

An Exploratory Look Aboard the Titanic

Marshall McQuillen

Guiding Question

Which passengers will survive the sinking of the Titanic?

Secondary Questions

1. What characteristics separate those who survived from those who died?
2. What characteristics make someone more likely to survive?
3. How do different characteristics of passengers vary with one another?

Data Overview

Looking at the training data from a bird-eye view, there are 891 observations representing passengers and 12 variables. Since some of the variable names are a little cryptic, a description for each is provided below.

| Variable Name | Description |
|---------------|---|
| PassengerId | Unique identifier for each passenger |
| Survived | Binary; 1 = Survived & 0 = Died |
| Pclass | Socio-economic status; 1 = Upper, 2 = Middle & 3 = Lower |
| Name | Passenger Name |
| Sex | Male or Female |
| Age | Passenger Age |
| SibSp | Number of siblings or spouse aboard ship |
| Parch | Number of parents or children aboard ship |
| Ticket | Ticket Number |
| Fare | Amount paid for ticket |
| Cabin | Cabin number |
| Embarked | The town from which the passenger boarded the ship; C = Cherbourg, Q = Queenstown & S = Southampton |

First and foremost, by running `str(training)` on the data, it is apparent that the first entries in the Cabin and Embarked columns are empty strings, indicating that the data is probably not perfectly clean (no surprises there). Checking to see where any Null's might be, it becomes clear that there are in fact no nulls, and that these spaces were intentionally left empty. In addition to null values, all the NA's are in the Age, accounting for roughly 20% of the values in that column. Both of these will need to be imputed intelligently when the time to create a predictive model comes around.

In addition to the missing values, it is important to note that some of the discrete attributes have been read in as continuous variables such as Pclass, Sibsp and Parch. Since these variables actually represent discrete characteristics of each passenger, changing them to be non-continuous will allow a more representative analysis.

Table 2: Attribute Null & NA Counts

| | PassengerId | Survived | Pclass | Name | Sex | Age |
|------------|-------------|----------|--------|------|-----|-----|
| Null Count | 0 | 0 | 0 | 0 | 0 | 0 |
| NA Count | 0 | 0 | 0 | 0 | 0 | 177 |

Table 3: Attribute Null & NA Counts (continued)

| | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|------------|-------|-------|--------|------|-------|----------|
| Null Count | 0 | 0 | 0 | 0 | 0 | 0 |
| NA Count | 0 | 0 | 0 | 0 | 0 | 0 |

2 Bayesian Survival

2.2 Does Money Sink or Swim?

By creating a table with the Pclass (which, recall refers to the socioeconomic status (SES) of the passenger) and Survived variables, I can get a good sense of the number of passengers that lived and died, based on their SES. Simple summation and division returns the probabilities of a passenger living given their respective SES

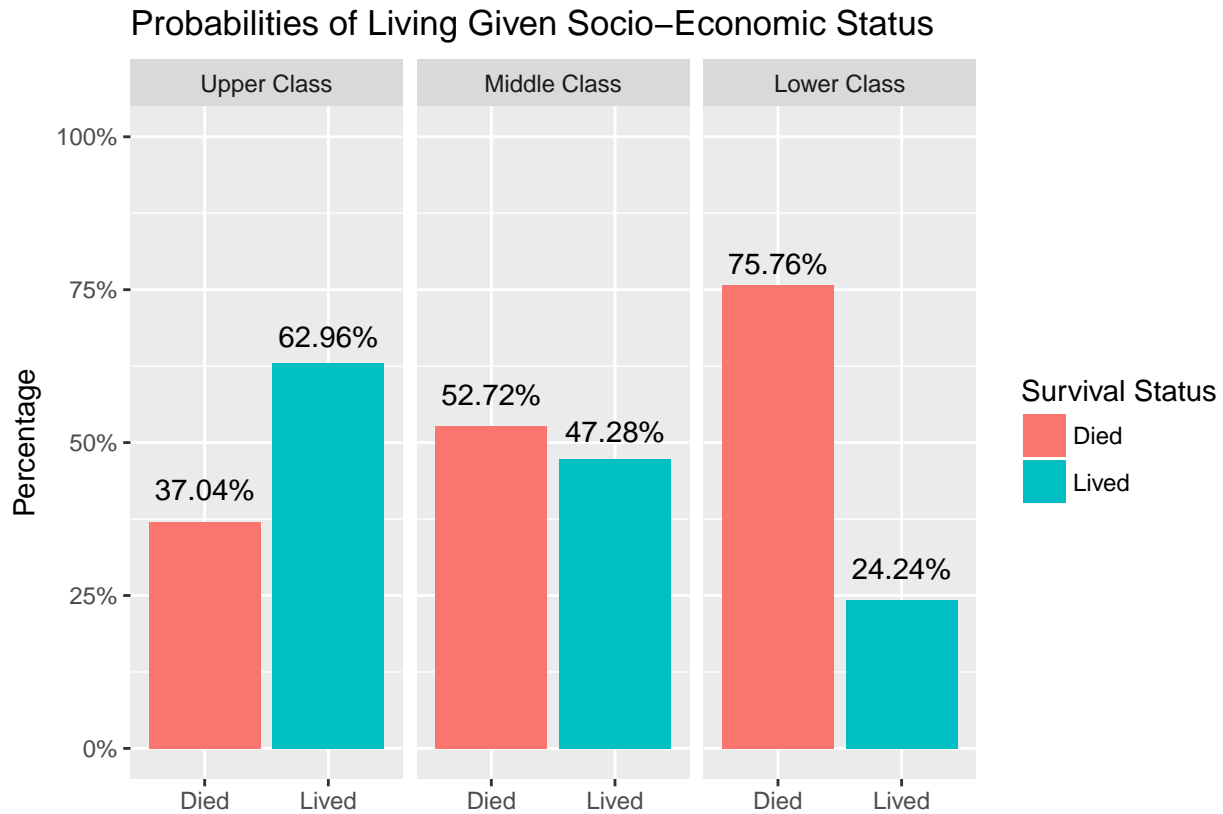
Table 4: Survival Counts by SES

| | Upper | Middle | Lower |
|----------|-------|--------|-------|
| Died | 80 | 97 | 372 |
| Survived | 136 | 87 | 119 |

Table 5: Survival Rates by SES

| | Probability of Surviving |
|--------------|--------------------------|
| Upper Class | 62.96% |
| Middle Class | 47.28% |
| Lower Class | 24.24% |

This same information is displayed visually below.



Illustrating Bayes Theorem with Survival Rates and Socio-Economic Status

This type of classification problem creates a great opportunity to illustrate Bayes' Theorem. Recall that Bayes Theorem is defined as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

...I can set **P(A)** = the probability of a passenger belonging to a defined SES (**X**) and **P(B)** = the probability of a passenger living. I can now rewrite the previously defined Theorem using my definitions as:

$$P(\text{"X class citizen"} | \text{"Lived"}) = \frac{P(\text{"Lived"} | \text{"X class citizen"}) P(\text{"X class citizen"})}{P(\text{"Lived"})}$$

The same information can be displayed visually as follows.

Now that I have found both **P("X class citizen")** (objects upper_prob, middle_prob and lower_prob) and **P("Lived")** (object prob_lived), and I have **P("Lived"|"X class citizen")** (objects upper_class, middle_class and lower_class), I can solve for **P("X class citizen"|"Lived")**...

$$P(\text{"Upper class citizen"} | \text{"Lived"}) = \frac{\text{upper_class} \times \text{upper_prob}}{\text{prob_lived}}$$

$$P(\text{"Middle class citizen"} | \text{"Lived"}) = \frac{\text{middle_class} \times \text{middle_prob}}{\text{prob_lived}}$$

$$P(\text{"Lower class citizen"} | \text{"Lived"}) = \frac{\text{lower_class} \times \text{lower_prob}}{\text{prob_lived}}$$

prop_lived

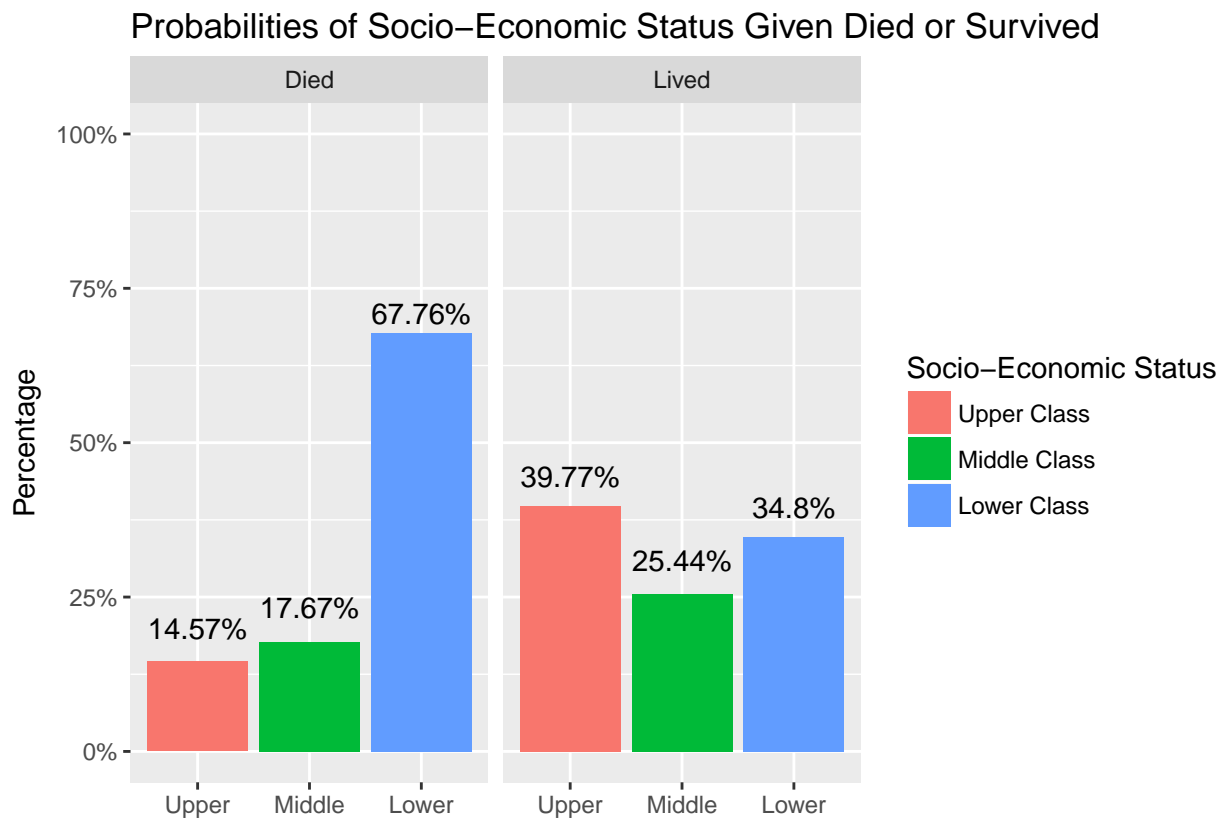
The “shorthand” way of finding these probabilities can be accomplished by dividing the the number of X class passengers that lived by the total number of passengers that lived using the pclass_table.

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

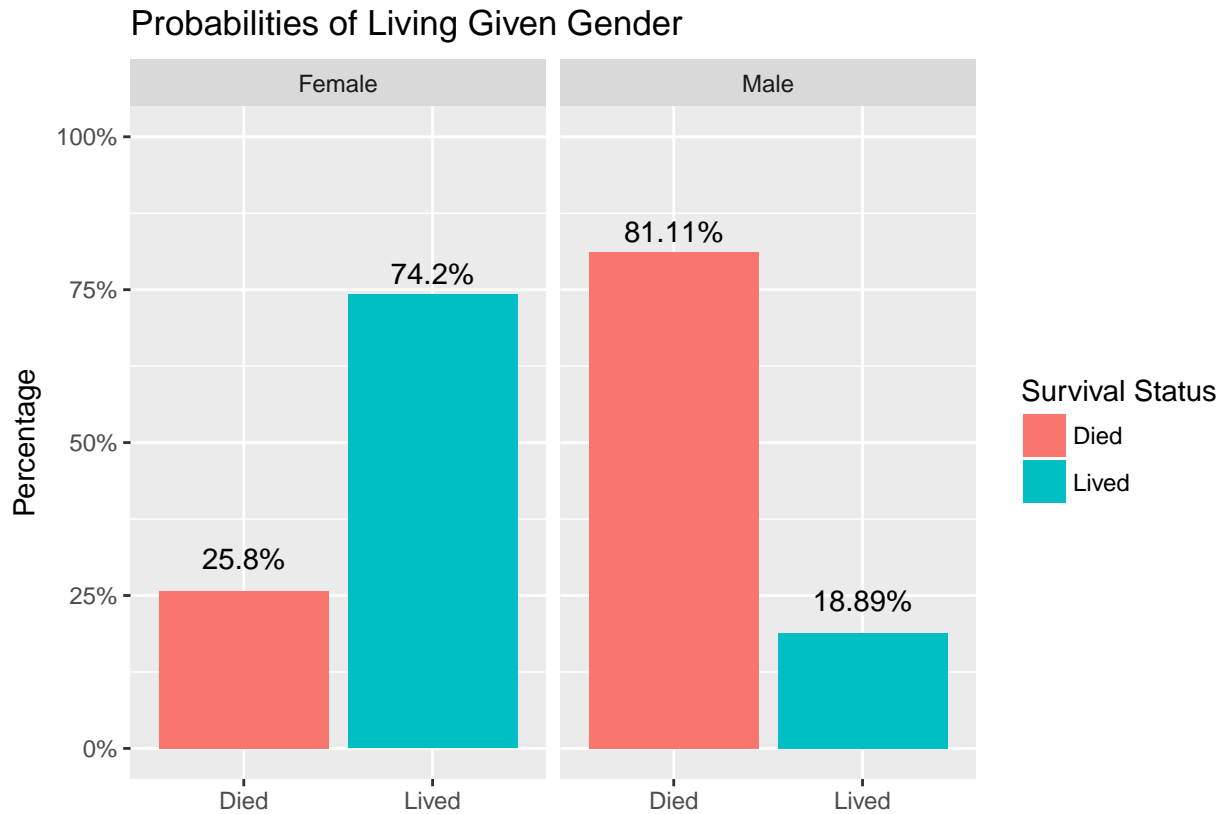
```
## [1] TRUE
```

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

A graphic illustrating said results.



Using the same code as above, with a few minor adjustments I can make similar graphs with other qualitative variables, such as the sex of the passenger.



3 Cabin Classification

It seems logical that looking at *where* each passenger was when the Titanic started sinking could provide some insight as to why some lived and others did not. The “Sinking” section on the Titanic Wikipedia Page states that the iceberg was struck at 11:40 pm. Considering the time of night, combined with the likely cold air temperature, I think it is safe to say that most passengers were inside, if not in their rooms sleeping.

Finding out where each passenger was will be a two fold process:

1. Subsetting on the Deck they were on, noted by the letter in the Cabin column.
2. Subsetting where on that deck they were, noted by the number in the Cabin column.

An important note is that the vast majority of the passengers did not have an entry in the Cabin column. (There aren’t any NA’s, the entries are not even filled with spaces, they are simply “nothing”). In order to subset these observations, I used the output from a “nothing” observation in the logical statement.

After subsetting, summing the number of rows in each subset, *which should equal 891, the total number of observations*, returns 894. A little searching led to finding the duplicates, show below.

```
## [1] 894

##      PassengerId Survived Pclass                                Name
## 76             76         0      3                      Moen, Mr. Sigurd Hansen
## 129            129         1      3                      Peter, Miss. Anna
## 700             700         0      3 Humblen, Mr. Adolf Mathias Nicolai Olsen
## 716             716         0      3 Soholt, Mr. Peter Andreas Lauritz Andersen
##      Sex Age SibSp Parch Ticket   Fare Cabin Embarked
## 76   male  25     0     0 348123  7.6500 F G73      S
## 129 female NA     1     1  2668 22.3583 F E69      C
## 700   male  42     0     0 348121  7.6500 F G63      S
```

```

## 716   male  19      0      0 348124  7.6500 F G73      S

##      PassengerId Survived Pclass
## 129          129         1      3
## 356          356         0      3
## 398          398         0      2
## 407          407         0      3
## 477          477         0      2
## 534          534         1      3
## 681          681         0      3
## 716          716         0      3
## 727          727         1      2
## 844          844         0      3
## 858          858         1      1
## 861          861         0      3

##                                     Name      Sex  Age SibSp Parch
## 129                                     Peter, Miss. Anna female   NA      1      1
## 356                                Vanden Steen, Mr. Leo Peter   male 28.0      0      0
## 398                                McKane, Mr. Peter David    male 46.0      0      0
## 407                                Widegren, Mr. Carl/Charles Peter   male 51.0      0      0
## 477                                Renouf, Mr. Peter Henry    male 34.0      1      0
## 534      Peter, Mrs. Catherine (Catherine Rizk) female   NA      0      2
## 681                                Peters, Miss. Katie female   NA      0      0
## 716      Soholt, Mr. Peter Andreas Lauritz Andersen    male 19.0      0      0
## 727      Renouf, Mrs. Peter Henry (Lillian Jefferys) female 30.0      3      0
## 844                                Lemberopolous, Mr. Peter L    male 34.5      0      0
## 858                                Daly, Mr. Peter Denis    male 51.0      0      0
## 861                                Hansen, Mr. Claus Peter    male 41.0      2      0

##      Ticket      Fare Cabin Embarked
## 129      2668 22.3583 F E69      C
## 356 345783  9.5000      S
## 398 28403 26.0000      S
## 407 347064  7.7500      S
## 477 31027 21.0000      S
## 534      2668 22.3583      C
## 681 330935  8.1375      Q
## 716 348124  7.6500 F G73      S
## 727 31027 21.0000      S
## 844      2683  6.4375      C
## 858 113055 26.5500  E17      S
## 861 350026 14.1083      S

```

To decide which subset to assign these observations too, looking at the Embarked and Ticket columns for those observations in the `g_class` subset, I can see that everyone in this cabin class embarked from Southampton and had similar ticket

```

##      1
## 0.4666667

##      1
## 0.7446809

##      1
## 0.5932203

##      1
## 0.7575758

```

```
##          1
## 0.7575758

##          1
## 0.6153846

##          1
## 0.2857143

##          1
## 0.2998544

##
##          0    1
## 0 445 233
## 1  53  65
## 2  40  40
## 3   2   3
## 4   4   0
## 5   4   1
## 6   1   0
```

