**Primary Objective:**

The objective of this case study is to analyse this data and come up with insights, highlighting the good things, bad things, cohorts of patterns found or cohorts of customers you can classify.

**Approaches:**

To segment the creators, I have used RFM scores along with clustering techniques such as K means++ and DB Scan. Although best clusters of k-means++ were not able to separate our best creators from rest so I focused more on DB Scan and the results are satisfying.

**Key Findings**

From Basic Stats:

* People those are here for a long time, are producing more videos. So this platform must be doing good for them.
* Number of GMV per month also increasing so people must be getting benefit.
* Watch per video is increasing on total but the mean and median are not increasing rather its declining which could have adverse effect on the creators since they are increasing their contents and they must be investing more on there quality of production if watch doesn’t increase then it may cause them disappointment. Reason must be viewers now have more options, but we should bring more customers as well.
* Courses seems to be live classes or have access period less than month. If business model doesn’t allow to sell recorded contents, then we should encourage them to produce more frequently.
* Top 20 Content creators must be having big studious and large man powers since the number of contents produced per month are quite high in number. Reason must be obvious it’s a B2B ED Tech company.

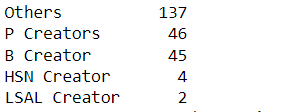
From DB SCAN

* 1 cluster separated by DB scan produces 99.2 % of total GMV and 96.17 % of videos but they are consisting of only 38.5% of total population. They have all the best creators from RFM segmentation. So, I consider them as the most important cluster to be focused on.
* Another cluster consisting 59.82% people don’t contribute to even .7 % of profit. Surprisingly there is one new high earning creator is among them so we can talk to him, and some potential creators to improve their level
* And rest of the 2% Falls in between these two clusters.

From RFM Segmentation

Using RFM scores 5 segments

* (B Creator) Best Creators
* (HSN Creator) High - Earning New Creators
* (LSAL Creator) Lowest-Profitable Active Loyal Creators (At Risk creator)
* (P Creators) Potential creators
* (Cb Creator) Churned Best creators



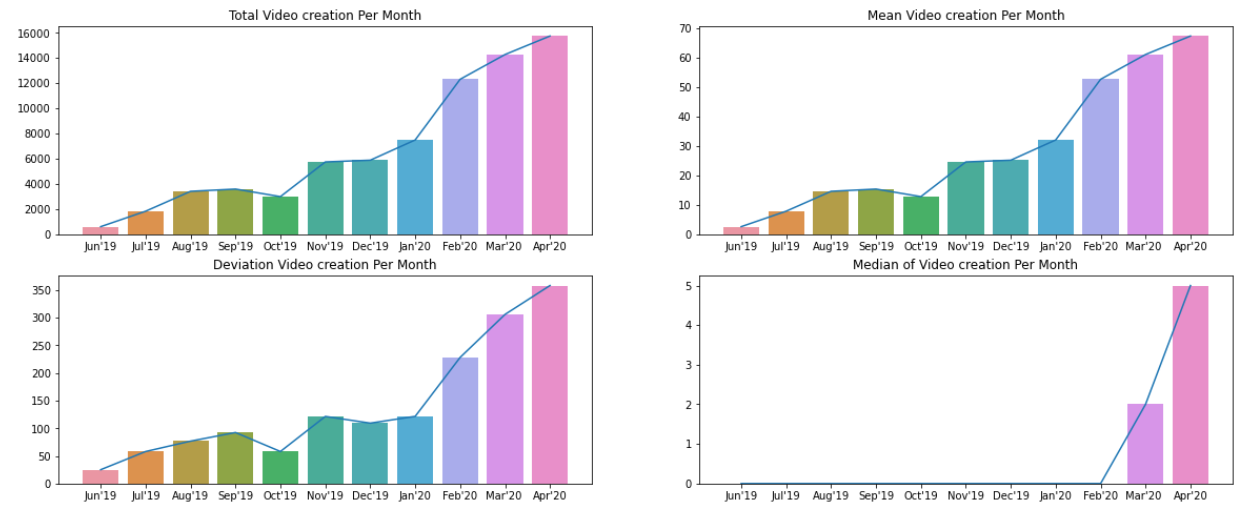
I’ve defined them briefly below. Here 2 LSAL creators with us for a long time they are even in DB scan’s important cluster but there GMV are not high or possibly their price of sell is high. We should communicate with them.

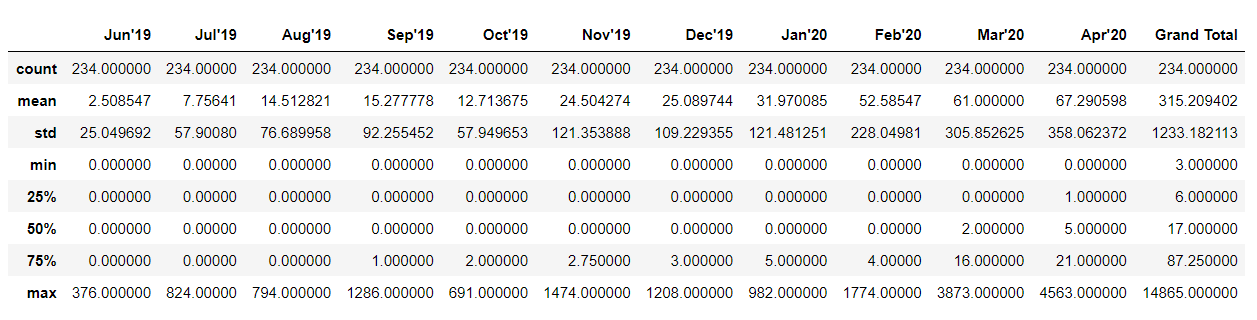
\*Detailed analysis is below or I urge one to have a look into attached jupyter notebook file

**In The data I was given two tables:**

1. Transactions

2. GMV

**Some Basic Stats and Plots of transaction table**

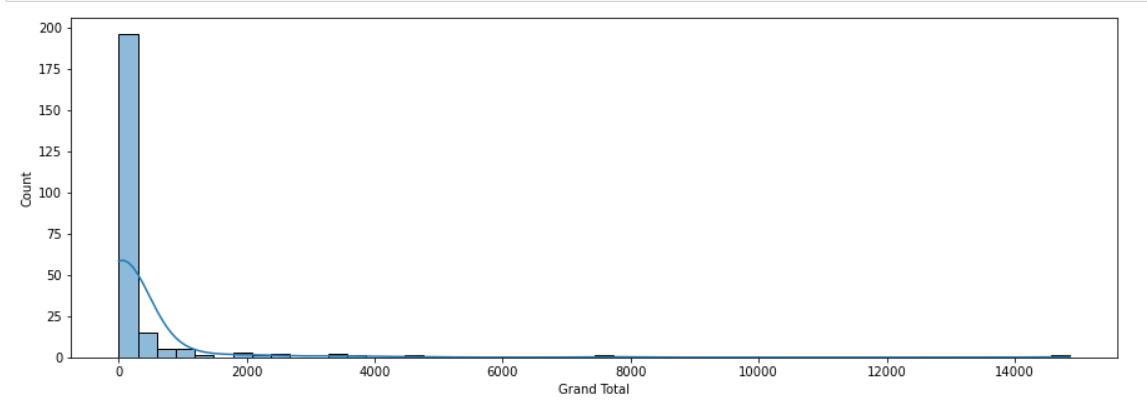


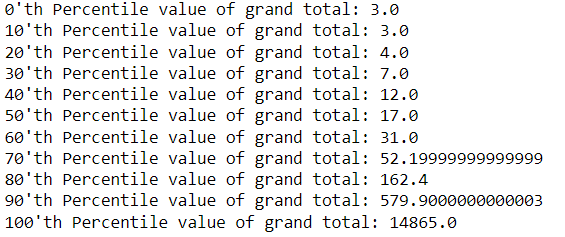
Observation:

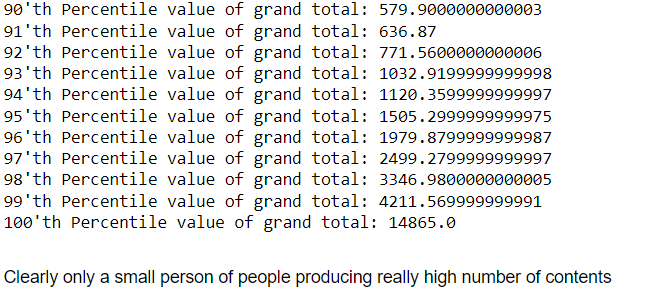
1. From total and mean video creation plot we can say with time the number video creation is increasing Which indicates this platform is doing good for the creators

2. Now from deviation plot since data deviation also increasing it could possibly because of those who are keep increasing their video production capacity than others. And they are more serious about it.

3. And from median video creation plot one interesting thing we notice that most of the creators like more than 75% (from above description table) were not even creating videos before feb20 or they were not even present in the platform.

****

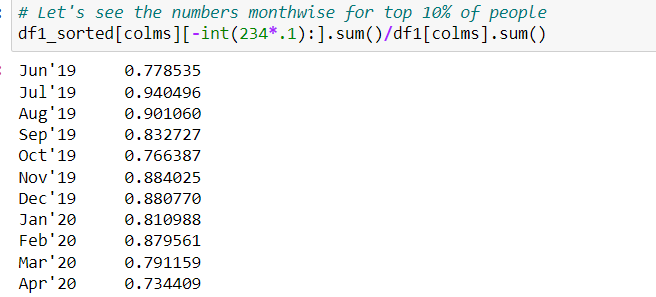
****

****

**Lets see if it has a tendency of following 80-20 principle or power law distribution**

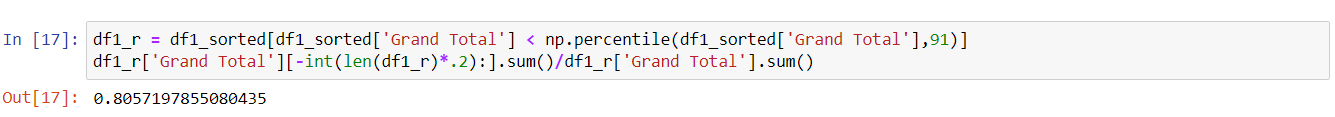
Total percentage of videos produced by top 20% of people - 0.935275695169403

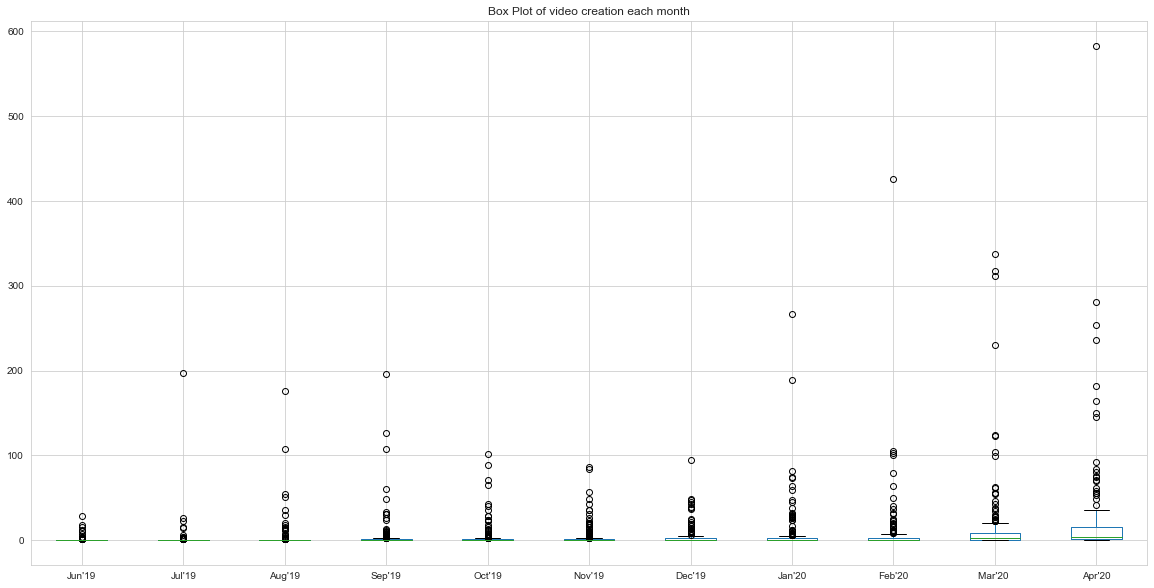
Total percentage of videos produced by top 10% of people - 0.819805040740791

****

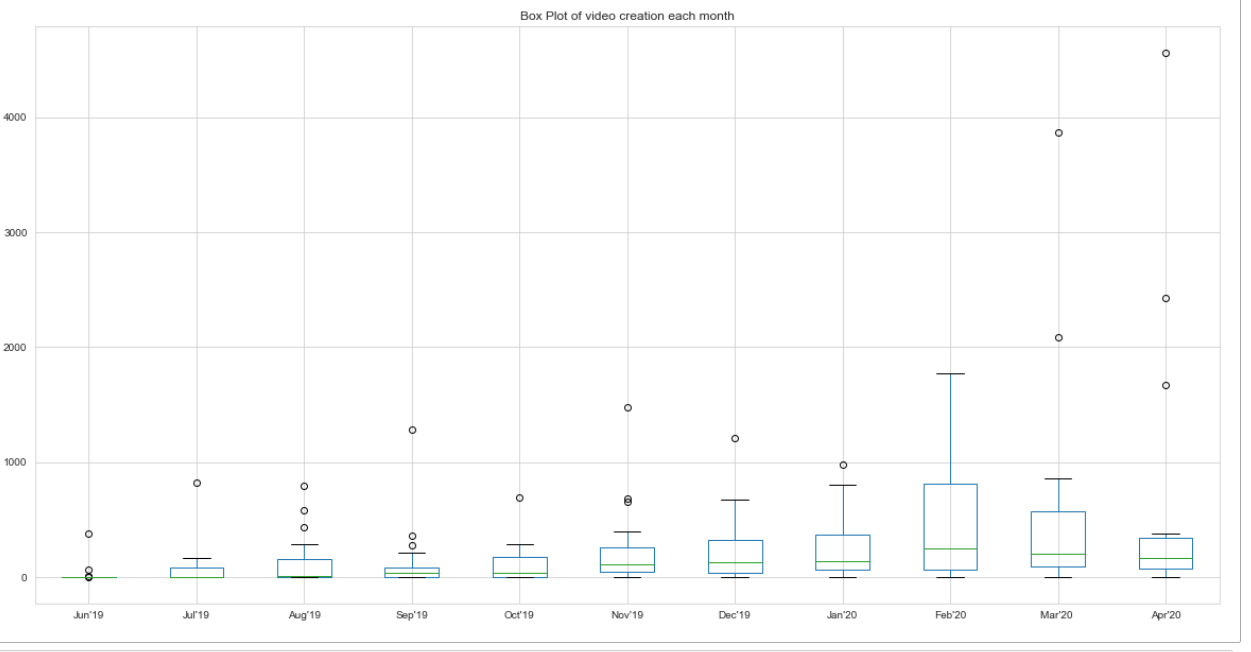
So, I would like to state that the dataset has a tendecy of following 80-20 rule, here instead 20, top 10 percent people are producing almost 80 percent of the videos.

**But if we remove top 9 percentile of creators then distribution very likely follows 80-20 rule**

****

****Lets See the box plot of least 81 % of content creator **.**

Now here I have selected creators having Grand Total less than around 636 but still some month has total videos close to 600 which I want to consider as outliers and can inspect them later.

****Now, the box plot for rest of the top 9%

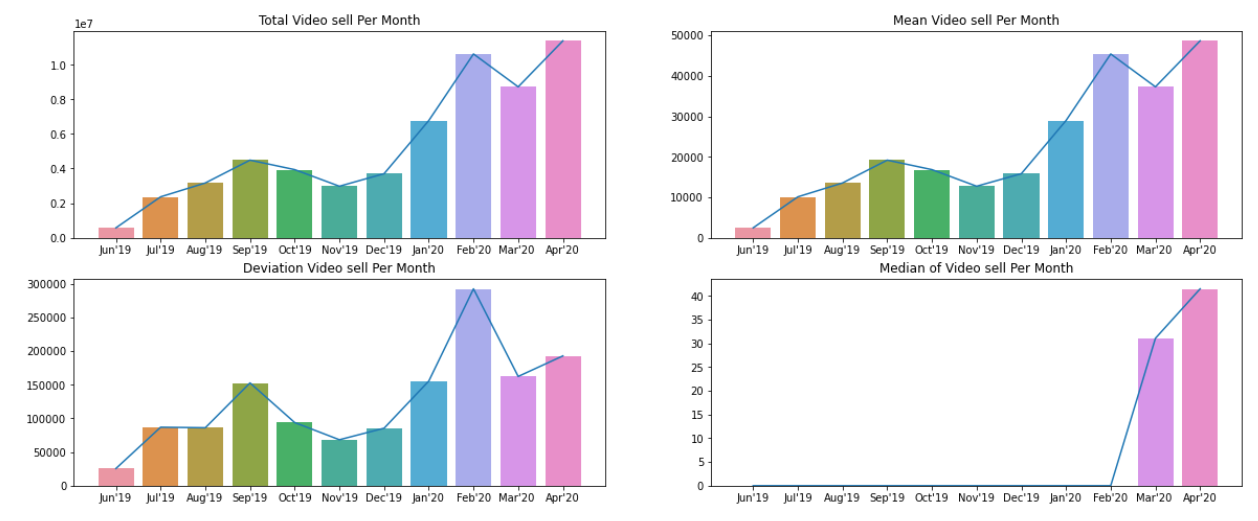
Conclusion:

1. These top 9 percent of creators seems to be here for a longer period of time and with time they seems to be increasing there contents, also the number of contents indicates they are not an individual but studios with a good capacity video production.

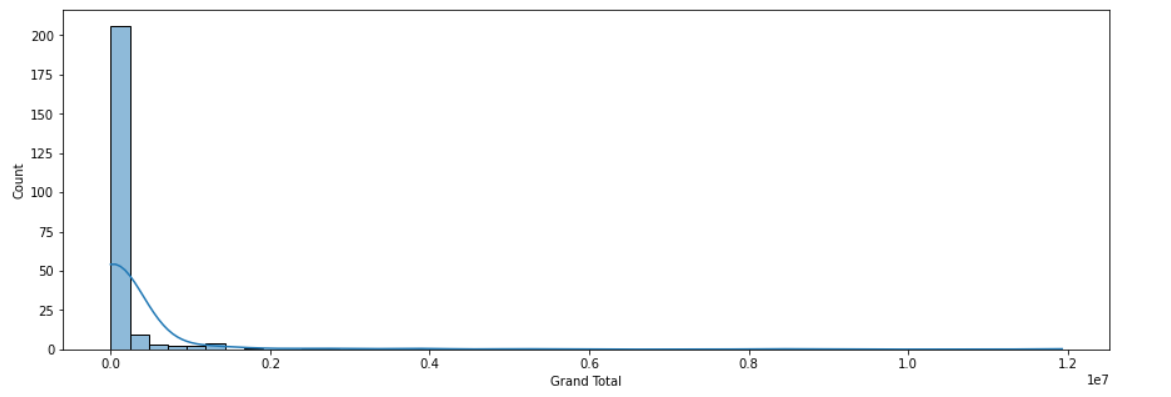
Since they are taking interest for a long period of time that means Classplus platform really working good for them. Which is a good sign.

2. Also they produce almost 80% of total videos, so they should be taken care of.

**Basic Stats for GMV sheet**

****

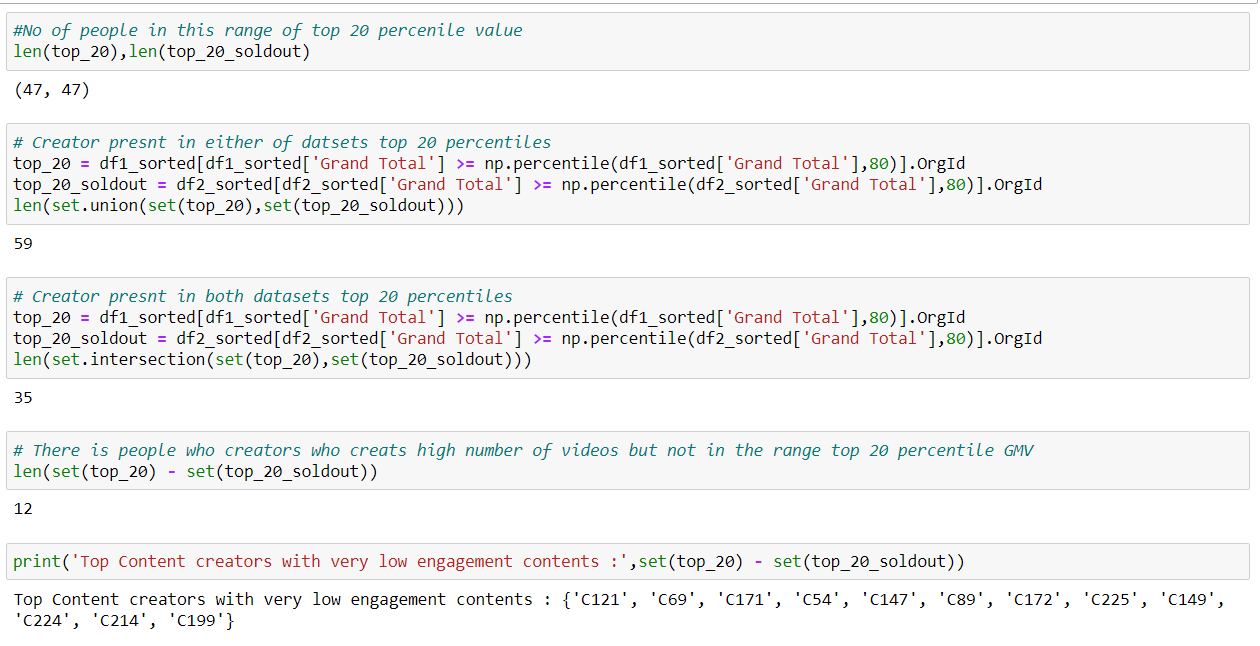
**GMV replicating the behaviours of transaction table which lead me to the conclusion that contents are very much engaging, since when they are making more content people are also sharing same interest, also content creators are keeping good track of their viewers .**

****

Total percentage of videos produced by top 20% of people - 0.975341069913418

Total percentage of videos produced by top 10% of people - 0.9033305604772583

#### Overlap between top 20 creator and top 20 percentile GMV organisation

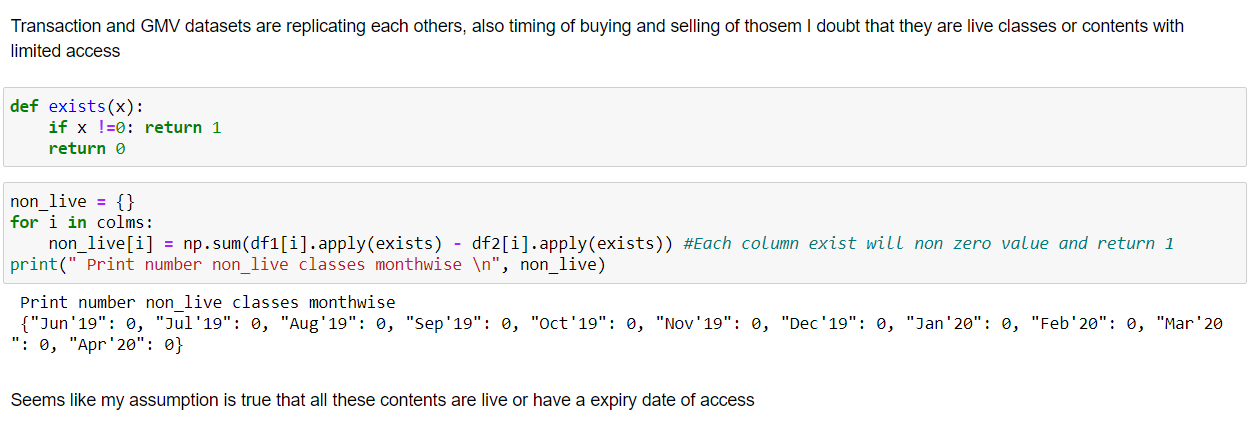


Top Content creators with very low engagement contents : {'C121', 'C69', 'C171', 'C54', 'C147', 'C89', 'C172', 'C225', 'C149', 'C224', 'C214', 'C199'}

These 12 people should review their contents since they are not in the that range where most of the videos are sold.

But They are in the range of most consistent one and if they are still producing gaining profit, then it indicates that they have high price of selling

**Check for non live classes**

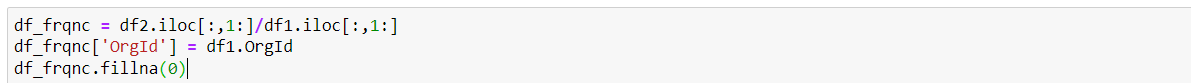
****

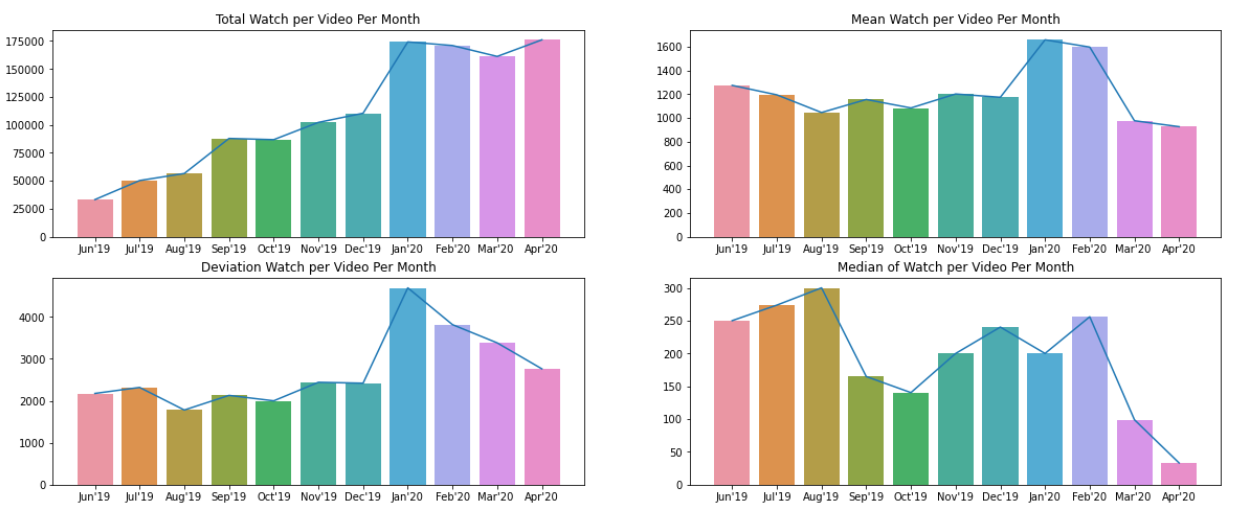
Print number non\_live classes monthwise

{"Jun'19": 0, "Jul'19": 0, "Aug'19": 0, "Sep'19": 0, "Oct'19": 0, "Nov'19": 0, "Dec'19": 0, "Jan'20": 0, "Feb'20": 0, "Mar'20 ": 0, "Apr'20": 0}

Seems like my assumption is true that all these contents are live or have a expiry date of access

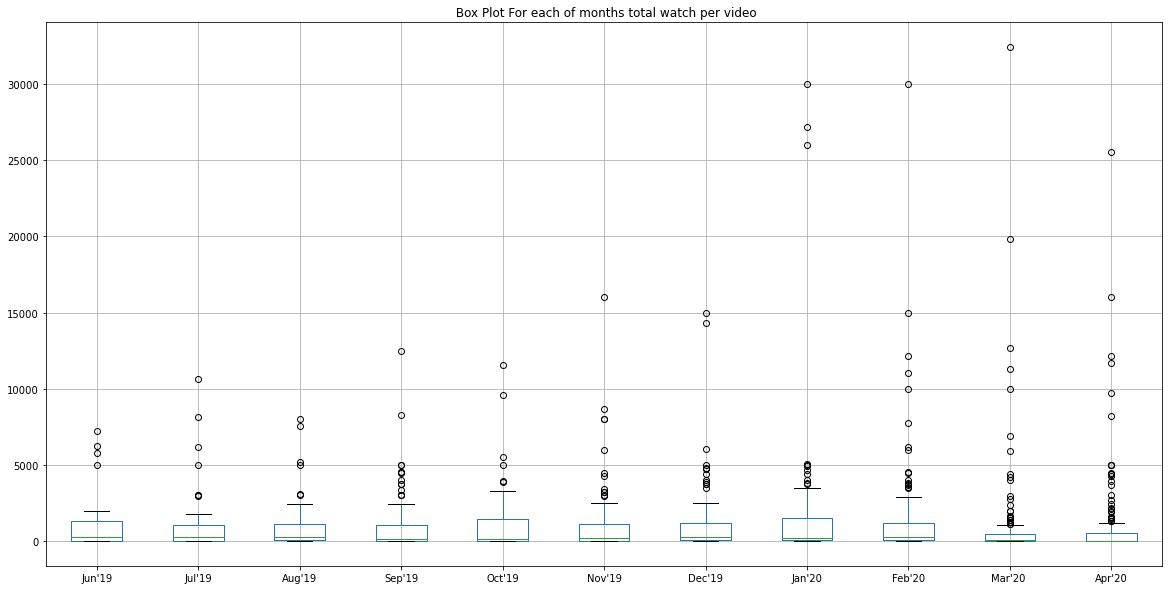
## Watch per video Table

****

****

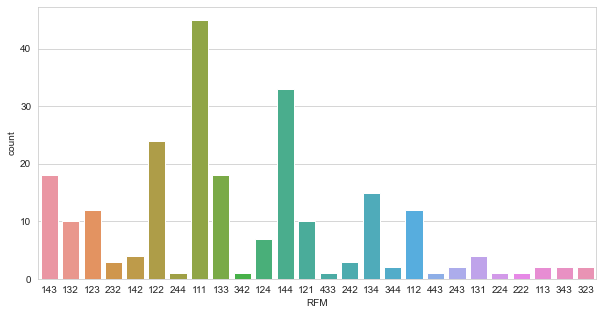
Median and mean for watch per videos are decreasing with time for reason must audience now have more number of options since there is too many courses to buy. They have more options now.

\*\*Also we should focus on bringing more customers or it may have some adverse effects on our content creators

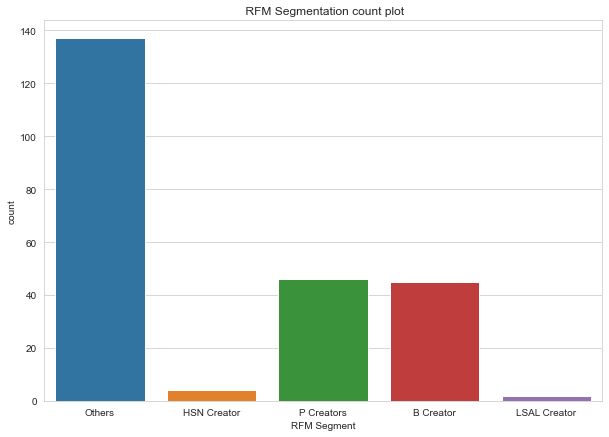
****

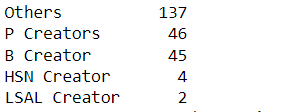
# **RFM Analysis**

**RFM Score’s Count plot**

****

**RFM Segmentation count plot**

****



**Defination of theses segmentaions**

**(B Creator) Best Creators** :– Communications with this group should make them feel valued and appreciated. These creators likely generate a disproportionately high percentage of overall revenues and thus focusing on keeping them happy should be a top priority. Further analysing their individual preferences and affinities will provide additional opportunities for even more personalized messaging.

**RFM SCORE**: 111

**(HSN Creator) High - Earning New Creators** :– It is always a good idea to carefully “incubate” all new creators, but because these new creators spent a lot on their first purchase, it’s even more important. Like with the Best Customers group, it’s important to make them feel valued and appreciated – and to give them terrific incentives to continue interacting with the brand.

**RFM SCORE**: 141, 142

**(LSAL Creator) Lowest-Profitable Active Loyal Creators (At Risk creator)** :– These frequent creators are active and loyal, but they are low soldout. Marketers should create campaigns for this group that make them feel valued, and incentivize them to increase their spend levels. As loyal customers, it often also pays to reward them with special offers if they spread the word about the brand to their friends, e.g., via social networks.

**RFM SCORE**: 113, 114

**(P Creators) Potential creators** :- They are quite promising and high in number. Revenues generated by them is in range 50 - 75 percentile which is not bad, they have good potential of going up in the ladder, communicating and understanding them could be helpful for both of us.

**RFM SCORE**: 112, 121, 122

**(Cb Creator) Churned Best creators** : – These are valuable creators who stopped transacting a long time ago. While it’s often challenging to re-engage churned creators, the high value of these creators makes it worthwhile trying. Like with the Best creators group, it’s important to communicate with them on the basis of their specific preferences, as known from earlier transaction data.

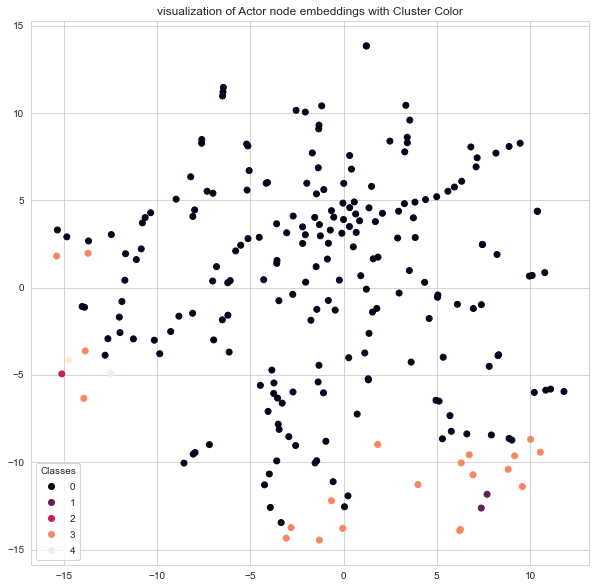
**RFM SCORE**: 411, 412, 421, 422

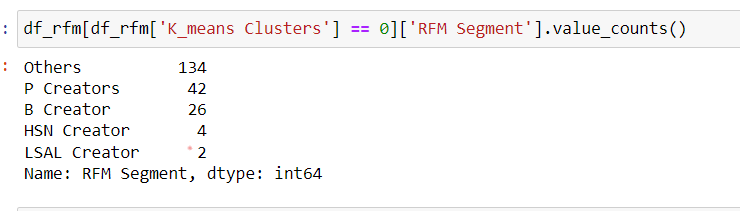
**And Others**: Those who are not among these and considered to avoid complexity and more advanced techniques are used get important creators among them

# Segmentation using unsupervised machine learning like K-Means and DB Scan

## K-Means Clustering

## Using k-means++ the best number of clusters are 5 which is shown in below tsne generated 2D plot

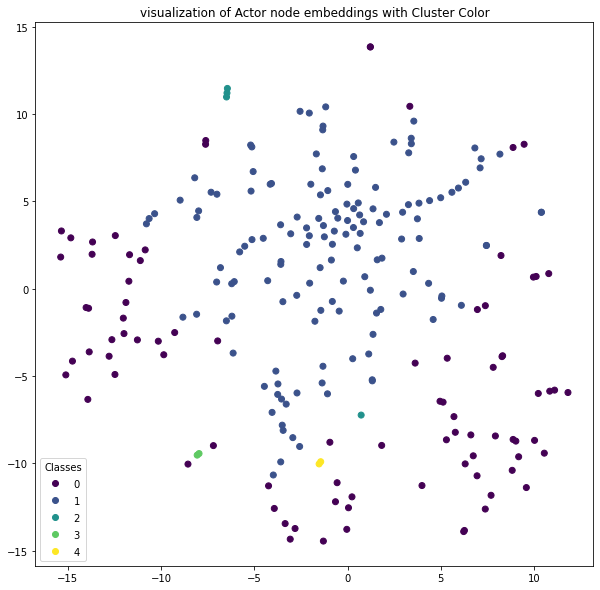




But these clusters are not separating best creators from rest not so important creators reason must be actual clusters are of different sizes.

# **DBSCAN Clustering**

To avoid the problems faced by K-means++ I have used DBSCAN, and results are very much confident.



label\_map : {-1: 0, 0: 1, 1: 2, 2: 3, 3: 4}

**People under label -1 cluster**

B Creator 45

P Creators 24

Others 13

HSN Creator 2

LSAL Creator 2

This group in population is 38.46% and produces 96.163 % contents, and they are making 99.2 % of GMV

### \*\*Seems like these 38 percent of people are the real asset of the company

We can see its able to clusters all the best customer and only 13 others creator are here with our important clients. So possibility is good that they have some good potential even if they are not, believing to this cluster won’t bring any harm.

**People under label 0 cluster**

Others 118

P Creators 21

HSN Creator 1

This group in population is 59.82% and produces 3.75 % contents, and they are making only .613 % of GMV

And these 59.82% of people belonging to cluster 0 are least important creators. They don't even contribute .7% of profit

\*\*Surprisingly there is one new high earning creator among them so we can talk to him, and some potential creators to improve their level.

And rest of 3 clusters consist of the 2% population makes around 1% of profit, falls in between these two clusters. We can consider them as a single segment.