se trata de un diagrama de dispersión o scatterplot donde se puede apreciar la distribución entre las dos variables. A priori no parece que tengan una relación excesivamente fuerte ni linear. Existen algunos outliers (pocos) en la parte superior, pero la mayor parte de los datos se distribuyen de forma bastante balanceada en la parte baja de la gráfica. Podemos concluir por tanto que no hay indicios para identificar una correlación específica entre la edad del pasajero y las tarifas pagadas:



Pregunta 6: ¿Cuál es la distribución de tarifa por número de pasajeros?
El código siguiente de R nos permite responder a dicha pregunta:

```
# 6.1. Histograma de Tarifas
qplot(tarifa, data = titanic_data, geom = 'histogram',
      binwidth = 20,
      xlim = c(-10,300),
      fill = 'rojo') +
  ggtitle("Distribucion por tarifa") +
  xlab("tarifa") +
  ylab("Numero de pasajeros")
```

Como se puede apreciar, es bastante llamativo que casi toda la distribución se acumula en el primer tramo de la gráfica, correspondiendo a las tarifas más bajos para el mayor número de pasajeros asociado, lo que hace pensar que el Titanic iba principalmente ocupado por pasajeros que pagaron tarifas bajas (tercera clasae probablemente). Por contra el número de pasajeros disminuye enormemente a medida que las tarifas van aumentando hacia la derecha:



5. Conclusiones

Reconozco que el caso de uso seleccionado para responder a la práctica no es excesivamente complejo desde el punto de vista de la investigación pura. De todos es conocido el caso 'Titanic' y los resultados de su investigación, por lo que es también sencillo llegar a conjeturar algunas de las aseveraciones incluidas en los análisis que forman parte de este estudio, sin necesidad de realizar complejas visualizaciones o análisis estadísticos. Considero sin embargo que el principal objetivo de esta práctica era mostrar el concimiento adquirido en ámbitos de limpieza de datos, análisis estadístico y visualizaciones, y en ese contexto considero que el dataset escogido es tremendamente práctico y útil, y mucho más rico que otros datasets que he encontrado asociados inicialmente a investigaciones más complejas, pero que presumo que me habrían dado menos 'juego' en relación a los objetivos de esta práctica.

En cualquier caso y a modo de resumen del trabajo expuesto en las secciones previas, resumo a continuación las conclusiones a las que he llegado tras el análisis efectuado:

1. La mayor frecuencia de pasajeros se corresponde a la franja de edad que va de los 30 a los 40 años, y dentro de ella, existe mayor densidad para el caso de hombres que para mujeres (que tienden a ser más jóvenes de media en el pasaje).
2. El mayor índice de no-supervivencia se corresponde a la franja de edad que va de los 30 a los 40 años.
3. La tarifa más frecuente con enorme diferencia se corresponde con el rango 0-50, que a su vez se corresponde con los pasajeros de tercera clase.
4. El ratio de supervivencia de los pasajeros de tercera clase es sensiblemente inferior al de las demás clases (esto es, teniendo en cuenta los valores absolutos por clase y no comparando unas clases con otras, menos pasajeros de tercera clase tendían a sobrevivir tras el desastre).
5. Sin tener en cuenta las clases, las edades con menor ratio de supervivencia se encontraban en el rango que va de 20 a 30 años de edad, seguido de cerca por el de 30 a 40.
6. Los ratios de supervivencia disminuyen concluyentemente en las tarifas bajas, especialmente por debajo de 50$, seguramente asociados a tercera clase.
7. En los gráficos de densidad, la distribución se acumula en el primer tramo de la gráfica, correspondiendo a las tarifas más bajos para el mayor número de pasajeros asociado, lo que hace pensar que el Titanic iba principalmente ocupado por pasajeros que pagaron tarifas bajas (tercera clase probablemente). Por contra el número de pasajeros disminuye enormemente a medida que las tarifas van aumentando hacia la derecha

En conclusión a los expuesto arriba, podeos decir que había tres variables determinantes en el hecho de que un pasajero sobreviviera o no:

- Sexo
- Edad
- Tarifa

En el caso de la tarifa podríamos incluso reemplazar el atributo por ‘clase’ y obtendríamos el mismo resultado, ya que es precisamente en primera clase donde las tarifas son más caras y en tercera donde son más baratas, y es en ambos casos donde la correlación entre supervivencia y ambos atributos es más elevada.