Andrea Reyes
20190265

Laboratorio 8

Parte 1

1. Reporte detallado de missing data para todas las columnas.
   El dataset tiene 12 variables/columnas de las cuales 6 cuentan con missing data:

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	25
SibSp	3
Parch	12
Ticket	0
Fare	8
Cabin	0
Embarked	12
missingSex	51

2. Para cada columna especificar qué tipo de modelo se utilizará y qué valores se le darán a todos los missing values.
   - Age: imputación de promedio porque así tenemos edades dentro del margen para no sesgar a los datos existentes.
   - SibSp: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.
   - Parch: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.
   - Fare: imputación de promedio porque así tenemos valores dentro del margen y no son muchos los faltantes.
   - Embarked: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.

3. Reporte de qué filas están completas
   El dataset tiene 12 variables/columnas de las cuales 6 están completas (0's en cuanto a missing values):

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	25
SibSp	3
Parch	12
Ticket	0
Fare	8
Cabin	0
Embarked	12
missingSex	51

4. Utilizar los siguientes métodos para cada columna que contiene missing values
   Dataframe sin missing values:

	PassengerId	Survived	Pclass	Name	Ticket	Cabin	Age	SibSp	Parch	Fare	Embarked
0	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	PC 17599	C85	38	1.0	0	71	C
1	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	113803	C123	35	1.0	0	53	S
2	7	0	1	McCarthy, Mr. Timothy J	17463	E46	54	0.0	0	51	S
3	11	1	3	Sandstrom, Miss. Marguerite Rut	PP 9549	G6	35	1.0	0	16	S
4	12	1	1	Bonnell, Miss. Elizabeth	113783	C103	58	0.0	0	26	S

5. Comparar los métodos del inciso 4 contra "titanic.csv"
   Age y Fare son los que menos se acercan a la realidad.

6. Conclusiones
   La cantidad de sobrevivientes fue de 123
   El dataframe tiene variables de distintos tipos y hay modelos que no son muy efectivos dependiendo de a que tipo de variable se aplique.
   Es importante tomar en cuenta la cantidad de missing values en cada columna y si hay correlación.

Parte 2

1. Luego del pre-procesamiento de la data con Missing Values, normalice las columnas numéricas por los métodos:
    a. Standarization
    b. MinMaxScaling
    c. MaxAbsScaler

a.

```
Standarization

Fare:
[[-0.10016537]
 [-0.33977665]
 [-0.36640012]
 [-0.83231094]
 [-0.69919356]
 [-0.87224615]
 [-0.57938792]
 [ 2.45568824]
 [-0.03360668]
 [-0.23328275]
 [ 0.05957548]

Fare2:
[[-9.71798041e-02]
 [-3.35997105e-01]
 [-3.52250282e-01]
 [-8.14070377e-01]
 [-6.84701648e-01]
 [-8.62665737e-01]
 [-5.67153412e-01]
 [ 2.42080454e+00]
 [-2.56540009e-02]
 [-2.19378747e-01]
 [ 6.29445343e-02]

Age:
[[ 0.16618153]
 [-0.04069752]
 [ 1.26953643]
 [-0.04069752]
 [ 1.54537516]
 [-0.1096572 ]
 [-0.04069752]
 [-1.14405242]
 [ 0.92473802]
 [ 2.02809293]
 [ 0.6488993 ]

Age2:
[[ 0.14906507]
 [-0.0432295 ]
 [ 1.17463611]
 [-2.03027338]
 [ 1.43102886]
 [-0.10732769]
 [-0.49191683]
 [-1.06880054]
 [ 0.85414516]
 [ 1.87971619]
 [ 0.5977524 ]
```

b.

```
MinMaxScaling

Fare:
[[0.13867188]
 [0.10351562]
 [0.09960938]
 [0.03125   ]
 [0.05078125]
 [0.02539062]
 [0.06835938]
 [0.51367188]
 [0.1484375 ]
 [0.11914062]
 [0.16210938]

Fare2:
[[0.13913574]
 [0.1036443 ]
 [0.10122886]
 [0.03259623]
 [0.05182215]
 [0.02537431]
 [0.06929139]
 [0.51334181]
 [0.14976542]
 [0.12097534]
 [0.16293235]

Age:
[[0.475 ]
 [0.4375]
 [0.675 ]
 [0.4375]
 [0.725 ]
 [0.425 ]
 [0.4375]
 [0.2375]
 [0.6125]
 [0.8125]
 [0.5625]

Age2:
[[0.46889226]
 [0.43095599]
 [0.67121902]
 [0.0389479 ]
 [0.72180071]
 [0.41831057]
 [0.34243804]
 [0.22862924]
 [0.60799191]
 [0.81031866]
 [0.55741022]
```

c.

```
MaxAbsScaler

Fare:
[[0.13867188]
 [0.10351562]
 [0.09960938]
 [0.03125   ]
 [0.05078125]
 [0.02539062]
 [0.06835938]
 [0.51367188]
 [0.1484375 ]
 [0.11914062]
 [0.16210938]

Fare2:
[[0.13913574]
 [0.1036443 ]
 [0.10122886]
 [0.03259623]
 [0.05182215]
 [0.02537431]
 [0.06929139]
 [0.51334181]
 [0.14976542]
 [0.12097534]
 [0.16293235]

Age:
[[0.475 ]
 [0.4375]
 [0.675 ]
 [0.4375]
 [0.725 ]
 [0.425 ]
 [0.4375]
 [0.2375]
 [0.6125]
 [0.8125]
 [0.5625]

Age2:
[[0.475  ]
 [0.4375 ]
 [0.675  ]
 [0.05   ]
 [0.725  ]
 [0.425  ]
 [0.35   ]
 [0.2375 ]
 [0.6125 ]
 [0.8125 ]
 [0.5625 ]
```

2. Compare los estadísticos que considere más importantes para su conclusión y compare contra la data completa de “titanic.csv” (deberán de normalizar también).

```
Standarization

Fare:
8.250837778635044e-17
1.140556987046609e-16

Age:
3.397403791202665e-17
-1.6501675557270087e-16


MinMaxScaling

0.1533683401639344
0.15357795115417788

Age:
0.4448770491803279
0.4394843984510395


MaxAbsScaler

0.1533683401639344
0.15357795115417786

Age:
0.44487704918032783
0.4459303278688525
```

Hay métodos que funcionan mejor dependiendo del tipo de variable. Como se observa en las imágenes de arriba, los mejores modelos son MinMaxScaling y MaxAbsScaler.

Código: https://github.com/AndreaNathalia/data-wrangling/blob/main/Laboratorio8/Lab8.ipynb