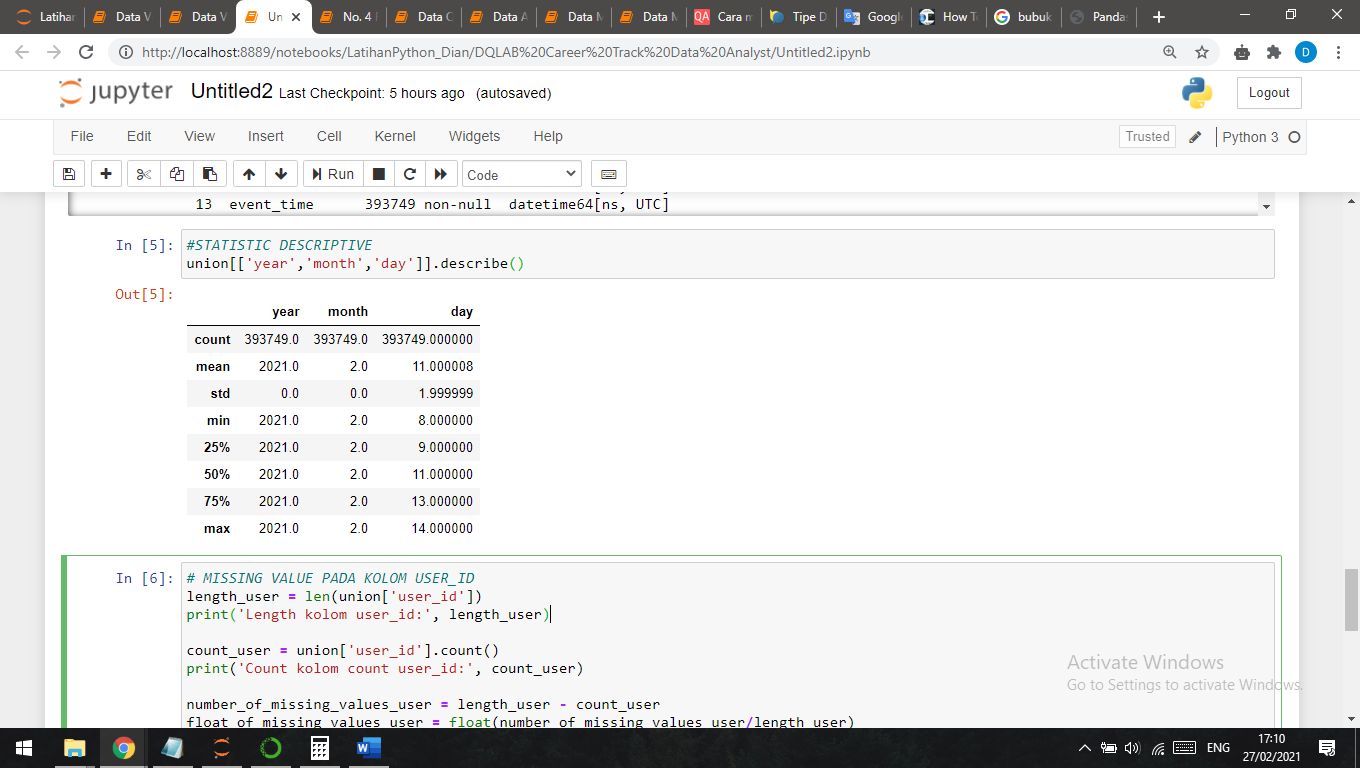
RESULT OF EDA ANALYSIS

I used the 1% sample data from 8th of February until 14th of February 2021.

There are total 19 variables, but I reduce it until it became 17 variables. Reasons: The column of device\_id have no values in each row and connection column has many null values so I have to delete these columns.

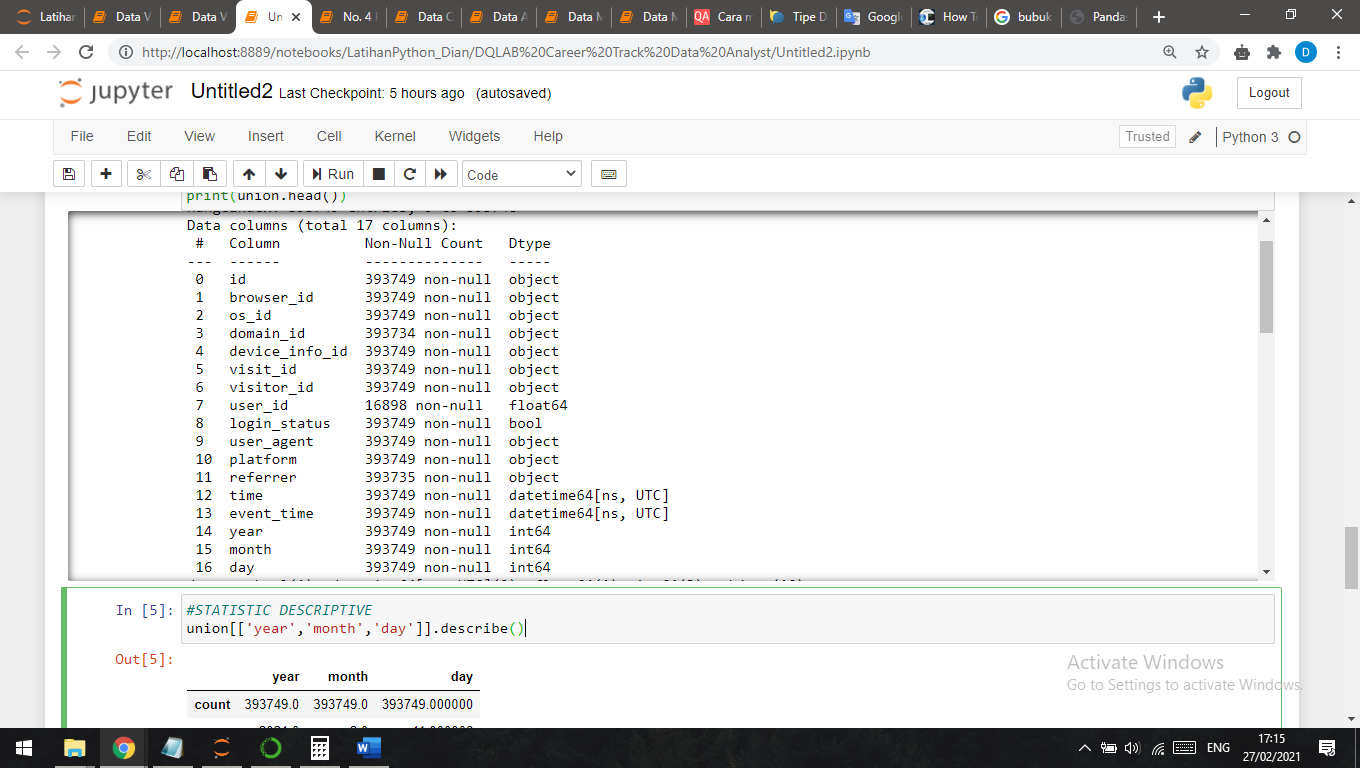
After doing “cleansing” to the data like deleting the unimportant columns, changing type of data from some columns that need to be changed (like time and event\_time to datetime type), I perform the EDA (Exploratory Data Analysis) using statistical method that python has already a feature to. First, I used .describe() to find count, mean, std, min, max for the columns that having int64 as its type (year, month, day) and I have these value to be presented:



Gambar 1.1

The calculation looks very simple, because this data is qualitative and no need to perform such a hard calculation.

And here’s the type of each column before I start performing EDA

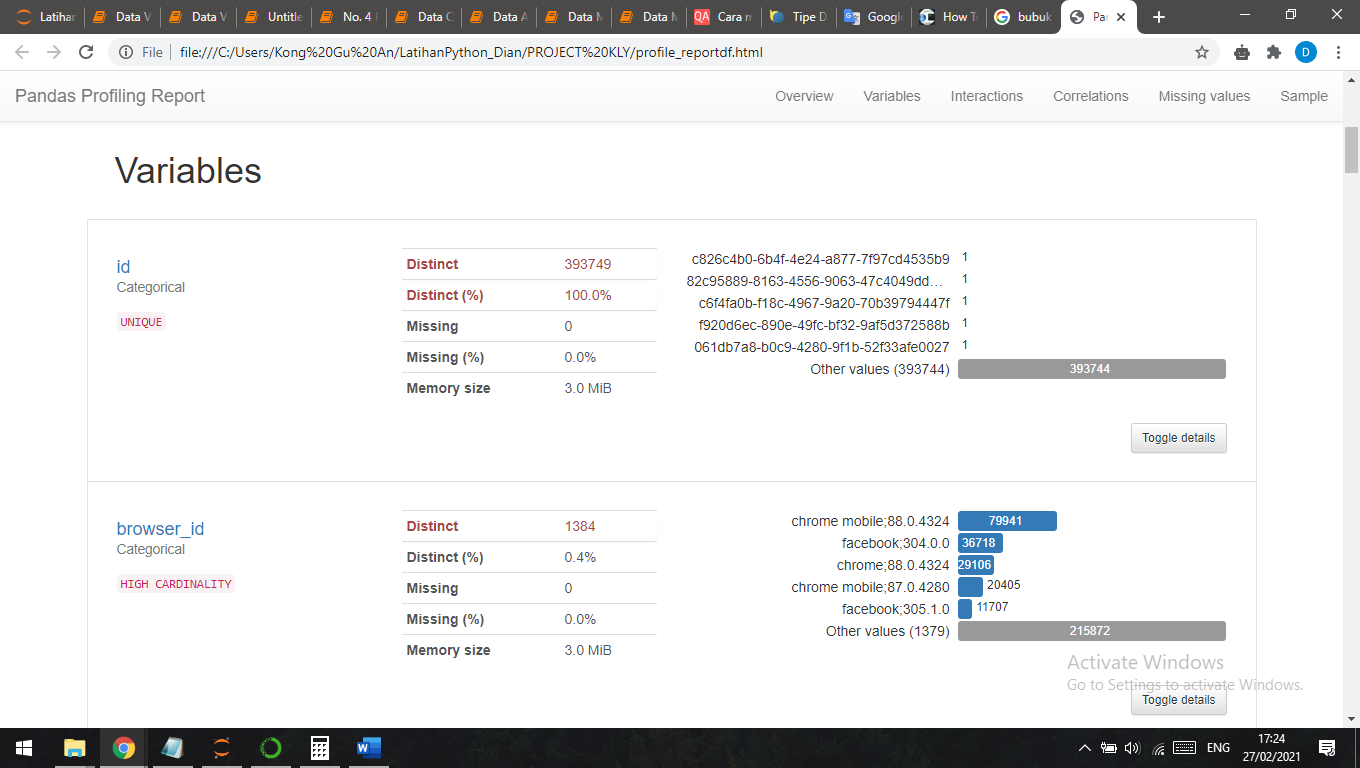


Gambar 1.2

Second, I used the library pandas profiling to resume all the calculation using statistical method in a file KLY\_Dian\_Nuryani\_question\_1.html that I have already upload it to my gdrive. And here’s the resume that I rewrite it again here to add some explanation that might be needed.

Start with the Variable Analysis

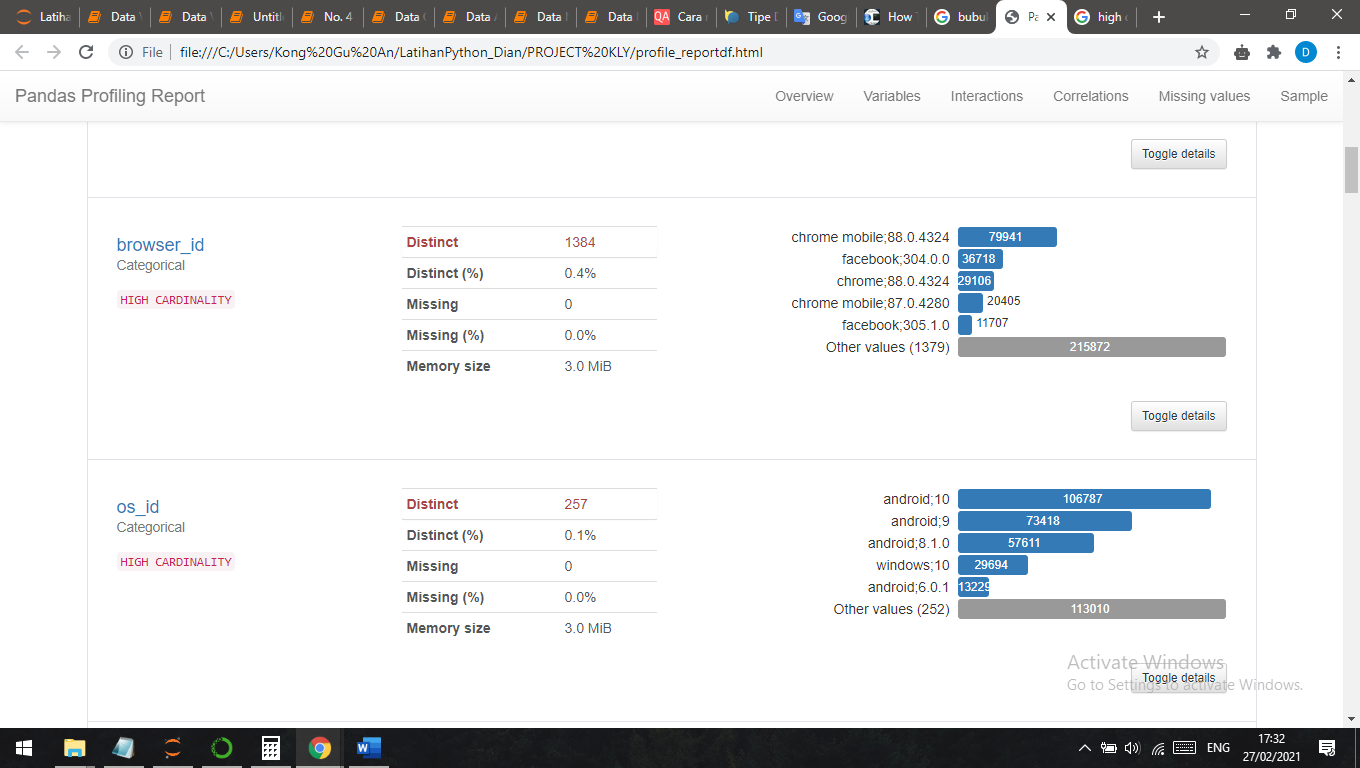
1. **id = record ID**



Gambar 1.3

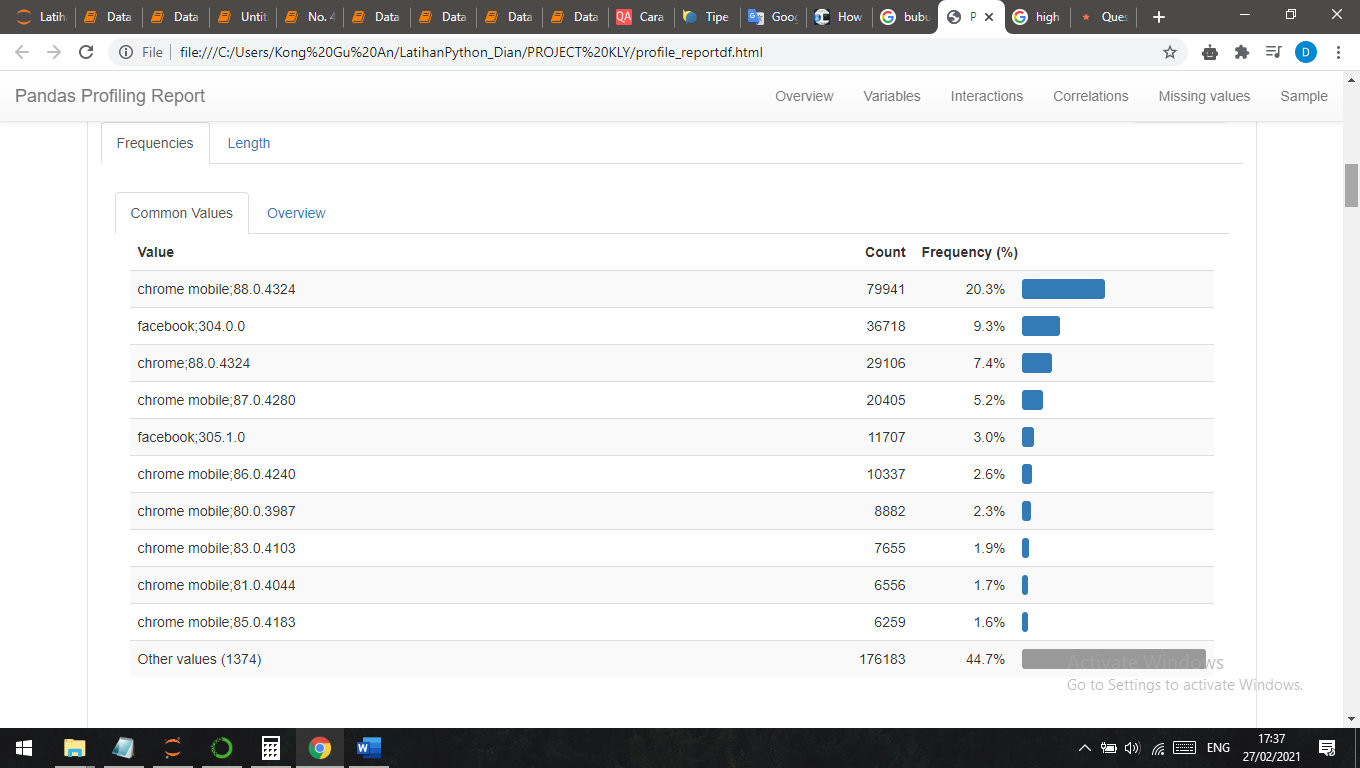
From the picture above we can see that all ids is totally **UNIQUE.** The distinct have total 393749 with percentage 100%, and no missing value.

1. **browser\_id = browser type**



Gambar 1.4

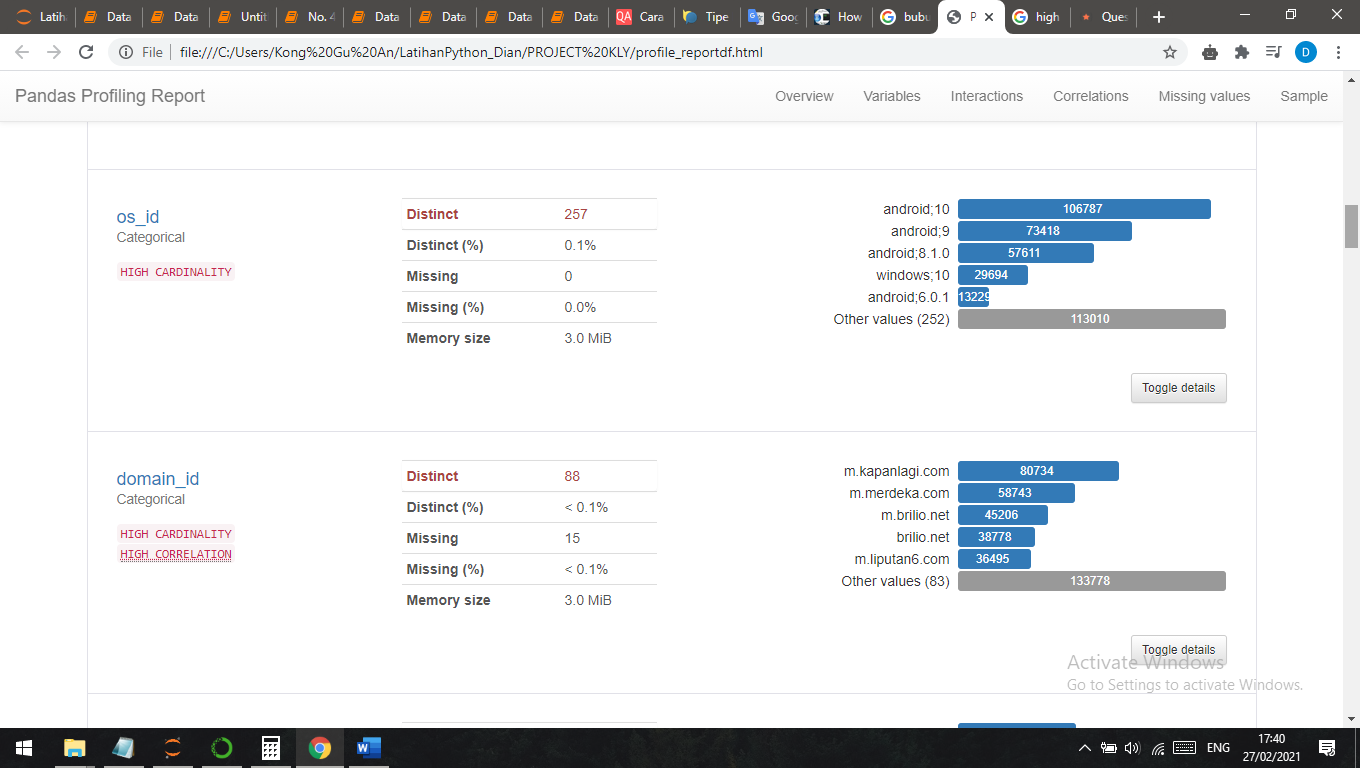
From the picture above, pandas profiling define browser\_id column as high cardinality. There are many ids who use the same type of browser, here’s the detail count of each browser type:



Gambar 1.5

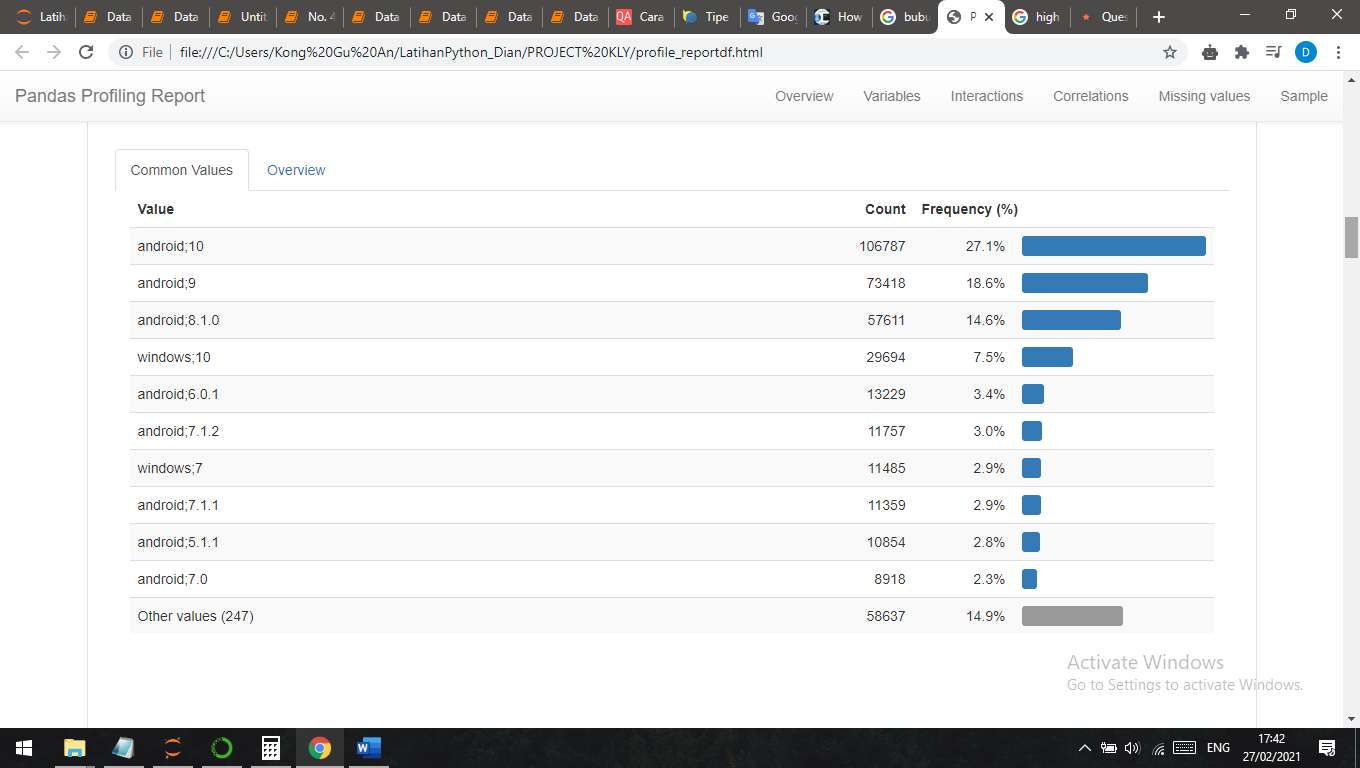
The top browser type is **Chrome mobile; 88.0.4324** with the frequency 20.3%

1. **os\_id = Operating System type**



Gambar 1.6

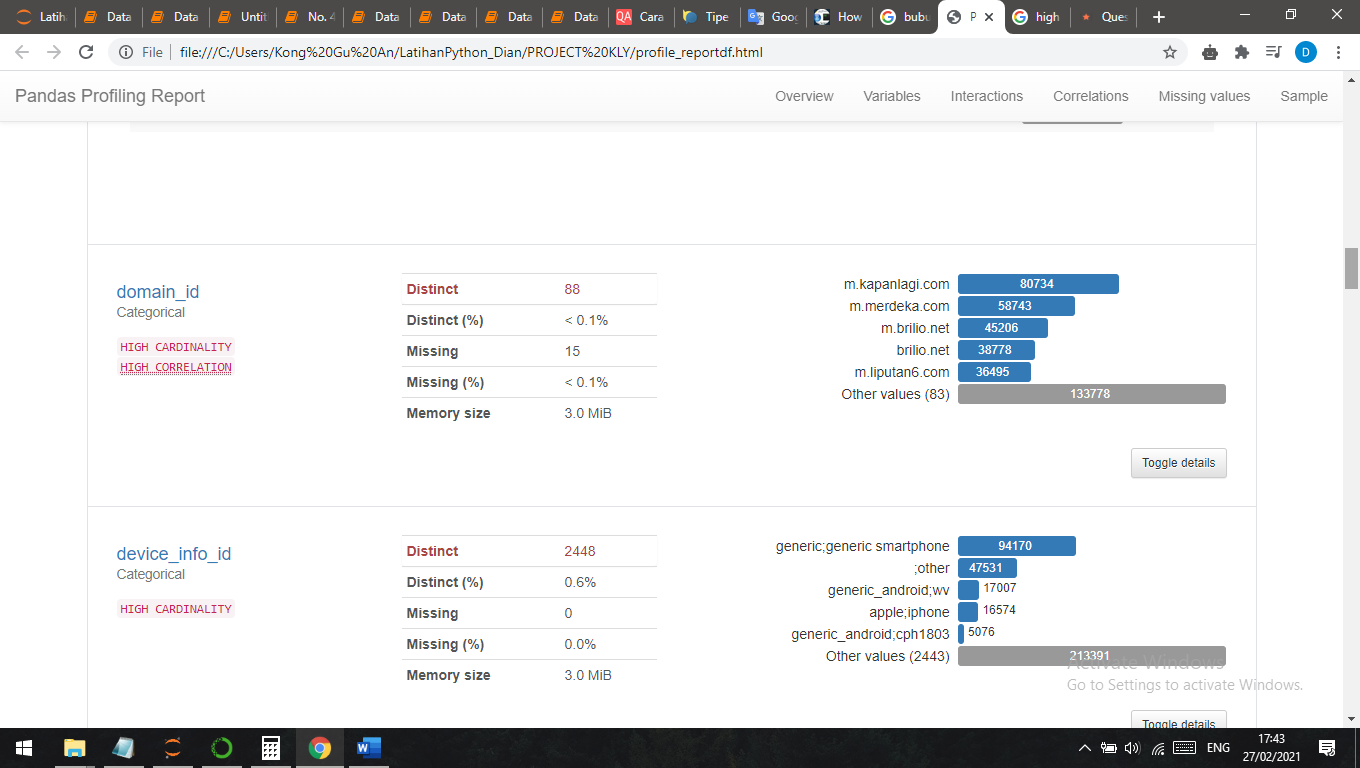
From the picture above, pandas profiling define os\_id column as high cardinality. There are many ids who use the same type of operating system, here’s the detail count of each operating system:



Gambar 1.7

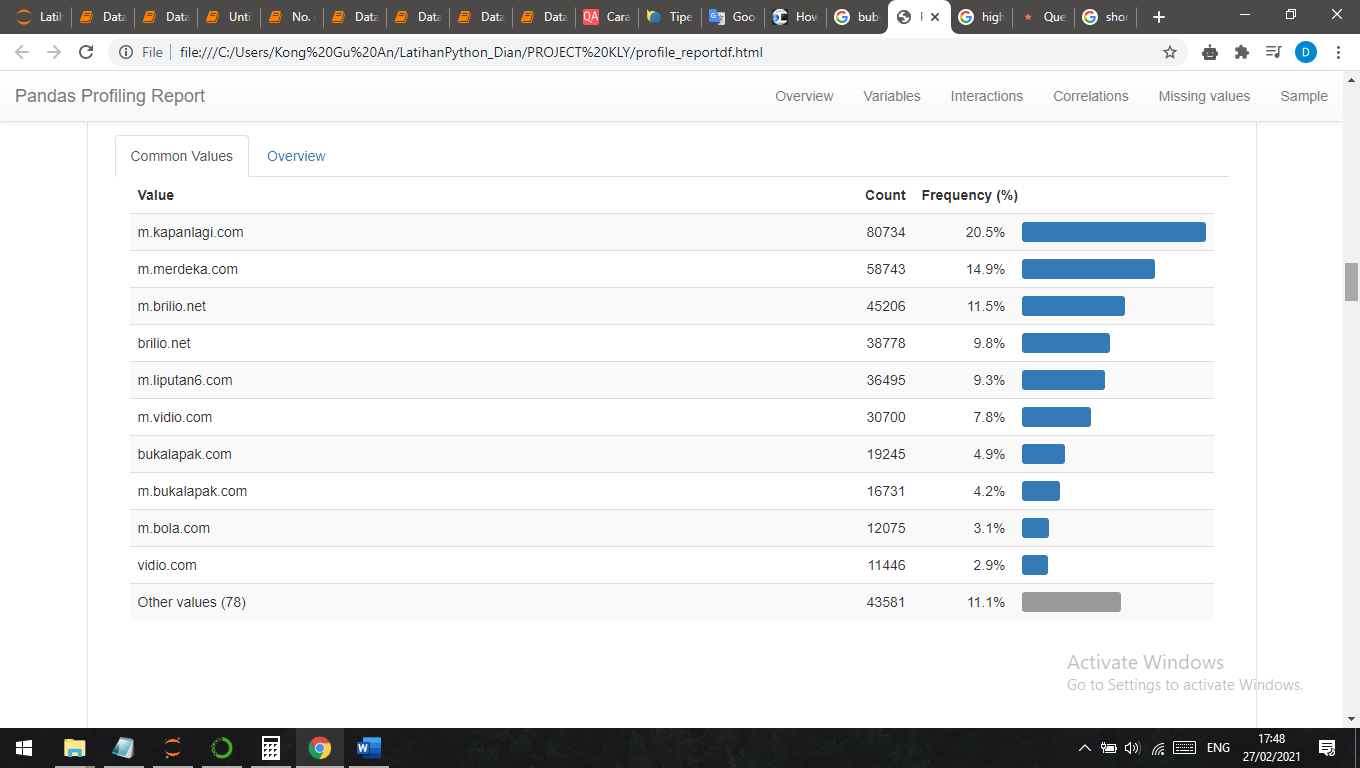
The top browser type is **Android; 10** with the frequency 27.1%

1. **domain\_id = domain or subdomain of webpage**



Gambar 1.8

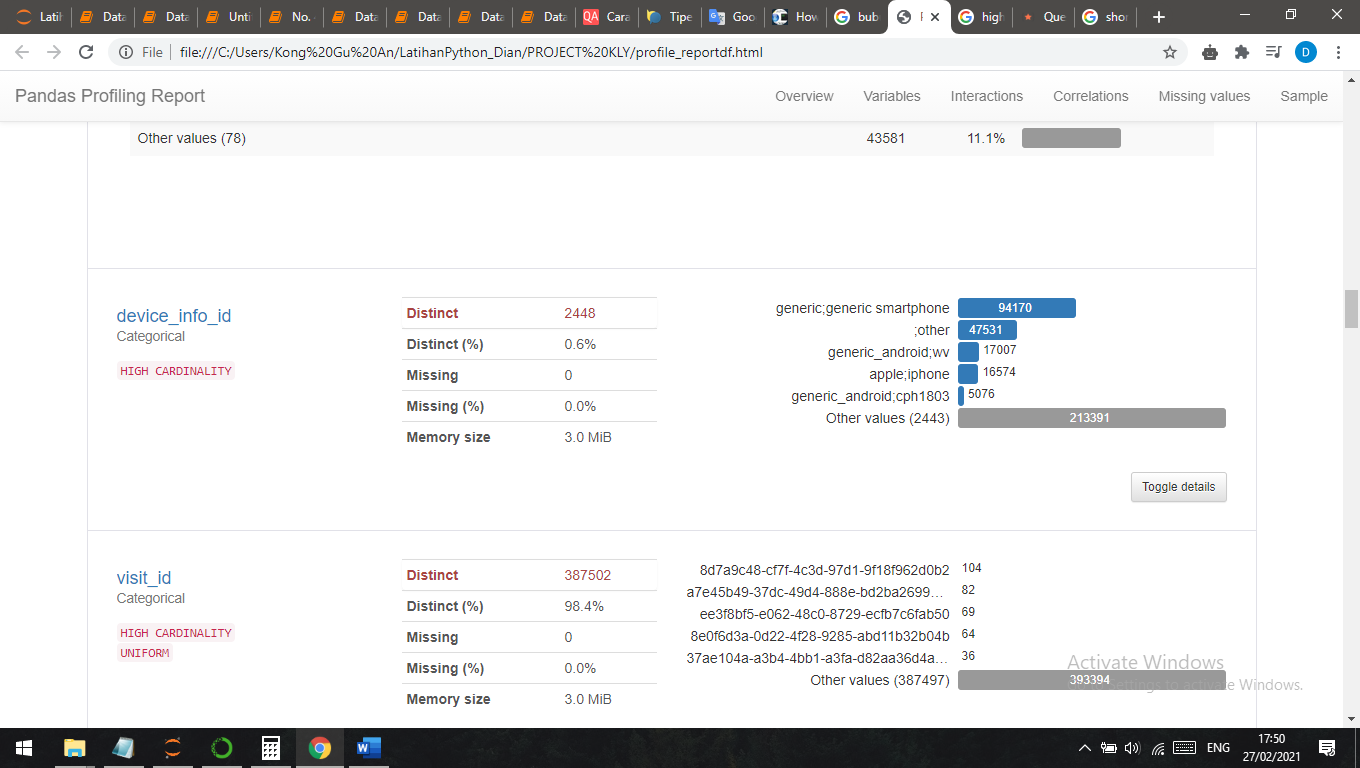
From the picture above, pandas profiling define domain\_id column as high cardinality and have high correlation with id. Domain\_id shows how much each id visit the domain, here’s the detail count of each domain\_id:



Gambar 1.9

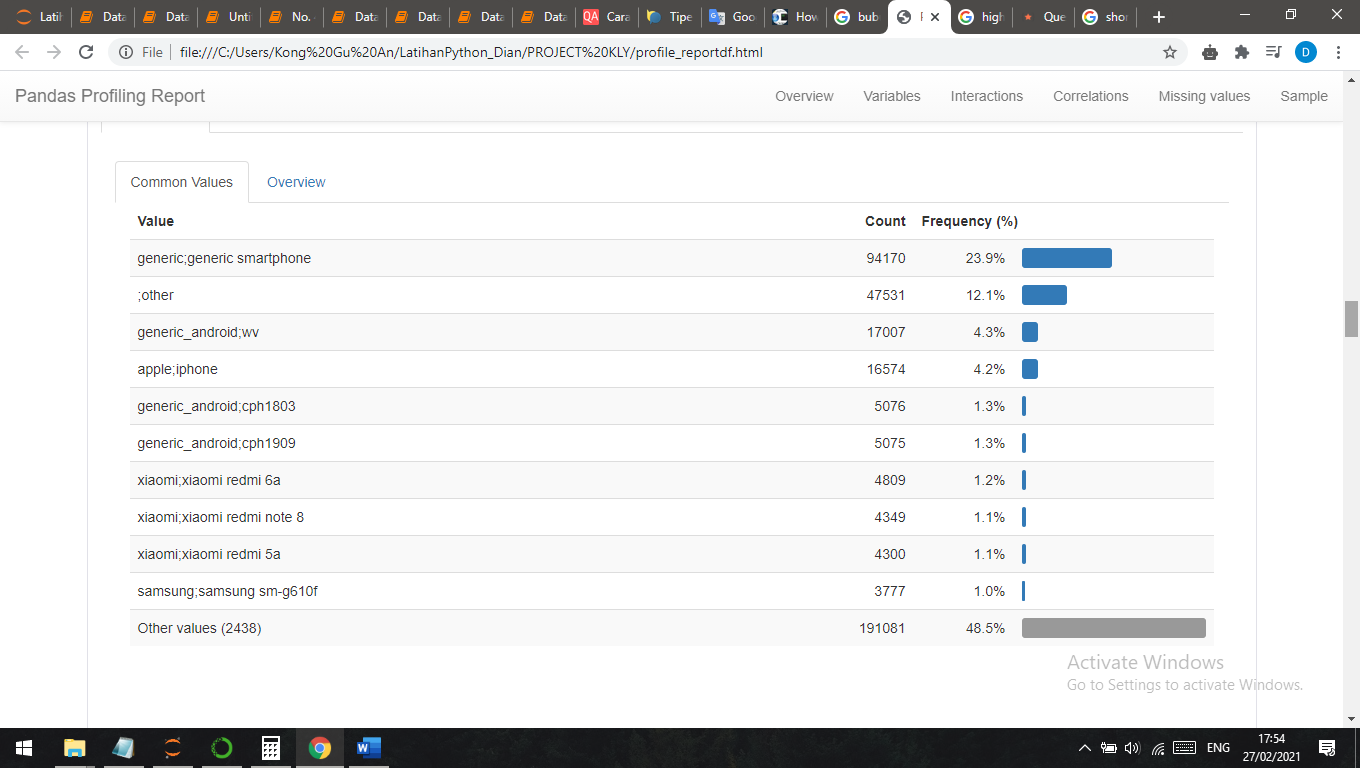
The top browser type is **m.kapanlagi.com** with the frequency 20.5%

1. **device\_info\_id = device detail information**



Gambar 1.10

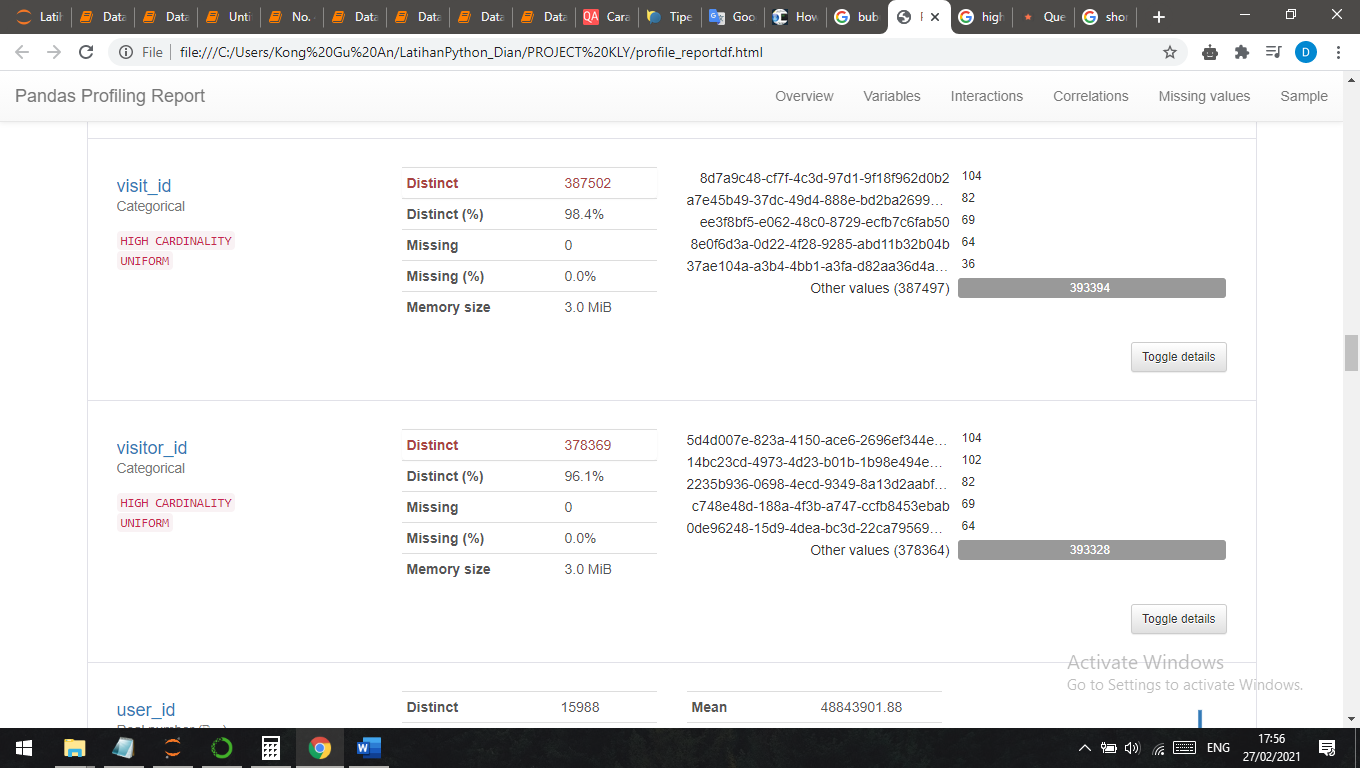
From the picture above, pandas profiling define device\_info\_id column as high cardinality. Device\_info\_id device detail information. Having distinct 0.6%, here’s the detail count of each device\_info\_id:



Gambar 1.11

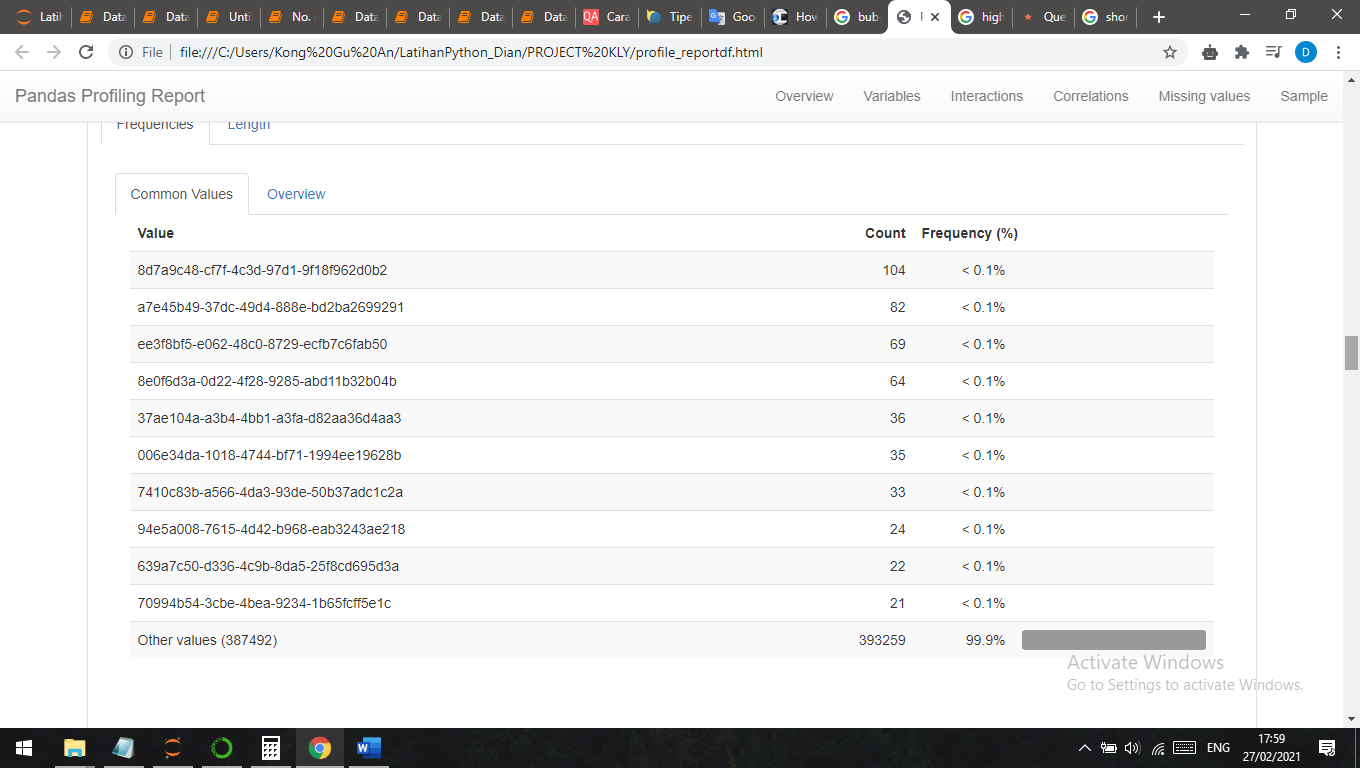
The top browser type is **generic;generic;smartphone** with the frequency 23.9%

1. **visit\_id = users session id**



Gambar 1.12

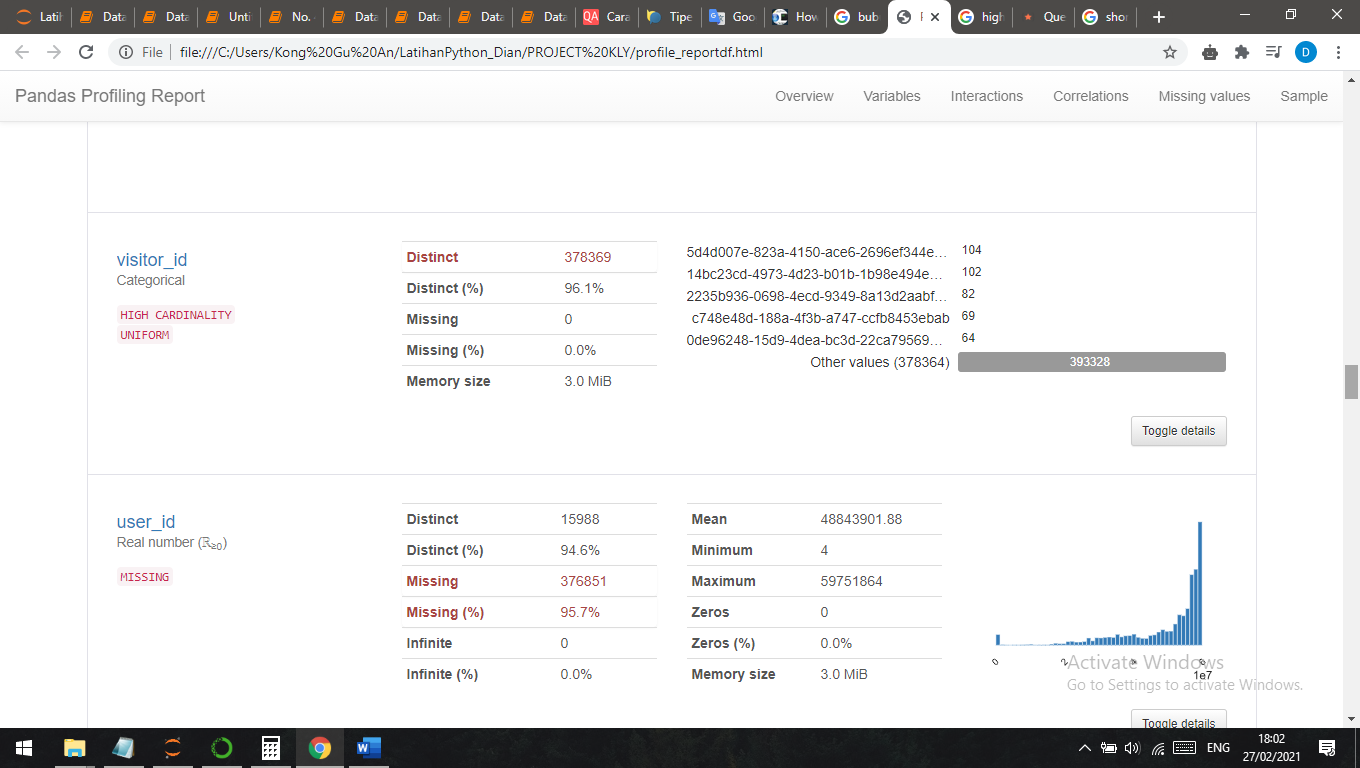
From the picture above, pandas profiling define visit\_id column as high cardinality and uniform distributed. **Almost have UNIQUE** values with the distinct is 98.4%, here’s the detail count of each visit\_id:



Gambar 1.13

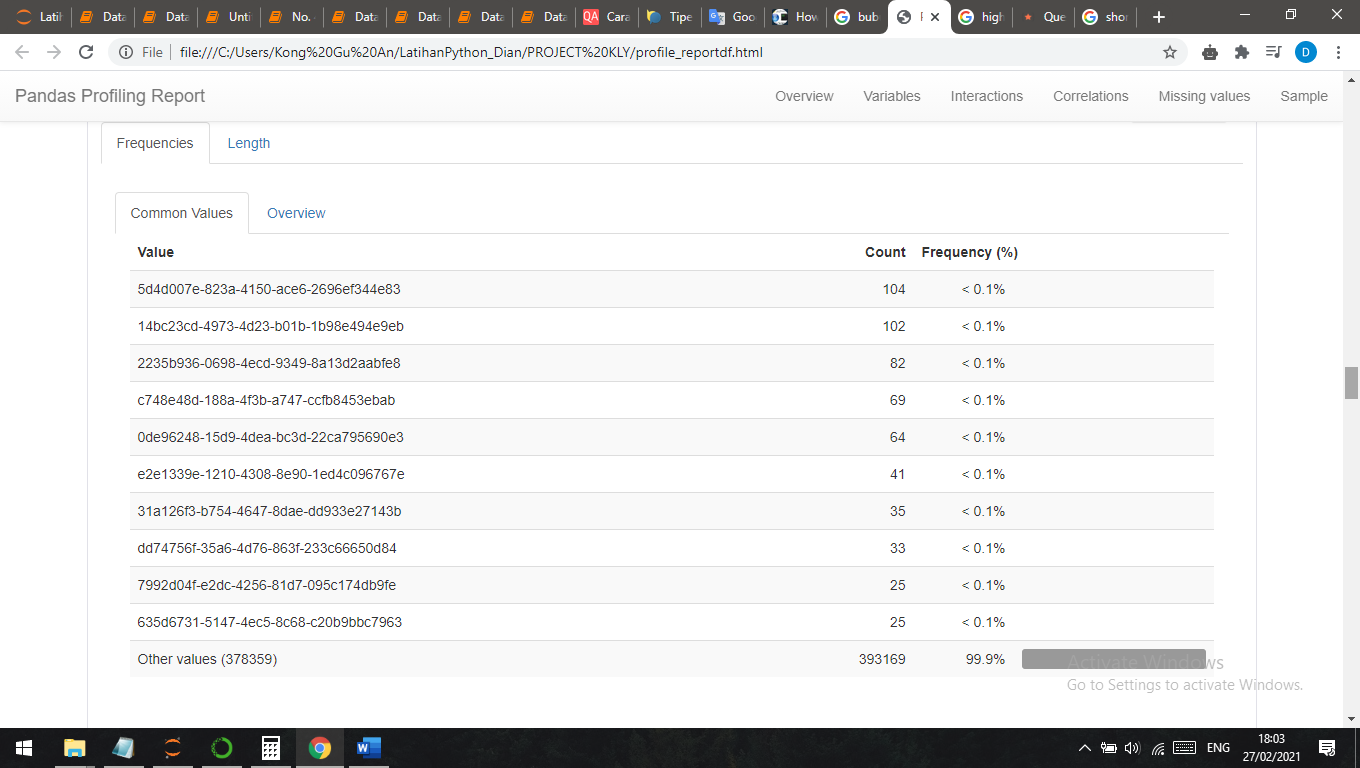
All these visit\_id have percentage below 0.1%. That’s why this variable **almost UNIQUE.**

1. **visitor\_id = unique user id**



Gambar 1.14

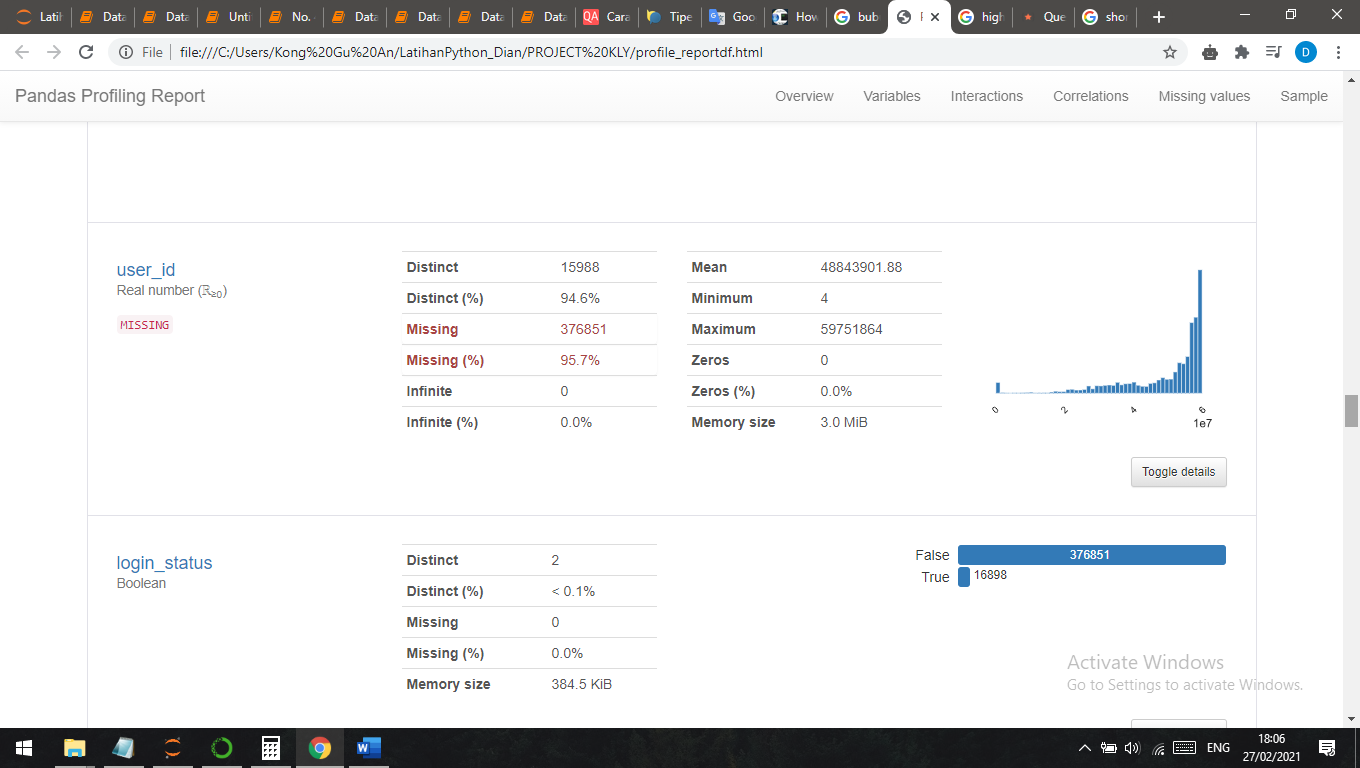
From the picture above, pandas profiling define visitor\_id column as high cardinality and uniform distributed. **Almost have UNIQUE** values with the distinct is 96.4%, here’s the detail count of each visitor\_id:



Gambar 1.15

Thevisitor\_id variable goes the same with visit\_id. All these visitor\_id have percentage below 0.1%. That’s why this variable **almost UNIQUE.**

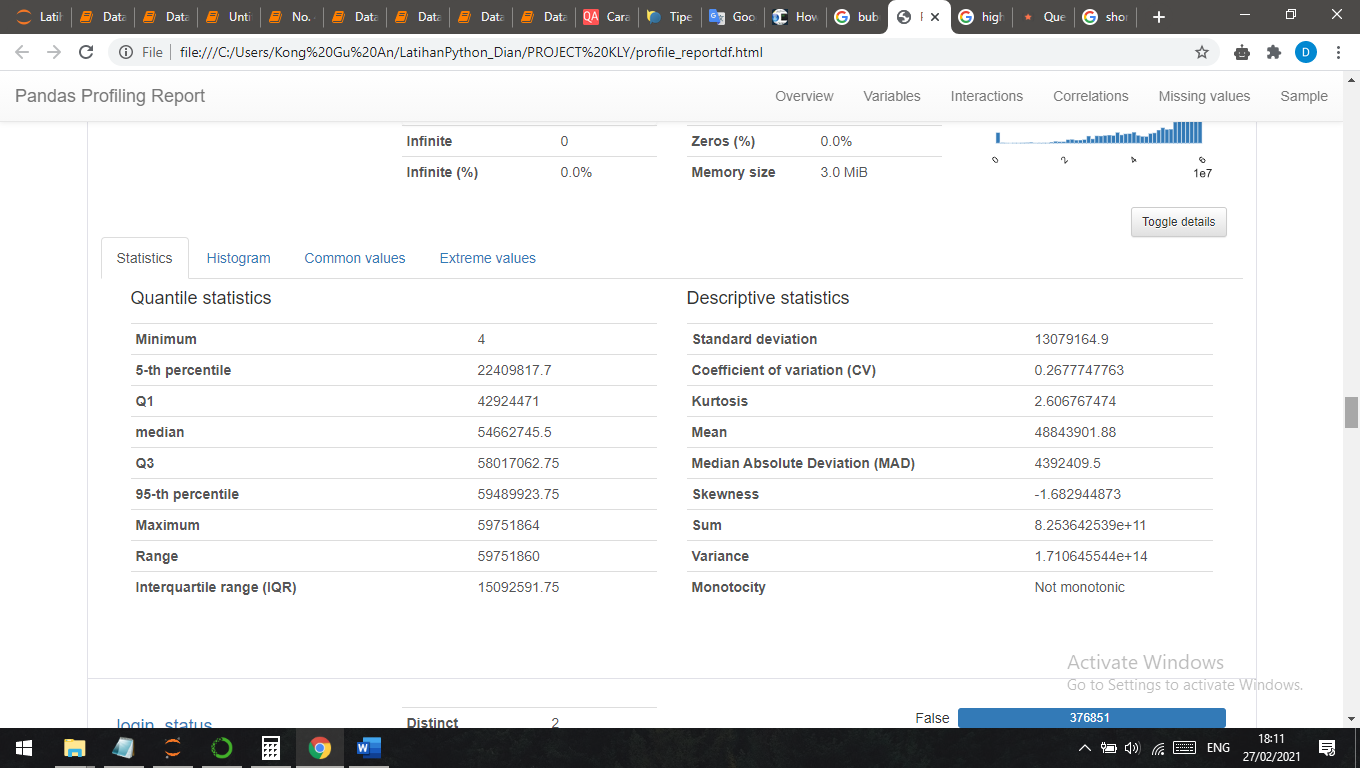
1. **user\_id = user login id**



Gambar 1.16

From the picture above, pandas profiling define user\_id column as missing. There are many missing values, it means, many users do not log in to the website page, This might be indicating that users are just scrolling or visit the website without sign in. Having distinct value in 96.7%, here’s the detail of each user\_id:

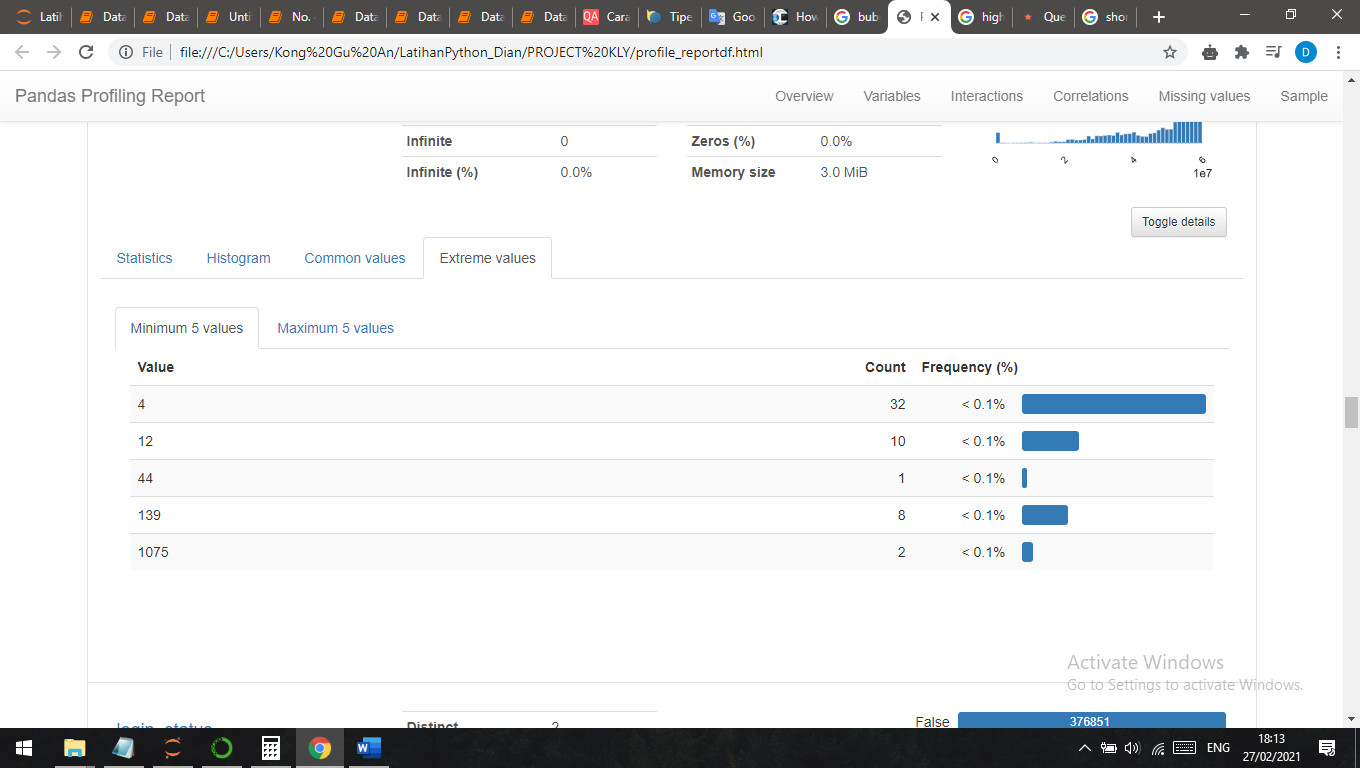
* Statistics



Gambar 1.17

The picture shows statistics calculation on this user\_id variable.

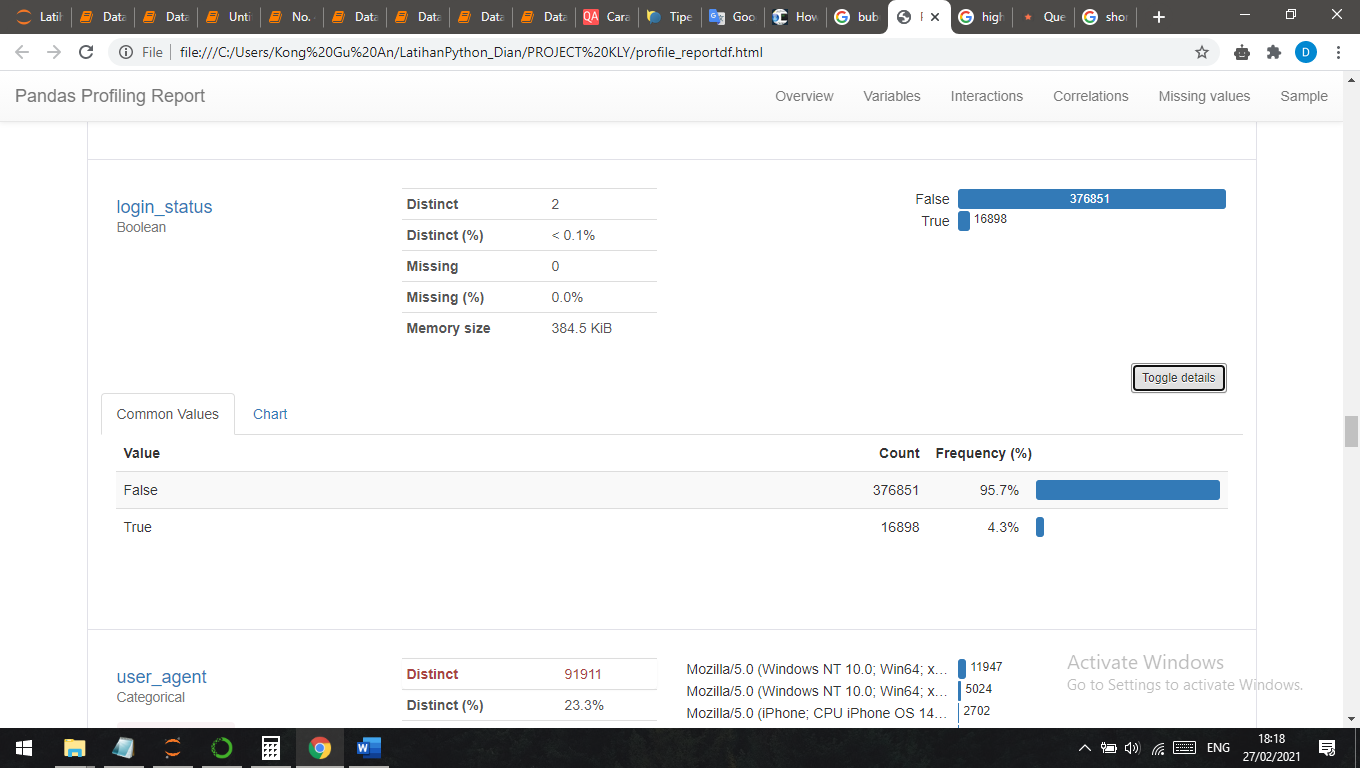
* Extreme values

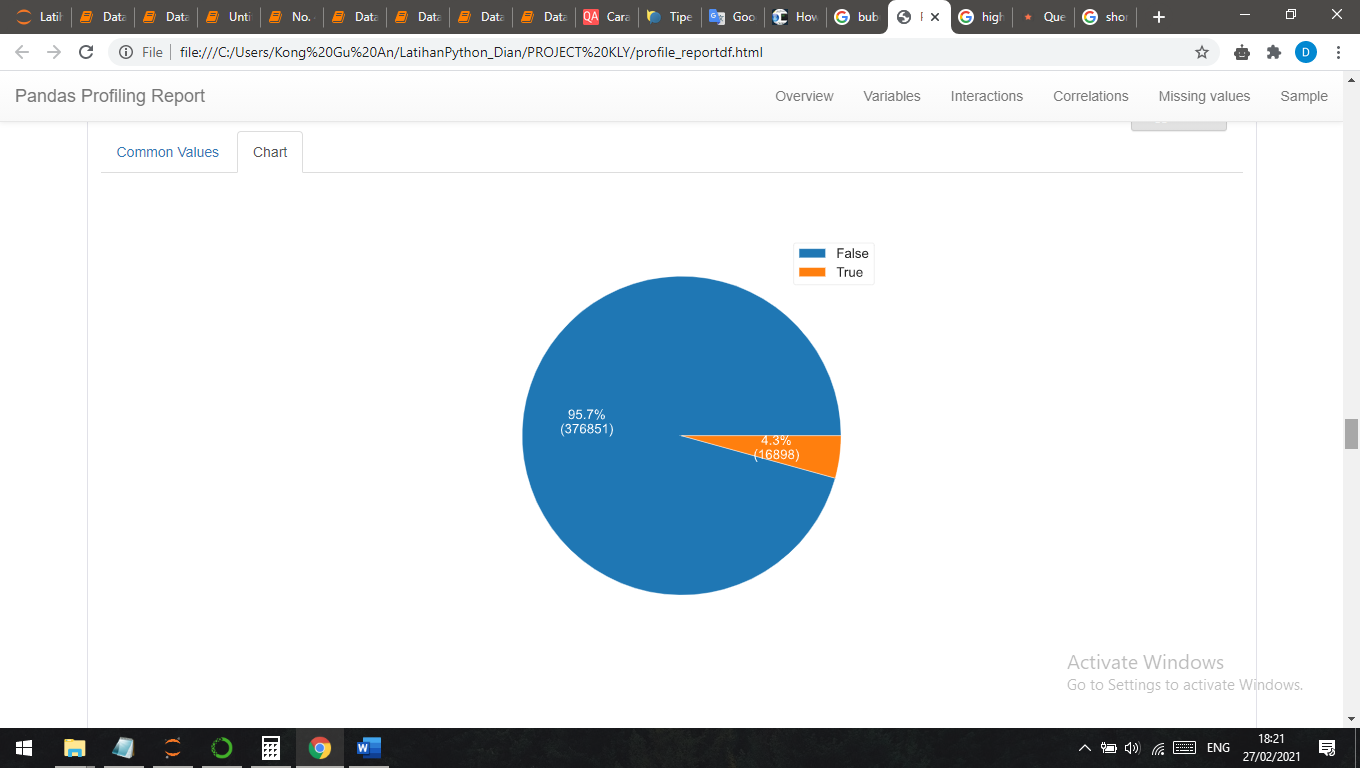


Gambar 1.18

Extreme values shows how much user\_id visit the website. The most often shows that user\_id = 4 have visited domain in 32 rimes.

1. **login\_status = boolean status of user login**

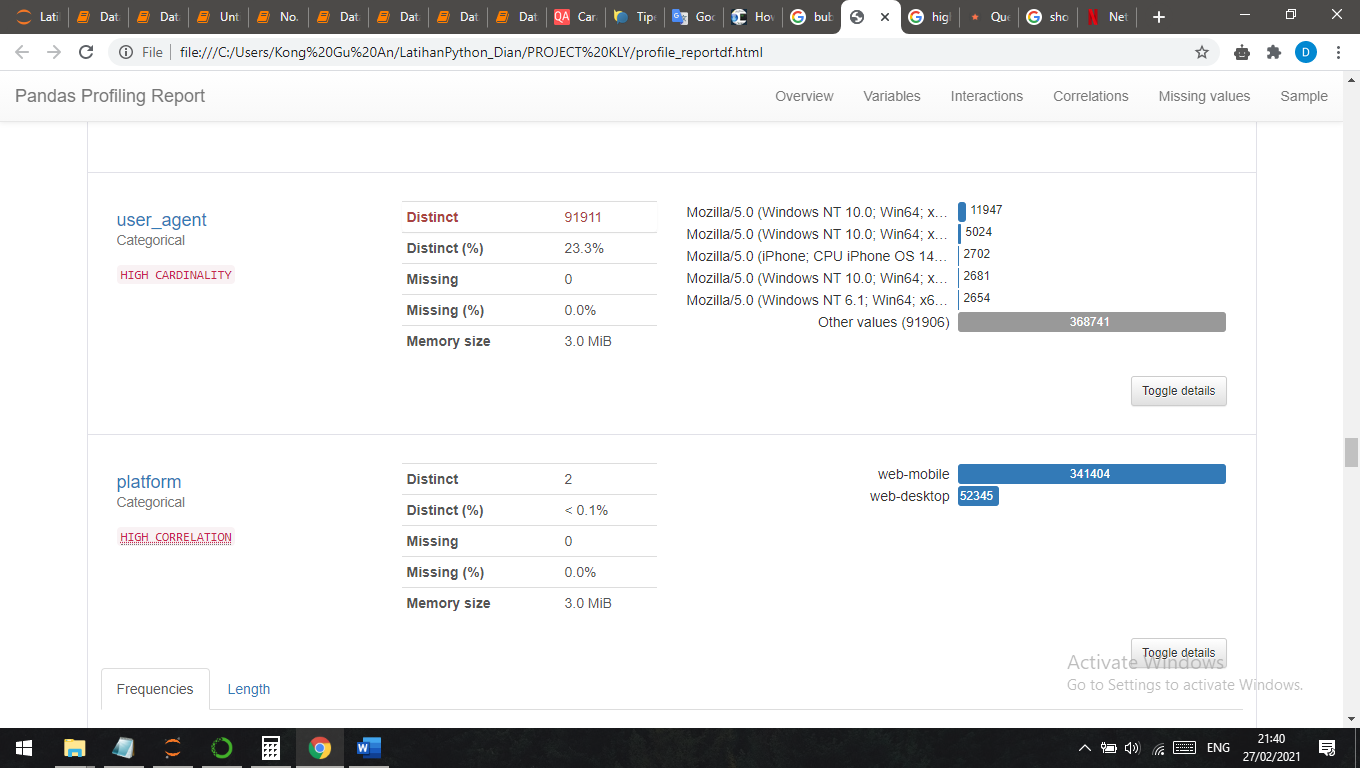




Gambar 1.19

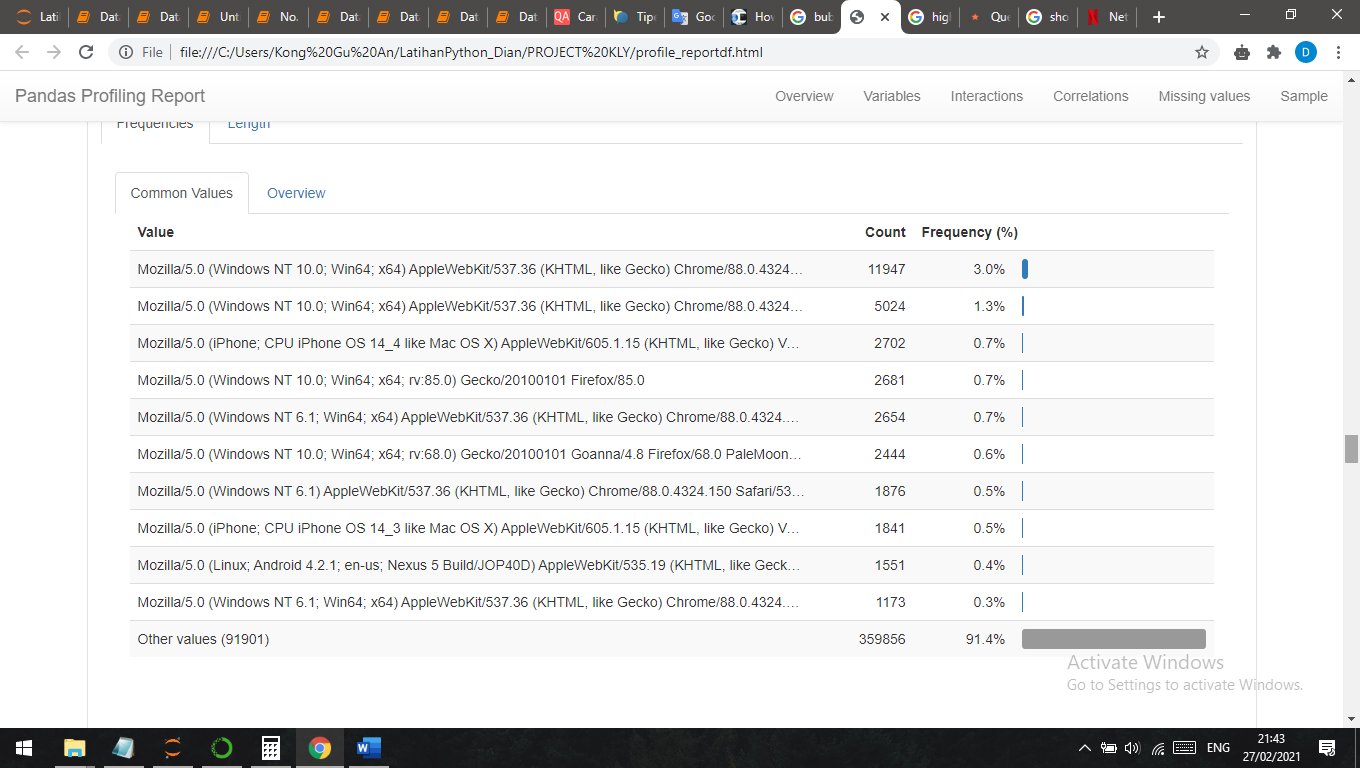
From the picture above we can conclude that many users do not log in when they visit the website. The percentage for the log in users is 4.3% and percentage for the “not log-in” users is 95.7%.

1. **user\_agent = browsers user agent details**



Gambar 1.20

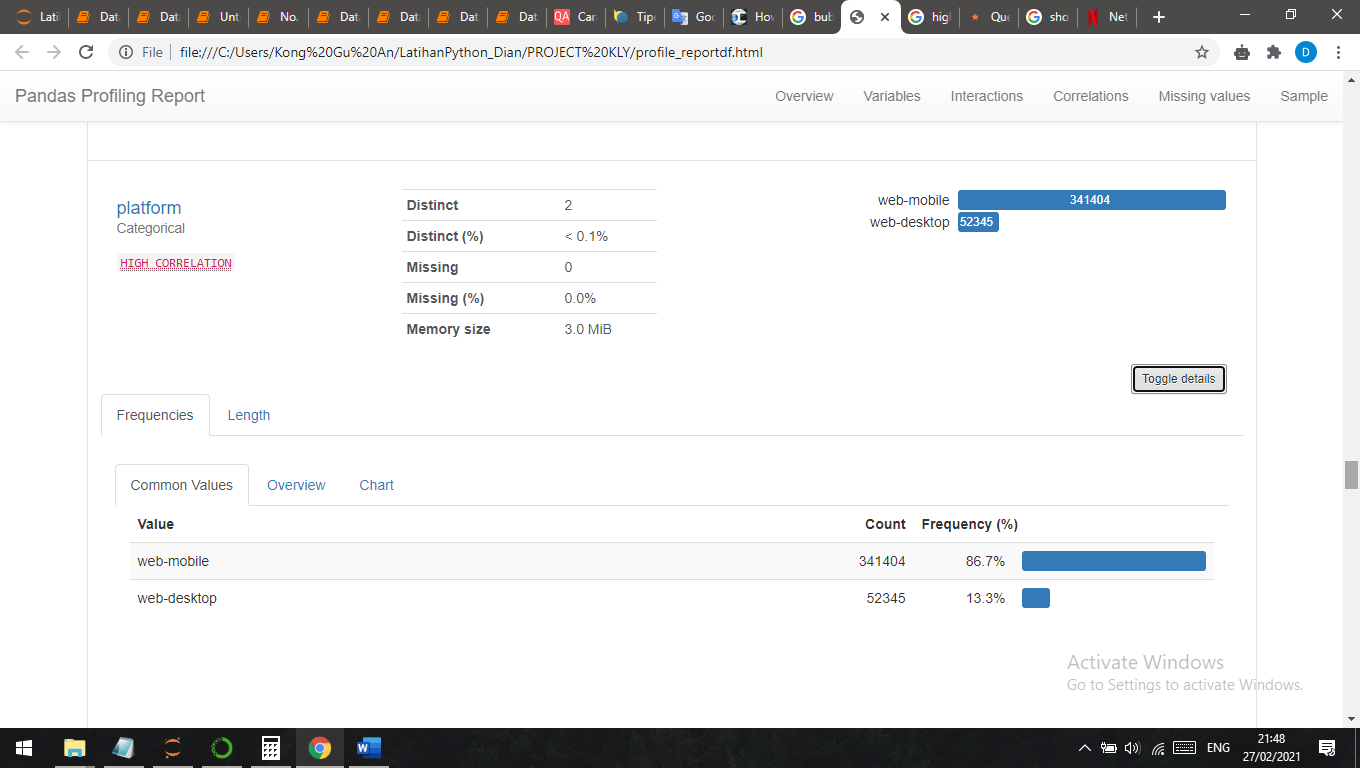
From the picture above, pandas profiling define user\_agent column as high cardinality. The distinct is 23.3%, here’s the details count for each user\_agent:

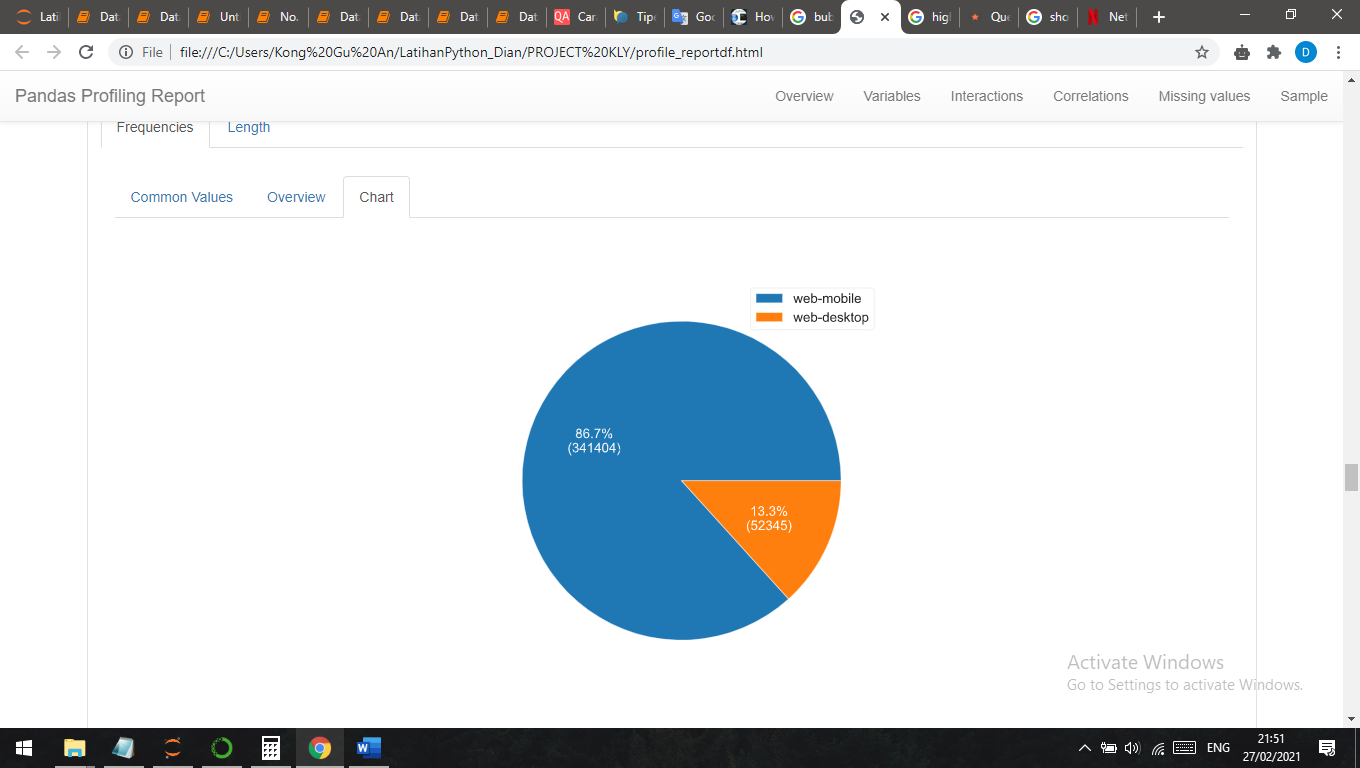


Gambar 1.21

The top browsers user agenty is **Mozilla/5.0** with frequency 3%.

1. **platform = device platform, ie desktop or mobile**

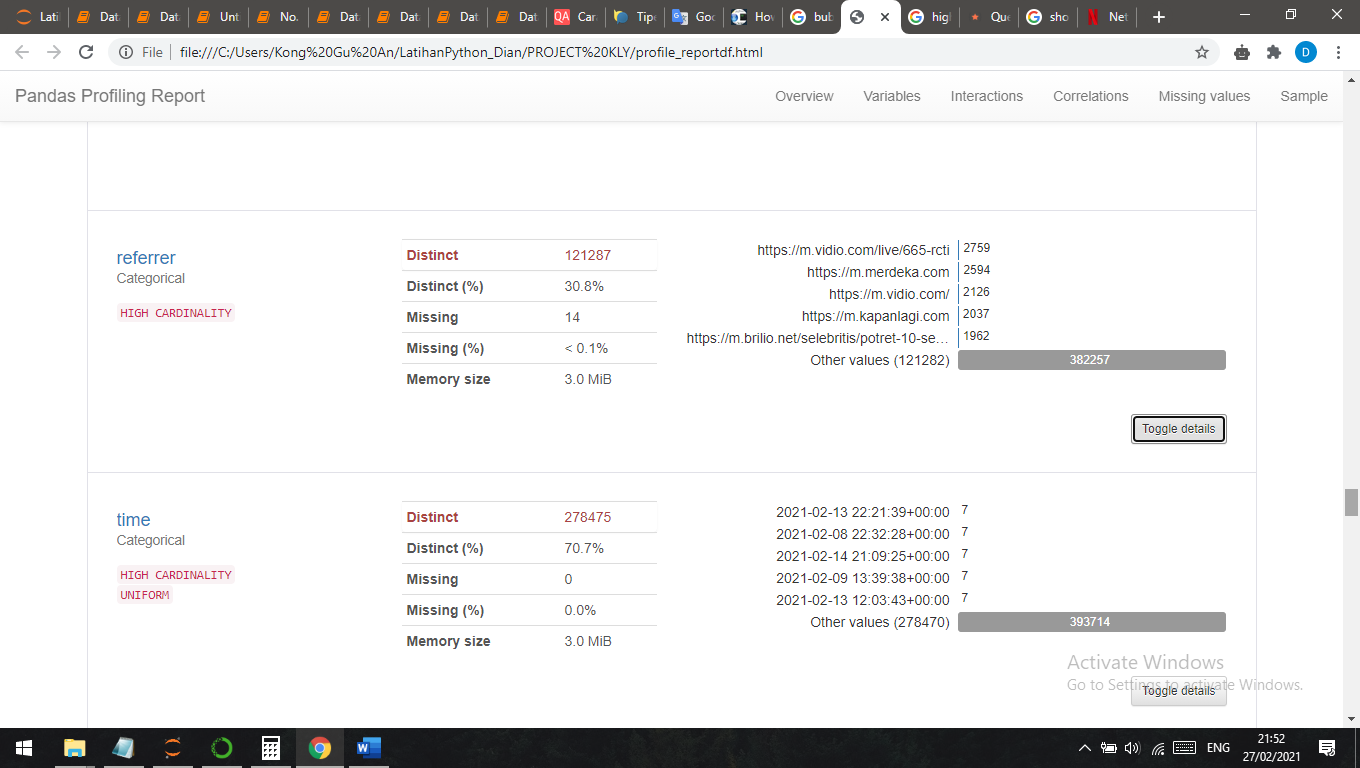




Gambar 1.22

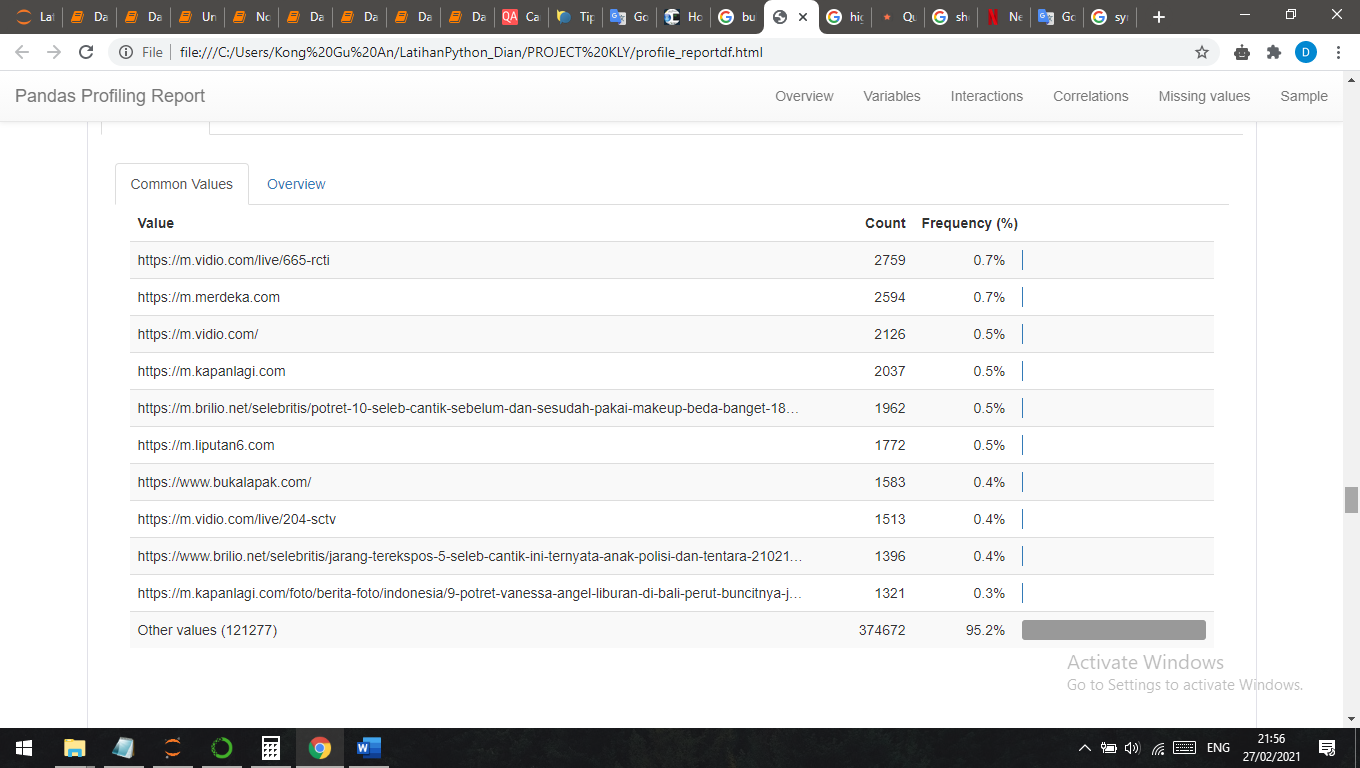
From the picture above we can see the flatform column have high correlation. Users have visited domains via web-mobile with 86.7% total and via web-desktop with 13.3%.

1. **referrer = attribution of the visit, source of visitor coming from**



Gambar 1.23

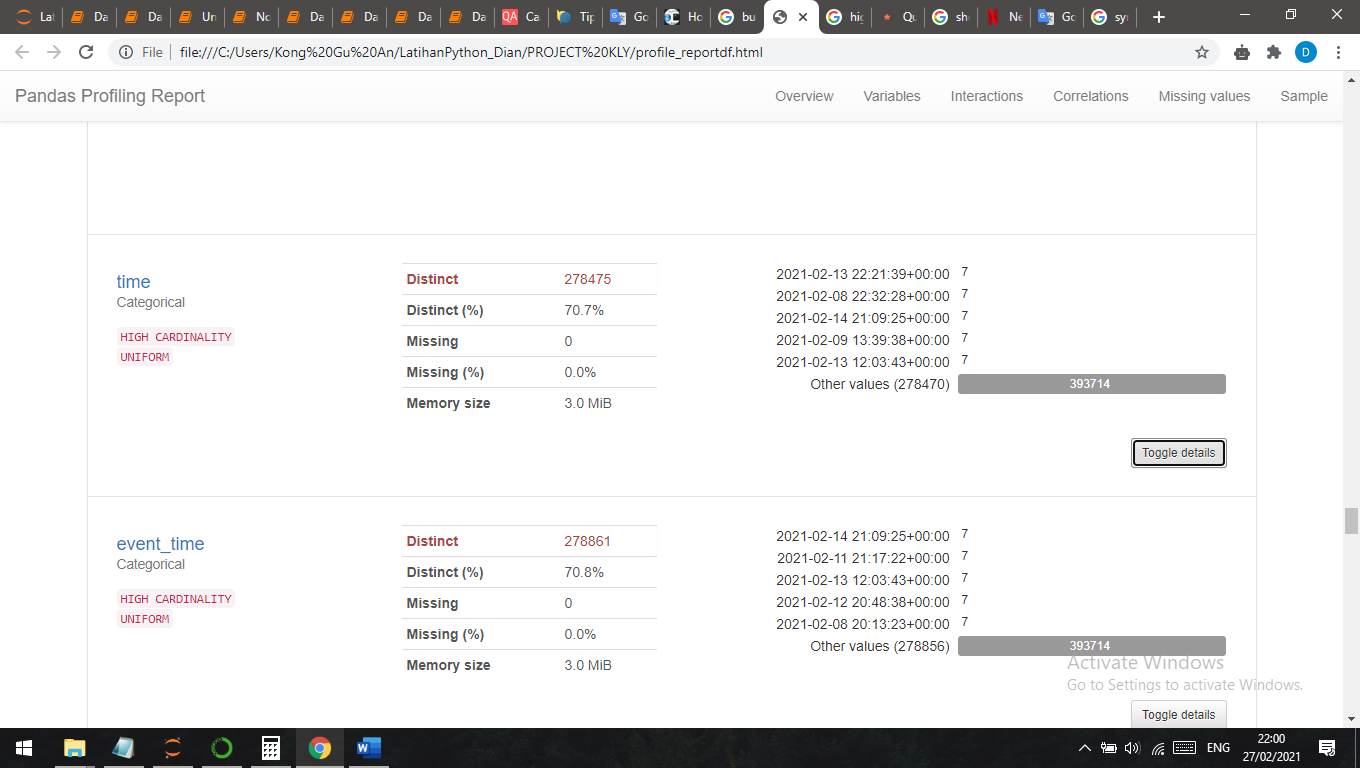
From the picture above, referrer has high cardinality, the distinct is 30.8%. We also can see that the source of visitor coming is very diverse. Here’s the detail count of each source:



Gambar 1.23

The top source is from **vidio.com/665-rcti** with frequency 0.7%. There is just a little difference between the first place and the second, it shows that the source is very evenly.

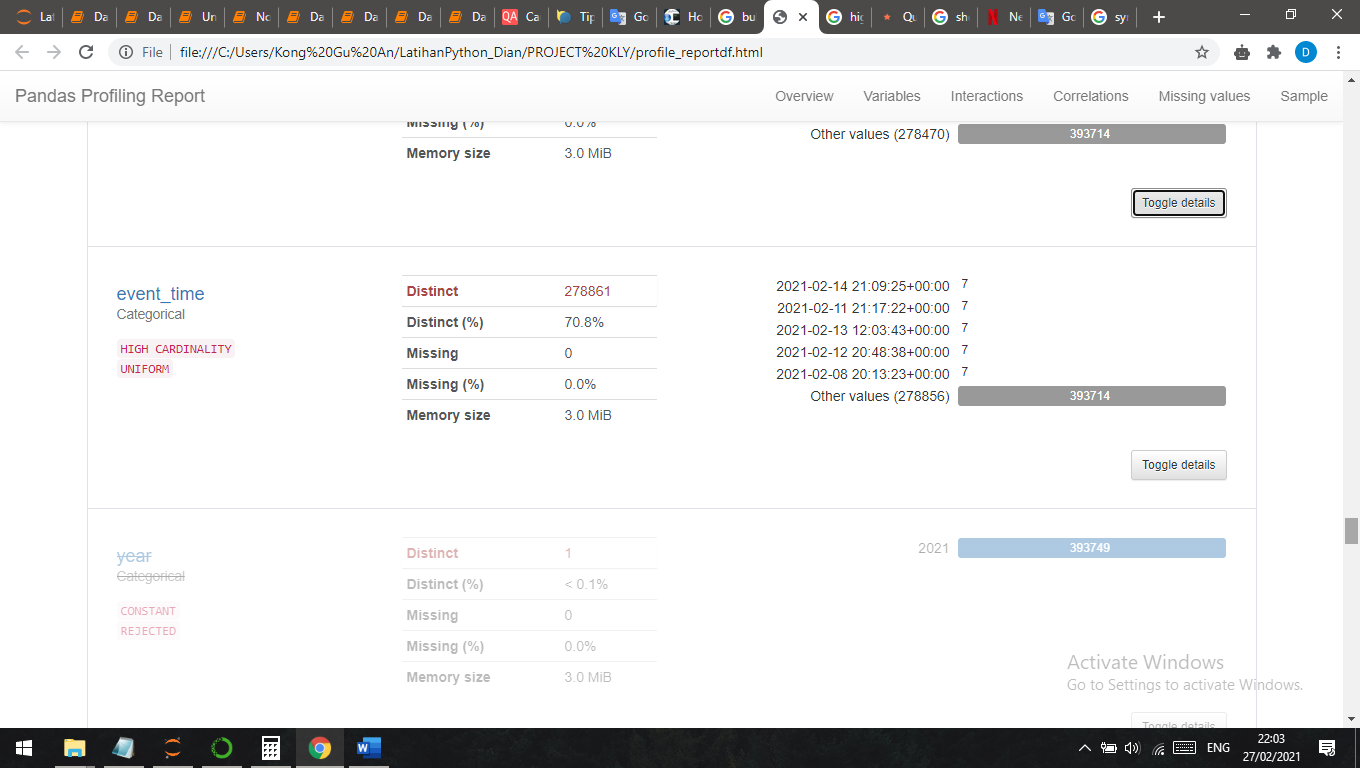
1. **time = users visit time**



Gambar 1.24

The picture above shows that time variable might have high chance to become unique, just a few users who have exactly visit the same time. The distinct is 70.7%, and this time column has high cardinality and uniform distributed.

1. **event\_time = logging time**

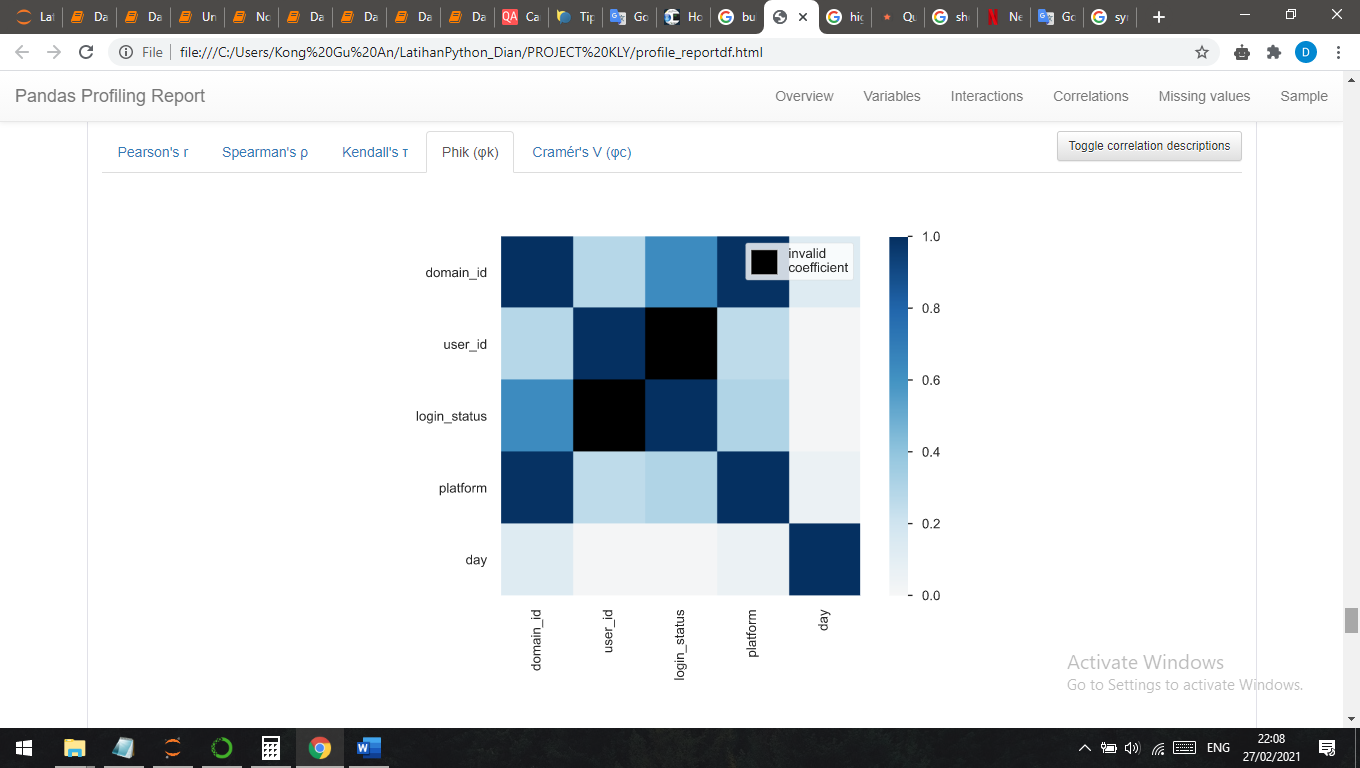


Gambar 1.25

Same for the time variable, the even\_time shows very diverse in time. just a few users who have exactly visit the same time. The distinct is 70.8%. Also this event\_time column has high cardinality and uniform distributed.

CORRELATION

Pandas profiling have feature to calculate the correlation using various method, but here I will present correlation between variables = domain\_id, user\_id, login\_status, platform, and day using Phik method.



Gambar 1.26