TITANIC

Target: **To predict who survived the disaster by who I mean (group –people/male/female/child…)**

Train data set contains – 12 variables

Test data set contains – 11 variables , the survived column is missing here we are predict it.

```
> test = read.csv("C:\\Users\\Jishu\\Desktop\\titanic\\test.csv")
> View(train)
> names(train)
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"
 [7] "SibSp"       "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
> names(test)
 [1] "PassengerId" "Pclass"      "Name"        "Sex"         "Age"         "SibSp"
 [7] "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
```

To take a look at the structure of the dataframe :

```
> str(train)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417
 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133
...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
>
```

Data types :

Int – integer type stores whole numbers

Num- numeric :hold decimals

Factor -R imports texts/strings as factors

```
#train <- read.csv("train.csv", stringsAsFactors=FALSE)
```

If we have a scenario, that we have to work with lot of text we can use the above command. It doesn't initialize the text as factors.

Checking how many survived and how many didn't in the train data:

```
> table(train$Survived)

  0   1
549 342
```

1 – survived

0 – didn't survive

TITANIC

Proportion of people survived: 38% people survived.

```
> prop.table(table(train$Survived))

        0         1
0.6161616 0.3838384
```

Reference :

```
> prop.table(table(train$Survived))

        0         1
0.6161616 0.3838384
> table(train$Survived)

  0   1
549 342
> 549/(549+342)
[1] 0.6161616
> 342/(549+342)
[1] 0.3838384
>
```

Assuming everyone died in the test data set we add a label column Survived to test dataset.

```
> test$Survived = rep(0,418)#adding a 0 vector column to test->survived
> ncol(test)
[1] 12
>
```

kaggle submission : PassengerId as well as our Survived predictions

```
> submit = data.frame(PassengerId=test$PassengerId,Survived = test$Survived)
> ?row.names
> ?write.csv
> write.csv(submit, file="C:\\Users\\Jishu\\Desktop\\titanic\\submit.csv",row.names = TRUE)
> write.csv(submit, file="C:\\Users\\Jishu\\Desktop\\titanic\\submit.csv",row.names = FALSE)
```

Rownames = TRUE : it provides an extra index -> 1,2,3…..

Rownames= FALSE : it doesn't provide any extra index, only two columns PassengerID and Survived are present.

Disaster – Priority : saving women and children first – taking look at age and sex variable.

```
> summary(train$Sex)
female   male
   314    577
```

```
> table(train$Sex,train$Survived)

            0   1
female  81 233
male   468 109
```

TITANIC

table(train$sex,train$Survived) # gives the result how many males and females survived and how many did not.

Prop.table(table(train$sex,train$Survived)) – gives the proportion of how many females/males survived or did not from the total population

Prop.table(table(train$sex,train$Survived),1) – gives the proportion of how many how many **females** survived or did not survive from **total female population**.

-gives the proportion of how many how many **males** survived or did not survive from **total male population**.

- Each sex that survived as separate groups:
- **(survived_female + not_survived_female) = total female population**
- **(survived_male + not_survived_male)= total male population**
- %Fem_survive from female : survived_female/(survived_female + not_survived_female)
- %Fem_not_survive from female: not_survived_female/(survived_female + not_survived_female)
- %men_survive from male : survived_male/(survived_male + not_survived_male)
- %men_not_survive from male: not_survived_male/(survived_male + not_survived_male)

```
Console ~/
> table(train$sex,train$survived)

            0   1
  female   81 233
  male    468 109
> ?prop.table()
> prop.table(table(train$sex,train$survived),1)

                0         1
  female 0.2579618 0.7420382
  male   0.8110919 0.1889081
> prop.table(table(train$sex,train$survived))

                0         1
  female 0.09090909 0.26150393
  male   0.52525253 0.12233446
> prop.table(table(train$sex,train$survived),2)

                0         1
  female 0.1475410 0.6812865
  male   0.8524590 0.3187135
> 0.09090909/(0.09090909+0.26150393)
[1] 0.2579618
> 0.26150393/(0.09090909+0.26150393)
[1] 0.7420382
>
```

Shows majority of females survived from the entire female population, whereas very few males survived from the male population.

TITANIC

Now lets start digging into Age data .

```
> summary(train$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.42   20.12   28.00   29.70   38.00   80.00     177
>
```

Our last actions were on categorical variables.

***Note on variables**: variables are of two types: Numeric and Categorical*

Numeric variable are those on which mathematical actions can be taken like addition, subtract…

Ex. Age, persons – height, weight, IQ , blood pressure.

Numerical variables can be sub divided – discrete, continuous

Discrete variables can be counted, whereas continuous variables cannot be counted.

Categorical variables are those on which limited number of categories can be identified, we cannot perform arithmetic operations on them.

Ex. Sex, Survived or not, which part of the hemisphere it is- north/south.

Refer: http://www.dummies.com/education/math/statistics/types-of-statistical-data-numerical-categorical-and-ordinal/

Note: Age is a continuous variable, and drawing proportion tables on continuous variables is useless, because its difficult to measure and put into table.

Since there are 177 NA, we assume that they fall at age limit >20 or end of 20s. It is difficult to work with dataset when no data is present or NA.

We create another column, child where condition of being a child is his/her age <18.child <- 1.

Note : Columns with NA will return 0, since NA does not work with Boolean test.

```
Console ~/
> train$Child = 0
> train$Child[train$Age<18]=1
>
```

Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Child
female	17.00	4	2	3101281	7.9250		S	1
male	26.00	2	0	315151	8.6625		S	0
male	32.00	0	0	C.A. 33111	10.5000		S	0
female	16.00	5	2	CA 2144	46.9000		S	1

TITANIC

Purpose of aggregating:

- Survive~Child+Sex : creates a subset which creates a subset of data from the train dataset -> child|sex|survived
- FUN=sum : sums the number of survivors in that group

```
> aggregate(Survived~Child+Sex,data = train,FUN=sum)
  Child    Sex Survived
1     0 female      195
2     1 female       38
3     0   male       86
4     1   male       23
```

Row1: number of survived female passengers whose age>18
Row2: number of survived female passengers whose age<18
Row3: number of survived male passengers whose age>18
Row4: number of survived female passengers whose age<18

- ➢ Total survived population:

```
> 195+38+86+23
[1] 342
```

- FUN=length: total number of passengers in that group

```
> aggregate(Survived~Child+Sex,data = train,FUN=length)
  Child    Sex Survived
1     0 female      259
2     1 female       55
3     0   male      519
4     1   male       58
> 259+55+519+58
[1] 891
>
```

- Row1: total number of female passengers in that group age>18
- Row2: total number of female passengers in that group age <18
- Row3: total number of male passengers in that group age>18
- Row4: total number of male passengers in that group age<18

Note the difference: length -> total includes both survived as well as not survived, whereas sum gives only the survived population.

```
> aggregate(Survived~Child+Sex,data = train,FUN=function(x) {sum(x)/length(x
)})
  Child    Sex  Survived
1     0 female 0.7528958
2     1 female 0.6909091
3     0   male 0.1657033
4     1   male 0.3965517
>
```

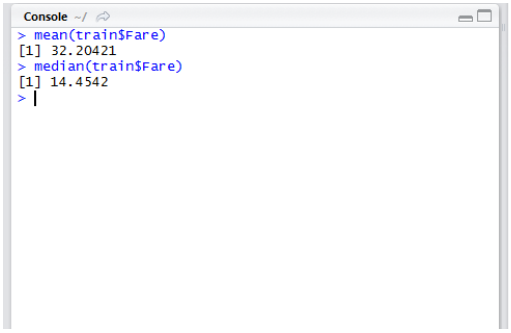
This is the ratio of survived passengers to total number of passengers. But this result also doesn't add difference to the analysis since this also says that more number of female passengers survived than male.

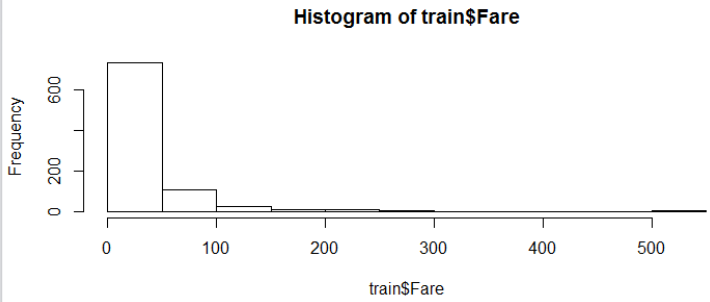
Taking a look at other variable: class the passengers are travelling and how much they paid for their ticket. Class can be evaluated easily because they have 1-3 classes.

TITANIC

Since fare is a continuous variable , it needs to be binned.

	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	male	22.00	1	0	A/5 21171	7.2500		S
ıayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
	female	26.00	0	0	STON/O2. 3101282	7.9250		S
	female	35.00	1	0	113803	53.1000	C123	S





We bin it in 3 slots as to tabulate well. We select the following bin values because the price range is mostly around these prices.

1.Less than $10,

2.Between $10- $20,

3.Between $20- $30.

```
train$Fare2 = '30+'
train$Fare2[train$Fare<10]='10-'
train$Fare2[train$Fare>10 & train$Fare<20]='10-20'
train$Fare2[train$Fare>20 & train$Fare<30]='20-30'
```

Fare	Cabin	Embarked	Child	Fare2
7.2500		S	0	10-
71.2833	C85	C	0	30+
7.9250		S	0	10-
53.1000	C123	S	0	30+
8.0500		S	0	10-
8.4583		Q	0	10-
51.8625	E46	S	0	30+

TITANIC

```
> aggregate(Survived ~ Fare2 + Pclass + Sex, data=train, FUN=function(x) {sum(x)/length(x)})
   Fare2 Pclass    Sex  Survived
1  20-30      1 female 0.8333333
2    30+      1 female 0.9772727
3  10-20      2 female 0.9142857
4  20-30      2 female 0.9000000
5    30+      2 female 1.0000000
6    10-      3 female 0.5937500
7  10-20      3 female 0.5813953
8  20-30      3 female 0.3333333
9    30+      3 female 0.1250000
10   10-      1   male 0.0000000
11 20-30      1   male 0.4000000
12   30+      1   male 0.3837209
13   10-      2   male 0.0000000
14 10-20      2   male 0.1587302
15 20-30      2   male 0.1600000
16   30+      2   male 0.2142857
17   10-      3   male 0.1115385
18 10-20      3   male 0.2368421
19 20-30      3   male 0.1250000
20   30+      3   male 0.2400000
```

From the result it seems that, the person who has survived are mostly female passengers of higher travelling classes and male passengers doesn't show a better result. Rather we understand from the result female passengers travelling in class 3 with $20-$30 or even more >$30 survived less**. Inferred missed life boats but it should not be the case for such expensive cabins. Analysis suggests that those cabins(class3) were close to iceberg hit places or far from stair case, hence resulted in instant deaths of the passengers.**

```
#inference from the above analysis
test$Survived <- 0
test$Survived[test$Sex == 'female'] <- 1
test$Survived[test$Sex == 'female' & test$Pclass == 3 & test$Fare >= 20] <- 0
```

We finally enter in the survived column of test data set gender those who are female survived , but since from our above analysis we see, there is a low chance of survivals for females travelling in class3 who has paid fare more than $20, we consider they did not survive. Hence set the condition female traveler of class3 with fare >20 ,they did not survive i.e 0.

Even the rate of female child survival of class 3 who has paid extremely high fares of more than $30 or $20-30 is extremely low, which is suggested in my future analysis of the data.

```
30 10-20      2     1    male 0.75000000
31 20-30      2     1    male 0.75000000
32   30+      2     1    male 1.00000000
33   10-      3     1    male 0.15384615
34 10-20      3     1    male 0.71428571
35 20-30      3     1    male 0.20000000
36   30+      3     1    male 0.07692308
```

```
10   30+      1     1 female 0.87500000
11 10-20      2     1 female 1.00000000
12 20-30      2     1 female 1.00000000
13   30+      2     1 female 1.00000000
14   10-      3     1 female 0.85714286
15 10-20      3     1 female 0.73333333
16 20-30      3     1 female 0.16666667
17   30+      3     1 female 0.14285714
```

TITANIC

This is my analysis of 100% survival of female child travelers of class 2

```
aggregate(Survived ~ Fare2 + Pclass + Sex, data=train, FUN=function(x) {sum(x)/length(x)})

aggregate(Survived ~ Fare2 + Pclass + Child + Sex, data=train, FUN=function(x) {sum(x)/length(x)})
summary(aggregate(Survived ~ Fare2 + Pclass + Child + Sex, data=train, FUN=function(x) {sum(x)/length(x)}))
```

```
> aggregate(Survived ~ Fare2 + Pclass + Child + Sex, data=train, FUN=function(x) {sum(x)/length(x)})
   Fare2 Pclass Child    Sex   Survived
1  20-30      1     0 female 0.83333333
2    30+      1     0 female 0.98750000
3  10-20      2     0 female 0.90625000
4  20-30      2     0 female 0.88000000
5    30+      2     0 female 1.00000000
6    10-      3     0 female 0.56140351
7  10-20      3     0 female 0.50000000
8  20-30      3     0 female 0.40000000
9    30+      3     0 female 0.11111111
10   30+      1     1 female 0.87500000
11 10-20      2     1 female 1.00000000
12 20-30      2     1 female 1.00000000
13   30+      2     1 female 1.00000000
14   10-      3     1 female 0.85714286
15 10-20      3     1 female 0.73333333
16 20-30      3     1 female 0.16666667
17   30+      3     1 female 0.14285714
```

Below is the summary of the Fare2,Pclass,child,sex subset:

```
> summary(aggregate(Survived ~ Fare2 + Pclass + Child + Sex, data=train, FUN=function(x) {sum(x)/length(x)}))
    Fare2              Pclass          Child             Sex
 Length:36          Min.   :1.00   Min.   :0.0000   female:17
 Class :character   1st Qu.:2.00   1st Qu.:0.0000   male  :19
 Mode  :character   Median :2.00   Median :0.0000
                    Mean   :2.25   Mean   :0.4444
                    3rd Qu.:3.00   3rd Qu.:1.0000
                    Max.   :3.00   Max.   :1.0000
    Survived
 Min.   :0.0000
 1st Qu.:0.1264
 Median :0.4583
 Mean   :0.5068
 3rd Qu.:0.8762
 Max.   :1.0000
```