TPP QUANT REVIEW
Coding Challenge
February, 11th 2021

STATISTICAL ANALYSIS The purpose of this task is to convert a csv file into a dataframe, and perform basic data manipulation on the dataframe.
A candy manufacturer at MIT produces Bertie Bott's Every Flavour Beans. Each jelly bag contains candies that ALL have the same flavor (hence color). Further, the manufacturer claims that each bag contains at least 49 candies.
TPP students collected a sample of candy bags over 8 months, to study the different candies flavor and to test whether the manufacturer's claim is honest. The data can be found in the $TPPQuant_Candies.csv$ file.

(a) Launch a Jupyter Notebook and open the TPP QUANT REVIEW.ipynb. Download the csv file, and open it in the Jupyter Notebook interface. What is the name of each column ? How many rows does the data have ? How many different flavors does the sample contain? What are the colors associated with the flavors?

(b) What are the maximum, minimum and average number of beans found in the sample of bags? How many bags of each color did the students found? What is the average number of candies in the blue bags? How about the green bags? How about the bags collected on April 1st?

(c) Group the data samples by their color, and create two vectors (one with the color's name and the other with the number of candy bags associated with the color). Plot a bar chart with the color's name in the x-axis and the bag number in the y-axis. Order the vectors and plot the bar chart again.

(d) You see that the average number of candies per bag is smaller than 49. Assume that the standard deviation found is 4.42. You want to check whether there is statistical evidence that the claim according to which bags have at least 49 candies is fallacious. What is the null hypothesis? What is the alternative hypothesis? The test is one-sided. What is the test value? Draw the rejection region. What is the threshold below which you would reject the null hypothesis? What is the p-value?

TPP NETWORK Let's study an imaginary network of students that get formed in a

school program. Students are connected if they met in person.

(a) Let's imagine that on September 1st 2020, 7 TPPers managed to reach campus (called students 0, 1, 2, 3, 4, 5 and 6). Two groups of three students met, and one student did not meet anyone. The TPPers network, called Graph 1, looks as in Figure 1. Write the adjacency matrix associated with the graph. Find the eigenvalues of the adjacency matrix. Any comment?

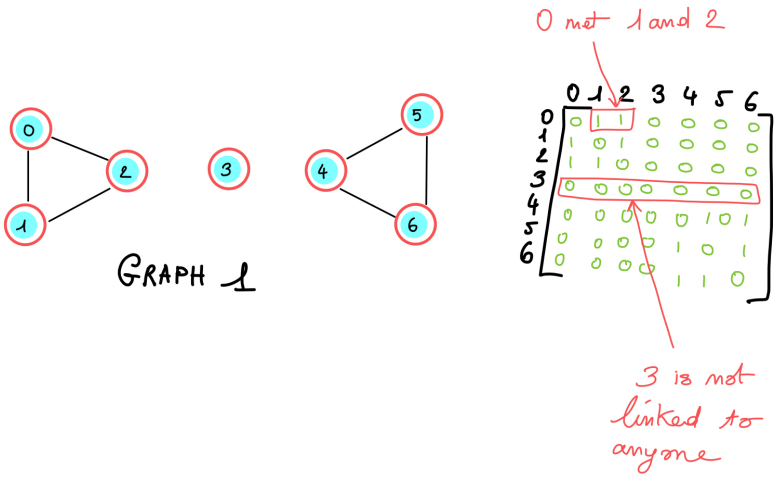

Figure 1: Graph 1

(b) Student 3 meets students 2 and 4. The new graph, called Graph 2, is represented in Figure 2. Write the adjacency matrix associated with Graph 2. Find the eigenvalues of the adjacency matrix. Compare them with those of Graph 1. Use the networkx package to create a graph object G. Draw the network. Compute the degree centrality, as well as the betweeness centrality. Which nodes are "central"?

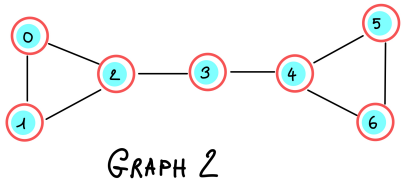

Figure 2: Graph 2

(c) Three new students reach campus. Students 7, 8 and 9. Student 6 meets all the new students, and student 5 only meets student 7. Add the nodes and the edges using the networkx built-in functiond. The new graph, called Graph 3. Draw the graph. Compute the degree centrality, as well as the betweeness centrality. Which nodes are "central"? The centrality scores are stored in an object called a dictionary. Create a dataframe (with 3 columns: node, degree centrality, betweeness centrality) from the two dictionaries. Print the dataframe. Export the dataframe.

CODE FUNCTIONS (OPTIONAL) Let's code a couple of short functions to get some practice.

(a) Code and test the function max_diff with the maximal difference between any pair of values in a given list.

(b) Code and test the function max_value that finds the maximum value in a list without using any built-in functions.

(c) Code and test the function $reverse$ that reverse a list (eg. return [2, 5, 3] form input [3, 5, 2]) without using any built-in functions.

(d) Code and test the function $sell\_stock$ that outputs the maximum profit possible, obtained form buying and selling a stock.

(e) Code and test the $alert1$ function, that outputs "ALERT at time t" if the value at time t is greater or equal than 1.5 time the average value of the past 5 data points (that is, the data between time t-5 and t-1). How would you change the code if you wanted to trigger an alert when the value at time t is greater or equal than 1.5 times the average value of the past 5 data points OR smaller or equal than 2/3 of the average value of the past 5 data points.