MODULE

D2

Data Analysis 2

“01100110 01110101 01100011 01101011 00100000 01101000 01100101 01110010 00100000 00100111 01110100 01101001 01101100 00100000 01110011 01101000 01100101 00100000 01100100 01101001 01100101 01100100 00100000 00101000 01101001 01110100 00100111 01110011 00100000 01101100 01101001 01110100 00101001”

- 01010100 00000000 01110010 00000000
01100001 00000000 01110110 00000000
01101001 00000000 01110011 00000000
00100000 00000000 01010011 00000000
01100011 00000000 01101111 00000000 01110100
00000000 01110100 00000000

Time Period and Situation

The time is now! But data analysis is forever.

Agenda

1. Homework review
2. Visualization
3. Functions
4. Apply() family

Student Objectives:

1. Students should build intuition in handling data frames and columns
2. Students should be able to crunch .CSVs into useful data ready for plotting using what they learn in this module.

Facilitation Notes

- Make sure to constantly ask to look at what students are doing, and ask everybody if they need help!

Homework review

Going over the answers

Resources:

1. n/a

CORE Crash Course (CCC):

Talk with them about the answers they got and how they got there.

Presentations

1. Have a couple students present how they solved the homework problems.
2. Take questions.
3. If necessary, show the students the proper way to do the homework.

Homework review

Going into visualization

Resources:

1. Use the avocado dataset
2. Show the students how to do a pie chart split into categories of year [column titled “year”] by number of Large Bags [column titled “Large Bags”]
3. Then ask the students to do two things for homework
   a. Create this same pie chart, but use small bags (literally changing one line)
   b. Create a scatter plot that compares the “average price” of avocados with the number of total bags sold (using plot() function)

CORE Crash Course (CCC):

Present the basic R plots like pie(), hist(), etc. Use the table() function to generate data for the pie plot, and use a time series from the avocado dataset to do the histogram.

Understanding functions

Resources:

1. https://www.tutorialspoint.com/r/r_functions.htm

CORE Crash Course (CCC):

They won’t be writing very complex functions but they should understand them. Functions work kind of the same as in other languages but there are some peculiarities of R.

The function:

1. Functions are defined using the function keyword.

```
function_name <- function(arg_1, arg_2, ...) {
   Function body
}
```

   a. Note how the function is also a variable with its own name.
   b. Arguments are interestingly duck typed.

2. Most of the functions that they will use are built in, like mean, sum, etc.

```
# Create a sequence of numbers from 32 to 44.
print(seq(32,44))

# Find mean of numbers from 25 to 82.
print(mean(25:82))
```

```
# Find sum of numbers from 41 to 68.
print(sum(41:68))
```

a. There's also max, min, median, mode.

Now you know how to do functions. If we have time, have them write some functions that implement basic operations.

apply, lapply, sapply, tapply

Resources:

1. https://www.guru99.com/r-apply-sapply-tapply.html

CORE Crash Course (CCC):

These functions run an input function on every element of a list, vector, or matrix and return the output. They are very useful for processing a lot of data at the same time.

Examples:

1. Let's say you had a column of locations as Illinois town names and you wanted a corresponding list of coordinates for each location. If you have a geolocation function that can take the town names and return coordinates, you would use something like:

```
apply(vector_of_names, geolocation_function)
```

i. Note that the function runs on each element, and not on the vector as a whole.

2. Another situation would be if you had multiple vectors, and you need to do some sort of operation on both and return a result. Imagine if you had a matrix (vector of vectors) of students' grades in a table,and you wanted to find each student's (row's) average grade. You could use:

```
table$average_grade <- apply(table, 1, mean)
```

i. The 1 in the middle tells the function to operate on the matrix's first dimension by iterating over rows. Iterating over columns is 2.

Practice:

- Add a column which is the year that the student entered IMSA, which would be three years before the graduation year.

Supplemental Content

References

1. All the csv files are in the Data Analysis folder.
2. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf is always useful