EDA Exercise 2

Lance Price

Exercise 1

Download 277 KB xml file associated with Reed courses

```
url <- "http://www.cs.washington.edu/research/xmldatasets/data/courses/reed.xml"

download.file(url, "~/Downloads/Reed_Courses.xml")
```

Read xml file into R as a dataframe

```
library(XML)

courses_df <- xmlToDataFrame("~/Downloads/Reed_Courses.xml")
```

Exercise 2

Count number of distinct subjects in document

```
distinct_subjects <- length(unique(courses_df[["subj"]]))
```

Print this number to console

```
cat("There are", distinct_subjects, "distinct subjects.")
```

```
## There are 31 distinct subjects.
```

Exercise 3

Count number of courses that have a NULL instructor listing

```
num_null_instructors <-length(which(courses_df$instructor == ""))
```

Print this number to console

```
cat("There are", num_null_instructors, "courses that have a NULL (blank) instructor listing.")
```

```
## There are 15 courses that have a NULL (blank) instructor listing.
```

Exercise 4

Count number of distinct instructors in document (it has a minus 1 in the computation because the NULL professor is counted in the list of unique instructors, but NULL is not a professor)

```
distinct_instructors <- length(unique(courses_df[["instructor"]])) - 1
```

Print this number to console

```
cat("There are", distinct_instructors, "distinct instructors.")
```

```
## There are 135 distinct instructors.
```