Task

On

Data Preprocessing

**Course**: Artificial Intelligence

(Machine Learning & Deep Learning)

Name: Abbas Shafi

Father Name: Dr Muhammad Shafi

Week: 08

Submitted to: Sir Syed Nazir Afridi



National Vocational & Technical Training Commission

National Center For Big Data & Cloud Computing

University of Engineering & Technology

Peshawar

[2]:

# Data Cleaning

* 1. **Missing value imputation using Scikit-Learn**
     1. **for Numeric and Categorical Variables/Data**

**import numpy as np import pandas as pd**

**from sklearn.impute import** SimpleImputer

[5]:

shape of train df = (1460, 81) shape of test df = (1459, 80)

[6]:

[6]: Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape \

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 60 | RL | 65.0 | | 8450 | Pave | NaN | Reg | |
| 1 | 2 | 20 | RL | 80.0 | | 9600 | Pave | NaN | Reg | |
| 2 | 3 | 60 | RL | 68.0 | | 11250 | Pave | NaN | IR1 | |
| 3 | 4 | 70 | RL | 60.0 | | 9550 | Pave | NaN | IR1 | |
| 4 | 5 | 60 | RL | 84.0 | | 14260 | Pave | NaN | IR1 | |
| LandContour Utilities … PoolArea PoolQC Fence MiscFeature MiscVal MoSold | | | | | | | | | | |
| 0 | Lvl AllPub | | … | 0 NaN | | NaN | NaN | | 0 | 2 |
| 1 | Lvl AllPub | | … | 0 NaN | | NaN | NaN | | 0 | 5 |
| 2 | Lvl AllPub | | … | 0 NaN | | NaN | NaN | | 0 | 9 |
| 3 | Lvl AllPub | | … | 0 NaN | | NaN | NaN | | 0 | 2 |
| 4 | Lvl AllPub | | … | 0 NaN | | NaN | NaN | | 0 | 12 |
| YrSold | | SaleType SaleCondition SalePrice | | | | |  |  |  |  |
| 0 | 2008 | WD | Normal | | 208500 | | | | | |
| 1 | 2007 | WD | Normal | | 181500 | | | | | |

\

|  |  |  |  |
| --- | --- | --- | --- |
| 2 2008 | WD | Normal | 223500 |
| 3 2006 | WD | Abnorml | 140000 |
| 4 2008 | WD | Normal | 250000 |

[5 rows x 81 columns]

[7]:

*#In ML Algo we have test and train*

*#X is capital because it consider it as matrix*

X\_train=train.drop(columns="SalePrice") *#we drop saleprice because it is our*␣

*‹→target value*

y\_train=train["SalePrice"] *#it is a vector* print("shape of X\_train df = ",X\_train.shape) print("shape of y\_train df = ",y\_train.shape)

shape of X\_train df = (1460, 80) shape of y\_train df = (1460,)

# Numerical Missing Value Imputation

[8]:

[9]:

[9]: Index(['Id', 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'MoSold', 'YrSold'],

dtype='object')

[10]:

|  |  |  |
| --- | --- | --- |
| [10]: | Id | 0 |
|  | MSSubClass | 0 |
|  | LotFrontage | 259 |
|  | LotArea | 0 |
|  | OverallQual | 0 |
|  | OverallCond | 0 |
|  | YearBuilt | 0 |
|  | YearRemodAdd | 0 |
|  | MasVnrArea | 8 |
|  | BsmtFinSF1 | 0 |

|  |  |
| --- | --- |
| BsmtFinSF2 | 0 |
| BsmtUnfSF | 0 |
| TotalBsmtSF | 0 |
| 1stFlrSF | 0 |
| 2ndFlrSF | 0 |
| LowQualFinSF | 0 |
| GrLivArea | 0 |
| BsmtFullBath | 0 |
| BsmtHalfBath | 0 |
| FullBath | 0 |
| HalfBath | 0 |
| BedroomAbvGr | 0 |
| KitchenAbvGr | 0 |
| TotRmsAbvGrd | 0 |
| Fireplaces | 0 |
| GarageYrBlt | 81 |
| GarageCars | 0 |
| GarageArea | 0 |
| WoodDeckSF | 0 |
| OpenPorchSF | 0 |
| EnclosedPorch | 0 |
| 3SsnPorch | 0 |
| ScreenPorch | 0 |
| PoolArea | 0 |
| MiscVal | 0 |
| MoSold | 0 |
| YrSold | 0 |
| dtype: int64 |  |

[11]:

[12]:

[12]: SimpleImputer()

[13]:

[13]: array([7.30500000e+02, 5.68972603e+01, 7.00499584e+01, 1.05168281e+04, 6.09931507e+00, 5.57534247e+00, 1.97126781e+03, 1.98486575e+03,

1.03685262e+02, 4.43639726e+02, 4.65493151e+01, 5.67240411e+02,

1.05742945e+03, 1.16262671e+03, 3.46992466e+02, 5.84452055e+00,

[14]:

1.51546370e+03, 4.25342466e-01, 5.75342466e-02, 1.56506849e+00,

3.82876712e-01, 2.86643836e+00, 1.04657534e+00, 6.51780822e+00,

6.13013699e-01, 1.97850616e+03, 1.76712329e+00, 4.72980137e+02,

9.42445205e+01, 4.66602740e+01, 2.19541096e+01, 3.40958904e+00,

1.50609589e+01, 2.75890411e+00, 4.34890411e+01, 6.32191781e+00,

2.00781575e+03])

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [14]: | array([[1.000e+00, 6.000e+01, | 6.500e+01, | …, | 0.000e+00, | 2.000e+00, |
|  | 2.008e+03], |  |  |  |  |
|  | [2.000e+00, 2.000e+01, | 8.000e+01, | …, | 0.000e+00, | 5.000e+00, |
|  | 2.007e+03], |  |  |  |  |
|  | [3.000e+00, 6.000e+01, | 6.800e+01, | …, | 0.000e+00, | 9.000e+00, |
|  | 2.008e+03], |  |  |  |  |
|  | …, |  |  |  |  |
|  | [1.458e+03, 7.000e+01, | 6.600e+01, | …, | 2.500e+03, | 5.000e+00, |
|  | 2.010e+03], |  |  |  |  |
|  | [1.459e+03, 2.000e+01, | 6.800e+01, | …, | 0.000e+00, | 4.000e+00, |
|  | 2.010e+03], |  |  |  |  |
|  | [1.460e+03, 2.000e+01, | 7.500e+01, | …, | 0.000e+00, | 6.000e+00, |
|  | 2.008e+03]]) |  |  |  |  |

[16]:

[17]:

[17]: Id 0

MSSubClass 0

LotFrontage 0

LotArea 0

OverallQual 0

OverallCond 0

YearBuilt 0

YearRemodAdd 0

MasVnrArea 0

BsmtFinSF1 0

BsmtFinSF2 0

BsmtUnfSF 0

TotalBsmtSF 0

[18]:

1stFlrSF 0

2ndFlrSF 0

LowQualFinSF 0

GrLivArea 0

BsmtFullBath 0

BsmtHalfBath 0

FullBath 0

HalfBath 0

BedroomAbvGr 0

KitchenAbvGr 0

TotRmsAbvGrd 0

Fireplaces 0

GarageYrBlt 0

GarageCars 0

GarageArea 0

WoodDeckSF 0

OpenPorchSF 0

EnclosedPorch 0

3SsnPorch 0

ScreenPorch 0

PoolArea 0

MiscVal 0

MoSold 0

YrSold 0

dtype: int64

[18]: Id 0

MSSubClass 0

LotFrontage 0

LotArea 0

OverallQual 0

OverallCond 0

YearBuilt 0

YearRemodAdd 0

MasVnrArea 0

BsmtFinSF1 0

BsmtFinSF2 0

BsmtUnfSF 0

TotalBsmtSF 0

1stFlrSF 0

2ndFlrSF 0

LowQualFinSF 0

GrLivArea 0

BsmtFullBath 0

BsmtHalfBath 0

FullBath 0

HalfBath 0

BedroomAbvGr 0

KitchenAbvGr 0

TotRmsAbvGrd 0

Fireplaces 0

GarageYrBlt 0

GarageCars 0

GarageArea 0

WoodDeckSF 0

OpenPorchSF 0

EnclosedPorch 0

3SsnPorch 0

ScreenPorch 0

PoolArea 0

MiscVal 0

MoSold 0

YrSold 0

dtype: int64

# Categorical Missing Value Imputation

[19]:

[19]: Index(['MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature', 'SaleType', 'SaleCondition'],

dtype='object')

[20]:

|  |  |
| --- | --- |
| [20]: MSZoning | 0 |
| Street | 0 |
| Alley | 1369 |
| LotShape | 0 |
| LandContour | 0 |
| Utilities | 0 |
| LotConfig | 0 |
| LandSlope | 0 |

|  |  |
| --- | --- |
| Neighborhood | 0 |
| Condition1 | 0 |
| Condition2 | 0 |
| BldgType | 0 |
| HouseStyle | 0 |
| RoofStyle | 0 |
| RoofMatl | 0 |
| Exterior1st | 0 |
| Exterior2nd | 0 |
| MasVnrType | 8 |
| ExterQual | 0 |
| ExterCond | 0 |
| Foundation | 0 |
| BsmtQual | 37 |
| BsmtCond | 37 |
| BsmtExposure | 38 |
| BsmtFinType1 | 37 |
| BsmtFinType2 | 38 |
| Heating | 0 |
| HeatingQC | 0 |
| CentralAir | 0 |
| Electrical | 1 |
| KitchenQual | 0 |
| Functional | 0 |
| FireplaceQu | 690 |
| GarageType | 81 |
| GarageFinish | 81 |
| GarageQual | 81 |
| GarageCond | 81 |
| PavedDrive | 0 |
| PoolQC | 1453 |
| Fence | 1179 |
| MiscFeature | 1406 |
| SaleType | 0 |
| SaleCondition | 0 |
| dtype: int64 |  |

[21]:

1. : SimpleImputer(strategy='most\_frequent')

[22]:

1. : SimpleImputer(strategy='most\_frequent')

[23]:

1. : array(['RL', 'Pave', 'Grvl', 'Reg', 'Lvl', 'AllPub', 'Inside', 'Gtl',

'NAmes', 'Norm', 'Norm', '1Fam', '1Story', 'Gable', 'CompShg',

'VinylSd', 'VinylSd', 'None', 'TA', 'TA', 'PConc', 'TA', 'TA',

'No', 'Unf', 'Unf', 'GasA', 'Ex', 'Y', 'SBrkr', 'TA', 'Typ', 'Gd',

'Attchd', 'Unf', 'TA', 'TA', 'Y', 'Gd', 'MnPrv', 'Shed', 'WD',

'Normal'], dtype=object)

[24]:

[25]:

1. : MSZoning 0

Street 0

Alley 0

LotShape 0

LandContour 0

Utilities 0

LotConfig 0

LandSlope 0

Neighborhood 0

Condition1 0

Condition2 0

BldgType 0

HouseStyle 0

RoofStyle 0

RoofMatl 0

Exterior1st 0

Exterior2nd 0

MasVnrType 0

ExterQual 0

ExterCond 0

Foundation 0

BsmtQual 0

BsmtCond 0

BsmtExposure 0

BsmtFinType1 0

BsmtFinType2 0

Heating 0

HeatingQC 0

CentralAir 0

Electrical 0

KitchenQual 0

[ ]:

[26]:

Functional 0

FireplaceQu 0

GarageType 0

GarageFinish 0

GarageQual 0

GarageCond 0

PavedDrive 0

PoolQC 0

Fence 0

MiscFeature 0

SaleType 0

SaleCondition 0

dtype: int64

1. : MSZoning 0

Street 0

Alley 0

LotShape 0

LandContour 0

Utilities 0

LotConfig 0

LandSlope 0

Neighborhood 0

Condition1 0

Condition2 0

BldgType 0

HouseStyle 0

RoofStyle 0

RoofMatl 0

Exterior1st 0

Exterior2nd 0

MasVnrType 0

ExterQual 0

ExterCond 0

Foundation 0

BsmtQual 0

BsmtCond 0

BsmtExposure 0

BsmtFinType1 0

BsmtFinType2 0

Heating 0

HeatingQC 0

CentralAir 0

Electrical 0

KitchenQual 0

Functional 0

FireplaceQu 0

GarageType 0

GarageFinish 0

GarageQual 0

GarageCond 0

PavedDrive 0

PoolQC 0

Fence 0

MiscFeature 0

SaleType 0

SaleCondition 0

dtype: int64

[27]:

[27]: 0

# One Hot Encoding

[3]:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [3]: | total\_bill | tip | sex | smoker | day | time | size |
|  | 0 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
|  | 1 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
|  | 2 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
|  | 3 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
|  | 4 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |
|  | .. … | … | … … | … | … … |  |  |
|  | 239 29.03 | 5.92 | Male | No | Sat | Dinner | 3 |
|  | 240 27.18 | 2.00 | Female | Yes | Sat | Dinner | 2 |
|  | 241 22.67 | 2.00 | Male | Yes | Sat | Dinner | 2 |
|  | 242 17.82 | 1.75 | Male | No | Sat | Dinner | 2 |
|  | 243 18.78 | 3.00 | Female | No | Thur | Dinner | 2 |

[244 rows x 7 columns]

[4]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [4]: |  | total\_bill | | | tip | | size | sex\_Female | | | sex\_Male | smoker\_No | | smoker\_Yes | \ |
|  | 0 | 16.99 | | | 1.01 | | 2 | 1 | | | 0 | 1 | | 0 |  |
|  | 1 | 10.34 | | | 1.66 | | 3 | 0 | | | 1 | 1 | | 0 |  |
|  | 2 | 21.01 | | | 3.50 | | 3 | 0 | | | 1 | 1 | | 0 |  |
|  | 3 | 23.68 | | | 3.31 | | 2 | 0 | | | 1 | 1 | | 0 |  |
|  | 4 | 24.59 | | | 3.61 | | 4 | 1 | | | 0 | 1 | | 0 |  |
|  | .. | … | | | … … | |  | … … | | | … | … | |  |  |
|  | 239 | 29.03 | | | 5.92 | | 3 | 0 | | | 1 | 1 | | 0 |  |
|  | 240 | 27.18 | | | 2.00 | | 2 | 1 | | | 0 | 0 | | 1 |  |
|  | 241 | 22.67 | | | 2.00 | | 2 | 0 | | | 1 | 0 | | 1 |  |
|  | 242 | 17.82 | | | 1.75 | | 2 | 0 | | | 1 | 1 | | 0 |  |
|  | 243 | 18.78 | | | 3.00 | | 2 | 1 | | | 0 | 1 | | 0 |  |
|  | | day\_Fri | | day\_Sat | | | day\_Sun | | day\_Thur | | time\_Dinner | | time\_Lunch | | |
| 0 | | 0 | | 0 | | | 1 | | 0 | | 1 | | 0 | | |
| 1 | | 0 | | 0 | | | 1 | | 0 | | 1 | | 0 | | |
| 2 | | 0 | | 0 | | | 1 | | 0 | | 1 | | 0 | | |
| 3 | | 0 | | 0 | | | 1 | | 0 | | 1 | | 0 | | |
| 4 | | 0 | | 0 | | | 1 | | 0 | | 1 | | 0 | | |
| .. … | | |  | … | |  | … |  | … |  | … | … |  | | |
| 239 | | | 0 |  | | 1 |  | 0 |  | 0 |  | 1 | 0 | | |
| 240 | | | 0 |  | | 1 |  | 0 |  | 0 |  | 1 | 0 | | |
| 241 | | | 0 |  | | 1 |  | 0 |  | 0 |  | 1 | 0 | | |
| 242 | | | 0 |  | | 1 |  | 0 |  | 0 |  | 1 | 0 | | |
| 243 | | | 0 |  | | 0 |  | 0 |  | 1 |  | 1 | 0 | | |

[244 rows x 13 columns]

[5]:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [5]: | total\_bill | tip | size | sex\_Male | smoker\_Yes | day\_Sat | day\_Sun | day\_Thur | \ |
|  | 0 16.99 | 1.01 | 2 | 0 | 0 | 0 | 1 | 0 |  |
|  | 1 10.34 | 1.66 | 3 | 1 | 0 | 0 | 1 | 0 |  |
|  | 2 21.01 | 3.50 | 3 | 1 | 0 | 0 | 1 | 0 |  |
|  | 3 23.68 | 3.31 | 2 | 1 | 0 | 0 | 1 | 0 |  |
|  | 4 24.59 | 3.61 | 4 | 0 | 0 | 0 | 1 | 0 |  |
|  | .. … | … … |  | … | … … | … | … |  |  |
|  | 239 29.03 | 5.92 | 3 | 1 | 0 | 1 | 0 | 0 |  |
|  | 240 27.18 | 2.00 | 2 | 0 | 1 | 1 | 0 | 0 |  |
|  | 241 22.67 | 2.00 | 2 | 1 | 1 | 1 | 0 | 0 |  |
|  | 242 17.82 | 1.75 | 2 | 1 | 0 | 1 | 0 | 0 |  |
|  | 243 18.78 | 3.00 | 2 | 0 | 0 | 0 | 0 | 1 |  |
|  | time\_Lunch  0 0 |  |  |  |  |  |  |  |  |
|  | 1 0 |  |  |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| 2 |  | 0 |
| 3 |  | 0 |
| 4 |  | 0 |
| .. | … |  |
| 239 |  | 0 |
| 240 |  | 0 |
| 241 |  | 0 |
| 242 |  | 0 |
| 243 |  | 0 |

[244 rows x 9 columns]

[6]:

[7]:

[8]:

|  |  |  |  |
| --- | --- | --- | --- |
| [8]: | array([[0., | 0., 0., 1., 0., | 0.], |
|  | [1., | 0., 0., 1., 0., | 0.], |
|  | [1., | 0., 0., 1., 0., | 0.], |
|  | …, |  |  |
|  | [1., | 1., 1., 0., 0., | 0.], |
|  | [1., | 0., 1., 0., 0., | 0.], |
|  | [0., | 0., 0., 0., 1., | 0.]]) |

[9]:

1. : Index(['total\_bill', 'tip', 'size', 'sex\_Male', 'smoker\_Yes', 'day\_Sat', 'day\_Sun', 'day\_Thur', 'time\_Lunch'],

dtype='object')

[10]:

1. : sex\_Male smoker\_Yes day\_Sat day\_Sun day\_Thur time\_Lunch

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 1 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 2 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 1.0 |  | 0.0 |  | 0.0 |  | 1.0 | 0.0 | 0.0 |
| 4 | 0.0 |  | 0.0 |  | 0.0 |  | 1.0 | 0.0 | 0.0 |
| .. | … | … |  | … |  | … | … | … |  |
| 239 | 1.0 |  | 0.0 |  | 1.0 |  | 0.0 | 0.0 | 0.0 |
| 240 | 0.0 |  | 1.0 |  | 1.0 |  | 0.0 | 0.0 | 0.0 |
| 241 | 1.0 |  | 1.0 |  | 1.0 |  | 0.0 | 0.0 | 0.0 |
| 242 | 1.0 |  | 0.0 |  | 1.0 |  | 0.0 | 0.0 | 0.0 |
| 243 | 0.0 |  | 0.0 |  | 0.0 |  | 0.0 | 1.0 | 0.0 |

[13]:

[244 rows x 6 columns]

# Label Encoding And Ordinal Encoding

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [13]: | PassengerId | Survived |  | Pclass \ |  | | | |
|  | 0 1 | 0 |  | 3 |  |  |  |  |
|  | 1 2 | 1 |  | 1 |  |  |  |  |
|  | 2 3 | 1 |  | 3 |  |  |  |  |
|  | 3 4 | 1 |  | 1 |  |  |  |  |
|  | 4 5 | 0 |  | 3 |  |  |  |  |
|  | .. … | … | … |  |  |  |  |  |
|  | 886 887 | 0 |  | 2 |  |  |  |  |
|  | 887 888 | 1 |  | 1 |  |  |  |  |
|  | 888 889 | 0 |  | 3 |  |  |  |  |
|  | 889 890 | 1 |  | 1 |  |  |  |  |
|  | 890 891 | 0 |  | 3 | Name | Sex Age | SibSp | \ |

1. Braund, Mr. Owen Harris male 22.0 1
2. Cumings, Mrs. John Bradley (Florence Briggs Th… female 38.0 1
3. Heikkinen, Miss. Laina female 26.0 0
4. Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1
5. Allen, Mr. William Henry male 35.0 0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| .. | … | | | … … | … |  |
| 886 | Montvila, Rev. Juozas | | | male | 27.0 | 0 |
| 887 | Graham, Miss. Margaret Edith | | | female | 19.0 | 0 |
| 888 | Johnston, Miss. Catherine Helen "Carrie" | | | female | NaN | 1 |
| 889 | Behr, Mr. Karl Howell | | | male | 26.0 | 0 |
| 890 | Dooley, Mr. Patrick | | | male | 32.0 | 0 |
| 0 | Parch  0 | Ticket A/5 21171 | Fare Cabin Embarked 7.2500 NaN S | | | |

1 0 PC 17599 71.2833 C85 C

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 0 | 373450 | 8.0500 | NaN | S |
| .. … |  | … | … … | … |  |
| 886 | 0 | 211536 | 13.0000 | NaN | S |
| 887 | 0 | 112053 | 30.0000 | B42 | S |
| 888 | 2 | W./C. 6607 | 23.4500 | NaN | S |
| 889 | 0 | 111369 | 30.0000 | C148 | C |
| 890 | 0 | 370376 | 7.7500 | NaN | Q |

[891 rows x 12 columns]

[14]:

[14]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],

dtype='object')

[19]:

[19]: Name Embarked

1. Braund, Mr. Owen Harris S
2. Cumings, Mrs. John Bradley (Florence Briggs Th… C
3. Heikkinen, Miss. Laina S
4. Futrelle, Mrs. Jacques Heath (Lily May Peel) S
5. Allen, Mr. William Henry S

.. … …

886 Montvila, Rev. Juozas S

887 Graham, Miss. Margaret Edith S 888 Johnston, Miss. Catherine Helen "Carrie" S 889 Behr, Mr. Karl Howell C

890 Dooley, Mr. Patrick Q

[891 rows x 2 columns]

[20]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [20]: | array([108, | 190, | 353, | 272, | 15, | 554, | 515, | 624, | 412, | 576, | 727, | 95, | 729, |
|  | 28, | 840, | 359, | 682, | 867, | 839, | 512, | 273, | 80, | 523, | 765, | 626, | 44, |
|  | 240, | 260, | 605, | 813, | 828, | 776, | 289, | 856, | 535, | 372, | 505, | 134, | 837, |
|  | 589, | 7, | 827, | 446, | 456, | 212, | 697, | 473, | 604, | 726, | 38, | 629, | 594, |
|  | 333, | 246, | 620, | 873, | 707, | 595, | 853, | 299, | 749, | 385, | 337, | 755, | 784, |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 560, | 596, | 185, | 25, | 440, | 396, | 300, | 376, | 161, | 87, | 546, | 778, | 563, |
| 127, | 223, | 842, | 737, | 519, | 141, | 386, | 52, | 255, | 764, | 259, | 149, | 159, |
| 31, | 150, | 209, | 184, | 740, | 292, | 308, | 219, | 422, | 656, | 658, | 858, | 405, |
| 312, | 542, | 724, | 559, | 675, | 552, | 667, | 882, | 67, | 417, | 46, | 645, | 178, |
| 826, | 70, | 24, | 362, | 551, | 575, | 848, | 857, | 588, | 525, | 499, | 653, | 234, |
| 225, | 165, | 695, | 851, | 772, | 685, | 585, | 271, | 618, | 282, | 99, | 597, | 319, |
| 116, | 32, | 586, | 29, | 254, | 580, | 121, | 68, | 643, | 532, | 889, | 612, | 866, |
| 287, | 181, | 767, | 716, | 186, | 847, | 82, | 129, | 628, | 293, | 157, | 760, | 69, |
| 486, | 832, | 680, | 408, | 752, | 769, | 444, | 464, | 388, | 320, | 474, | 717, | 650, |
| 41, | 76, | 441, | 700, | 601, | 698, | 101, | 823, | 664, | 136, | 20, | 579, | 113, |
| 495, | 533, | 611, | 497, | 881, | 834, | 721, | 403, | 878, | 166, | 789, | 51, | 9, |
| 140, | 94, | 12, | 131, | 649, | 288, | 437, | 582, | 375, | 390, | 72, | 338, | 791, |
| 105, | 306, | 581, | 379, | 83, | 531, | 492, | 243, | 467, | 339, | 460, | 754, | 43, |
| 479, | 329, | 370, | 172, | 648, | 383, | 883, | 567, | 169, | 500, | 47, | 539, | 481, |
| 322, | 78, | 147, | 673, | 790, | 782, | 487, | 702, | 819, | 808, | 156, | 844, | 636, |
| 771, | 42, | 799, | 340, | 356, | 674, | 631, | 652, | 305, | 91, | 125, | 817, | 530, |
| 577, | 350, | 33, | 482, | 634, | 681, | 2, | 227, | 614, | 887, | 221, | 770, | 779, |
| 886, | 570, | 378, | 177, | 65, | 90, | 477, | 315, | 541, | 478, | 324, | 17, | 714, |
| 71, | 431, | 518, | 411, | 428, | 868, | 16, | 248, | 647, | 3, | 263, | 348, | 710, |
| 449, | 355, | 342, | 590, | 423, | 555, | 859, | 775, | 211, | 200, | 762, | 128, | 722, |
| 877, | 598, | 60, | 295, | 363, | 517, | 637, | 304, | 838, | 265, | 210, | 642, | 117, |
| 192, | 93, | 578, | 258, | 170, | 733, | 262, | 111, | 768, | 207, | 183, | 215, | 608, |
| 870, | 237, | 37, | 879, | 835, | 104, | 270, | 522, | 545, | 759, | 888, | 64, | 40, |
| 599, | 5, | 845, | 562, | 400, | 48, | 328, | 863, | 75, | 690, | 625, | 536, | 454, |
| 862, | 85, | 314, | 86, | 572, | 811, | 373, | 665, | 205, | 298, | 118, | 715, | 468, |
| 144, | 393, | 313, | 583, | 728, | 404, | 613, | 524, | 623, | 821, | 592, | 6, | 418, |
| 318, | 616, | 274, | 861, | 687, | 88, | 465, | 732, | 343, | 540, | 191, | 792, | 528, |
| 226, | 742, | 513, | 829, | 281, | 155, | 884, | 198, | 701, | 872, | 162, | 662, | 249, |
| 663, | 92, | 810, | 491, | 420, | 744, | 143, | 253, | 688, | 261, | 447, | 344, | 323, |
| 659, | 679, | 402, | 216, | 529, | 735, | 55, | 661, | 854, | 316, | 257, | 290, | 644, |
| 392, | 537, | 436, | 815, | 603, | 21, | 557, | 280, | 538, | 502, | 296, | 132, | 766, |
| 731, | 54, | 430, | 124, | 855, | 401, | 787, | 164, | 676, | 107, | 425, | 365, | 297, |
| 268, | 706, | 825, | 89, | 466, | 381, | 435, | 774, | 182, | 317, | 871, | 547, | 39, |
| 781, | 880, | 242, | 738, | 18, | 796, | 130, | 133, | 607, | 450, | 501, | 646, | 670, |
| 106, | 610, | 455, | 195, | 849, | 521, | 705, | 168, | 843, | 472, | 709, | 35, | 641, |
| 651, | 841, | 448, | 364, | 427, | 244, | 689, | 245, | 725, | 366, | 669, | 818, | 236, |
| 654, | 123, | 341, | 120, | 461, | 692, | 266, | 188, | 26, | 27, | 73, | 222, | 587, |
| 74, | 622, | 294, | 203, | 804, | 736, | 600, | 463, | 609, | 874, | 228, | 693, | 800, |
| 885, | 558, | 751, | 593, | 746, | 527, | 204, | 786, | 627, | 217, | 416, | 336, | 36, |
| 251, | 432, | 708, | 639, | 277, | 745, | 135, | 419, | 160, | 805, | 224, | 703, | 640, |
| 798, | 395, | 267, | 284, | 566, | 691, | 783, | 238, | 100, | 154, | 830, | 470, | 409, |
| 98, | 229, | 391, | 761, | 335, | 816, | 374, | 483, | 424, | 199, | 458, | 741, | 30, |
| 394, | 568, | 377, | 110, | 357, | 197, | 488, | 77, | 279, | 875, | 439, | 573, | 327, |
| 103, | 794, | 442, | 490, | 96, | 602, | 66, | 494, | 777, | 635, | 757, | 206, | 469, |
| 173, | 633, | 809, | 397, | 723, | 758, | 252, | 53, | 332, | 180, | 747, | 865, | 780, |
| 544, | 218, | 421, | 606, | 352, | 360, | 671, | 102, | 232, | 584, | 264, | 57, | 171, |
| 167, | 484, | 361, | 119, | 699, | 179, | 802, | 114, | 202, | 543, | 864, | 846, | 231, |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 730, | 824, | 302, | 137, | 655, | 345, | 615, | 301, | 112, | 457, | 632, | 194, | 247, |
| 498, | 213, | 426, | 452, | 713, | 850, | 153, | 433, | 565, | 803, | 382, | 45, | 743, |
| 63, | 275, | 326, | 556, | 434, | 126, | 163, | 561, | 514, | 443, | 801, | 459, | 307, |
| 773, | 241, | 822, | 520, | 410, | 331, | 399, | 286, | 368, | 151, | 617, | 677, | 507, |
| 115, | 387, | 14, | 346, | 445, | 84, | 820, | 869, | 256, | 476, | 389, | 574, | 347, |
| 148, | 711, | 526, | 788, | 187, | 1, | 748, | 511, | 176, | 852, | 549, | 833, | 415, |
| 358, | 321, | 138, | 58, | 806, | 704, | 276, | 591, | 62, | 146, | 233, | 369, | 109, |
| 506, | 553, | 309, | 480, | 398, | 496, | 235, | 367, | 569, | 812, | 239, | 438, | 694, |
| 50, | 214, | 489, | 414, | 13, | 330, | 753, | 683, | 208, | 310, | 429, | 278, | 719, |
| 380, | 201, | 621, | 462, | 619, | 384, | 831, | 666, | 174, | 142, | 807, | 351, | 406, |
| 34, | 660, | 534, | 152, | 10, | 475, | 763, | 23, | 814, | 269, | 354, | 504, | 371, |
| 756, | 349, | 493, | 678, | 550, | 630, | 250, | 453, | 503, | 516, | 785, | 876, | 686, |
| 712, | 49, | 19, | 175, | 638, | 750, | 158, | 508, | 11, | 564, | 734, | 471, | 189, |
| 0, | 720, | 509, | 334, | 291, | 22, | 795, | 97, | 485, | 145, | 8, | 860, | 196, |
| 56, | 672, | 325, | 283, | 797, | 718, | 285, | 122, | 230, | 696, | 890, | 407, | 59, |
| 79, | 139, | 836, | 4, | 571, | 311, | 657, | 451, | 668, | 739, | 510, | 193, | 61, |
| 793, | 684, | 548, | 303, | 413, | 81, | 220]) |  |  |  |  |  |  |

[22]:

<ipython-input-22-7ac73fbe6c42>:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas- docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy

df2["Name\_lab\_enc"]=le.fit\_transform(df2["Name"])

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [22]: |  | Name | Embarked | Name\_lab\_enc |
|  | 0 | Braund, Mr. Owen Harris | S | 108 |
|  | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th… | C | 190 |
|  | 2 | Heikkinen, Miss. Laina | S | 353 |
|  | 3 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | S | 272 |
|  | 4 | Allen, Mr. William Henry | S | 15 |
|  | .. | … | … | … |
|  | 886 | Montvila, Rev. Juozas | S | 548 |
|  | 887 | Graham, Miss. Margaret Edith | S | 303 |
|  | 888 | Johnston, Miss. Catherine Helen "Carrie" | S | 413 |
|  | 889 | Behr, Mr. Karl Howell | C | 81 |
|  | 890 | Dooley, Mr. Patrick | Q | 220 |

[891 rows x 3 columns]

[23]:

1. : <bound method IndexOpsMixin.value\_counts of 0 Braund, Mr. Owen Harris
   1. Cumings, Mrs. John Bradley (Florence Briggs Th…
   2. Heikkinen, Miss. Laina
   3. Futrelle, Mrs. Jacques Heath (Lily May Peel)
   4. Allen, Mr. William Henry

…

1. Montvila, Rev. Juozas
2. Graham, Miss. Margaret Edith
3. Johnston, Miss. Catherine Helen "Carrie"
4. Behr, Mr. Karl Howell
5. Dooley, Mr. Patrick

Name: Name, Length: 891, dtype: object>

[24]:

[24]: S 644

C 168

Q 77

Name: Embarked, dtype: int64

[25]:

[27]:

[28]:

<ipython-input-27-d02e487440a2>:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas- docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy

df2['Embarked\_ord\_enc']=df2['Embarked'].map(order\_label)

|  |  |  |
| --- | --- | --- |
| [28]: | 0 | 108 |
|  | 1 | 190 |
|  | 2 | 353 |
|  | 3 | 272 |
|  | 4 | 15 |
|  | 5 | 554 |
|  | 6 | 515 |
|  | 7 | 624 |
|  | 8 | 412 |
|  | 9 | 576 |
|  | 10 | 727 |

[ ]:

11 95

12 729

13 28

14 840

15 359

16 682

17 867

18 839

19 512

20 273

21 80

22 523

23 765

24 626

25 44

26 240

27 260

28 605

29 813

30 828

31 776

32 289

33 856

34 535

35 372

36 505

37 134

38 837

39 589

40 7

41 827

42 446

43 456

44 212

45 697

46 473

47 604

48 726

49 38

Name: Name\_lab\_enc, dtype: int32