databricks Data Analysis using spark

Importing and loading Data

```
#df =  spark.read.format("mongo").load()
df =  spark.read.format("mongo").option("uri",
"mongodb://54.197.41.201/final-project.twitter-collection").load()

df2 =  spark.read.format("mongo").option("uri",
"mongodb://54.197.41.201/final-project.youtube-collection").load()
```

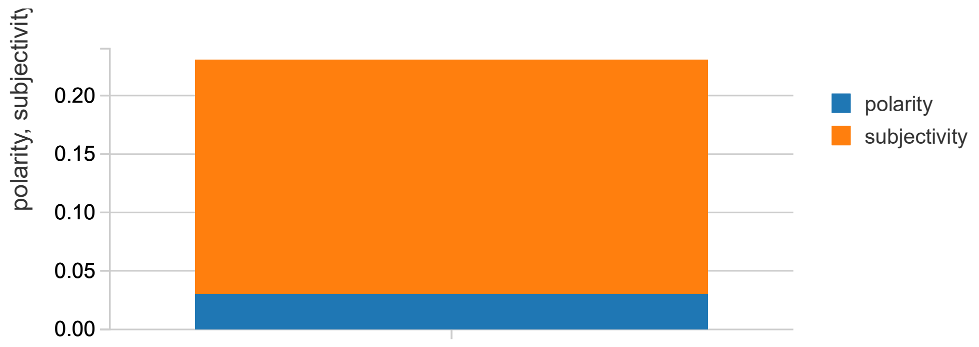
Checking if the data is loaded properly

```
print(df.count())
print(df2.count())
```

```
158557
45
```

Twitter sentiment analysis in graphical form

```
df.show()
display(df.select("subjectivity", "polarity"))
```



Showing sample based on the first 1000 rows.

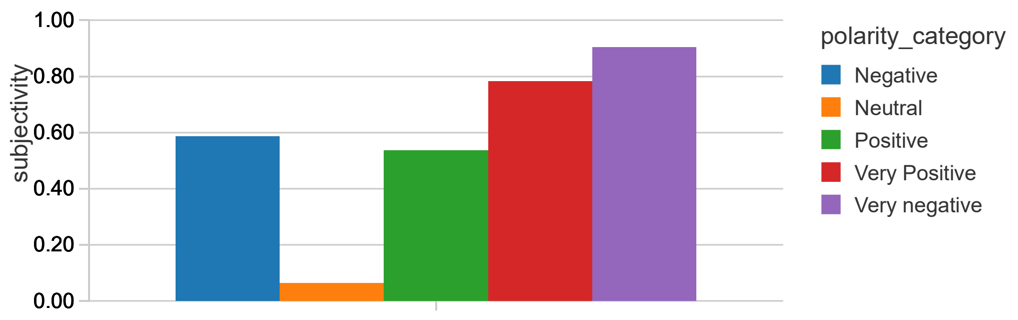
Binning continous variable polarity into categories

```python
from pyspark.sql.functions import udf
from pyspark.sql.types import *
def categorizer(polarity):
  if polarity < -0.6:
    return "Very negative"
  elif polarity < -0.1:
    return "Negative"
  elif polarity <0.1:
    return "Neutral"
  elif polarity <0.6:
    return "Positive"
  else:
    return "Very Positive"

bucket_udf = udf(categorizer, StringType() )
bucketed = df.withColumn("polarity_category", bucket_udf("polarity"))
bucketed.show()
```

```
+--------------------+-------------------+-----------------------------------+---------------------+
---------------+------------------+-------------------------+--------------------+-----------------
+
|                 _id|         created_at|                           location|             polarity|
screen_name|      subjectivity|                     text|           tweet_id|polarity_category|
+--------------------+-------------------+-----------------------------------+---------------------+
---------------+------------------+-------------------------+--------------------+-----------------
+
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                          Argentina|                  0.0|
SillviaBak|               0.0|     A este paso no va...|1250191033234849792|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                               null|                  0.0|
villaverdeh|               0.0|     RT @CiroGomezL: #...|1250191033704632321|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                        Philippines|                 0.35|
skipDLC|              0.65|     RT @iamkarendavil...|1250191033624748032|         Positive|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|東京都港区。竹島は日本国領土です。|                  0.0|
nippon1_|               0.0|RT @nikkei: 中小企業の...|1250191034144952322|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                               DM04|                  0.0|
ai6yrham|               0.0|                  #Hawaii|1250191034241318913|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                               null|                  0.3|
madamshome|               0.9|     RT @marxdeane: Br...|1250191034258259969|         Positive|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                  San Borja, Lima, ...|                  0.0|
DavidPbPrz10|               0.0|     RT @sigridbazan: ...|1250191034446802944|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                          Sri Lanka|                  0.0|
sumanebot|               0.0|     RT @easwaranrutna...|1250191034539220993|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                         Permeating|-0.075000000000000001|
NZSilentRage|               0.4|     Empty hospitals

...|1250191034631389184|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                     Ankara, Türkiye|                  0.0|
kilincbahadir|               0.0|     RT @CemilSahince:...|1250191034698682378|          Neutral|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                               null| -0.6999999999999998|
PleaseThink1776|0.6666666666666666|     RT @PrincessBrava...|1250191035122081794|    Very negativ
e|
|[5e963ac51e0ecc6d...|2020-04-14 22:35:44|                               null|                  0.0|
nahonak|               0.0|     RT @Drmartyufml: ...|1250191035243905026|          Neutral|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:44|                               null|-0.48333333333333334|
okaymarnie|0.7166666666666667|     RT @tanyaboozary:...|1250191035277459463|         Negative|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:44|                           Scotland|                  0.4|
hoppinghaggis|               1.0|     RT @JamieWoodhous...|1250191035294154760|         Positive|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:44|              Sunny Southern ...|                 0.0|
SDSUgrad1983|               0.0|     Hey”Dude” #Hydrox...|1250191035113697280|          Neutral|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:44|                            null| 0.05000000000000002|
fernandovd21|               0.5|     RT @tumbaburross:...|1250191035688349697|          Neutral|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:45|              味噌ラーメン不毛の地|                 0.0|
gomadanngou|               0.0|     RT @sukimaweb: WH...|1250191035893874689|          Neutral|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:45|                           Egypt|              0.1875|
YElkabaniBPharm|            0.1875|     RT @mgranovetter:...|1250191036019806216|         Positiv
e|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:45|                            null|-0.16666666666666666|
BobBack15|0.5666666666666668|     RT @monica_sassy:...|1250191035965333504|         Negative|
|[5e963ac61e0ecc6d...|2020-04-14 22:35:45|                            null|                -0.8|
SandDollar04|               1.0|     RT @SandraBundy: ...|1250191036036583424|    Very negative|
+--------------------+-------------------+--------------------------------+--------------------+
---------------+------------------+------------------------+-------------------+-----------------
+
only showing top 20 rows
```

Subjectivity score for each polarity category

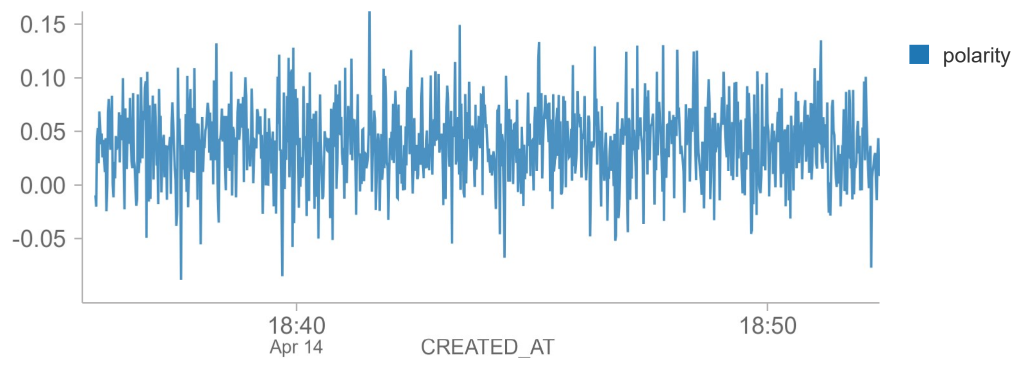
```
display(bucketed.select("polarity_category","subjectivity"))
```



Aggregated (by avg) in the backend.

Checking variation in polarity overtime

```
display(bucketed.select("created_at","polarity"))
```



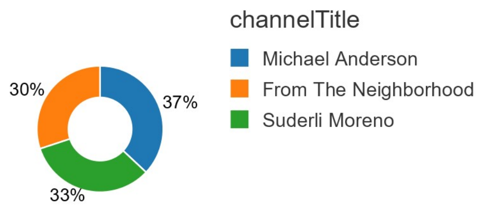
Aggregated (by avg) in the backend.
Showing the first 1000 rows.

Casting the type to float (str to float)

```
yt_df = df2.withColumn("viewCount",df2["viewCount"].cast('float'))
yt_df = yt_df.withColumn("likeCount",yt_df["likeCount"].cast('float'))
yt_df = yt_df.withColumn("dislikeCount",yt_df["dislikeCount"].cast('float'))
yt_df = yt_df.withColumn("commentCount",yt_df["commentCount"].cast('float'))
```

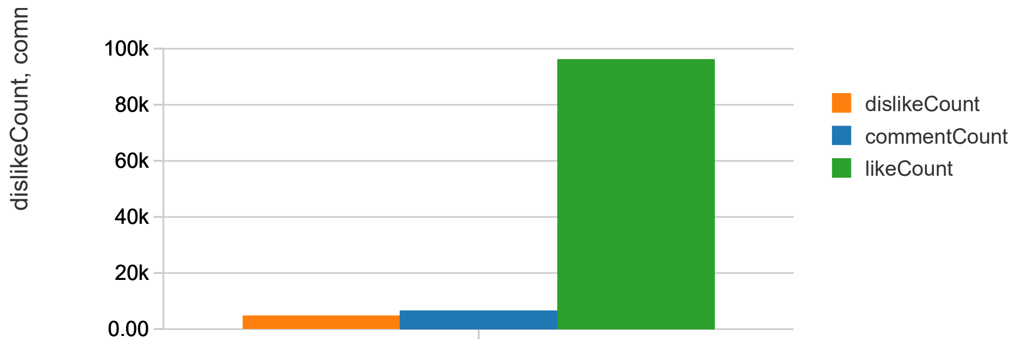
Average of view count based on channel title

```
yt_grouped = yt_df.select("viewCount","channelTitle").groupBy("channelTitle").avg().limit(3)
display(yt_grouped)
```



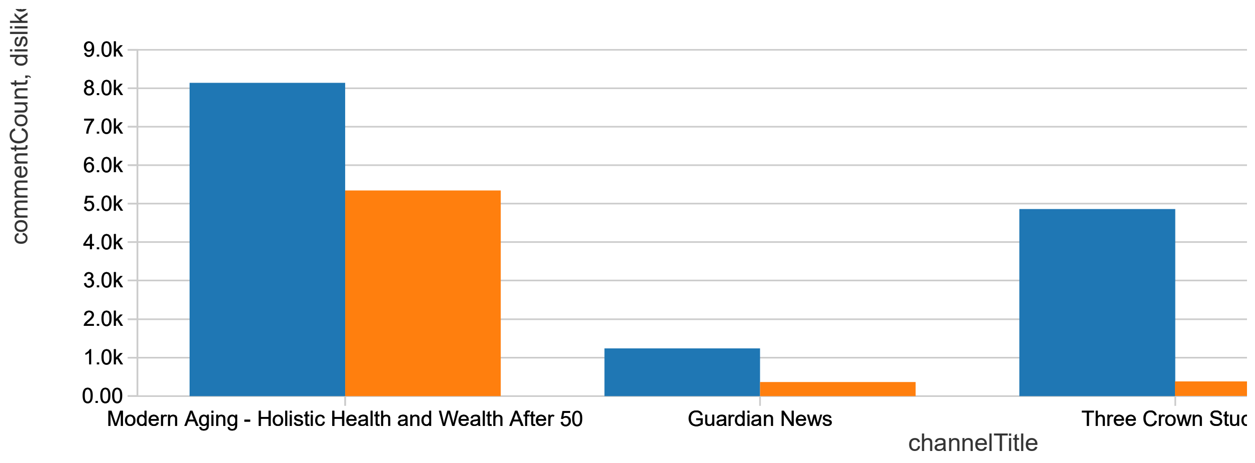
Comparing like count, dislike count, view count and commen

```
yt_ld_ratio = yt_df.select("likeCount", "dislikeCount","commentCount", "viewCount")
display(yt_ld_ratio)
```



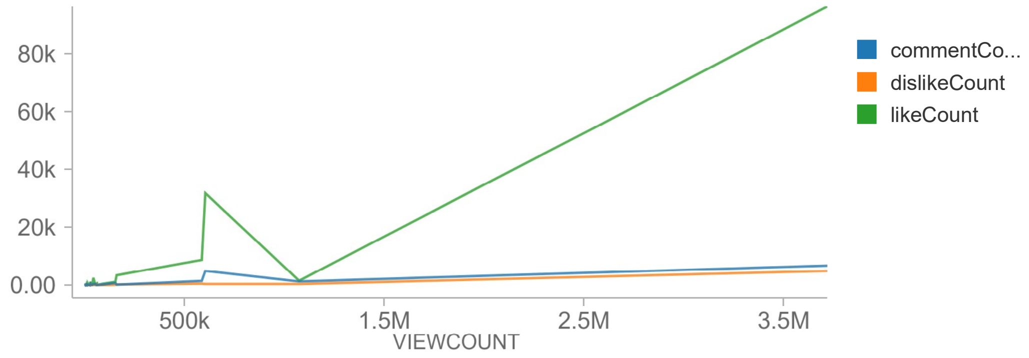
Comparing dislike count and comment count for youtube ch

```
views_likes_comments = yt_df.select("likeCount", "viewCount",
"commentCount","dislikeCount","channelTitle")
vlc_limited = views_likes_comments.limit(5)
display(vlc_limited)
```



Checking the growth of like, dislikes and comments with res

```
display(views_likes_comments)
```



Applying linear regression to predict "Likes"

```
Coefficients: [0.0143575858110491,5.49635697215436]
Intercept: -1084.6353744766473
RMSE: 535.135422
r2: 0.991494
+-------------------+---------+--------------+
|         prediction|likeCount|      features|
+-------------------+---------+--------------+
| 265.77877294633345|    499.0| [4377.0,53.0]|
|-7.3924354398246805|     51.0| [5616.0,17.0]|
| -104.1540122977456|     60.0|  [5762.0,4.0]|
|   571.276383110188|     49.0| [6224.0,96.0]|
|  648.9613042216656|    278.0|[6771.0,107.0]|
+-------------------+---------+--------------+
only showing top 5 rows

R Squared (R2) on test data = 0.420769
```