Data Refinement

https://github.com/EmanuelGaw/SIwB https://mybinder.org/v2/gh/EmanuelGaw/SIwB/HEAD

Packages

Titanic Data

Load Titanic data

```
titanicMessData <- read.table('TitanicMess.tsv',
    header = TRUE,
    sep = "\t",
    row.names = NULL)

summary(titanicMessData)

##   PassengerId       Survived          Pclass          Name          
##  Min.   :   1.0   Min.   :0.0000   Min.   :1.000   Length:892        
##  1st Qu.: 223.8   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
##  Median : 444.5   Median :0.0000   Median :3.000   Mode  :character  
##  Mean   : 445.8   Mean   :0.3868   Mean   :2.307                     
##  3rd Qu.: 668.2   3rd Qu.:1.0000   3rd Qu.:3.000                     
##  Max.   :1000.0   Max.   :1.0000   Max.   :3.000                     
##      Sex                Age                SibSp            Parch       
##  Length:892         Length:892         Min.   :0.0000   Min.   :0.0000  
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.0000  
##                                        Mean   :0.5258   Mean   :0.3711  
##                                        3rd Qu.:1.0000   3rd Qu.:0.0000  
##                                        Max.   :8.0000   Max.   :5.0000  
##     Ticket              Fare              Cabin             Embarked        
##  Length:892         Length:892         Length:892         Length:892        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      ship          
##  Length:892        
##  Class :character  
##  Mode  :character  
```

Dataset refinement

We have 13 dataset attributes, but we can expect, that not every attribute would be useful in data exploration.

Handling duplicates

It can happen, that the dataset contians duplicated data.
First we want to remove rows containing duplicated data from the dataset.

We can use the "distinct()" function from package "dplyr":

```
titanicMessData <- distinct(titanicMessData)
summary(titanicMessData)
```

```
##   PassengerId      Survived          Pclass          Name          
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:889        
##  1st Qu.: 225   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
##  Median : 446   Median :0.0000   Median :3.000   Mode  :character  
##  Mean   : 447   Mean   :0.3847   Mean   :2.307                     
##  3rd Qu.: 669   3rd Qu.:1.0000   3rd Qu.:3.000                     
##  Max.   :1000   Max.   :1.0000   Max.   :3.000                     
##      Sex                Age                SibSp            Parch       
##  Length:889         Length:889         Min.   :0.0000   Min.   :0.0000  
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.0000  
##                                        Mean   :0.5242   Mean   :0.3701  
##                                        3rd Qu.:1.0000   3rd Qu.:0.0000  
##                                        Max.   :8.0000   Max.   :5.0000  
##     Ticket              Fare              Cabin             Embarked        
##  Length:889         Length:889         Length:889         Length:889        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      ship          
##  Length:889        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

We had 892 rows of data, now the dataset has 889 rows.

Defining valuable attributes

Now we want to take a short look onto the dataset, to see the specificity of each column:

```
head(titanicMessData, 40)
```

```
##    PassengerId Survived Pclass
## 1            1        0      3
## 2            2        1      1
## 3            3        1      3
## 4            4        1      1
## 5            5        0      3
## 6            6        0      3
## 7            7        0      1
## 8            8        0      3
## 9            9        1      3
## 10          10        1      2
## 11          11        1      3
## 12          12        1      1
## 13          13        0      3
## 14          15        0      3
## 15          16        1      2
## 16          17        0      3
## 17          18        1      2
## 18          19        0      3
## 19          20        1      3
## 20          21        0      2
## 21          22        1      2
## 22          23        1      3
## 23          25        0      3
## 24          26        1      3
## 25          27        0      3
## 26          28        0      1
## 27          29        1      3
## 28          30        0      3
## 29          31        0      1
## 30          32        1      1
## 31          33        1      3
## 32          34        0      2
## 33          35        0      1
## 34          36        0      1
## 35          37        1      3
## 36          38        0      3
## 37          39        0      3
## 38          40        1      3
## 39          41        0      3
## 40          42        0      2
##                                                         Name    Sex Age
SibSp
## 1                                     Braund, Mr. Owen Harris   male  22
1
## 2         Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38
1
## 3                                      Heikkinen, Miss. Laina female  26
0
## 4                Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35
```

```
1
## 5                                   Allen, Mr. William Henry   male  35
0
## 6                                           Moran, Mr. James   male    
0
## 7                                    McCarthy, Mr. Timothy J   male  54
0
## 8                             Palsson, Master. Gosta Leonard   male   2
3
## 9          Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female  27
0
## 10                       Nasser, Mrs. Nicholas (Adele Achem) female  14
1
## 11                        Sandstrom, Miss. Marguerite Ru&5$$ female   4
1
## 12                                  Bonnell, Miss. Elizabeth female  58
0
## 13                            Saundercock, Mr. William Henry   male  20
0
## 14                      Vestrom, Miss. Hulda Amanda Adolfina female  14
0
## 15                           Hewlett, Mrs. (Mary D Kingcome) female  55
0
## 16                                      Rice, Master. Eugene   male   2
4
## 17                              Williams, Mr. Charles Eugene   male    
0
## 18   Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) female  31
1
## 19                                   Masselmani, Mrs. Fatima female    
0
## 20                                      Fynney, Mr. Joseph J   male  35
0
## 21                                     Beesley, Mr. Lawrence  malef  34
0
## 22                               McGowan, Miss. Anna "Annie" female  15
0
## 23                             Palsson, Miss. Torborg Danira female   8
3
## 24 Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson) female  38
1
## 25                                   Emir, Mr. Farred Chehab   male    
0
## 26                            Fortune, Mr. Charles Alexander   male  19
3
## 27                             O'Dwyer, Miss. Ellen "Nellie" female    
0
## 28                                       Todoroff, Mr. Lalio   male    
0
## 29                                  Uruchurtu, Don. Manuel E   male  40
```

```
0
## 30            Spencer, Mrs. William Augustus (Marie Eugenie) female  .9
1
## 31                                  Glynn, Miss. Mary Agatha female
0
## 32                                     Wheadon, Mr. Edward H   male  66
0
## 33                                   Meyer, Mr. Edgar Joseph   male  28
1
## 34                            Holverson, Mr. Alexander Oskar   male  42
1
## 35                                          Mamee, Mr. Hanna   male
0
## 36                                  Cann, Mr. Ernest Charles   male  21
0
## 37                        Vander Planke, Miss. Augusta Maria female  18
2
## 38                               Nicola-Yarred, Miss. Jamila female  14
1
## 39            Ahlin, Mrs. Johan (Johanna Persdotter Larsson) female  40
1
## 40  Turpin, Mrs. William John Robert (Dorothy Ann Wonnacott) female  27
1
##    Parch           Ticket      Fare       Cabin Embarked    ship
## 1      0        A/5 21171      7,25                    S Titanic
## 2      0         PC 17599  71,2833         C85        C Titanic
## 3      0 STON/O2. 3101282    7,925                    S Titanic
## 4      0           113803      53,1        C123        S Titanic
## 5      0           373450      8,05                    S Titanic
## 6      0           330877    8,4583                    Q Titanic
## 7      0            17463  51,8625         E46        S Titanic
## 8      1           349909    21,075                    S Titanic
## 9      2           347742  11,1333                    S Titanic
## 10     0           237736  30,0708                    C Titanic
## 11     1          PP 9549      16,7          G6        S Titanic
## 12     0           113783     26,55        C103        S Titanic
## 13     0        A/5. 2151      8,05                    S Titanic
## 14     0           350406    7,8542                    S Titanic
## 15     0           248706        16                    S Titanic
## 16     1           382652    29,125                    Q Titanic
## 17     0           244373        13                   So Titanic
## 18     0           345763        18                    S Titanic
## 19     0             2649     7,225                    C Titanic
## 20     0           239865        26                    S Titanic
## 21     0           248698        13         D56        S Titanic
## 22     0           330923    8,0292                    Q Titanic
## 23     1           349909    21,075                    S Titanic
## 24     5           347077  31,3875                    S Titanic
## 25     0             2631     7,225                    C Titanic
## 26     2            19950       263 C23 C25 C27        S Titanic
## 27     0           330959   7,8792                    Q Titanic
## 28     0           349216   7,8958                    S Titanic
## 29     0         PC 17601  27,7208                    C Titanic
## 30     0         PC 17569 146,5208         B78        C Titanic
## 31     0           335677     7,75                    Q Titanic
## 32     0       C.A. 24579     10,5                    S Titanic
## 33     0         PC 17604  82,1708                    C Titanic
## 34     0           113789       52                    S Titanic
## 35     0             2677   7,2292                    C Titanic
## 36     0       A./5. 2152     8,05                    S Titanic
## 37     0           345764       18                    S Titanic
## 38     0             2651  11,2417                    C Titanic
## 39     0             7546    9,475                    S Titanic
## 40     0            11668       21                    S Titanic
```

Taking a short look we grasp, that column PassengerId and Ship are not useful.
We know that the ship is Titanic, and PassengerId's are unique. We don't need this columns.

We remove columns "PassengersId" (col id 1) and "Ship" (col id 13):

```
titanicMessData <- titanicMessData[, -c(1,13)]
head(titanicMessData, 12)
```

```
##    Survived Pclass                                                Name
Sex
## 1         0      3                             Braund, Mr. Owen Harris
male
## 2         1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
female
## 3         1      3                              Heikkinen, Miss. Laina
female
## 4         1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel)
female
## 5         0      3                            Allen, Mr. William Henry
male
## 6         0      3                                    Moran, Mr. James
male
## 7         0      1                             McCarthy, Mr. Timothy J
male
## 8         0      3                      Palsson, Master. Gosta Leonard
male
## 9         1      3   Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
female
## 10        1      2                 Nasser, Mrs. Nicholas (Adele Achem)
female
## 11        1      3                  Sandstrom, Miss. Marguerite Ru&5$$
female
## 12        1      1                            Bonnell, Miss. Elizabeth
female
##    Age SibSp Parch           Ticket    Fare Cabin Embarked
## 1   22     1     0        A/5 21171    7,25              S
## 2   38     1     0         PC 17599 71,2833   C85        C
## 3   26     0     0 STON/O2. 3101282   7,925              S
## 4   35     1     0           113803     53,1  C123        S
## 5   35     0     0           373450     8,05              S
## 6          0     0           330877   8,4583              Q
## 7   54     0     0            17463 51,8625   E46        S
## 8    2     3     1           349909   21,075              S
## 9   27     0     2           347742 11,1333              S
## 10  14     1     0           237736 30,0708              C
## 11   4     1     1          PP 9549     16,7    G6        S
## 12  58     0     0           113783    26,55  C103        S
```

We removed columns, which values are not helpful in the analysis.

Handling missing values

Many datasets often contain data rows with some missing attribute values.
These missing values can be marked as: " “,”Na“,”NaN", etc…

Let’s explicitly mark them out in our dataset:

```
titanicMessData[
  titanicMessData == ''
  | titanicMessData == ' '
  | titanicMessData == "Na"
  | titanicMessData == "NaN"
  ] <- NA
```

Let’s now check the count of missing values for each column containing missing values:

```
colSums(is.na(titanicMessData))
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        0        0        0        0      173        0        0        0
##     Fare    Cabin Embarked
##        0      685        2
```

Out of almost 900 rows at least 685 contain missing values.
In order to solve the problem of missing values we can insert mode values,
or remove the rows.

However, in this case there is also other possibility, which is removing of the “Cabin” column. We choose to remove this column, for it would give much noise in dataset:
- replacing makes the analysis results wander from the truth,
- removing rows would leave a small part of a dataset.

Removing “Cabin”:

```
titanicMessData <- titanicMessData[, -10]
summary(titanicMessData)
```

```
##     Survived          Pclass          Name               Sex
##  Min.   :0.0000   Min.   :1.000   Length:889         Length:889
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
##  Median :0.0000   Median :3.000   Mode  :character   Mode  :character
##  Mean   :0.3847   Mean   :2.307
##  3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :3.000
##      Age               SibSp            Parch            Ticket
##  Length:889         Min.   :0.0000   Min.   :0.0000   Length:889
##  Class :character   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :0.0000   Median :0.0000   Mode  :character
##                     Mean   :0.5242   Mean   :0.3701
##                     3rd Qu.:1.0000   3rd Qu.:0.0000
##                     Max.   :8.0000   Max.   :5.0000
##      Fare             Embarked
##  Length:889         Length:889
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

We remember, there are still two more columns with missing values:

```
colSums(is.na(titanicMessData))

## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        0        0        0        0      173        0        0        0
##     Fare Embarked
##        0        2
```

Now the number of rows containing missing values is around 20%.
Therefore we decide to remove the rows with missing values:

```
titanicMessData <- na.omit(titanicMessData)
head(titanicMessData, 12)

##    Survived Pclass                                               Name
Sex
## 1         0      3                            Braund, Mr. Owen Harris
male
## 2         1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
female
## 3         1      3                             Heikkinen, Miss. Laina
female
## 4         1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
female
## 5         0      3                           Allen, Mr. William Henry
male
## 7         0      1                            McCarthy, Mr. Timothy J
male
```

```
## 8         0      3                      Palsson, Master. Gosta Leonard
male
## 9         1      3   Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
female
## 10        1      2                 Nasser, Mrs. Nicholas (Adele Achem)
female
## 11        1      3                  Sandstrom, Miss. Marguerite Ru&5$$
female
## 12        1      1                            Bonnell, Miss. Elizabeth
female
## 13        0      3                      Saundercock, Mr. William Henry
male
##    Age SibSp Parch           Ticket    Fare Embarked
## 1   22     1     0        A/5 21171    7,25        S
## 2   38     1     0         PC 17599 71,2833        C
## 3   26     0     0 STON/O2. 3101282   7,925        S
## 4   35     1     0           113803    53,1        S
## 5   35     0     0           373450    8,05        S
## 7   54     0     0            17463 51,8625        S
## 8    2     3     1           349909  21,075        S
## 9   27     0     2           347742 11,1333        S
## 10  14     1     0           237736 30,0708        C
## 11   4     1     1          PP 9549    16,7        S
## 12  58     0     0           113783   26,55        S
## 13  20     0     0        A/5. 2151    8,05        S
```

Now the dataset is set free from the rows including missing values.

```
head(titanicMessData, 40)
```

```
##    Survived Pclass
Name
## 1         0      3                                   Braund, Mr. Owen
Harris
## 2         1      1       Cumings, Mrs. John Bradley (Florence Briggs
Thayer)
## 3         1      3                                    Heikkinen, Miss.
Laina
## 4         1      1              Futrelle, Mrs. Jacques Heath (Lily May
Peel)
## 5         0      3                                  Allen, Mr. William
Henry
## 7         0      1                                   McCarthy, Mr. Timothy
J
## 8         0      3                            Palsson, Master. Gosta
Leonard
## 9         1      3         Johnson, Mrs. Oscar W (Elisabeth Vilhelmina
Berg)
## 10        1      2                       Nasser, Mrs. Nicholas (Adele
Achem)
## 11        1      3                        Sandstrom, Miss. Marguerite
```

```
Ru&5$$
## 12        1      1                                  Bonnell, Miss.
Elizabeth
## 13        0      3                            Saundercock, Mr. William
Henry
## 14        0      3                      Vestrom, Miss. Hulda Amanda
Adolfina
## 15        1      2                           Hewlett, Mrs. (Mary D
Kingcome)
## 16        0      3                                      Rice, Master.
Eugene
## 18        0      3   Vander Planke, Mrs. Julius (Emelia Maria
Vandemoortele)
## 20        0      2                                      Fynney, Mr. Joseph
J
## 21        1      2                                     Beesley, Mr.
Lawrence
## 22        1      3                               McGowan, Miss. Anna
"Annie"
## 23        0      3                             Palsson, Miss. Torborg
Danira
## 24        1      3 Asplund, Mrs. Carl Oscar (Selma Augusta Emilia
Johansson)
## 26        0      1                            Fortune, Mr. Charles
Alexander
## 29        0      1                                  Uruchurtu, Don. Manuel
E
## 30        1      1            Spencer, Mrs. William Augustus (Marie
Eugenie)
## 32        0      2                                     Wheadon, Mr. Edward
H
## 33        0      1                                   Meyer, Mr. Edgar
Joseph
## 34        0      1                            Holverson, Mr. Alexander
Oskar
## 36        0      3                                  Cann, Mr. Ernest
Charles
## 37        0      3                        Vander Planke, Miss. Augusta
Maria
## 38        1      3                               Nicola-Yarred, Miss.
Jamila
## 39        0      3            Ahlin, Mrs. Johan (Johanna Persdotter
Larsson)
## 40        0      2  Turpin, Mrs. William John Robert (Dorothy Ann
Wonnacott)
## 42        1      2                  Laroche, Miss. Simonne Marie Anne
Andree
## 43        1      3                             Devaney, Miss. Margaret
Delia
## 48        0      3             Arnold-Franchi, Mrs. Josef (Josefine
Franchi)
## 49        0      3                                Panula, Master. Juha
Niilo
## 50        0      3                              Nosworthy, Mr. Richard
Cater
## 51        1      1                  Harper, Mrs. Henry Sleeper (Myna
Haxtun)
## 52        1      2        Faunthorpe, Mrs. Lizzie (Elizabeth Anne
Wilkinson)
## 53        0      1                            Ostby, Mr. Engelhart
Cornelius
##       Sex Age SibSp Parch           Ticket     Fare Embarked
## 1    male  22     1     0        A/5 21171     7,25        S
## 2  female  38     1     0         PC 17599  71,2833        C
## 3  female  26     0     0 STON/O2. 3101282    7,925        S
## 4  female  35     1     0           113803     53,1        S
## 5    male  35     0     0           373450     8,05        S
## 7    male  54     0     0            17463  51,8625        S
## 8    male   2     3     1           349909   21,075        S
## 9  female  27     0     2           347742  11,1333        S
## 10 female  14     1     0           237736  30,0708        C
## 11 female   4     1     1          PP 9549     16,7        S
## 12 female  58     0     0           113783    26,55        S
## 13   male  20     0     0        A/5. 2151     8,05        S
## 14 female  14     0     0           350406   7,8542        S
## 15 female  55     0     0           248706       16        S
## 16   male   2     4     1           382652   29,125        Q
## 18 female  31     1     0           345763       18        S
## 20   male  35     0     0           239865       26        S
## 21  malef  34     0     0           248698       13        S
## 22 female  15     0     0           330923   8,0292        Q
## 23 female   8     3     1           349909   21,075        S
## 24 female  38     1     5           347077  31,3875        S
## 26   male  19     3     2            19950      263        S
## 29   male  40     0     0         PC 17601  27,7208        C
## 30 female  .9     1     0         PC 17569 146,5208        C
## 32   male  66     0     0       C.A. 24579     10,5        S
## 33   male  28     1     0         PC 17604  82,1708        C
## 34   male  42     1     0           113789       52        S
## 36   male  21     0     0       A./5. 2152     8,05        S
## 37 female  18     2     0           345764       18        S
## 38 female  14     1     0             2651  11,2417        C
## 39 female  40     1     0             7546    9,475        S
## 40 female  27     1     0            11668       21        S
## 42 female   3     1     2    SC/Paris 2123  41,5792        C
## 43 female  19     0     0           330958   7,8792        Q
## 48 female  18     1     0           349237     17,8        S
## 49   male   7     4     1          3101295  39,6875        S
## 50   male  21     0     0       A/4. 39886      7,8        S
## 51 female  49     1     0         PC 17572  76,7292        C
## 52 female  29     1     0             2926       26        S
## 53   male  65     0     1           113509  61,9792        C
```

```
summary(titanicMessData)
```

```
##     Survived          Pclass          Name                Sex           
##   Min.   :0.0000   Min.   :1.000   Length:714         Length:714        
##   1st Qu.:0.0000   1st Qu.:1.000   Class :character   Class :character  
##   Median :0.0000   Median :2.000   Mode  :character   Mode  :character  
##   Mean   :0.4062   Mean   :2.237                                        
##   3rd Qu.:1.0000   3rd Qu.:3.000                                        
##   Max.   :1.0000   Max.   :3.000                                        
##      Age                SibSp           Parch          Ticket          
##   Length:714         Min.   :0.000   Min.   :0.000   Length:714        
##   Class :character   1st Qu.:0.000   1st Qu.:0.000   Class :character  
##   Mode  :character   Median :0.000   Median :0.000   Mode  :character  
##                      Mean   :0.514   Mean   :0.416                     
##                      3rd Qu.:1.000   3rd Qu.:1.000                     
##                      Max.   :5.000   Max.   :5.000                     
##      Fare             Embarked        
##   Length:714         Length:714        
##   Class :character   Class :character  
##   Mode  :character   Mode  :character  
##                                        
##                                        
##                                        
```

Done. The dataset after the refinement has 714 data records and is ready for further analysis and processing.

Saving refined dataset

```
write.table(titanicMessData,
    file = "TitanicCleaned.tsv",
    sep = "\t",
    row.names=FALSE)
```