Introduction to R for Data Science

Week 1

Why R?

R is a language and an environment for statistical computing. It allows for robust data analysis and is the industry-standard tool in the field of data science.

Installation

- » Download **R** - https://cran.r-project.org/
- » Download **RStudio** - https://www.rstudio.com/

R is a language and environment for statistical computing and graphics.

RStudio is an integrated development environment (IDE) for R.

Importing Data to R

- Download dataset.
  http://stat-computing.org/dataexpo/2009/the-data.html
- Open RStudio and create a new file '`something.r`'.

```
# Importing data in R is easy. Here, we are using `read.csv` command to
# read a csv file and assigning it to a variable called `myDataFile`.
myDataFile <- read.csv('C://Users/Saugat/Downloads/2008.csv')
```

Executing Your Code

- Set your cursor on the line `read.csv()` and press the `Run` button.
- Another way to do this is it to set your cursor on the line you want to execute and hit '`Ctrl + Enter`'.

```
# When you execute the line, R will start importing the data into `myDataFile`
# variable. This will take some time if the data is large. The console window
# in RStudio is where the execution takes place.
```

Extracting Head and Tail of a Dataset

- In the console window of RStudio do the following:

```
# The head command will return the first 6 rows of the dataset.
> head(myDataFile)

# The tail command will return the last 6 rows of the dataset.
> tail(myDataFile)
```

Extracting Properties from Dataset

```
# The $ symbol is used to extract column properties from a dataset.
> head(myDataFile$Dest)

[1] TPA TPA BWI BWI BWI JAX   --> Output
```

This command returns the first 6 rows with only `Dest` (destination) column values from our airline dataset.

```
# What does this command return?
> head(myDataFile$Dest == 'IND')
```

Sum

- The sum command is used to sum the number of rows returned by an expression.

```
> sum(myDataFile$Dest == 'IND')
[1] 42732    --> Output

# Basically, R will check if the destination column (Dest) matches `IND` Indiana
# and sums up the total, which yields the number of flights departing from Indiana
```

Subset

- » Creating a subset of data from the original dataset.

```
# This will store a subset of data into the variable called `tup2008` which
# satisfies the expression.
> tup2008 <- subset(myDataFile, myDataFile$Origin == 'TUP' & myDataFile$Year == 2008)

# Sum the departure delays in Tupelo
> sum(tup2008$DepDelay)

[1] -38   --> Output
```

Caveats

- Comments start with # in R
- To find more about what a command does put a ? in front of it.

```
# RStudio will show show documentation on the right side of the screen.
> ?sum

# Sometimes your data will not have appropriate values in required columns. In
# cases like these you can ignore these values by making the second parameter for
# sum command to TRUE. This will ignore the N/A (not available) values in the data.
atlToLax <- subset(myDataFile, myDataFile$Origin == 'ATL' & myDataFile$Dest == 'LAX')
sum(atlToLax$DepTime < 1200, na.rm = TRUE)

[1] 2133   --> Output
```

End of Week 1