# Titanic: Exploratory Analysis

*Richard Kublik*

*April 3, 2017*

## Introduction and Setup

As an introductory data science project, I have chosen the Titanic: Machine Learning from Disaster competition hosted by Kaggle. As the goal of the competition is to build a machine learning model that will predict whether a passenger will survive, Kaggle has split the data set into 2 subsets. Both data sets contain information about the gender, travel class, age, etc. for each passenger. The training data also indicates if the passenger survived or not, while the test data set does not.

For this exploratory analysis, I am interested in learning about all of the passengers, and will be working with the combined data set, titanic_data:

```
# Merge datasets, initial pre-processing
titanic_data <- train_data %>%
  select(-Survived) %>%
  bind_rows(test_data) %>%
  mutate(Pclass = factor(Pclass),
         Sex = factor(Sex),
         Embarked = factor(Embarked),
         Title = factor(str_extract(Name, "[a-zA-z]+\\.")),
         age_est = 0)

# Convert variable names to lowercase
names(titanic_data) <- tolower(names(titanic_data))
```

Examining the structure, we find the variables:

```
str(titanic_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1309 obs. of  13 variables:
##  $ passengerid: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ sibsp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ cabin      : chr  NA "C85" NA "C123" ...
##  $ embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ title      : Factor w/ 18 levels "Capt.","Col.",..: 14 15 11 15 14 14 14 10 15 15 ...
##  $ age_est    : num  0 0 0 0 0 0 0 0 0 0 ...
```

While many of these are self-explanatory, we define the remainder below:

- **pclass:** The passengers ticket class, a proxy for socio-economic status:
  - 1: Upper
  - 2: Middle
  - 3: Lower

- **parch:** The number of parents/children travelling with the individual. Some children traveled with only a nanny, therefore parch = 0 for them.
- **sibsp:** The number of siblings/spouses travelling with the individual.
- **embarked:** The port of embarkation:
  - $C$: Cherbourg
  - $Q$: Queenstown
  - $S$: Southampton

## Dealing with Missing Data

In this case, the data is quite complete, with data missing for only a few variables:

```
sapply(titanic_data, function(df){sum(is.na(df))})
```

```
## passengerid      pclass        name         sex         age       sibsp
##           0           0           0           0         263           0
##       parch      ticket        fare       cabin    embarked       title
##           0           0           1        1014           2           0
##     age_est
##           0
```

We will impute values for the missing ages, fare, and embarkation location.

### Age

We begin by examining the number of missing ages for each title.

```
tb <- cbind(titanic_data$age, titanic_data$title)

# get the mean ages for each title
(age_dist <- titanic_data %>%
  group_by(title) %>%
  summarize(n = n(),
            n_missing = sum(is.na(age)),
            perc_missing = 100*n_missing/n,
            mean_age = mean(age, na.rm = TRUE),
            sd_age = sd(age, na.rm = TRUE)) %>%
  filter(n_missing > 0))
```

```
## # A tibble: 6 × 6
##     title     n n_missing perc_missing  mean_age    sd_age
##    <fctr> <int>     <int>        <dbl>     <dbl>     <dbl>
## 1     Dr.     8         1     12.50000 43.571429 11.731115
## 2 Master.    61         8     13.11475  5.482642  4.161554
## 3   Miss.   260        50     19.23077 21.774238 12.249077
## 4     Mr.   757       176     23.24967 32.252151 12.422089
## 5    Mrs.   197        27     13.70558 36.994118 12.901767
## 6     Ms.     2         1     50.00000 28.000000       NaN
```

In many cases, the number of missing values is substantial, and filling the missing values with the mean age for the title may significantly skew our results. We will assume that the ages for each title are normally distributed with the mean and standard deviation provided in the table.

```
for (key in c("Dr.", "Master.", "Miss.", "Mr.", "Mrs.")) {
  idx_na <- which(titanic_data$title == key & is.na(titanic_data$age))
```

```r
  age_idx <- which(age_dist$title == key)
  titanic_data$age[idx_na] <- rnorm(length(idx_na),
                                    age_dist$mean_age[age_idx],
                                    age_dist$sd_age[age_idx])
  titanic_data$age_est[idx_na] <- 1
}
# impute single missing Ms. value to be the mean:
idx_na <- which(titanic_data$title == "Ms." & is.na(titanic_data$age))
age_idx <- which(age_dist$title == "Ms.")
titanic_data$age[idx_na] <- age_dist$mean_age[age_idx]
titanic_data$age_est[idx_na] <- 1

titanic_data %>%
  group_by(title) %>%
  summarize(n = n(),
            mean_age = mean(age))
```

```
## # A tibble: 18 × 3
##        title     n  mean_age
##       <fctr> <int>     <dbl>
## 1      Capt.     1 70.000000
## 2       Col.     4 54.000000
## 3  Countess.     1 33.000000
## 4       Don.     1 40.000000
## 5      Dona.     1 39.000000
## 6        Dr.     8 46.884789
## 7  Jonkheer.     1 38.000000
## 8      Lady.     1 48.000000
## 9     Major.     2 48.500000
## 10   Master.    61  5.430484
## 11     Miss.   260 22.106866
## 12     Mlle.     2 24.000000
## 13      Mme.     1 24.000000
## 14       Mr.   757 31.948086
## 15      Mrs.   197 36.638373
## 16       Ms.     2 28.000000
## 17      Rev.     8 41.250000
## 18      Sir.     1 49.000000
```

### Fare

As there is only a single missing fare, we will impute the mean fare for that passenger's travel class.

```r
missing_fare <- titanic_data %>%
  filter(is.na(fare))

mean_fare <- titanic_data %>%
  filter(pclass == missing_fare$pclass) %>%
  summarize(mean_fare = mean(fare, na.rm = TRUE))

titanic_data$fare[which(titanic_data$passengerid == missing_fare$passengerid)] <- mean_fare$mean_fare[1
```

### Embarkation Port

Again, there are only a few missing values for port of embarkation, we will again impute these with the most common value, "S".

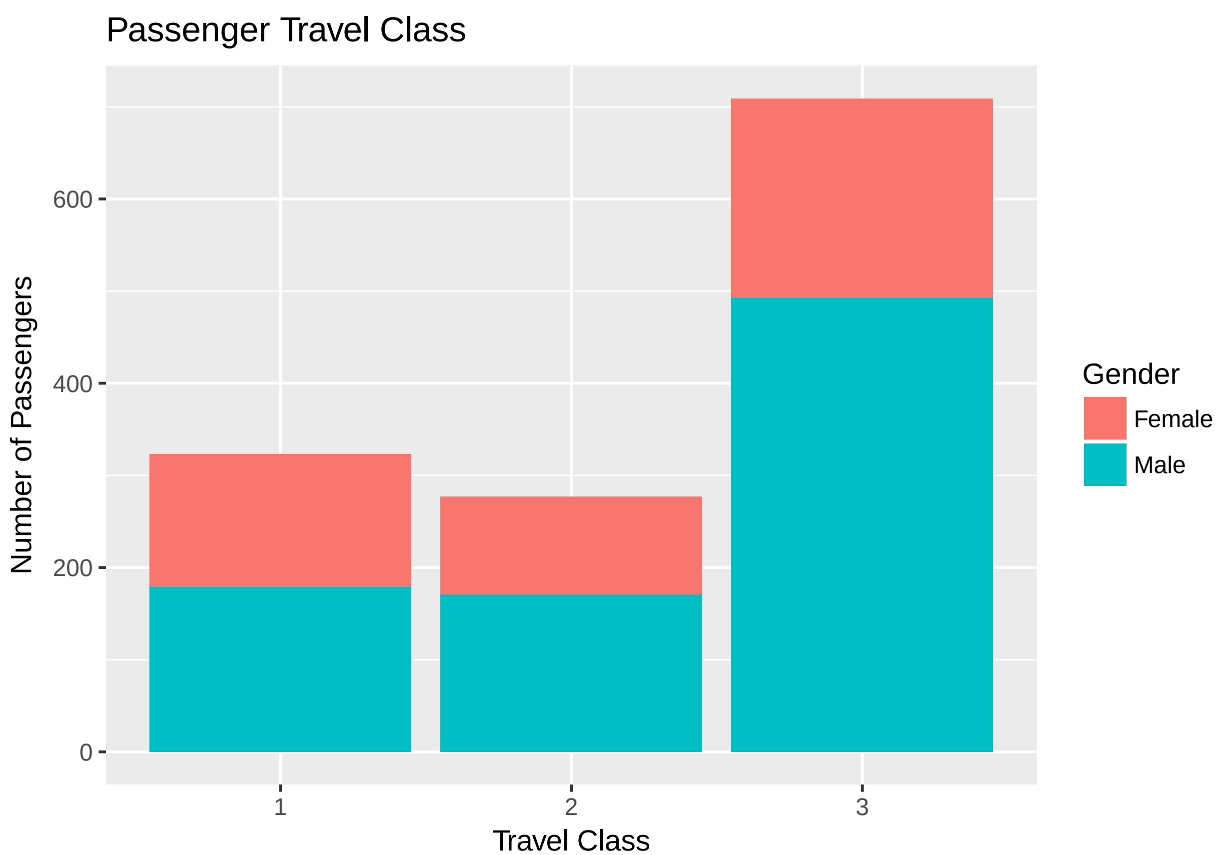
```
table(titanic_data$embarked, useNA = "always")
```

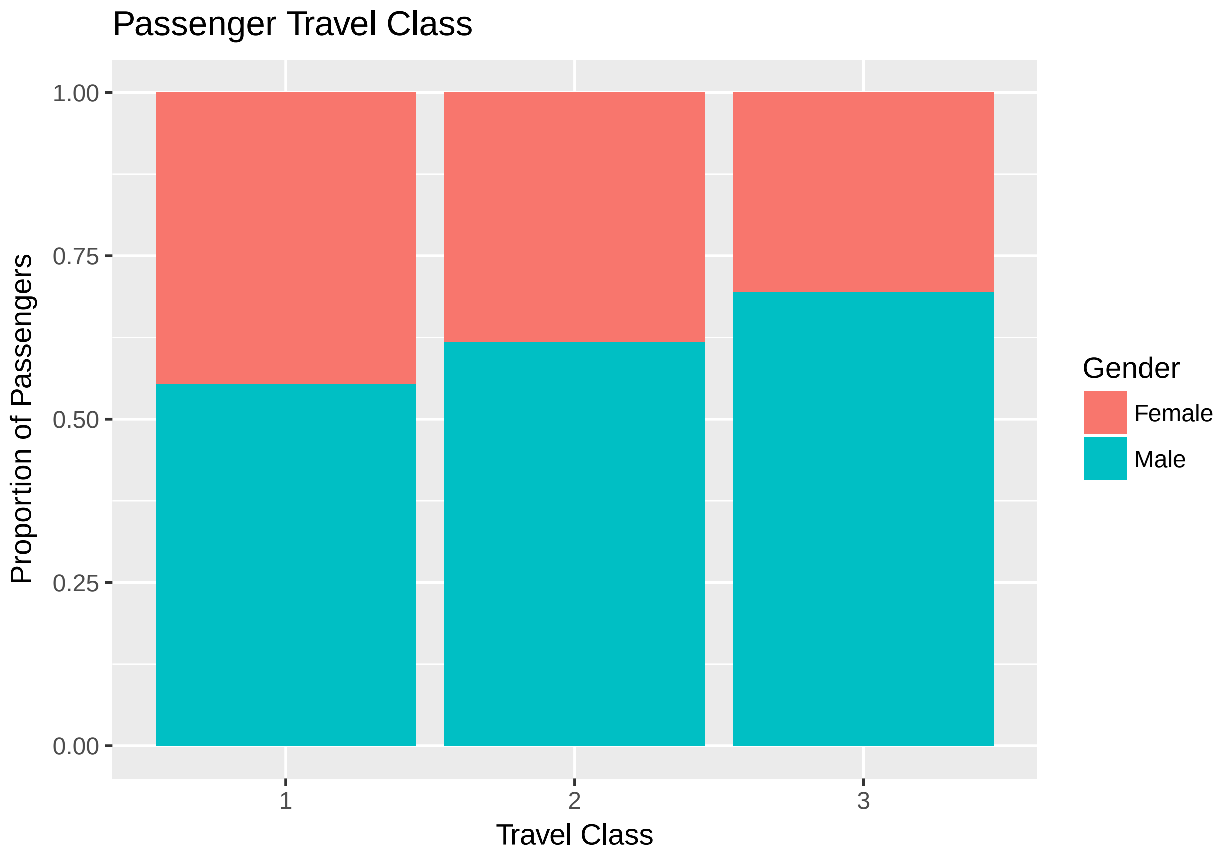
```
##
##    C    Q    S <NA>
##  270  123  914    2
```

```
# set missing data to S, as the most common.
titanic_data$embarked[which(is.na(titanic_data$embarked))] <- "S"
```

## Exploratory Analysis

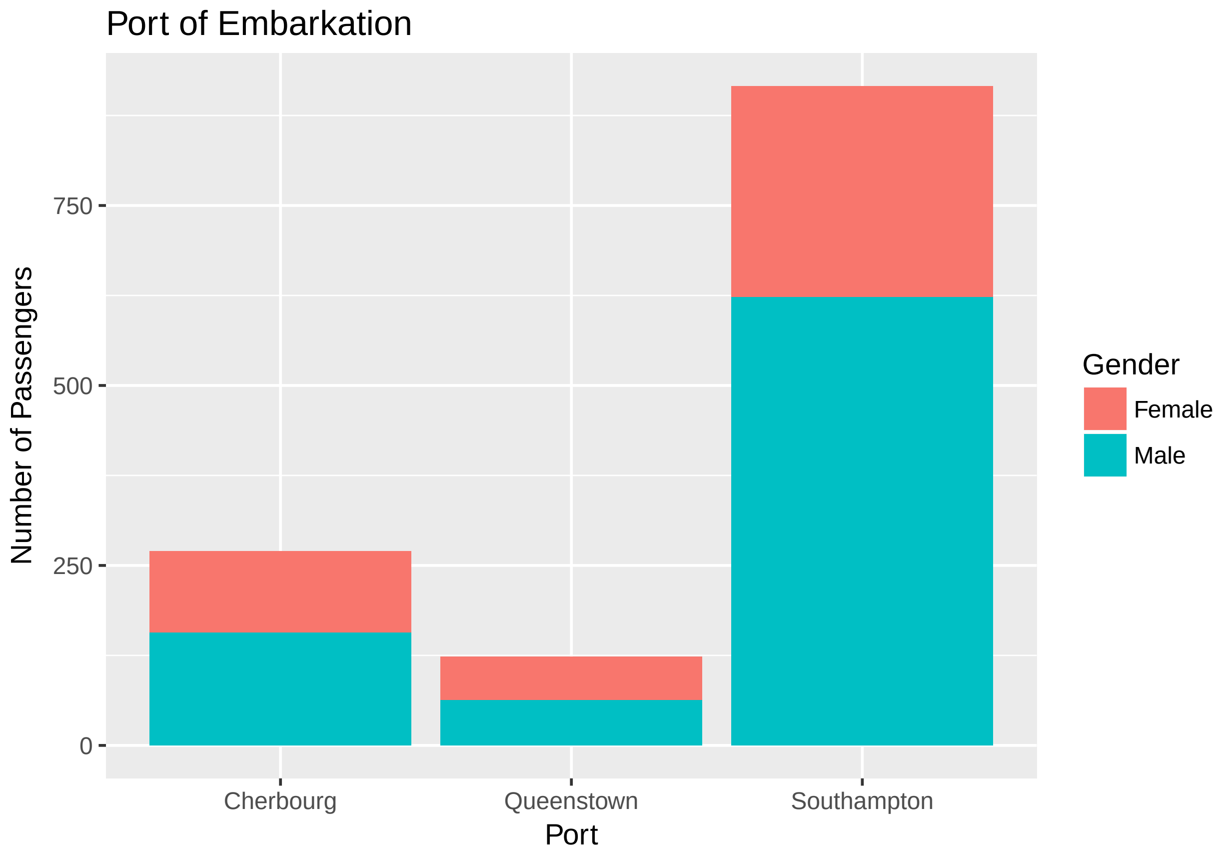
Now that we have completed the missing data, we can explore who was on the Titanic. We begin by exploring the passenger class, broken down by gender:



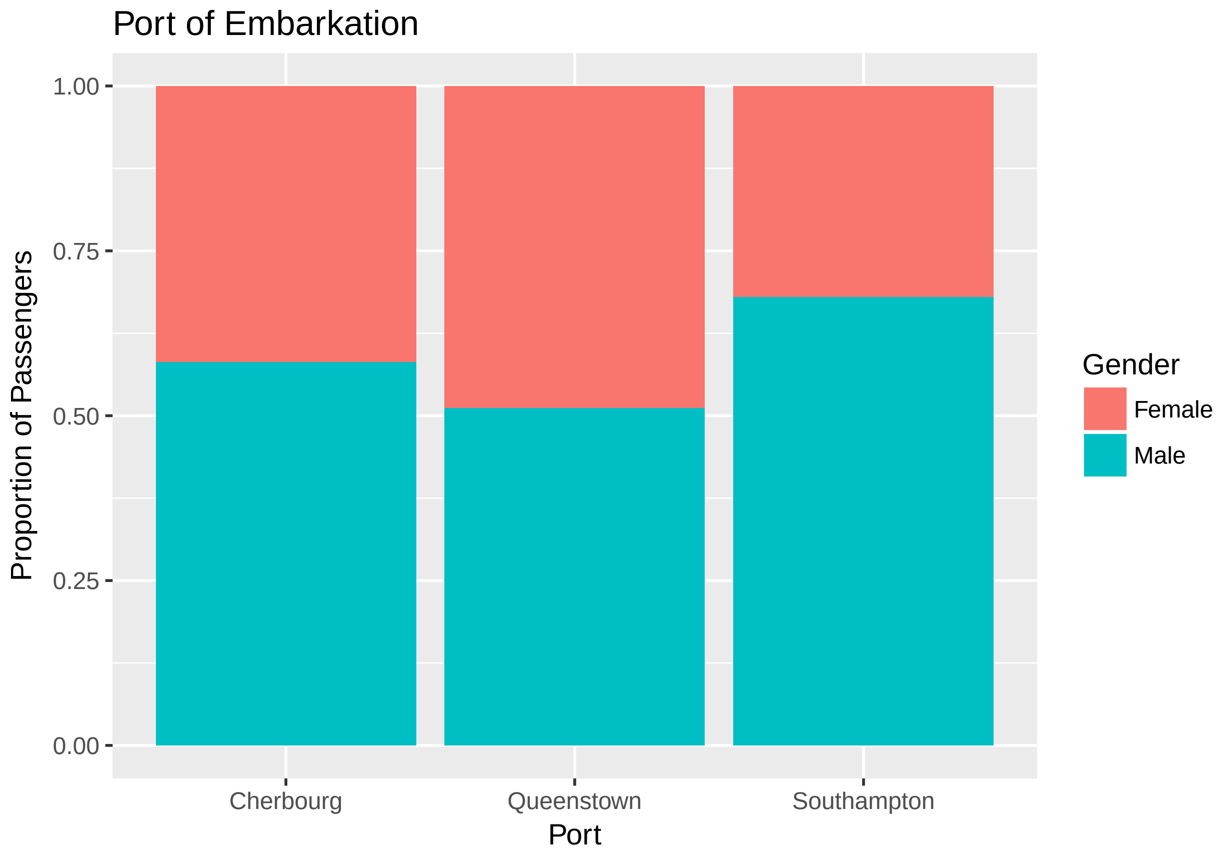



Here we notice that the greatest number of passengers were in third class, and the least number of passengers in second class. We can also see that there are more males than females within each class.

As we already mentioned, most of the passengers embarked at Southampton:

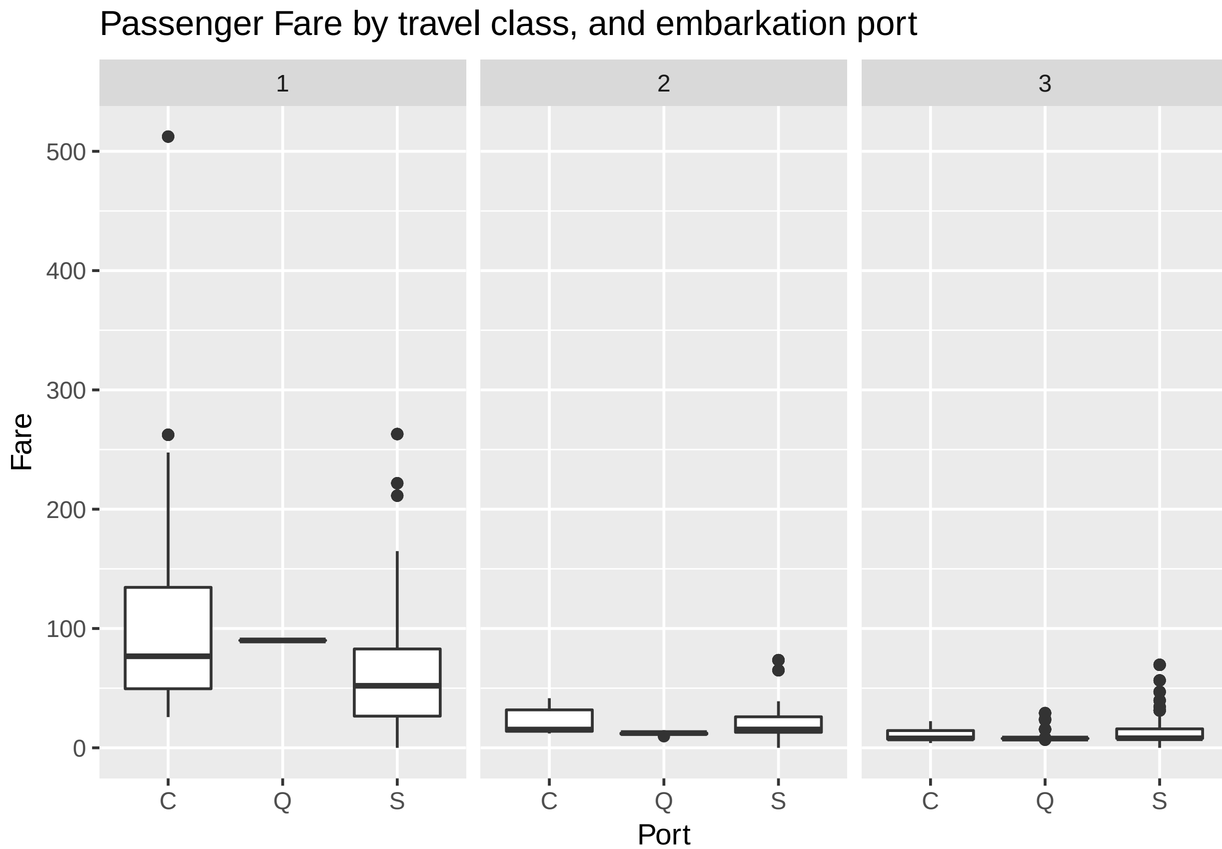


Port of Embarkation
Number of Passengers
750
500
250
0
Cherbourg
Queenstown
Southampton
Port
Gender
Female
Male



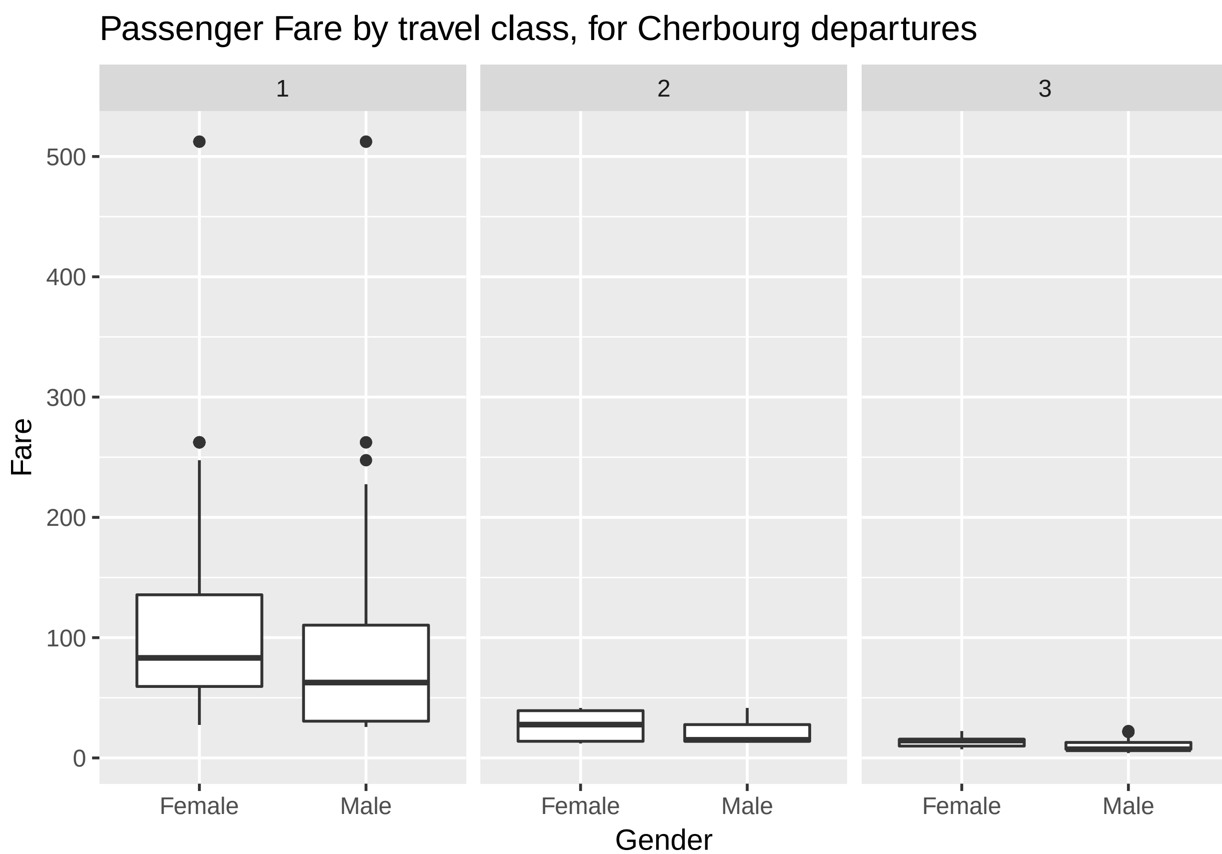

**Fares Paid by Passengers (or, what's going on at Southampton)**

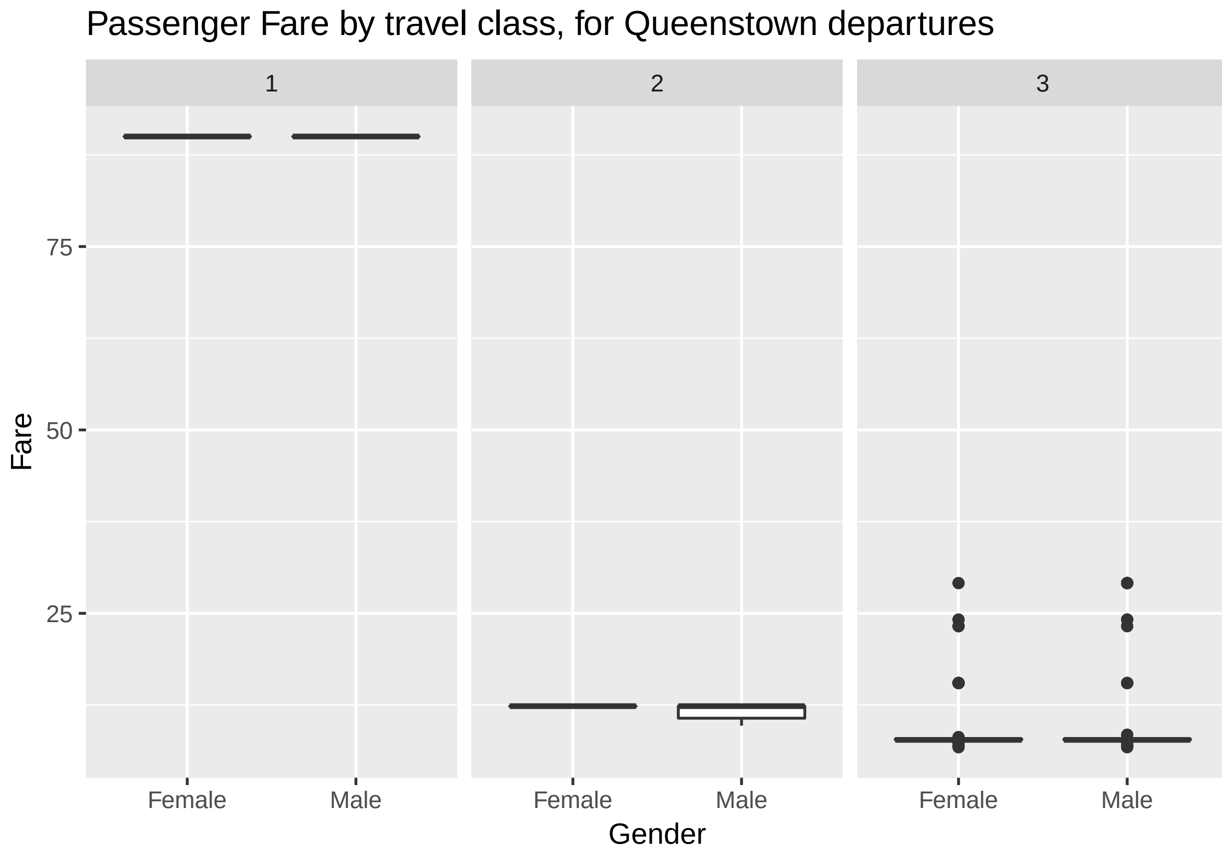
We know that the airlines change ticket prices dramatically in response to consumer activity, and track our online browsing to inflate the price for trips we're interested in. Apparently this has been going on for a long time. Consider the ticket prices paid by the Titanic Passengers:



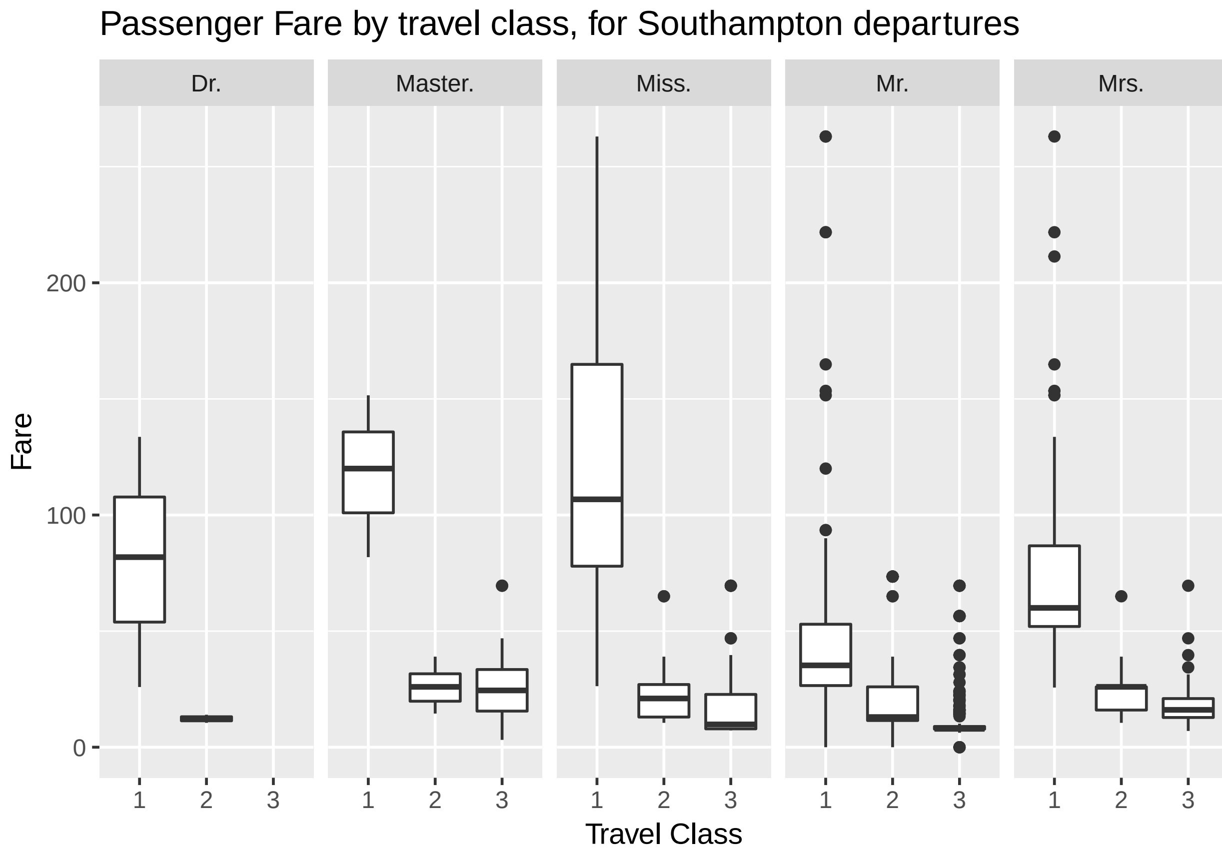

Here we see something interesting: some passengers boarding at Southampton paid more for second and third class travel than the median first class ticket from the same port. If we just look at the Southampton data, and break it down by gender:


Interestingly, female passengers tended to pay more for Southampton departures, and with greater variability. This doesn't seem to be the case at the other ports:


Passenger Fare by travel class, for Cherbourg departures
1
2
3
Fare
500
400
300
200
100
0
Female
Male
Female
Male
Female
Male
Gender



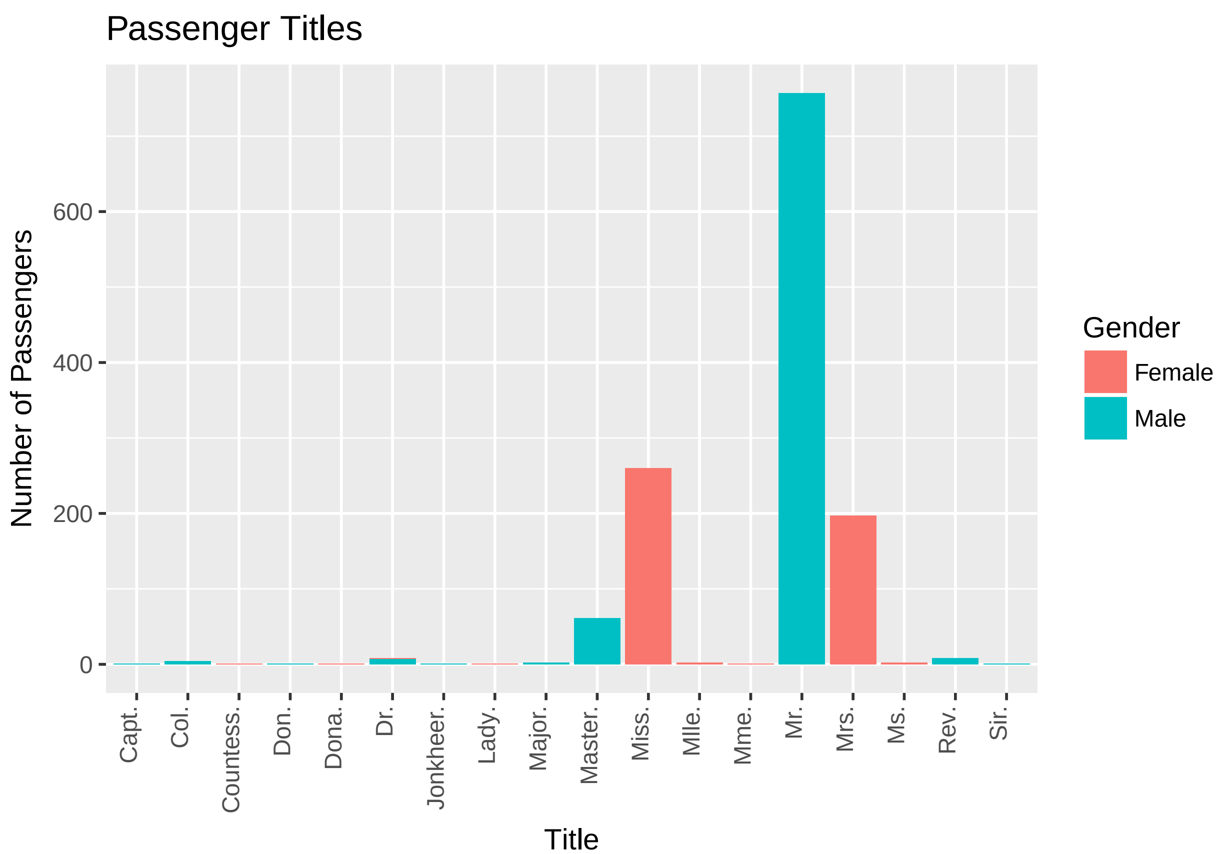

So what happened in Southampton? Perhaps looking at the fare as determined by title:



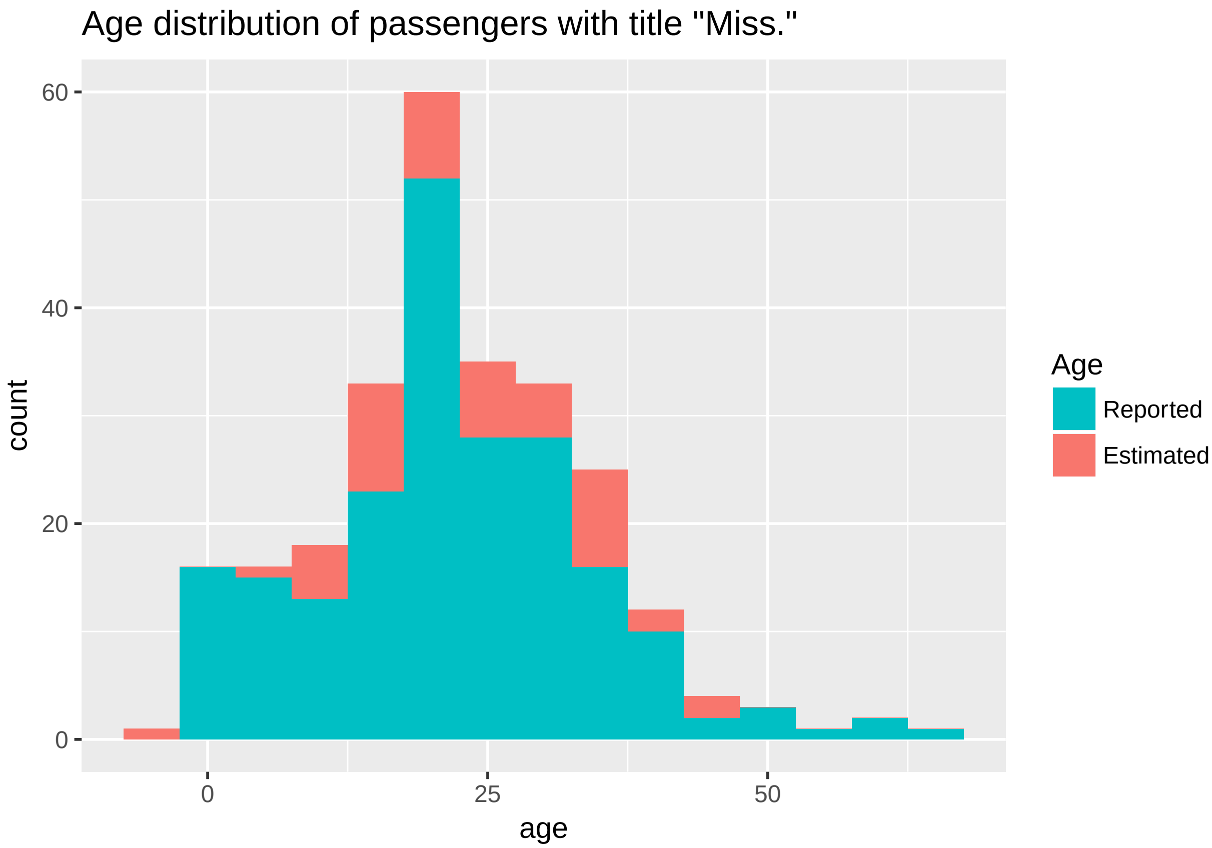

I'm not seeing a pattern here that makes any sense to me, other than that a large number of people were likely taken advantage of. For example, The group with the highest average fare is "Master", or young boys less than 14.5 years old.

**Why so many unmarried women?**

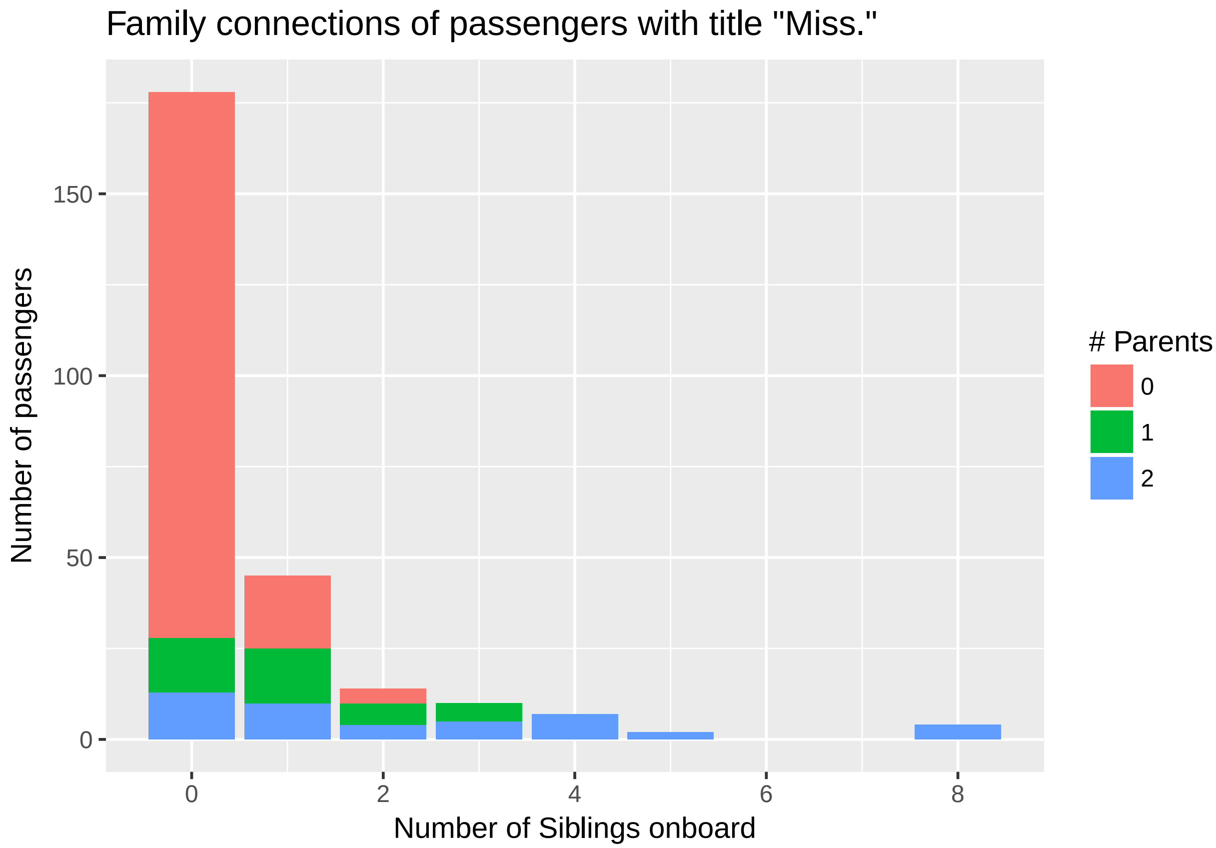
We already saw that there were more males than females on the Titanic, but each passenger also has a title (Mr., Mrs., etc.) Looking at the number of passengers with each title reveals something surprising:




The largest group of women have the title Miss. We would expect these to be young, unmarried women traveling with their family.




They do indeed seem to be young women, with an average age of 22, with the ages $< 0$ due to our method of imputing missing values. Let's look at the number of women travelling with immediate family members:




The majority of young unmarried women are not travelling with any immediate family member. At that time (1912) it seems unlikely that unmarried women would travel unaccompanied, so who were these women travelling with?

## Conclusions

As this was an exploratory analysis, there are no concrete conclusions to draw at this point. We found two surprising trends in the data that suggest new avenues of investigation.

1. **Ticket Prices for Southampton Departures:** While there were no obvious trends in the ticket price data, further analysis using machine learning techniques (cluster analysis, or regression models) may reveal underlying patterns.
2. **Who are the unmarried women traveling with?** There are online resources, such as the Encyclopedia Titanica that contain biographies of each passenger. A quick investigation of two unmarried women found that one was travelling with her aunt and uncle, and the other with close family friends. I suspect that this is the case for many of the young women travelling without immediate family. A future project involving text mining could investigate this further.

### Next Steps

Since the Titanic data set was obtained from the Kaggle competition, the next step will be to develop a model to predict survival of Titanic passengers.