1. In the training set, the following features are available:
    - *PassengerID*
    - *Survived*
    - *Pclass*
    - *Sex*
    - *Age*
    - *Sibsp*
    - *Parch*
    - *Ticket*
    - *Fare*
    - *Cabin*
    - *Embarked*

2. In the training set, the following features are categorical:
    - *Sex*
    - *Embarked*

3. In the training set, the following features are numerical:
    - *PassengerID*
    - *Survived*
    - *Pclass*
    - *Age*
    - *Sibsp*
    - *Parch*

4. In the training set, the following features are mixed data types:
    - *Ticket*

5. In the training set, the following features contain blank, null, and/or empty values:
    - *Age*
    - *Cabin*
    - *Embarked*

6. In the training set, the data types for the features are as follows:
    - *PassengerID* = Int
    - *Survived* = Int
    - *Pclass* = Int
    - *Sex* = String
    - *Age* = Float
    - *Sibsp* = Int
    - *Parch* = Int
    - *Ticket* = String
    - *Fare* = Float

- *Cabin* = String
- *Embarked* = String

7. The distribution of the features in the training dataset is shown below in Fig. 1:

| | 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 |
| **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 |

***Figure 1: Distribution of Features in Training Dataset***

8. The feature information for the categorical variables is shown below:
   - *Sex*

```
count      891
unique       2
top       male
freq       577
Name: Sex, dtype: object
```

***Figure 2: Feature Information for Variable* Sex**

   - *Embarked*

```
count     889
unique      3
top         S
freq      644
Name: Embarked, dtype: object
```

***Figure 3: Feature Information for Variable* Embarked**

9. Yes, there is a correlation between *Pclass* and *Survived*. Thus, I will use this feature in my predictive model.

10. In the training set, **women** were more likely to have survived.

11. Fig. 4 is shown below:

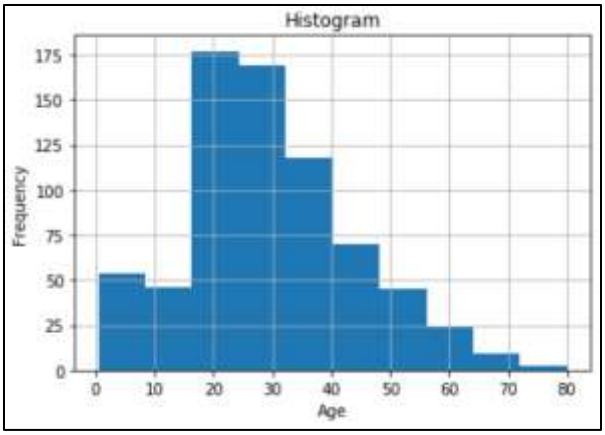


***Figure 4: Histogram for Feature* Age**

- Infants do have a relatively high survival rate.
- The oldest passengers do survive.
- Large number of 15 – 25 yr. old's do not survive.

We should consider Age in our model training, and we should band the age groups – they help to visualize the data distribution.

12. Based upon the plots below, the following answers were obtained:
    - Most passengers in *Pclass* = 3 did not survive.
    - Most infant passengers in *Pclass* = 2 and *Pclass* = 3 did survive.
    - Most passengers in *Pclass* = 1 did survive.
    - *Pclass* does vary in *Age* distribution of passengers.
    - We should consider *Pclass* in model training.

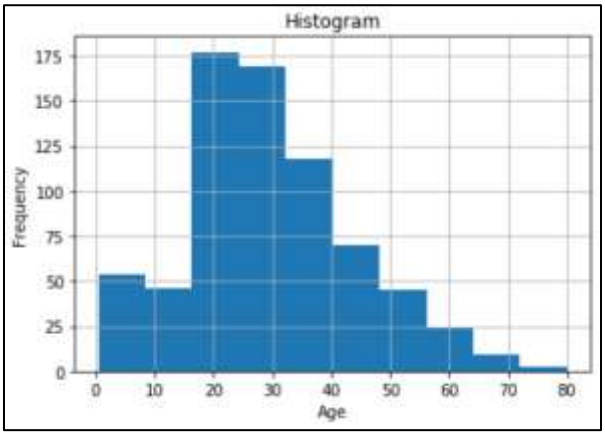


***Figure 5: Histogram for Feature* Age**

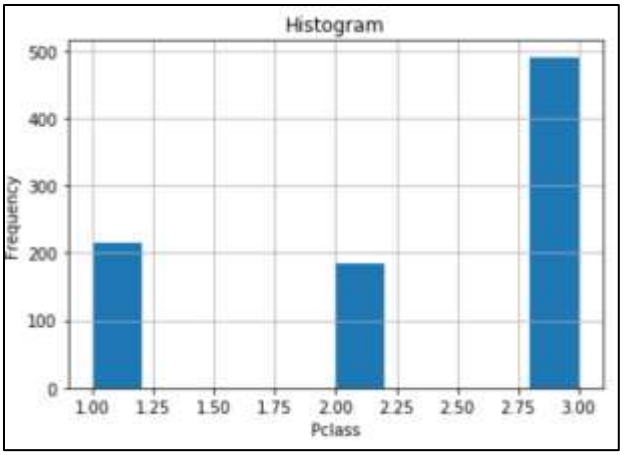


***Figure 6: Histogram for Feature* Pclass**

13. The answers are below:
    - Higher fare-paying passengers did have better survival.
    - We should band fare.

14. There were 210 duplicate values within the *Ticket* feature. However, there is no correlation between *Ticket* and survival; thus, we should drop the *Ticket* feature.

15. No, the Cabin feature is not complete. There are 688 null values in the *Cabin* feature for the training set, and 328 null values in the *Cabin* feature for the test set. Thus, we should drop the *Cabin* feature.

16. I added the *Gender* feature to represent the *Sex* feature in a numerical format.

17. In the training set, I used K-Nearest Neighbors as a predictive model.

18. I filled the *Embarked* feature in the training set with the **S** value (it was the most common value).

19. I filled the *Fare* feature in the testing set with the **35.627188** value (it was the most common value).