

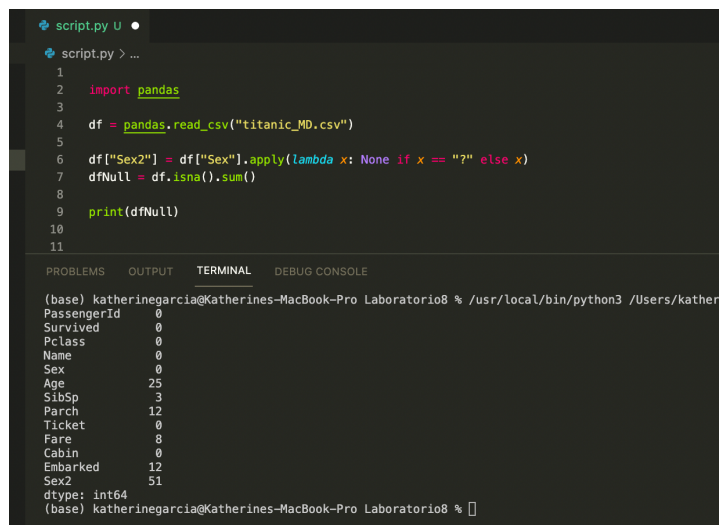
## Laboratorio 8

<https://github.com/katherineggs/dataWrangling#laboratorio-8---python>

### Parte 1

#### 1. Missing data

- En el dataframe de titanic\_MD hay 6 columnas con missing values. Los cuales son; Age, SibSp, Parch, Fare, Embarked, Sex2.
- Sex2 es una columna creada para los pasajeros que tenían un '?' en la especificación de género.



```
script.py U
script.py > ...
1
2 import pandas
3
4 df = pandas.read_csv("titanic_MD.csv")
5
6 df["Sex2"] = df["Sex"].apply(lambda x: None if x == "?" else x)
7 dfNull = df.isna().sum()
8
9 print(dfNull)
10
11
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
(base) katherinegarcia@Katherines-MacBook-Pro Laboratorio8 % /usr/local/bin/python3 /Users/katherinegarcia/...
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             25
SibSp            3
Parch           12
Ticket           0
Fare             8
Cabin            0
Embarked        12
Sex2             51
dtype: int64
(base) katherinegarcia@Katherines-MacBook-Pro Laboratorio8 %
```

#### 2. Modelo a usar para los missing values

- Edad
  - o Regresión Lineal
  - o Ya que la edad pudo afectar en la supervivencia de la persona, se buscará hacer de manera mas exacta con este método. Se tomara en cuenta si sobrevivió, la tarifa y la clase del ticket.
- SibSp
  - o Imputacion de la moda
  - o Al tener unicamente dos resultados en esta columna y pocos datos faltantes se puede colocar el valor mas repetido sin sesgar demasiado la data.
- Parch
  - o Imputacion de la moda

- De igual manera, al tener unicamente dos resultados en esta columna y pocos datos faltantes se puede colocar el valor mas repetido sin sesgar demasiado la data.
- Tarifa
  - Imputación del promedio
  - Se utilizará este modelo debido a los pocos datos con los que no se cuenta y que no se desea tener tarifas que sesguen demasiado los datos, sino que se mantengan dentro de el rango.
- Embarque
  - Imputación de la moda
  - Esta variable no se considera una que sesgue de manera significativa el hecho de si la persona sobrevivió o no. Por lo que se rellenarán los valores con la moda.
- Género
  - Regresion lineal
  - Debido a que el sexo pudo afectar en la supervivencia, se tomará en cuenta la clase y si sobrevivió o no.

### 3. Filas completas

- Las filas que cuentan con los datos completos son PassengerId, Survived, Pclass, Name, Ticket, Cabin.
- Estas columnas se utilizarán de esta manera, sin modificaciones.

### 4. Resultados de los modelos

- Se creó un dataframe con las columnas que se encontraban llenas y luego se llenó con los datos ya modificados por los modelos.
- La columna edad se trabajo por medio de imputación de promedio y regresión lineal.

[183 rows x 16 columns]													
	ID	Sobrevivio?	Clase	Nombre	Ticket	Cabina	Edad	SibSp	Parch	Tarifa	Embarque	Genero	EdadLr
0	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	PC 17599	C85	38	1.0	0	71	C	female	38
1	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	113803	C123	35	1.0	0	53	S	female	35
2	7	0	1	McCarthy, Mr. Timothy J	17463	E46	54	0.0	0	51	S	male	54
3	11	1	3	Sandstrom, Miss. Marguerite Rut	PP 9549	G6	35	1.0	0	16	S	female	13
4	12	1	1	Bonnell, Miss. Elizabeth	113783	C103	58	0.0	0	26	S	female	58
..	...	...	...	...	...	...	...	...	...	...	...	...	...
178	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	11751	D35	47	1.0	1	78	S	female	47
179	873	0	1	Carlsson, Mr. Frans Olof	695	B51 B53 B55	35	0.0	0	5	S	male	43
180	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	11767	C50	56	0.0	0	83	S	female	56
181	888	1	1	Graham, Miss. Margaret Edith	112053	B42	19	0.0	0	30	S	female	19
182	890	1	1	Behr, Mr. Karl Howell	111369	C148	35	0.0	0	30	C	female	34

## 5. Comparación con los datos originales

Comparacion contra data original	
Columna ---	
True	167
False	16
Name: MatchGenero, dtype: int64	
Columna ---	
True	157
False	26
Name: MatchEdad, dtype: int64	
Media Data Modificada	35.59
Media Data Original	35.67
Columna ---	
True	156
False	27
Name: MatchEdadLr, dtype: int64	
Media Data Modificada	35.07
Media Data Original	35.67
Columna ---	
True	181
False	2
Name: MatchSibSp, dtype: int64	
Media Data Modificada	0.45
Media Data Original	0.46
Columna ---	
True	177
False	6
Name: MatchParch, dtype: int64	
Media Data Modificada	0.43
Media Data Original	0.48
Columna ---	
False	147
True	36
Name: MatchTarifa, dtype: int64	
Media Data Modificada	78.52
Media Data Original	78.68
Columna ---	
True	177
False	6
Name: MatchEmbarque, dtype: int64	

Se realizó una comparación sobre los datos modificados y los originales. Se presenta la cantidad de datos que fueron acertados y los que no.

Las columnas que menos acertaron fueron Tarifa y Edad con ambos modelos de manejo.

Se considera que la cantidad de aciertos erróneos que los modelos generaron es debido a que estas dos casillas presentan números que varían fuertemente.

La columna de edad, con regresión lineal, a pesar de tener la misma cantidad de datos erróneos que los missing values, se mantiene al lado del promedio de los datos originales.

## 6. Conclusiones

- Dependiendo del tipo de variable, categorica o numerica, las predicciones con regresión lineal pueden ser más o menos acertadas.
- Por ejemplo, para predecir la edad, no se cuentan con muchas variables que tengan una correlación alta con esta variable. Lo que provoca que la predicción no sea muy confiable.
- En otras variables como género, sibSp o embarque, la certeza de los modelos que se utilizaron fue mejor ya que los valores no eran muchos y no había mucho margen de error.
- Considero que, no existe una manera perfecta para lidiar con missing data. Depende de la percepción de la persona y su enfoque en los datos. Ya que todo esto afecta sobre qué variables son más importantes para el tipo de análisis que se esté conduciendo. Así que, el mejor modelo es el que se adapte mejor al tipo de variable e investigación que se esté haciendo.
- Con respecto a la data
- Se cuenta con un dataset con los datos de 183 personas que estuvieron a bordo del Titanic. Por un accidente en las hojas de papel, en donde se guardaban estos datos, se perdieron algunos datos. Luego de un proceso realizado a la data para que esta fuera funcional nuevamente se obtuvieron algunas conclusiones.
- De 123 personas que sobrevivieron 96 eran mujeres y 5 eran niños de género masculino. Se observa como sí fue real la frase "mujeres y niños primero".

```
223 # Sobrevivio siendo Mujer o niño
224 cond1 = dfRenewed["Sobrevivio?"] == 1
225 cond2 = dfRenewed["Genero"] == "female"
226 cond3 = dfRenewed["Genero"] != "female"
227 cond4 = dfRenewed["Edad"] < 18
228 # print(dfRenewed.where(cond1).sum())
229 print(dfRenewed.where(cond1 & cond2).sum()) # Mujer
230 print(dfRenewed.where(cond3 & cond4).sum()) # niño
231
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
print(dfRenewed.where(cond1 & cond2).sum()) # Mujer
ID 45268.0
Sobrevivio? 96.0
Clase 115.0
Edad 3182.0
SibSp 50.0
Parch 39.0
Tarifa 8329.0
EdadLr 3093.0
dtype: float64
/Users/katherinegarcia/Documents/Semestre6/DataWrangling/dataWrangling/Lab
n this will raise TypeError. Select only valid columns before calling the
print(dfRenewed.where(cond3 & cond4).sum()) # niño
ID 3864.0
Sobrevivio? 5.0
Clase 10.0
Edad 40.0
SibSp 2.0
Parch 10.0
Tarifa 484.0
EdadLr 40.0
dtype: float64
```

- Arriba de una tarifa de 83 todos estaban en 1era clase. Así que se podría concluir que la tarifa y la clase no están relacionadas de manera directa. Probablemente la variable Clase no se refiere a el tipo de servicio que se recibirá a bordo del barco.

```

232 # tarifa y clase
233 meanTarifa = dfRenewed["Tarifa"].mean()
234 cond1 = dfRenewed["Tarifa"] >= meanTarifa
235 dfnew = dfRenewed.where(cond1)
236 dfnew = dfnew.dropna()
237 print("\n\nPromedio tarifas", meanTarifa)
238 print(dfnew)
239 print(dfnew["Clase"].unique())
240
PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

Promedio tarifas 78.52459016393442
7 ID Sobrevivio? Clase Nombre Ticket Cabina Edad SibSp Parch Tarifa Embarque Genero EdadLr
10 63.0 0.0 1.0 Fortune, Mr. Charles Alexander 19950 C23 C25 C27 19.0 3.0 2.0 263.0 S male 19.0
13 89.0 1.0 1.0 Harris, Mr. Henry Birkhardt 36973 C83 45.0 1.0 0.0 83.0 S male 45.0
19 119.0 0.0 1.0 Fortune, Miss. Mabel Helen 19950 C23 C25 C27 35.0 3.0 2.0 263.0 S female 34.0
24 140.0 0.0 1.0 Baxter, Mr. Quigg Edmond PC 17558 B58 B60 24.0 0.0 1.0 247.0 C male 24.0
166 790.0 0.0 1.0 Giglio, Mr. Victor PC 17593 B86 24.0 0.0 0.0 79.0 C male 24.0
168 803.0 1.0 1.0 Guggenheim, Mr. Benjamin PC 17593 B82 B84 35.0 0.0 0.0 79.0 C male 43.0
171 821.0 1.0 1.0 Carter, Master. William Thornton II 113760 B96 B98 11.0 1.0 2.0 120.0 S male 11.0
173 836.0 1.0 1.0 Hays, Mrs. Charles Melville (Clara Jennings Gr... 12749 B69 52.0 1.0 1.0 93.0 S female 52.0
180 880.0 1.0 1.0 Compton, Miss. Sara Rebecca PC 17756 E49 39.0 1.0 1.0 83.0 C female 39.0
Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) 11767 C50 56.0 0.0 0.0 83.0 S female 56.0

[62 rows x 13 columns]
[1.]

```

## Parte 2

### 1. Normalice las columnas numéricas

- Las columnas numéricas de este dataset son Edad y Tarifa. Las demás variables no se consideran numéricas ya que representan una categoría.

- Edad

#### MinMax

--- NORMALIZACION ---	Original - AGE
MD - EDAD	
[0.475 ]	[0.46889226]
[0.4375]	[0.43095599]
[0.675 ]	[0.67121902]
[0.4375]	[0.0389479 ]
[0.725 ]	[0.72180071]
[0.425 ]	[0.41831057]
[0.4375]	[0.34243804]
[0.2375]	[0.22862924]
[0.6125]	[0.60799191]
[0.8125]	[0.81031866]
[0.5625]	[0.55741022]
[0.3625]	[0.35508346]
[0.3125]	[0.30450177]
[0.4375]	[0.27921093]
[0.575 ]	[0.57005564]
[0.8875]	[0.8861912 ]
[0.2875]	[0.27921093]
[0.2625]	[0.25392008]
[0.5875]	[0.58270106]
[0.3 ]	[0.29185635]
[0.4375]	[0.39934244]
[0.675 ]	[0.67121902]
[0.2375]	[0.22862924]
[0.4625]	[0.45624684]
[0.3 ]	[0.29185635]
[0.3 ]	[0.44992413]

#### Standardization

MD - EDAD	Original - AGE
[0.475 ]	[0.46889226]
[0.4375]	[0.43095599]
[0.675 ]	[0.67121902]
[0.4375]	[0.0389479 ]
[0.725 ]	[0.72180071]
[0.425 ]	[0.41831057]
[0.4375]	[0.34243804]
[0.2375]	[0.22862924]
[0.6125]	[0.60799191]
[0.8125]	[0.81031866]
[0.5625]	[0.55741022]
[0.3625]	[0.35508346]
[0.3125]	[0.30450177]
[0.4375]	[0.27921093]
[0.575 ]	[0.57005564]
[0.8875]	[0.8861912 ]
[0.2875]	[0.27921093]
[0.2625]	[0.25392008]
[0.5875]	[0.58270106]
[0.3 ]	[0.29185635]
[0.3 ]	[0.44992413]

#### MaxAbsScaler

MD - EDAD	Original - AGE
[0.475 ]	[0.475 ]
[0.4375]	[0.4375 ]
[0.675 ]	[0.675 ]
[0.4375]	[0.05 ]
[0.725 ]	[0.725 ]
[0.425 ]	[0.425 ]
[0.4375]	[0.35 ]
[0.2375]	[0.2375 ]
[0.6125]	[0.6125 ]
[0.8125]	[0.8125 ]
[0.5625]	[0.5625 ]
[0.3625]	[0.3625 ]
[0.3125]	[0.3125 ]
[0.4375]	[0.2875 ]
[0.575 ]	[0.575 ]
[0.8875]	[0.8875 ]
[0.2875]	[0.2875 ]
[0.2625]	[0.2625 ]
[0.5875]	[0.5875 ]
[0.3 ]	[0.3 ]
[0.4375]	[0.40625]
[0.675 ]	[0.675 ]
[0.2375]	[0.2375 ]
[0.4625]	[0.4625 ]
[0.3 ]	[0.3 ]
[0.45 ]	[0.45625]
[0.275 ]	[0.275 ]
[0.7625]	[0.7625 ]



- Tarifa  
MinMax

--- NORMALIZACION ---	[0.05859375]]
MD - TARIFA_____	Original - FARE_
[0.13867188]	[0.13913574]
[0.10351562]	[0.1036443 ]
[0.09960938]	[0.10122886]
[0.03125 ]	[0.03259623]
[0.05078125]	[0.05182215]
[0.02539062]	[0.02537431]
[0.06835938]	[0.06929139]
[0.51367188]	[0.51334181]
[0.1484375 ]	[0.14976542]
[0.11914062]	[0.12097534]
[0.16210938]	[0.16293235]
[0.01953125]	[0.02049464]
[0.15234375]	[0.01493181]
[0.51367188]	[0.51334181]
[0.11914062]	[0.11940565]
[0.06640625]	[0.06764049]
[0.12304688]	[0.12366717]
[0.15039062]	[0.15085515]
[0.1015625 ]	[0.10149724]
[0.48242188]	[0.48312843]
[0.02539062]	[0.02537431]
[0.15039062]	[0.15085515]

Standarization

MD - TARIFA_____	Original - FARE_
[[-0.10016537]	[[-9.71798041e-02]
[-0.33977665]	[-3.35997105e-01]
[-0.36640012]	[-3.52250282e-01]
[-0.83231094]	[-8.14070377e-01]
[-0.69919356]	[-6.84701648e-01]
[-0.87224615]	[-8.62665737e-01]
[-0.57938792]	[-5.67153412e-01]
[ 2.45568824]	[ 2.42080454e+00]
[-0.03360668]	[-2.56540009e-02]
[-0.23328275]	[-2.19378747e-01]
[ 0.05957548]	[ 6.29445343e-02]
[-0.91218136]	[-8.95500440e-01]
[-0.00698321]	[-9.32932001e-01]
[ 2.45568824]	[ 2.42080454e+00]
[-0.23328275]	[-2.29941015e-01]
[-0.59269966]	[-5.78262049e-01]
[-0.20665927]	[-2.01265812e-01]
[-0.02029494]	[-1.83213551e-02]
[-0.35308838]	[-3.50444374e-01]
[ 2.24270044]	[ 2.21750257e+00]
[-0.87224615]	[-8.62665737e-01]
[ 0.02029494]	[ 1.83213551e-02]

MaxAbsScaler

MD - TARIFA_____	Original - FARE_
[0.13867188]	[0.13913574]
[0.10351562]	[0.1036443 ]
[0.09960938]	[0.10122886]
[0.03125 ]	[0.03259623]
[0.05078125]	[0.05182215]
[0.02539062]	[0.02537431]
[0.06835938]	[0.06929139]
[0.51367188]	[0.51334181]
[0.1484375 ]	[0.14976542]
[0.11914062]	[0.12097534]
[0.16210938]	[0.16293235]
[0.01953125]	[0.02049464]
[0.15234375]	[0.01493181]
[0.51367188]	[0.51334181]
[0.11914062]	[0.11940565]
[0.06640625]	[0.06764049]
[0.12304688]	[0.12366717]
[0.15039062]	[0.15085515]
[0.1015625 ]	[0.10149724]
[0.48242188]	[0.48312843]
[0.02539062]	[0.02537431]
[0.15039062]	[0.15085515]
[0.05078125]	[0.05130158]
[0.10351562]	[0.1036443 ]

## 2. Comparación contra la data original normalizada

- En cuanto a los valores de la mayoría de los datos en el proceso de la normalización en ambas columnas, se mantuvieron cercanos y sin fluctuaciones mayores.
- Sin embargo, en el promedio se puede observar que el modelo menos acertado es la estandarización. Se cree que es debido a que esta oscila del -1 al 1 y por esto los valores fluctúan más.
- Para la variable Edad el mejor tipo de normalización es Max Abs Scaler.
- Por otro lado, para la variable Tarifa el mejor modelo es Min Max Scaling.
- Recalcando que ambos modelos, minMax y maxAbs, tienen un nivel de precisión casi exacto.

### Edad

-- Promedio min max scaling MD edad	-- Promedio Z value MD edad	-- Promedio Max AbsScaler MD edad
0.4448770491803279	3.397403791202665e-17	0.44487704918032783
Original age	Original age	Original age
0.4394843984510395	-1.6501675557270087e-16	0.4459303278688525

### Tarifa

-- Promedio min max scaling MD Tarifa	-- Promedio Z value MD Tarifa	-- Promedio Max AbsScaler MD Tarifa
0.1533683401639344	8.250837778635044e-17	0.1533683401639344
Original Fare	Original Fare	Original Fare
0.15357795115417788	1.140556987046609e-16	0.15357795115417786