Titanic Data Analysis

Eugen Buehler

November 28, 2016

Importing the Data

For my class project, I decided to use survival data from a Kaggle Competition, Titanic: Machine Learning from Disaster. A free Kaggle account is required to access the data. I downloaded the data file "train.csv" and imported it:

```
train <- read.csv("train.csv", stringsAsFactors = FALSE)
```

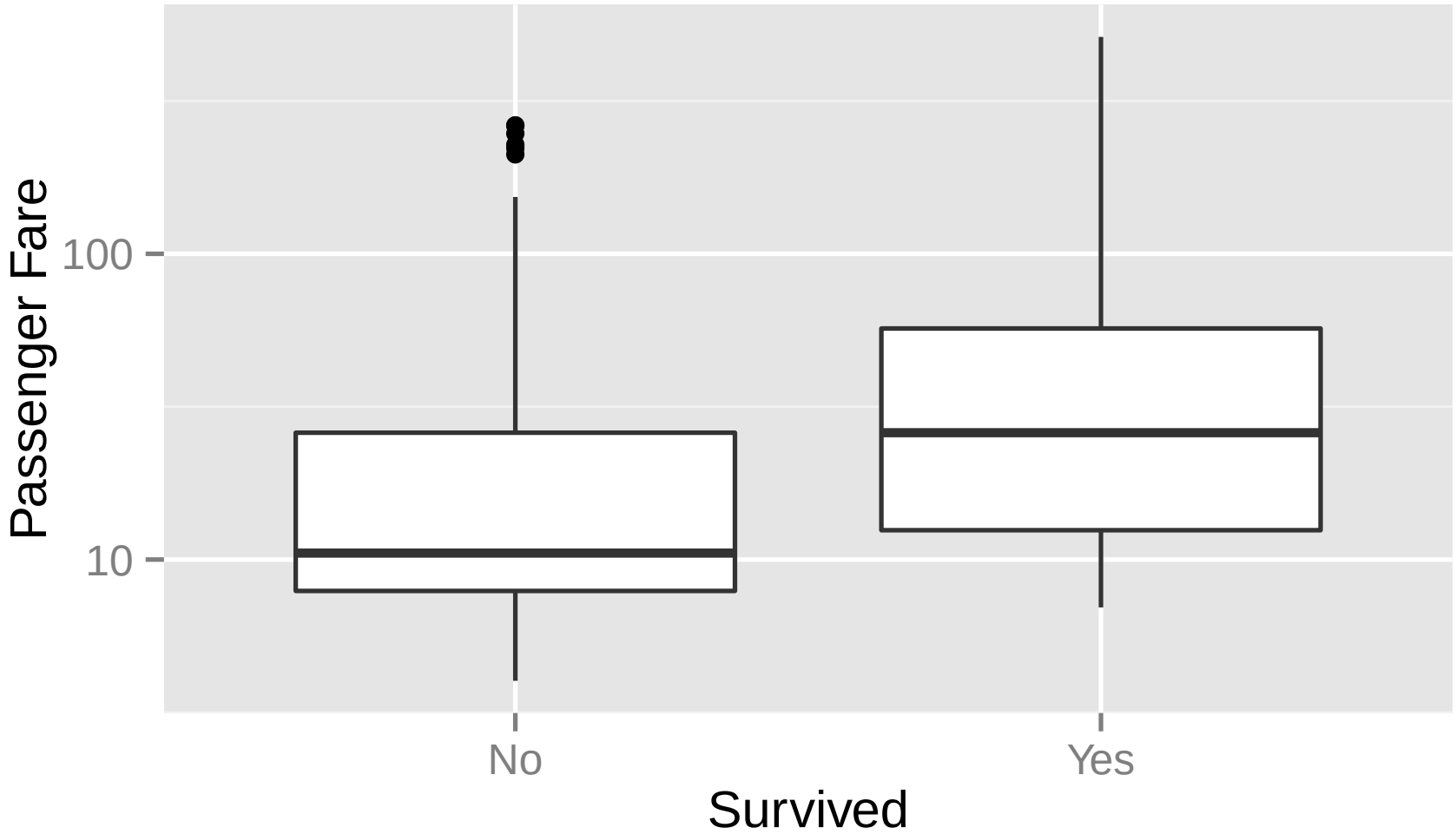
Data Preparation/Manipulation

The data includes age, gender, and fare for passengers, as well as whether they survived. I decided to focus on the relationship between fare and survival. Before analyzing the data, I needed to change "survival" to be a factor (because it was encoded as 0s and 1s, R treated it as a numeric). I also decided to excluded passengers that paid no fare, assuming that there must be something unusual about them (crew members, press?).

```
train$Survived <- as.factor(train$Survived)
train <- train[! train$Fare == 0,]
```

Graphing Fare versus Survival

```
library(ggplot2)
ggplot(train, aes(x=Survived, y=Fare)) +
  geom_boxplot() + scale_y_log10("Passenger Fare") +
  scale_x_discrete(labels = c("No", "Yes"))
```



Statistical Analysis

Because the fares paid by paid by passengers were non-normally distributed, I decided to use a non-parametric statistical test to determine if there is a relationship between the fare paid and a passenger's survival.

```
wilcox.test(Fare ~ Survived, data=train)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Fare by Survived
## W = 57264, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal
```

Since the probability of the null hypothesis that survival and fare are uncorrelated is very small, we can rule out the null hypothesis and conclude that the two variables are correlated.