**Female survived in Titanic**

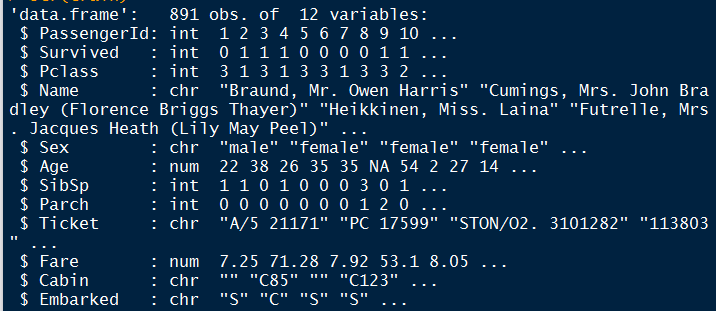
Framework



1. import training data

> train<-read.csv('F:/GraduateStudy/2016Fall/DatAna/Assignment/3/train.csv',stringsAsFactors=F)

> str(train)



2. Missing value and abnormal value processing

> sum(is.na(train))

[1] 177

Choose that complete and create a new dataset

> train1=train[complete.cases(train),]



I don’t know how to deal with the abnormal value here. Sorry….

3. Calculate single factor in Female surviving

1) Pclass

> table(train1$Sex,train1$Pclass,train1$Survived)

, , = 0

1 2 3

female 3 6 55

male 61 84 215

, , = 1

1 2 3

female 82 68 47

male 40 15 38

The female survived in the Titanic most from first and second class, the number of female in the first and second class died in Titanic is less than 10. This is a very important factor.

2) Embarked

> table(train1$Sex,train1$Embarked,train1$Survived)

, , = 0

C Q S

female 0 6 5 53

male 0 45 15 300

, , = 1

C Q S

female 2 55 7 133

male 0 24 1 68

The female survived in the Titanic most from Southampton, the number of female from Queenstown survived is only 7. But the total female from Southampton is more than from Queenstown. We should use proportion to do some deep calculation.

3) Cabin

> table(train1$Sex,train1$Cabin,train1$Survived)

, , = 1

A10 A16 A20 A23 A24 A26 A31 A34 A36 A5 A6 A7 B101

female 113 0 1 0 0 0 0 0 0 0 0 0 0 0

male 52 0 0 1 1 0 1 1 1 0 0 1 0 1

B18 B19 B20 B22 B28 B3 B30 B35 B37 B38 B39 B4 B41 B42

female 2 0 1 1 2 1 0 2 0 0 1 1 0 1

male 0 0 1 0 0 0 0 0 0 0 0 0 1 0

B49 B5 B50 B51 B53 B55 B57 B59 B63 B66 B58 B60 B69 B71

female 1 2 0 0 2 1 1 0

male 1 0 1 1 0 0 0 0

B73 B77 B79 B80 B82 B84 B86 B94 B96 B98 C101 C103 C104

female 1 2 1 1 0 0 0 2 1 1 0

male 0 0 0 0 0 0 0 2 0 0 1

C110 C111 C118 C123 C124 C125 C126 C148 C2 C22 C26

female 0 0 0 1 0 2 0 0 1 0

male 0 0 0 0 0 0 1 1 0 1

C23 C25 C27 C30 C32 C45 C46 C49 C50 C52 C54 C62 C64 C65

female 2 0 1 1 0 0 1 0 1 1 1

male 0 0 0 0 0 0 0 1 0 0 0

C68 C7 C70 C78 C82 C83 C85 C86 C87 C90 C91 C92 C93 C99

female 1 1 0 1 0 1 1 0 0 1 0 0 1 1

male 0 0 1 0 0 0 0 0 0 0 0 1 1 0

D D10 D12 D11 D15 D17 D19 D20 D26 D28 D30 D33 D35 D36

female 2 0 1 1 2 0 2 0 1 0 1 1 2

male 0 1 0 0 0 1 0 0 0 0 1 1 0

D37 D46 D47 D48 D49 D50 D56 D6 D7 D9 E10 E101 E12 E121

female 1 0 1 0 0 0 0 0 1 1 0 2 0 1

male 0 0 0 0 1 0 1 0 0 0 1 0 1 1

E17 E24 E25 E31 E33 E34 E36 E38 E40 E44 E46 E49 E50 E58

female 0 0 0 0 1 1 1 0 1 1 0 1 0 0

male 1 2 2 0 0 0 0 0 0 0 0 0 1 0

E63 E67 E68 E77 E8 F G63 F G73 F2 F33 F4 G6 T

female 0 1 1 0 1 0 0 0 3 1 2 0

male 0 0 0 0 1 0 0 2 0 1 0 0

I can’t find any clues from this result. I’m thinking to make a graph to show every cabin how many female survived. And I also want to sum up the total female surviving number by cabin with the same first letter.

4) Parch

> table(train1$Sex,train1$Parch,train1$Survived)

, , = 0

0 1 2 3 4 5 6

female 33 11 13 1 2 3 1

male 302 38 16 1 2 1 0

, , = 1

0 1 2 3 4 5 6

female 120 44 29 3 0 1 0

male 66 17 10 0 0 0 0

Women without any children or parents are more likely to survive while the more parch they have, the less probability they can survive. (just a assumption, I think this some regression or proportion to identify)

5) SibSp

> table(train1$Sex,train1$SibSp,train1$Survived)

, , = 0

0 1 2 3 4 5

female 29 23 3 4 4 1

male 267 63 11 4 11 4

, , = 1

0 1 2 3 4 5

female 115 67 9 4 2 0

male 60 30 2 0 1 0

Women without any siblings or spouses are more likely to survive while the more siblings they have, the less probability they can survive. ( just an assumption, I think this needs some regression or proportion to identify)

6) title (use approaches on Kaggle. https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic)

Grab title from passenger names

> train1$Title<-gsub('(.\*,)|(\\..\*)','',train1$Name)

Show result

> table(train1$Sex,train1$Title,train1$Survived)

, , = 0

Capt Col Don Dr Jonkheer Lady Major Master Miss

female 0 0 0 0 0 0 0 0 41

male 1 1 1 3 1 0 1 15 0

Mlle Mme Mr Mrs Ms Rev Sir the Countess

female 0 0 0 23 0 0 0 0

male 0 0 331 0 0 6 0 0

, , = 1

Capt Col Don Dr Jonkheer Lady Major Master Miss

female 0 0 0 1 0 1 0 0 105

male 0 1 0 2 0 0 1 21 0

Mlle Mme Mr Mrs Ms Rev Sir the Countess

female 2 1 0 85 1 0 0 1

male 0 0 67 0 0 0 1 0

The survived unmarried women is more than married women, which is accordance with the result about SibSp. Generally, the number of women survived is more than the women died.

Work to do:

I think this method can specify the female characteristics in surviving, but we need to do more analysis.

1. The analysis about age and fare need to classified the numbers first. For example, 20-30 is a same kind, 31-41 is another kind. But how to identify different kinds, I’m not very familiar with and I also think the kind needs to be classified according to average and mean. So, I think we need to discuss about this.
2. The importance of some factors should use proportion but numbers to measure. I will add proportions to every single factor later.
3. I think sometimes two or three factors together will affect the number of survived female. I knew how to calculate this in SPSS. But for R, I’m not very familiar with.
4. I didn’t do the final visualization because I think we need to discuss what kind of graph suitable for what kind of data.