Práctica 2: Limpieza y análisis de datos

Hugo Martín Méndez

Marc Subira Zurita

Tipología y ciclo de vida de los datos

09 de Junio de 2020

1 Descripción del dataset

El conjunto de datos pertenece a los datos de los pasajeros del titanic. Los datos contienen datos de 887 pasajeros. Cada fila representa una persona. Las columnas describen diferentes atributos acerca de la persona, si sobrevivió o no, su edad, la clase en la que viajaba, el genero y el coste del billete.

Los datos han sido dividido en dos:

- training set (train.csv)
- test set (test.csv)

El training set se utiliza en el caso de querer construir un modelo de machine learning. El test set, se utiliza para comprobar que bien trabaja el modelo que se ha realizado. Las columnas para el set test:

- **PassengerId.** Id del pasajero
- **Pclass.** Clase del billete. 1, primera clase, 2, segunda clase, 3, tercera clase.
- **Sex.** Genero del pasajero, hombre o mujer.
- **Age.** Edad.
- **SibSp.** Numero de gemelos.
- **Parch.** Numero de padres con hijos a bordo.
- **Fare.**
- **Embarked.** Puerto de embarque. C=Cherbourg, Q=Queenstown, S=Southampton
- **Title.** Titulo de la persona.

Columnas del set train:

- **Survived.** 1 si sobrevivió, 0 si murió.
- **Pclass.** Clase del billete. 1, primera clase, 2, segunda clase, 3, tercera clase.
- **Sex.** Genero del pasajero, hombre o mujer.
- **Age.** Edad.
- **SibSp.** Numero de gemelos.
- **Parch.** Numero de padres con hijos a bordo.
- **Fare.**
- **Embarked.** Puerto de embarque. C=Cherbourg, Q=Queenstown, S=Southampton
- **Title.** Titulo de la persona.

Accedemos a los datos imprimendo unicamente los 5 primeros registros, para ver como se organizan los datos.

```
1 test.head(5)
```

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title
0	892	3	0	34	0	0	7.8292	Q	1
1	893	3	1	47	1	0	7.0000	S	3
2	894	2	0	62	0	0	9.6875	Q	1
3	895	3	0	27	0	0	8.6625	S	1
4	896	3	1	22	1	1	12.2875	S	3

```
1 train.head(5)
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title
0	0	3	0	22	1	0	7.2500	S	1
1	1	1	1	38	1	0	71.2833	C	3
2	1	3	1	26	0	0	7.9250	S	2
3	1	1	1	35	1	0	53.1000	S	3
4	0	3	0	35	0	0	8.0500	S	1

Dado que vamos a necesitar la columna de supervivientes, trabajaremos para la parte de análisis con el set train.

Para tener una visión general del set usamos describe.

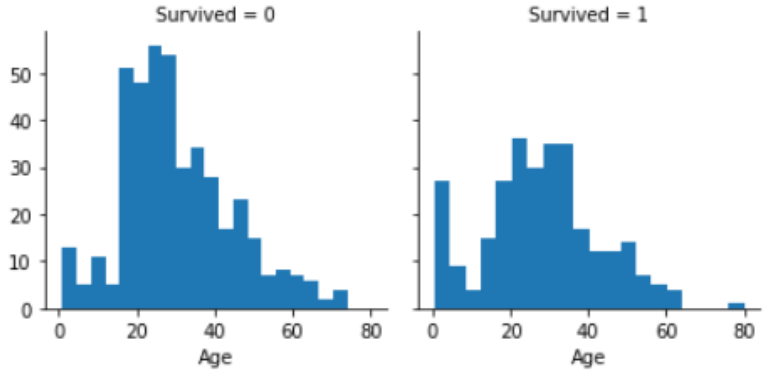
```
1 train.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

2 Integracion y selección de los datos de interés a analizar

Siempre hemos oído la frase “primero mujeres y niños” cuando tiene lugar algún accidente o alguna catástrofe. Por eso vamos a tener en consideración las columnas genero y edad, para ver si ha habido alguna diferencia considerable en cuanto a los supervivientes dependiendo de su genero o edad.

A continuación mostraremos la distribución de los pasajeros, por rango de edad, tanto los que han sobrevivido como los que no.



Con este resultado debemos considerar la columna edad para llegar a algun tipo de conclusion. Además, dado la importancia de esta columna, tenemos que buscar por datos perdidos y completar, en el caso que falten.

Ademas, podríamos pensar que la gente que ha pagado mas por su billete, podría haber tenido algún tipo de privilegio a la hora de ser rescatado o si tuvieron acceso a los botes salvavidas por encima del resto. Si esto tuvo lugar de esta forma, habrá alguna forma de verlo tras analizar los datos. Por lo tanto, para esta segunda suposición, tendremos en cuenta, la clase y cuanto ha pagado el pasajero por su billete.

3 Limpieza de los datos

3.1 Los datos contienen ceros o elementos vacíos.

Analizando en nuestro dataframe encontramos que existen tres columnas con valores perdidos.

```
1 val_perdidos(train)
```

```
El dataset tiene 9 columnas.
Existen 2 columnas con valores perdidos.
```

	Valores Perdidos	% del Total
Age	177	19.9
Embarked	2	0.2

Los datos correspondiente a cabina representan un 78.2% del total de los datos, sin embargo, no todos los pasajeros disponían de un camarote, por lo tanto, tiene sentido esa gran cantidad de datos perdidos.

Por otro lado, vemos que la columna edad (*Age*) posee 86 valores perdidos, que representan una perdida del 20.6% del total de los datos. Debido a la importancia de esta columna para nuestro análisis, procederemos a continuación a completar esta columna.

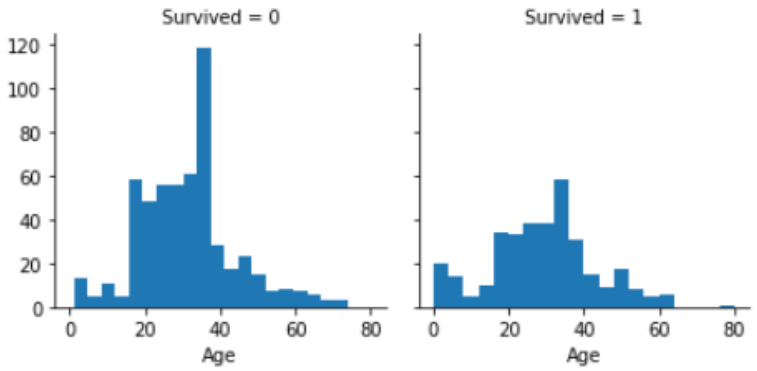
Podríamos considerar tres métodos para completar una característica numérica continua como lo es la edad.

Una de las posibles formas de completar los datos que faltan sería generar números aleatorios entre la media y la desviación estándar de los datos disponibles.

Una forma más precisa de terminar de completar los valores perdidos es utilizar otras funciones correlacionadas. En nuestro caso, para la columna genero, se ha observado una diferencia estadística significativa entre las dos poblaciones. Además, en nuestro estudio estadístico hemos que el gráfico muestra que la mayor correlación la encontramos entre las variables Pclass y fare, como es intutivo pensar que a mayor precio del bilete, mejor la clase. En nuestro estudio estadístico, también hemos observado que en las correlaciones con la clase Survived, vemos que los factores mas correlacionados son el genero y el título de la persona, que entre otros cosas refleja datos del género y la edad.

Entonces para completar los datos de la edad, usaremos números aleatorios entre la media y la desviación estándar, basados en la clase y el genero.

Una vez completada la columna edad, volvemos a realizar el histograma.



Vemos en el grafico anterior, resultado de haber completado la columna edad, como se mantiene la distribución de los datos para los diferentes rangos de edad.

4 Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar

Las principales hipótesis que queremos validar son:

- Gender. Las mujeres tiene mejor ratio de supervivencia que los hombres
- Age. Los niños tiene preferencia sobre los adultos.
- Passenger Class. Los pasajeros de 1a clase son los primeros en abandonar el barco en botes.
- Fare. Los pasajeros con coste del billete mayor, estan por delante de los pasajeros de su misma clase.
- Solos o en groupo -- Los grupos icentivan al resto a buscarles, mientras que los pasajeros que viajan solos tiene mas numeros a ser olvidados.

Factores irrelevantes:

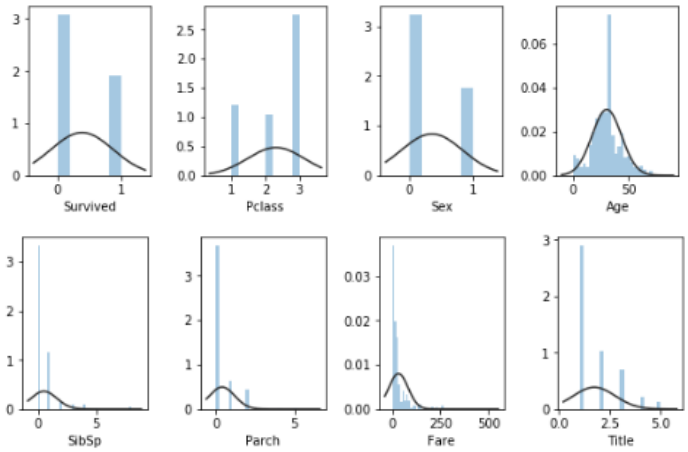
- Cabin #. No tiene relevancia para este análisis.
- Ticket #. No tiene relevancia para este análisis.
- Name. No tiene relevancia para este análisis.

Posibles variables determiantes:

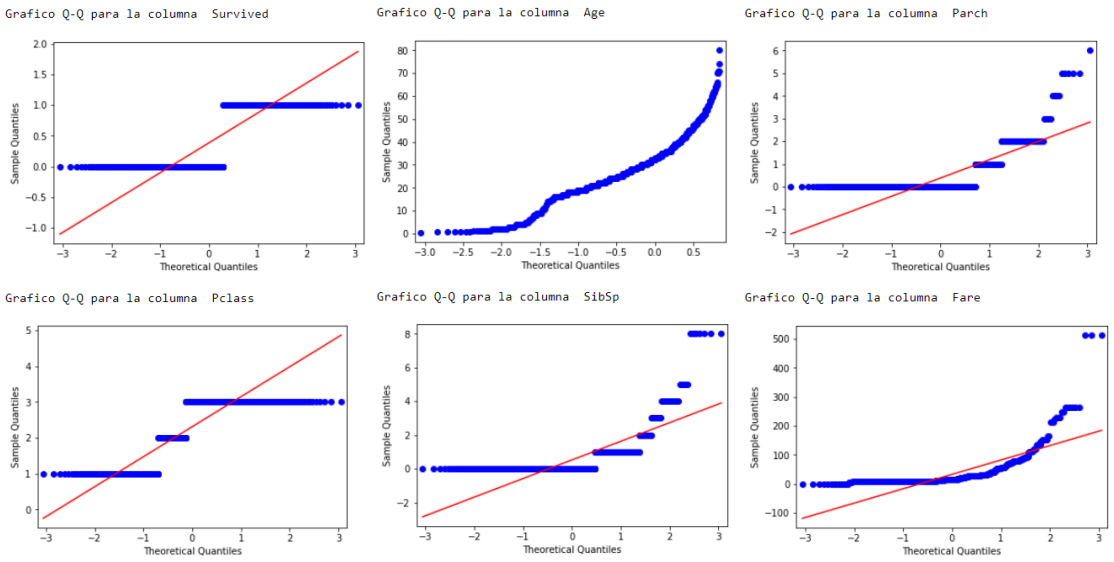
- Título -- Miss, Master, Mr, Mrs, otros.
- Puerto de embarque.
- Niños vs adultos.

4.2 Comprobacion de la normalidad y homogeneidad de la varianza.

Generamos gráficos con las distribuciones de las variables.



Los gráficos Q-Q ("Q" viene de cuantil) es un método gráfico para comparar la distribución de nuestra muestra contra una distribución normal teórica.



Aplicamos ahora el test Shapiro-Wilk y validamos los resultados con un alpha de 0.05.

- p <= alpha: rechazamos H0, no podemos asumir normalidad.
- p > alpha: No podemos rechazar H0, normalidad.

```python
for i, col in enumerate(train.select_dtypes(include=np.number).columns):
    stat, p = shapiro(train[col])
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    # interpretar
    alpha = 0.05
    if p > alpha:
        print(col, 'sigue una distribución normal (aceptamos la H0)\n')
    else:
        print('No podemos asumir normalidad para ',col, '(rechazamos la H0)\n')
```

```
Statistics=0.617, p=0.000
No podemos asumir normalidad para  Survived (rechazamos la H0)

Statistics=0.718, p=0.000
No podemos asumir normalidad para  Pclass (rechazamos la H0)

Statistics=nan, p=1.000
Age sigue una distribución normal (aceptamos la H0)

Statistics=0.513, p=0.000
No podemos asumir normalidad para  SibSp (rechazamos la H0)

Statistics=0.533, p=0.000
No podemos asumir normalidad para  Parch (rechazamos la H0)

Statistics=0.522, p=0.000
No podemos asumir normalidad para  Fare (rechazamos la H0)
```

Conclusión: Despues de analizar la normalidad de las distintas variables, solo podemos assumir normalidad para la variable Age.

Analizamos ahora la homogeneidad de la varianza para para los supervivientes vs los fallecidos. Para ello, utilizaremos el Test de Levene. Podriamos utilizar el F-test para la variable, ya que hemos asumido normalidad en su distribución.

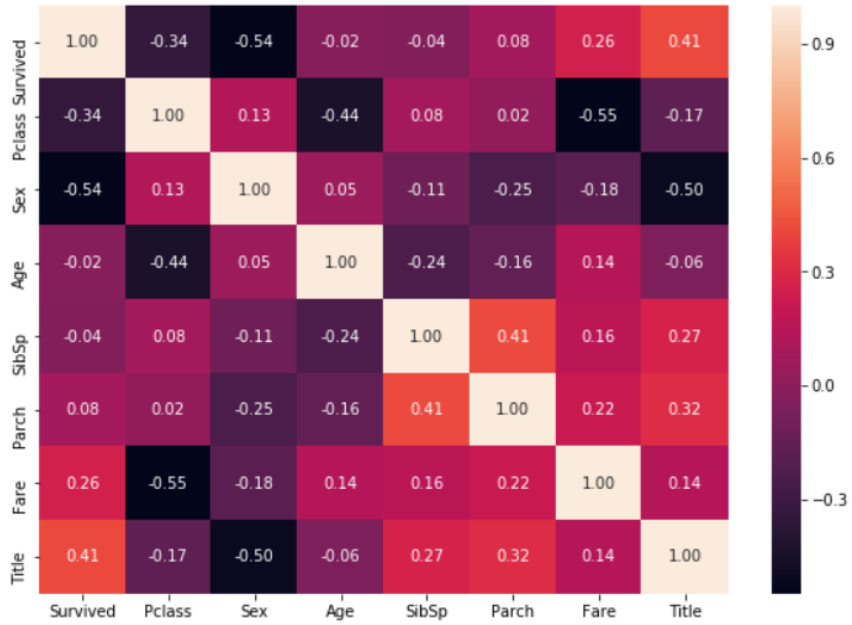
```python
from scipy.stats import levene

#separamos la muestra ne dos poblaciones segun la variable survided:
train.fillna(30, inplace = True)

train_1= train.loc[train['Survived'] == 1]
train_0= train.loc[train['Survived'] == 0]


for i, col in enumerate(train.select_dtypes(include=np.number).columns):
    stat, p = levene(train_1[col],train_0[col])
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    # interpretar
    alpha = 0.05
    if p > alpha:
        print(col, ' asume igualdad de varianzas (aceptamos la H0)\n')
    else:
        print('diferencia entre las varianzas muestrales para la variable ',col, '(rechazamos la H0)\n')
```

```
Statistics=nan, p=nan
diferencia entre las varianzas muestrales para la variable  Survived (rechazamos la H0)

Statistics=39.897, p=0.000
diferencia entre las varianzas muestrales para la variable  Pclass (rechazamos la H0)

Statistics=5.511, p=0.019
diferencia entre las varianzas muestrales para la variable  Age (rechazamos la H0)

Statistics=1.111, p=0.292
SibSp  asume igualdad de varianzas (aceptamos la H0)

Statistics=5.963, p=0.015
diferencia entre las varianzas muestrales para la variable  Parch (rechazamos la H0)

Statistics=45.100, p=0.000
diferencia entre las varianzas muestrales para la variable  Fare (rechazamos la H0)
```

Observamos que solo para la variable SibSp no podemos asumir igualdad de varianzas.

4.3 Aplicacion de pruebas estadisticas para comparar los grupos de datos.

Mapa de correlaciones



Vemos que el gráfico muestra que la mayor correlación la encontramos entre las variables Pclass y fare, como es intuitivo pensar que a mayor precio del billete, mejor la clase. Si nos fijamos en las correlaciones con la clase Survived, vemos que los fatores mas correlacionados son el genero y el título de la persona, que entre otros cosas refleja datos del género y la edad.

Para las variables categoricas, utilizamos la funciona get_dumies para crear columnas para cada unos de los valores categoricos.

Estadistico de contraste

Queremos analizar ahora el ratio de supervivencia por género, i determinar a partir de un estadístico de contraste si la diferencia es significativa (ya esperamos como hemos indicado en nuestras hipótesis, que le ratio de supervivencia de las mujeres es superior al de los hombres).

```
Ratio de supervivencia de hombres: 0.7420382165605095
Ratio de supervivencia de mujeres: 0.18890814558058924
T-value is -18.67183317725917
P-value is [[       nan 0.99983235 0.98296867 ...        nan        nan
0.98296867]
 [0.98296867 0.98296867        nan ... 0.98296867        nan        nan]
 [0.98296867 0.99983235        nan ...        nan        nan 0.98296867]
 ...
 [       nan 0.99983235        nan ...        nan        nan 0.98296867]
 [0.98296867 0.98296867 0.98296867 ... 0.98296867        nan        nan]
 [       nan 0.99983235 0.98296867 ...        nan 0.98296867        nan]]
```

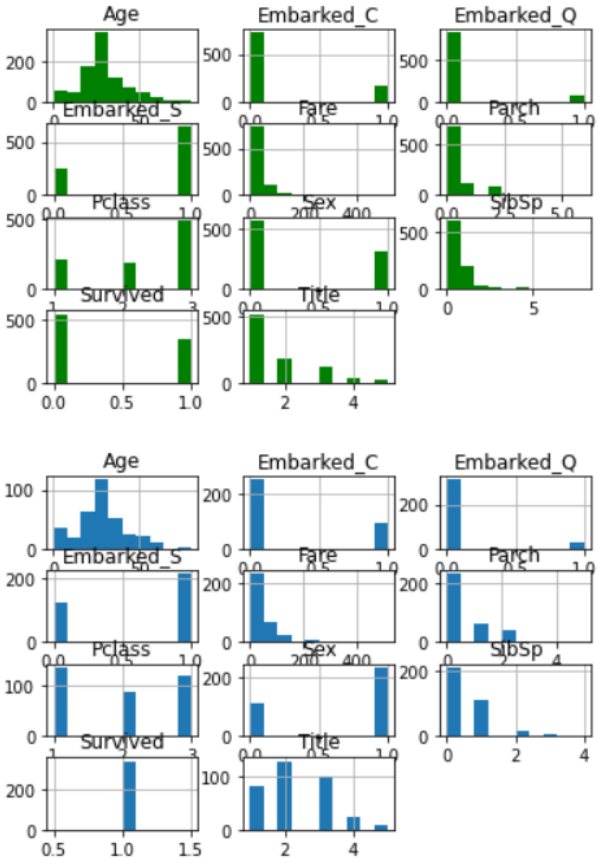
Regresion logística

Regresiones logísticas nos permiten crear un modelo para predecir el resultado de una variable categórica.

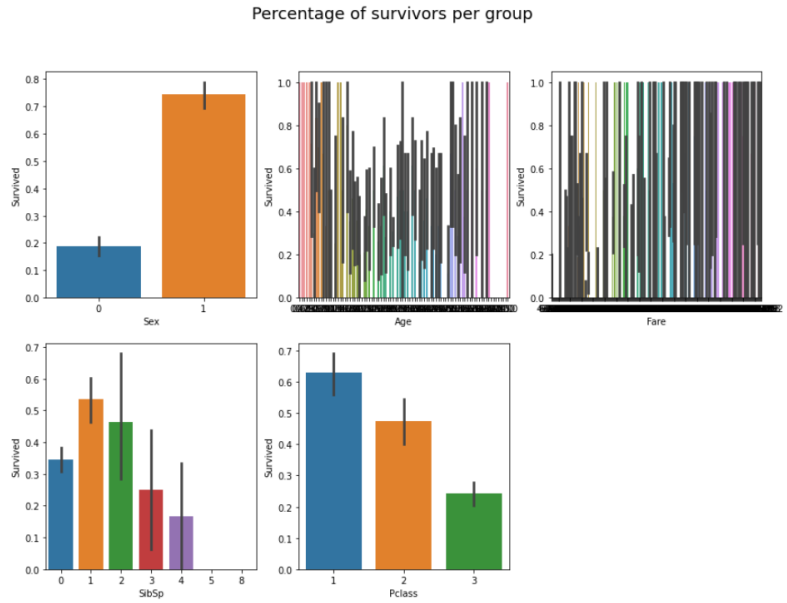
	precision	recall	f1-score	support
0	0.80	0.90	0.85	154
1	0.83	0.70	0.76	114
micro avg	0.81	0.81	0.81	268
macro avg	0.82	0.80	0.80	268
weighted avg	0.82	0.81	0.81	268

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

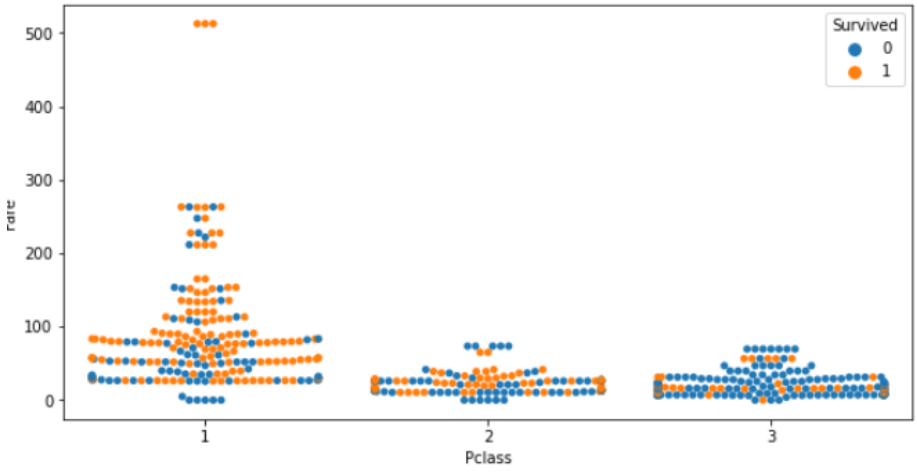
Vamos a visualizar primero los ratios de supervivencia para cada una de las variables:



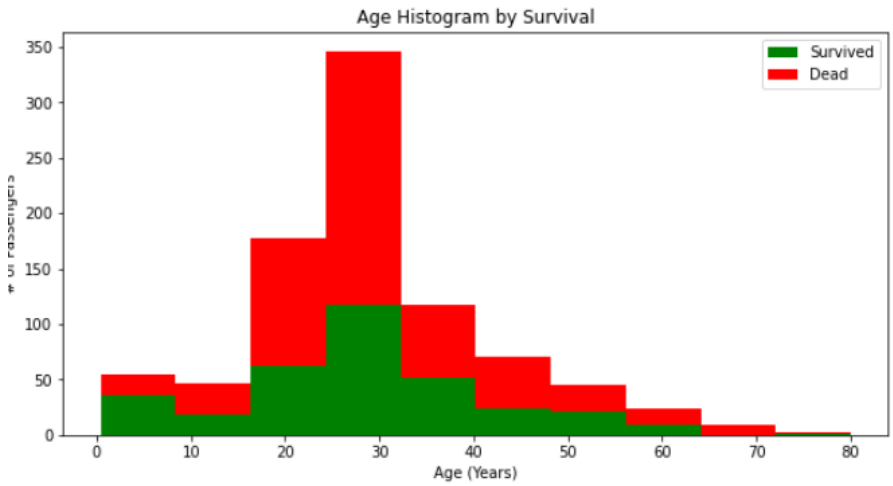
% Sobrevivientes por grupos



Comparamos la relación entre Pclass y Fare segun si han sobrevivido.



Veamos ahora una distribución de los supervivientes por edad.



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

Después de analizar los datos desde distintas perspectivas podemos validar la mayoría de las hipótesis iniciales planteadas en nuestro modelo:

- Gender – La variable gender nos indica el género de los tripulantes del Titanic. Hemos podido confirmar que las mujeres tienen mejor ratio de supervivencia que los hombres.
- Age – Inicialmente postulamos que los niños tendrían preferencia para salvarse. Nuestros análisis han validado que, en edades menores, el ratio de supervivencia es ligeramente mejor, aunque no es un factor determinante por si solo.
- Passenger Class -- Los pasajeros de 1a clase son los primeros en abandonar el barco en botes. Esta hipótesis se ha podido confirmar ampliamente, siendo uno de los factores mas determinantes junto al precio pagado por el billete (Fare)
- Fare – La variable Fare en conjunción con la clase ilustra de forma muy clara en la última representación como afecta claramente a la ratio de supervivencia, favoreciendo a las clases altas con billetes mas caros,
- Solos o en group – En el análisis de las distribuciones de cada variable, la variable SibSp muestra que aquellas personas viajando solas tuvieron una menor ratio de supervivencia.

Nuestra conclusión final que perfiles de pasajeros mujeres, o con familia incluidas mujeres, de clases altas y con precios de billetes altos fueron las mas favorecidas en el desastre del Titanic. Intuitivamente podemos deducir que dicha población dispuso de ciertos privilegios privados a otros segmentos de pasajeros.

Contribuciones	Firma
Investigación prèvia	HMM, MSZ
Redacción de las respuestas	HMM, MSZ
Desarrollo código	HMM, MSZ