Loading Data & Exploration

Sam Mason

Loading Data

Most data that we'll use during this practicum will be stored in .csv files. We use the `read.csv()` function to read these types of files into our R environment.

```
data <- read.csv(file = "/Users/sam/Desktop/titanic.csv")
```

Looking at the help documentation for this function, you'll notice under **Usage** that the `read.csv()` has some helpful default behavior. For example, unless we explicitly include the argument `header = FALSE`, the first row of the .csv will be read in as the column names. Having the `sep =` argument set to "," by default just means that R will split the data into columns every time it sees a comma.

Task 1

Read the "titanic.csv" file into R. The `read.csv()` function creates an object of class `data.frame`.

Task 2

Call the `str()` and `head()` function on this data frame to get an idea of what the data looks like.

Survived: Whether (1) or not (0) the individual survived the sinking of the Titanic
Pclass: The passenger class (1st, 2nd, 3rd)
Fare: How much the passenger paid to board the ship
SibSp: The number of siblings or spouses the individual had aboard
Parch: The number of parents or children the individual had aboard

---

Some Useful Functions for Data Exploration

Plotting

```
# Histograms to investigate the spread and central tendency of a variable
hist(data$Fare, breaks = 20) # play around with breaks = argument to make more bins

# Boxplots to compare the spread and central tendency among different groups
boxplot(data$Fare ~ data$Pclass) # boxplot() takes a formula y ~ x
```

```r
# Scatterplots to look for relationships between continuous variables
plot(data$Fare ~ data$Age) # again, response ~ predictor
```

Summarizing

```r
# Calculate the mean for all numeric columns on the basis of sex
by_sex <- aggregate(data, by = list(data$Sex), FUN = mean)

# Calculate the mean for only a subset of the numeric columns on the basis of sex
by_sex_sub <- aggregate(data[, c(3, 4, 11)], by = list(data$Sex), FUN = mean)

# This could accomplished using the function's formula syntax
by_sex_form <- aggregate(cbind(Survived, Pclass, Fare) ~ Sex, FUN = mean, data = data)
# You must specify the data = argument to tell R where objects like Survived and
# Pclass come from, because they do not independently exist in the environment

# Grouping by multiple elements
age_by <- aggregate(Age ~ Pclass + Sex, FUN = mean, data = data)
```