**Profiling Internet Users:**

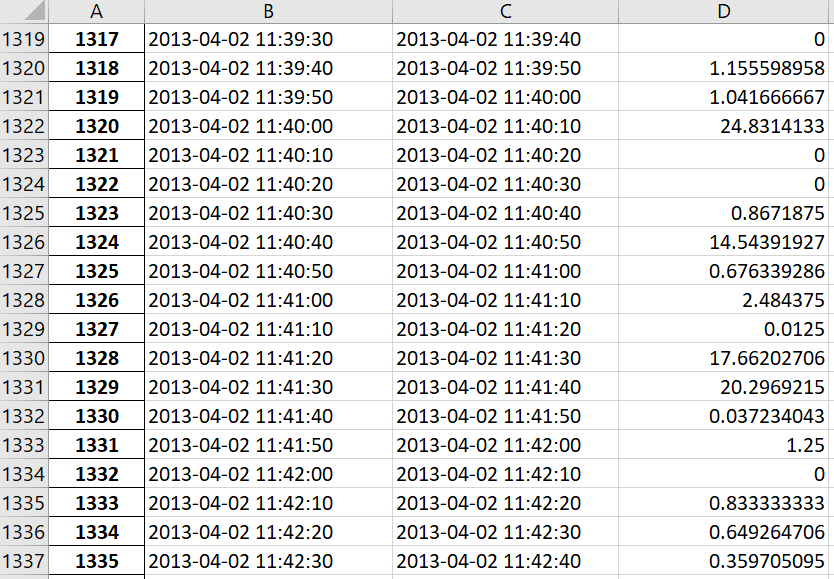
**Objective:** To demonstrate whether the internet usage of each subject is statistically indistinguishable when compared to the same user over time, being distinguishable to the internet usage of other users.

**Procedure:**

* Given 54 xlsx (excel) files which represent internet usage of 54 users. We have to find out the amount packets used ny each user, so we calculate doctet/duration for each packet. But first we have to create 10 seconds, 227 seconds and 5 minutes time frames and see the usage (doctet/duration) in each time slot. It’s given that we have to only observe on weekdays (Mon-Fri) from 8AM to 5PM.
* The first step is to convert the epoch time into real time. Then we will see if there are any packets in the first time slot (Mon 8:00:00 – 8:00:10 in 10 sec window). So, in this way we check all the packets that fall these time slots. Then we calculate doctet/duration for each packet and sum it and find its average for each time slot. This step is done for 1st week and 2nd week.
* Then we calculate spearman correlation values r1a2a (Week 1 user a, week 2 user a), r1a2b (Week 1 user a, week 2 user b) and r2a2b (Week 2 user a, week 2 user b). Using these three correlation values we can calculate the Z value of the MRR-Z test. With this Z value we can calculate the final P-Value which tells us the distinguishability between the 54 users internet usage.

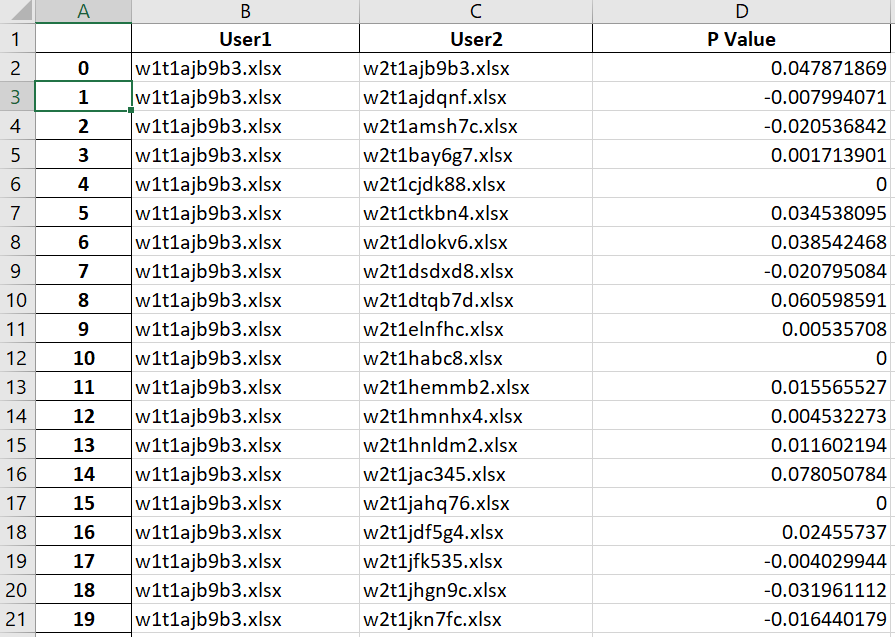
**Intermediate Output Samples:**

1. First, I had split the time and seen the doctet/duration is each time slot.



This is a sample of user (zws2p8) of 1st week for 10 second window.

1. Using the above files data I calculated the spearman correlation values r1a2a , r1a2b and r2a2b.

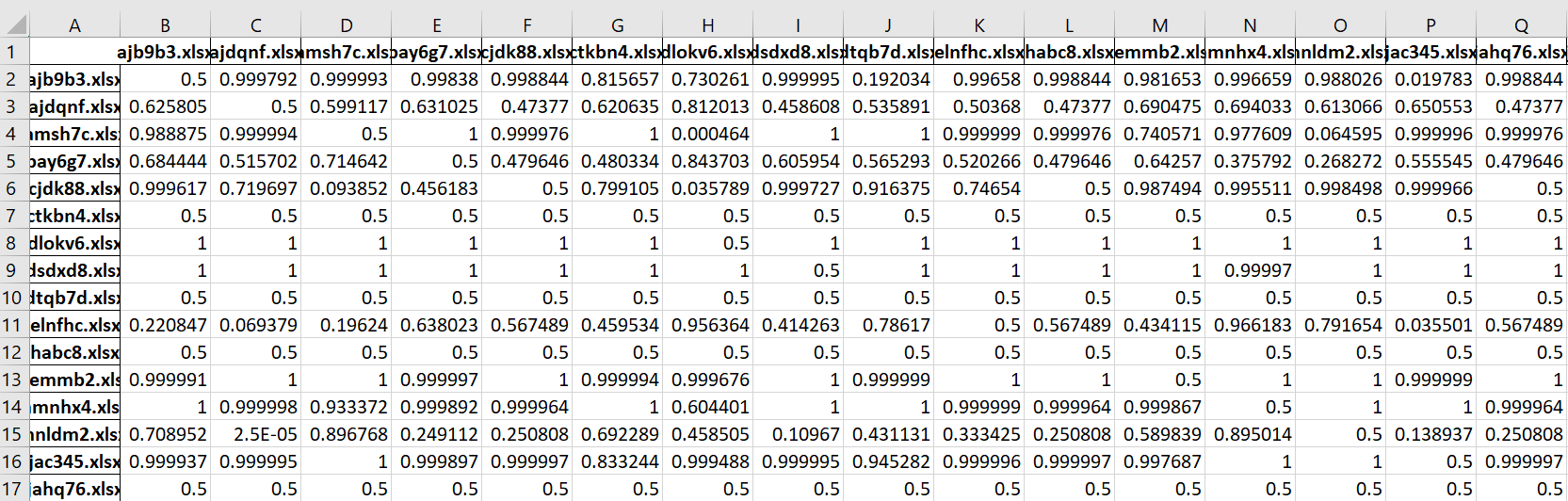


This is a sample of r1a2b values.

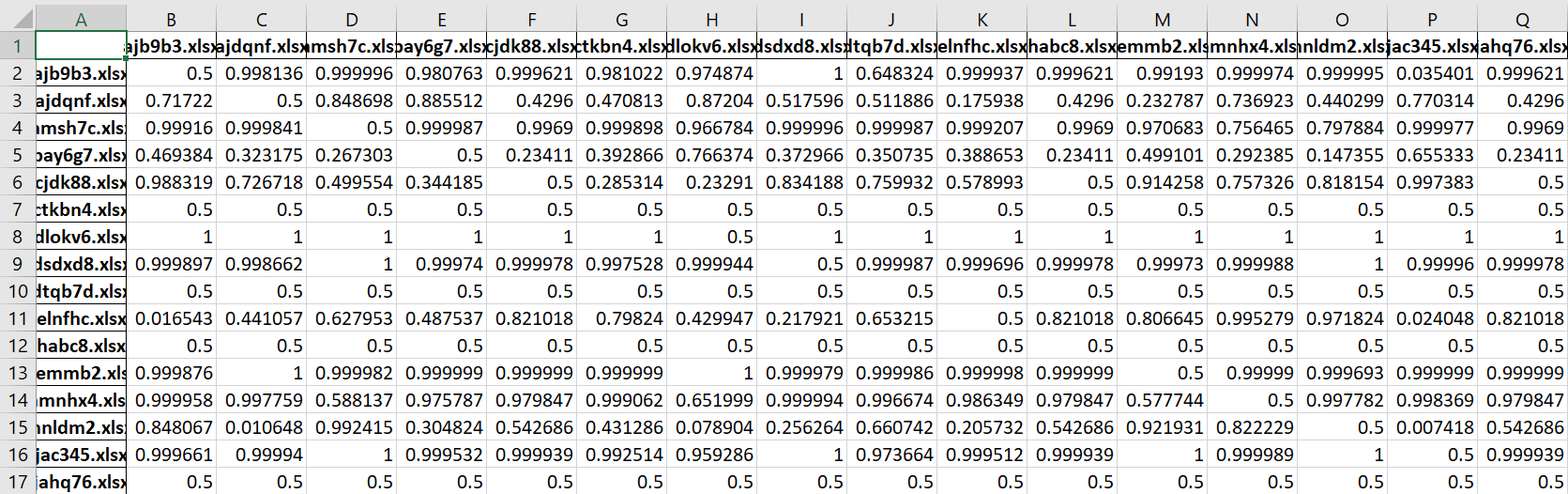
1. Finally, the Z value and P-Value is calculated using the three correlation values (r1a2a, r1a2b and r2a2b), which is the final output.

**Output Tables:**

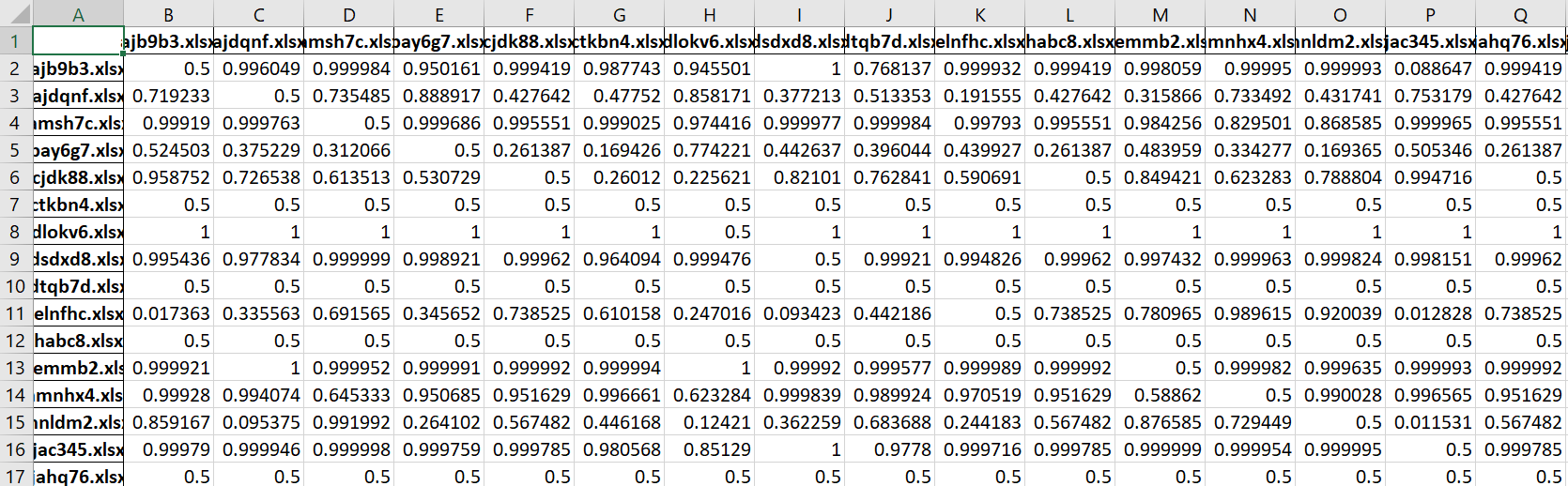
1. 10 Seconds Window:



1. 227 Seconds Window:

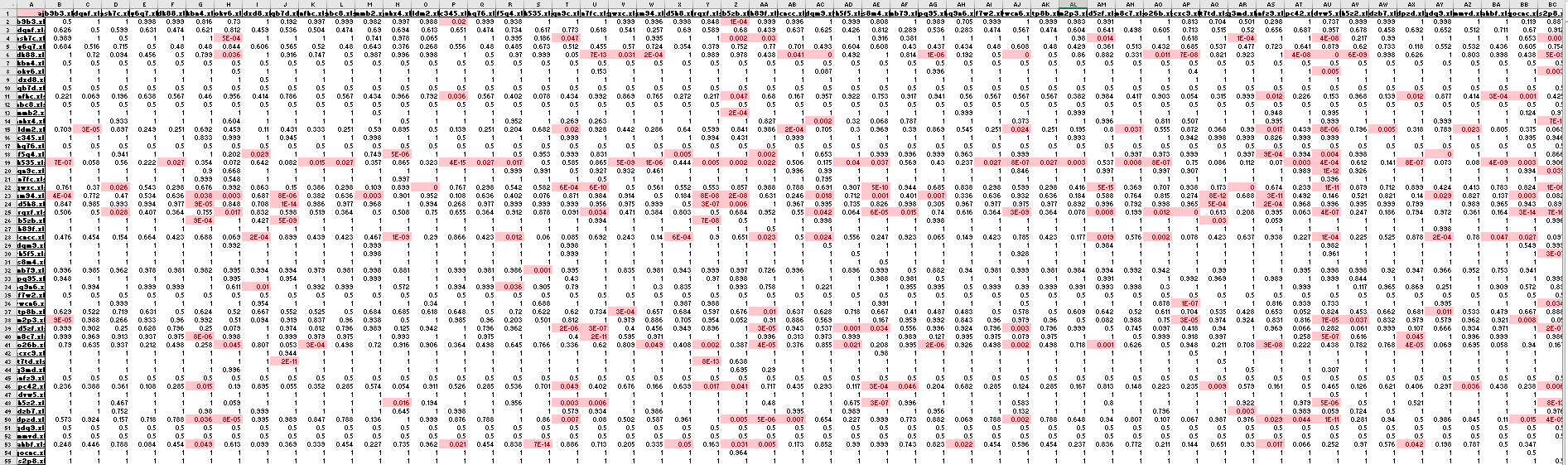


1. 5 Minutes Window:

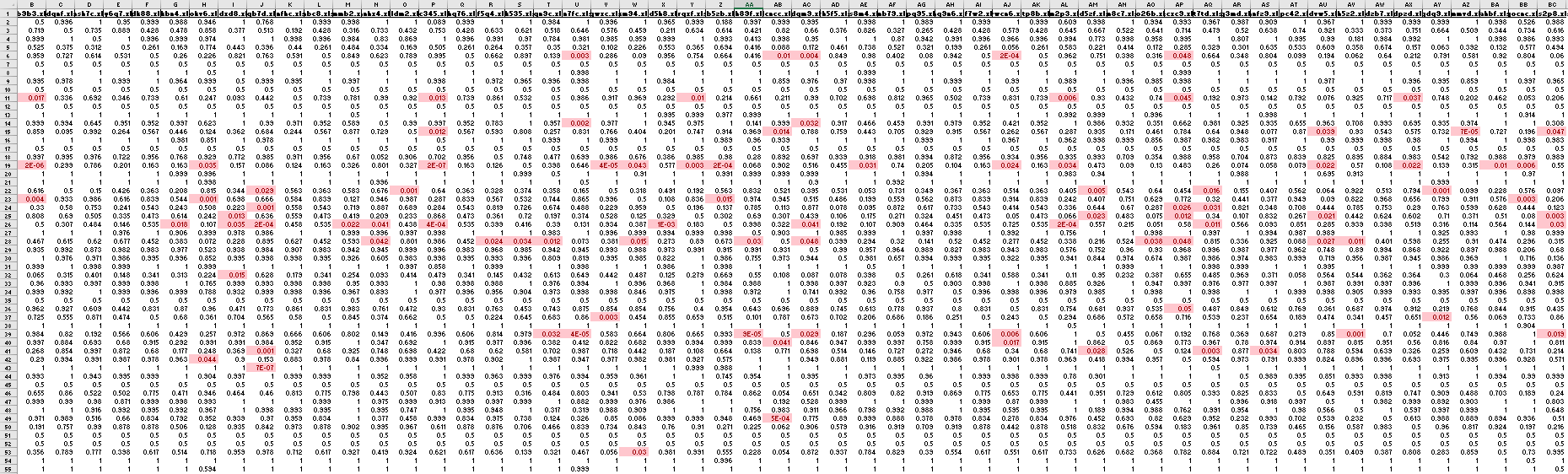


**Observations from the above table:**

10 Seconds:



5 minute:



All the highlighted red cells are the P-values which are less than 0.05. There are many red cells in the 10 seconds window, which tells us that this window is good for authentication as there are more number of distinguishable cells when compared to 5 minute window.

In 5 minute window we can see that there is lot of indistinguishability between the user data when compared to the 10 second window. So, the 5minute window is not good for authentication.

**File Descriptions:**

Since we had to correlation for 3 time slots (10s, 227s & 5mins), I had divided the files each respective time slot and placed them in different folders T1, T2 and T3 representing 10s, 227s and 5mins. Each of those folders accomplishes the same thing for their time slot. An example folder directory and its description is as follows,

T1 files include:

* Infow1t1.py: It divides the input into 10 second windows for the 1st week and stores it in W1T1 sub – folder.
* Infow2t1.py: It divides the input into 10 second windows for the 2nd week and stores it in W1T1 sub – folder.
* W1T1: Contains the output of infow1t1
* W2T1: Contains the output of infow2t1
* core1.py: It calculates the Sprearman Correlation of W1T1 and W1T1 files.
* core2.py: It calculates the Sprearman Correlation of W1T1 and W2T1 files
* core3.py: It calculates the Sprearman Correlation of W2T1 and W1T1 files.
* core4.py: It calculates the Sprearman Correlation of W2T1 and W2T1 files.
* Spear Correlation: Holds the outputs of core1.py, core2.py, core3.py and core4.py
* Zfinal.py: It is used to calculate the final P value using the files in Spear Correlation sub folder.
* Final Output: It holds the final output for T1 directory which is generated with the Zfinal.py

In the same manner T2 and T3 are structured.

***Note1:*** All the file directories used in the code are in accordance to my system. If you want to run the code, you just change the directories in those files.

***Note2:*** This entire project can be done in a single py file, but I had chosen this method so that I could cross check my intermediate outputs and see where I would be going wrong. It’s like a checkpoint method. But either way is fine.

***Note3:*** Calculation of Spearman Correlation without using built in one line method is also included in the comments of core python files, in the subfolders.