Phase 4,5 : Cleaning Data

- **We Clean the data Manually and with Python Code:**

In column : (What is your income level ?):

We Convert each Text Value to Numerical For example ,

Low=1 , Medium=2 , Lower medium=3 , Higher medium=4 , High=5

Setting a numerical value as a weight for each level of income making it easier to be analyzed in the future using a regression technique or any other technique that depends on a numerical values For it's algorithm to work.

In column : (Current Career of Your Future Career Plans/Field):

We Replaced the Miss Value With Null Value .

In column: (if you got the chance to rebuy the main computer device (laptop , pc)

you use at moment what would you choose instead , (or if you going to stick with your current device choose your device) and column(Which of these devices do you own?):

1- We used the data of these two columns to create a new column(would he change his device) To know who wants to change his device

2- If he does not change his device, we put a zero ,else put 1.

In column:(What is the MAIN usage of your device?):

We took the first answer for those who chose more than one option because we want to know what is the main use of their devices.

In column:(choose your processor) :

We replace the Intel values (core i3 , core i5 , core i7) with value 1 and we replace the other types of processor like(Intel xeon, Ryzon 7, Ryzen 5, Ryzen 3, AMd 10 ,AMD a8, AMD a6,AMD a4, Ryzen 9, snap dragon) with 0 value And we put the new values at the(is it Intel) column.

We did that to group it by generation

as the Intel processor starts from 1 to 13 th generation and the other types like AMD it starts from 1 to 7 th generation.

In Column: (the price of the device) :

There was small values that do not fit to be prices for devices such as 15 and 8, and it is likely that the person who entered you the number forgot the rest of the numbers, so we set them up to be 15000 and 8000, and there are also illogical values such as 1, so we deleted them and replaced them with null, in order to make it easier for us to deal with the numbers and the numbers are real and logical.

In column: (Enter your ID):

If ID don't start with 2020,2021,2022 or not understand value ,We put null.

In this column :(Sum of Min):

We added all the min value of how many hours do you use your devices for many tasks for this 5 columns (How often do you use your device for gaming daily ? , How often do you use your device for work daily ? , How often do you use your device for academic study daily ? , How often do you use your device for self study and courses daily ? , How often do you use your device for entertainment (other than gaming) daily ? And put them in one column.

In Column :(which of these devices do you own):

The data in the column represents types of devices each survey participant has.

we asked him/her through a list of choices which include laptop desktop and tablet and they can select the devices they own from it .

We created Three

columns(laptop, Desktop, Tablet) corresponds This column sothat we can express the data inside The column by (yes or No)

:If someone have laptop and Desktop only the corresponding value for the same person in column of “laptop” & ”Desktop” will be “Yes” but the column of tablet will be “No”…etc.

In column: (If you own both laptop and desktop computer ,which of them do you use more):

Some people choose that they own one of the two devices, the computer or laptop, and do not own the other in the column of “Which of these devices do you own?” At the same time, they choose that they prefer to use one of the two devices over the other in the column of If “you own both laptop and desktop computer, which of them do you use more" - This is because they did not read all the options, so they did not notice this option, so they chose another option that they thought was the closest to their situation. - That is why we amended this column: “If you own both laptop and desktop computer ,which of them do you use more” by comparing it with the column: “Which of these devices do you own” so that everyone who wrote that he owns one of the two devices and does not own the other in the column : "Which of these devices do you own" and did not choose that he does not own both in the column: "if you own both laptop and desktop computer ,which of them do you use more" We modified his answer so that he does not have both

In column :(What is the MAIN usage of your device?) :

so that each cell contains only the main use of the device that the user owns instead of many several usages, And I considered that the first use the user wrote is actually the main use , I did this modification since the most important thing for us in the analysis phase is the main use , it is done by python .

In column:(Use the best choice to describe your device's performance.)

where each cell includes only the first 3 or 4 words instead of a long sentence, as these words are sufficient to express or describe the performance of the device and we do not need the whole sentence in the analysis.