CS181 Final Project

Spotify API and Billboard Webscraping

Alex Tubbs and Hieu Nguyen Notebook 3

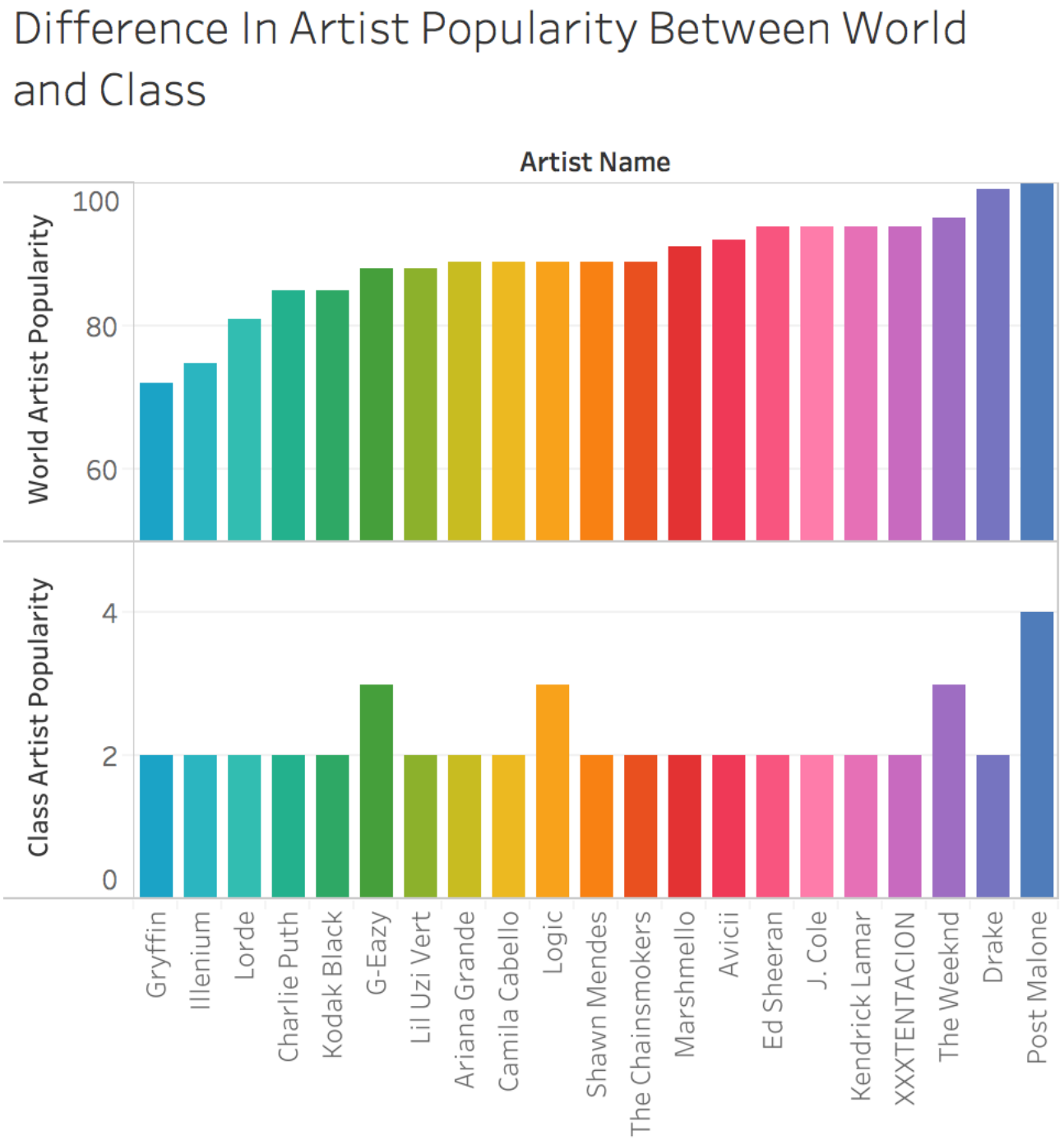
In this project, we explored the Spotify usage of some of our classmates, as well as Billboard and a few other websites to gather information on public listening habits. We started by getting our classmates to authorize our use of their Spotify information. We turned that information into usable tables that we could then make into mySQL tables. We also used the Billboard website and a Spotify article to "webscrape" or extract data from it using the HTML code behind the website. This gave us information on popular artists, songs, and genres. Using both sets of data, we could then compare our classmates' listening habits to those of the general public, looking for similarities and differences.

Questions

- Do our classmates have the same favorite artists as the general public?
- Do our classmates listen to genres at the same rate as the average person?
- Do certain characteristics about a song or artist affect it's overall popularity?
- Do number of songs an artist published have any effect on his/her popularity?

Question 1.

Our first question is whether or not our class listens to popular artists with the same frequency as other Spotify users. In order to do this, we used our table of Artist Genres, and counted the number of Artist and Genre combinations that were repeated, which would mean that more than one person in our class listened to that same artist. We then connected that table to our Artist table, to get the name of the artist that multiple people listened to. The graph below shows all of the artists that at least two out of the six classmates had in their top 50 artists on Spotify, which means that multiple people in our class listen to that artist often. This is compared to the artist's popularity score, which is a number assigned by Spotify between 1 and 100 to categorize how popular artists are, with 100 being the most popular on Spotify. Comparing the two will allow us to see if our class' music preferences differ from the preferences of all Spotify users.

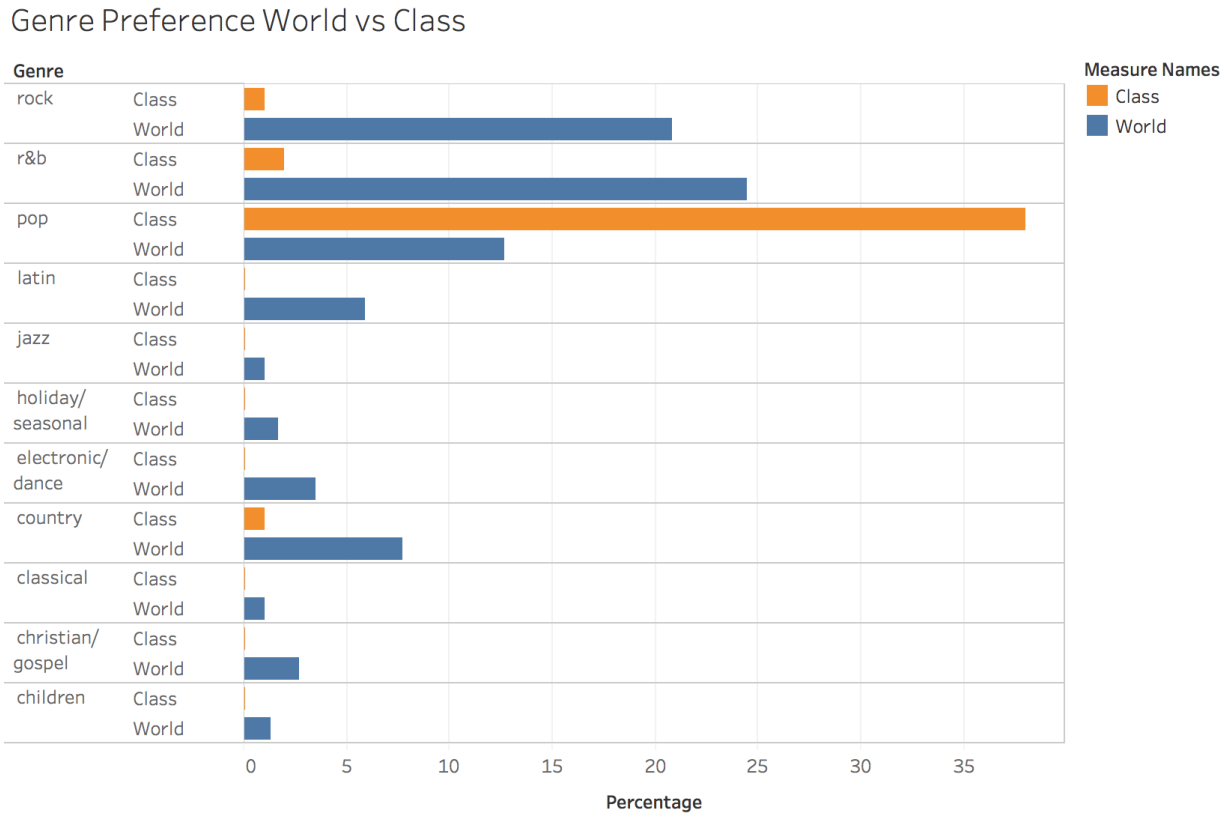


As you can see in the graph above, Post Malone is the most popular artist based on Spotify popularity rankings, with a raanking of 100. This would then make sense that he is also the artist thatthe most people in our class have as a part of the top 50, with 4 out of 6 people having him in their top 50. Beyond Post Malone, it's hard to find any trends about our class versus the world's ranking of artist, because we have so few people. It is worth noting that all artists that were included have a popularity score above 70, which implies they are all pretty popular among Spotify, so with a small number of people the only matches will likely be with popular artists. If we had more data it would be interesting to see if the rankings would follow the popularity scores more closely, beyond the top artist.

Question 2.

The next question that we wanted to answer was whether the genres of music that we as a class listen to are similar to those of everyone who uses Spotify. We found a report of the most listened to genres of 2017, which is what we used for our data about the world. For our class data, we found every time one of these genres came up in out Artist Genre table and counted them. This helped to avoid accidentally repeating an artist too many times, since each artist would only have 'pop' as a genre once, but may

have also had 'indie pop' or similar genres. We then had to join that table with our world table in an outer join, which is a way that makes sure all the information from our Genre table stayed, but we only kept information from our Artist Genre table that matched a Genre label. This allowed us to keep the percentages of the whole world, and compare it to what our class listens to.



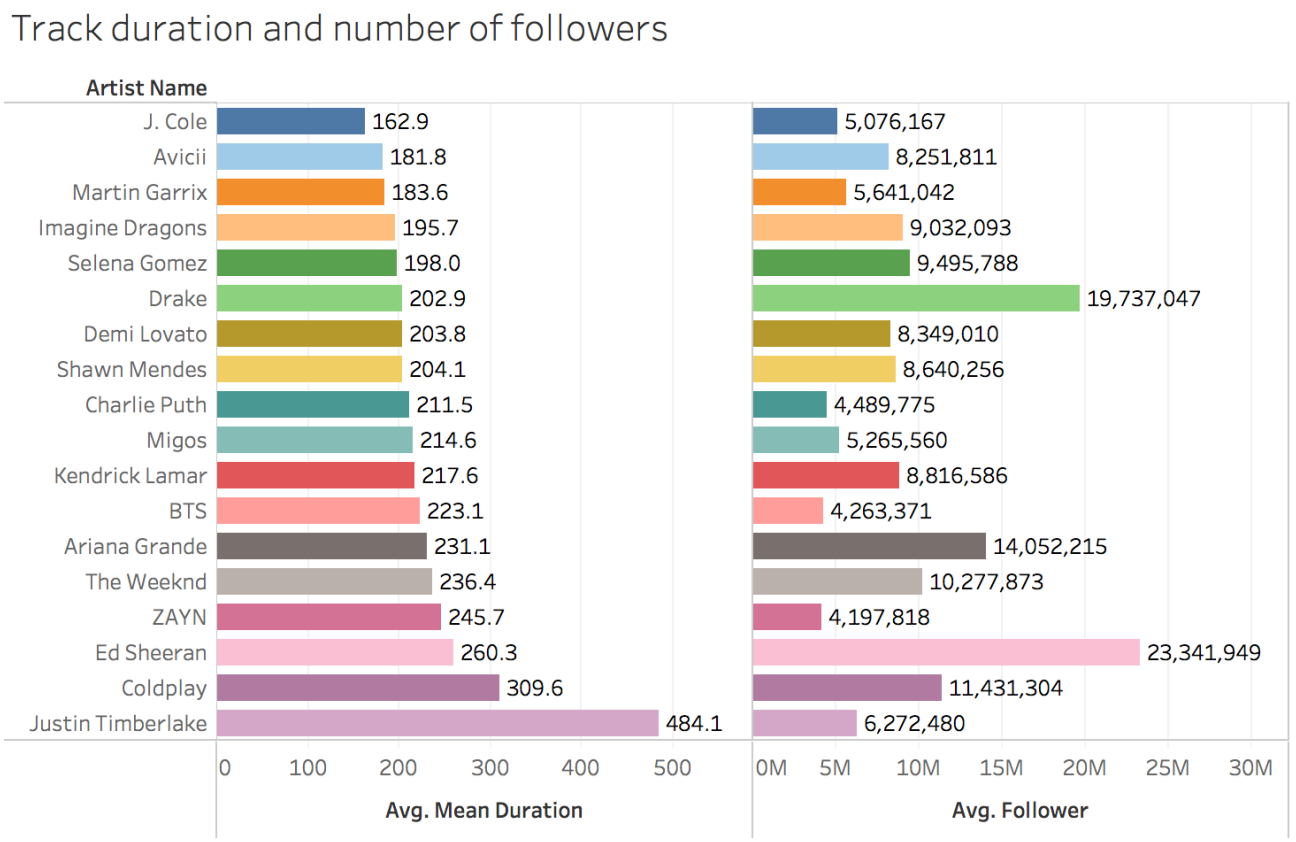
Class and World for each Genre. Color shows details about Class and World.

The above graph shows that our class greatly favors pop over any other genre, which is interesting because pop is only the third highest in the world rankings. The next three do fall in the order we would expect, with R&B/hip hop, then rock and country close together. However those are all incredibly small percentages compared to the pop percentage, and compared to the percentages they are supposed to be based on the world percentages. Part of this could be Spotify's fairly limited genre labeling, not using labels such as children's music or christian/gospel, so there may have been some of that music that simply was overlooked, but based on the demographic that we sampled, it does make sense that the highest genres would be pop, rock, r&b and country. Overall, it would be interesting to do this study with more classmates to see if the disparities in genre would continue to be high, or if more people would bring them closer to world averages.

Question 3.

Part 1.

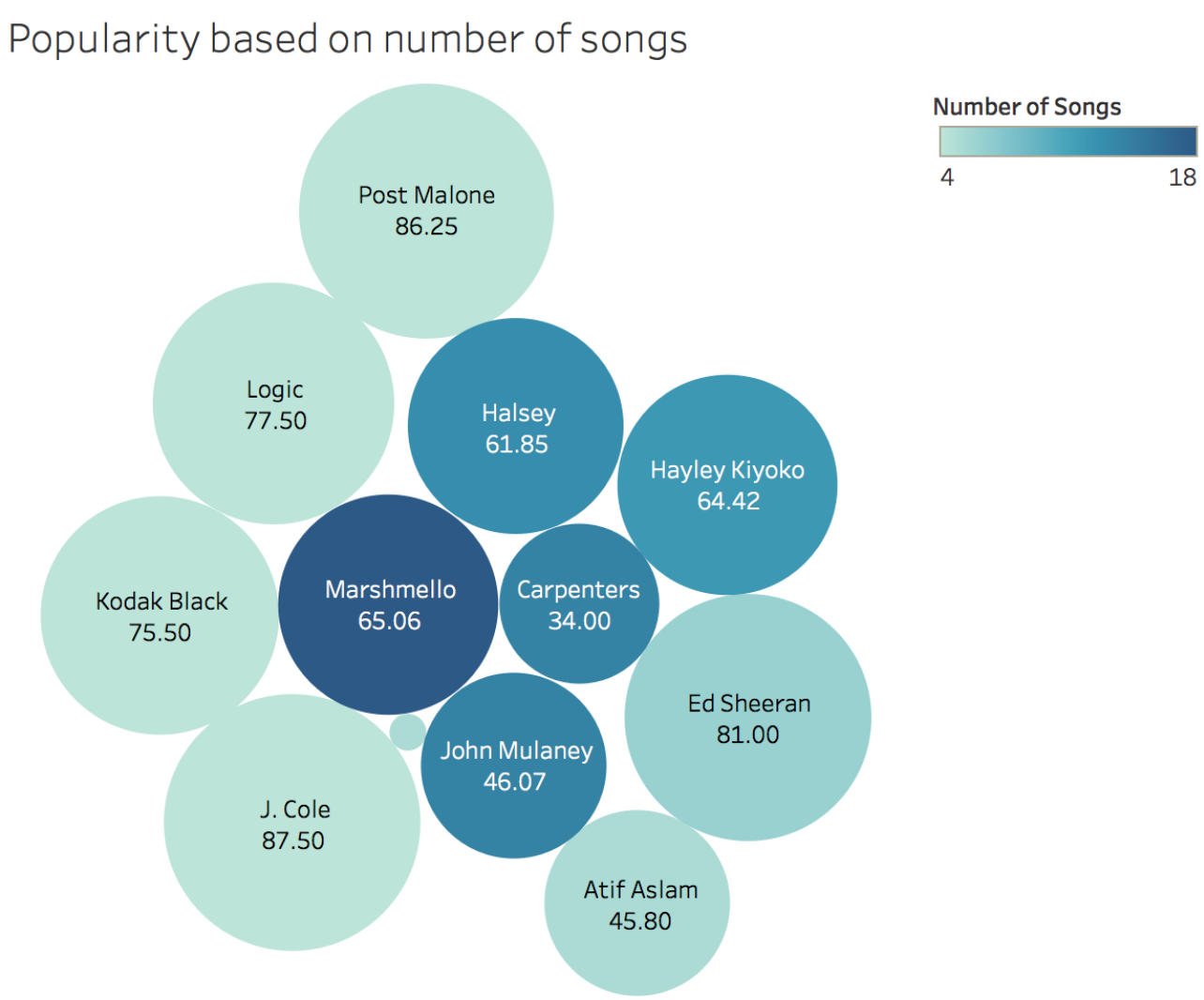
The third question that we wanted to answer was whether there were characteristics that could predict an artists popularity or number of followers on Spotify. The factor that we looked at was the average track duration, or how long an artist's songs tended to be. We compared this to an artist's number of followers on Spotify, which would theoretically be another good indicator of an artist's popularity, because if you like an artist, you're more likely to follow them. In the graph below, we filtered out artists to only include those that have more than 4 million followers, so we're only looking at very popular artists. We then have the track duration average in seconds, compared to the number of followers.



From the above graph, we can see that there does not appear to be any correlation between number of followers and the length of an artist's songs. Justin Timberlake has the longest songs by a large margin, but is in the middle to lower end in terms of follower count. From this you may think that longer songs have less followers, but Ed Sheeran has by far the most followers, and also has the third longest average songs. From these things, we can conclude that there is no real correlation between followers and song length.

Part 2.

In this final question, we wanted to examine the correlation between the number of songs published by an artist in our class top 50 songs, and his/her popularity. The number of songs an artist had in our top 50 was obtained by grouping by distinct artists and monitoring their emerging frequencies in the dataset. The individual popularity is an index which Spotify scales from 0 to 100 with 100 to be the maximum popular scores. Shown below is our graph on the relationship between artists' popularities and their number of songs



In the graph, the color intensification indicates higher frequency of artists in users' playlists, whereas the area of the circles shown in numerical values displays the individual Spotify popularity index. Marshmello has the most number of songs in user's playlist but his popularity point is only 65.06. Other than that, other common names like Ed Sheeran, J. Cole, or Post Malone have a decent point of popularity, considering that they have less songs than other artists. One surprising finding is that among these famous artists, Carpenters has a relatively small popularity number of 34. However, this is reasonable because Carpenters was an old vocal duo famous in the 60s, whereas Spotify users are young and more into other type of musics

Conclusion

In conclusion, we found that our class follows similar patterns to the average spotify user in terms of popular artists, as all artists that appeared for multiple people had a fairly high popularity score. However, the genres that our class listens to does not follow the average person, with the pop category being insanely high for us. This may be due to the small demographic we pulled from, though. We also found that track length cannot predict the number of followers that a certain artist may have.. And finally, as a class we listen to popular artists, but the songs we listen to the most are by less popular artists, generally in the 30-60 range, as opposed to ones we listen to a bit less, who generlly had popularity ratings above 75.