Data Science Final Portfolio

The Data Science Life Cycle

By: Charlotte Malbouef

- Question -

According to the amount of hit songs and total number of streams on Spotify, who is currently the most listened to artist on Spotify?

Why I chose this question: I listen to spotify almost on a daily basis. Listening to different singers and the kinds of art they are able to produce is fascinating. At first, I was genuinely curious on what the most popular music genre is amongst the public because I thought the answer would differ from my own opinion. Though, my original question was to broad and hard to find data that wouldn't contradict. This eventually evolved my question into the one I have now, which is, "Who is the most listened to artist?".

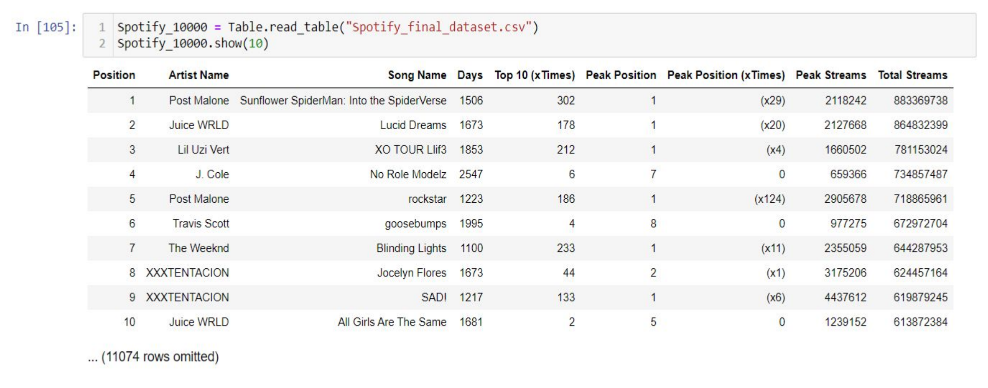
Scientific goal: My scientific goal for answering this question is so I can obtain a general understanding on what or who is deemed the most listened to artist in this day and age. Answering this question could help This data could then be used as a basis to answering many other complex questions. For example, "Does the amount of streams and number of hits songs correlate to an artist popularity or relevancy"

- Getting the Data -

- ## What Data do I need? -
  - find a recent dataset of the top streamed songs on Spotify
  - analyze this data and get (at most) the top 1000 streamed songs
  - use most relevant artists to determine # of songs in top 1000
  - use combined stream total from all of their songs, and then compare them to other artists. In result, I would Analyze the data from the top 10 Artist's to determine the answer to my question.
- ## Resources -
  - Data Source: https://www.kaggle.com/datasets/rakkesharv/spotify-top-10000-streamed-songs by, Rakkesh Aravind G (Kaggle)
  - This dataset from Kaggle. Scraped from Spotify API to rank them 1 - 10,000 overall on Spotify.
- ## Scope - What am I focusing on?
  - use the top 1000 songs in the dataset to find the amount of top songs from one Artist
  - combine the streams from all of these songs to determine total streams
  - Other data in this set, like peak streams, position, and days, will not be used as a factor in answering the question
- ## Ethics - What are the Ethics of Narrowing Data?
  - When deducting an answer from a smaller, more narrow subset, rather than the whole dataset, it is important to take into consideration inaccuracies that could be produced when not looking at all variables/outliers.
  - These factors could play a big role on how the data is perceived. There is a chance that these factors would even change the entire perspective of the question
  - example: my subset (only looks at top 1000 songs)
  - could affect artists total streams putting them at different position in the ranking compared to the results I would get with my data. Results vary depending on the scope.
  - Consider variables that are being left out with a smaller subset.(Maybe an artist has had the most songs reach the number one peak position, but their name might not even appear on the chart)
  - There are so many factors that play into the ethical side of asking and answering these questions and its important to be considerate of this and how it can affect your data
  - For this same reason, be considerate of the data you receive as biases and the that is used can alternate data possibly to how it wants to be perceived.

- Exploring the Data -

Raw Data:



```
In [105]: Spotify_10000 = Table.read_table("Spotify_final_dataset.csv")
          Spotify_10000.show(10)
```

Position	Artist Name	Song Name	Days	Top 10 (xTimes)	Peak Position	Peak Position (xTimes)	Peak Streams	Total Streams
1	Post Malone	Sunflower SpiderMan: Into the SpiderVerse	1506	302	1	(x29)	2118242	883369738
2	Juice WRLD	Lucid Dreams	1673	178	1	(x20)	2127668	864832399
3	Lil Uzi Vert	XO TOUR Llif3	1853	212	1	(x4)	1660502	781153024
4	J. Cole	No Role Modelz	2547	6	7	0	659366	734857487
5	Post Malone	rockstar	1223	186	1	(x124)	2905678	718865961
6	Travis Scott	goosebumps	1995	4	8	0	977275	672972704
7	The Weeknd	Blinding Lights	1100	233	1	(x11)	2355059	644287953
8	XXXTENTACION	Jocelyn Flores	1673	44	2	(x1)	3175206	624457164
9	XXXTENTACION	SAD!	1217	133	1	(x6)	4437612	619879245
10	Juice WRLD	All Girls Are The Same	1681	2	5	0	1239152	613872384

... (11074 rows omitted)

Data Wrangling

Column Information:

Artist Name - The name of the Artist that sang/created the song

Song Name - The name of the song

Total Streams - The total amount of streams the song has received

Removed:

~~**Position** - What rank the Song is based off stream total~~

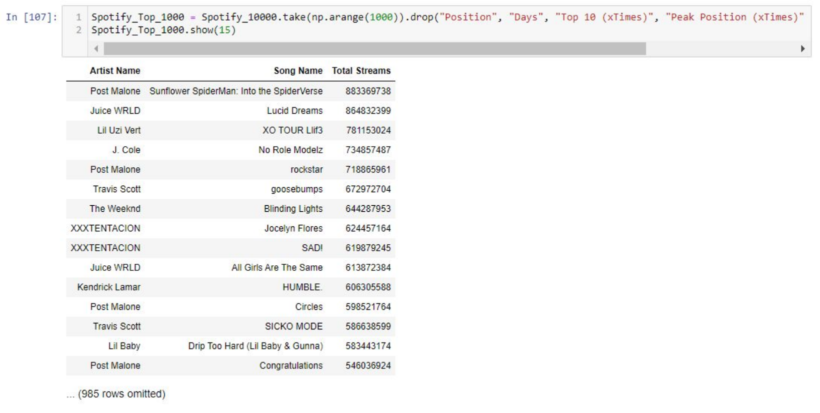
~~**Days** - The amount of days since the song has been released/made~~

~~**Top 10 (x Times)** - The amount of days/time the song has been in the top 10 charts~~

~~**Peak Position** - The highest position reached in the top 10 (1 being the highest)~~

~~**Peak Position (x Times)** - How many times the song has reached its peak position~~

~~**Peak Streams** - How many streams the song received when in peak position~~



```
In [107]: Spotify_Top_1000 = Spotify_10000.take(np.arange(1000)).drop("Position", "Days", "Top 10 (xTimes)", "Peak Position (xTimes)"
          Spotify_Top_1000.show(15)
```

Artist Name	Song Name	Total Streams
Post Malone	Sunflower SpiderMan: Into the SpiderVerse	883369738
Juice WRLD	Lucid Dreams	864832399
Lil Uzi Vert	XO TOUR Llif3	781153024
J. Cole	No Role Modelz	734857487
Post Malone	rockstar	718865961
Travis Scott	goosebumps	672972704
The Weeknd	Blinding Lights	644287953
XXXTENTACION	Jocelyn Flores	624457164
XXXTENTACION	SAD!	619879245
Juice WRLD	All Girls Are The Same	613872384
Kendrick Lamar	HUMBLE.	606305588
Post Malone	Circles	598521764
Travis Scott	SICKO MODE	586638599
Lil Baby	Drip Too Hard (Lil Baby & Gunna)	583443174
Post Malone	Congratulations	546036924

... (985 rows omitted)

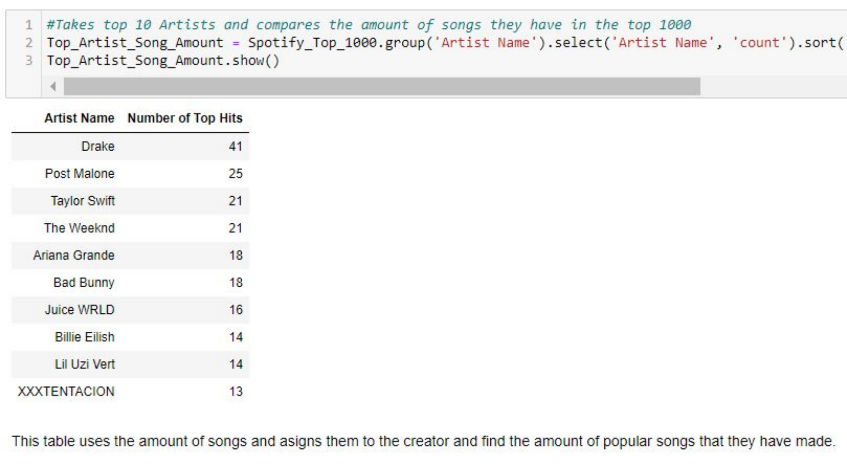
To tidy the data, I removed the columns that were not necessary for drawing a conclusion for my data and didn't focus on the scope. I also narrowed the dataset down to 1,000 songs. I only need the above columns and some additional math functions to achieve the data for my answer.

Data Wrangling

- Exploring the Data -

# of Top Hits

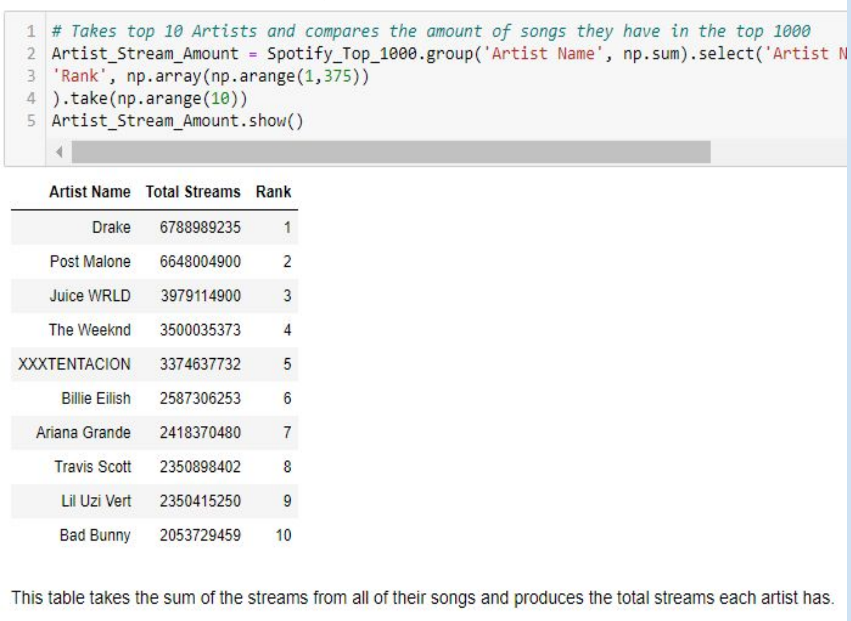
# of Total Streams



```
#Takes top 10 Artists and compares the amount of songs they have in the top 1000
Top_Artist_Song_Amount = Spotify_Top_1000.group('Artist Name').select('Artist Name', 'count').sort(
Top_Artist_Song_Amount.show()
```

Artist Name	Number of Top Hits
Drake	41
Post Malone	25
Taylor Swift	21
The Weeknd	21
Ariana Grande	18
Bad Bunny	18
Juice WRLD	16
Billie Eilish	14
Lil Uzi Vert	14
XXXTENTACION	13

This table uses the amount of songs and asigns them to the creator and find the amount of popular songs that they have made.



```
# Takes top 10 Artists and compares the amount of songs they have in the top 1000
Artist_Stream_Amount = Spotify_Top_1000.group('Artist Name', np.sum).select('Artist N
'Rank', np.array(np.arange(1,375))
).take(np.arange(10))
Artist_Stream_Amount.show()
```

Artist Name	Total Streams	Rank
Drake	6788989235	1
Post Malone	6648004900	2
Juice WRLD	3979114900	3
The Weeknd	3500035373	4
XXXTENTACION	3374637732	5
Billie Eilish	2587306253	6
Ariana Grande	2418370480	7
Travis Scott	2350898402	8
Lil Uzi Vert	2350415250	9
Bad Bunny	2053729459	10

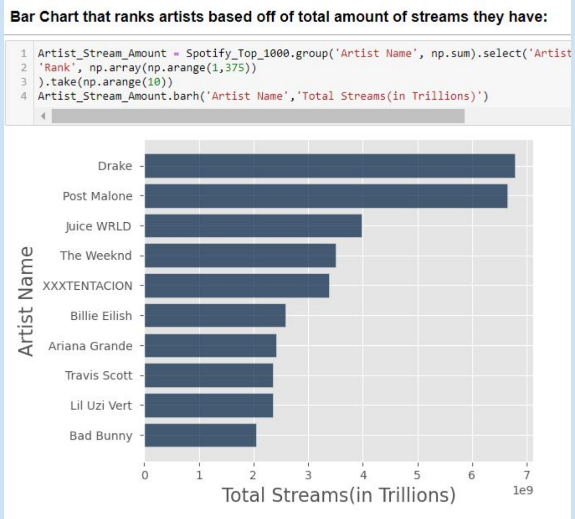
This table takes the sum of the streams from all of their songs and produces the total streams each artist has.

Further, I separated the table into two additional categories, one that finds the number of top hits, and one that finds the total streams. I did this by grouping data to artists. This will allow me to make a correlation between the two. Listeners like the artist if many of their songs are in the top 1000, and they are also listened to if they have the highest amount of streams.

Hypotheses/Observations:

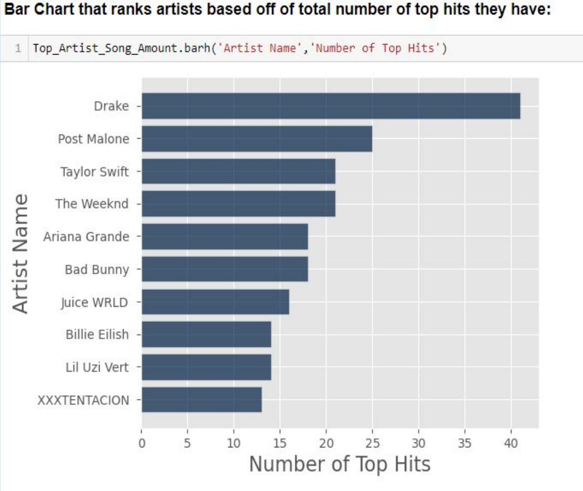
- 8/10 artists are in the top 10 chart
- half of them are in the same position
- rest displaced by one or two
- there will be correlation between top artist and total streams

- Visualization / Analysis -



Bar Chart that ranks artists based off of total amount of streams they have:

```python
Artist_Stream_Amount = Spotify_Top_1000.group('Artist Name', np.sum).select('Artist
'Rank', np.array(np.arange(1,375))
).take(np.arange(10))
Artist_Stream_Amount.barh('Artist Name','Total Streams(in Trillions)')
```



Bar Chart that ranks artists based off of total number of top hits they have:

```python
Top_Artist_Song_Amount.barh('Artist Name','Number of Top Hits')
```

-8/10 of the top ten artists appear on both charts.

-Half of those artists being in the exact same position in both bar graphs

-The others being shifted 1 or 2 places either up or down.

- correlation between the amount of hit songs made and the total amount of streams each Artist has obtained with all of their songs

- One could gather that the more hit songs an Artist has, the higher their stream count will be

- drake for example has significantly higher amount of hit songs, and also has the most streams

- According to the data, Drake is considered the most listened to artist

-he had over 40 songs in the top 1000 streamed songs. Which is a lot.

-The total streams of all of his hit songs were the highest amongst anyone.

- Using the objective answer gather from the data, one could consider that his music style is liked by a large majority of the population due to this data and is the most listened to artist on Spotify

- this is followed by the Weekend and Post Malone.

- Conclusion -

Takeaways -

- According to the data, Drake was/is the most listened to artist when looking at the top 1000 streamed songs
- I didn’t really expect these answers. I was surprised that Taylor Swift wasn’t even on the highest streams chart.
- The data should be taken with a grain of salt, it is an older dataset, so the answer are more for 1-2 years ago rather than currently. Drake does make more sense with this timeline. (Post Malone?)

What question does answering this data produce? -

- Answering this question opens the door for many more questions to be asked. Earlier: “ Do these results correlate to artist popularity or relevancy?” Big names - Beyonce, Taylor Swift, Harry Styles, Billie Eilish, Rihanna....
- Other questions: What will this data look like in 20 years? What *would* this data look like 10 years ago? Particularly interesting, could one use this data to develop some kind of pattern or algorithm to producing material that many will like? - used to generating revenue, popularity, etc.