Reflective Document

Student Id: 24045464

## Part 1

**Describe your approach to this part of the assessment:**

Since we cannot use anything other than core python, our main source of info comes from markdown files of lectures. The main challenge here is how to open a CSV file cause we can use readline() method to read its lines. We will use open() function in core python here to open our file.

**Did you find it easy or difficult? Why?**

One of the advantages of python is its libraries and how it makes work easier. Comparing to R the process for python is rather detailed and needs us to read every line. But when we go through the reading materials of the module, we find examples on how to implement this part easily.

**What problems did you encounter?**

At first, I encountered memory issues as I forgot to close my file after each time of running it. But when I read about this problem, I noticed I can use with to simplify the resource and connection management. This keyword has significantly made my code simpler.

**How did you overcome them?**

First, I started reading about reading files using core python in communities. Most of the solutions suggested that I close my file after opening it. So, the first version of the code was longer than the final version.

**Identify any strengths/weaknesses of the approach you took**

While with manages resources effectively we haven’t run tests to find out if it’s truly as efficient as close() function or not. But the strength in it is that we don’t have to add additional codes to close the file or face memory overload if we forget to close our file.

**How could this approach be improved?**

For our Pearson’s correlation function. We can use encoding instead of removing the non-numeric values so we can also calculate the correlation for categorical columns as well.

**Suggest an alternative approach you could have taken**

I would suggest that we also normalize the values and do z method on them. This will create a normalize scale in which we can compare different variables.

## Part 2

**Describe your approach to this part of the assessment**

When we look at our columns, we see that we have a lot of non-numeric values such as age group and region. Our two only numeric columns are score and click\_events which is in another file. First we create a clean data frame to include our columns. While age, region and gender can be a predictor for students’ scores we want to find a variable that will give us actionable insights into this data. While it is very valuable to find out that male population are earning higher grades than female population as a hypothesis we’re going to need systematic changes to make changes for future observation. But we can see that we can see the click\_events of each student along with their score and final\_result. This will help us see if using the online education platform helps students get better results.

**Did you find it easy or difficult? Why?**

The main challenge we find with this task is that “How to define an outlier?”. So statistically we can set up different boundaries with different methods, but we cannot be for sure that these datums are outliers or not.

**What problems did you encounter?**

When writing the z-score function, I encountered the problem with missing values. It would give division by zero sometimes and empty all of our data frame.

**How did you overcome them?**

I used the function dropna() to drop the rows that have missing values in them.

**Identify any strengths/weaknesses of the approach you took**

While this approach is good for fast analysis. If we have a lot of missing values we will lose a big chunk of our data because we are dropping the rows. It is better if we could use other methods to handle the missing values without dropping the row completely.

**How could this approach be improved?**

For example we could replace the missing values with median of those columns. We could also find other libraries that will offer unpaired tests and correlation and handle the outliers within themselves.

**Suggest an alternative approach you could have taken**

Another approach would be looking at data from region and higher education status. We can see how our score compares against people from different regions or education backgrounds. This will also help us localize or simplify our educational content.