CAP- 5610 Machine Learning

Homework 1

Name: Nabila Shahnaz Khan

NID: 5067496

Data Preprocessing and Analyze by pivoting features of Titanic dataset

### Q1: In training set, which features are available?
Answer: In training set, the available features are: “PassengerId”, “Survived”, “Pclass”, “Name”, “Sex”, “Age”, “SibSp”, “Parch”, “Ticket”, “Fare”, “Cabin”, “Embarked”.

### Q2: In training set, which features are categorical?
Answer: The categorical features are: “Survived”, “Pclass”, “Sex”, “Embarked”.

### Q3: In training set, which features are numerical (e.g., discrete, continuous, or time series based)?
Answer: The numerical features are: “PassengerId”, “Age”, “SibSp”, “Parch”, “Fare”.

### Q4: In training set, which features are mixed data types?
Answer: The mixed data type is “Ticket”.

### Q5: In training set, which features contain blank, null or empty values? In test set, which features contain blank, null or empty values?
Answer: In training set, columns with missing values are: “Age”, “Cabin”, “Embarked”. In testing set, columns with missing values are: “Age”, “Fare”, “Cabin”.

### Q6: In training set, what are the data types (e.g., integer, floats or strings) for various features?
Answer: In training set, the data types for various features are:

Feature	**Data Type**
PassengerId	integer
Survived	integer
Pclass	integer
Name	string
Sex	string
Age	float
SibSp	integer
Parch	integer
Ticket	string
Fare	float
Cabin	string
Embarked	string

### Q7: In training set, to understand the distribution of numerical feature values across the samples, please list the properties, including count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max, of numerical features?

Answer: The properties of the numerical features have been listed below:

	PassengerId	**Age**	**SibSp**	**Parch**	**Fare**
count	891	714	891	891	891
mean	446	29.7	0.5	0.38	32.2
std	257.35	14.5	1.1	0.8	49.7
min	1	0.42	0	0	0
25% percentile	223.5	20.13	0	0	7.9
50% percentile	446	28	0	0	14.5
75% percentile	668	38	1	0	31
max	891	80	8	6	512.3

### Q8: In training set, to understand the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please list the properties, including count, unique, top, freq, of categorical features?

Answer: The properties of the categorical features have been shown below:

	Survived	**Pclass**	**Sex**	**Embarked**
count	891	891	891	889
unique	2	3	2	3
top	0	3	male	S
freq	549	491	577	644

### Q9: In training set, can you observe significant correlation (average survived ratio>0.5) among the group of Pclass=1 and Survived? If Pclass has significant correlation with Survived, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?

Answer: I can see a significant correlation between “Pclass = 1” and “Survived”. For “Pclass = 1”, survival rate is 0.63 (> 0.5) which is significantly high. So, “Pclass” has significant relation with “Survived”. Hence, we should include this feature in the predictive model.

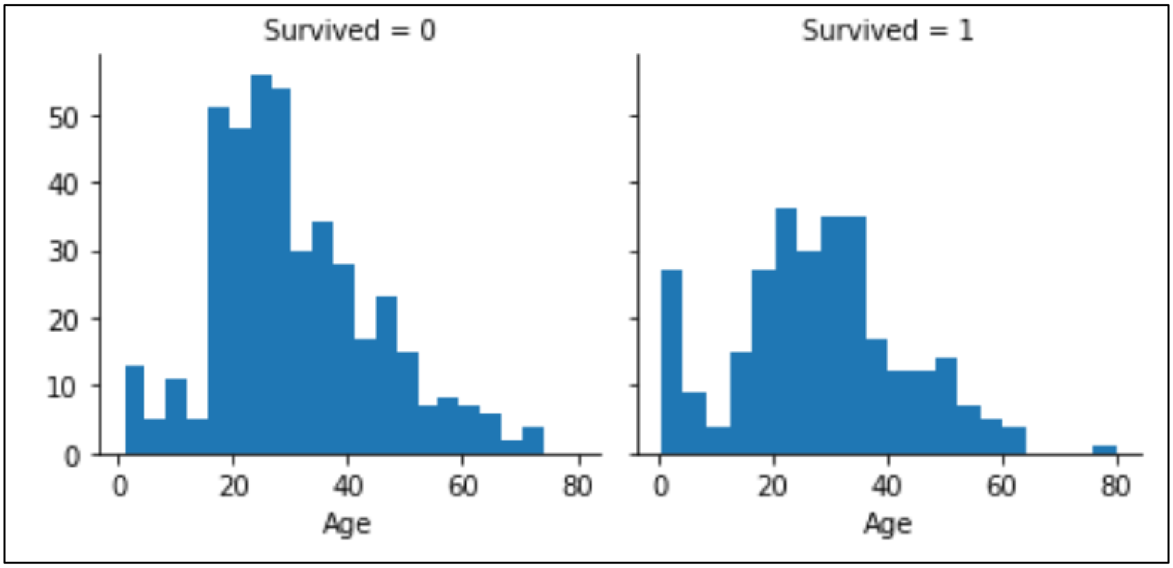
### Q10: In training set, are Women (Sex=female) were more likely to have survived?

Answer: In the training dataset, among 314 female, 233 female survived. So women were more likely to have survived with a survival rate of 74.2%.

### Q11: In training set, let us start by understanding correlations between a numeric feature (Age) and our predictive goal (Survived). A histogram chart is useful for analyzing continuous numerical variables like Age where banding or ranges will help identify useful patterns. The histogram can indicate distribution of samples using automatically defined bins or equally ranged bands. This helps us answer questions relating to specific bands (e.g., infants, old). Please plot the histograms between ages and Survived (Figure 1 is an example), and answer the following questions:

• Do infants (Age <=4) have high survival rate?
• Do oldest passengers (Age = 80) survive?
• Do large number of 15-25 year olds not survive?
Based on your analysis of the figures,
• Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)
• Should we should band age groups?

Answer: The histograms between "Age" and "Survived" using bin size 20 is given below:



- Infants (Age <=4) have high survival rate.
- Oldest passengers (Age = 80) survive.
- Large number of 15-25 year olds do not survive.

Based on my analysis of the figures,

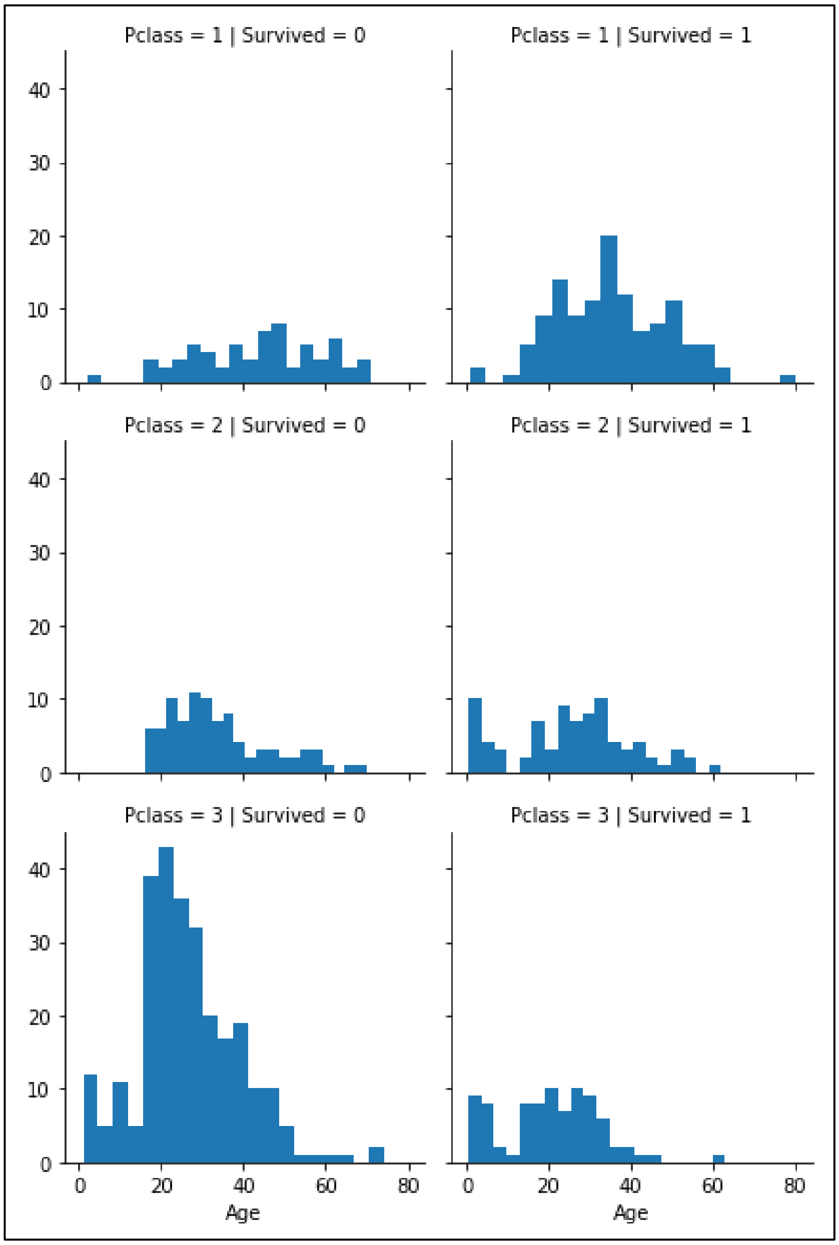
- We should consider Age in our model training (The Age feature for null values has been completed).
- We should band age groups.

### Q12: In training set, we can combine three features (age, Pclass, and survivied) for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values.

Please plot the plot using python, and answer the following questions:

- Does Pclass=3 have most passengers, however most did not survive?
- Do infant passengers in Pclass=2 and Pclass=3 mostly survive?
- Do most passengers in Pclass=1 survive?
- Does Pclass vary in terms of Age distribution of passengers?
- Should we consider Pclass for model training?

Answer: Plots combining three features (age, Pclass, and survivied) are given below:



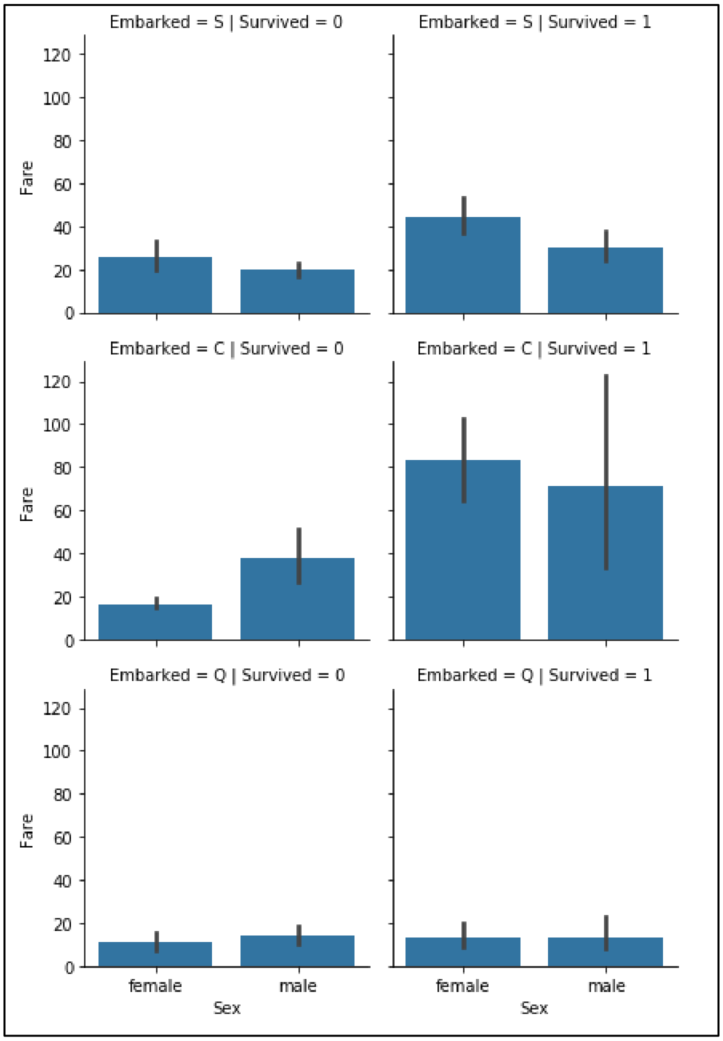
Based on the plots, the answers to the given questions are:

- It is quite evident from the plots that “Pclass = 3” have most passengers, and most of them did not survive.
- For “Pclass=2”, all the infants (Age <= 4) survived. In case of “Plass=3”, the difference between the number of surviving and not surviving infants is not that high. Still, number of survival is lower for “Pclass=3”.
- Yes, most passengers in “Pclass=1” survive.
- Yes, “Pclass” vary in terms of “Age” distribution of passengers.
- We should consider “Pclass” for model training.

### Q13: In training set, we want to correlate categorical features (with non-numeric values) and numeric features. We can consider correlating Embarked (Categorical non-numeric), Sex (Categorical non-numeric), Fare (Numeric continuous), with Survived (Categorical numeric). Please plot a figure to illustrate the correlations of Embarked, Sex, Fare, and Survivied. Here is a sample plot (Figure 3):

- Do higher fare-paying passengers have better survival?
- Should we consider the banding fare feature?

Answer: A plot to illustrate the correlations of Embarked, Sex, Fare, and Survivied is given below:



- Yes, higher fare-paying passengers have better survival.
- Yes, we should consider the banding fare feature.

### Q14: In training set, what is the rate of duplicates for the Ticket feature? Is there a correlation between Ticket and survival? Should we drop the Ticket feature?

Answer: “Ticket” features has total 891 values. Among them 681 are unique features. So, the rate of duplication is 23.6%. No clear correlation is visible between “Ticket” and “Survived” features. So we should drop “Ticket” feature.

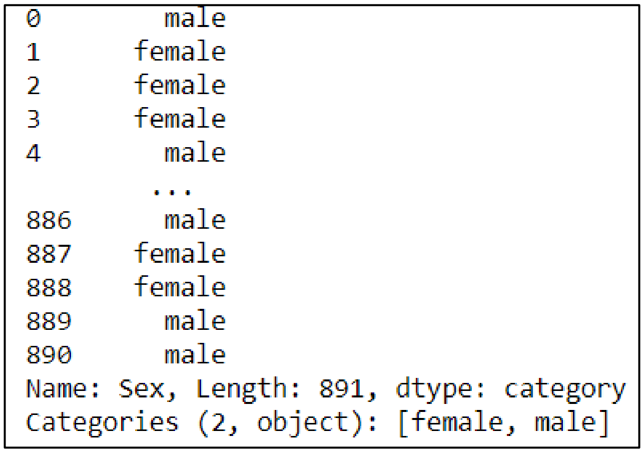
### Q15: In the training set, Is the Cabin feature complete? How many null values there are in the Cabin features of the combined dataset of training and test dataset? Should we drop the Cabin feature?

Answer: In the training set, “Cabin” feature is not complete, there are still 687 null values. There are 327 null values of “Cabin” feature in the test dataset. In total there are 1014 null values. So we should drop “Cabin” feature.

### Q16: In the training set, we can convert features which contain strings to numerical values. This is required by most model algorithms. Doing so will also help us in achieving the feature completing goal. In this question ,please convert Sex feature to a new feature called Gender where female=1 and male=0.

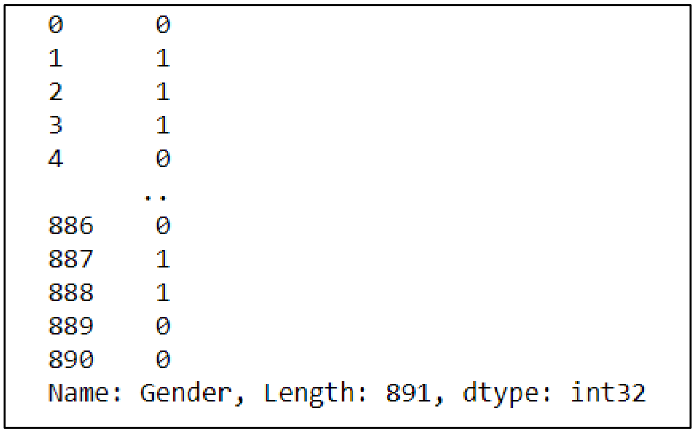
Answer: Categorical feature “Sex” has been converted into a numerical feature “Gender” where female=1 and male=0.

Categorical feature ‘Sex’



```
0        male
1      female
2      female
3      female
4        male
        ...
886      male
887    female
888    female
889      male
890      male
Name: Sex, Length: 891, dtype: category
Categories (2, object): [female, male]
```

Numerical feature ‘Gender’

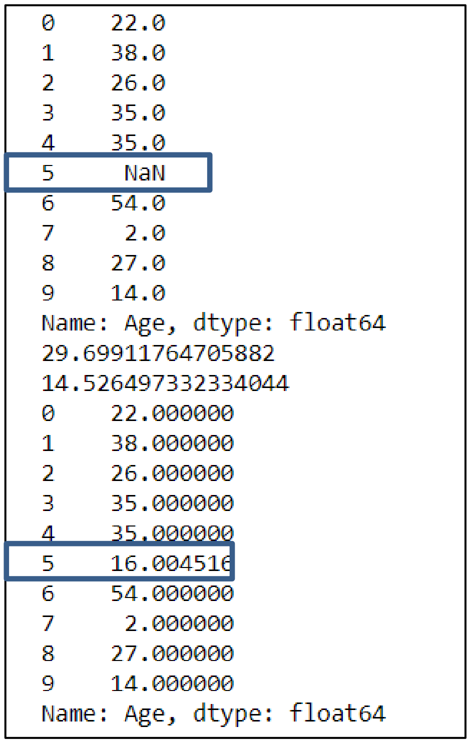


```
0      0
1      1
2      1
3      1
4      0
      ..
886    0
887    1
888    1
889    0
890    0
Name: Gender, Length: 891, dtype: int32
```

### Q17: In the training set, we start estimating and completing features with missing or null values. We will first do this for the Age feature. We can consider three methods to complete a numerical continuous feature. A simple way is to generate random numbers between mean and standard deviation. More accurate way of guessing missing values is to use the K-Nearest Neighbor algorithm to select the top-K most similar data points, and then use the top-K most similar data points to impute the missing values of ages.

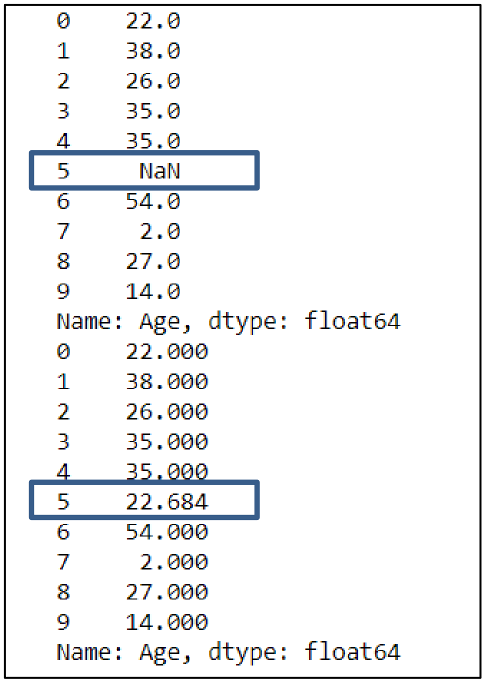
Answer: Here, the missing values of "Age" has been estimated using the following methods (finally used the 2nd method):

1. Generating random numbers between mean and standard deviation



```
0     22.0
1     38.0
2     26.0
3     35.0
4     35.0
5      NaN
6     54.0
7      2.0
8     27.0
9     14.0
Name: Age, dtype: float64
29.69911764705882
14.526497332334044
0     22.000000
1     38.000000
2     26.000000
3     35.000000
4     35.000000
5     16.004516
6     54.000000
7      2.000000
8     27.000000
9     14.000000
Name: Age, dtype: float64
```

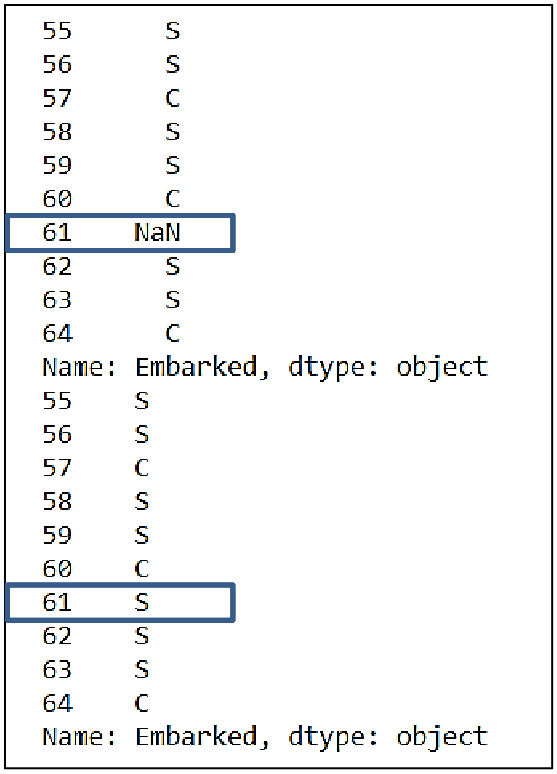
2. Using the K-Nearest Neighbor algorithm (K = 5)



```
0     22.0
1     38.0
2     26.0
3     35.0
4     35.0
5      NaN
6     54.0
7      2.0
8     27.0
9     14.0
Name: Age, dtype: float64
0     22.000
1     38.000
2     26.000
3     35.000
4     35.000
5     22.684
6     54.000
7      2.000
8     27.000
9     14.000
Name: Age, dtype: float64
```

### Q18: In the training set, complete a categorical feature: Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has some missing values. Please simply fill these with the most common occurrences.

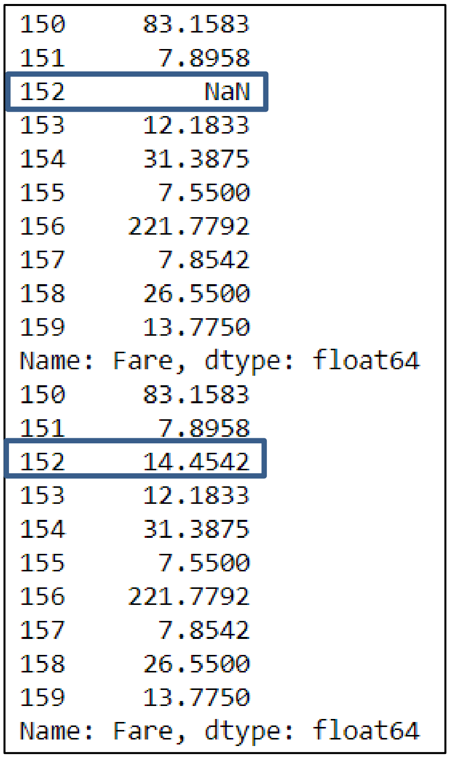
Answer: Categorical feature “Embarked” has been completed with the most commonly occurring value ‘S’.



```
55      S
56      S
57      C
58      S
59      S
60      C
61    NaN
62      S
63      S
64      C
Name: Embarked, dtype: object
55    S
56    S
57    C
58    S
59    S
60    C
61    S
62    S
63    S
64    C
Name: Embarked, dtype: object
```

### Q19: In the training set, complete and convert a numeric feature. Please complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature.

Answer: Here, “Fare” feature have been completed for single missing value in both test and train dataset using mode to get the value that occurs most frequently for this feature.

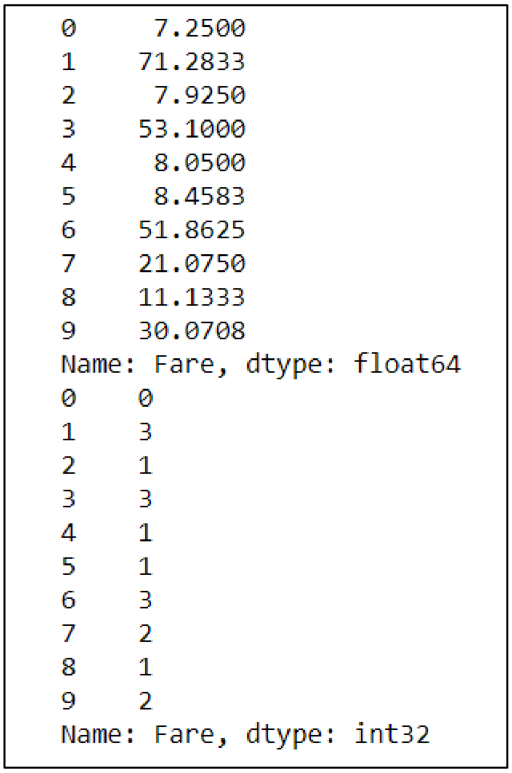


```
150      83.1583
151       7.8958
152          NaN
153      12.1833
154      31.3875
155       7.5500
156     221.7792
157       7.8542
158      26.5500
159      13.7750
Name: Fare, dtype: float64
150      83.1583
151       7.8958
152      14.4542
153      12.1833
154      31.3875
155       7.5500
156     221.7792
157       7.8542
158      26.5500
159      13.7750
Name: Fare, dtype: float64
```

### Q20: In the training set, convert the Fare feature to ordinal values based on the FareBand defined as follows:

Ordinal Fare Indicator	FareBand	Survivied
0	(-0.001, 7.91]	0.197309
1	(7.91, 14.454]	0.303571
2	(14.454, 31.0]	0.454955
3	(31.0, 512.329]	0.581081

Answer: Fare feature has been converted to ordinal values based on the FareBand defined in the given table.



```
0     7.2500
1    71.2833
2     7.9250
3    53.1000
4     8.0500
5     8.4583
6    51.8625
7    21.0750
8    11.1333
9    30.0708
Name: Fare, dtype: float64
0    0
1    3
2    1
3    3
4    1
5    1
6    3
7    2
8    1
9    2
Name: Fare, dtype: int32
```

Code Link: https://github.com/NabilaKhan/CAP-5610-Machine-Learning-/blob/main/CAP-5610-HW1.ipynb

[Please let me know if you are having any issue to find the code]