

Week 5 worksheet - RMS Titanic

On April 15, 1912, during her maiden voyage from Southampton to New York, the four-funnelled ocean liner *RMS Titanic* hit an iceberg and sank to the bottom of the ocean in what is now one of the most infamous maritime disasters in history. For aesthetic reasons, and because the ship was widely considered unsinkable, many lifeboats were removed before the launch. This unfortunate decision contributed to the death of 1502 out of 2224 passengers and crew.

The [Encyclopedia Titanica](#) maintain a database of the names of passengers on the Titanic, which includes other information such as age, sex, and whether the passenger survived. A subset of these data are available freely through [Kaggle](#), an online data repository that also hosts regular data science competitions. Extensive analyses of the Titanic data set have shown that, while luck played its part, some groups of passengers did appear to have a better chance of surviving than others.

The Titanic passenger dataset from Kaggle can be found in the `data` folder of the course materials, with the name `titanic.csv`. Let's use `pandas` to interrogate it!

Here is the key to the data:

```
| Variable | Definition | Key | |:-:-----|:-:-----|:-:-----| | | | | |
| PassengerID | Unique numeric identifier for the passenger | NA | |
| Survival | Survival | 0 = No, 1 = Yes | | Pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd | | Sex | Sex | NA | |
| Age | Age in years | NA | | SibSp | Number of siblings / spouses aboard the Titanic | NA | | Parch |
| Number of parents / children aboard the Titanic | NA | | Ticket | Ticket number | NA | | Fare | Passenger
| fare | NA | | Cabin | Cabin number | NA | | Embarked | Port of Embarkation | C = Cherbourg, Q =
| Queenstown, S = Southampton |
```

Notes:

`Pclass` : a proxy for socio-economic status

- 1st = Upper
- 2nd = Middle
- 3rd = Lower

`Age` : fractional if less than 1. If the age is estimated, it is in the form of `xx.5`

`SibSp` : The dataset defines family relations as follows:

- Sibling = brother, sister, stepbrother, stepsister
- Spouse = husband, wife (mistresses and fiancés were ignored)

`Parch` : The dataset defines family relations as follows:

- Parent = mother, father
- Child = daughter, son, stepdaughter, stepson
- Some children travelled only with a nanny, therefore parch=0 for them.

1. Import `pandas` and read `data/titanic.csv` into a `DataFrame`

In []:

2. Display the **first 6 rows of the DataFrame**

In []:

3. Display the **last 6 rows of the DataFrame**

In []:

4. Assign PassengerID as the index

In []:

5. Print out all of the available information for the passenger with 567 as their PassengerID

In []:

6. Generate some descriptive statistics for the numeric data

In []:

7. How many male and female passengers are there in this dataset? Use pandas to show this information in a simple bar chart.

In []:

8. What were the names of the oldest and youngest passengers?

In []:

9. How many of the 5 oldest female passengers traveled first class?

In []:

10. How many passengers embarked at Southampton?

In []:

11. Who paid the most for their ticket?

In []:

12. Of the top ten passengers who **paid the most for their ticket, how**

many survived?

In []:

13. Of the top ten passengers who ***paid the least*** for their ticket, how many survived?

In []:

14. Of all surviving passengers, how many were male and how many were female?

In []:

15. Make a boxplot describing the age distribution for males and females, showing the mean age as well as the median.

In []:

16. Get the information for all passengers with 'Frank' in their name

In []:

17. What was the surname of the largest family on board the Titanic?

In []:

18. What was the name, sex and age of the youngest passenger to embark at Cherbourg?

In []:

19. Who stayed in cabin D56 ? Can you find out anything interesting about this person on the [Encyclopedia Titanica](#)?

In []:

20. There was a Countess aboard the Titanic. What information can you find about her?

In []:

21. Make two new columns containing the first and last name for each passenger.

In []:

22. Make a bar chart showing the proportion of males and females who survived in each class

In []: