

```
from google.colab import drive  
drive.mount("/content/gdrive/")
```

Mounted at /content/gdrive/

```
import pandas as pd  
import csv
```

```
profile_data = pd.read_csv("/content/gdrive/MyDrive/MSING055_Amazon_Review-master/final_datas  
profile_data
```



$$[0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1]$$


```
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but..	5.0	These headphones are pretty damn good for the ...
1	https://www.amazon.co.uk/gp/review/R3M2146VNTD...	Best Price for Great Workout / Running Headphones	5.0	These are amazing for £10. AMAZING. They stick...
2	https://www.amazon.co.uk/gp/review/R3U00QINFHP...	Amazed at how good these sound for the price.	5.0	They fit very comfortably in my ear and have p...
3	https://www.amazon.co.uk/gp/review/R5BZDAV3JC42E/	Earbuds that give you great sound for a small ...	5.0	These earbuds will change the way you look at ...
4	https://www.amazon.co.uk/gp/review/R1AS6K8FBV8...	Great price	5.0	These headphones are a great value for this pr...
...
7647	https://www.amazon.co.uk/gp/review/R13ECIUNXJW...	Bootleg width too large	3.0	Rockport Women's Jاليايا Buckle Tall Heels Boo...
7648	https://www.amazon.co.uk/gp/review/R126KTBSZV2...	Size problems	2.0	I order my usual size (36C) and this is much ...
7649	https://www.amazon.co.uk/gp/review/R2KO58OARSW...	Be aware about size differences from garments ...	4.0	The garment is OK however one needs to be aware...
7650	https://www.amazon.co.uk/gp/review/R1Q1317KGI15	Different	2.0	The colour on the page does not

▼ cleaning helpful column

```
profile_data['helpful'] = profile_data['helpful'].fillna(0)
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but	5.0	These headphones are pretty

```
profile_data['product_link'] = profile_data['product_link'].fillna(0)
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but..	5.0	These headphones are pretty damn good for the ...
1	https://www.amazon.co.uk/gp/review/R3M2146VNTD	Best Price for Great Workout /	5.0	These are amazing for £10

▼ cleaning date column

```
new_date=profile_data['date'].tolist()
new_date
```

```
['25 April 2017',
 '3 May 2017',
 '28 April 2017',
 '25 April 2017',
 '25 March 2017',
 '20 June 2017',
 '20 June 2017',
 '20 June 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '9 March 2017',
 '16 October 2016',
 '29 July 2015',
 '28 November 2014',
 '24 March 2014',
 '24 March 2014',
 '24 March 2014',
 '4 May 2017',
 '27 April 2017',
 '24 April 2017',
 '31 March 2017',
 '2 July 2017',
 '3 May 2017',
 '26 April 2017',
 '4 May 2017',
 '28 April 2017',
 '25 April 2017',
 '3 May 2017',
 '28 April 2017',
 '26 April 2017',
 '4 May 2017',
 '27 April 2017',
```

Amazed at

they're very


```
'24 April 2017',
'5 April 2017',
'4 May 2017',
'25 April 2017',
'6 April 2017',
'3 May 2017',
'28 April 2017',
'25 April 2017',
'4 May 2017',
'28 April 2017',
'25 April 2017',
'3 May 2017',
'30 April 2017',
'28 April 2017',
'25 April 2017',
'5 May 2017',
'3 May 2017',
'25 April 2017',
'3 May 2017',
'28 April 2017',
'25 April 2017',
... ..
```

```
date_reformat=[(str(item))[3:] for item in new_date]
date_reformat
```

```
['April 2017',
'ay 2017',
'April 2017',
'April 2017',
'March 2017',
'June 2017',
'June 2017',
'June 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'arch 2017',
'October 2016',
'July 2015',
'November 2014',
'March 2014',
'March 2014',
'March 2014',
'ay 2017',
'April 2017',
'April 2017',
'March 2017',
'uly 2017',
'ay 2017',
'April 2017',
'ay 2017',
```

```
'April 2017',  
'April 2017',  
'ay 2017',  
'April 2017',  
'April 2017',  
'ay 2017',  
'April 2017',  
'April 2017',  
'pril 2017',  
'ay 2017',  
'April 2017',  
'pril 2017',  
'ay 2017',  
'April 2017',  
'April 2017',  
'ay 2017',  
'April 2017',  
'April 2017',  
'ay 2017',  
'April 2017',  
'April 2017',  
'ay 2017',  
'ay 2017',  
'April 2017',  
'ay 2017',  
'April 2017',  
'April 2017'.
```

```
profile_data['date'] = date_reformat  
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but..	5.0	These headphones are pretty damn good for the ...
1	https://www.amazon.co.uk/gp/review/R3M2146VNTD...	Best Price for Great Workout / Running Headphones	5.0	These are amazing for £10. AMAZING. They stick...
2	https://www.amazon.co.uk/gp/review/R3U00QINFHP...	Amazed at how good these sound for the price.	5.0	They fit very comfortably in my ear and have p...
3	https://www.amazon.co.uk/gp/review/R5BZDAV3JC42E/	Earbuds that give you great sound for a small ...	5.0	These earbuds will change the way you look at ...
4	https://www.amazon.co.uk/gp/review/R1AS6K8FBV8...	Great price	5.0	These headphones are a great value for this pr...
...

▼ reformat stars

```
stars=profile_data['ratings'].tolist()
new_stars=[str((str(item))[:-17]) for item in stars]
new_stars
```

[illegible]

```
profile_data['ratings'] = new_stars
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but..		These headphones are pretty damn good for the ...
1	https://www.amazon.co.uk/gp/review/R3M2146VNTD...	Best Price for Great Workout / Running Headphones		These are amazing for £10. AMAZING. They stick...
2	https://www.amazon.co.uk/gp/review/R3U00QINFHP...	Amazed at how good these sound for the price.		They fit very comfortably in my ear and have p...
3	https://www.amazon.co.uk/gp/review/R5BZDAV3JC42E/	Earbuds that give you great sound for a small ...		These earbuds will change the way you look at ...
4	https://www.amazon.co.uk/gp/review/R1AS6K8FBV8...	Great price		These headphones are a great value for this pr...
...
7647	https://www.amazon.co.uk/gp/review/R13ECIUNXJW...	Bootleg width too large		Rockport Women's Jalcia Buckle Tall Heels Boo...
7648	https://www.amazon.co.uk/gp/review/R126KTBSZV2...	Size problems		I order my usual size (36C) and this is much ...
7649	https://www.amazon.co.uk/gp/review/R2KO58OARSW...	Be aware about size differences from garments ...		The garment is OK however one needs to be awar...

▼ most reviews made in a day

```
import datetime
profile_data = pd.read_csv("/content/drive/MyDrive/MSTNG055 Amazon Review-master/final_data.csv")
https://colab.research.google.com/drive/1bRqYQzm8-5s7Af_T0BPriuS-t_2swuQ8#scrollTo=NIRsYGGEuBt-&printMode=true
```

```
profile_data = pd.read_csv('content/guidive/nydrive/MSING055_Amazon_Review-master/final_data')  
profile_id = profile_data['profile_id']  
date = (profile_data['date'])  
profile_data
```

	url	review_bold	ratings	review
0	https://www.amazon.co.uk/gp/review/R18JE8EYCYX...	I rarely use them but..	5	These headphones are pretty damn good

```

i = 1
pid_seq={}

while i < (len(profile_data)):
    print(i)
    if (profile_data['profile_id'][i] == profile_data['profile_id'][i-1]):
        if (profile_data['date'][i] == profile_data['date'][i-1]):
            seq += 1
        else:
            if (profile_data['profile_id'][i] in pid_seq.keys()):
                if (seq>pid_seq[profile_data['profile_id'][i]]):
                    pid_seq[profile_data['profile_id'][i]] = seq
            else:
                pid_seq[profile_data['profile_id'][i]] = seq
            seq=1
        i+=1
    else:
        seq=1
        i+=1

for pid in profile_id:
    if len(pid)> 30:
        profile_id.remove(pid)
    if pid not in pid_seq.keys():
        pid_seq[pid] = 1

```

Streaming output truncated to the last 5000 lines.

2652
 2653
 2654
 2655
 2656
 2657
 2658
 2659
 2660
 2661
 2662
 2663
 2664
 2665
 2666
 2667
 2668
 2669
 2670

2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709

pid_seq

```
{ 'AE26RGNSOWQRIYHGXJDXEVSU6CPA': 2,  
  'AE2FSLQFCDSLVIY6Q574V46BBKTEA': 1,  
  'AE20BNXZ67NBBWYMOKY4PJPUJUIUQ': 1,  
  'AE32RCULY67ZIXMFY5HWFQR4SM7Q': 1,  
  'AE3CTNGABNIKGHCCEMK7S3VDWM7A': 1,  
  'AE3M5HTIVEKQPX3SKDXCGMFT6HWA': 1,  
  'AE3SIYGRZ3YEFYLYQCMNPINWLHFA': 2,  
  'AE3V3CL7B3YHGDUEBUJNLYVDLS2Q': 3,  
  'AE4CQGU7ESRGDMUBQPHCHEFIWHYQ': 3,  
  'AE4F5ME3L607IYCEKNKYVCIF7GCA': 1,  
  'AE4JEHDKV6W2WKJ4DBQVZ4JCVKA': 1,  
  'AE5AAFOEDWPFRV5LDNTXKXVS6SQ': 7,  
  'AE5K3763I3CI3W32LUC2R2V6CE3Q': 1,  
  'AE5OK2L53CJP2VFG5JF62B4A5Z7Q': 9,  
  'AE5PV5LYPIPEK7IAEYF7GOWRQLFQ': 1,
```



```
'AE5S7SZSP5MRBZ22EDLKZWW2BQ7Q': 1,
'AE5SKJZZKXLJYGVND25IG3MSIJCQ': 6,
'AE5T34URKFAGLWEI50QA7RKI605A': 1,
'AE63RKFLGD4G7ZGB6QUECK5KWM6Q': 1,
'AE6CW27SNPAOMEXGVLJU6HKWDSOA': 1,
'AE6RBG4WOBIDNIQ6FFJZRJ2J6AFA': 1,
'AE6SEYKIL3QJVUOAXORY7IIFRPZA': 1,
'AE73WUPMEWCBZQEUGFBETNQFRFDA': 1,
'AE7NXQID7WJDWYL3GHWHCJ5WSBQ': 1,
'AE7RTSGM2RA4DXTZW5RK7QU6JMTQ': 1,
'AEAACDHQSGOR2EADHCFQX6M3ZPTQ': 1,
'AEACAJEHAXWWP55AP5IIMAYWT20Q': 1,
'AEAN2IWQCYQMMDOC4HT7HKR6PQDQ': 1,
'AEBEEUVBQJT435ZVOTTM30BCBKWQ': 1,
'AEBFOB2P6WTTTPNMA75QVJVXNCZ6Q': 1,
'AEBLH43UPSHZMFLOCQUFLAD2U3A': 8,
'AEBQYDMIWSRWTKF53JKJEDKOR26Q': 1,
'AEBVYOHTFWTRBTBZG4BFTZQKI7SQ': 1,
'AEBYLF3H5LYR6KUK6TMTOVQ4EG3A': 1,
'AEBZAT5K5V3VMXX5355BZTLGULJA': 1,
'AECB76FI7FVIWZMGAGEAPFJUNTIA': 1,
'AECIEFFWJUYPKBNIVOMJB7QX65FQ': 1,
'AECOIPXPCC6RWWCUPHYLB5ZUYIZQ': 1,
'AECTTTW5R43TU5LI7EQATTYKZHHQ': 1,
'AED5F4QMMBYMKSY6YGGC2XFLJWQ': 1,
'AED7A6DMQ7Q3YFRWH57X4QYPE5JA': 18,
'AEDBEFYUBFJZIH5A22HDJ3WQSTBA': 1,
'AEDFVXS6VQGIXXRULTS7PCVREFYA': 9,
'AEDKQ6WVHTKI6A6FCOAZ34HV7QCQ': 4,
'AEDKV5UBHTBCFIV43VJWPF74ZZWA': 1,
'AEDNL64E37MBAB7HXND7LVZHQ3LA': 8,
'AEDZVEFBHHV2PZWV7IMV7ZJV2ZTA': 1,
'AEE3JCKNBPY44PGKBHVDDP3536GA': 9,
'AEKALPNG3S7L42ZY6JOM5QXQ4HQ': 1,
'AEEL45A20VFRRABGASWTIIIOAFGDQ': 1,
'AEELFCZCXZTJYFA7HSED7PZ5B4UQ': 1,
'AEESBYJK5UOB4QE04KGF6REQR4A': 1,
'AEF45VY5VB07YNBGXP4E2N4OCHPA': 24,
'AEF6QLRWU4TOPOG2BIUMQNTBYC5Q': 1,
'AEFG3OVK6RW5J5IATODROHNPIIBQ': 19,
'AEFGA7GDPFBF3QZ6QROQG7RDZVQA': 1,
'AEFUCWDGFDNLPCGXAHY04UZ4TQA': 6,
'AEFVMCCBMY5LRMND7GZYRZSNEI5A': 1,
'AEGIMBCSRSEWL5YH5JFOQ75WXVQ': 1,
```

```
len(pid_seq)
pid=[]
count_rev=[]
for k,v in pid_seq.items():
    pid.append(k)
    count_rev.append(v)
df_count = pd.DataFrame(
    {'profile_id': pid,
     'most_rev': count_rev,
    })
df_count.to_csv("rev_aggregate.csv")
```

df_count

	profile_id	most_rev
0	AEMS5UDZTZN4L3BIJV57COUNV2FA	1
1	AG57EELG6RNTZWK7RDHDO53RE4A	9
2	AHZI3QY65DLTDUDWWIPI4SQQBKVA	1
3	AHJ4TATYP5OGXMMFDJ2XZSTKJYDQ	1
4	AHPBJJIJAYS65CAV372NAHS6IZXA	1
...
652	AF6NQFFHVVDDBOKIRBIU2E6J3QILA	1
653	AGLQYA4GZRJRMLEAMMQLX752DD4A	1
654	AFXHJQMHYB4VVAZPNXVKNYAGVDNQ	1
655	AHYUW7NWBjVNRIJFR4RGSTFTGR2A	1
656	AEI27ABNXLFARFVKFJXZQPDS7XQ	1

657 rows × 2 columns

▼ saving the cleaned data to csv

```
import csv
profile_data.to_csv("/content/gdrive/MyDrive/MSING055_Amazon_Review-master/final_dataset/fina
```

```
profile_data['url']
```

```
0      https://www.amazon.co.uk/gp/review/R18JE8EYCYX...
1      https://www.amazon.co.uk/gp/review/R3M2146VNTD...
2      https://www.amazon.co.uk/gp/review/R3U00QINFHP...
3      https://www.amazon.co.uk/gp/review/R5BZDAV3JC42E/
4      https://www.amazon.co.uk/gp/review/R1AS6K8FBV8...
...
7647   https://www.amazon.co.uk/gp/review/R13ECIUNXJW...
7648   https://www.amazon.co.uk/gp/review/R126KTBSZV2...
7649   https://www.amazon.co.uk/gp/review/R2K058OARSW...
7650   https://www.amazon.co.uk/gp/review/R10J3IZKGU5...
7651   https://www.amazon.co.uk/gp/review/R2ICU27DP40...
Name: url, Length: 7652, dtype: object
```

