Quiz 1

Regular Expressions

`strsplit()` This function split the elements of a character vector x into substrings according to the matches to substring split within them.

Question 1

The American Community Survey distributes downloadable data about United States communities.

Download the 2006 microdata survey about housing for the state of Idaho using `download.file()` from here:

`https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv`

and load the data into `R`. The code book, describing the variable names is here:

`https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf`

Apply `strsplit()` to split all the names of the data frame on the characters "`wgtp`".

What is the value of the 123 element of the resulting list?

1. "" "15"
2. "wgtp"
3. "w" "15"
4. "wgtp" "15"

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
file <- file.path(getwd(), "ss06hid.csv")
download.file(url, file, method = "curl")
dt <- data.table(read.csv(file))
```

```
varNames <- names(dt)
varNamesSplit <- strsplit(varNames, "wgtp")
varNamesSplit[[123]]
[1] ""   "15"
```

Question 2

Load the Gross Domestic Product data for the 190 ranked countries in this data set:

`https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv`

Remove the commas from the GDP numbers in millions of dollars and average them. What is the average?

(i) 377652.4

(ii) 387854.4

(iii) 381615.4

(iv) 293700.3

Original data sources:

`http://data.worldbank.org/data-catalog/GDP-ranking-table`

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
file <- file.path(getwd(), "GDP.csv")
download.file(url, file, method = "curl")
```

```
dtGDP <- data.table(read.csv(file, skip = 4, nrows = 215,
   stringsAsFactors = FALSE))
dtGDP <- dtGDP[X != ""]
dtGDP <- dtGDP[, list(X, X.1, X.3, X.4)]
setnames(dtGDP, c("X", "X.1", "X.3", "X.4"), c("CountryCode", "rankingGDP", "Long.Name",
gdp <- as.numeric(gsub(",", "", dtGDP$gdp))
```

```
> mean(gdp, na.rm = TRUE)
[1] 377652.4
```

Question 3

In the data set from Question 2 what is a regular expression that would allow you to count the number of countries whose name **begins** with "United"?

Assume that the variable with the country names in it is named `countryNames`. How many countries begin with United?

1. `grep("*United",countryNames), 2`
2. `grep("^United",countryNames), 3`
3. `grep("*United",countryNames), 5`
4. `grep("^United",countryNames), 4`

```
> isUnited <- grepl("^United", dtGDP$Long.Name)
> summary(isUnited)
```

```
Mode   FALSE    TRUE    NA's
logical     211       3       0
```

Question 4

Load the Gross Domestic Product data for the 190 ranked countries in this data set:

`https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv`

Load the educational data from this data set:

`https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv`

Match the data based on the country shortcode. Of the countries for which the end of the fiscal year is available, how many end in June?

(i) 7

(ii) 31

(iii) 15

(iv) 13

Original data sources: http://data.worldbank.org/data-catalog/GDP-ranking-table
http://data.worldbank.org/data-catalog/ed-stats

Downloading the Data

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv"
file <- file.path(getwd(), "EDSTATS_Country.csv")
download.file(url, file, method = "curl")
```

```
dtEd <- data.table(read.csv(file))
 dt <- merge(dtGDP, dtEd, all = TRUE, by = c("CountryCode"))
 isFiscalYearEnd <- grepl("fiscal year end", tolower(dt$Special.Notes))
 isJune <- grepl("june", tolower(dt$Special.Notes))
 table(isFiscalYearEnd, isJune)
```

Quantmod R package

Question 5

You can use the quantmod (http://www.quantmod.com/) package to get historical stock prices for publicly traded companies on the NASDAQ and NYSE.

Use the following code to download data on Amazon's stock price and get the times the data was sampled.

```
library(quantmod)
amzn = getSymbols("AMZN",auto.assign=FALSE)
sampleTimes = index(amzn)
```

How many values were collected in 2012? How many values were collected on Mondays in 2012?

1. 250, 47
2. 365, 52
3. 251, 47
4. 251, 51

```
amzn <- getSymbols("AMZN", auto.assign = FALSE)
sampleTimes <- index(amzn)
addmargins(table(year(sampleTimes),       weekdays(sampleTimes)))
```