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:

	year	month	day
count	393749.0	393749.0	393749.000000
mean	2021.0	2.0	11.000008
std	0.0	0.0	1.999999
min	2021.0	2.0	8.000000
25%	2021.0	2.0	9.000000
50%	2021.0	2.0	11.000000
75%	2021.0	2.0	13.000000
max	2021.0	2.0	14.000000

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

```
Data columns (total 17 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   id              393749 non-null  object
 1   browser_id      393749 non-null  object
 2   os_id           393749 non-null  object
 3   domain_id       393734 non-null  object
 4   device_info_id  393749 non-null  object
 5   visit_id        393749 non-null  object
 6   visitor_id      393749 non-null  object
 7   user_id         16898 non-null   float64
 8   login_status    393749 non-null  bool
 9   user_agent      393749 non-null  object
 10  platform        393749 non-null  object
 11  referrer        393735 non-null  object
 12  time            393749 non-null  datetime64[ns, UTC]
 13  event_time      393749 non-null  datetime64[ns, UTC]
 14  year            393749 non-null  int64
 15  month           393749 non-null  int64
 16  day             393749 non-null  int64
```

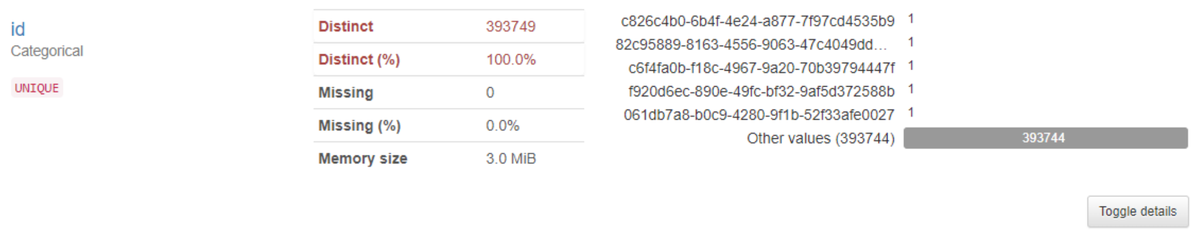
Gambar 1.2

No.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

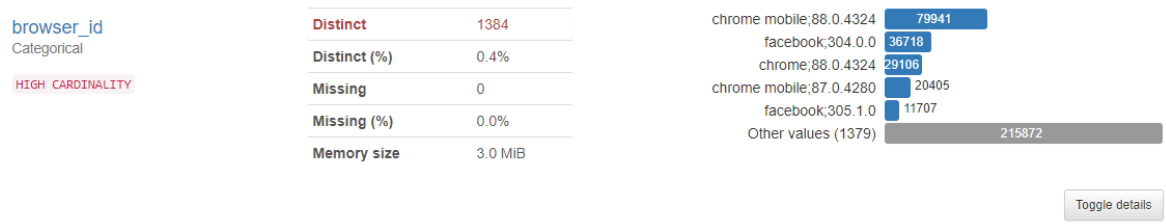
1. **id = record ID**



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.

2. **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:

Common Values | Overview

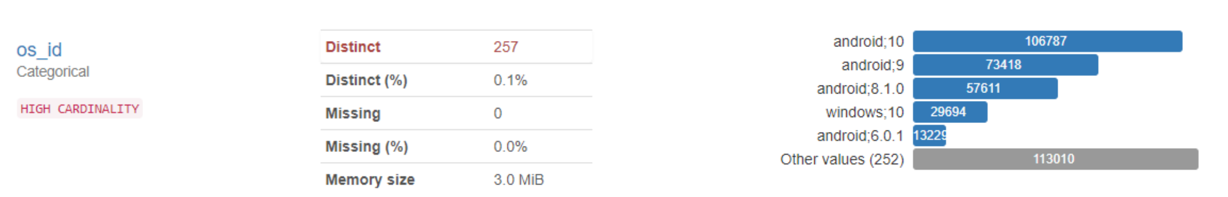
Value	Count	Frequency (%)
chrome mobile;88.0.4324	79941	20.3%
facebook;304.0.0	36718	9.3%
chrome;88.0.4324	29106	7.4%
chrome mobile;87.0.4280	20405	5.2%
facebook;305.1.0	11707	3.0%
chrome mobile;86.0.4240	10337	2.6%
chrome mobile;80.0.3987	8882	2.3%
chrome mobile;83.0.4103	7655	1.9%
chrome mobile;81.0.4044	6556	1.7%
chrome mobile;85.0.4183	6259	1.6%
Other values (1374)	176183	44.7%

Gambar 1.5

No.2

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

3. 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:

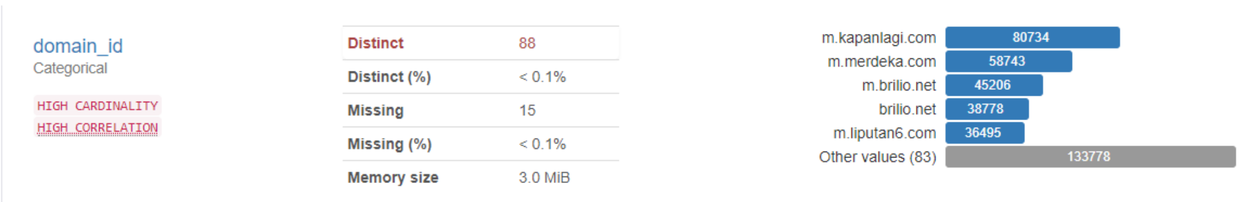
Common Values | Overview

Value	Count	Frequency (%)
android;10	106787	27.1%
android;9	73418	18.6%
android;8.1.0	57611	14.6%
windows;10	29694	7.5%
android;6.0.1	13229	3.4%
android;7.1.2	11757	3.0%
windows;7	11485	2.9%
android;7.1.1	11359	2.9%
android;5.1.1	10854	2.8%
android;7.0	8918	2.3%
Other values (247)	58637	14.9%

Gambar 1.7

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

4. 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:

No.2

Common Values | Overview

Value	Count	Frequency (%)
m.kapanlagi.com	80734	20.5%
m.merdeka.com	58743	14.9%
m.brilio.net	45206	11.5%
brilio.net	38778	9.8%
m.liputan6.com	36495	9.3%
m.vidio.com	30700	7.8%
bukalapak.com	19245	4.9%
m.bukalapak.com	16731	4.2%
m.bola.com	12075	3.1%
vidio.com	11446	2.9%
Other values (78)	43581	11.1%

Gambar 1.9

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

5. device_info_id = device detail information

device_info_id
Categorical

HIGH CARDINALITY

Distinct	2448
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

generic;generic smartphone	94170
;other	47531
generic_android;wv	17007
apple;iphone	16574
generic_android;cph1803	5076
Other values (2443)	213391

Toggle details

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:

Common Values | Overview

Value	Count	Frequency (%)
generic;generic smartphone	94170	23.9%
;other	47531	12.1%
generic_android;wv	17007	4.3%
apple;iphone	16574	4.2%
generic_android;cph1803	5076	1.3%
generic_android;cph1909	5075	1.3%
xiaomi;xiaomi redmi 6a	4809	1.2%
xiaomi;xiaomi redmi note 8	4349	1.1%
xiaomi;xiaomi redmi 5a	4300	1.1%
samsung;samsung sm-g610f	3777	1.0%
Other values (2438)	191081	48.5%

Gambar 1.11

No.2

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

6. **visit_id = users session id**

visit_id
Categorical

HIGH CARDINALITY
UNIFORM

Distinct	387502
Distinct (%)	98.4%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

8d7a9c48-cf7f-4c3d-97d1-9f18f962d0b2	104
a7e45b49-37dc-49d4-888e-bd2ba2699...	82
ee3f8bf5-e062-48c0-8729-ecfb7c6fab50	69
8e0f6d3a-0d22-4f28-9285-abd11b32b04b	64
37ae104a-a3b4-4bb1-a3fa-d82aa36d4a...	36
Other values (387497)	393394

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:

Common Values | Overview

Value	Count	Frequency (%)
8d7a9c48-cf7f-4c3d-97d1-9f18f962d0b2	104	< 0.1%
a7e45b49-37dc-49d4-888e-bd2ba2699291	82	< 0.1%
ee3f8bf5-e062-48c0-8729-ecfb7c6fab50	69	< 0.1%
8e0f6d3a-0d22-4f28-9285-abd11b32b04b	64	< 0.1%
37ae104a-a3b4-4bb1-a3fa-d82aa36d4aa3	36	< 0.1%
006e34da-1018-4744-bf71-1994ee19628b	35	< 0.1%
7410c83b-a566-4da3-93de-50b37adc1c2a	33	< 0.1%
94e5a008-7615-4d42-b968-eab3243ae218	24	< 0.1%
639a7c50-d336-4c9b-8da5-25f8cd695d3a	22	< 0.1%
70994b54-3cbe-4bea-9234-1b65fcff5e1c	21	< 0.1%
Other values (387492)	393259	99.9%

Gambar 1.13

All these visit_id have percentage below 0.1%. That's why this variable **almost UNIQUE.**

7. **visitor_id = unique user id**

visitor_id
Categorical

HIGH CARDINALITY
UNIFORM

Distinct	378369
Distinct (%)	96.1%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

5d4d007e-823a-4150-ace6-2696ef344e...	104
14bc23cd-4973-4d23-b01b-1b98e494e...	102
2235b936-0698-4ecd-9349-8a13d2aabf...	82
c748e48d-188a-4f3b-a747-ccfb8453ebab	69
0de96248-15d9-4dea-bc3d-22ca79569...	64
Other values (378364)	393328

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:

No.2

Common Values | Overview

Value	Count	Frequency (%)
5d4d007e-823a-4150-ace6-2696ef344e83	104	< 0.1%
14bc23cd-4973-4d23-b01b-1b98e494e9eb	102	< 0.1%
2235b936-0698-4ecd-9349-8a13d2aabfe8	82	< 0.1%
c748e48d-188a-4f3b-a747-ccfb8453ebab	69	< 0.1%
0de96248-15d9-4dea-bc3d-22ca795690e3	64	< 0.1%
e2e1339e-1210-4308-8e90-1ed4c096767e	41	< 0.1%
31a126f3-b754-4647-8dae-dd933e27143b	35	< 0.1%
dd74756f-35a6-4d76-863f-233c66650d84	33	< 0.1%
7992d04f-e2dc-4256-81d7-095c174db9fe	25	< 0.1%
635d6731-5147-4ec5-8c68-c20b9bbc7963	25	< 0.1%
Other values (378359)	393169	99.9%

Gambar 1.15

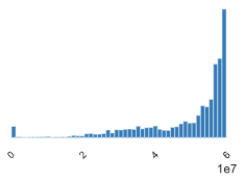
Thevisitor_id variable goes the same with visit_id. All these visitor_id have percentage below 0.1%. That’s why this variable **almost UNIQUE.**

8. user_id = user login id

user_id
Real number ($\mathbb{R}_{\geq 0}$)
MISSING

Distinct	15988
Distinct (%)	94.6%
Missing	376851
Missing (%)	95.7%
Infinite	0
Infinite (%)	0.0%

Mean	48843901.88
Minimum	4
Maximum	59751864
Zeros	0
Zeros (%)	0.0%
Memory size	3.0 MiB



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

Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	4
5-th percentile	22409817.7
Q1	42924471
median	54662745.5
Q3	58017062.75
95-th percentile	59489923.75
Maximum	59751864
Range	59751860
Interquartile range (IQR)	15092591.75

Descriptive statistics

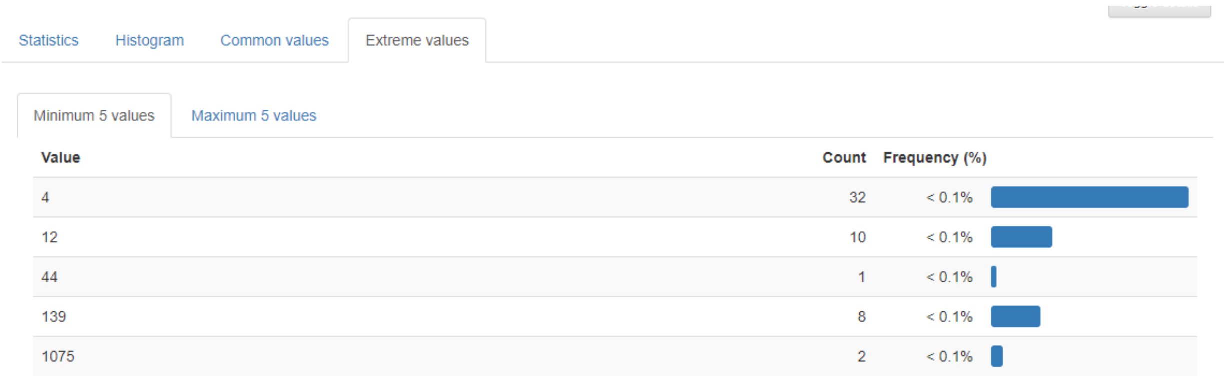
Standard deviation	13079164.9
Coefficient of variation (CV)	0.2677747763
Kurtosis	2.606767474
Mean	48843901.88
Median Absolute Deviation (MAD)	4392409.5
Skewness	-1.682944873
Sum	8.253642539e+11
Variance	1.710645544e+14
Monotocity	Not monotonic

Gambar 1.17

The picture shows statistics calculation on this user_id variable.

No.2

- Extreme values

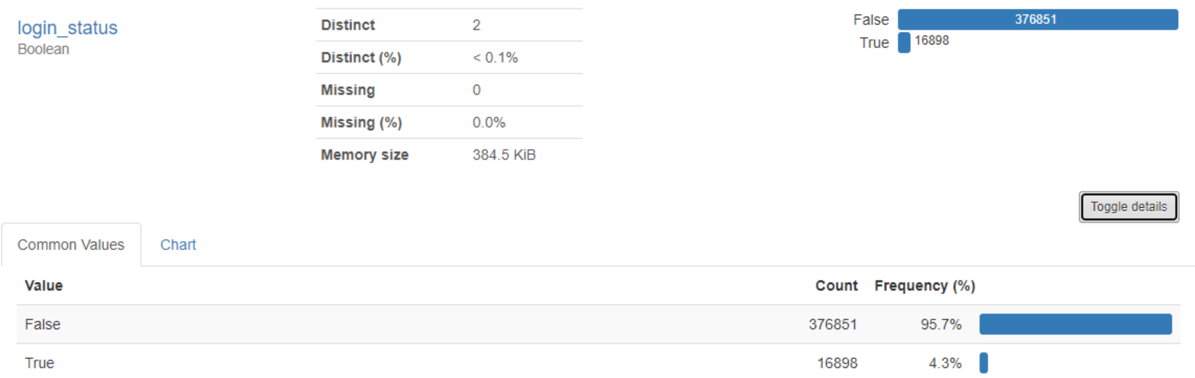


Value	Count	Frequency (%)
4	32	< 0.1%
12	10	< 0.1%
44	1	< 0.1%
139	8	< 0.1%
1075	2	< 0.1%

Gambar 1.18

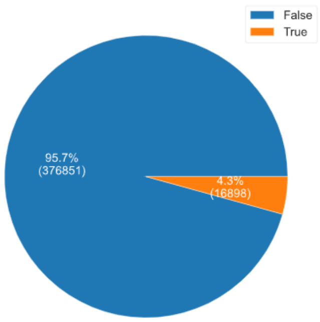
Extreme values shows how much user_id visit the website. The most often shows that user_id = 4 have visited domain in 32 rimes.

9. **login_status = boolean status of user login**



Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	384.5 KiB

Value	Count	Frequency (%)
False	376851	95.7%
True	16898	4.3%

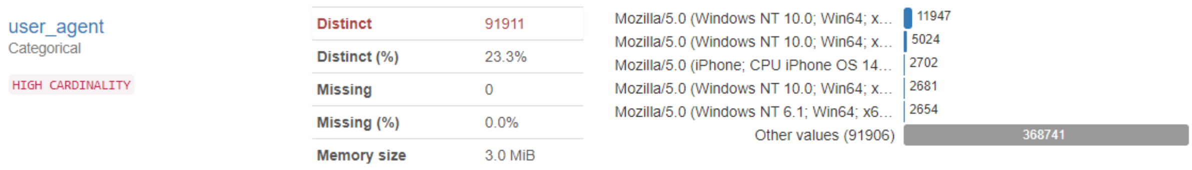


Gambar 1.19

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%.

10. **user_agent = browsers user agent details**

No.2



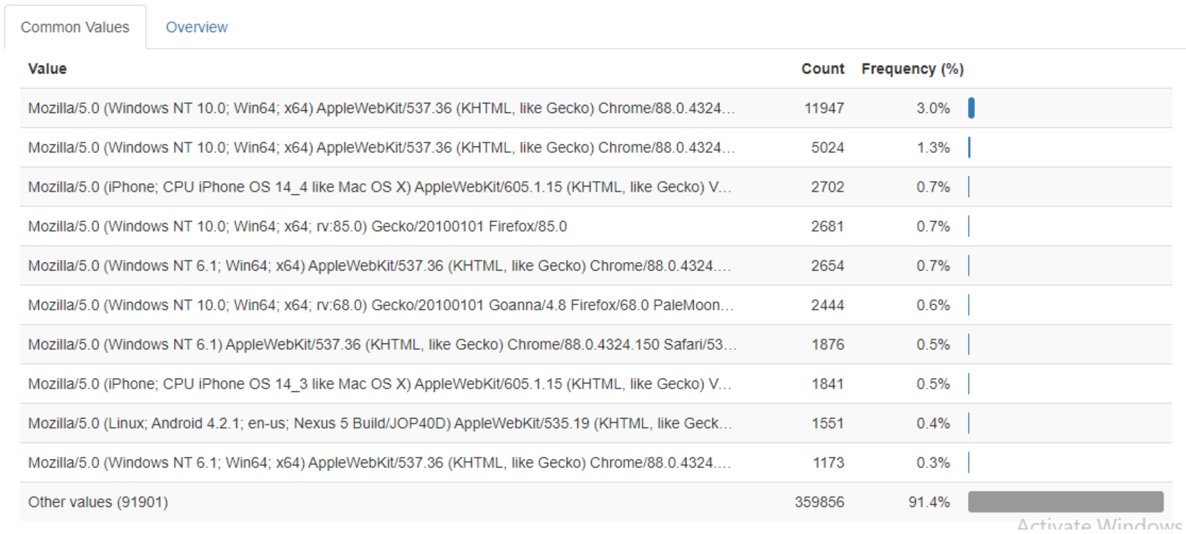
user_agent
Categorical
HIGH CARDINALITY

Distinct	91911
Distinct (%)	23.3%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

Mozilla/5.0 (Windows NT 10.0; Win64; x…	11947
Mozilla/5.0 (Windows NT 10.0; Win64; x…	5024
Mozilla/5.0 (iPhone; CPU iPhone OS 14…	2702
Mozilla/5.0 (Windows NT 10.0; Win64; x…	2681
Mozilla/5.0 (Windows NT 6.1; Win64; x6…	2654
Other values (91906)	368741

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:



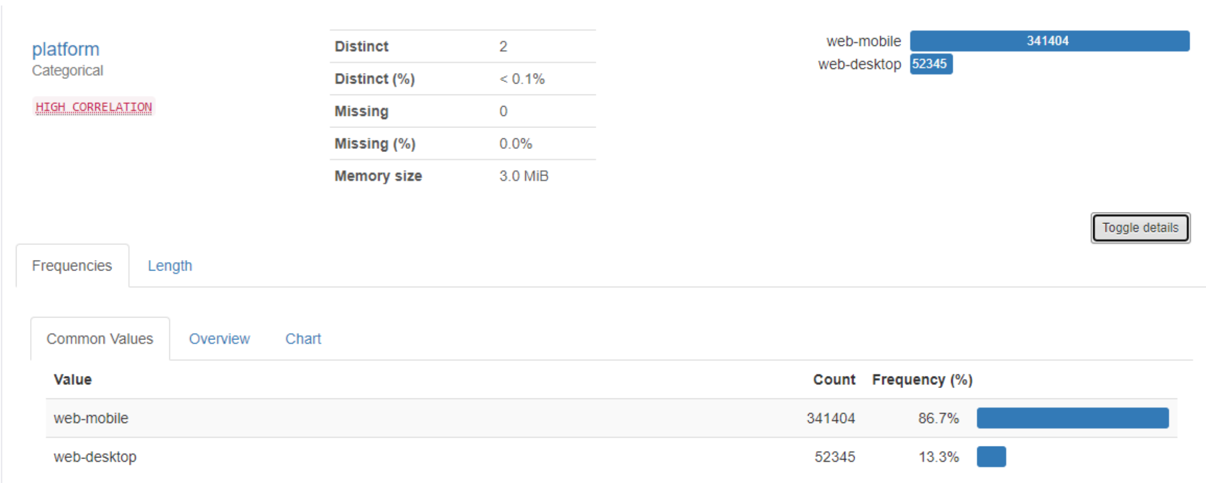
Common Values | Overview

Value	Count	Frequency (%)
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324…	11947	3.0%
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324…	5024	1.3%
Mozilla/5.0 (iPhone; CPU iPhone OS 14_4 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) V…	2702	0.7%
Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:85.0) Gecko/20100101 Firefox/85.0	2681	0.7%
Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.…	2654	0.7%
Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:68.0) Gecko/20100101 Goanna/4.8 Firefox/68.0 PaleMoon…	2444	0.6%
Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.150 Safari/53…	1876	0.5%
Mozilla/5.0 (iPhone; CPU iPhone OS 14_3 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) V…	1841	0.5%
Mozilla/5.0 (Linux; Android 4.2.1; en-us; Nexus 5 Build/JOP40D) AppleWebKit/535.19 (KHTML, like Geck…	1551	0.4%
Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.…	1173	0.3%
Other values (91901)	359856	91.4%

Gambar 1.21

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

11. platform = device platform, ie desktop or mobile



platform
Categorical
HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

web-mobile	341404
web-desktop	52345

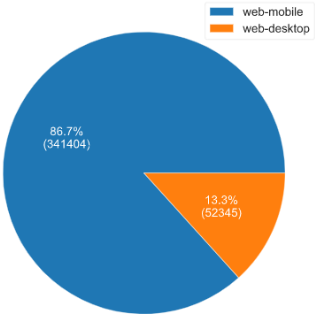
Toggle details

Frequencies | Length

Common Values | Overview | Chart

Value	Count	Frequency (%)
web-mobile	341404	86.7%
web-desktop	52345	13.3%

No.2



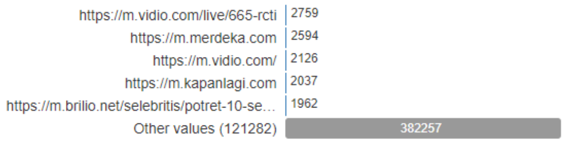
Gambar 1.22

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%.

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

referrer
Categorical
HIGH CARDINALITY

Distinct	121287
Distinct (%)	30.8%
Missing	14
Missing (%)	< 0.1%
Memory size	3.0 MiB



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:

Common Values | Overview

Value	Count	Frequency (%)
https://m.vidio.com/live/665-rcti	2759	0.7%
https://m.merdeka.com	2594	0.7%
https://m.vidio.com/	2126	0.5%
https://m.kapanlagi.com	2037	0.5%
https://m.brilio.net/selebritis/potret-10-seleb-cantik-sebelum-dan-sesudah-pakai-makeup-beda-banget-18…	1962	0.5%
https://m.liputan6.com	1772	0.5%
https://www.bukalapak.com/	1583	0.4%
https://m.vidio.com/live/204-sctv	1513	0.4%
https://www.brilio.net/selebritis/jarang-terekspos-5-seleb-cantik-ini-ternyata-anak-polisi-dan-tentara-21021…	1396	0.4%
https://m.kapanlagi.com/foto/berita-foto/indonesia/9-potret-vanessa-angel-liburan-di-bali-perut-buncitnya-j…	1321	0.3%
Other values (121277)	374672	95.2%

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.

13. time = users visit time

No.2

time
Categorical
HIGH CARDINALITY
UNIFORM

Distinct	278475
Distinct (%)	70.7%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

2021-02-13 22:21:39+00:00	7
2021-02-08 22:32:28+00:00	7
2021-02-14 21:09:25+00:00	7
2021-02-09 13:39:38+00:00	7
2021-02-13 12:03:43+00:00	7
Other values (278470)	393714

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.

14. event_time = logging time

event_time
Categorical
HIGH CARDINALITY
UNIFORM

Distinct	278861
Distinct (%)	70.8%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

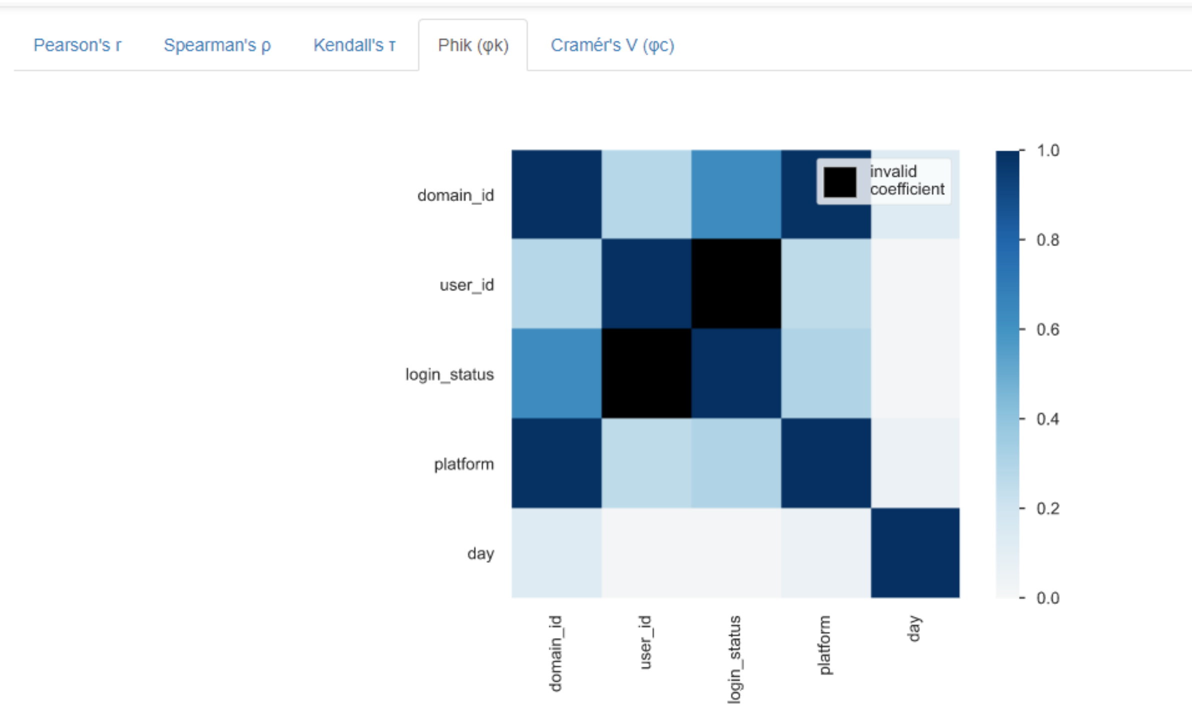
2021-02-14 21:09:25+00:00	7
2021-02-11 21:17:22+00:00	7
2021-02-13 12:03:43+00:00	7
2021-02-12 20:48:38+00:00	7
2021-02-08 20:13:23+00:00	7
Other values (278856)	393714

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.



No.2

Gambar 1.26