1. PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
2. Survived, Pclass, Sex, Embarked
3. Age, parch , SibSp, fare
4. ticket
5. Age, Cabin, Embarked
6. In order of column of training data:

    integer

    integer

    string

    string

    float

    integer

    integer

    string

    float

    string

    string

7.

```
Age:   714, 29.699, 14.526, 0.42, 20.125, 28, 38, 80
SibSp: 891, 0.523, 1.103, 0, 0, 0, 1, 8
Parch: 891, 0.382, 0.806, 0, 0, 0, 0, 6
Fare:  891, 32.204, 49.693, 0, 7.910, 14.454, 31, 512.329
```

8.

```
Survived:     891, 2, 0, 549
Pclass:       891, 3, 3, 491
Sex:          891, 2, male, 577
Embarked:     891, 3, S, 644
```

9. my results were:

```
first class:  136/216=0.629630
second class:  87/184=0.472826
third class:  119/491=0.242363
```

therefore, the category should be added since it seems that there is a relationship between passenger class and survival

10. overwhelmingly yes

```
male:  109/577=0.188908
female:  233/314=0.742038
```

11.

a. Infants have a high survival rate

b. The graph shows that 80 year olds survived

c. 15-25 year olds are the largest range of people that did not survive

Age seems to be an important factor and should be completed.

Banding seems like a worthwhile endeavor since there are about 900 entries It'll be clearer with less bars on the graph

12.

a. Pclass 3 had the most passengers and had the least survive

b. The majority of Pclass 2 and 3 infants do survive

c. more Pclass 1 passengers survived than did not

d. The higher the Pclass, the lower the avg mean age

e. Pclass was very relevant to the training data

13.

a. For the majority of the data, it seems that fare does not affect the survival rate that much

b. Banding fare might be useful on a graph with less comparisons. This graph makes it hard to band data. However it would be easier on a less active graph.

14.

```
rate of duplicates:  0.235690
```

Since everyone had a ticket and I am not sure what kind of corollary you would find since over 75% of the tickets are unique, I would say there is none. Therefore, we do not need the Ticket feature

15.

Cabin is not complete.

There are about 1000 missing cabin values. Therefore, there is so little information for this feature that it is not worth keeping.

16 - 20 will be submitted with the other code I used in a notebook (.ipynb file)

https://github.com/JoshuaFoldes2/CAP5610-Homework