VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

PROJECT REPORT

Data Analysis and Visualization on YOUTUBE

CSE3020

DATA VISUALIZATION

Faculty : Prof. RAJKUMAR R

JUNE 2020

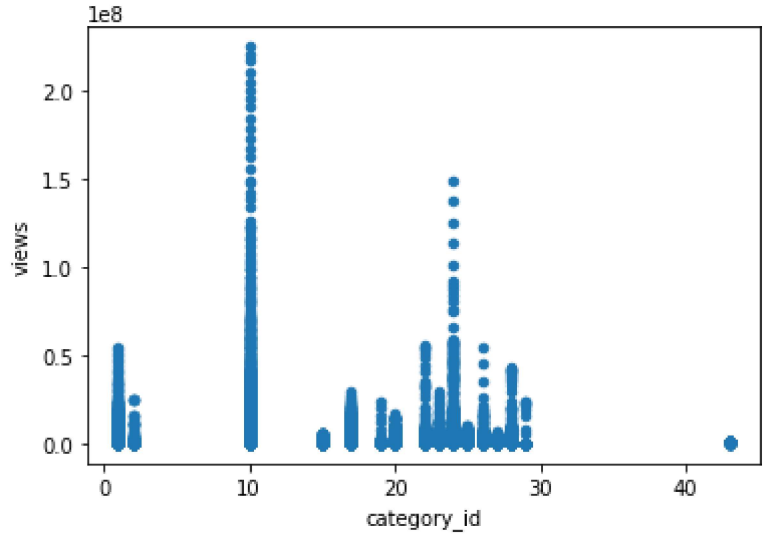
BY – MANISH RAJ

17BCE0447

School of Computer Sciences and Engineering (SCOPE)

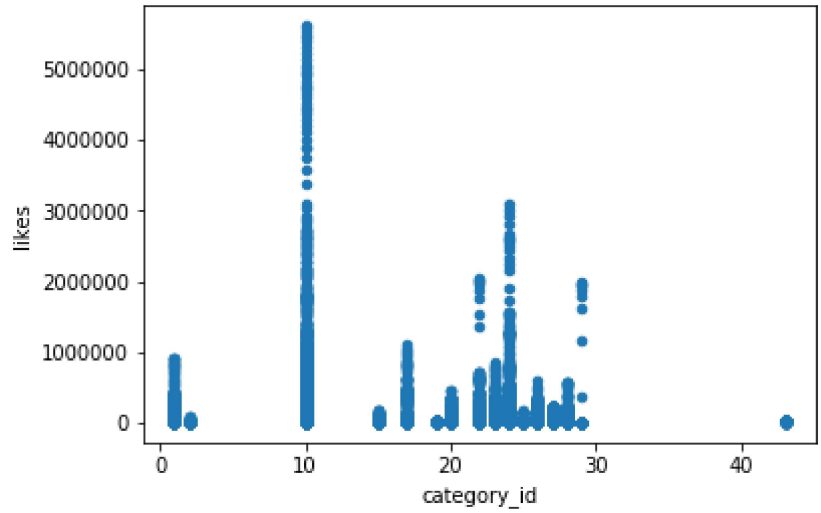
In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv('F:\\r\\USvideos.csv')
df.plot(kind='scatter',x='category_id',y='views')
plt.show()
```



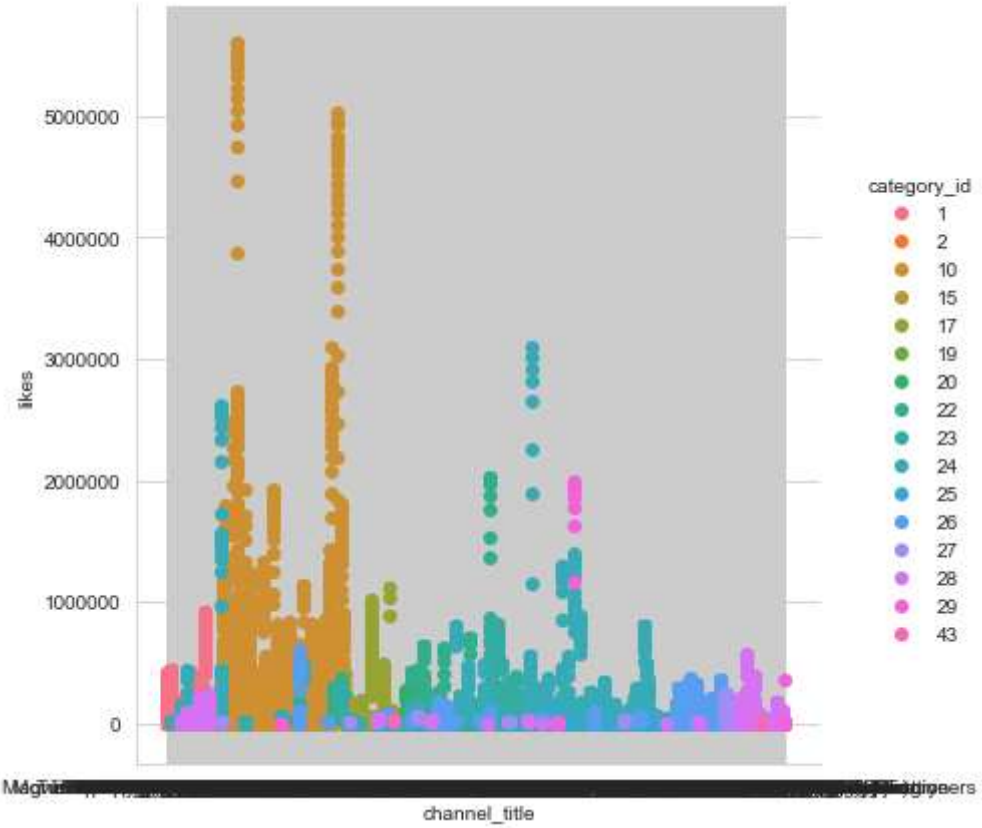
In [34]:
```
#2-D SCATTER PLOT The above plotting is between the category_id that is to which
#category 10 mostly has the highest number of views, the category_id 10 is music
```

In [4]:
```
df.plot(kind='scatter',x='category_id',y='likes')
plt.show()
```



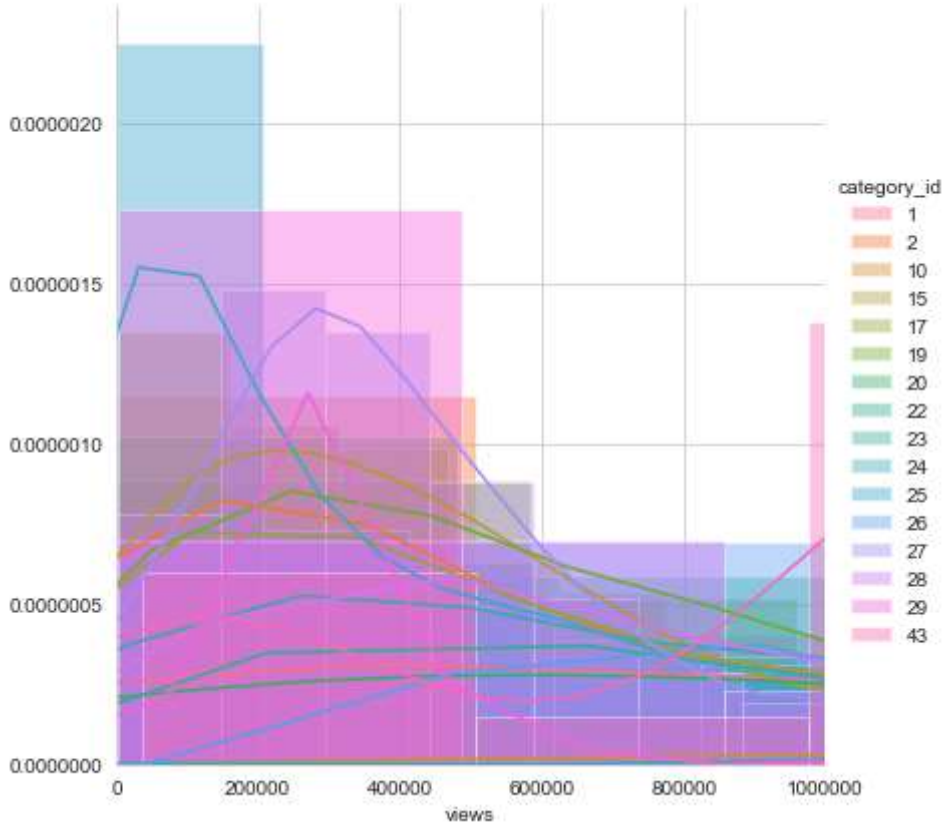
In [35]:
```
#2-D SCATTER PLOT The above plotting is between the category_id that is to which
#category 10 mostly has the highest number of likes, the category_id 10 is music
```

In [25]:
```
sns.set_style("whitegrid")
sns.FacetGrid(df,hue="category_id",height=6)\
   .map(plt.scatter,"channel_title","likes")\
   .add_legend();
plt.show()
```



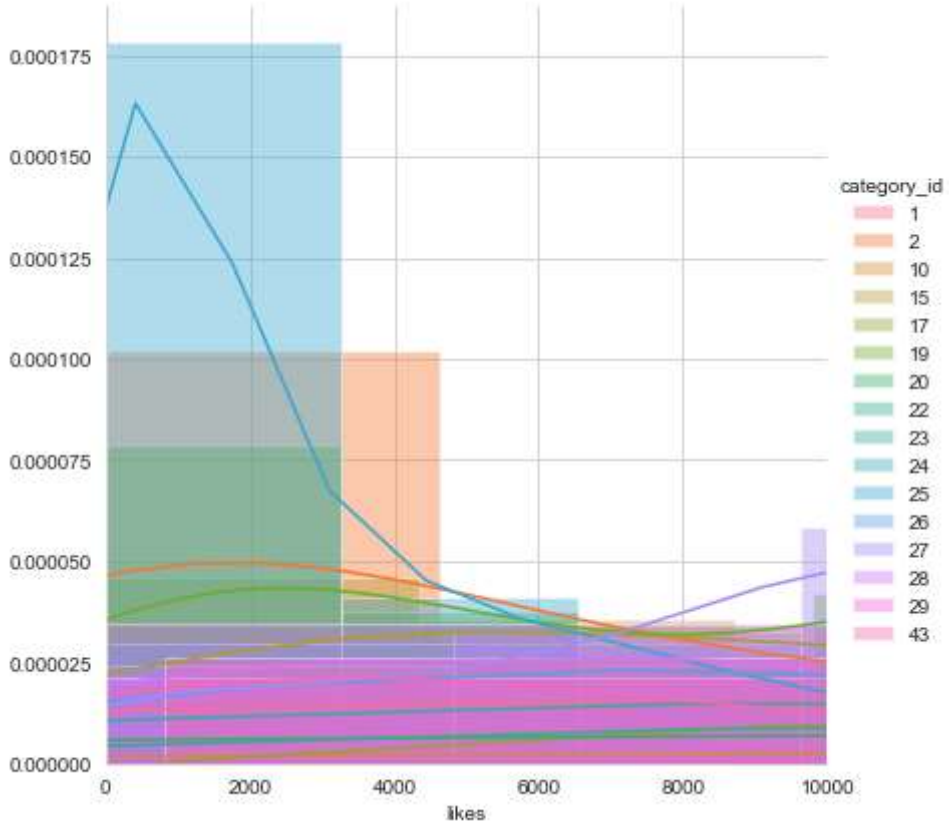
In [7]:
```
#2-D Scatter Plot with color Coding- we can see the different video categories tl
#category 10 which is music video and also the x axis which contains the channel
#music channels recieve the most number of likes
```

In [32]:
```
sns.FacetGrid(df,hue="category_id",height=6)\
   .map(sns.distplot,"views")\
   .add_legend();
plt.xlim(0, 1000000)
plt.ylim(0, None)
plt.show()
```



In [ ]:
```
#histogram or distribution plot - This is a distribution plot of views in the x
#gives us the counts of the various points which is belonging to the number of v
#comedy and entertainment videos which that have views between 0-200000 and then
#of views between 0-500000 and also we can observe that the comedy and entertainm
#number of views of the videos.The line is for the PDF
#height of the histogram tells us how often this paticular number of view occurs
```

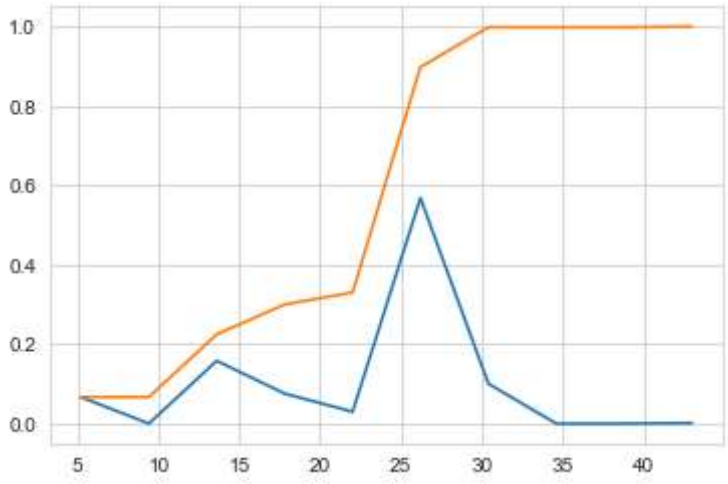
In [33]:
```
sns.FacetGrid(df,hue="category_id",height=6)\
   .map(sns.distplot,"likes")\
   .add_legend();
plt.xlim(0, 10000)
plt.ylim(0, None)
plt.show()
```



In [36]:
```
#histogram or distribution plot - this is a distribution plot of the number of l
#this plot tells us that the count of number of likes for comedy and entertainme
#we can also observe that there are more film videos that are recieving various
#good number of likes in all the grid
```

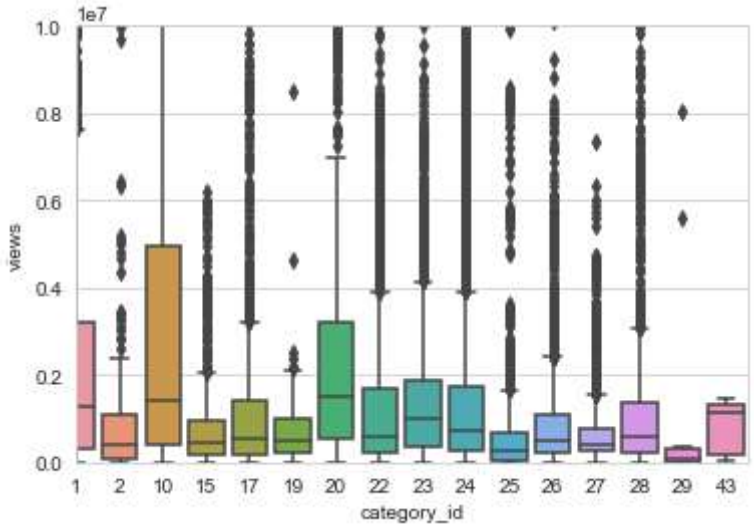
In [40]:
```
counts,bin_edges=np.histogram(df['category_id'],bins=10,density=True)
pdf=counts/(sum(counts))

#compute CDF
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
```

Out[40]:
```
[<matplotlib.lines.Line2D at 0x2d32ed5d780>]
```



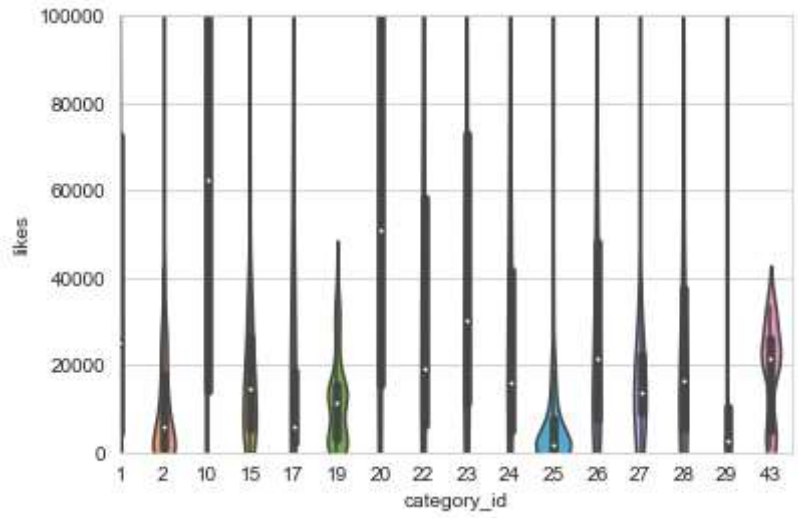
In [42]:
```
#PDF and CDF- The blue line here gives the PDF and the orange line here gives the
#The x axis represents the category_id and the y axis represents the probability
#particular category video. We can observe that the Style category(id=26) occurs
#probability is 58% and from the CDF we can observe that that 99% of the videos
```

In [52]:
```
sns.boxplot(x='category_id',y='views', data=df)
plt.ylim(0, 10000000)
plt.xlim(0, None)
plt.show()
```



In [55]:
```
#Box Plot- The boxplot shows us that for each category of video, how the number o
#category(id=10), the number of views vary from 1000000-5000000 and it also,the m
#25th percentile,the middle one gives us the 50th percentile and the topmost end
#For the music category, there are 25% of videos that have views less than 10000(
```

In [74]:
```
sns.violinplot(x='category_id',y='likes',data=df,size=8)
plt.ylim(0, 100000)
plt.xlim(0, None)
plt.show()
```



In [75]:
```
#violin plot-This violin plot is plotted for the categories and the number of li
#box plot and and the white color inseide it shows the 25th and 75th percentile.
```

In [ ]:
```

```