Problem Set - 3

Data Science with R

Images in R

Images are broken down into small square boxes, called pixels. Pixels in color images are made from the three primary colors: red, green, blue (rgb). Every pixel in the image can written as a combination of these three colors. Images in greyscale are **not** saved as rgb and rather as just one numeric value (indicating how much white there is).

We will use a package in `R` to study images. `R` packages are a collection of functions written to perform tasks. To install any package, the command is

```
install.packages("name of package in quotations")
```

We will use package `imager`. Your first task is to install this package:

```
install.packages("imager")
```

Packages only needs to be installed once on a computer. After that, for any `R` session you only need to load the package using

```
library(imager) #now without quotes
```

Images in package `imager` are represented as 4D numeric arrays. The four dimensions are labelled x, y, z, c. The first two are the usual spatial dimensions, the third one will usually corresponds to depth or time (for gifs), and the fourth one is colour for each channel: red, blue, and green. Arrays are structured similar to matrices and follow the same rules of subsetting.

Let's first load, plot, and check dimension of an image.

First, make sure you download the all the images in the Week - 3 folder in your local machine, and change your working directory to the location of the images.

```
dog <- load.image("dog.jpeg")
dim(dog) # stored as RGB
plot(dog)  # plot image
```

You will see that the dimension of the image is $640 \times 635 \times 1 \times 3$. That means, there are 640×635 pixels in the image (640 columns and 635 rows), only one time (so it is not a gif) and 3 color channels, one for each primary color. We can also obtain the grayscale version of the image.

```r
graydog <- grayscale(dog)
plot(graydog)
dim(graydog)
```

To extract the raw matrix or array of the image, we may do the following

```r
# Extract the black and white image as matrix
gray.mat <- as.matrix(graydog[,,1,1])
dim(gray.mat)

# Extracts the array will all three rgb channels
col.mat <- as.array(dog[, ,1, ])
dim(col.mat)
```

We may now do manipulations on the arrays and then save the updated arrays as image

```r
# Vertical cropping
cropped.mat <- col.mat[1:300, , ]
crop.dog <- as.cimg(cropped.mat)
plot(crop.dog)
```

Using all this information, and the basics of the `R` that you know, write code for the following tasks:

1. Find the "purest green" part of the image and mark that with a red point on the dog image. You may have to use the `which( …, arr.ind = TRUE)` command and the `points()` function.

2. Repeat the previous part for the "purest red" and "purest blue".

3. Images "col1.png", "col2.png", "col3.png" are images of the primary colors. Without opening the files on the computer, write an R program that guesses which file is which color.

4. Write an `R` function that takes a image as an input and outputs a prediction on whether the image has a lot of snow or not. Test this function on "land1.jpeg" and "land2.jpeg".

5. Write an `R` function that takes an `imager` image as input and outputs the `imager` image rotated by 180 degrees.

6. Write an `R` function that takes an `imager` image as input and outputs the `imager` image rotated clockwise by 90 degrees.

7. Write an `R` function that takes an `imager` image as input and outputs the `imager` image rotated anti-clockwise by 90 degrees.

8. Crop the image of the dog to a 600×600 pixel file. Write an algorithm to convert the image to a 300×300 pixel `imager` image. The reduced image should still have the complete dog. Save the `imager` image in a `jpeg` file using command `save.image()`. What is the size of this new file?

9. Repeat the above for making a 60×60 `imager` image of the dog.

10. The above is an example of a simple image compression. Can you think of other ways of compressing the image, using tools from linear algebra?

Introduction to Python

Python is equally powerful language. Some of you might be interested in learning python. Go to https://colab.research.google.com/drive/1DCFqRTUbshVYJ-7q7SfMOMQjcudLSjtL?usp=sharing . This is containing some basics of syntax of python. If you want to learn more here is the resource to learn Data Science in Python https://jakevdp.github.io/PythonDataScienceHandbook/

Basic Visualization in R

So far, we have learned how to collect data, clean and process it. Go to :

https://www.r-bloggers.com/2019/05/how-to-save-and-load-datasets-in-r-an-overview/ and learn how datasets can be saved in R. we will discuss this during our session.

However, a crucial component of data analysis is data visualization. This can often help ask interesting questions about the data.

We can think of visualizations as:

- single variable visualization - histogram, boxplots
- Multivariable visualization - scatterplot, side-by-side boxplot.

Note: In any visualization one must be very clear about what we are trying to visualize. Further axes should be clearly described and legends should be

1. You can find the data in `IMDB_movies.Rdata` in the week - 3 folder. Download the `.Rdata` in your working directory and load the file using the `load()` function.

2. **Histogram**: read the documentation for the `hist()` function that makes a histogram.

   a. Make a histogram of the ratings for the top 250 movies. Using the argument `main`, set the title of the histogram to be "Histogram of Ratings" and the label on the x-axis as "Ratings".

   b. Make the histogram again so that the bars are white in color.

3. **Boxplot**: read the documentation for the `boxplot()` function that makes a boxplot. The line in the middle is the median of the observations, the bottom box the 25\% percentile, the top box the 75\% percentile.

a. Make a boxplot of the ratings of the top 250 movies. Make sure to assign an appropriate title.

b. Make the boxplot again so that the bars are pink in color.

4. **Side-by-Side Boxplots:** Reading the help page for `boxplot()`, make a side-by-side boxplot of men's ratings and women's ratings. Make sure the axis labels are appropriate.

5. **Overlapped histograms:** Make a plot of histograms of men's ratings and women's ratings, overlayed on top of each other. You may use the `col = adjustcolor("red", alpha.f = .5)` option to make colors transparent.

   Use `legend()` command to add a legend to the plot.

6. **Scatterplot:** Scatterplots can help explain the relationship between two quantitative variables. Using the `plot()` function, make a scatterplot of number of votes on the x-axis and the ratings on the y-axis.

7. **Text:** Using the `text()` function, write down the names of the movies in the above plot whose ratings are more than 8.9.

8. **3 variable plots in two dimensions:** Visualizing a third variable in the above plot is possible when the third variable is categorical.
   IMDB.com was established in the year 1996. In the plot in Part (6), color the movies released before 1996 in a different color from the movies released after 1996. Make sure to add a legend.

9. Using such visualizations can you identify some potential biases or unfairness in the movie rankings?

10. Create an animation using `Sys.sleep()` and `points()` function, of presenting one data-point at a time in Park (8).

We will learn Advance Visualization techniques next week, Make sure you have done with this problem set timely.