Data Understanding - Stats

Setup

```
import pandas as pd
import numpy as np
```

Load the data

```
df = pd.read_csv("/content/Spotify_Youtube_Sample.csv")
df.head()
```

	Artist	Track	Album	Album_type	Views	Likes	Comments	License
0	Gorillaz	Feel Good Inc.	Demon Days	album	693555221.0	6220896.0	169907.0	Tru
1	Gorillaz	Rhinestone Eyes	Plastic Beach	album	72011645.0	1079128.0	31003.0	Tru
2	Gorillaz	New Gold (feat. Tame Impala and Bootie Brown)	New Gold (feat. Tame Impala and Bootie Brown)	single	8435055.0	282142.0	7399.0	Tru
3	Gorillaz	On Melancholy Hill	Plastic Beach	album	211754952.0	1788577.0	55229.0	Tru
4	Gorillaz	Clint Eastwood	Gorillaz	album	618480958.0	6197318.0	155930.0	Tru

General Idea

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20718 entries, 0 to 20717
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Artist          20718 non-null  object 
 1   Track           20718 non-null  object 
 2   Album           20718 non-null  object 
 3   Album_type      20718 non-null  object 
 4   Views           20248 non-null  float64
 5   Likes           20177 non-null  float64
 6   Comments        20149 non-null  float64
 7   Licensed        20248 non-null  object 
 8   official_video  20248 non-null  object 
 9   Stream          20142 non-null  float64
dtypes: float64(4), object(6)
memory usage: 1.6+ MB
```

```
df.describe()
```

	Views	Likes	Comments	Stream
count	2.024800e+04	2.017700e+04	2.014900e+04	2.014200e+04
mean	9.393782e+07	6.633411e+05	2.751899e+04	1.359422e+08
std	2.746443e+08	1.789324e+06	1.932347e+05	2.441321e+08
min	0.000000e+00	0.000000e+00	0.000000e+00	6.574000e+03
25%	1.826002e+06	2.158100e+04	5.090000e+02	1.767486e+07
50%	1.450110e+07	1.244810e+05	3.277000e+03	4.968298e+07
75%	7.039975e+07	5.221480e+05	1.436000e+04	1.383581e+08
max	8.079649e+09	5.078865e+07	1.608314e+07	3.386520e+09

No-Numerical Attributes

```
df['Artist'].value_counts()
```

```
Gorillaz                 10
Die drei !!!             10
Hollywood Undead         10
Empire of the Sun        10
White Noise for Babies   10
                         ..
NewJeans                  6
Alfonso Herrera           6
Jimin                     3
Stars Music Chile         1
Bootie Brown              1
Name: Artist, Length: 2079, dtype: int64
```

```
df['Artist'].unique()
```

```
array(['Gorillaz', 'Red Hot Chili Peppers', '50 Cent', ..., 'LE SSERAFIM',
       'ThxSoMch', 'SICK LEGEND'], dtype=object)
```

```
df['Artist'].nunique()
```

```
2079
```

```
nonnumericalcols = ['Artist', 'Track', 'Album', 'Album_type', 'Licensed', 'officia
df[nonnumericalcols].nunique()
```

```
Artist             2079
Track             17841
Album             11937
Album_type            3
Licensed              2
official_video        2
dtype: int64
```

Categorical Attributes

```
album_type = pd.DataFrame({'Album_type' : df['Album_type'].value_counts()})
album_type
```

	Album_type
album	14926
single	5004
compilation	788

```
licensed = pd.DataFrame({'Licensed' : df['Licensed'].value_counts()})
licensed
```

	Licensed
True	14140
False	6108

```
official_video = pd.DataFrame({'official_video' : df['official_video'].value_count
official_video
```

	official_video
True	15723
False	4525

Numerical Attributes

Central Tendency

> min, max, median, mode, midrange

```
col = 'Views'
min = df[col].min()
max = df[col].max()
median = df[col].median()
mode = df[col].mode()[0]
midrange = (max - min)/2
print('col:',col,
      '\n\tmin:', min,
      'max:',max,
      'median:', median,
      'mode:', mode,
      'midrange:', midrange)
```

```
col: Views
        min: 0.0 max: 8079649362.0 median: 14501095.0 mode: 6639.0 midrange:
```

```
def getCentralTendency(col):
    min = df[col].min()
    max = df[col].max()
    median = df[col].median()
    mode = df[col].mode()[0]
    midrange = (max - min)/2
    print('col:',col,
      '\n\tmin:', min,
      'max:',max,
      'median:', median,
      'mode:', mode,
      'midrange:', midrange)

numericalcols = ['Views', 'Likes', 'Comments', 'Stream']

for col in numericalcols:
    getCentralTendency(col)
```

```
col: Views
        min: 0.0 max: 8079649362.0 median: 14501095.0 mode: 6639.0 midrange:
col: Likes
        min: 0.0 max: 50788652.0 median: 124481.0 mode: 0.0 midrange: 2539432
col: Comments
        min: 0.0 max: 16083138.0 median: 3277.0 mode: 0.0 midrange: 8041569.0
col: Stream
        min: 6574.0 max: 3386520288.0 median: 49682981.5 mode: 169769959.0 mi
```

Dispersion

> range, quantiles, var, std

```
col = 'Views'
range = df[col].max() - df[col].min()
quantiles = df[col].quantile([0.25, 0.5, 0.75])
IQR = quantiles[0.75] - quantiles[0.25]
var = df[col].var()
std = df[col].std()

print('col:',col,
      '\n\trange:', range,
      'Q1:',quantiles[0.25],
      'Q2:', quantiles[0.5],
      'Q3:', quantiles[0.75],
      'IQR:', IQR,
      'var:', var,
      'std:', std)
```

```
col: Views
        range: 8079649362.0 Q1: 1826001.5 Q2: 14501095.0 Q3: 70399749.0 IQR:
```

```
def getDispersion(col):
    range = df[col].max() - df[col].min()
    quantiles = df[col].quantile([0.25, 0.5, 0.75])
    IQR = quantiles[0.75] - quantiles[0.25]
    var = df[col].var()
    std = df[col].std()
    print('col:',col,
      '\n\trange:', range,
      'Q1:',quantiles[0.25],
      'Q2:', quantiles[0.5],
      'Q3:', quantiles[0.75],
      'IQR:', IQR,
      'var:', var,
      'std:', std)
numericalcols = ['Views', 'Likes', 'Comments', 'Stream']

for col in numericalcols:
    getDispersion(col)
```

```
col: Views
        range: 8079649362.0 Q1: 1826001.5 Q2: 14501095.0 Q3: 70399749.0 IQR:
col: Likes
        range: 50788652.0 Q1: 21581.0 Q2: 124481.0 Q3: 522148.0 IQR: 500567.0
col: Comments
        range: 16083138.0 Q1: 509.0 Q2: 3277.0 Q3: 14360.0 IQR: 13851.0 var:
col: Stream
        range: 3386513714.0 Q1: 17674864.25 Q2: 49682981.5 Q3: 138358065.25 I
```

Correlation

```
df[numericalcols].corr()
```

	Views	Likes	Comments	Stream
Views	1.000000	0.891101	0.431185	0.601905
Likes	0.891101	1.000000	0.631670	0.654247
Comments	0.431185	0.631670	1.000000	0.267737
Stream	0.601905	0.654247	0.267737	1.000000

Colab paid products - Cancel contracts here