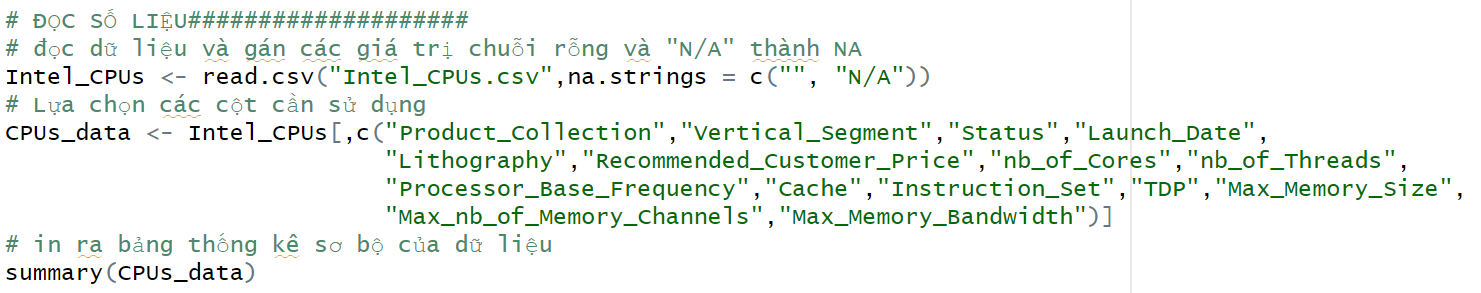
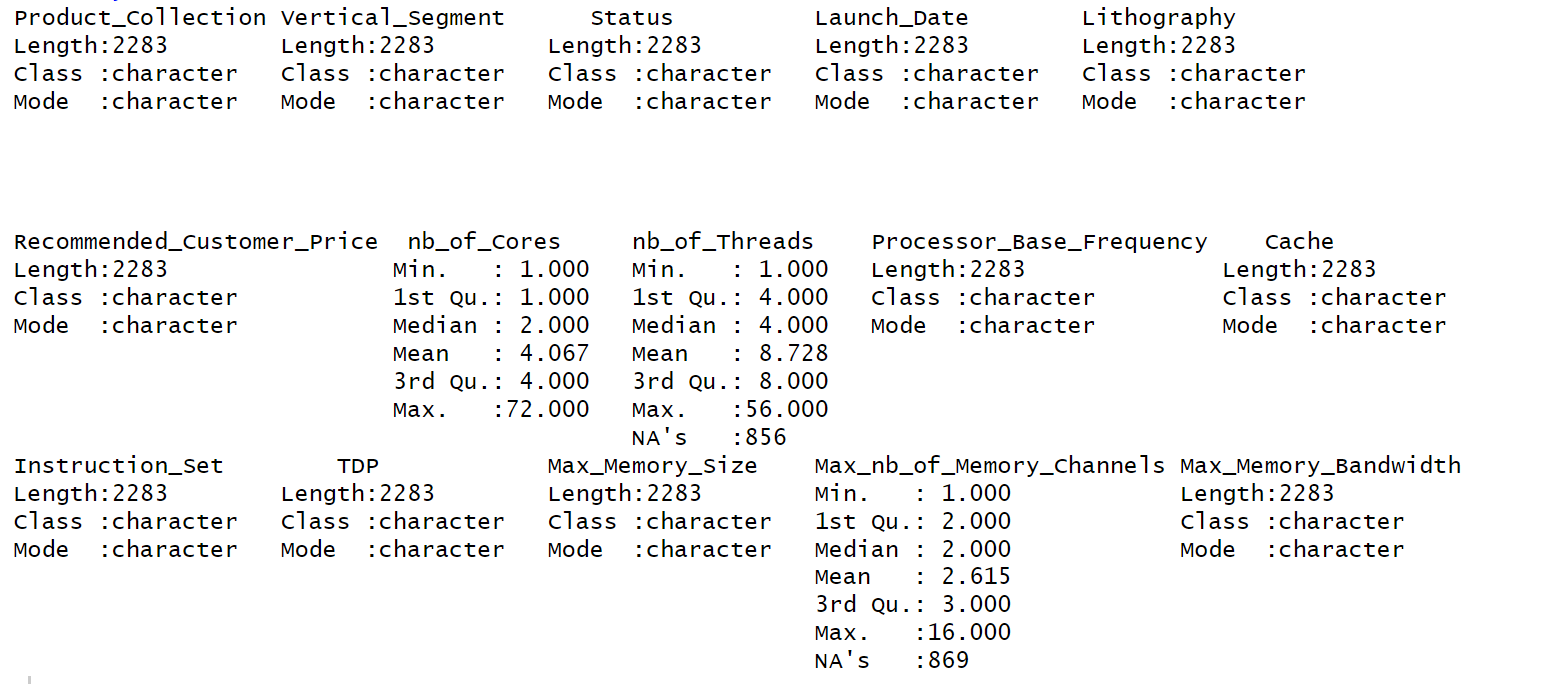
**3. Tiền xử lý dữ liệu**

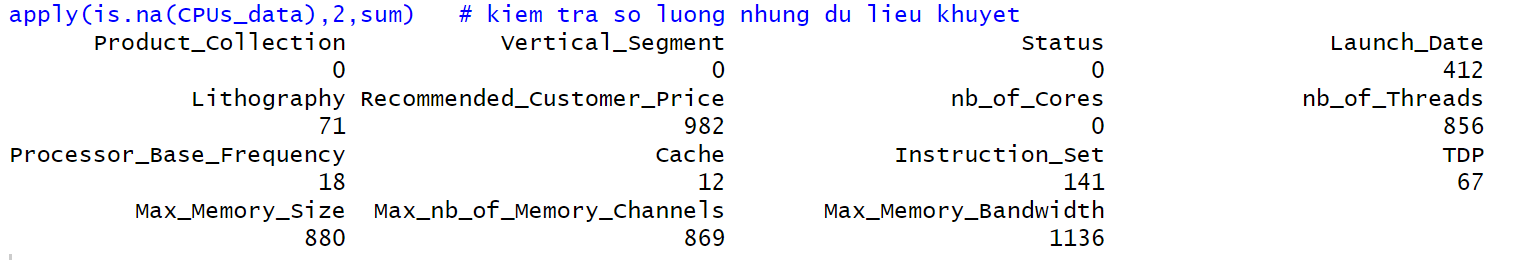
**3.1. Đọc dữ liệu**

Đọc dữ liệu từ file Intel\_CPUs.csv bằng hàm *read.csv()* và gán các giá trị chuỗi rỗng hay “N/A” thành NA bằng cách sử dụng tham số na.string(). Đồng thời chọn các cột cần sử dụng và in ra các thống kê sơ bộ.





**3.2. Xử lý dữ liệu khuyết**



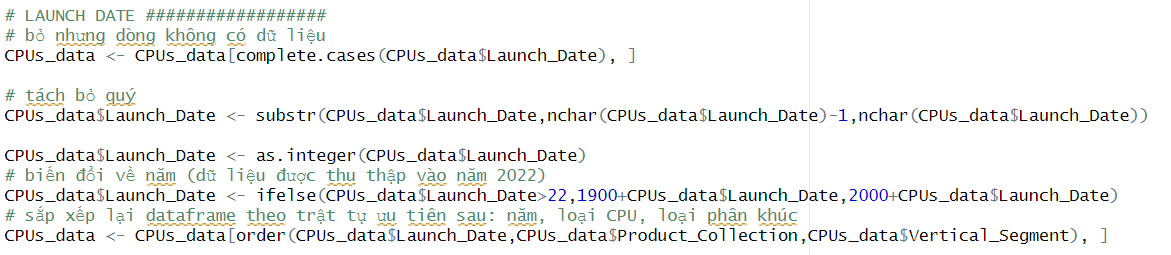
Nhận thấy nhiều cột bị khuyết dữ liệu, như Launch\_Date, Lithography,

Recommended Customer Price, Processor Base Frequency, Instruction Set, Max\_nb\_of\_Memory, Max\_Memory\_Bandwidth. Ta có thể lấp vào đó nhưng dữ liệu mới bằng cách sự dụng mean (giá trị trung bình), median (trung vị) hoặc tìm hiểu các mối quan hệ để điền vào.

- Launch\_Data

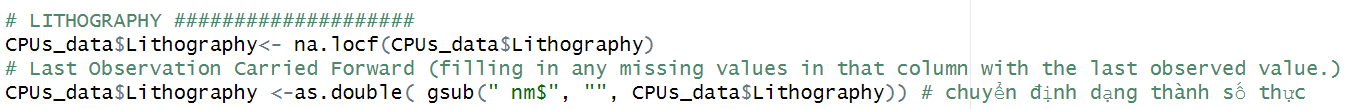
Sử dụng hàm *substr()* để trích xuất chuỗi con từ chuỗi các ký tự, nchar() để tích độ dài của chuỗi. Ở đây ta trích xuất hai ký tự cuối cùng trong dãy.

Chuyển chúng về năm, dữ liệu được thu thập vào năm 2022 nên mọi giá trị lớn hơn 22 sẽ được xem là được phát triển vào thập niên 20



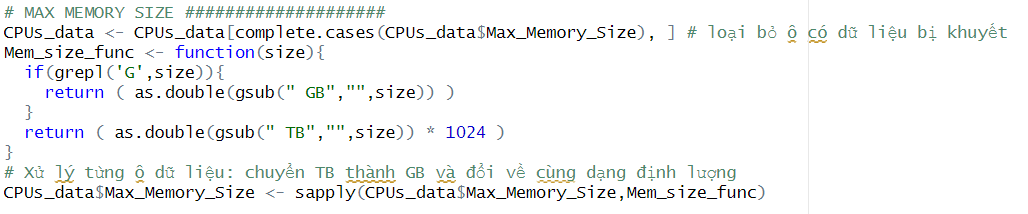
- Lithography

Lithography giảm dần theo từng năm cho các loại CPU hay phân khúc vì sự tiến bộ của công nghệ nên có thể điền dữ liệu thiếu bằng dữ liệu của năm trước. Sử dụng hàm *na.locf()* để điền vào những ô có giá trị còn thiếu bằng giá trị quan sát được ngay trước nó. Chuyển giá trị của Lithography về dạng định lượng của số thực và cắt đi phần đơn vị.

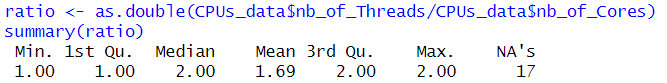


- Max memory size

Dữ liệu vào những năm nhỏ hơn 2009 mất đồng loạt nên việc lấp đầy là không khả thi. Ta bỏ tất cả những ô dữ liệu bị khuyết. Chuyển đổi giá trị của các ô có đơn vị TB thành GB bằng cách nhân cho 1024. Sau đó bỏ đơn vị và chuyển các gái trị của cột về dạng định lượng



- Number of threads

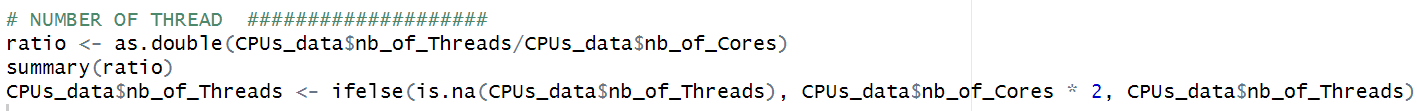


Dựa vào dữ liệu thu được, ta có trung bị của tỉ lệ giữa số lượng nhân và luồng là 2 có nghĩa rằng một nhân thường có xác suất cao sẽ có 2 luồng đi với nó.

IMG_263

Sử dụng phương pháp thống kê như hệ số tương quan Pearson ta nhận được hệ số tương quan cao.

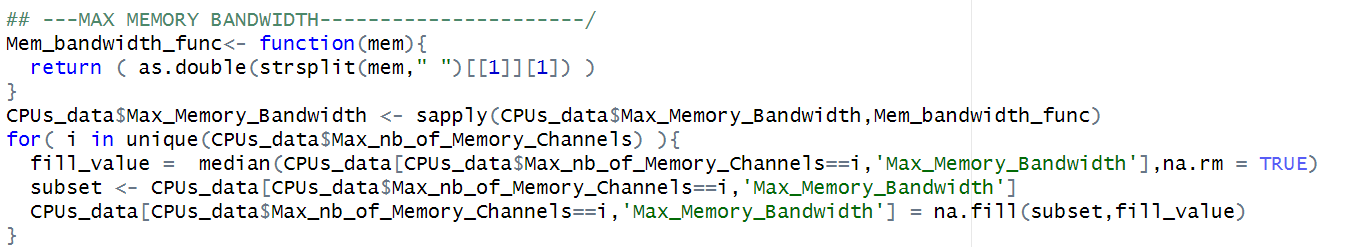
Vì dữ liệu cột nb\_of\_Cores không bị mất và do có độ tương quang cao nên có thể dễ dàng điền vào nb\_of\_Threads bằng cách nhân 2 lần số nb\_of\_Cores.



- Max memory bandwidth

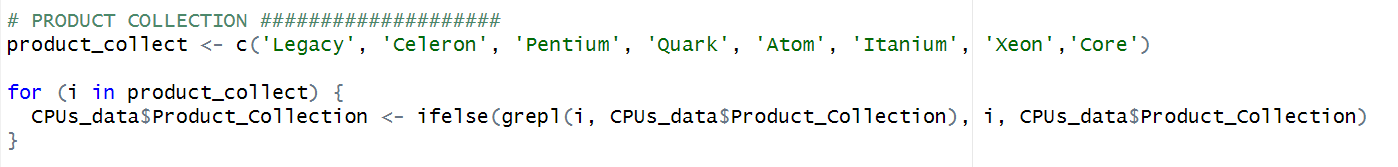
IMG_265

Sử dụng phương pháp thống kê như hệ số tương quan Pearson ta nhận được hệ số tương quan cao. Thế nênk hi tăng số lượng kênh tối đa sẽ dẫn đến tăng bandwitdth, vì vậy ta sẽ điền dữ liệu theo median của Max\_Memory\_Bandwidth theo Max\_nb\_of\_Memory\_Channels



- Product collection

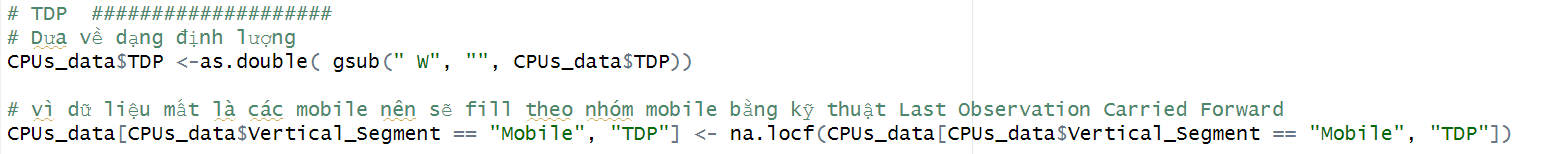
Các dòng CPU có thể được chia thành các nhóm: Legacy, Celeron, Pentinum, Quark, Atom, Itanium, Xeon, Core



- TDP (Thermal design power)

Xoá đi đơn vị và đưa về định dạng số thực.

Dữ liệu chỉ bị mất ở nhóm mobile nên ta điền dữ liệu thiếu theo TDP mobile của năm trước năm trước bằng kỹ thuật Last Observation Carried Forward.

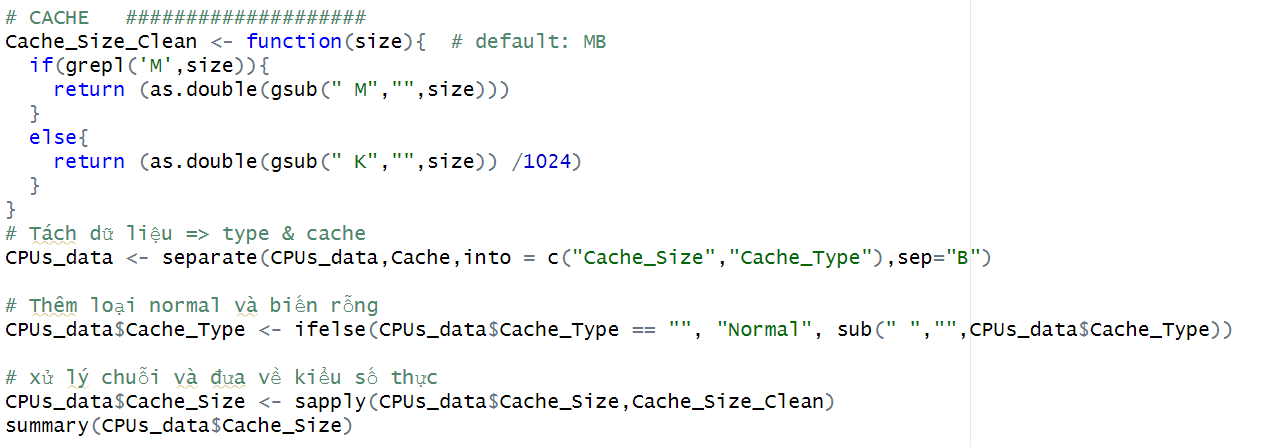


- Cache

Vì cache có 2 thành phần là loại cache và size của nó nên ta sẽ tách thành 2 cột: Cache\_Type và Cache\_Size.

Đối với Cache\_Size ta đưa về dạng số bằng cách xóa đi đơn vị của nó và chuyển sang định dạng số thực.

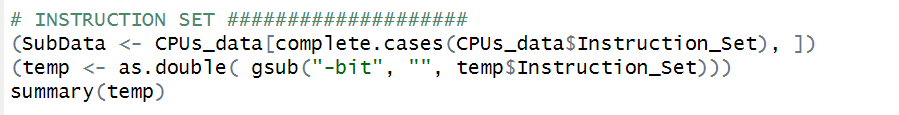
Đối với Cache\_Type, có một loại biến không chứa cache cụ thể (như: SmartCache, L2) nên ta thay thế nó bằng normal.

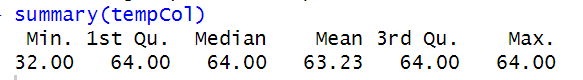


- Instruction Set

Theo bản số liệu thì instruction set chỉ có 2 loại dữ liệu là 32-bit và 64-bit.

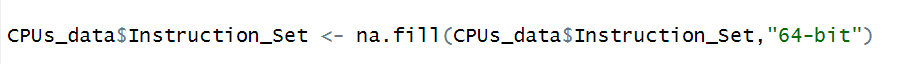
Thực hiện loại bỏ nhưng biến bị khuyết, ta lưu dữ liệu tạm thời vào SubData đồng thời chuyển dữ liệu về dạng số thực bằng cách bỏ đi đơn vị bit.





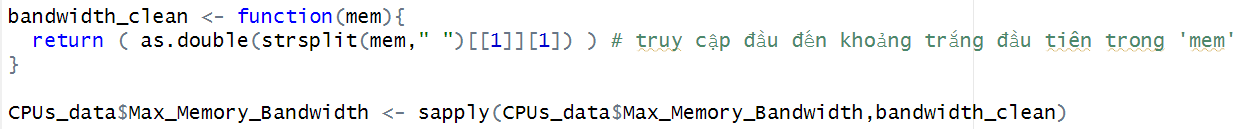
 Theo kết quả thu được, ba giá trị của khoảng tứ phân vị đều đạt giá trị 64 nên Mode của cột dữ liệu này sẽ là 64.

Vậy ta sẽ điền các dữ liệu còn thiếu là 64-bit với lý do tập 64-bit được ưu chuộng hơn.



- Max\_Memory\_Banwidth

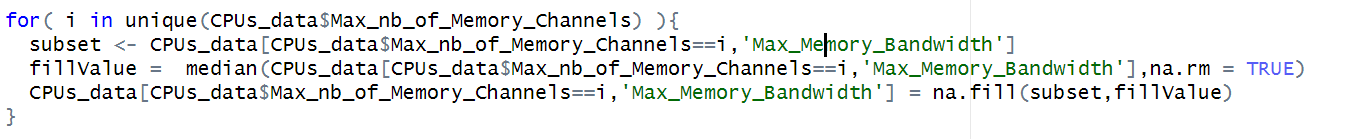
Tất cả dữ liệu đều ở dạng GB/s, ta xử lý chuỗi và đưa dữ liệu về dạng số thực bằng cách bỏ đi đơn vị



Max memory bandwidth có tương quan cao với max number memory channels, khi tăng số lượng kênh tối đa sẽ dẫn đến tăng bandwitdth.

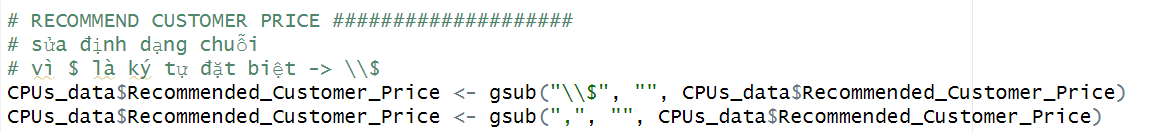
IMG_274

Vậy nên ta sẽ điền dữ liệu bị khuyết theo median của số Max Memory Banwidth theo từng nhóm của Max\_nb\_of\_Memory\_Channels.

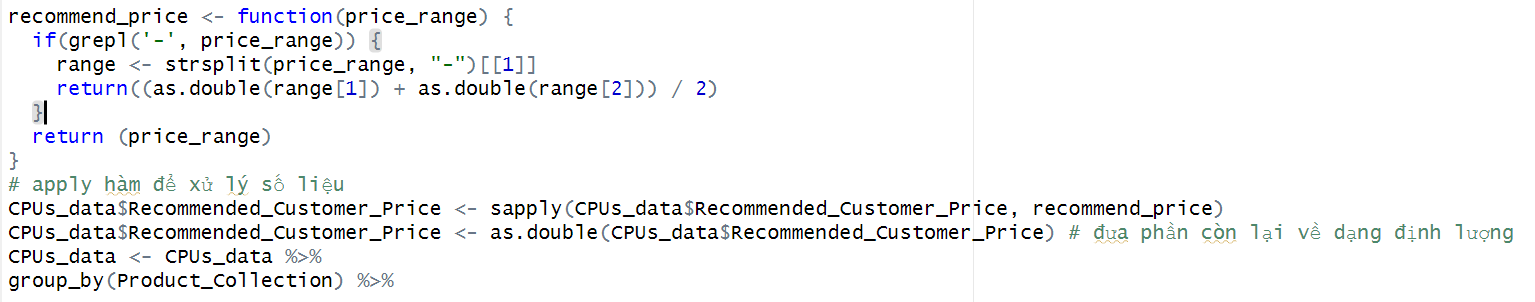


- Recommended Customer Price

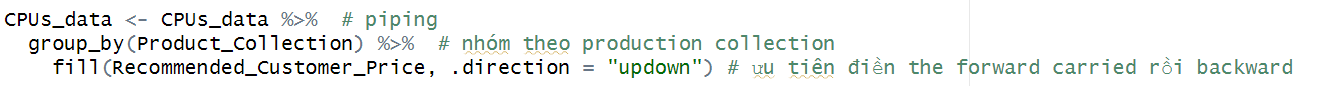
Xử lý định dạng chuỗi: xoá đi các kí tự ‘$’ và dấu phân chia hàng nghìn ‘,’.



Một số dữ liệu ở dạng khoảng giá thì lấy giá trị trung bình 2 đầu của khoảng.



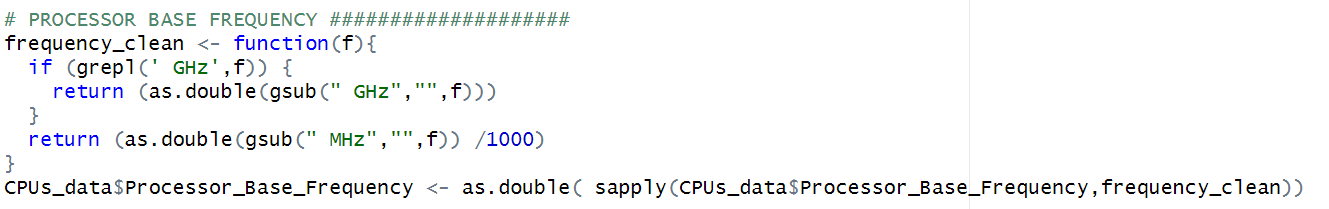
Giá tiền đề xuất được chia rõ ràng theo thời gian với các dòng CPU cùng loại, vì vậy ta áp dụng điền dữ liệu với từng loại CPU theo năm trước đó rồi theo năm sau đó



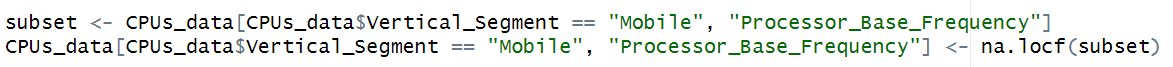
**- Processor Base Frequency**

Đưa từ dạng chuỗi về cùng một đơn vị GHz và bỏ đi đơn vị của tần số để đưa về

dạng số thực.

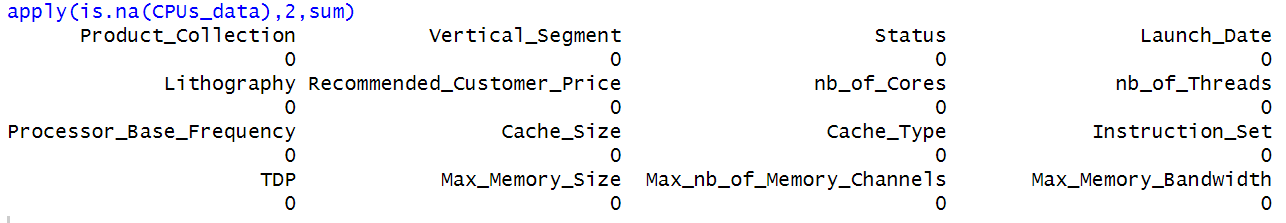


Dữ liệu bị thiếu ở loại mobile vì vậy ta điền dữ liệu thiếu theo tần số của mobile năm trước.



3.3. Kiểm tra lại dữ liệu

Dữ liệu đã không còn thiếu



Dữ liệu đã được đưa về đúng định dạng

