Practical No 01

Data Wrangling, I Perform the following operations using Python on any open source dataset (e.g., data.csv) Import all the required Python Libraries. Locate an open source data from the web (e.g. [https://www.kaggle.com](https://www.kaggle.com/)). Provide a clear description of the data and its source (i.e., URL of the web site).

Import all the python Libraries

**In [2]:**

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

**import matplotlib.pyplot as plt import seaborn as sns**

Load the Dataset into pandas data frame.

**In [3]:**

**df = pd.read\_csv('titanic\_train.csv') df**

**Out [3]:**

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

1 2 1 1

Cumings, Mrs. John Bradley (Florence Briggs Th...

female 38.0 1 0 PC 17599 71.2833 C85 C

3 4 1 1

Futrelle, Mrs. Jacques Heath (Lily May Peel)

female 35.0 1 0 113803 53.1000 C123 S

... ... ... ... ... ... ... ... ... ... ... ... ...

887 888 1 1 Graham, Miss. Margaret Edith

female 19.0 0 0 112053 30.0000 B42 S

889 890 1 1 Behr, Mr. Karl Howell

male 26.0 0 0 111369 30.0000 C148 C

891 rows × 12 columns

**In [4]:**

**Out [4]:**

**df.head() # It's showing top 5 result**

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

1 2 1 1

Cumings, Mrs. John Bradley (Florence Briggs Th...

female 38.0 1 0 PC 17599 71.2833 C85 C

0 1

0

3

Braund, Mr.

Owen Harris

male

22.0 1

0

A/5 21171 7.2500 NaN S

2 3

1

3

Heikkinen,

Miss. Laina

female 26.0 0

0

STON/O2.

3101282 7.9250 NaN S

4 5

0

3

Allen, Mr.

William Henry

male

35.0 0

0

373450

8.0500 NaN S

886 887

0

2

Montvila, Rev.

Juozas

male

27.0 0

0

211536

13.0000 NaN S

888 889

0

3

Johnston,

Miss. Catherine female NaN 1 Helen "Carrie"

2

W./C. 6607 23.4500 NaN S

890 891

0

3

Dooley, Mr.

Patrick

male 32.0 0

0

370376

7.7500 NaN Q

0 1

0

3

Braund, Mr. Owen

Harris

male

22.0 1

0

A/5 21171 7.2500 NaN S

2 3

1

3

Heikkinen, Miss.

Laina

female 26.0 0

0

STON/O2.

3101282 7.9250 NaN S

3 4 1 1

Futrelle, Mrs. Jacques Heath (Lily May Peel)

female 35.0 1 0 113803 53.1000 C123 S

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

**In [5]:**

**Out [5]:**

**df.tail() # It's showing bottom 5 result**

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked Juozas

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 886 887 0 2 Montvila, Rev. male | | | | | 27.0 | 0 | 0 | 211536 | 13.00 | NaN | S |
| 887 888 1 1 Graham, Miss. female | | | | | 19.0 | 0 | 0 | 112053 | 30.00 | B42 | S |
| Margaret Edith  Johnston, Miss.  888 889 0 3 Catherine Helen female | | | | | NaN | 1 | 2 | W./C. 6607 | 23.45 | NaN | S |
|  |  |  | "Carrie" |  |  |  |  |  |  |  |  |
| 889 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.00 | C148 | C |
| 890 891 | 0 | 3 | Dooley, Mr. Patrick | male | 32.0 | 0 | 0 | 370376 | 7.75 | NaN | Q |

Data Preprocessing: check for missing values in the data using pandas insult(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

**In [6]:**

**df.isnull().sum() # Caluclating the Null values**

|  |  |
| --- | --- |
| **Out [6]: PassengerId** | **0** |
| **Survived** | **0** |
| **Pclass** | **0** |
| **Name** | **0** |
| **Sex** | **0** |
| **Age** | **177** |
| **SibSp** | **0** |
| **Parch** | **0** |
| **Ticket** | **0** |
| **Fare** | **0** |
| **Cabin** | **687** |
| **Embarked** | **2** |
| **dtype: int64** |  |

**In [7]:**

# sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis') # Finding Null values by usin



4 5

0

3

Allen, Mr. William

Henry

male

35.0 0

0

373450

8.0500 NaN S

**Out [7]: <AxesSubplot:>**

**In [8]:**

**df['Age'].isnull().sum() # Calcualting the Null Values in AGE Columns**

**Out [8]: 177**

**In [9]:**

**df['Cabin'].isnull().sum() # Calcualting the Null Values in Cabin Columns**

**Out [9]: 687**

**In [10]:**

**df.describe() # Get some initial statistics.**

**Out [10]:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

PassengerId Survived Pclass Age SibSp Parch Fare

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

**In [11]:**

**df.info() # Getting some informatation about dataset**

**<class 'pandas.core.frame.DataFrame'> RangeIndex: 891 entries, 0 to 890 Data columns (total 12 columns):**

**# Column Non-Null Count Dtype**

1. **PassengerId 891 non-null int64**
2. **Survived 891 non-null int64**
3. **Pclass 891 non-null int64**
4. **Name 891 non-null object**
5. **Sex 891 non-null object**
6. **Age 714 non-null float64**
7. **SibSp 891 non-null int64**
8. **Parch 891 non-null int64**
9. **Ticket 891 non-null object**
10. **Fare 891 non-null float64**
11. **Cabin 204 non-null object**
12. **Embarked 889 non-null object dtypes: float64(2), int64(5), object(5) memory usage: 83.7+ KB**

**In [12]:**

**df.dtypes # Finding Data Types**

**Out [12]: PassengerId int64**

**Survived int64**

**Pclass int64**

**Name object**

**Sex object**

**Age float64**

**SibSp int64**

**Parch int64**

**Ticket object**

**Fare float64**

**Cabin object**

**Embarked object dtype: object**

**In [13]:**

**df.shape # Finding Dimensions of the data frame.**

**Out [13]: (891, 12)**

Making Impute function for ﬁlling Null values

**In [14]:**

**def impute\_age(cols): Age = cols[0] Pclass = cols[1]**

**if pd.isnull(Age): if Pclass == 1:**

**return 37**

**elif Pclass == 2: return 29**

**else:**

**return 24**

**else:**

**return Age**

**In [15]:**

**df['Age'] = df[['Age','Pclass']].apply(impute\_age,axis=1) # Appyling the function**

**In [16]:**

# df.drop('Cabin',axis=1,inplace=True) # Droping Cabin Column becasue here lots of null values so it

**In [17]:**

**df.dropna(inplace=True)**

**In [18]:**

**df.head()**

**Out [18]:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Embarked |
| 0 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | S |
| 1 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C |
|  |  |  | Briggs Th... |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 3  3 4 | | 1  1 | 3  1 | Heikkinen, Miss. Laina  Futrelle, Mrs. Jacques | female  female | 26.0  35.0 | 0  1 | 0 STON/O2. 7.9250 S  3101282  0 113803 53.1000 S | | | |
|  |  |  |  | Heath (Lily May Peel) |  |  |  |  |  |  |  |
|  | 4 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | S |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **In [19]:** | **df.tail()** |  |  |  |  |  |  |  |  |  |  |

**Out [19]:**

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Embarked

886 887 0 2 Montvila, Rev. Juozas male 27.0 0 0 211536 13.00 S

887 888 1 1 Graham, Miss. Margaret

Edith

female 19.0 0 0 112053 30.00 S

888 889

0

3

Johnston, Miss. Catherine

Helen "Carrie"

female 24.0 1

2

W./C.

6607

23.45 S

889 890 1 1 Behr, Mr. Karl Howell male 26.0 0 0 111369 30.00 C

890 891 0 3 Dooley, Mr. Patrick male 32.0 0 0 370376 7.75 Q

**In [20]:**

**df.isnull().sum()**

**Out [20]: PassengerId** **0**

**Survived** **0**

**Pclass** **0**

**Name 0**

**Sex 0**

**Age 0**

**SibSp** **0**

**Parch 0**

**Ticket** **0**

**Fare 0**

**Embarked** **0**

**dtype: int64**

Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

**In [21]:**

**df['Age'] = df['Age'].astype('int') # Data Type Coversion**

**In [22]:**

**df.dtypes**

**Out [22]: PassengerId int64**

**Survived int64**

**Pclass int64**

**Name object**

**Sex object**

**Age int64**

**SibSp int64**

**Parch int64**

**Ticket object**

**Fare float64**

**Embarked object dtype: object**

**In [23]:**

**df['Age'] = df['Age'].round(0).astype('int') # Data Type Coversion**

**In [24]:**

**df.dtypes**

**Out [24]: PassengerId int64**

**Survived int64**

**Pclass int64**

**Name object**

**Sex object**

**Age int64**

**SibSp int64**

**Parch int64**

**Ticket object**

**Fare float64**

**Embarked object dtype: object**

**In [25]:**

**df.head()**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Out [25]:** PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Embarked   1. 1 0 3 Braund, Mr. Owen Harris male 22 1 0 A/5 21171 7.2500 S   Cumings, Mrs. John   1. 2 1 1 Bradley (Florence Briggs female 38 1 0 PC 17599 71.2833 C   Th...   1. 3 1 3 Heikkinen, Miss. Laina female 26 0 0 STON/O2. 7.9250 S   3101282   1. 4 1 1 Futrelle, Mrs. Jacques female 35 1 0 113803 53.1000 S   Heath (Lily May Peel)   1. 5 0 3 Allen, Mr. William Henry male 35 0 0 373450 8.0500 S Converting Categorical Variables to Quantitative Variables   **In [31]: cat = pd.get\_dummies(df, columns=['Sex']) # Converting Categorical Variables to Quantitative Varia**  **In [32]: cat.head()**  **Out [32]:** PassengerId Survived Pclass Name Age SibSp Parch Ticket Fare Embarked female Sex\_female Sex\_m | | | | | | | | | | | | | | |
|  | 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | 22 | 1 | 0 | A/5 21171 | 7.2500 | S | 0 | 0 | 1 |
|  | 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | 38 | 1 | 0 | PC 17599 | 71.2833 | C | 1 | 1 | 0 |
|  | 2 | 3 | 1 | 3 | Heikkinen, Miss.  Laina | 26 | 0 | 0 | STON/O2. 3101282 | 7.9250 | S | 1 | 1 | 0 |
|  | 3 | 4 | 1 | 1 | Futrelle, Mrs.  Jacques Heath (Lily May Peel) | 35 | 1 | 0 | 113803 | 53.1000 | S | 1 | 1 | 0 |
|  | 4 | 5 | 0 | 3 | Allen, Mr. William Henry | 35 | 0 | 0 | 373450 | 8.0500 | S | 0 | 0 | 1 |
| **In [33]:** | **cat['Sex\_female'] # Female = 0** | | | | |  |  |  |  |  |  |  |  |  |
| **Out [33]:** | **0 0** | | | | |  |  |  |  |  |  |  |  |  |
| **1 1** | | | | | | | | | | | | | | |
| **2 1** | | | | | | | | | | | | | | |
| **3 1** | | | | | | | | | | | | | | |
| **4 0** | | | | | | | | | | | | | | |
| **..** | | | | | | | | | | | | | | |
| **886 0** | | | | | | | | | | | | | | |
| **887 1** | | | | | | | | | | | | | | |
| **888 1** | | | | | | | | | | | | | | |
| **889 0** | | | | | | | | | | | | | | |
| **890 0** | | | | | | | | | | | | | | |
| **Name: Sex\_female, Length: 889, dtype: uint8** | | | | | | | | | | | | | | |
| **In [34]:** | **cat['Sex\_male'] # Male = 1** | | | | |  |  |  |  |  |  |  |  |  |
| **Out [34]:** | **0 1** | | | | |  |  |  |  |  |  |  |  |  |
| **1 0** | | | | | | | | | | | | | | |
| **2 0** | | | | | | | | | | | | | | |
| **3 0** | | | | | | | | | | | | | | |
| **4 1** | | | | | | | | | | | | | | |
| **..** | | | | | | | | | | | | | | |
| **886 1** | | | | | | | | | | | | | | |
| **887 0** | | | | | | | | | | | | | | |
| **888 0** | | | | | | | | | | | | | | |
| **889 1** | | | | | | | | | | | | | | |
| **890 1** | | | | | | | | | | | | | | |
| **Name: Sex\_male, Length: 889, dtype: uint8** | | | | | | | | | | | | | | |
| --------- END --------- | | | | | | | | | | | | | | |