For code, please visit: https://github.com/lumalav/CAP5610/blob/master/HW1/HW1.ipynb

Q1: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

Q2: Survived, Pclass, Name, Sex, SibSp, Parch, Ticket, Cabin, Embarked

Q3: PassengerId, Age, SibSp, Parch, Fare

Q4: Age and Fare: sometimes they use complete values as integers and other times fractioned values as floats, but at the end everything is casted to a float

Ticket: sometimes it uses numerical values as integers and other times combinations of numbers and letters as strings, but at the end everything is casted to a string

Q5: Train Set: Age, Cabin, Embarked | Test Set: Age, Fair, Cabin (See Section #Q5 of HW1.ipynb)

Q6:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Integer	Integer	Integer	string	String	Float	integer	integer	String	Float	String	String

Q7:

	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

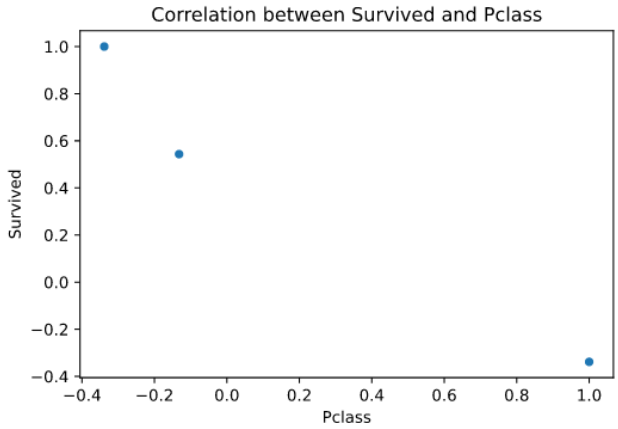
(See Section #Q7 of HW1.ipynb)

Q8:

	Survived	Pclass	Sex	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	891.0	891.0	891	891.0	891.0	891	891.00	204	889
unique	2.0	3.0	2	7.0	7.0	681	248.00	147	3
top	0.0	3.0	male	0.0	0.0	1601	8.05	B96 B98	S
freq	549.0	491.0	577	608.0	678.0	7	43.00	4	644

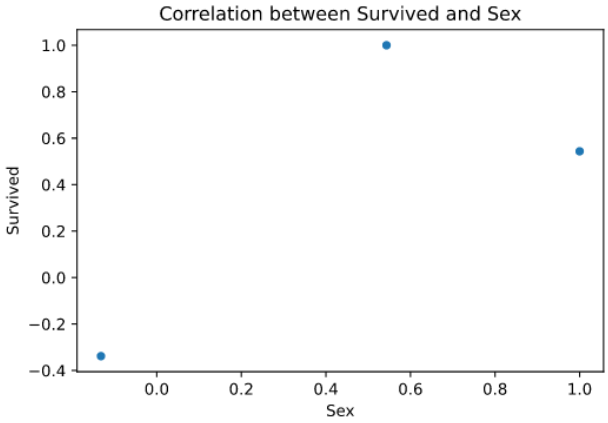
(See Section #Q8 of HW1.ipynb)

Q9: Yes. The following figure displays high survival rate among the most privileged classes. I would choose this feature to be part of the predictive model based on this high correlation.



(See Section #Q9 of HW1.ipynb)

Q10: Yes. For the following chart, I converted the Sex feature to be a numerical value (female = 1, male = 0) and calculated the correlation with the Survived feature. As we can see, there's higher chance for survival among women than for men.



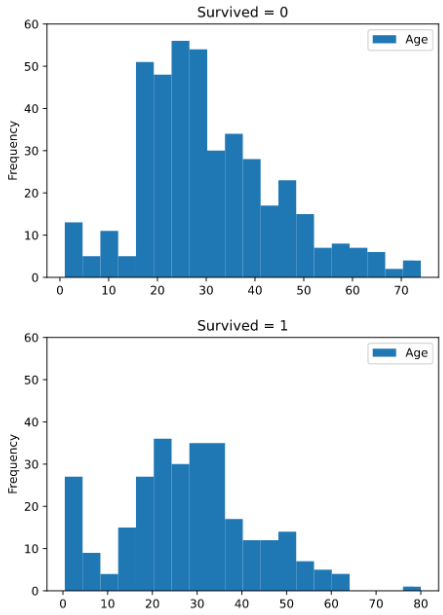
(See Section #Q10 of HW1.ipynb)

Q11: a) Yes. Infants age <= 4 have a high survival rate (>0.5)

b) Yes. One 80-year-old passenger survived.

c) Yes. Death rate is high among passengers of ages between 15 and 25

d) Yes. Age could be a useful feature when used with other ones like Pclass. I think we should consider it adding to the model after properly replacing the null values.



(For code, see Section #Q11 of HW1.ipynb)

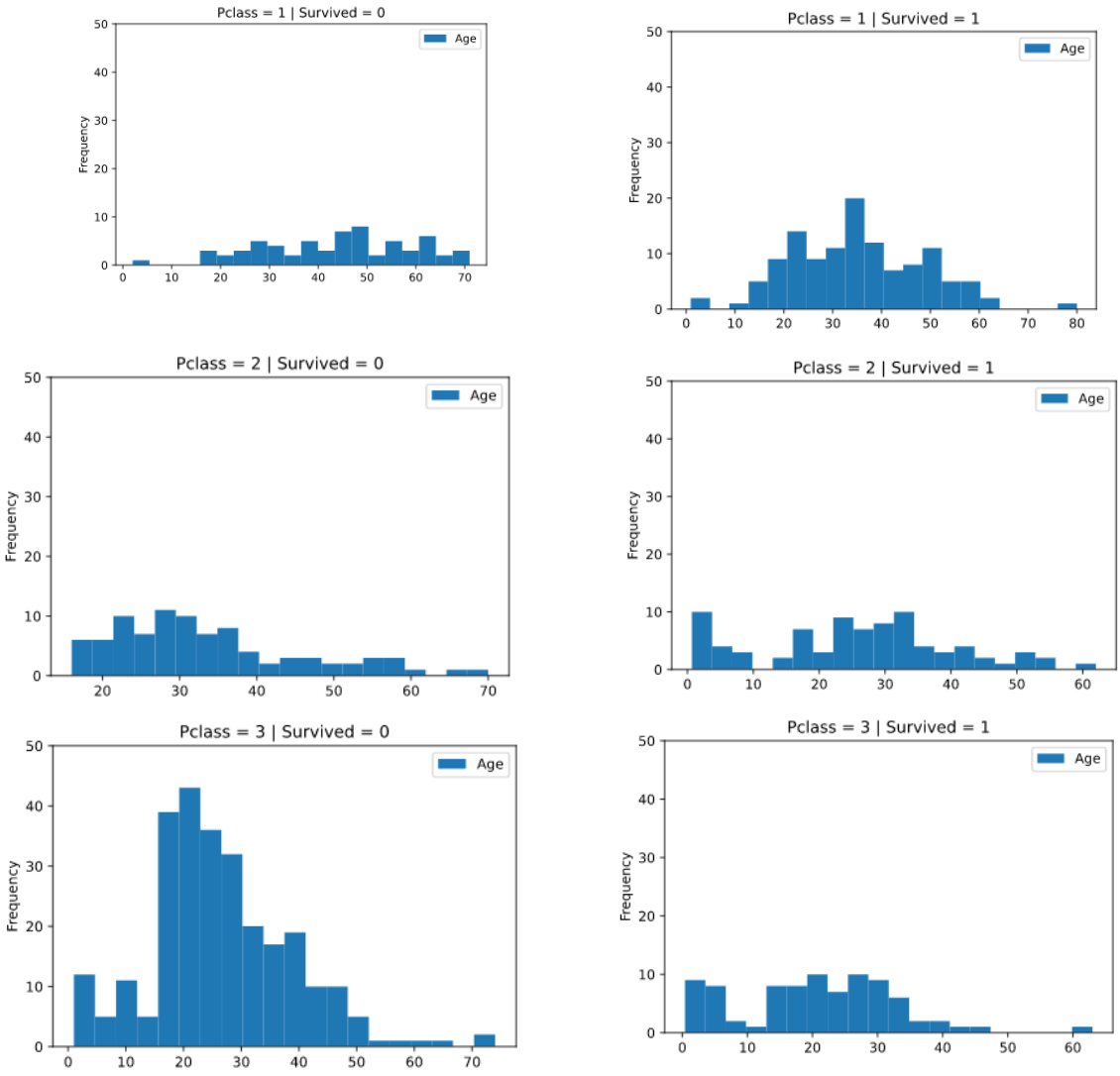
Q12: a) Yes. Most of the passengers in Pclass=3 did not survive. It was also the class with more people.

b) No, for Pclass=2 most of the infants survived. However, for Pclass=3 more infants died than those that survived.

c) Yes. Most of the passengers in Pclass=1 survived.

d) Yes. Pclass=3 had a lot of younger people (20-40 years-old) and their death rate was significantly higher than other classes.

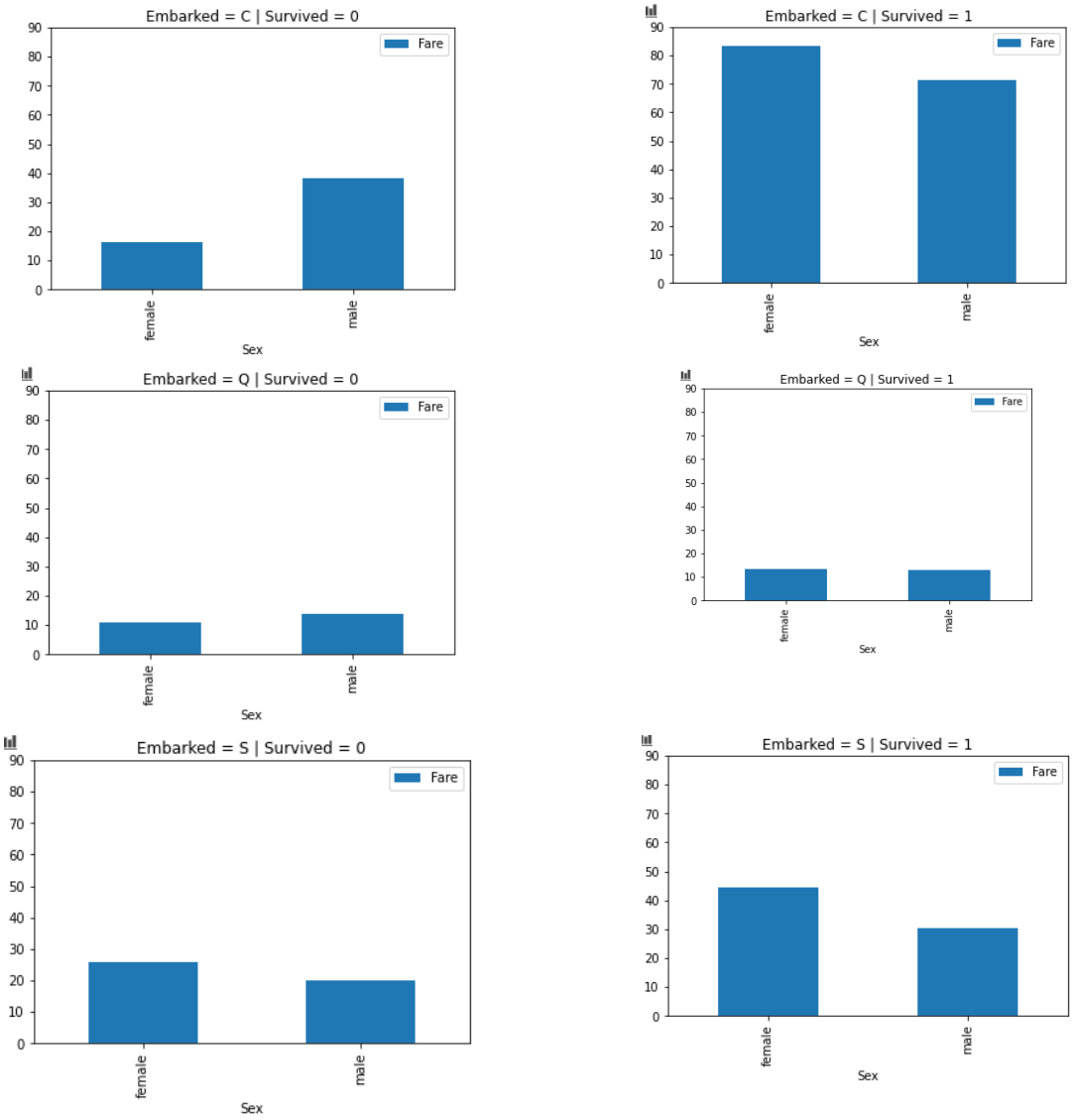
e) Yes. Pclass and Survival rate are highly correlated. I would add it for model training.



(For code, see Section #Q12 of HW1.ipynb)

Q13: a) Yes. People who paid more for their ticket had better chance of survival for embarkation ports C, and S. On the other hand, for port Q, there was not significant difference between the survival rate and fare.

b) I think it could be beneficial banding the fare into a few buckets to make the algorithm more efficient. The same principle is applied to the Age feature on the last question.



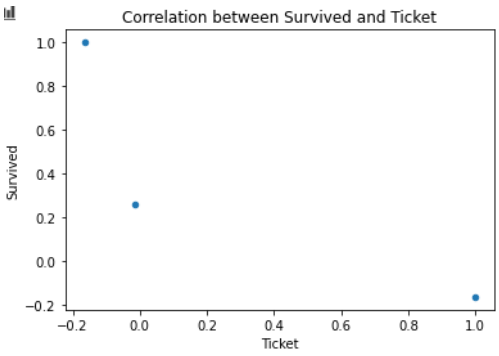
(For code, see Section #Q13 of HW1.ipynb)

Q14: Unique tickets: 681

Tickets duplicated: 134

Rate of duplicates: 0.19676945668135096

There is a correlation between ticket price (Fare) and survival. Consequently, there exist same correlation between ticket a survival. I would probably drop the ticket feature after some analysis. However, I think some interesting charts could be made based on some of those duplicates.



(For code, see Section #Q14 of HW1.ipynb)

Q15: The cabin feature is not complete. There are a total of 1014 missing values for Cabin across both datasets. I think we should drop the Cabin feature.

```
unique cabins: 186
total records: 1309
missing cabins: 1014
missing cabins ratio: 0.774637127578304
```

(For code, see Section #Q15 of HW1.ipynb)

Q16: Exactly to what I did in Q10, please see Section #Q16 of HW1.ipynb

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	P
0	1	0.0	3	Braund, Mr. Owen Harris	0	22.0	1	
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	
2	3	1.0	3	Heikkinen, Miss. Laina	1	26.0	0	
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	
4	5	0.0	3	Allen, Mr. William Henry	0	35.0	0	

Q17: Please see section #Q17 of HW1.ipynb

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Par
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	
4	5	0	3	Allen, Mr. William Henry	male	35	0	

Q18:

```
Most common port of embarkation: S, with: 644 occurrences
Number of records with null Embarked values: 2
New number of records with null Embarked values: 0
```

Q19:

```
Most common Fare: 8.05, with: 60 occurrences
Number of records with null Fare values: 1
New number of records with null Fare values: 0
```

Q20:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	0	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	3	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	1	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	3	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	1	Na

Please see Section #Q18, #Q19, and #Q20 of HW1.ipynb