Post 1: Cleaning Data for Analysis

Soham Kudtarkar

Post 1: Cleaning Data for Analysis

Introduction

The aim of this Rmd Notebook is to apply data cleaning techniques in order to wrangle messy data into a useful format for analysis. Overall, the content of this post will be: (1) importing data and doing exploration to see what it looks like, (2) combining data sets, (3) removing columns not as relevant to analysis, (4) restructuring qualitative entries to quantitative entries, and (5) Removing all incomplete entries. All references are included at the bottom of the notebook.

Cleaning Data

We will be analyzing data from the Titanic data set [1]. This data records information on passengers on the famously sunken Titanic ship. Columns include, among others, the names of the passengers, whether they survived, whether they were male or female, and what their fare was.

Importing Data and Initial Exploration

We will begin by importing the `train.csv` and `test.csv` files. Both files are contained in the `data` folder.

```
train <- read.csv('../data/train.csv')
test <- read.csv('../data/test.csv')
```

We will then initially explore the data. Here is what some of the `train.csv` data looks like.

Here is what some of the `train.csv` data looks like [5].

```
head(train, 20)
```

```
##    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          14        0      3
## 15          15        0      3
## 16          16        1      2
## 17          17        0      3
## 18          18        1      2
## 19          19        0      3
## 20          20        1      3
##                                                        Name    Sex Age
## 1                                   Braund, Mr. Owen Harris   male  22
## 2       Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38
## 3                                    Heikkinen, Miss. Laina female  26
## 4              Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35
## 5                                  Allen, Mr. William Henry   male  35
## 6                                          Moran, Mr. James   male  NA
## 7                                   McCarthy, Mr. Timothy J   male  54
## 8                            Palsson, Master. Gosta Leonard   male   2
## 9         Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female  27
## 10                     Nasser, Mrs. Nicholas (Adele Achem) female  14
## 11                         Sandstrom, Miss. Marguerite Rut female   4
## 12                                Bonnell, Miss. Elizabeth female  58
## 13                          Saundercock, Mr. William Henry   male  20
## 14                             Andersson, Mr. Anders Johan   male  39
## 15                    Vestrom, Miss. Hulda Amanda Adolfina female  14
## 16                        Hewlett, Mrs. (Mary D Kingcome)  female  55
## 17                                    Rice, Master. Eugene   male   2
## 18                            Williams, Mr. Charles Eugene   male  NA
## 19 Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) female  31
## 20                                 Masselmani, Mrs. Fatima female  NA
##    SibSp Parch           Ticket    Fare Cabin Embarked
## 1      1     0        A/5 21171  7.2500              S
## 2      1     0         PC 17599 71.2833   C85        C
## 3      0     0 STON/O2. 3101282  7.9250              S
## 4      1     0           113803 53.1000  C123        S
## 5      0     0           373450  8.0500              S
## 6      0     0           330877  8.4583              Q
## 7      0     0            17463 51.8625   E46        S
## 8      3     1           349909 21.0750              S
## 9      0     2           347742 11.1333              S
## 10     1     0           237736 30.0708              C
## 11     1     1          PP 9549 16.7000    G6        S
## 12     0     0           113783 26.5500  C103        S
## 13     0     0        A/5. 2151  8.0500              S
## 14     1     5           347082 31.2750              S
## 15     0     0           350406  7.8542              S
## 16     0     0           248706 16.0000              S
## 17     4     1           382652 29.1250              Q
## 18     0     0           244373 13.0000              S
## 19     1     0           345763 18.0000              S
## 20     0     0             2649  7.2250              C
```

Here are some attributes on train [4].

```
ncol(train)
```

```
## [1] 12
```

```
nrow(train)
```

```
## [1] 891
```

Here is what some of the `test.csv` data looks like [5].

```
head(test, 20)
```

```
##    PassengerId Pclass
## 1          892      3
## 2          893      3
## 3          894      2
## 4          895      3
## 5          896      3
## 6          897      3
## 7          898      3
## 8          899      2
## 9          900      3
## 10         901      3
## 11         902      3
## 12         903      1
## 13         904      1
## 14         905      2
## 15         906      1
## 16         907      2
## 17         908      2
## 18         909      3
## 19         910      3
## 20         911      3
##                                                        Name    Sex  Age
## 1                                          Kelly, Mr. James   male 34.5
## 2                          Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3                                 Myles, Mr. Thomas Francis   male 62.0
## 4                                          Wirz, Mr. Albert   male 27.0
## 5              Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6                                Svensson, Mr. Johan Cervin   male 14.0
## 7                                      Connolly, Miss. Kate female 30.0
## 8                              Caldwell, Mr. Albert Francis   male 26.0
## 9                 Abrahim, Mrs. Joseph (Sophie Halaut Easu) female 18.0
## 10                                  Davies, Mr. John Samuel   male 21.0
## 11                                         Ilieff, Mr. Ylio   male   NA
## 12                               Jones, Mr. Charles Cresson   male 46.0
## 13           Snyder, Mrs. John Pillsbury (Nelle Stevenson) female 23.0
## 14                                     Howard, Mr. Benjamin   male 63.0
## 15 Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood) female 47.0
## 16           del Carlo, Mrs. Sebastiano (Argenia Genovesi) female 24.0
## 17                                        Keane, Mr. Daniel   male 35.0
## 18                                        Assaf, Mr. Gerios   male 21.0
## 19                             Ilmakangas, Miss. Ida Livija female 27.0
## 20                   Assaf Khalil, Mrs. Mariana (Miriam")" female 45.0
##    SibSp Parch           Ticket    Fare Cabin Embarked
## 1      0     0           330911  7.8292              Q
## 2      1     0           363272  7.0000              S
## 3      0     0           240276  9.6875              Q
## 4      0     0           315154  8.6625              S
## 5      1     1          3101298 12.2875              S
## 6      0     0             7538  9.2250              S
## 7      0     0           330972  7.6292              Q
## 8      1     1           248738 29.0000              S
## 9      0     0             2657  7.2292              C
## 10     2     0        A/4 48871 24.1500              S
## 11     0     0           349220  7.8958              S
## 12     0     0              694 26.0000              S
## 13     1     0            21228 82.2667   B45        S
## 14     1     0            24065 26.0000              S
## 15     1     0      W.E.P. 5734 61.1750   E31        S
## 16     1     0    SC/PARIS 2167 27.7208              C
## 17     0     0           233734 12.3500              Q
## 18     0     0             2692  7.2250              C
## 19     1     0 STON/O2. 3101270  7.9250              S
## 20     0     0             2696  7.2250              C
```

Here are some attributes on test [4].

```
ncol(test)
```

```
## [1] 11
```

```
nrow(test)
```

```
## [1] 418
```

It appears that there are 12 columns and 891 rows in `train.csv` and 11 columns and 418 rows in `test.csv`. The missing column in test appears to be the Survived column. This makes sense as the source of the data is from a Kaggle competition page where the `test.csv` file is intended to be used to train a machine learning algorithm in order to predict the survivability of passengers in the `test.csv` table. Since predicting survivability is beyond the end goal of this post (and would be a task worthy of its own post entirely), we will ignore this column for our educational purposes.

Combining Data Sets

In this section, we will combine the train and test data sets to create a more complete source of data to analyze.

First, let us add a Survived column to the test table [2].

```
test$Survived <- rep(NA, 418)
```

Then, we will combine the tables vertically [3].

```
dat <- rbind(test, train)
head(dat, 20)
```

```
##    PassengerId Pclass
## 1          892      3
## 2          893      3
## 3          894      2
## 4          895      3
## 5          896      3
## 6          897      3
## 7          898      3
## 8          899      2
## 9          900      3
## 10         901      3
## 11         902      3
## 12         903      1
## 13         904      1
## 14         905      2
## 15         906      1
## 16         907      2
## 17         908      2
## 18         909      3
## 19         910      3
## 20         911      3
##                                                       Name    Sex  Age
## 1                                         Kelly, Mr. James   male 34.5
## 2                         Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3                                Myles, Mr. Thomas Francis   male 62.0
## 4                                         Wirz, Mr. Albert   male 27.0
## 5             Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6                               Svensson, Mr. Johan Cervin   male 14.0
## 7                                     Connolly, Miss. Kate female 30.0
## 8                             Caldwell, Mr. Albert Francis   male 26.0
## 9                Abrahim, Mrs. Joseph (Sophie Halaut Easu) female 18.0
## 10                                 Davies, Mr. John Samuel   male 21.0
## 11                                        Ilieff, Mr. Ylio   male   NA
## 12                              Jones, Mr. Charles Cresson   male 46.0
## 13           Snyder, Mrs. John Pillsbury (Nelle Stevenson) female 23.0
## 14                                    Howard, Mr. Benjamin   male 63.0
## 15 Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood) female 47.0
## 16           del Carlo, Mrs. Sebastiano (Argenia Genovesi) female 24.0
## 17                                       Keane, Mr. Daniel   male 35.0
## 18                                       Assaf, Mr. Gerios   male 21.0
## 19                            Ilmakangas, Miss. Ida Livija female 27.0
## 20                  Assaf Khalil, Mrs. Mariana (Miriam")" female 45.0
##    SibSp Parch           Ticket    Fare Cabin Embarked Survived
## 1      0     0           330911  7.8292              Q       NA
## 2      1     0           363272  7.0000              S       NA
## 3      0     0           240276  9.6875              Q       NA
## 4      0     0           315154  8.6625              S       NA
## 5      1     1          3101298 12.2875              S       NA
## 6      0     0             7538  9.2250              S       NA
## 7      0     0           330972  7.6292              Q       NA
## 8      1     1           248738 29.0000              S       NA
## 9      0     0             2657  7.2292              C       NA
## 10     2     0        A/4 48871 24.1500              S       NA
## 11     0     0           349220  7.8958              S       NA
## 12     0     0              694 26.0000              S       NA
## 13     1     0            21228 82.2667   B45        S       NA
## 14     1     0            24065 26.0000              S       NA
## 15     1     0      W.E.P. 5734 61.1750   E31        S       NA
## 16     1     0    SC/PARIS 2167 27.7208              C       NA
## 17     0     0           233734 12.3500              Q       NA
## 18     0     0             2692  7.2250              C       NA
## 19     1     0 STON/O2. 3101270  7.9250              S       NA
## 20     0     0             2696  7.2250              C       NA
```

```
ncol(dat)
```

```
## [1] 12
```

```
nrow(dat)
```

```
## [1] 1309
```

The dat table contains 12 rows and 1309 rows.

Removing Irrelevant Columns

Next, we will remove columns that will not be useful to us in our analysis.

As mentioned above, we will remove the Survived column [2].

```
dat$Survived <- NULL
head(dat, 20)
```

```
##    PassengerId Pclass
## 1          892      3
## 2          893      3
## 3          894      2
## 4          895      3
## 5          896      3
## 6          897      3
## 7          898      3
## 8          899      2
## 9          900      3
## 10         901      3
## 11         902      3
## 12         903      1
## 13         904      1
## 14         905      2
## 15         906      1
## 16         907      2
## 17         908      2
## 18         909      3
## 19         910      3
## 20         911      3
##                                                        Name    Sex  Age
## 1                                          Kelly, Mr. James   male 34.5
## 2                          Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3                                 Myles, Mr. Thomas Francis   male 62.0
## 4                                          Wirz, Mr. Albert   male 27.0
## 5             Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6                                Svensson, Mr. Johan Cervin   male 14.0
## 7                                      Connolly, Miss. Kate female 30.0
## 8                              Caldwell, Mr. Albert Francis   male 26.0
## 9                 Abrahim, Mrs. Joseph (Sophie Halaut Easu) female 18.0
## 10                                  Davies, Mr. John Samuel   male 21.0
## 11                                         Ilieff, Mr. Ylio   male   NA
## 12                               Jones, Mr. Charles Cresson   male 46.0
## 13           Snyder, Mrs. John Pillsbury (Nelle Stevenson) female 23.0
## 14                                     Howard, Mr. Benjamin   male 63.0
## 15 Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood) female 47.0
## 16           del Carlo, Mrs. Sebastiano (Argenia Genovesi) female 24.0
## 17                                        Keane, Mr. Daniel   male 35.0
## 18                                        Assaf, Mr. Gerios   male 21.0
## 19                             Ilmakangas, Miss. Ida Livija female 27.0
## 20                    Assaf Khalil, Mrs. Mariana (Miriam")" female 45.0
##    SibSp Parch           Ticket    Fare Cabin Embarked
## 1      0     0           330911  7.8292              Q
## 2      1     0           363272  7.0000              S
## 3      0     0           240276  9.6875              Q
## 4      0     0           315154  8.6625              S
## 5      1     1          3101298 12.2875              S
## 6      0     0             7538  9.2250              S
## 7      0     0           330972  7.6292              Q
## 8      1     1           248738 29.0000              S
## 9      0     0             2657  7.2292              C
## 10     2     0        A/4 48871 24.1500              S
## 11     0     0           349220  7.8958              S
## 12     0     0              694 26.0000              S
## 13     1     0            21228 82.2667   B45        S
## 14     1     0            24065 26.0000              S
## 15     1     0      W.E.P. 5734 61.1750   E31        S
## 16     1     0    SC/PARIS 2167 27.7208              C
## 17     0     0           233734 12.3500              Q
## 18     0     0             2692  7.2250              C
## 19     1     0 STON/O2. 3101270  7.9250              S
## 20     0     0             2696  7.2250              C
```

The PassengerId column is not really necessary since it does not give us any new information on the passengers.

```
dat$PassengerId <- NULL
head(dat, 20)
```

```
##    Pclass                                                    Name    Sex
## 1       3                                        Kelly, Mr. James   male
## 2       3                        Wilkes, Mrs. James (Ellen Needs) female
## 3       3                               Myles, Mr. Thomas Francis   male
## 4       3                                        Wirz, Mr. Albert   male
## 5       3            Hirvonen, Mrs. Alexander (Helga E Lindqvist) female
## 6       3                              Svensson, Mr. Johan Cervin   male
## 7       3                                    Connolly, Miss. Kate female
## 8       2                            Caldwell, Mr. Albert Francis   male
## 9       3               Abrahim, Mrs. Joseph (Sophie Halaut Easu) female
## 10      3                                 Davies, Mr. John Samuel   male
## 11      3                                        Ilieff, Mr. Ylio   male
## 12      1                              Jones, Mr. Charles Cresson   male
## 13      1           Snyder, Mrs. John Pillsbury (Nelle Stevenson) female
## 14      2                                    Howard, Mr. Benjamin   male
## 15      1 Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood) female
## 16      2           del Carlo, Mrs. Sebastiano (Argenia Genovesi) female
## 17      2                                       Keane, Mr. Daniel   male
## 18      3                                       Assaf, Mr. Gerios   male
## 19      3                            Ilmakangas, Miss. Ida Livija female
## 20      3                   Assaf Khalil, Mrs. Mariana (Miriam")" female
##     Age SibSp Parch           Ticket    Fare Cabin Embarked
## 1  34.5     0     0           330911  7.8292              Q
## 2  47.0     1     0           363272  7.0000              S
## 3  62.0     0     0           240276  9.6875              Q
## 4  27.0     0     0           315154  8.6625              S
## 5  22.0     1     1          3101298 12.2875              S
## 6  14.0     0     0             7538  9.2250              S
## 7  30.0     0     0           330972  7.6292              Q
## 8  26.0     1     1           248738 29.0000              S
## 9  18.0     0     0             2657  7.2292              C
## 10 21.0     2     0        A/4 48871 24.1500              S
## 11   NA     0     0           349220  7.8958              S
## 12 46.0     0     0              694 26.0000              S
## 13 23.0     1     0            21228 82.2667   B45        S
## 14 63.0     1     0            24065 26.0000              S
## 15 47.0     1     0      W.E.P. 5734 61.1750   E31        S
## 16 24.0     1     0    SC/PARIS 2167 27.7208              C
## 17 35.0     0     0           233734 12.3500              Q
## 18 21.0     0     0             2692  7.2250              C
## 19 27.0     1     0 STON/O2. 3101270  7.9250              S
## 20 45.0     0     0             2696  7.2250              C
```

Similar logic could be applied to the Name and Ticket columns.

```
dat$Name <- NULL
dat$Ticket <- NULL
head(dat, 20)
```

```
##    Pclass    Sex  Age SibSp Parch    Fare Cabin Embarked
## 1       3   male 34.5     0     0  7.8292              Q
## 2       3 female 47.0     1     0  7.0000              S
## 3       2   male 62.0     0     0  9.6875              Q
## 4       3   male 27.0     0     0  8.6625              S
## 5       3 female 22.0     1     1 12.2875              S
## 6       3   male 14.0     0     0  9.2250              S
## 7       3 female 30.0     0     0  7.6292              Q
## 8       2   male 26.0     1     1 29.0000              S
## 9       3 female 18.0     0     0  7.2292              C
## 10      3   male 21.0     2     0 24.1500              S
## 11      3   male   NA     0     0  7.8958              S
## 12      1   male 46.0     0     0 26.0000              S
## 13      1 female 23.0     1     0 82.2667   B45        S
## 14      2   male 63.0     1     0 26.0000              S
## 15      1 female 47.0     1     0 61.1750   E31        S
## 16      2 female 24.0     1     0 27.7208              C
## 17      2   male 35.0     0     0 12.3500              Q
## 18      3   male 21.0     0     0  7.2250              C
## 19      3 female 27.0     1     0  7.9250              S
## 20      3 female 45.0     0     0  7.2250              C
```

The Cabin column is fairly incomplete and includes many complexities. We will remove this column as well so as to not complicate our analysis with the complexity of the column.

```
dat$Cabin <- NULL
head(dat, 100)
```

```
##     Pclass    Sex  Age SibSp Parch     Fare Embarked
## 1        3   male 34.5     0     0   7.8292        Q
## 2        3 female 47.0     1     0   7.0000        S
## 3        2   male 62.0     0     0   9.6875        Q
## 4        3   male 27.0     0     0   8.6625        S
## 5        3 female 22.0     1     1  12.2875        S
## 6        3   male 14.0     0     0   9.2250         S
## 7        3 female 30.0     0     0   7.6292         Q
## 8        2   male 26.0     1     1  29.0000         S
## 9        3 female 18.0     0     0   7.2292         C
## 10       3   male 21.0     2     0  24.1500         S
## 11       3   male   NA     0     0   7.8958         S
## 12       1   male 46.0     0     0  26.0000         S
## 13       1 female 23.0     1     0  82.2667         S
## 14       2   male 63.0     1     0  26.0000         S
## 15       1 female 47.0     1     0  61.1750         S
## 16       2 female 24.0     1     0  27.7208         C
## 17       2   male 35.0     0     0  12.3500         Q
## 18       3   male 21.0     0     0   7.2250         C
## 19       3 female 27.0     1     0   7.9250         S
## 20       3 female 45.0     0     0   7.2250         C
## 21       1   male 55.0     1     0  59.4000         C
## 22       3   male  9.0     0     1   3.1708         S
## 23       1 female   NA     0     0  31.6833         S
## 24       1   male 21.0     0     1  61.3792         C
## 25       1 female 48.0     1     3 262.3750         C
## 26       3   male 50.0     1     0  14.5000         S
## 27       1 female 22.0     0     1  61.9792         C
## 28       3   male 22.5     0     0   7.2250         C
## 29       1   male 41.0     0     0  30.5000         S
## 30       3   male   NA     2     0  21.6792         C
## 31       2   male 50.0     1     0  26.0000         S
## 32       2   male 24.0     2     0  31.5000         S
## 33       3 female 33.0     1     2  20.5750         S
## 34       3 female   NA     1     2  23.4500         S
## 35       1   male 30.0     1     0  57.7500         C
## 36       3   male 18.5     0     0   7.2292         C
## 37       3 female   NA     0     0   8.0500         S
## 38       3 female 21.0     0     0   8.6625         S
## 39       3   male 25.0     0     0   9.5000         S
## 40       3   male   NA     0     0  56.4958         S
## 41       3   male 39.0     0     1  13.4167         C
## 42       1   male   NA     0     0  26.5500         S
## 43       3   male 41.0     0     0   7.8500         S
## 44       2 female 30.0     0     0  13.0000         S
## 45       1 female 45.0     1     0  52.5542         S
## 46       3   male 25.0     0     0   7.9250         S
## 47       1   male 45.0     0     0  29.7000         C
## 48       3   male   NA     0     0   7.7500         Q
## 49       1 female 60.0     0     0  76.2917         C
## 50       3 female 36.0     0     2  15.9000         S
## 51       1   male 24.0     1     0  60.0000         S
## 52       2   male 27.0     0     0  15.0333         C
## 53       2 female 20.0     2     1  23.0000         S
## 54       1 female 28.0     3     2 263.0000         S
## 55       2   male   NA     0     0  15.5792         C
## 56       3   male 10.0     4     1  29.1250         Q
## 57       3   male 35.0     0     0   7.8958         S
## 58       3   male 25.0     0     0   7.6500         S
## 59       3   male   NA     1     0  16.1000         S
## 60       1 female 36.0     0     0 262.3750         C
## 61       3   male 17.0     0     0   7.8958         S
## 62       2   male 32.0     0     0  13.5000         S
## 63       3   male 18.0     0     0   7.7500         S
## 64       3 female 22.0     0     0   7.7250         Q
## 65       1   male 13.0     2     2 262.3750         C
## 66       2 female   NA     0     0  21.0000         S
## 67       3 female 18.0     0     0   7.8792         Q
## 68       1   male 47.0     0     0  42.4000         S
## 69       1   male 31.0     0     0  28.5375         C
## 70       1 female 60.0     1     4 263.0000         S
## 71       3 female 24.0     0     0   7.7500         Q
## 72       3   male 21.0     0     0   7.8958         S
## 73       3 female 29.0     0     0   7.9250         S
## 74       1   male 28.5     0     0  27.7208         C
## 75       1 female 35.0     0     0 211.5000         C
## 76       1   male 32.5     0     0 211.5000         C
## 77       3   male   NA     0     0   8.0500         S
## 78       1 female 55.0     2     0  25.7000         S
## 79       2   male 30.0     0     0  13.0000         S
## 80       3 female 24.0     0     0   7.7500         Q
## 81       3   male  6.0     1     1  15.2458         C
## 82       1   male 67.0     1     0 221.7792         S
## 83       1   male 49.0     0     0  26.0000         S
## 84       3   male   NA     0     0   7.8958         S
## 85       2   male   NA     0     0  10.7083         Q
## 86       3   male   NA     1     0  14.4542         C
## 87       3 female 27.0     0     0   7.8792         Q
## 88       3 female 18.0     0     0   8.0500         S
## 89       3 female   NA     0     0   7.7500         Q
## 90       2   male  2.0     1     1  23.0000         S
## 91        3 female 22.0     1     0  13.9000         S
## 92        3   male   NA     0     0   7.7750         S
## 93        1 female 27.0     1     2  52.0000         S
## 94        3   male   NA     0     0   8.0500         S
## 95        1   male 25.0     0     0  26.0000         C
## 96        3   male 25.0     0     0   7.7958         S
## 97        1 female 76.0     1     0  78.8500         S
## 98        3   male 29.0     0     0   7.9250         S
## 99        3 female 20.0     0     0   7.8542         S
## 100       3   male 33.0     0     0   8.0500         S
```

Finally, we will also remove the Fare column as the Pclass column already captures highly similar information, given that the two are related (i.e. if a passenger spent more money on the fare, they were likely in the higher class).

```
dat$Fare <- NULL
head(dat, 100)
```

```
##     Pclass    Sex  Age SibSp Parch Embarked
## 1        3   male 34.5     0     0        Q
## 2        3 female 47.0     1     0        S
## 3        2   male 62.0     0     0        Q
## 4        3   male 27.0     0     0        S
## 5        3 female 22.0     1     1        S
## 6        3   male 14.0     0     0        S
## 7        3 female 30.0     0     0        Q
## 8        2   male 26.0     1     1        S
## 9        3 female 18.0     0     0        C
## 10       3   male 21.0     2     0        S
## 11       3   male   NA     0     0        S
## 12       1   male 46.0     0     0        S
## 13       1 female 23.0     1     0        S
## 14       2   male 63.0     1     0        S
## 15       1 female 47.0     1     0        S
## 16       2 female 24.0     1     0        C
## 17       2   male 35.0     0     0        Q
## 18       3   male 21.0     0     0        C
## 19       3 female 27.0     1     0        S
## 20       3 female 45.0     0     0        C
## 21       1   male 55.0     1     0        C
## 22       3   male  9.0     0     1        S
## 23       1 female   NA     0     0        S
## 24       1   male 21.0     0     1        C
## 25       1 female 48.0     1     3        C
## 26       3   male 50.0     1     0        S
## 27       1 female 22.0     0     1        C
## 28       3   male 22.5     0     0        C
## 29       1   male 41.0     0     0        S
## 30       3   male   NA     2     0        C
## 31       2   male 50.0     1     0        S
## 32       2   male 24.0     2     0        S
## 33       3 female 33.0     1     2        S
## 34       3 female   NA     1     2        S
## 35       1   male 30.0     1     0        C
## 36       3   male 18.5     0     0        C
## 37       3 female   NA     0     0        S
## 38       3 female 21.0     0     0        S
## 39       3   male 25.0     0     0        S
## 40       3   male   NA     0     0        S
## 41       3   male 39.0     0     1        C
## 42       1   male   NA     0     0        S
## 43       3   male 41.0     0     0        S
## 44       2 female 30.0     0     0        S
## 45       1 female 45.0     1     0        S
## 46       3   male 25.0     0     0        S
## 47       1   male 45.0     0     0        C
## 48       3   male   NA     0     0        Q
## 49       1 female 60.0     0     0        C
## 50       3 female 36.0     0     2        S
## 51       1   male 24.0     1     0        S
## 52       2   male 27.0     0     0        C
## 53       2 female 20.0     2     1        S
## 54       1 female 28.0     3     2        S
## 55       2   male   NA     0     0        C
## 56       3   male 10.0     4     1        Q
## 57       3   male 35.0     0     0        S
## 58       3   male 25.0     0     0        S
## 59       3   male   NA     1     0        S
## 60       1 female 36.0     0     0        C
## 61       3   male 17.0     0     0        S
## 62       2   male 32.0     0     0        S
## 63       3   male 18.0     0     0        S
## 64       3 female 22.0     0     0        Q
## 65       1   male 13.0     2     2        C
## 66       2 female   NA     0     0        S
## 67       3 female 18.0     0     0        Q
## 68       1   male 47.0     0     0        S
## 69       1   male 31.0     0     0        C
## 70       1 female 60.0     1     4        S
## 71       3 female 24.0     0     0        Q
## 72       3   male 21.0     0     0        S
## 73       3 female 29.0     0     0        S
## 74       1   male 28.5     0     0        C
## 75       1 female 35.0     0     0        C
## 76       1   male 32.5     0     0        C
## 77       3   male   NA     0     0        S
## 78       1 female 55.0     2     0        S
## 79       2   male 30.0     0     0        S
## 80       3 female 24.0     0     0        Q
## 81       3   male  6.0     1     1        C
## 82       1   male 67.0     1     0        S
## 83       1   male 49.0     0     0        S
## 84       3   male   NA     0     0        S
## 85       2   male   NA     0     0        Q
## 86       3   male   NA     1     0        C
## 87       3 female 27.0     0     0        Q
## 88       3 female 18.0     0     0        S
## 89       3 female   NA     0     0        Q
## 90       2   male  2.0     1     1        S
## 91       3 female 22.0     1     0        S
## 92       3   male   NA     0     0        S
## 93       1 female 27.0     1     2        S
## 94       3   male   NA     0     0        S
## 95       1   male 25.0     0     0        C
## 96       3   male 25.0     0     0        S
## 97       1 female 76.0     1     0        S
## 98       3   male 29.0     0     0        S
## 99       3 female 20.0     0     0        S
## 100      3   male 33.0     0     0        S
```

We are finally left with 6 columns: Pclass, Sex, Age, SibSp, Parch, Embarked. This simplifies the complexity of our data.

Restructuring Qualitative Entries to Quantitative Entries

One further step in making the data useful is to restructure qualitative entries to quantitative entries. This makes analyses easier to create.

We will begin by changing the Sex column to numeric values [6].

```
dat$Sex <- as.numeric(factor(dat$Sex, levels = c('male', 'female')))
head(dat, 20)
```

```
##    Pclass Sex  Age SibSp Parch Embarked
## 1       3   1 34.5     0     0        Q
## 2       3   2 47.0     1     0        S
## 3       2   1 62.0     0     0        Q
## 4       3   1 27.0     0     0        S
## 5       3   2 22.0     1     1        S
## 6       3   1 14.0     0     0        S
## 7       3   2 30.0     0     0        Q
## 8       2   1 26.0     1     1        S
## 9       3   2 18.0     0     0        C
## 10      3   1 21.0     2     0        S
## 11      3   1   NA     0     0        S
## 12      1   1 46.0     0     0        S
## 13      1   2 23.0     1     0        S
## 14      2   1 63.0     1     0        S
## 15      1   2 47.0     1     0        S
## 16      2   2 24.0     1     0        C
## 17      2   1 35.0     0     0        Q
## 18      3   1 21.0     0     0        C
## 19      3   2 27.0     1     0        S
## 20      3   2 45.0     0     0        C
```

We will do the same for the Embarked column [6].

```
dat$Embarked <- as.numeric(factor(dat$Embarked, levels = c('Q', 'S', 'C')))
head(dat, 20)
```

```
##    Pclass Sex  Age SibSp Parch Embarked
## 1       3   1 34.5     0     0        1
## 2       3   2 47.0     1     0        2
## 3       2   1 62.0     0     0        1
## 4       3   1 27.0     0     0        2
## 5       3   2 22.0     1     1        2
## 6       3   1 14.0     0     0        2
## 7       3   2 30.0     0     0        1
## 8       2   1 26.0     1     1        2
## 9       3   2 18.0     0     0        3
## 10      3   1 21.0     2     0        2
## 11      3   1   NA     0     0        2
## 12      1   1 46.0     0     0        2
## 13      1   2 23.0     1     0        2
## 14      2   1 63.0     1     0        2
## 15      1   2 47.0     1     0        2
## 16      2   2 24.0     1     0        3
## 17      2   1 35.0     0     0        1
## 18      3   1 21.0     0     0        3
## 19      3   2 27.0     1     0        2
## 20      3   2 45.0     0     0        3
```

We now have a numerically populated table.

Removing All Incomplete Entries

In this final section, we will remove all rows that contain NA values.

First, we will check if this removal is necessary by examininy if entries with NA exist in the dat table.

```
any(is.na(dat))
```

```
## [1] TRUE
```

It appears that such entries exist.

```
nrow(dat)
```

```
## [1] 1309
```

The dat table currently has 1309 rows.

Now, let us remove the entries that contain NA.

```
dat <- na.omit(dat)
head(dat, 20)
```

```
##    Pclass Sex  Age SibSp Parch Embarked
## 1       3   1 34.5     0     0        1
## 2       3   2 47.0     1     0        2
## 3       2   1 62.0     0     0        1
## 4       3   1 27.0     0     0        2
## 5       3   2 22.0     1     1        2
## 6       3   1 14.0     0     0        2
## 7       3   2 30.0     0     0        1
## 8       2   1 26.0     1     1        2
## 9       3   2 18.0     0     0        3
## 10      3   1 21.0     2     0        2
## 12      1   1 46.0     0     0        2
## 13      1   2 23.0     1     0        2
## 14      2   1 63.0     1     0        2
## 15      1   2 47.0     1     0        2
## 16      2   2 24.0     1     0        3
## 17      2   1 35.0     0     0        1
## 18      3   1 21.0     0     0        3
## 19      3   2 27.0     1     0        2
## 20      3   2 45.0     0     0        3
## 21      1   1 55.0     1     0        3
```

```
nrow(dat)
```

```
## [1] 1044
```

The dat table now contains only 1044 columns.

Conclusion

We have successfully created a clean table ready for analysis. We began with multiple tables with a maximum of 12 columns and a sum of 1309 rows to a clean table containing only numeric entries and no NA entries with 6 columns and 1044 rows.

References

[1] Source of data–both the `train.csv` and `test.csv` files were downloaded from Kaggle. https://www.kaggle.com/c/titanic/data.

[2] Adding a new column. Deleting a column. https://github.com/ucb-stat133/stat133-fall-2017/blob/master/labs/lab04-data-frame-basics.Rmd.

[3] Combining tables vertically. https://www.statmethods.net/management/merging.html.

[4] Finding the number of columns and rows. http://stat.ethz.ch/R-manual/R-devel/library/base/html/nrow.html.

[5] Finding the first 20 entries of a table. https://stackoverflow.com/questions/2667673/select-first-4-rows-of-a-data-frame-in-r.

[6] Converting a string factor to numeric values. https://stackoverflow.com/questions/23103223/converting-factors-to-numeric-values-in-r.

[7] Finding if there exist NA entries in a table. https://stackoverflow.com/questions/6551825/r-fastest-way-to-detect-if-vector-has-at-least-1-na.

[8] Removing all entries with NA. https://stackoverflow.com/questions/4862178/remove-rows-with-nas-missing-values-in-data-frame.