Case Study I: Titanic Dataset Analysis

Aslihan Demirkaya

Get into groups of 3 and answer the following questions:

Before you begin, make sure that you have R and R Studio properly installed. Also make sure that you understand how to use the `knitr` package and R Markdown. You are asked to submit both the R Markdown file and its pdf output.

The dataset is a list of passengers. The second column of the dataset is the label for each person indicating whether that person survived (1) or did not survive (0).

Here is the Kaggle page with more information on the dataset: https://www.kaggle.com/c/titanic

STEPS TO FOLLOW:

1. On the corresponding Kaggle website, download the dataset. You will see `train.csv` and `test.csv`. Download the `train.csv` to your computer.
2. When you load it, name it as `titanic_data`.

QUESTIONS:

1. How many passengers are in our passenger list?
2. Are there any missing values?
3. What is the overall survival rate?
4. How many male passengers were onboard?
5. How many female passengers were onboard?
6. What is the overall survival rate of male passengers?
7. What is the overall survival rate of female passengers?
8. What is the average age of all passengers onboard?
   - How did you calculate the average age?
   - Note that some of the passengers do not have an age value. How did you deal with this?
9. What is the average age of passengers who survived?
10. What is the average age of passengers who did not survive?
11. How many passengers are in each of the three classes of service (e.g. First, Second, and Third?)
12. What is the survival rate for passengers in each of the three classes of service?
13. Based on your data analysis, what can you conclude?
14. Use visualization techniques for few variables of interest. Specially, visualize association of different variables with survival rate.