Facebook Live Sellers Dataset Analysis

Done by, Adyasha Choudhury

CONTENTS

01

Data pre-processing

02

Elbow method– to find best value of ‘k’

03

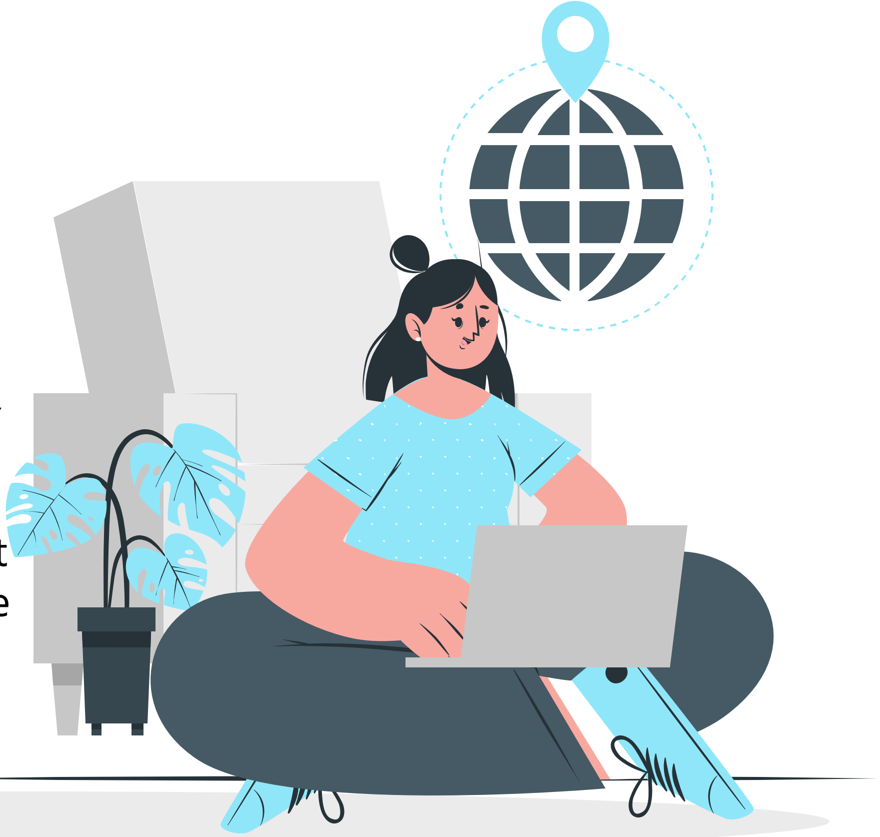
Kmeans Clustering

04

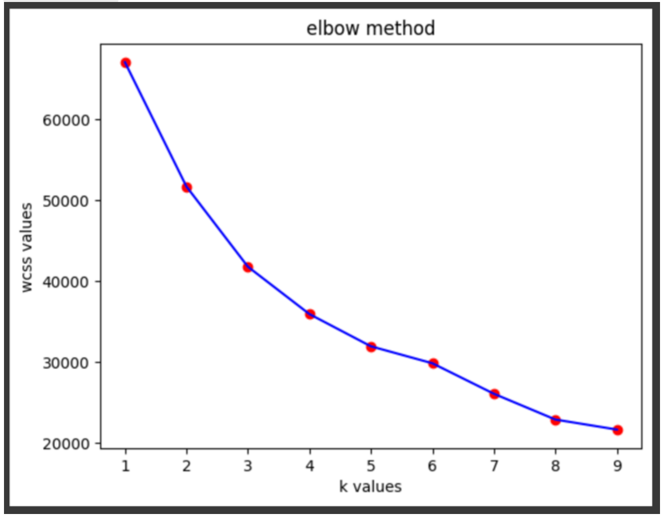
Analysis

1. Data Preprocessing

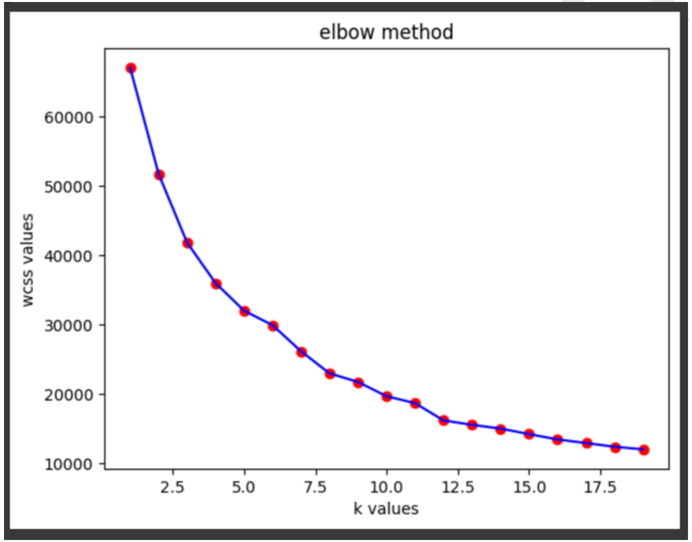
- **Eliminating empty columns** - Eliminated last 4 columns, as all had nan values, and were irrelevant
- **Missing attributes in datapoints -** Checking if any other datapoints have missing attributes. If yes, use imputer.
- **Onehot encoding -** Status type is a string, which is either ‘photo’, ‘video’, ‘link’ or ‘status’ . This column is onehot encoded.
- **Feature scaling** – All the attributes except the first four (onehot encoded) are feature scaled using StandardScaler() object.



2. Elbow Method



This graph plots k values from 1 to 10 against wcss values (within cluster sum of squares)

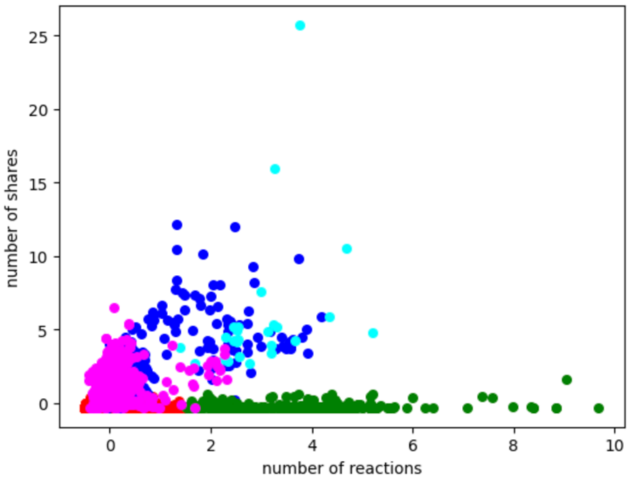


The previous graph couldn't give proper clarity on when does the change in wcss stop reducing significantly with k. Here k = 1 to 20. It is eident that **after k = 5**, wcss stops reducing significantly .

3. Clustering

Number of shares vs
Number of reactions

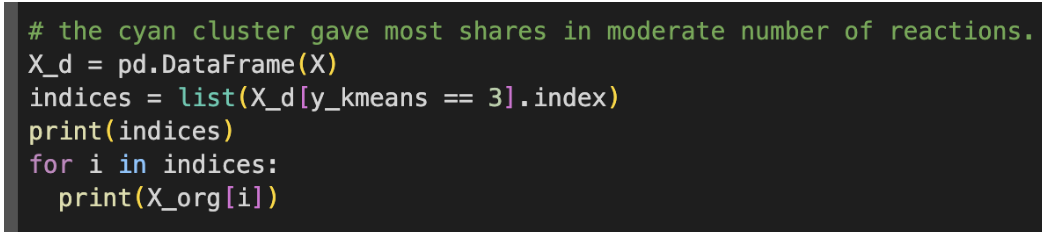
- **Cluster 1 (red) –** less number of reactions and less number of shares (it's hidden behind magenta)
- **Cluster 2 (blue) –** moderate number of reactions and moderate number of shares
- **Cluster 3 (green) –** many reactions but less shares.
- **Cluster 4 (cyan) –** moderate to many reactions and many shares.
- **Cluster 5 (magenta) –** very less reactions but comparatively more shares



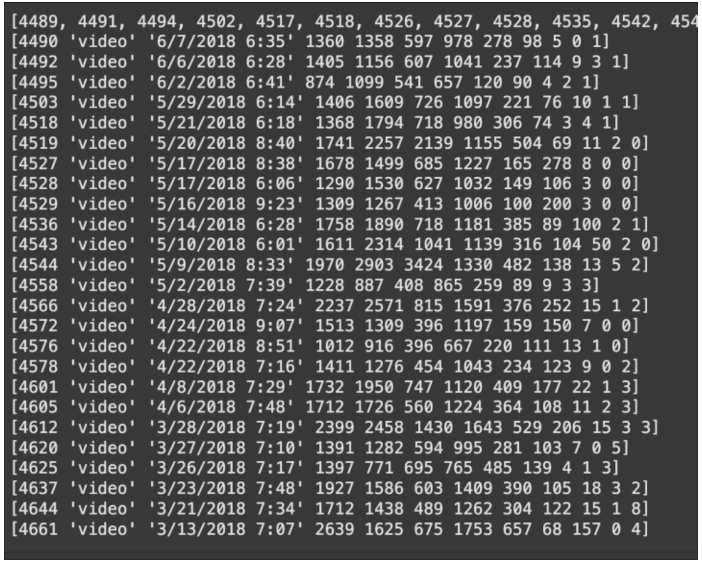
The datapoints in each cluster can be observed to find what type of status or other attributes gives more shares.

4. Analysis

From the scatter plot we saw that the cyan cluster gave us large number of shares with moderate number of reactions. To increase sales, we shall see what datapoints does this cluster consist of -



```
# the cyan cluster gave most shares in moderate number of reactions.
X_d = pd.DataFrame(X)
indices = list(X_d[y_kmeans == 3].index)
print(indices)
for i in indices:
  print(X_org[i])
```



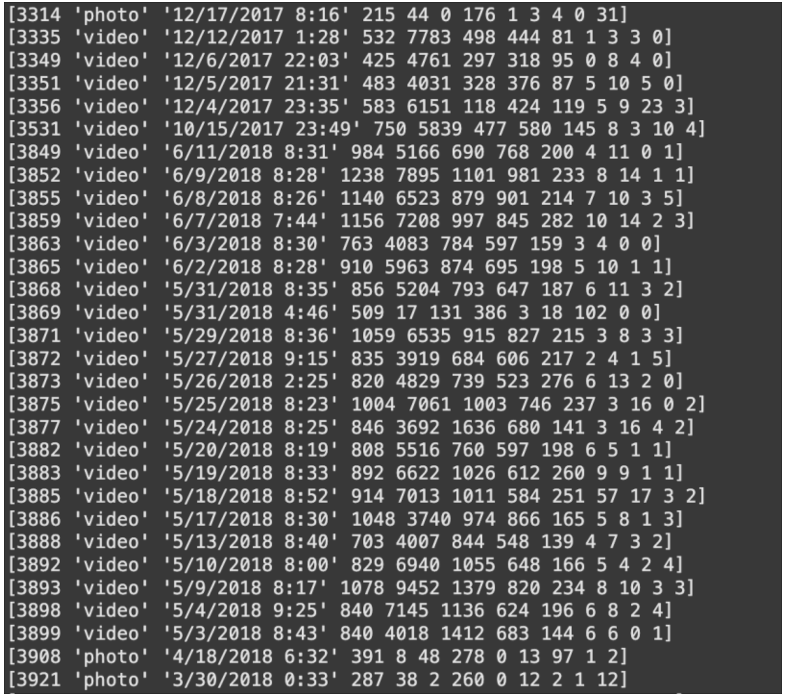
```
[4489, 4491, 4494, 4502, 4517, 4518, 4526, 4527, 4528, 4535, 4542, 454
[4490 'video' '6/7/2018 6:35' 1360 1358 597 978 278 98 5 0 1]
[4492 'video' '6/6/2018 6:28' 1405 1156 607 1041 237 114 9 3 1]
[4495 'video' '6/2/2018 6:41' 874 1099 541 657 120 90 4 2 1]
[4503 'video' '5/29/2018 6:14' 1406 1609 726 1097 221 76 10 1 1]
[4518 'video' '5/21/2018 6:18' 1368 1794 718 980 306 74 3 4 1]
[4519 'video' '5/20/2018 8:40' 1741 2257 2139 1155 504 69 11 2 0]
[4527 'video' '5/17/2018 8:38' 1678 1499 685 1227 165 278 8 0 0]
[4528 'video' '5/17/2018 6:06' 1290 1530 627 1032 149 106 3 0 0]
[4529 'video' '5/16/2018 9:23' 1309 1267 413 1006 100 200 3 0 0]
[4536 'video' '5/14/2018 6:28' 1758 1890 718 1181 385 89 100 2 1]
[4543 'video' '5/10/2018 6:01' 1611 2314 1041 1139 316 104 50 2 0]
[4544 'video' '5/9/2018 8:33' 1970 2903 3424 1330 482 138 13 5 2]
[4558 'video' '5/2/2018 7:39' 1228 887 408 865 259 89 9 3 3]
[4566 'video' '4/28/2018 7:24' 2237 2571 815 1591 376 252 15 1 2]
[4572 'video' '4/24/2018 9:07' 1513 1309 396 1197 159 150 7 0 0]
[4576 'video' '4/22/2018 8:51' 1012 916 396 667 220 111 13 1 0]
[4578 'video' '4/22/2018 7:16' 1411 1276 454 1043 234 123 9 0 2]
[4601 'video' '4/8/2018 7:29' 1732 1950 747 1120 409 177 22 1 3]
[4605 'video' '4/6/2018 7:48' 1712 1726 560 1224 364 108 11 2 3]
[4612 'video' '3/28/2018 7:19' 2399 2458 1430 1643 529 206 15 3 3]
[4620 'video' '3/27/2018 7:10' 1391 1282 594 995 281 103 7 0 5]
[4625 'video' '3/26/2018 7:17' 1397 771 695 765 485 139 4 1 3]
[4637 'video' '3/23/2018 7:48' 1927 1586 603 1409 390 105 18 3 2]
[4644 'video' '3/21/2018 7:34' 1712 1438 489 1262 304 122 15 1 8]
[4661 'video' '3/13/2018 7:07' 2639 1625 675 1753 657 68 157 0 4]
```

As we can see, all the datapoints have one major thing in common – **they are all videos**.

This draws an important conclusion, that videos are engaging for sales and marketing as they not only receive reactions, but also many shares in a good reaction to share ratio.

4. Analysis

From the scatter plot we saw that red cluster had less number of reactions and shares.



```
[3314 'photo' '12/17/2017 8:16' 215 44 0 176 1 3 4 0 31]
[3335 'video' '12/12/2017 1:28' 532 7783 498 444 81 1 3 3 0]
[3349 'video' '12/6/2017 22:03' 425 4761 297 318 95 0 8 4 0]
[3351 'video' '12/5/2017 21:31' 483 4031 328 376 87 5 10 5 0]
[3356 'video' '12/4/2017 23:35' 583 6151 118 424 119 5 9 23 3]
[3531 'video' '10/15/2017 23:49' 750 5839 477 580 145 8 3 10 4]
[3849 'video' '6/11/2018 8:31' 984 5166 690 768 200 4 11 0 1]
[3852 'video' '6/9/2018 8:28' 1238 7895 1101 981 233 8 14 1 1]
[3855 'video' '6/8/2018 8:26' 1140 6523 879 901 214 7 10 3 5]
[3859 'video' '6/7/2018 7:44' 1156 7208 997 845 282 10 14 2 3]
[3863 'video' '6/3/2018 8:30' 763 4083 784 597 159 3 4 0 0]
[3865 'video' '6/2/2018 8:28' 910 5963 874 695 198 5 10 1 1]
[3868 'video' '5/31/2018 8:35' 856 5204 793 647 187 6 11 3 2]
[3869 'video' '5/31/2018 4:46' 509 17 131 386 3 18 102 0 0]
[3871 'video' '5/29/2018 8:36' 1059 6535 915 827 215 3 8 3 3]
[3872 'video' '5/27/2018 9:15' 835 3919 684 606 217 2 4 1 5]
[3873 'video' '5/26/2018 2:25' 820 4829 739 523 276 6 13 2 0]
[3875 'video' '5/25/2018 8:23' 1004 7061 1003 746 237 3 16 0 2]
[3877 'video' '5/24/2018 8:25' 846 3692 1636 680 141 3 16 4 2]
[3882 'video' '5/20/2018 8:19' 808 5516 760 597 198 6 5 1 1]
[3883 'video' '5/19/2018 8:33' 892 6622 1026 612 260 9 9 1 1]
[3885 'video' '5/18/2018 8:52' 914 7013 1011 584 251 57 17 3 2]
[3886 'video' '5/17/2018 8:30' 1048 3740 974 866 165 5 8 1 3]
[3888 'video' '5/13/2018 8:40' 703 4007 844 548 139 4 7 3 2]
[3892 'video' '5/10/2018 8:00' 829 6940 1055 648 166 5 4 2 4]
[3893 'video' '5/9/2018 8:17' 1078 9452 1379 820 234 8 10 3 3]
[3898 'video' '5/4/2018 9:25' 840 7145 1136 624 196 6 8 2 4]
[3899 'video' '5/3/2018 8:43' 840 4018 1412 683 144 6 6 0 1]
[3908 'photo' '4/18/2018 6:32' 391 8 48 278 0 13 97 1 2]
[3921 'photo' '3/30/2018 0:33' 287 38 2 260 0 12 2 1 12]
```

This time, **contradictory to the previous result**, videos received least number of reactions and shares. Upon comparing the previous result with this, it is observed that videos started gaining more attention from the second half of the year 2018. in most of 2017, they received lesser attention.

This could have been due to the following reasons –

- The social media page started gaining followers in 2018.
- The quality of videos improved in 2018.
- Social media strategies improved in 2018.

Apart from these, to get better analysis we can also **analyse only video datapoints** to gain better insights. **Adding parameters like ‘video length’** can help us check if customers get bored of long videos, which is the reason of failure with videos.

Clustering can help us analyse our data and find new patterns. This would help in making sales and marketing more targeted and efficient

Thank you !