STAT 509 2017 Summer Course Project
Instructor: Shiwen Shen

1 Motivation

According to Wikipedia, **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms. **Data scientist** has become a popular occupation with Harvard Business Review dubbing it "The Sexiest Job of the 21st Century" and *McKinsey* projecting a global excess demand of 1.5 million new data scientists.

Here I quote are some potential relationship between data science and statistics:

- A data scientist is a statistician who lives in San Francisco.
- Data Science is statistics on a Mac.
- A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistican.

The arriving of the personal computer (1965) changes the way how we live. Now, with some level of statistics knowledge, a college student is able to do data analysis, which is unimaginable even 20 years ago.

In this short semester, you have learnt many concepts, which are very important in statistics field. Additionally, you have taste the flavor of `R`, the best programming language for statisticians, which is used by data scientists every single day. Perhaps it is time for you to practice your data analysis skills!

2 Your Mission

A training data and testing data are given in the course webapge. You need to use the training data to make statistical inference about the testing data. See the detailed data description in the next section. You can use any methods you have learnt / will learn in this course, and you can consult your best friend, Google, to learn more sophisticated and powerful tools. Be creative and open-minded. By the end of the day, you need to **provide your estimation results and make a in-class presentation to demonstrate how you achieve your estimation**.

All graduate students are mandatory to attend the course project alone. Undergraduate students can choose either do the project or not, and for any undergraduate student, who intends to do the project, has an option to do it alone or with another student. That means a team with two students is allowed for undergraduate students. **Undergraduate students have to inform me your intention before the end of June 8.**

3 Data Background

Wikipedia: "**RMS Titanic** was a British passenger liner that sank in the North Atlantic Ocean in the early morning of 15 April 1912, after colliding with an iceberg during her maiden voyage from Southampton to New York City. Of the 2,224 passengers and crew aboard, more than 1,500 died, making it one of the deadliest maritime disasters in modern history."

Here are two YouTube videos containing memorable (very old) pictures and the process of the wreck of Titanic:

- `https://youtu.be/8wTlureUMP8`
- `https://youtu.be/9xoqXVjBEF8`

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

You will be provided with two datasets: training and testing. There are 791 passengers' information in the training data with multiple variables, and most importantly whether the passengers survived the disaster. Additionally, there are 100 passengers' information in the testing data, however, survival status is not given. Your job is to use the information in the training/testing data to estimate the survival status of those 100 passengers in the testing data.

4 Variable Descriptions

There are 11 variables in the training data and 11 variable in the testing data (exclude Survived). Here is a list of variable descriptions:

1. Survived: Survival status (1 = Yes; 0 = No).
2. Pclass: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd).
3. Name: Passenger name.
4. Sex: Passenger sex.
5. Age: Passenger age.
6. SibSp: Number of siblings/spouses aboard.
7. Parch: Number of parents/children aboard.
8. Ticket: Ticket number.
9. Fare: Passenger fare.
10. Cabin: Cabin number (if possible).
11. Embarked: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).

Here are some remarks:

- Pclass is a proxy for socio-economic status (1st = Upper; 2nd = Middle; 3rd = Lower).
- Age is in Years; Fractional if Age less than One (If the Age is Estimated, it is in the form xx.5).
- With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch:
  - Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
  - Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
  - Parent: Mother or Father of Passenger Aboard Titanic
  - Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

  Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

5 Data

Two datasets `traindata.csv` and `testdata.csv` can be downloaded from the course webpage. You can use Microsoft Excel to open it for a brief look. All your analysis should be done based on R, and the R code you can borrow to load two datasets is here:

```
trainurl <- "http://people.stat.sc.edu/sshen/courses/17smstat509/project/traindata.csv"
testurl <- "http://people.stat.sc.edu/sshen/courses/17smstat509/project/testdata.csv"
traindata <- read.csv(trainurl, header=TRUE)
testdata <- read.csv(testurl, header=TRUE)

traindata
names(traindata) ## show variable names of the traindata
dim(traindata) ## show the dimension of the traindata
summary(traindata)

testdata
names(testdata) ## show variable names of the testdata
dim(testdata) ## show the dimension of the testdata
summary(testdata)
```