DATA QUALITY ISSUES

SYRACUSE UNIVERSITY
School of Information Studies

DATA QUALITY ISSUES

Noise

Outliers

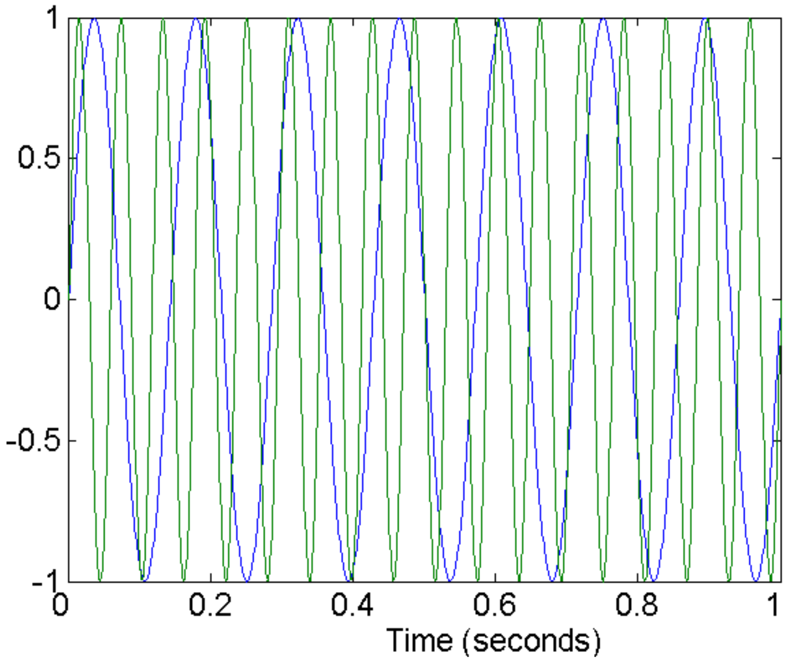
Missing values

Duplicate data

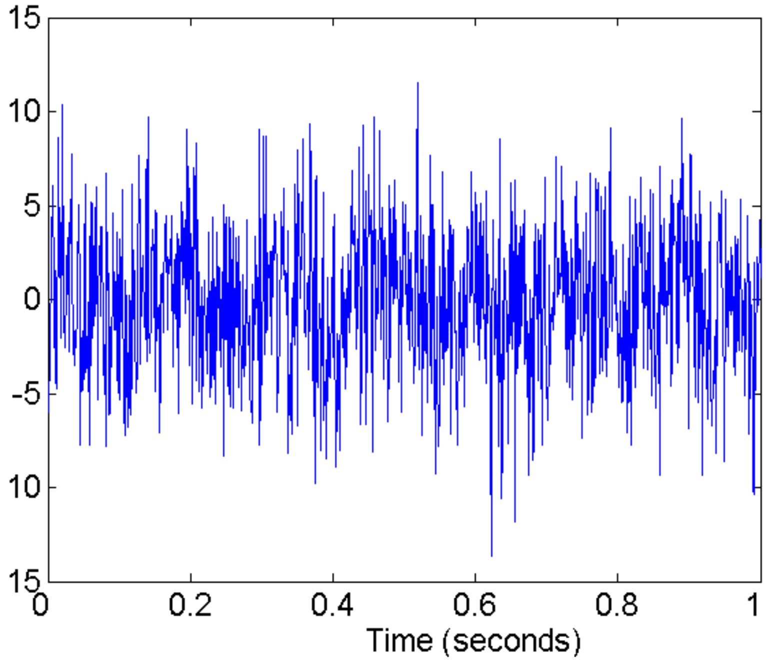
NOISE

Noise refers to modification of original values.

- Examples: Distortion of a person’s voice when talking on a poor-quality phone and “snow” on television screen



Two Sine Waves

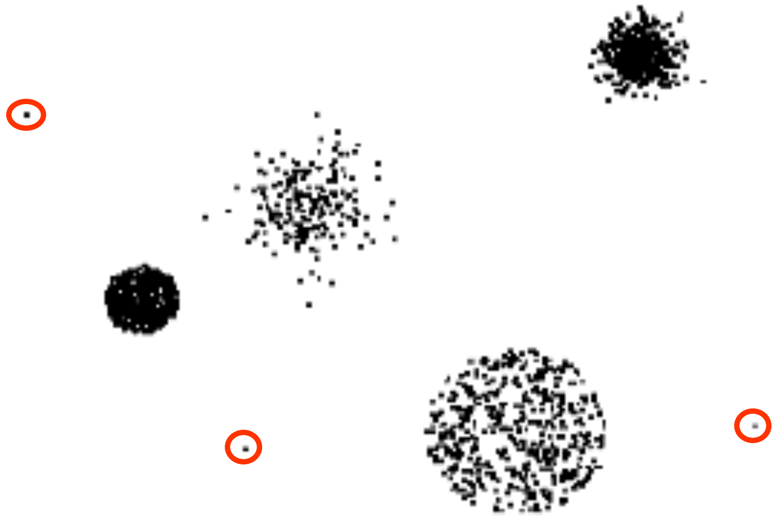


Two Sine Waves + Noise

OUTLIERS

Outliers are data objects with characteristics that are considerably different from most of the other data objects in the data set.

E.g., 250 would be an outlier for variable "people's age."



OUTLIERS SHOULD BE DETECTED AND ANALYZED CAREFULLY

Each year, satellites measure the ozone level over Antarctica.

In the early 1980s, however, scientists were so astounded in detecting a dramatic seasonal drop in ozone levels over Antarctica by a flyover that they spent two years rechecking their satellite data.

They discovered that satellites had dutifully been recording the ozone collapse, but the computers had not raised an alert because they were programmed to reject such extreme data as anomalies.

http://www.stat.wmich.edu/s160/book/node8.html

MISSING VALUES

Why are values missing?

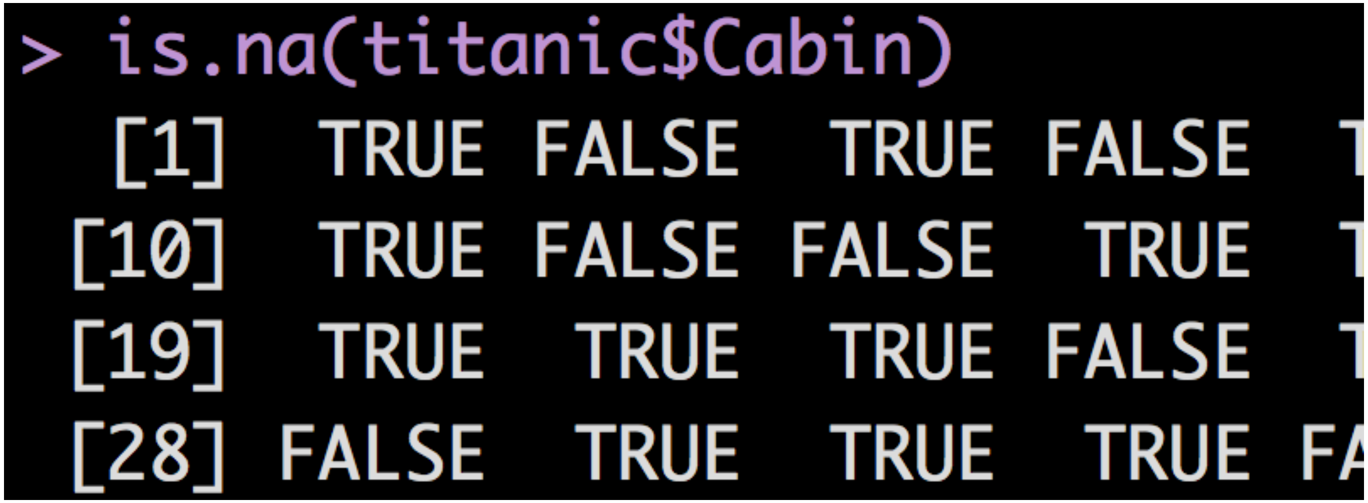
- Information is not collected.
(E.g., people decline to give their age and weight.)
- Attributes may not be applicable to all cases.
(E.g., annual income is not applicable to most children.)

Handling missing values:

- Eliminate data objects.
- Ignore the missing value during analysis.
- Estimate missing values and replace them.

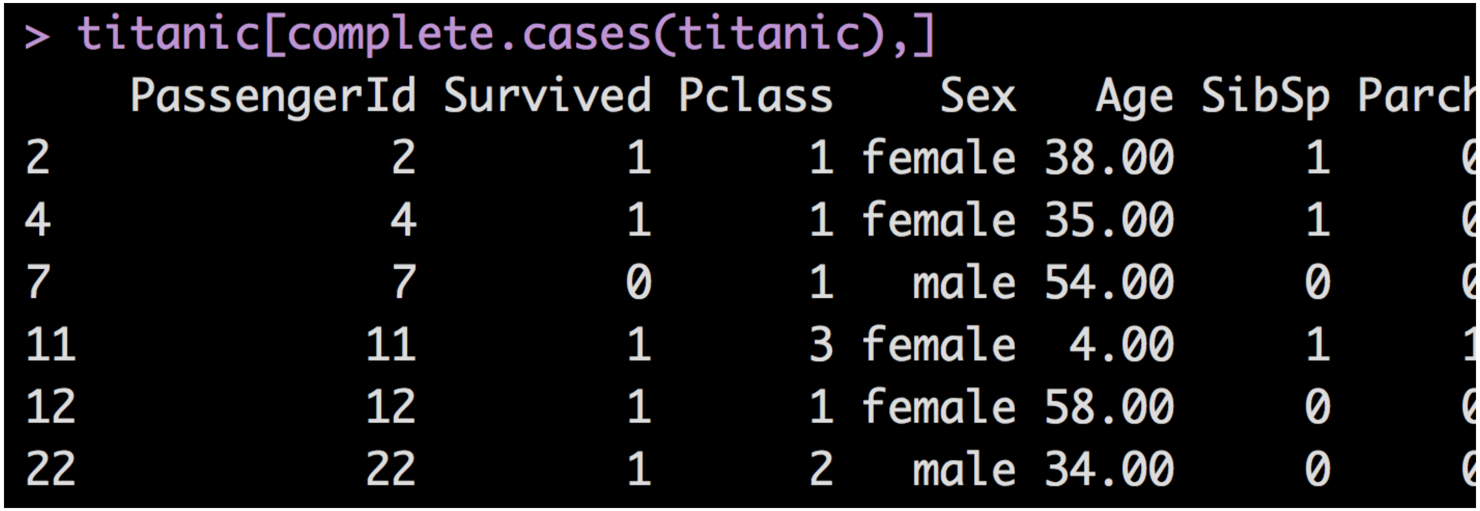
CHECK MISSING VALUES IN R

```
>is.na(titanic)
>is.na(titanic$Cabin)
```

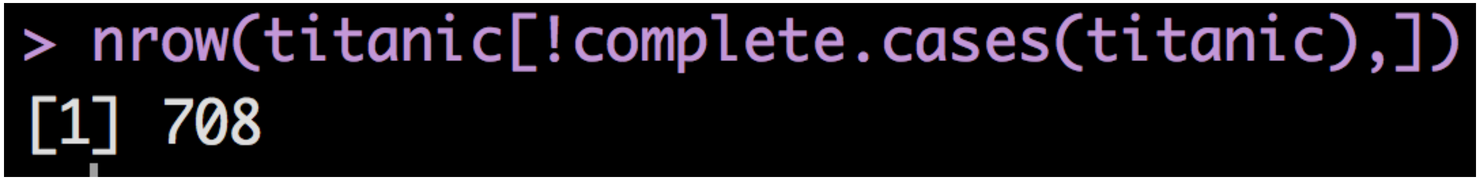


```
> is.na(titanic$Cabin)
  [1]  TRUE FALSE  TRUE FALSE  T
 [10]  TRUE FALSE FALSE  TRUE  T
 [19]  TRUE  TRUE  TRUE FALSE  T
 [28] FALSE  TRUE  TRUE  TRUE FA
```

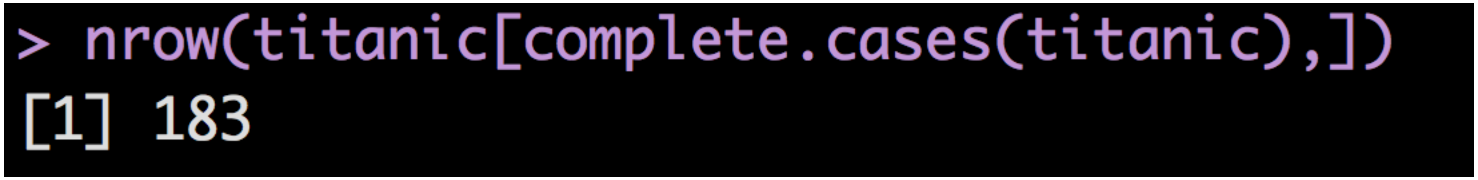
FIND COMPLETE RECORDS



```
> titanic[complete.cases(titanic),]
    PassengerId Survived Pclass    Sex   Age SibSp Parch
2             2        1      1 female 38.00     1     0
4             4        1      1 female 35.00     1     0
7             7        0      1   male 54.00     0     0
11           11        1      3 female  4.00     1     1
12           12        1      1 female 58.00     0     0
22           22        1      2   male 34.00     0     0
```

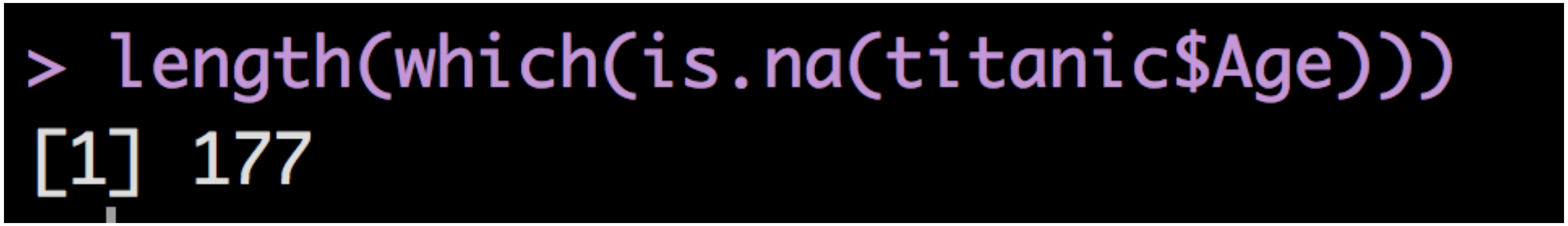


```
> nrow(titanic[!complete.cases(titanic),])
[1] 708
```



```
> nrow(titanic[complete.cases(titanic),])
[1] 183
```

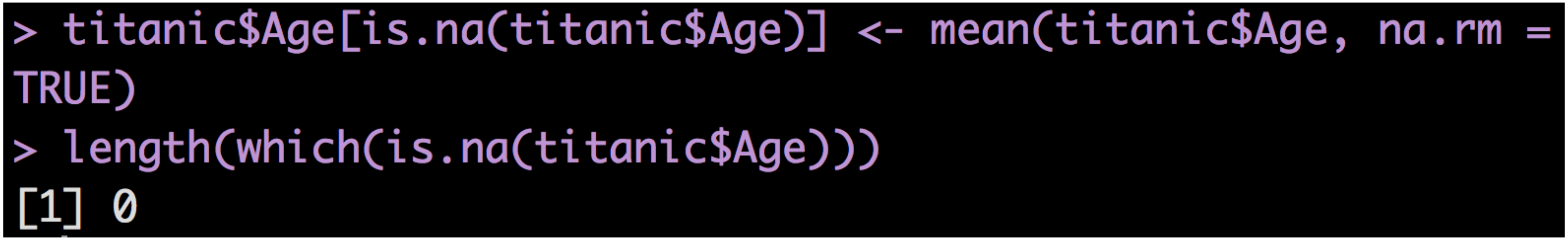
COUNT NUMBER OF MISSING VALUES



```
> length(which(is.na(titanic$Age)))
[1] 177
```

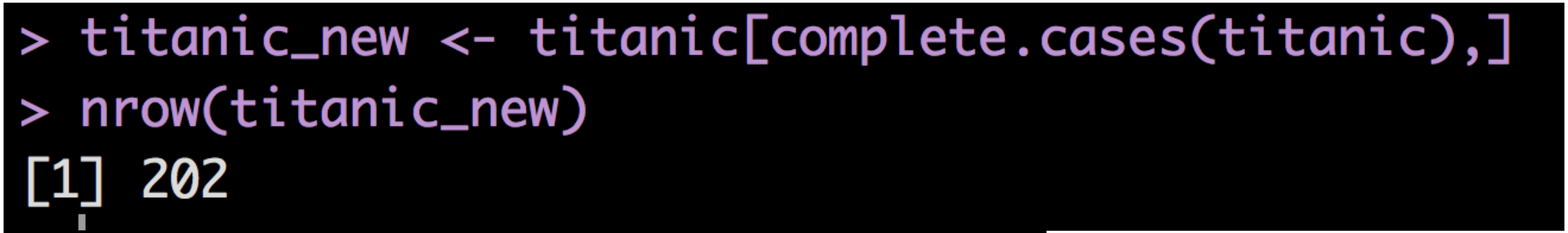
Is “age” still a useful variable for predicting survivors?

ESTIMATE AND REPLACE MISSING VALUES

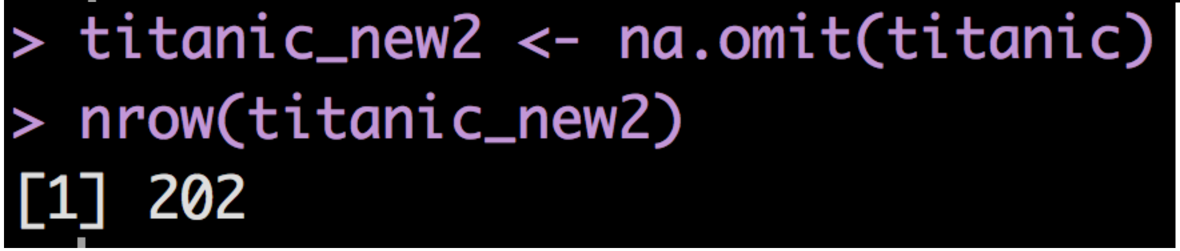


```
> titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm =
TRUE)
> length(which(is.na(titanic$Age)))
[1] 0
```

REMOVE RECORDS WITH MISSING VALUES



```
> titanic_new <- titanic[complete.cases(titanic),]
> nrow(titanic_new)
[1] 202
```

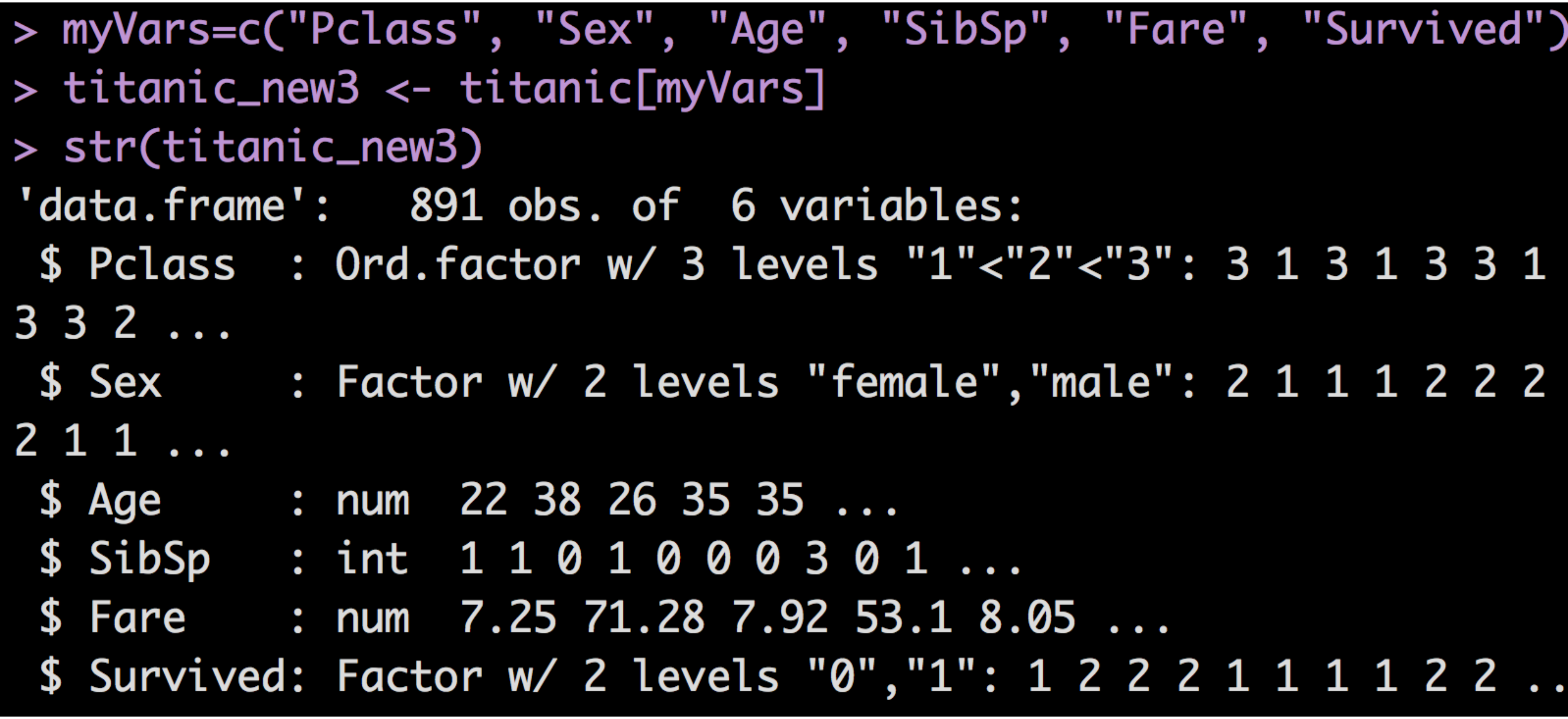


```
> titanic_new2 <- na.omit(titanic)
> nrow(titanic_new2)
[1] 202
```

Isn’t the number of complete cases 183?

It was, but remember the missing values in “age” have been replaced by average age.

REMOVE VARIABLES WITH MISSING VALUES



```
> myVars=c("Pclass", "Sex", "Age", "SibSp", "Fare", "Survived")
> titanic_new3 <- titanic[myVars]
> str(titanic_new3)
'data.frame':   891 obs. of  6 variables:
 $ Pclass  : Ord.factor w/ 3 levels "1"<"2"<"3": 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 ...
 $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ..
```

DUPLICATE DATA

Data set may include data objects that are duplicates or almost duplicates of one another.

- Major issue when merging data from heterogenous sources

Examples:

- Same person with multiple e-mail addresses

Data cleaning:

- Process of dealing with duplicate data issues

AN EXAMPLE OF DUPLICATE DATA

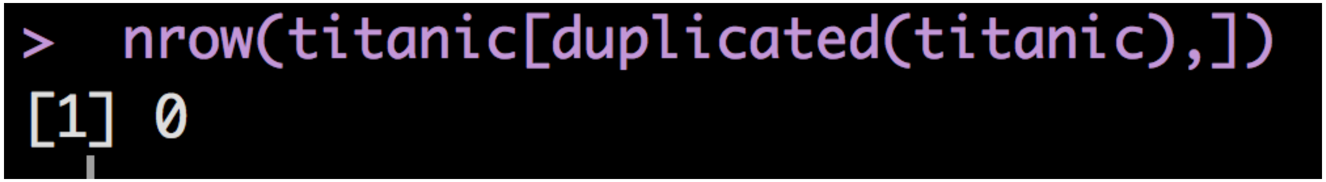
An Amazon Mechanical Turk worker set up two accounts and finished a task twice in order to get double payment.

Two identical records were sent to the data collector.

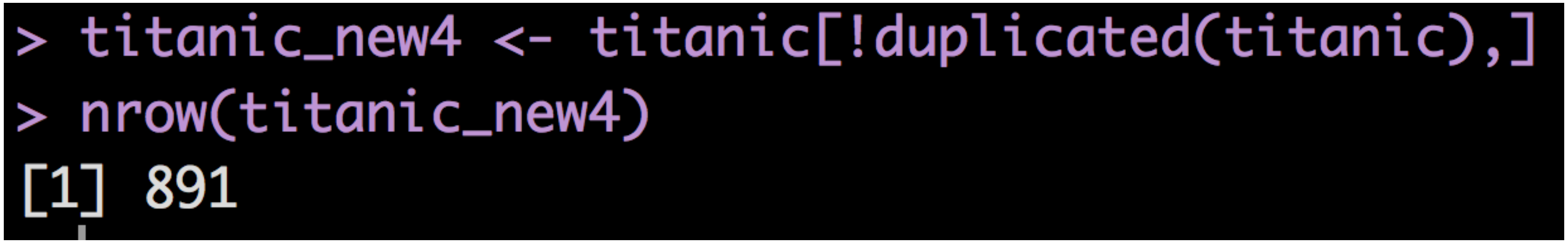
How to identify them?

- Check IP address.
- Compare similarity between records.

CHECK AND REMOVE DUPLICATED RECORDS



```
>  nrow(titanic[duplicated(titanic),])
[1] 0
```



```
> titanic_new4 <- titanic[!duplicated(titanic),]
> nrow(titanic_new4)
[1] 891
```

REVIEW DATA QUALITY ISSUES

Noise

Outliers

Missing values

Duplicate data