**Yahya Alhinai**

**EE5373**

**September 17th, 2019**

**Lab 02**

**OBJECTIVE FOR THIS LAB:**

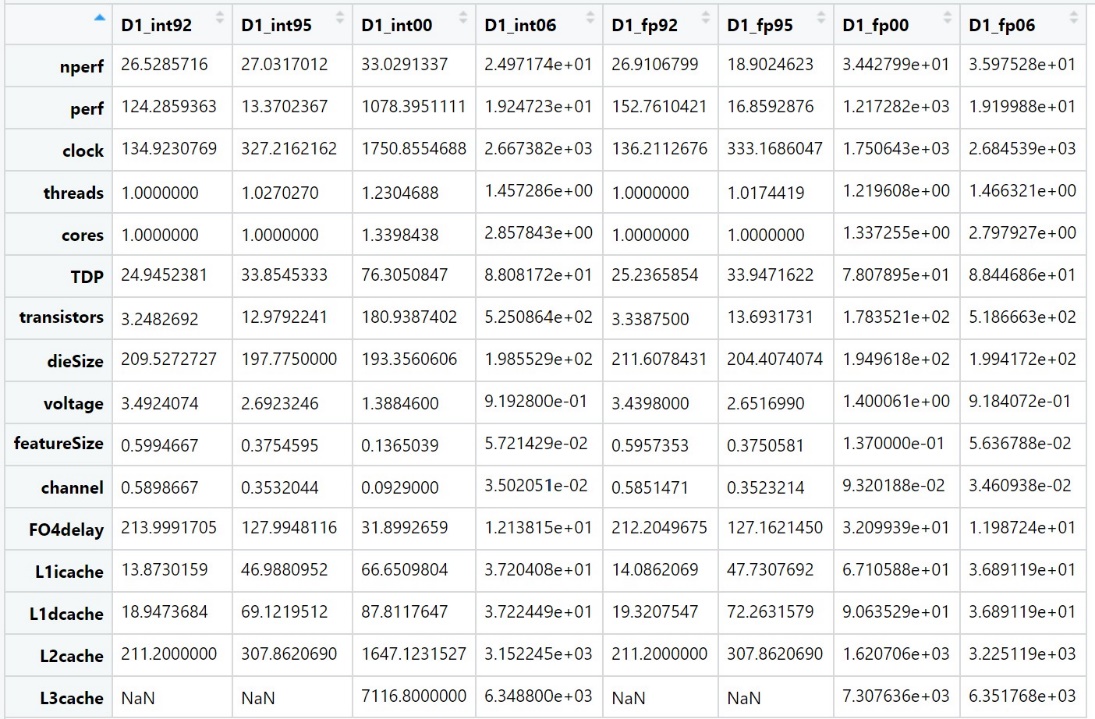
* We will be conducting various analyze for the data provided. Analyze consist of look for unusual patterns, outliers, and missing data points. Multiple procedures have been conducted for each of data set.

**TESTED THE COLUMNS:**

**Mean:**

* Several built-in functions were used to obtain the mean of every column of every processors data type. The following functions were used:
  + mean(): return the mean value for the given data.
  + sapply (): to apply mean for every column of any data table.
  + cbind(): combining all the resulted tables into one big table.
  + na.rm = TRUE: to remove all the NA value from the resulted table

the following table is the results of obtained mean for every column of all the provided data:

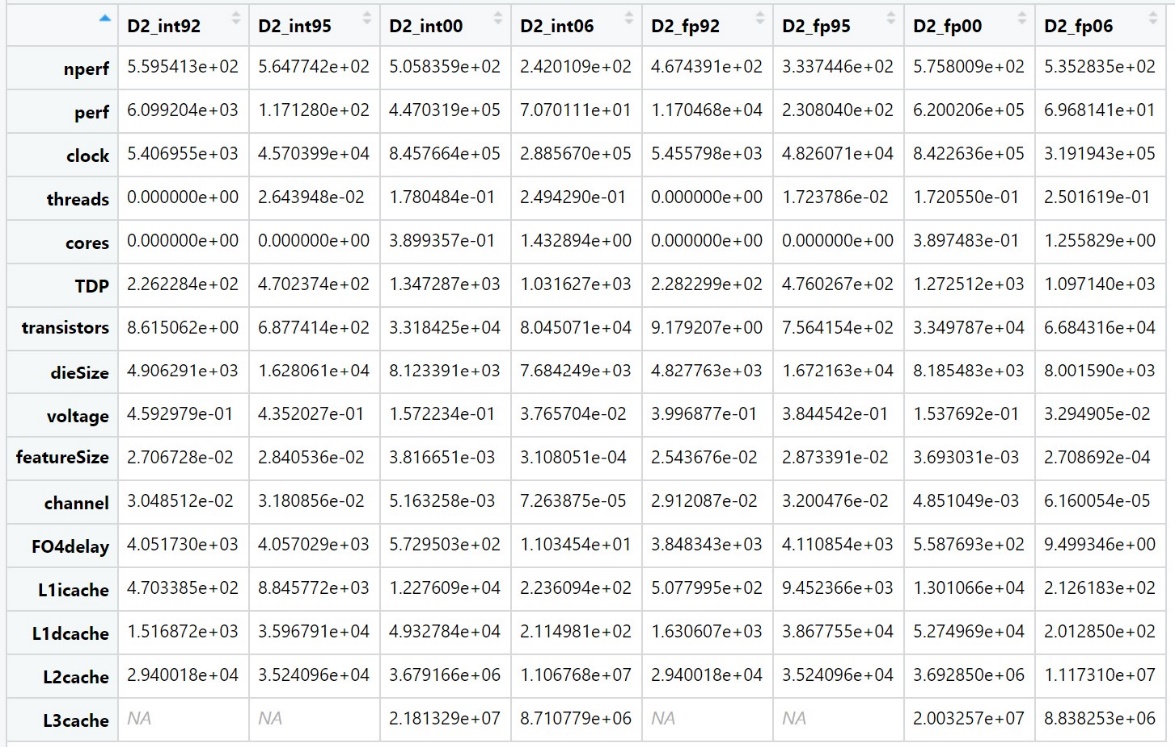
****

Removing all the NA values is crucial step to implement; Otherwise, if any column has a single NA value, the whole column would be valued to NA when mean() is called.

**Variance:**

* the same procedures that is implemented to obtain the mean for all of the provided data, variance is conducted similarly. We replace the mean with the following:
  + var(): returns the variance for any set of data provided.

The following table is the result of the produced results:

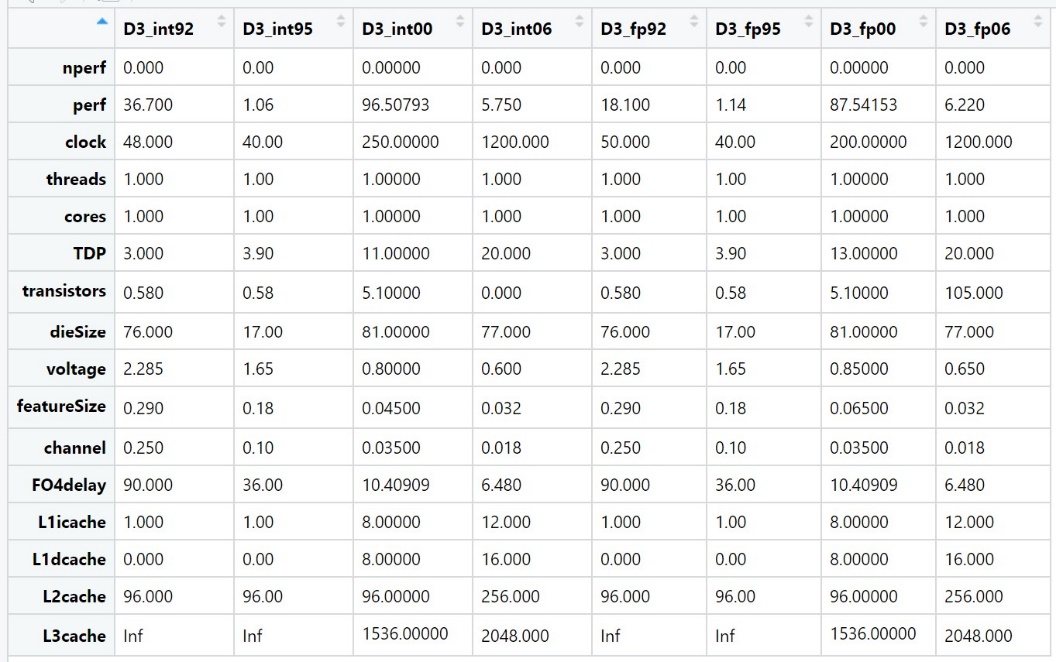


All NA values were removed for the same reason explained in the last section. One thing worth mentioning is that L2chahe did not exists in processors in the year of 1992 and 1995; therefore, its labeled as NA because no single value is found in the those column.

**Minimum:**

* similar to Mean and variance sections, set of built-in function were used to obtain the desirable value. To get the minimum value the following function was used:
  + min(): out of all the data points given to this function, it retunes the minimum value of the data.

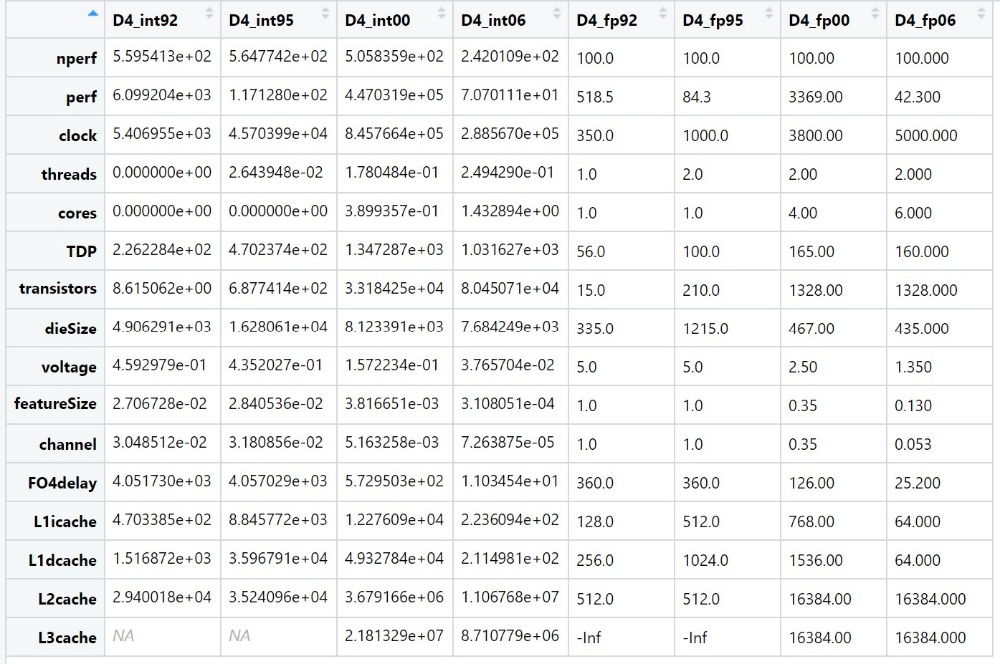
The following table shows the value returned after calling min function after NA value were not computed:



**Maximum:**

* Like minimum section, maximum, built-in function, is used here to return back the maximum value in every column. To get the maximum value the following function was used:
  + Max(): return back the highest value from the data given.

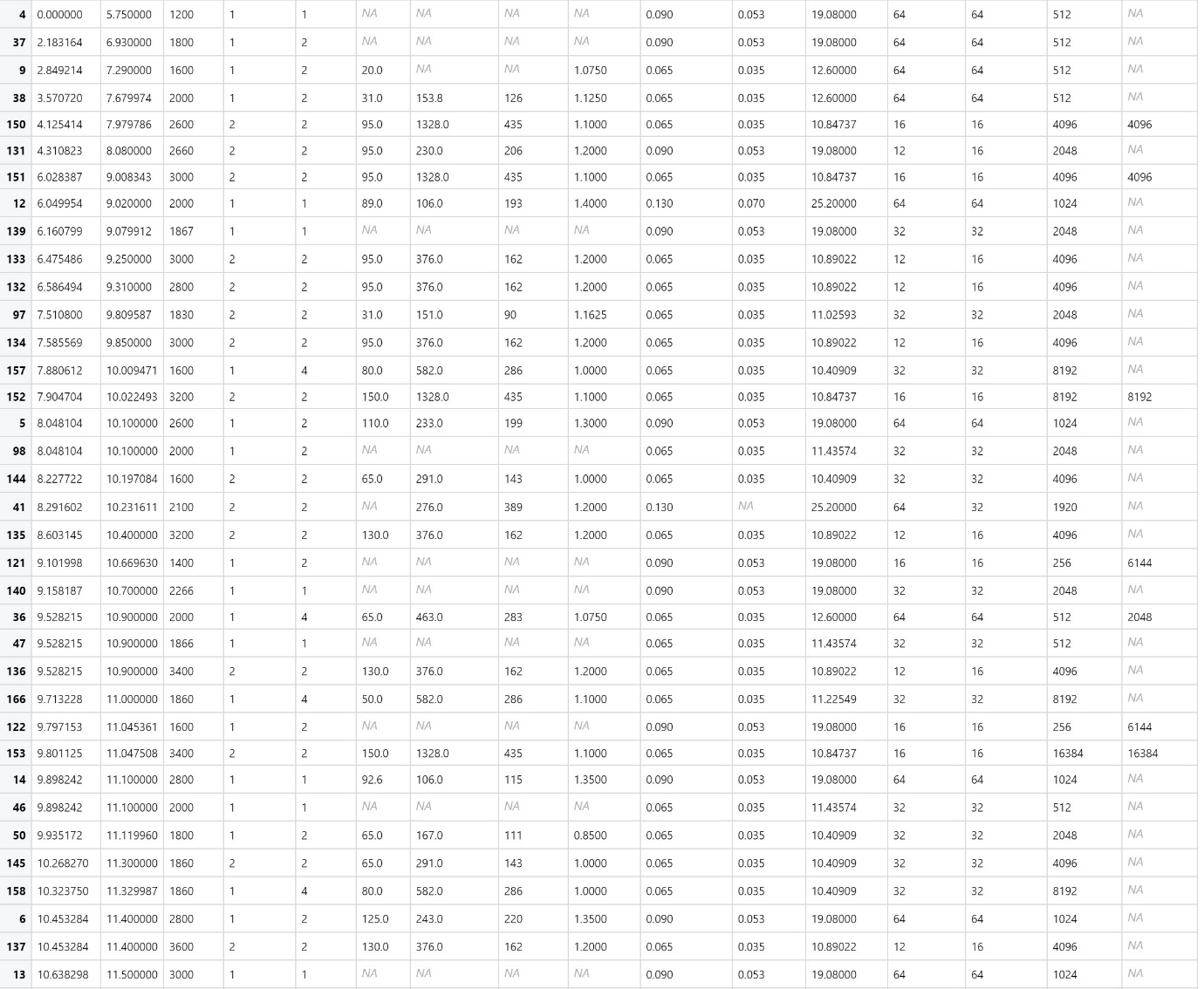
After NA were removed from data computed, the following table is describing the maximum value of all columns of all set of data:



**Sorting:**

* The RStudio interface has a build-in feature in the table view where you can sort each column at time. By having elements sorted, it is easy to look for outliers if there is any.

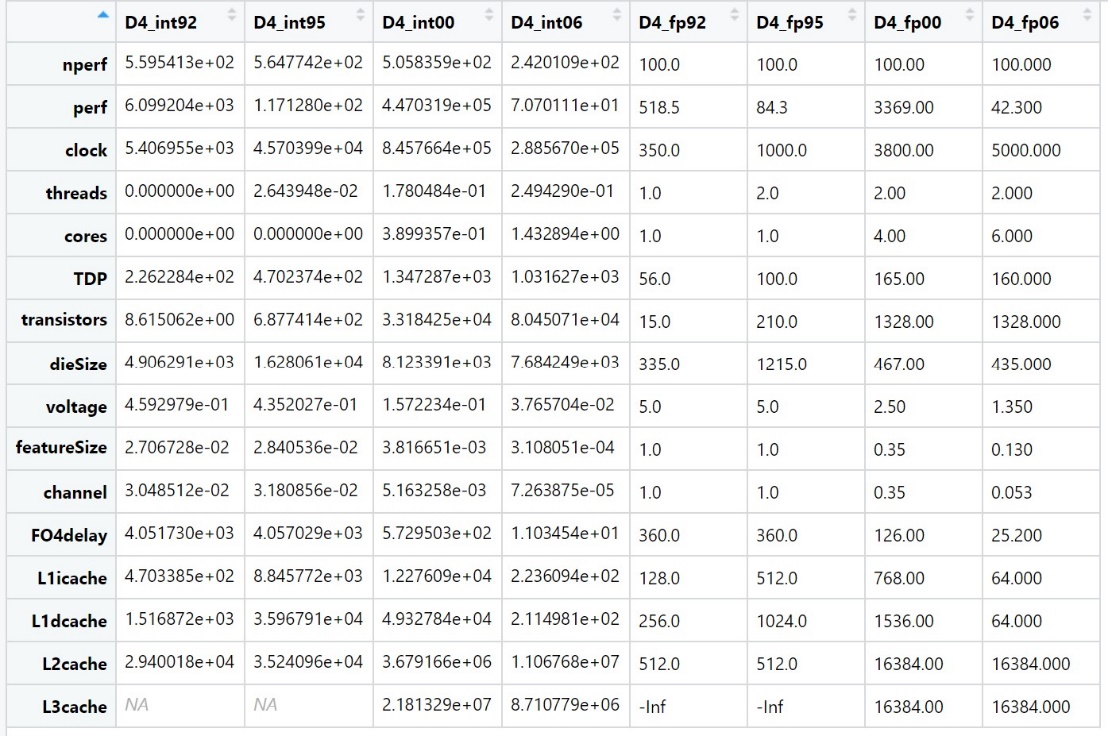
An example of part of a sorted column is shown below. The table is sorting “nperf” in the set of data “int06.dat”. we can see the non-descending order for npref column with all the rest information next to it.



**Fraction of NA values:**

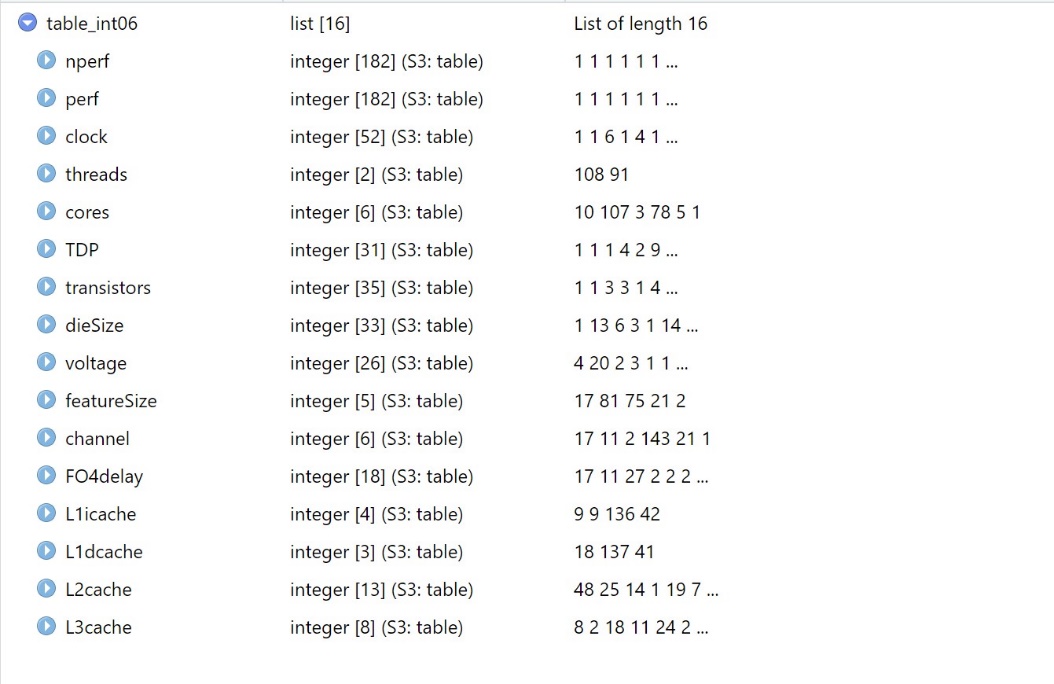
* This section is meant to identify the percentage values of the data missing (NA) to identify if there is enough data to draw a conclusion. The following function were used to compute the missing value:
  + colMeans(is.na()): this is actually two different functions. Is.na() will return the percentage of data points that is value as NA out of all data points. colMeans() will apply whatever function (in this case is.na() function) is pass into it to be applied to all column.

The following table provide the parentage of the data missing in every column of all the data sets:



**Table() distributions:**

* table() is a built-in function that shows how frequent a specific value shows in every data set. This is useful in identifying if there is a tend or a value that is mostly repeated. An example is the following table which describe the frequency of values in every column of data set “int06.dat”:



**IDENTIFIED AND DESCRIBED UNUSUAL COLUMNS:**

**Mean:**

* The clock seems to grow exponentially between the years 1992 to 2006 for integer and floating point. This is correlated with the number of transistors where it grows exponentially.
* The voltage dropped significantly for all possessors from 3.5V in 1992 to 0.9V in 2006.
* In 1992 and 1995, L3 did not exist. It started in the year 2000.
* dieSize stayed the same in between the year 1992 to 2006 for all possessors.

**Variance:**

* The most widely distributed result is in transistors section. In 1992, the variance is within 1 digit number. In 1995, the variance is 3-digit number. In 2000 and 2006, the variance within 5-digit number.
* The variance in channel decreases drastically. In 1992, it was 0.03; whereas in 2006, the variance was 0.00007 for integer possessors. Meaning the difference in channel between two possessors is diminishing as time goes by. The same trend found in floating point possessors.

**Minimum:**

* The most appealing observation goes to the clock. The minimum speed of the clock was 48 in 1996. On the other hand, minimum speed of the clock was 1200. The minimum speed of the clock increased 25 times between the year 1992 and 2006

**Maximum:**

* The most interesting thing to notice is the cashed L1, and L3. As L3 cashed emerged in 2000, there was a drastic deceased in the maximum size of the L1 cashes. The decrease is in the maximum size of L1 caches is 55 times between the year 2000 and 2006 for integer and 12 time decreases for the floating point possessors.

**Sort the column to look for outliers or unusual patterns:**

* From the sorted columns, the most noticeable columnis clock. It seems the most preferred clock speed is 3000 across all possessors and throughout the years 1992 to 2006.

**Compute the fraction of NA values to see if there are enough values:**

* Cashes have the highest NA fraction especially in L2 and L3 in the year of 1992 and 1995 because the face that they were new technology and they were not widely implemented.
* Also TDP has very high NA fraction.
* Otherwise, it seems that there are enough values to look for patterns.

**Use the table() function to determine if the distribution of values appears to have any anomalies:**

* The most important observation to be made is in the core column. As we know, multi-core processors have started in the early 2000s. We can see that all processors before 2000 were single core processors. In 2000, processors started to have dual and quadrat cores but single-core processors were still the most widely used. In 2006, single- core processors were the least to be used as dual-core processors were the predominant as well as 4-core processors.
* The voltage has an interesting trend. The most common used voltage in the year of 1992 and 1995 was 3.3V. Then it dropped to 1.2V the most common voltage in the year 2000. In 2006, the most widely used voltage was 0.85V.

**DESCRIBED ANOMALIES IN THE DATA:**

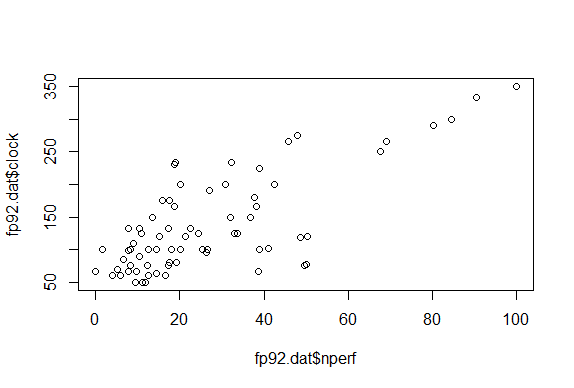
* The missing values (NA) from the origin data frame is an anomaly. We are not able to compute variance, mean, minimum, maximum, sorting, and distributions of the value with the present of NA values. This is was a major issue because if a single element in a column was missing then the whole column will be disregarded.
* Another thing showed up when analyzing the data was the warnings when trying to call min/max function to compute a specific column. It shows inf/-inf warning.

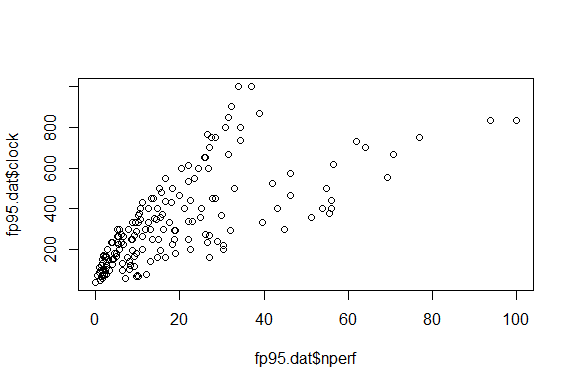
**EXPLAINED HOW TO FIX ANOMALOUS DATA:**

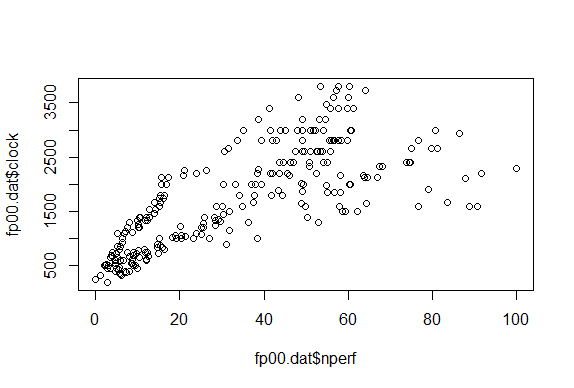
* To get by this issue of anomalous data, we added “na.rm = TRUE” for every function call. This will ignore the missing value as they are not part of the table. It will only take into account the non-missing value.
* To fix the warring when min and max function is called the instruction na.rm = TRUE” was assigned to all of function call which will be produce in the resulting data frame when I executed the function. This step will replace the non-existing value with “1” to all unvalued elements.

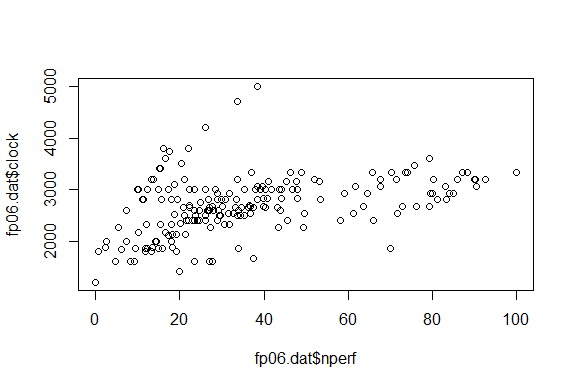
**YOU MUST INCLUDE ONE INTERESTING TYPE OF ANALYSIS BEYOND THE LIST ABOVE:**

* An interesting type of analysis I notices was the performance vs the clock speed. Before the real use of multi-core processors in the years 1992 to 2000, the clock speed was linearly linked to the performance. In the year of 2006, the increase in clock speed did not respond to a linear increase in the performance. The saturation of performance is because of the shit to multi-core processors.
* The Graph for performance vs clock for all given processors from the year 1992 to 2006 are shown below.

****

****

****

****

