Ben Simpson

CS5402

Homework 1

https://github.com/bmsr56/cs5402/blob/master/hw1/code/hw1.py

Task 1

a. Discrete, ordinal
b. Continuous, ratio
c. Discrete, ratio
d. Discrete, nominal
e. Continuous, ratio
f. Discrete, ordinal
g. Continuous, interval

Task 2

a. Euclidean distance would be better to use for grouping the boxes based on length to width ratio. This is because the square root of the squared difference of length and width values (the definition of Euclidean distance applied to this case) would always result in 0 if the box was square, yet would be greater as the ratio of length to the width (or vice-versa) increased.

b. Correlation

Task 3

1. Passenger ID, ticket class, sex, name, age, number of siblings and spouses, number of parents and children, ticket number, fare cost, cabin number, port of embarkation
2. Sex, name, cabin number, port of embarkation, passenger ID, ticket class
3. Age, number of siblings and spouses, number of parents and children, fare cost
4. Ticket number
5. Age, cabin number
6.

Passenger ID	Integer
Ticket class	Integer
Sex	String
Name	String
Age	Decimal
# siblings and spouses	Integer
# of parents and children	Integer
Ticket number	String
Fare cost	Decimal
Cabin number	string
Port of embarkation	string

7.

	Age	Fare	Parch	PassengerId	Pclass	\
count	1046.000000	1308.000000	1309.000000	1309.000000	1309.000000	
mean	29.881138	33.295479	0.385027	655.000000	2.294882	
std	14.413493	51.758668	0.865560	378.020061	0.837836	
min	0.170000	0.000000	0.000000	1.000000	1.000000	
25%	21.000000	7.895800	0.000000	328.000000	2.000000	
50%	28.000000	14.454200	0.000000	655.000000	3.000000	
75%	39.000000	31.275000	0.000000	982.000000	3.000000	
max	80.000000	512.329200	9.000000	1309.000000	3.000000	

	SibSp	Survived
count	1309.000000	891.000000
mean	0.498854	0.383838
std	1.041658	0.486592
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	1.000000	1.000000
max	8.000000	1.000000

8.

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Johansson, Mr. Erik	male	1601	C23 C25 C27	S
freq	1	577	7	4	644

9.

	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

The correlation between Pclass 1 and survival rate is < 50%, so I will not include it in the predictive model.

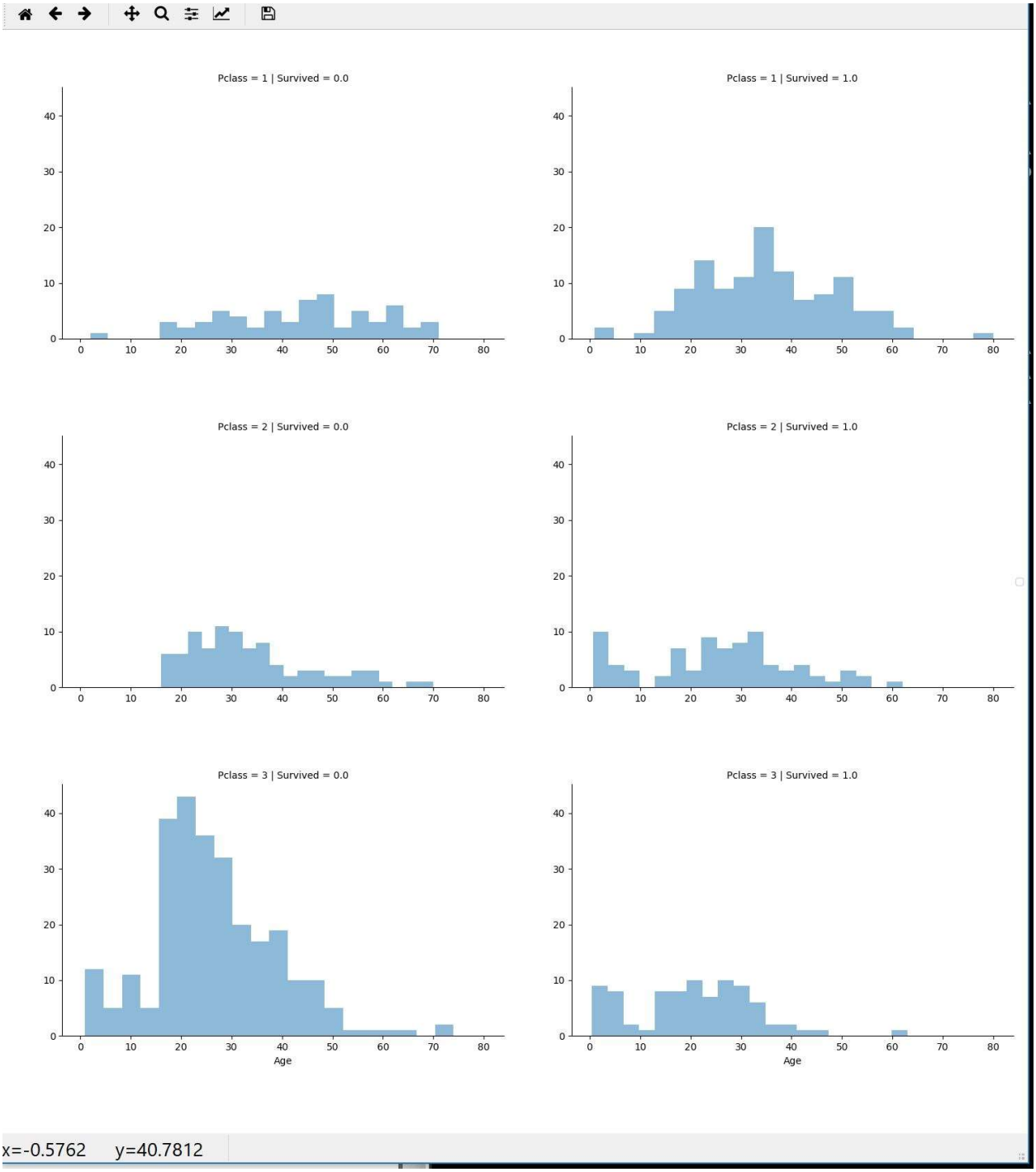
10.

	Sex	Survived
0	female	0.742038
1	male	0.188908

Yes, women are more likely to have survived.

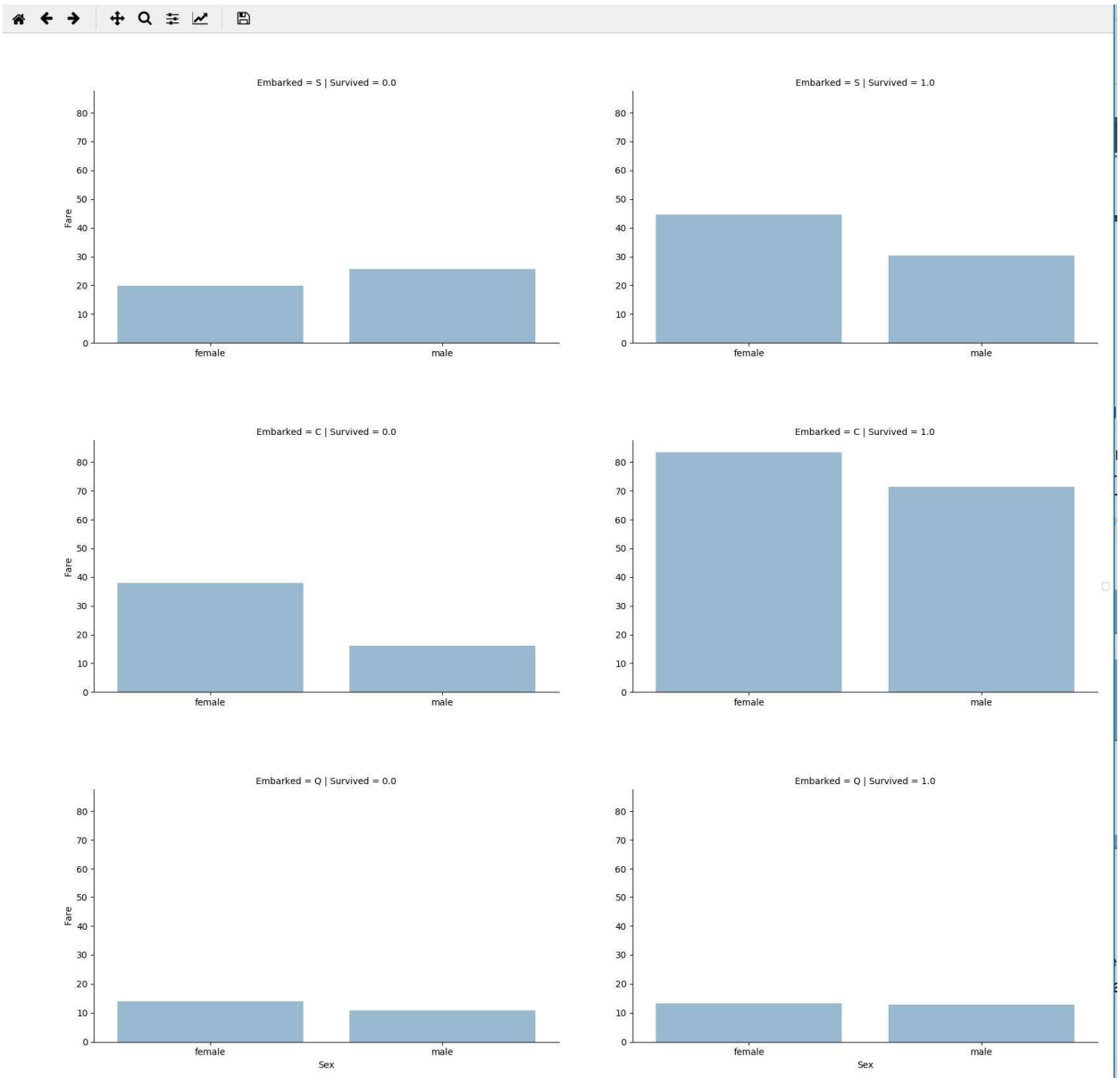
11.

a. Relatively, yes they did. They had the highest survival rate of anyone under 15.
b. Yes
c. Yes
d. Yes
e. Yes



12.

a. Yes, Pclass = 3 had the most passengers and most did not survive.
b. All infants survived in Pclass = 2. However, about the survival rate in Pclass = 3 was only about ~50%.
c. Yes.
d. Yes. Pclass = 1 seems to have mostly middle aged people, whereas the others tend to have slightly younger people.
e. Yes.



13.

a. For Embarked = {S, C}, yes they do. For Embarked = Q, no they do not.
b. Yes, point of embarkation correlates to survival rates.
c. Yes

14. From the categorical uniqueness analysis:
    Ticket Number
    unique values: 929
    total count: 1309

    rate of duplicates = (1309 – 929) / 1309 = ~29%

    Ticket feature is not related to survival rate and there are many duplicates, so we should drop it.

15. No, it is not complete. ~77% percent (1,014) of them are missing. Regardless of whether it might produce a correlation, there is too much missing data so we should drop this feature.

16. See code
17. See code
18. See code
19. See code
20. See code