# CONVERT ATTRIBUTE TYPE IN R

SYRACUSE UNIVERSITY
School of Information Studies

# CONVERT DATA TYPE IN R

When reading data into tools like R, the tool might not interpret the data types correctly.

Examine data definitions in R:

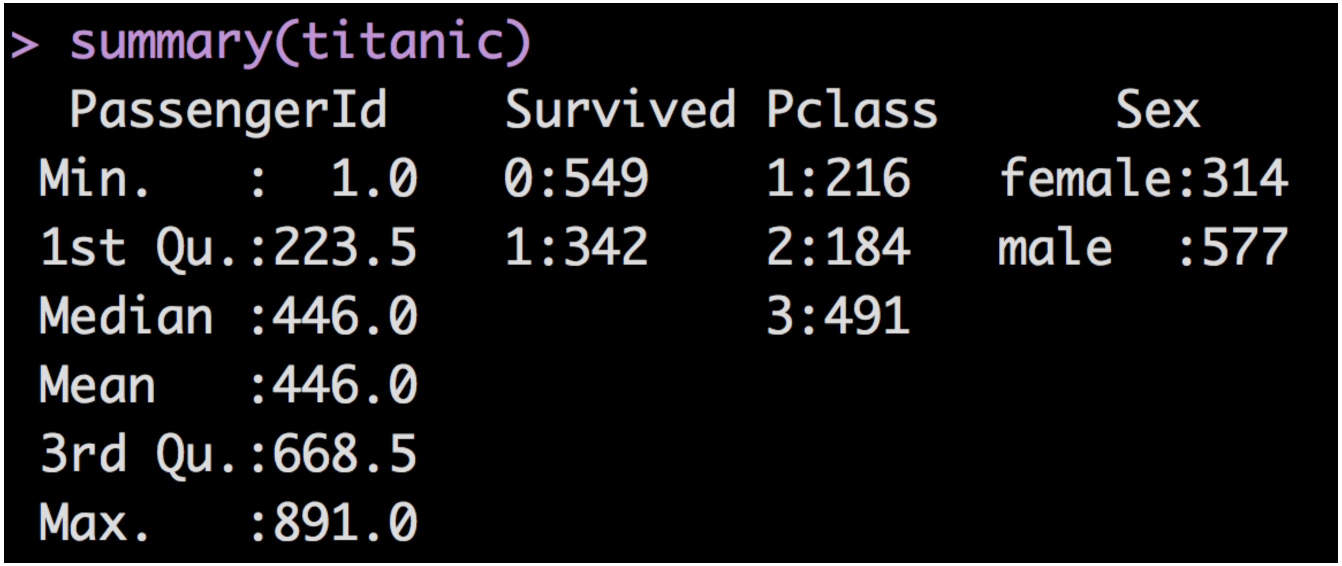
```
> titanic <- read.csv("/Users/byu/Desktop/Data/titanic-train.csv",
na.string = c(""))

> str(titanic)
```

# OUTPUT

```
'data.frame':            891 obs. of  11 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 ...

 $ 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 50 473 345 133 ...

 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...

 $ Cabin      : Factor w/ 147 levels "A10","A14","A16",..: NA 82 NA 56 NA NA 130 NA ...

 $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

# PROBLEM WITH WRONG DATA TYPE

When R misinterpreted nominal variable “PassengerId” as numeric, it would calculate the mean and variance of passenger IDs, which does not make sense.



```
> summary(titanic)
  PassengerId     Survived Pclass      Sex     
 Min.   :  1.0   0:549    1:216   female:314  
 1st Qu.:223.5   1:342    2:184   male  :577  
 Median :446.0            3:491               
 Mean   :446.0                                
 3rd Qu.:668.5                                
 Max.   :891.0                                
```

# CONVERT DATA TYPE IN R

R treats nominal variables as “factors”:

```
> titanic$Survived=factor(titanic$Survived)

> str(titanic)
```

Output:

```
…
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
…
```

# CONVERT DATA TYPE IN R

R treats ordinal variables as “ordered factors”:

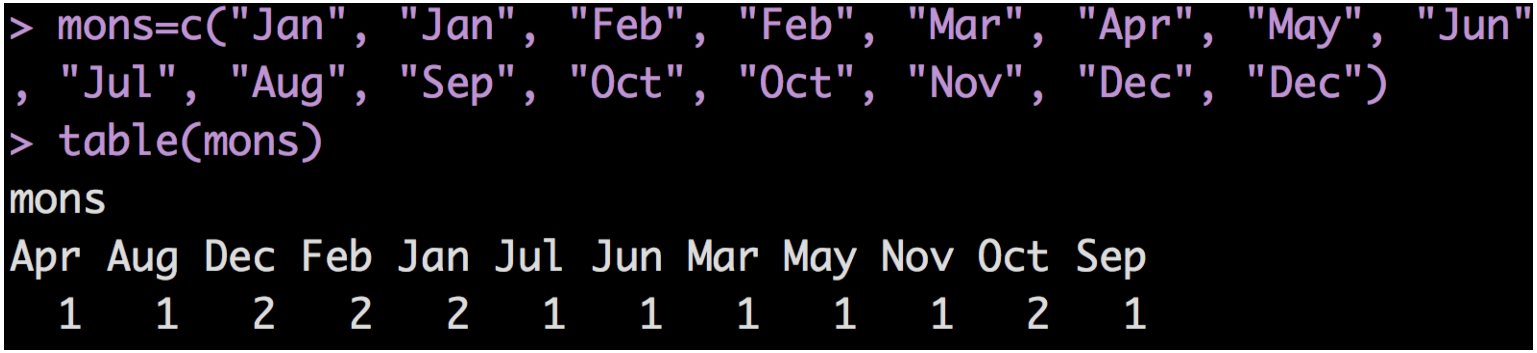
```
> titanic$Pclass=ordered(titanic$Pclass)
> str(titanic)
```

Output:

```
…
 $ Pclass     : Ord.factor w/ 3 levels "1"<"2"<"3": 3 1 3 1 3 3 1 3 3 2 ...
…
```

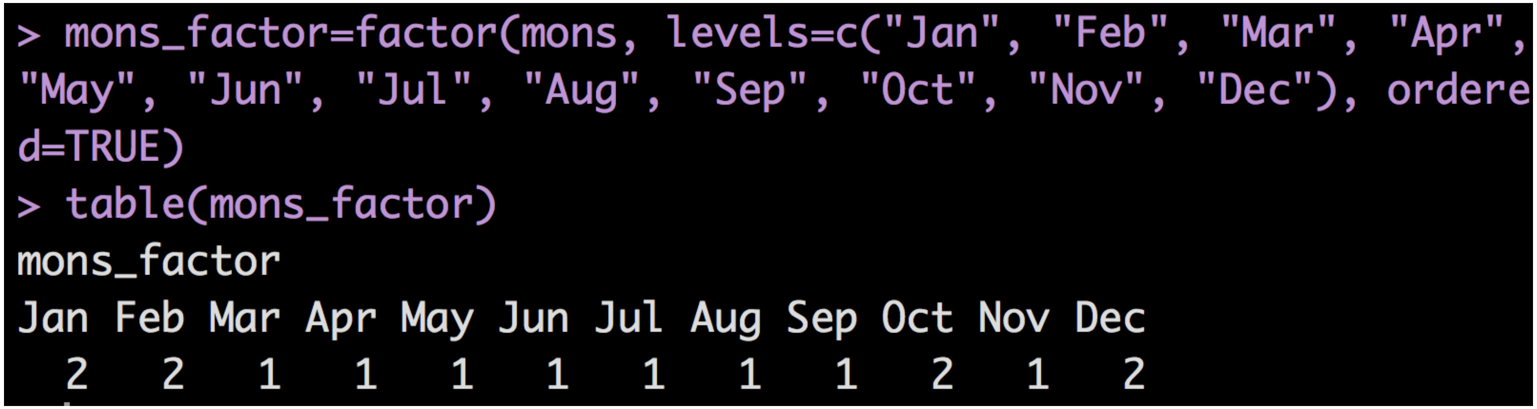
# ORDERED FACTOR IN R

Month defined as a list:



```
> mons=c("Jan", "Jan", "Feb", "Feb", "Mar", "Apr", "May", "Jun"
, "Jul", "Aug", "Sep", "Oct", "Oct", "Nov", "Dec", "Dec")
> table(mons)
mons
Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
  1   1   2   2   2   1   1   1   1   1   2   1
```

Month defined as an ordered factor:



```
> mons_factor=factor(mons, levels=c("Jan", "Feb", "Mar", "Apr",
"May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"), ordere
d=TRUE)
> table(mons_factor)
mons_factor
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
  2   2   1   1   1   1   1   1   1   2   1   2
```