Activity - 3

Data Preprocessing

Task : 1

NUMERIC DATA:

In [1]:
```
#import libraries
import pandas as pd
import numpy as np
```

In [2]:
```
df1=pd.read_csv("popularity.csv")
df1
```

Out[2]:

	Unnamed: 0	avg_shares	avg_comments	avg_expert	popularity_score
0	19	147.3	23.9	19.1	14.6
1	91	28.6	1.5	33.0	7.3
2	166	17.9	37.6	21.6	8.0
3	196	94.2	4.9	8.1	9.7
4	42	293.6	27.7	1.8	20.7
...	...	...	...	...	...
195	155	4.1	11.6	5.7	3.2
196	80	76.4	26.7	22.3	11.8
197	181	218.5	5.4	27.4	12.2
198	145	140.3	1.9	9.0	10.3
199	36	266.9	43.8	5.0	25.4

200 rows × 5 columns

1.1 Remove the first column of ‘Unnamed: 0’

In [3]:
```
df1.drop(df1.columns[[0]], axis = 1,inplace = True)
df1
```

Out[3]:

	avg_shares	avg_comments	avg_expert	popularity_score
0	147.3	23.9	19.1	14.6
1	28.6	1.5	33.0	7.3
2	17.9	37.6	21.6	8.0
3	94.2	4.9	8.1	9.7
4	293.6	27.7	1.8	20.7
...	...	...	...	...
195	4.1	11.6	5.7	3.2
196	76.4	26.7	22.3	11.8
197	218.5	5.4	27.4	12.2
198	140.3	1.9	9.0	10.3
199	266.9	43.8	5.0	25.4

200 rows × 4 columns

1.2 Detect missing values, and replace them with the mean.

In [4]:
```
#detect
df1.isnull().sum()
```

Out[4]:
```
avg_shares          1
avg_comments        4
avg_expert          0
popularity_score    0
dtype: int64
```

In [6]:
```
df1.head(20)
```

Out[6]:

	avg_shares	avg_comments	avg_expert	popularity_score
0	147.3	23.9	19.1	14.6
1	28.6	1.5	33.0	7.3
2	17.9	37.6	21.6	8.0
3	94.2	4.9	8.1	9.7
4	293.6	27.7	1.8	20.7
5	137.9	46.4	59.0	19.2
6	199.8	2.6	21.2	10.6
7	168.4	NaN	12.8	11.7
8	280.2	10.1	21.4	14.8
9	19.4	16.0	22.3	6.6
10	107.4	14.0	10.9	11.5
11	177.0	9.3	6.4	12.8
12	296.4	36.3	100.9	23.8
13	237.4	27.5	11.0	18.9
14	232.1	8.6	8.7	13.4
15	206.9	8.4	26.4	12.9
16	131.1	42.8	28.9	18.0
17	191.1	28.7	18.2	17.3
18	151.5	41.3	58.5	18.5
19	NaN	7.6	7.2	9.7

In [10]:
```
#replace missing value by mean
df2 = df1.fillna(df1.mean())
```

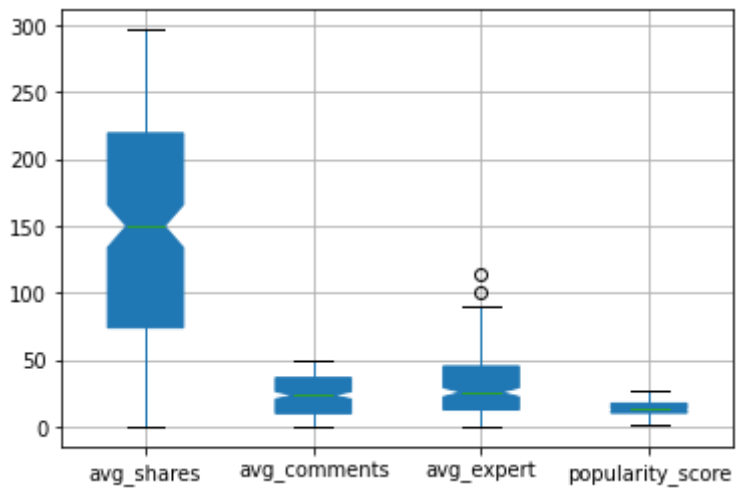
In [11]:
```
df2.head(20)
```

Out[11]:

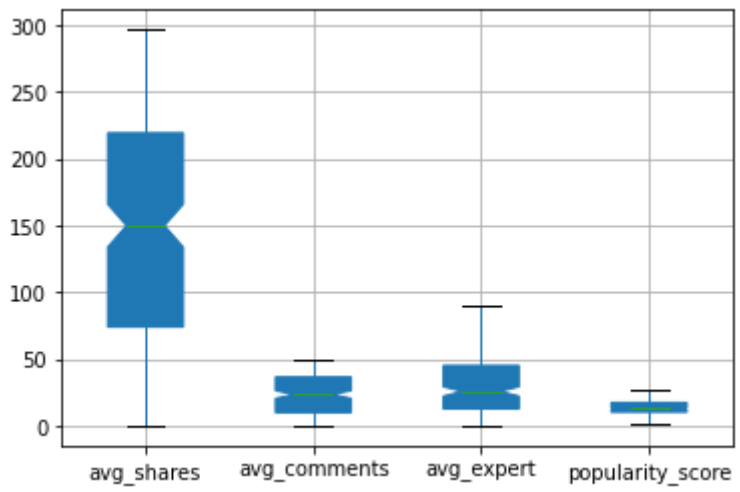
	avg_shares	avg_comments	avg_expert	popularity_score
0	147.300000	23.900000	19.1	14.6
1	28.600000	1.500000	33.0	7.3
2	17.900000	37.600000	21.6	8.0
3	94.200000	4.900000	8.1	9.7
4	293.600000	27.700000	1.8	20.7
5	137.900000	46.400000	59.0	19.2
6	199.800000	2.600000	21.2	10.6
7	168.400000	23.319388	12.8	11.7
8	280.200000	10.100000	21.4	14.8
9	19.400000	16.000000	22.3	6.6
10	107.400000	14.000000	10.9	11.5
11	177.000000	9.300000	6.4	12.8
12	296.400000	36.300000	100.9	23.8
13	237.400000	27.500000	11.0	18.9
14	232.100000	8.600000	8.7	13.4
15	206.900000	8.400000	26.4	12.9
16	131.100000	42.800000	28.9	18.0
17	191.100000	28.700000	18.2	17.3
18	151.500000	41.300000	58.5	18.5
19	147.291457	7.600000	7.2	9.7

1.3 Draw box-plots for each attribute to detect if there are any outliers. If there are outliers, ignore

In [13]:
```
# Boxplot with outliers

boxplot = df1.boxplot(column = ['avg_shares','avg_comments','avg_expert','popularity_score'],patch_artist = True,
                      notch ='True')
```



In [15]:
```
# Boxplot without outliers

boxplot = df1.boxplot(column = ['avg_shares','avg_comments','avg_expert','popularity_score'],patch_artist = True,
                      notch ='True',showfliers=False)
```



1.4 Normalize all attributes within the range of 0 to 1.

In [20]:
```
from sklearn import preprocessing

# Normalization refers to rescaling real valued numeric attributes into the range 0 and 1.
normalized_X = preprocessing.normalize(df2)
normalized_X
```

Out[20]:
```
array([[0.97452547, 0.15812056, 0.12636413, 0.09659248],
       [0.64559655, 0.03385996, 0.7449191 , 0.16478513],
       [0.37613592, 0.79009557, 0.45388469, 0.16810544],
       [0.98980656, 0.05148675, 0.08511076, 0.10192276],
       [0.99311693, 0.09369666, 0.00608859, 0.0700188 ],
       [0.87182477, 0.29334786, 0.37300697, 0.12138532],
       [0.99295394, 0.01292132, 0.10535848, 0.05267924],
       [0.98543442, 0.13645919, 0.07490238, 0.06846546],
       [0.99507553, 0.03586818, 0.07599792, 0.05255931],
       [0.56638775, 0.46712392, 0.65105396, 0.19268862],
       [0.98116689, 0.12789885, 0.09957839, 0.10505977],
       [0.99538328, 0.0522998 , 0.03599126, 0.07198252],
       [0.9376841 , 0.11483783, 0.31920488, 0.07529312],
       [0.98922489, 0.11459008, 0.04583603, 0.07875464],
       [0.99695836, 0.03694029, 0.03736983, 0.05755813],
       [0.98926685, 0.04016356, 0.12622835, 0.06167976],
       [0.92291315, 0.30130193, 0.2034492 , 0.12671576],
       [0.98066522, 0.14727939, 0.09339669, 0.08877817],
       [0.89863215, 0.24497365, 0.34699657, 0.10973396],
       [0.99533835, 0.05135784, 0.0486548 , 0.06554883],
       [0.97684844, 0.15928643, 0.09427156, 0.10727454],
       [0.68975276, 0.42729463, 0.5617244 , 0.1616358 ],
       [0.99765272, 0.01767097, 0.02978821, 0.05907154],
       [0.98941316, 0.06408652, 0.11287497, 0.06491344],
       [0.73514827, 0.48616231, 0.44974934, 0.14466773],
       [0.94784942, 0.12344487, 0.27365052, 0.1070431 ],
       [0.9785445 , 0.11268671, 0.15270127, 0.08021649],
       [0.93051956, 0.1764892 , 0.30424631, 0.10207398],
       [0.78255531, 0.04553049, 0.58905072, 0.19635024],
       [0.92078137, 0.08918563, 0.3326383 , 0.18319211],
       [0.99207159, 0.01451306, 0.07671191, 0.09848151],
       [0.98625233, 0.11990668, 0.08059953, 0.08020248],
       [0.96687433, 0.21271235, 0.09244262, 0.106592  ],
       [0.49072941, 0.26490703, 0.82512025, 0.09047371],
       [0.96264997, 0.15731019, 0.20151915, 0.08915472],
       [0.98620851, 0.1236968 , 0.04543964, 0.10013551],
       [0.94726118, 0.21953455, 0.20847488, 0.10506691],
       [0.98317871, 0.14022707, 0.04449861, 0.10823986],
       [0.9973804 , 0.03032435, 0.03613998, 0.05483308],
       [0.99096398, 0.06892586, 0.0945151 , 0.06562402],
       [0.97812708, 0.16920498, 0.07542872, 0.0945917 ],
       [0.98226944, 0.09490688, 0.14487552, 0.07176704],
       [0.94448464, 0.05146809, 0.31061445, 0.09390669],
       [0.83571882, 0.3638193 , 0.39121992, 0.12710842],
       [0.97117493, 0.14180428, 0.17164017, 0.08514961],
       [0.82613828, 0.27287591, 0.47742836, 0.12286759],
       [0.97251747, 0.06763106, 0.21359429, 0.06335061],
       [0.94788063, 0.05662119, 0.30617383, 0.06763086],
       [0.94986744, 0.273769  , 0.0776912 , 0.12948534],
       [0.97864426, 0.13795866, 0.11927676, 0.09484658],
       [0.19514017, 0.61544208, 0.75387485, 0.12175412],
       [0.80834817, 0.37447439, 0.43473464, 0.13171168],
       [0.44112274, 0.4516675 , 0.76098066, 0.14938419],
       [0.79305681, 0.15967587, 0.57350249, 0.12907133],
       [0.16747749, 0.92732907, 0.29153489, 0.16437606],
       [0.9702547 , 0.2158079 , 0.01409358, 0.1087848 ],
       [0.98928164, 0.09459267, 0.03547225, 0.10543141],
       [0.95651988, 0.14301236, 0.23879738, 0.08713776],
       [0.46881629, 0.11175272, 0.86131364, 0.16081489],
       [0.95119426, 0.01467892, 0.29357848, 0.09394511],
       [0.91880711, 0.1014068 , 0.36567908, 0.10857698],
       [0.96024957, 0.14220903, 0.22414852, 0.0863412 ],
       [0.94435972, 0.1277918 , 0.28049216, 0.11479602],
       [0.99766786, 0.03719761, 0.02245893, 0.05263812],
       [0.99133041, 0.10441186, 0.02134815, 0.07685334],
       [0.96017123, 0.19345555, 0.17489961, 0.10028104],
       [0.85933018, 0.15028644, 0.47269583, 0.12459645],
       [0.98733273, 0.08286126, 0.116706  , 0.06846752],
       [0.97315527, 0.16091208, 0.1365832 , 0.09176683],
       [0.70090598, 0.68204101, 0.12334784, 0.16833353],
       [0.98591237, 0.14196807, 0.02193676, 0.08567752],
       [0.09640672, 0.54187225, 0.83109241, 0.07978487],
       [0.99720169, 0.0484654 , 0.01290086, 0.05543884],
       [0.791417  , 0.45679102, 0.36472461, 0.17882129],
       [0.98719689, 0.10752828, 0.09086338, 0.07499205],
       [0.93179413, 0.14360689, 0.3185762 , 0.09821391],
       [0.98937435, 0.10846586, 0.05988219, 0.07607675],
       [0.9652571 , 0.14711607, 0.19805776, 0.08607388],
       [0.98856933, 0.07317924, 0.11344948, 0.06711706],
       [0.97962155, 0.01474699, 0.18679522, 0.07233048],
       [0.97960789, 0.15401365, 0.09402939, 0.0883552 ],
       [0.74624162, 0.21550579, 0.62031863, 0.10902058],
       [0.97448374, 0.06468556, 0.19405668, 0.09240794],
       [0.99623118, 0.01654024, 0.05280152, 0.06679711],
       [0.99541964, 0.00827418, 0.08526001, 0.04245013],
       [0.58179707, 0.59663884, 0.50461991, 0.22559478],
       [0.98338183, 0.1116875 , 0.11700595, 0.08243601],
       [0.56678378, 0.36674245, 0.70923746, 0.20307226],
```

```
[0.01724902, 0.9758017 , 0.21438068, 0.03942633],
[0.2443368 , 0.29431478, 0.91811402, 0.10365803],
[0.82162253, 0.30133582, 0.46907447, 0.11875303],
[0.90629693, 0.21026089, 0.33351727, 0.15225788],
[0.96680392, 0.14858965, 0.18792221, 0.08886244],
[0.97104618, 0.14620741, 0.16690544, 0.08851713],
[0.95046906, 0.12075868, 0.27387376, 0.08377995],
[0.59113524, 0.52205877, 0.59910561, 0.13815296],
[0.9202487 , 0.23283401, 0.29583022, 0.10684158],
[0.9709041 , 0.19907271, 0.08560806, 0.10191436],
[0.98841702, 0.09047777, 0.09416   , 0.07732693],
[0.97307034, 0.12503759, 0.17505263, 0.08280267],
[0.99614132, 0.01326168, 0.07388648, 0.0454686 ],
[0.96922989, 0.10234104, 0.21141039, 0.07365722],
[0.98863827, 0.04895644, 0.1303157 , 0.05670494],
[0.98849988, 0.11395078, 0.06068111, 0.07874648],
[0.95427216, 0.13332276, 0.25070476, 0.09347085],
[0.97494555, 0.10015604, 0.17930033, 0.08544781],
[0.84982063, 0.20751434, 0.09881635, 0.47431849],
[0.99043367, 0.11071452, 0.01845242, 0.08026803],
[0.98223352, 0.132388  , 0.01281174, 0.132388  ],
[0.93125232, 0.18020559, 0.30167329, 0.09637578],
[0.60995025, 0.26837811, 0.7246209 , 0.17566567],
[0.75947839, 0.49410509, 0.39528407, 0.1510074 ],
[0.95662176, 0.18051507, 0.20915967, 0.09241939],
[0.9848822 , 0.00854704, 0.15976394, 0.06640394],
[0.99360537, 0.02498427, 0.09321056, 0.05861695],
[0.98695873, 0.01687461, 0.15187146, 0.05062382],
[0.99505523, 0.07492886, 0.00112393, 0.06518811],
[0.90292831, 0.27849194, 0.30654149, 0.11486958],
[0.65865602, 0.69852204, 0.20626333, 0.18893028],
[0.28857631, 0.93443759, 0.07214408, 0.19581964],
[0.93192596, 0.18847941, 0.2737439 , 0.14509923],
[0.96157105, 0.1852935 , 0.17596217, 0.10042285],
[0.96794801, 0.19600054, 0.12143997, 0.09956292],
[0.9188941 , 0.03435118, 0.36755764, 0.13912228],
[0.96969416, 0.1712461 , 0.14638214, 0.09455306],
[0.99352285, 0.03624916, 0.06879942, 0.08285522],
[0.99110994, 0.09740871, 0.05059578, 0.07518439],
[0.92573471, 0.22992104, 0.27925632, 0.11030625],
[0.76567808, 0.59795812, 0.05347593, 0.23091879],
[0.94211621, 0.15476746, 0.28333091, 0.09048574],
[0.94829126, 0.13795389, 0.27326209, 0.0839062 ],
[0.96788989, 0.03514923, 0.2421868 , 0.05743899],
[0.95342138, 0.22433444, 0.17051227, 0.10764435],
[0.12067351, 0.60182045, 0.78283072, 0.10210835],
[0.94034552, 0.08955672, 0.31977588, 0.07398164],
[0.96405624, 0.19416589, 0.15262342, 0.09798604],
[0.95127289, 0.14638329, 0.25780937, 0.08477122],
[0.98913837, 0.11113322, 0.05003629, 0.08216485],
[0.99354663, 0.02134409, 0.09835024, 0.05231395],
[0.98979282, 0.0597701 , 0.10519537, 0.07531032],
[0.88337213, 0.2381468 , 0.38126951, 0.13256448],
[0.87252881, 0.08849492, 0.46110508, 0.13507119],
[0.99682665, 0.01523926, 0.05871596, 0.05154454],
[0.92901149, 0.08151689, 0.34012221, 0.12086988],
[0.98718776, 0.11640422, 0.07094163, 0.08293176],
[0.56409553, 0.69459524, 0.40623298, 0.1852254 ],
[0.80725471, 0.46048617, 0.32992149, 0.16566271],
[0.7936371 , 0.4749644 , 0.35013402, 0.14817266],
[0.96778837, 0.04263055, 0.23409409, 0.08226949],
[0.9876338 , 0.        , 0.11329465, 0.1083688 ],
[0.97871361, 0.14000142, 0.07777856, 0.12833463],
[0.97608673, 0.04473253, 0.20186988, 0.06709879],
[0.98572315, 0.1380106 , 0.04350843, 0.08608119],
[0.93924469, 0.01361805, 0.3396501 , 0.04766316],
[0.9400113 , 0.13229789, 0.30391651, 0.08068942],
[0.96605297, 0.22973604, 0.03871956, 0.11164139],
[0.97606179, 0.10241576, 0.17371071, 0.08147994],
[0.20142628, 0.53752711, 0.81156054, 0.10891072],
[0.96825675, 0.13625788, 0.19199974, 0.08395688],
[0.96328744, 0.24255439, 0.02079038, 0.11319205],
[0.96442483, 0.00320052, 0.24750726, 0.09281522],
[0.98344616, 0.09448182, 0.13363175, 0.07777785],
[0.96106683, 0.12189355, 0.23498612, 0.0792088 ],
[0.98366648, 0.01526209, 0.17034465, 0.05612511],
[0.95983344, 0.15472082, 0.21632699, 0.08935245],
[0.93412001, 0.22803237, 0.25289124, 0.10708438],
[0.80858017, 0.33777328, 0.46514588, 0.1254856 ],
[0.43272528, 0.45266114, 0.76928939, 0.1266513 ],
[0.99850702, 0.01408283, 0.02919611, 0.04396591],
[0.96719362, 0.22963656, 0.00889253, 0.10827965],
[0.14352099, 0.5524575 , 0.81394094, 0.10813225],
[0.32177962, 0.37141584, 0.86264323, 0.11981156],
[0.78088684, 0.47034812, 0.37228326, 0.17433753],
[0.91240594, 0.14037015, 0.36894585, 0.10812295],
[0.30163094, 0.44484227, 0.83518819, 0.11659683],
[0.91456638, 0.16714326, 0.35695002, 0.09065397],
```

```
[0.96060072, 0.22719602, 0.11319942, 0.11319942],
[0.92213905, 0.06401626, 0.37114192, 0.08840341],
[0.98751271, 0.10279206, 0.0723089 , 0.09499404],
[0.97461348, 0.01020538, 0.18879947, 0.11991318],
[0.9506225 , 0.22017142, 0.16707126, 0.14116874],
[0.1668177 , 0.43135702, 0.88245577, 0.08587657],
[0.99619845, 0.01067355, 0.06937811, 0.05158885],
[0.91339982, 0.32920701, 0.19153863, 0.14365397],
[0.99292131, 0.08702896, 0.01356295, 0.07968236],
[0.67744531, 0.70488614, 0.09947298, 0.18522555],
[0.91887889, 0.27221123, 0.24299832, 0.15004814],
[0.81672379, 0.23975424, 0.51459446, 0.10330874],
[0.52773971, 0.80397846, 0.19171794, 0.19584091],
[0.98595032, 0.01078151, 0.15986374, 0.04721557],
[0.95156355, 0.26745816, 0.09585206, 0.11749607],
[0.75858354, 0.35353427, 0.5323216 , 0.12727234],
[0.97437846, 0.01883915, 0.21818367, 0.05126002],
[0.44795292, 0.01367795, 0.87538891, 0.18123286],
[0.95790787, 0.10495184, 0.25615364, 0.07604561],
[0.29428701, 0.8326169 , 0.40913072, 0.22968742],
[0.90123546, 0.31496056, 0.26305695, 0.13919605],
[0.99041284, 0.02447702, 0.12419822, 0.05529994],
[0.9951909 , 0.01347728, 0.06383976, 0.07306106],
[0.98231137, 0.16120359, 0.01840224, 0.09348336]])
```

Task : 2

CATEGORICAL DATA:

In [21]:
```
df3=pd.read_csv('test.csv')
df3
```

Out[21]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows × 11 columns

2.1 Detect missing values in the age column, and replace them with the mean.

In [22]:
```
df3.isnull().sum()
```

Out[22]:
```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [23]:
```
df3['Age'] = df3['Age'].fillna(df3['Age'].mean())
df3
```

Out[23]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.50000	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.00000	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.00000	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.00000	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.00000	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	30.27259	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.00000	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.50000	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	30.27259	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	30.27259	1	1	2668	22.3583	NaN	C

418 rows × 11 columns

2.2 Encode each variable in columns-sex and embarkeed to integers.

In [24]:
```
from sklearn.preprocessing import LabelEncoder
df4=pd.read_csv('test.csv')
df4['Sex'] = df4['Sex'].fillna('None')
label_encoder = LabelEncoder()
df4['Sex'] = label_encoder.fit_transform(df4['Sex'])
df4['Embarked'] = label_encoder.fit_transform(df4['Embarked'])
```

In [25]:
```
df4
```

Out[25]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	1	34.5	0	0	330911	7.8292	NaN	1
1	893	3	Wilkes, Mrs. James (Ellen Needs)	0	47.0	1	0	363272	7.0000	NaN	2
2	894	2	Myles, Mr. Thomas Francis	1	62.0	0	0	240276	9.6875	NaN	1
3	895	3	Wirz, Mr. Albert	1	27.0	0	0	315154	8.6625	NaN	2
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	0	22.0	1	1	3101298	12.2875	NaN	2
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	1	NaN	0	0	A.5. 3236	8.0500	NaN	2
414	1306	1	Oliva y Ocana, Dona. Fermina	0	39.0	0	0	PC 17758	108.9000	C105	0
415	1307	3	Saether, Mr. Simon Sivertsen	1	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	2
416	1308	3	Ware, Mr. Frederick	1	NaN	0	0	359309	8.0500	NaN	2
417	1309	3	Peter, Master. Michael J	1	NaN	1	1	2668	22.3583	NaN	0

418 rows × 11 columns

Drop the columns which are not required and in the last cell print the first 5 rows of the data after performing all mentioned tasks on it.

In [28]:
```
df4.drop(df4.columns[[0,1,2,5,6,7,8,9]], axis = 1,inplace = True)
df4.head()
```

Out[28]:

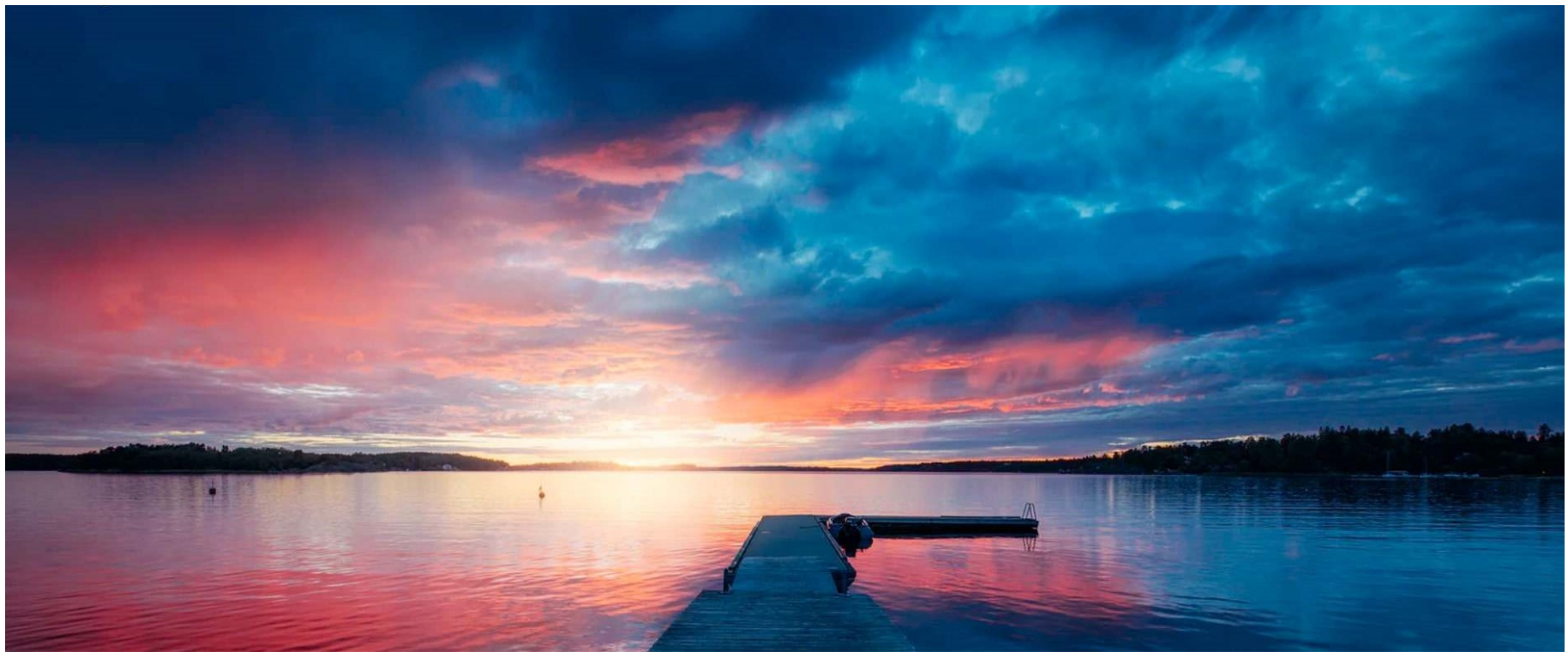
	Sex	Age	Embarked
0	1	34.5	1
1	0	47.0	2
2	1	62.0	1
3	1	27.0	2
4	0	22.0	2

Task : 3

IMAGE DATA:

In [29]:
```
from IPython.display import Image
Image(filename="SampleImage.jpg")
```

Out[29]:



3.1 Convert image to Black and white.

In [30]:
```
import PIL
```

In [31]:
```
from PIL import Image
```

In [32]:
```
image = Image.open("SampleImage.jpg")
gray_img = image.convert(mode='L')
```

In [33]:
```
gray_img
```

Out[33]:

In [34]:
```
gray_img.save("SampleImage1.jpg")
```

In [36]:
```
image2 = Image.open("SampleImage1.jpg")
```

In [37]:
```
#save in photos
image.show()
image2.show()
```

3.2 Resize image to 100*100.

In [38]:
```
print(image2.size)
```

```
(1884, 779)
```

In [39]:
```
img_resized = image2.resize((100,100))
img_resized
```

Out[39]:

In [40]:
```
from matplotlib import image
from matplotlib import pyplot
```

In [41]:
```
img_resized.show()
```

In [42]:
```
img_resized.save("img_resize.jpg")
```

In [52]:
```
from os import listdir
img = image.imread('SampleImage.jpg')
data=image.imread('img_resize.jpg')
```

In [53]:
```
print(data.dtype)
print(data.shape)
```

```
uint8
(100, 100)
```

In [54]:
```
data.size, data.shape, data.ndim
```

Out[54]:
```
(10000, (100, 100), 2)
```

3.3 Convert given image into a numpy array.

In [56]:
```
pyplot.imshow(img)
```

Out[56]:
```
<matplotlib.image.AxesImage at 0x1afe51c26a0>
```

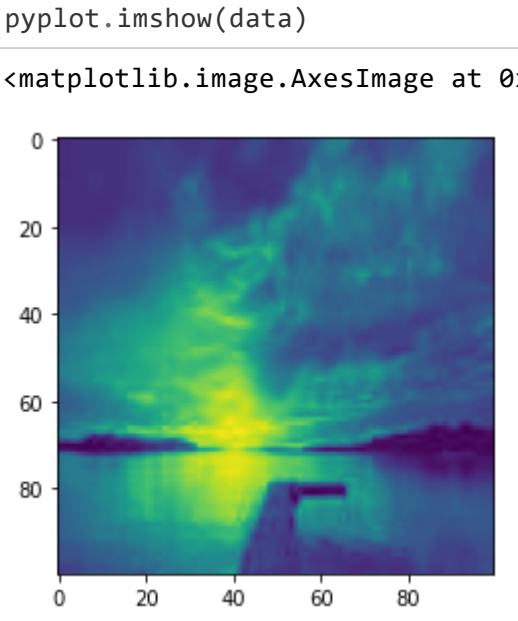
In [57]:
```
img
```

Out[57]:
```
array([[[  0,  38,  87],
        [  0,  38,  87],
        [  0,  38,  87],
        ...,
        [  0,  65, 121],
        [  0,  65, 121],
        [  0,  65, 121]],

       [[  0,  38,  87],
        [  0,  38,  87],
        [  0,  38,  87],
        ...,
        [  0,  64, 120],
        [  0,  64, 120],
        [  0,  64, 120]],

       [[  0,  38,  87],
        [  0,  38,  87],
        [  0,  38,  87],
        ...,
        [  0,  64, 120],
        [  0,  64, 120],
        [  0,  64, 120]],

       ...,

       [[ 88,  39,  60],
        [ 90,  41,  62],
        [ 92,  43,  64],
        ...,
        [  4,  82, 128],
        [  4,  82, 128],
        [  4,  82, 128]],

       [[ 94,  45,  66],
        [ 94,  45,  66],
        [ 95,  46,  67],
        ...,
        [  3,  81, 127],
        [  3,  81, 127],
        [  3,  81, 127]],

       [[ 95,  46,  67],
        [ 95,  46,  67],
        [ 95,  46,  67],
        ...,
        [  2,  79, 125],
        [  2,  79, 125],
        [  2,  79, 125]]], dtype=uint8)
```

Print the numpy array after step 3.3.

In [49]:
```
pyplot.imshow(data)
```

Out[49]:
```
<matplotlib.image.AxesImage at 0x1afe4fe45f8>
```



In [50]:
```
data
```

Out[50]:
```
array([[34, 34, 33, ..., 64, 59, 55],
       [34, 34, 33, ..., 68, 61, 56],
       [34, 34, 33, ..., 70, 62, 56],
       ...,
       [54, 56, 57, ..., 59, 60, 61],
       [49, 50, 51, ..., 60, 63, 63],
       [49, 50, 50, ..., 58, 61, 62]], dtype=uint8)
```

Task : 4

In [58]:
```
data=pd.read_csv("a3-Q4.csv")
data
```

Out[58]:

	tweet_id	sentiment	author	content
0	1956967341	empty	xoshayzers	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	wannamama	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	coolfunky	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	czareaquino	wants to hang out with friends SOON!
4	1956968416	neutral	xkilljoyx	@dannycastillo We want to trade with someone w...
...	...	...	...	...
39995	1753918954	neutral	showMe_Heaven	@JohnLloydTaylor
39996	1753919001	love	drapeaux	Happy Mothers Day All my love
39997	1753919005	love	JenniRox	Happy Mother's Day to all the mommies out ther...
39998	1753919043	happiness	ipdaman1	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...
39999	1753919049	love	Alpharalpha	@mopedronin bullet train from tokyo the gf ...

40000 rows × 4 columns

In [60]:
```
#drop unwanted attribute
data.drop(data.columns[[0,1,2]], axis = 1,inplace = True)
data
```

Out[60]:

	content
0	@tiffanylue i know i was listenin to bad habi...
1	Layin n bed with a headache ughhhh...waitin o...
2	Funeral ceremony...gloomy friday...
3	wants to hang out with friends SOON!
4	@dannycastillo We want to trade with someone w...
...	...
39995	@JohnLloydTaylor
39996	Happy Mothers Day All my love
39997	Happy Mother's Day to all the mommies out ther...
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...
39999	@mopedronin bullet train from tokyo the gf ...

40000 rows × 1 columns

4.1 Split into Words

In [67]:
```
data['split_content'] = data['content'].str.split()
```

In [68]:
```
data
```

Out[68]:

	content	split_content
0	@tiffanylue i know i was listenin to bad habi...	[@tiffanylue, i, know, i, was, listenin, to, b...
1	Layin n bed with a headache ughhhh...waitin o...	[Layin, n, bed, with, a, headache, ughhhh...wa...
2	Funeral ceremony...gloomy friday...	[Funeral, ceremony...gloomy, friday...]
3	wants to hang out with friends SOON!	[wants, to, hang, out, with, friends, SOON!]
4	@dannycastillo We want to trade with someone w...	[@dannycastillo, We, want, to, trade, with, so...
...	...	...
39995	@JohnLloydTaylor	[@JohnLloydTaylor]
39996	Happy Mothers Day All my love	[Happy, Mothers, Day, All, my, love]
39997	Happy Mother's Day to all the mommies out ther...	[Happy, Mother's, Day, to, all, the, mommies, ...
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!...
39999	@mopedronin bullet train from tokyo the gf ...	[@mopedronin, bullet, train, from, tokyo, the,...

40000 rows × 2 columns

4.2 Filter out Punctuation

In [69]:
```
import re
import string
def remove_punctuations(text):
    for punctuation in string.punctuation:
        text = text.replace(punctuation,'')
    return text
```

In [70]:
```
data['clean_content'] = data['content'].apply(remove_punctuations)
data
```

Out[70]:

	content	split_content	clean_content
0	@tiffanylue i know i was listenin to bad habi...	[@tiffanylue, i, know, i, was, listenin, to, b...	tiffanylue i know i was listenin to bad habit...
1	Layin n bed with a headache ughhhh...waitin o...	[Layin, n, bed, with, a, headache, ughhhh...wa...	Layin n bed with a headache ughhhhwaitin on y...
2	Funeral ceremony...gloomy friday...	[Funeral, ceremony...gloomy, friday...]	Funeral ceremonygloomy friday
3	wants to hang out with friends SOON!	[wants, to, hang, out, with, friends, SOON!]	wants to hang out with friends SOON
4	@dannycastillo We want to trade with someone w...	[@dannycastillo, We, want, to, trade, with, so...	dannycastillo We want to trade with someone wh...
...	...	...	...
39995	@JohnLloydTaylor	[@JohnLloydTaylor]	JohnLloydTaylor
39996	Happy Mothers Day All my love	[Happy, Mothers, Day, All, my, love]	Happy Mothers Day All my love
39997	Happy Mother's Day to all the mommies out ther...	[Happy, Mother's, Day, to, all, the, mommies, ...	Happy Mothers Day to all the mommies out there...
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!...	niariley WASSUP BEAUTIFUL FOLLOW ME PEEP OUT ...
39999	@mopedronin bullet train from tokyo the gf ...	[@mopedronin, bullet, train, from, tokyo, the,...	mopedronin bullet train from tokyo the gf a...

40000 rows × 3 columns

4.3 Filter out stop words

In [71]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\lenovo\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[71]: True

In [72]:
```
from nltk.corpus import stopwords
stopwords = stopwords.words('english')
```

In [73]:
```
def remove_stopwords(txt_tokenized):
    txt_clean = [word for word in txt_tokenized if word not in stopwords]
    return txt_clean

data['stop_content'] = data['split_content'].apply(lambda x: remove_stopwords(x))
data
```

Out[73]:

	content	split_content	clean_content	stop_content
0	@tiffanylue i know i was listenin to bad habi...	[@tiffanylue, i, know, i, was, listenin, to, b...	tiffanylue i know i was listenin to bad habit...	[@tiffanylue, know, listenin, bad, habit, earl...
1	Layin n bed with a headache ughhhh...waitin o...	[Layin, n, bed, with, a, headache, ughhhh...wa...	Layin n bed with a headache ughhhhwaitin on y...	[Layin, n, bed, headache, ughhhh...waitin, cal...
2	Funeral ceremony...gloomy friday...	[Funeral, ceremony...gloomy, friday...]	Funeral ceremonygloomy friday	[Funeral, ceremony...gloomy, friday...]
3	wants to hang out with friends SOON!	[wants, to, hang, out, with, friends, SOON!]	wants to hang out with friends SOON	[wants, hang, friends, SOON!]
4	@dannycastillo We want to trade with someone w...	[@dannycastillo, We, want, to, trade, with, so...	dannycastillo We want to trade with someone wh...	[@dannycastillo, We, want, trade, someone, Hou...
...	...	...	...	...
39995	@JohnLloydTaylor	[@JohnLloydTaylor]	JohnLloydTaylor	[@JohnLloydTaylor]
39996	Happy Mothers Day All my love	[Happy, Mothers, Day, All, my, love]	Happy Mothers Day All my love	[Happy, Mothers, Day, All, love]
39997	Happy Mother's Day to all the mommies out ther...	[Happy, Mother's, Day, to, all, the, mommies, ...	Happy Mothers Day to all the mommies out there...	[Happy, Mother's, Day, mommies, there,, woman,...
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!...	niariley WASSUP BEAUTIFUL FOLLOW ME PEEP OUT ...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!...
39999	@mopedronin bullet train from tokyo the gf ...	[@mopedronin, bullet, train, from, tokyo, the,...	mopedronin bullet train from tokyo the gf a...	[@mopedronin, bullet, train, tokyo, gf, visiti...

40000 rows × 4 columns

4.4 Stem words

In [75]:
```
from nltk.stem import PorterStemmer
ps = PorterStemmer()
dir(ps)
```

Out[75]:
```
['MARTIN_EXTENSIONS',
 'NLTK_EXTENSIONS',
 'ORIGINAL_ALGORITHM',
 '__abstractmethods__',
 '__class__',
 '__delattr__',
 '__dict__',
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__lt__',
 '__module__',
 '__ne__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
 '__subclasshook__',
 '__weakref__',
 '_abc_impl',
 '_apply_rule_list',
 '_contains_vowel',
 '_ends_cvc',
 '_ends_double_consonant',
 '_has_positive_measure',
 '_is_consonant',
 '_measure',
 '_replace_suffix',
 '_step1a',
 '_step1b',
 '_step1c',
 '_step2',
 '_step3',
 '_step4',
 '_step5a',
 '_step5b',
 'mode',
 'pool',
 'stem',
 'vowels']
```

In [76]:
```
pd.set_option('display.max_colwidth',100)
```

In [77]:
```
def stemming(tokenized_text):
    text = [ps.stem(word) for word in tokenized_text]
    return text

data['stemmed'] = data['stop_content'].apply(lambda x: stemming(x))
data
```

Out[77]:

	content	split_content	clean_content	stop_content	stemmed
0	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[	[@tiffanylue, i, know, i, was, listenin, to, bad, habit, earlier, and, i, started, freakin, at, ...	tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part	[@tiffanylue, know, listenin, bad, habit, earlier, started, freakin, part, =[]	[@tiffanylu, know, listenin, bad, habit, earlier, start, freakin, part, =[]
1	Layin n bed with a headache ughhhh...waitin on your call...	[Layin, n, bed, with, a, headache, ughhhh...waitin, on, your, call...]	Layin n bed with a headache ughhhhwaitin on your call	[Layin, n, bed, headache, ughhhh...waitin, call...]	[layin, n, bed, headach, ughhhh...waitin, call...]
2	Funeral ceremony...gloomy friday...	[Funeral, ceremony...gloomy, friday...]	Funeral ceremonygloomy friday	[Funeral, ceremony...gloomy, friday...]	[funer, ceremony...gloomi, friday...]
3	wants to hang out with friends SOON!	[wants, to, hang, out, with, friends, SOON!]	wants to hang out with friends SOON	[wants, hang, friends, SOON!]	[want, hang, friend, soon!]
4	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.	[@dannycastillo, We, want, to, trade, with, someone, who, has, Houston, tickets,, but, no, one, ...	dannycastillo We want to trade with someone who has Houston tickets but no one will	[@dannycastillo, We, want, trade, someone, Houston, tickets,, one, will.]	[@dannycastillo, we, want, trade, someon, houston, tickets,, one, will.]
...	...	...	...	...	...
39995	@JohnLloydTaylor	[@JohnLloydTaylor]	JohnLloydTaylor	[@JohnLloydTaylor]	[@johnlloydtaylor]
39996	Happy Mothers Day All my love	[Happy, Mothers, Day, All, my, love]	Happy Mothers Day All my love	[Happy, Mothers, Day, All, love]	[happi, mother, day, all, love]
39997	Happy Mother's Day to all the mommies out there, be you woman or man as long as you're 'momma' t...	[Happy, Mother's, Day, to, all, the, mommies, out, there,, be, you, woman, or, man, as, long, as...	Happy Mothers Day to all the mommies out there be you woman or man as long as youre momma to som...	[Happy, Mother's, Day, mommies, there,, woman, man, long, 'momma', someone, day!]	[happi, mother', day, mommi, there,, woman, man, long, 'momma', someon, day!]
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEEP OUT MY NEW HIT SINGLES WWW.MYSPACE.COM/IPSOHOT I...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!, PEEP, OUT, MY, NEW, HIT, SINGLES, WWW.MYSPACE.CO...	niariley WASSUP BEAUTIFUL FOLLOW ME PEEP OUT MY NEW HIT SINGLES WWWMYSPACECOMIPSOHOT I DEF WAT ...	[@niariley, WASSUP, BEAUTIFUL!!!, FOLLOW, ME!!, PEEP, OUT, MY, NEW, HIT, SINGLES, WWW.MYSPACE.CO...	[@niariley, wassup, beautiful!!!, follow, me!!, peep, out, my, new, hit, singl, www.myspace.com/...
39999	@mopedronin bullet train from tokyo the gf and i have been visiting japan since thursday vac...	[@mopedronin, bullet, train, from, tokyo, the, gf, and, i, have, been, visiting, japan, since, t...	mopedronin bullet train from tokyo the gf and i have been visiting japan since thursday vaca...	[@mopedronin, bullet, train, tokyo, gf, visiting, japan, since, thursday, vacation/sightseeing, ...	[@mopedronin, bullet, train, tokyo, gf, visit, japan, sinc, thursday, vacation/sightse, gaijin, ...

40000 rows × 5 columns

In [ ]:
```

```