Spark – LW - 3 RMS Titanic

Data Description:
Passenger ID
Survived: Weather Survived or not: 0 = No, 1 = Yes
Pclass: Ticket class: 1 = 1st, 2 = 2nd, 3 = 3rd
Name: Name of the Passenger
Sex: Gender
Age: Age in Years
SibSp: No. of siblings / spouses aboard the Titanic
Parch: No. of parents / children aboard the Titanic
Ticket: Ticket number
Fare: Passenger fare
Cabin: Cabin number
EmbarkedPort of Embarkation:C = Cherbourg, Q = Queenstown, S = Southampton

Answer the following questions using the Spark DataFrames API.

1. What is the mean of ticket fare?
2. Provide the six point summary of age based on the survivability.
3. What is the rate of survival of passengers, if they have siblings vs not having siblings.
4. hat is the probability of survival based on the gender.
5. Based on the age, which group managed to survive more relativly?
   - You can take 10 years per age group. Say 0 - 10; 11-20 so on
   - Do you think that females outlived male all the age group. Enumerate your learnings.-
6. What is the average survival rate based on the Embarked City?
7. *"A passenger from first class is more likely to sucummb then the passenger from 3rd Class"*

   Prove or disprove the hypothesis with the data and 95% of confidence.
8. Which passenger group has the highest survival rate based on the age group, gender, class and boarding city? Find the least survival group as well.
9. How are you planning to handle missing data.
10. ***Please specify the number of actions and transformations used along with your approach to the question***
11. *Assume you are asked to solve the same set of questions in hadoop compare your experience with Spark.*