Unit 1: Basic Use of R and RStudio

Instructions

This document is organized such that problems 1-4 are intended to be done on your own time, ideally before the live synchronous session (or for review afterwards). Problem 5 will be used as the basis for the live synchronous session.

Problem 1: Basic R commands

part 1

Define:

```
x<-c(4,2,6)
y<-c(1,0,-1)
```

Decide what the result will be of the following:

1. `length(x)`
2. `sum(x)`
3. `x+y`
4. `x*y`
5. `x-2`
6. `x^2`

part 2

Decide what the following sequences produce in R:

1. `7:11`
2. `seq(2,9)`
3. `seq(4,10,by=2)`
4. `rep(NA,10)`

part 3

If `x = c(5,9,2,3,4,6,7,0,8,12,2,9)` decide what each of the following is and use R to check your answers:

1. `x[2]`
2. `x[2:4]`
3. `x[c(2,3,5:7)]`
4. `x[-(10:12)]`

Problem 2: 100 meter dash data and misleading graphs

We have data from 1896 to 2016 on the winning time of the Mens 100 Meter dash in the Olympic Games. Read the data into R as follows:

```
dashdata=read.csv("mens100.csv")
```

1. How many variables are in this data set? How many observations are in this data set?
2. One can create a barplot in R as follows (the strange `las=2` rotates the x-axis labels)

```r
barplot(dashdata$time,names.arg=dashdata$year,las=2)
```

3. One can create misleading graphs by changing the starting points on the axes. Compare the following graphs of the same data. The par command tells R to plot two graphs, one on top of each other:

```r
barplot(dashdata$time,names.arg=dashdata$year,las=2)
barplot(dashdata$time,names.arg=dashdata$year,las=2,ylim=c(9.5,12.5),xpd=FALSE)
```

Problem 3: Playing around with the cars data in R

This question uses an old data set on cars from Consumer Reports. To load the data into R enter the following command

```r
cardata=read.csv("cars10.csv")
```

1. How many variables are in the data set? How many observations are in this data set?
2. How many cars in the data set are Domestic? How many are Foreign?
3. Which car has the lowest mpg? The command `which.min(cardata$mpg)` might prove helpful here.
4. Which car has the highest price? The command `which.max(cardata$price)` might prove useful here.
5. Make a histogram of the price of cars. What shape does the histogram take? (Is it symmetric? Skewed?)
6. Create seperate histograms for the mpg of domestic cars and the mpg of foreign cars. Discuss any obvious differences in the two histograms.
7. Make a scatter plot of the variables weight and length. Does there appear to be any association between the variables?

Problem 4: Finance data via the `quantmod` package

An important feature of R is that it is highly extendable. Indeed there are over 800 user contributed packages that add all sorts of functionality to R, from analysis of genetic data to pricing options or fitting neural networks. A quick read about R packages may be found here: >https://www.datacamp.com/community/tutorials/r-packages-guide

For example, the `quantmod` package reads financial data from Yahoo into R. Report on what the following code does. For those curious as to what is returned by the `getSymbols` function call, look at the object `AAPL` to see what it consists of.

```r
# If you are using your own install of RStudio on your machine (not using Rstudio.cloud),
# run the line of code below the first time you want to use 'quantmod':
# install.packages("quantmod")
library(quantmod)
getSymbols("AAPL",from="2016-01-01")
getSymbols("SPY",from="2016-01-01")
aaplret=as.numeric(monthlyReturn(AAPL))
spyret=as.numeric(monthlyReturn(SPY))
par(mfrow=c(2,2))
plot(Ad(AAPL))
plot(Ad(SPY))
hist(Vo(AAPL)/1000)
plot(spyret,aaplret)
```

Problem 5: Airline data (this will be covered in the live session)

We have a subset of a larger data set on airline ontime performance of domestic flights operated by large air carriers. The information was compiled from the Bureau of Transportation Statistics. We will only be analyzing the data from randomly selected flights from November 2008 which is in the data set airline2008Nov.csv. The variable names and definitions are listed in the file airline2008_dataset_definition.pdf. Read this data set into R and answer the following questions. You may read the data set into R using the command:

```
mydata=read.csv("airline2008Nov.csv")
```

1. Explore this data set. How many variables does it contain? How many obesrvations? What airline has the most flights in the data set?
2. Write two questions that may be answered from this data set.
3. We want to be able to deal with missing entries (represented as `NA`s in the data set).

a) Are there missing values (NA) in the dataset? If so, remove the rows that contain NAs in the dataset. Save this data file to be used again:

The following command prints out all the rows with missing data
`mydata[!complete.cases(mydata),]`

The following command creates a data set without misising data
`newdata <- na.omit(mydata)`

The following command writes a csv file `write.csv(newdata,"cleanairline.csv")`

b) How many observations were removed? How many observations are there in the new data set?

c) What might go wrong if we analyze this data set and ignore the fact that we removed some observations?

4. Suppose we now only want to work with flights that have as destination ATL, JFK, DFW and MSP.

a) Create a smaller data set using the `subset` command as follows: > `smalldata=subset(mydata,Dest=="ATL" | Dest == "JFK" | Dest=="DFW" | Dest == "MSP")}`

For some reason we need to reset the factor variable `Dest` since we removed many levels of it.
Do the following command in R `smalldata$Dest=factor(smalldata$Dest)`

b) For readability, we want to transform airport codes to their full name in the variable "Dest". Change the name of the four airports listed below using the following abbreviated names. You can see why the three letter code of abbreviations was started!

ATL → Atlanta
JFK → NYKennedy
DFW → DallasFtWorth
MSP → MinneapolisStPaul

c) Create a pie chart on the destination variable. You will need to use the `table` command in conjunction with the `pie` command. What do you notice?

5. We are going to see if the variable "ActualElapsedTime" can be calculated from other variables in the data set.

a) Write down a mathematical equation to calculate "ActualElapsedTime" from "AirTime", "TaxiIn" and "Taxiout".

b) Code your expression using R by creating a new variable.

c) Show that your code is correct by displaying the original variable "ActualElapsedTime" and the variable that you calculated. Please only print out the first 6 rows.

6. Let's explore the reduced data set further to describe several variables.

a) How many flights in our reduced data set had a weather delay? What proportion of flights?

b) How many weather delay flights were there going into JFK? What proportion of flights were delayed?

c) In our reduced data set, what was the maximum departure delay? Which flight was this?

d) In our reduced data set, which day of the week had the lowest number of flights?

e) Create a scatter plot of TaxiIn on the x axis and TaxiOut on the Y axis. Is this a useful plot?