Ben Simpson

CS5402

Homework 1

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

Task 1

1. Discrete, ordinal
2. Continuous, ratio
3. Discrete, ratio
4. Discrete, nominal
5. Continuous, ratio
6. Discrete, ordinal
7. Continuous, interval

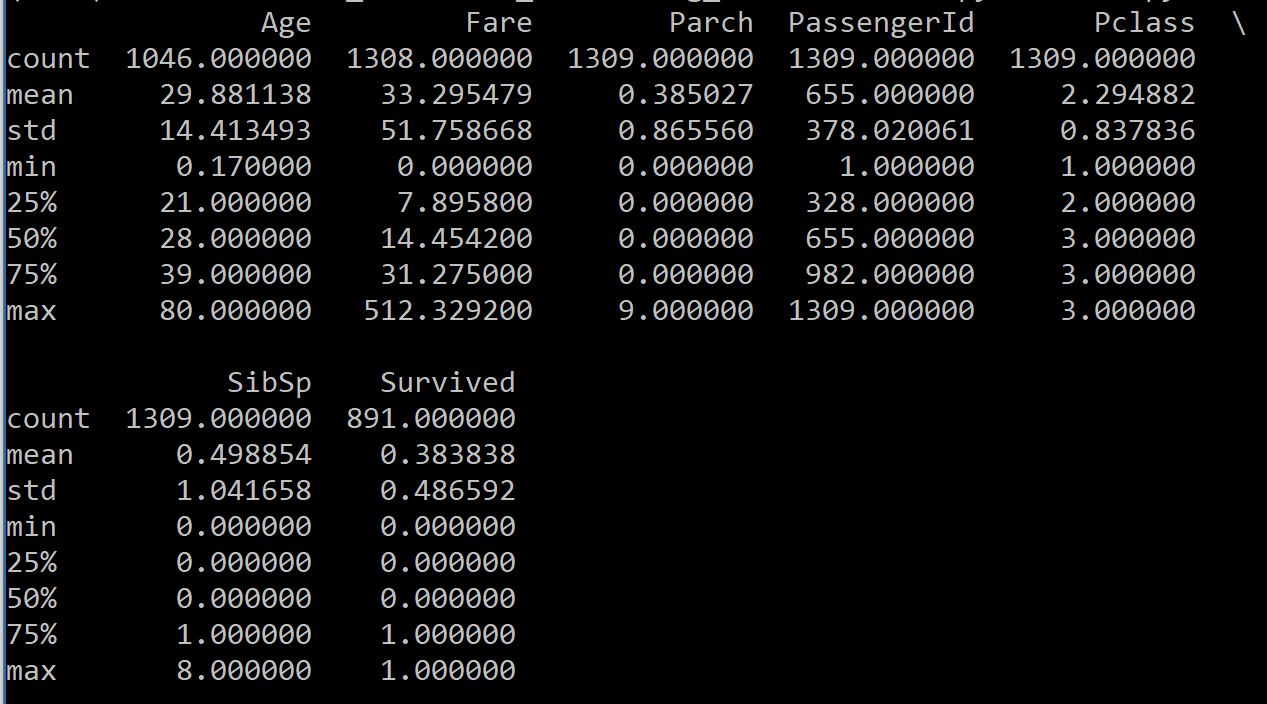
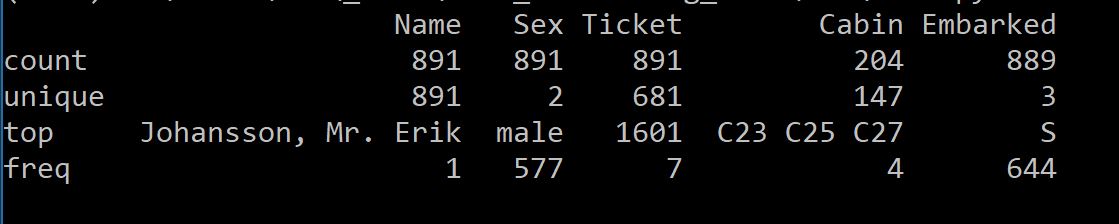
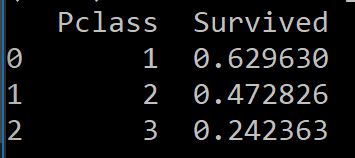
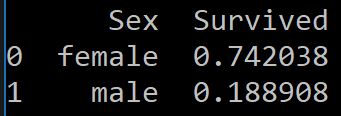
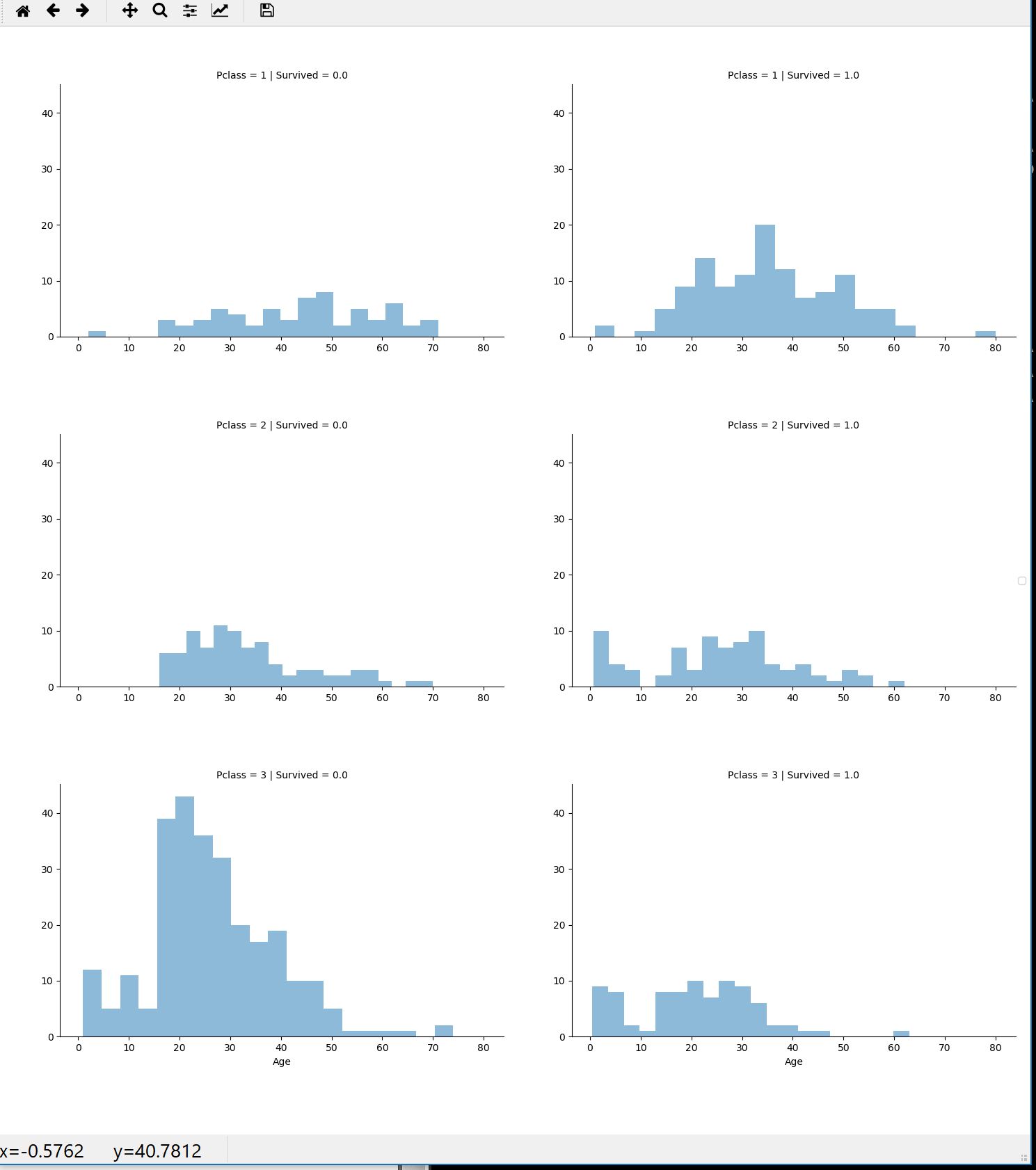
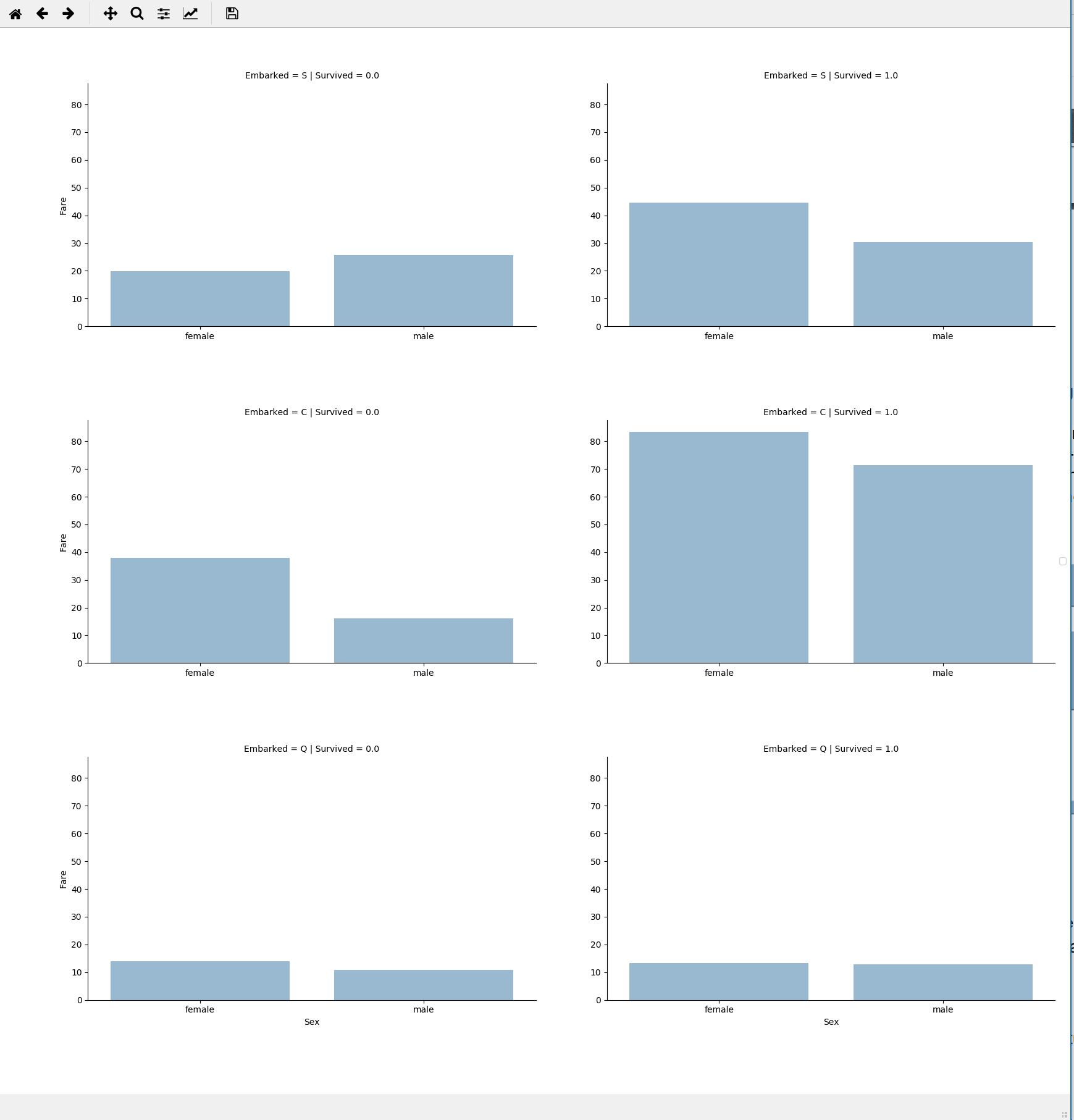
Task 2

1. 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.
2. 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

|  |  |
| --- | --- |
| 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 |

1. 
2. 
3.   
     
   The correlation between Pclass 1 and survival rate is < 50%, so I will not include it in the predictive model.
4.   
     
   Yes, women are more likely to have survived.
   1. Relatively, yes they did. They had the highest survival rate of anyone under 15.
   2. Yes
   3. Yes
   4. Yes
   5. Yes
5. 
   1. Yes, Pclass = 3 had the most passengers and most did not survive.
   2. All infants survived in Pclass = 2. However, about the survival rate in Pclass = 3 was only about ~50%.
   3. Yes.
   4. Yes. Pclass = 1 seems to have mostly middle aged people, whereas the others tend to have slightly younger people.
   5. Yes.
6. 
   1. For Embarked = {S, C}, yes they do. For Embarked = Q, no they do not.
   2. Yes, point of embarkation correlates to survival rates.
   3. Yes
7. 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.
8. 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.
9. See code
10. See code
11. See code
12. See code
13. See code