**Sayma Sultana**

**Id: 5068163**

**Q9:**

|  |  |
| --- | --- |
| Pclass | Survival ratio |
| 1st class | 0.629630 |
| 2nd class | 0.472826 |
| 3rd class | 0.242363 |

So, people from 1st class have the highest survival ratio and passenger class is an important feature, so I will include this feature in the model.

**Q10:**

|  |  |
| --- | --- |
| Sex | Survival ratio |
| Female | 0.742038 |
| Male | 0.188908 |

So, female have a very higher survival ratio than male. They are more likely to have survived.

**Q11.**

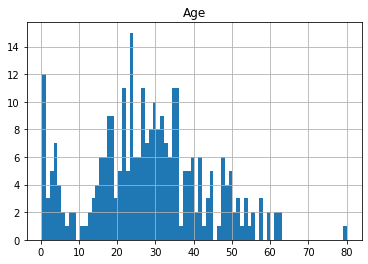


Figure : Histogram of survived = 1

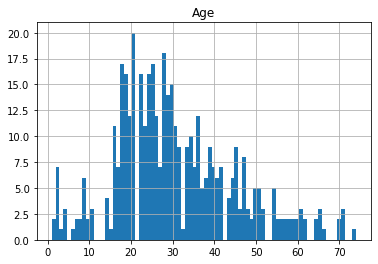


Figure : Histogram of survived = 0

From the figures above,

a. Infants of age <= 4 have the second highest survival rate.

b. People of age 80 survived.

c. People of age 15-20 have a high fatality rate, so most of them did not survive.

d. yes, we should include age in our model.

e. yes. We should band age group.

**12.**

Histograms for Survived = 1, Pclass = 1 and Survived = 0 and Pclass = 1 are given below respectively:

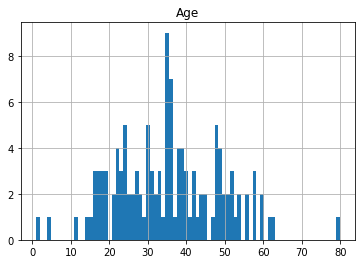


Figure : Histogram for survived = 1 and plcass = 1

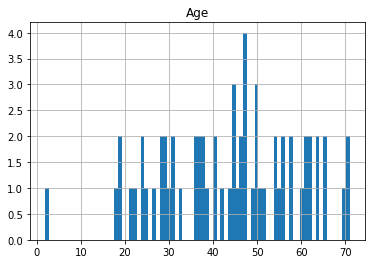


Figure : Histogram for Survived = 0 and plcass = 1

Histograms for Survived = 1, Pclass = 2 and Survived = 0, Pclass = 2 are given below respectively:

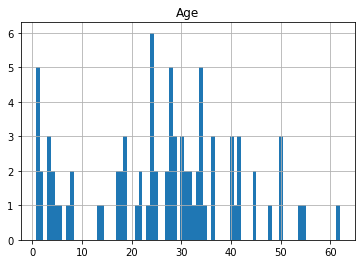


Figure : Histogram for survived = 1 and plcass = 2

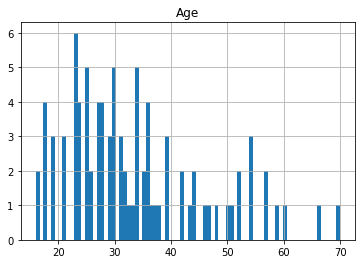


Figure : Histogram for survived = 0 and pclass = 2

Histograms for Survived = 1, Pclass = 3 and Survived = 0, Pclass = 3 are given below respectively:

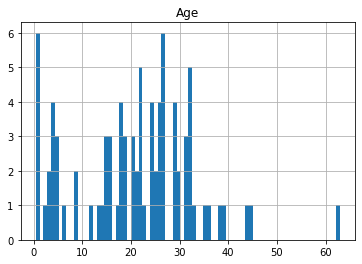


Figure : HIstogram for survived = 1 and pclass = 3

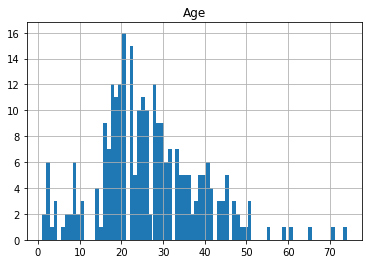
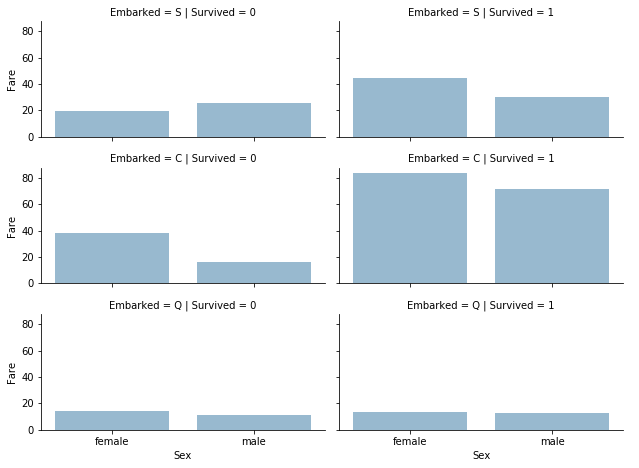


Figure : Histogram for survived = 0 and pclass = 3

1. Yes. Pclass = 3 have most passengers and they have very high fatality . So most them did not survived.
2. Yes, infant passengers in pclass = 2 and 3 have high survival rate, so most of them survived.
3. Yes, most passengers in pclass = 1 survived.
4. Yes, pclass vary in terms of age.
5. Yes, we should consider pclass for model training.

**13.**



1. Yes, higher paying passengers survived more.
2. No, port of embarkation is not related to survival rate because ratios of survival and fatality are almost same for three cases.
3. Yes. We should consider banding fare feature.

**14.**

a. Ticket has unique 681 values. Duplicate rate is 0.2356902356902357.

b. No, there is no relation between survival and ticket.

c. Yes, we should drop the ticket feature.

**15.**

a. No. Cabin feature is not complete.

b. There are 1014 null values in cabin column.

c. Yes, we should drop the cabin column.

**16.** Please follow the link.

**17.** Please follow the link.

**18.** Please follow the link.

**19.** Please follow the link.

**20.** Please follow the link.

**Source code:** <https://github.com/Sayma23/ML_Spring2020/blob/master/temp.py>