Part 2: Your Project
Question 3

I started looking backwards for an empirical question to work with as part of my final project. And I've heard Kaggle is a really good place for finding quality datasets to work with.

So I opened Kaggle and looked for datasets that just seemed interesting to me by the title, types of features, reviews and ratings under the dataset. As I was searching through datasets in Kaggle sorted by popularity, I found lots of interesting datasets that were useful for conducting some research on but there were two datasets that caught my attention immediately. One is Spotify top hit playlist (2010 - 2022) because it's an app that I use from time to time personally and so it stood out to me as something I'd like to potentially work with. After downloading and going through the data, I see that it has information for artists and their popularity along with their music and their popularity. Additionally, there were other attributes like duration of the song and what genres it comes under. But the set of attributes that really stood out to me were the musical features like danceability, speechiness, instrumentalness and so on. When I saw these, I was immediately interested in trying to make some meaningful analysis out of it.

To start off, I had checked to see if there was any correlation between the duration of the song and the track's popularity. Turns out, there is an interesting observation that can be made from the graph we see in the code, which is that a major portion of popular tracks are ones that have a duration of around 3 minutes. This could mean that artists are more likely to have hit a song if the duration is for 3 minutes. But we also know that not always is this case because track's duration is not the only factor influencing the popularity. There are a lot of assumptions being made about how good the actual song's background music or the lyrics are, for example. So to better understand this dataset we're dealing with, we could try measuring the track's popularity based on features like liveliness, danceability and so on.

For now, I couldn't figure out the code logic to plot track's popularity against all the musical attributes, but for the sake of analysis at least to some degree, I plotted the track's popularity against specific features individually. The first was against danceability, which turned out to be playing a huge role in a song's popularity. Next was loudness and this was an undisputed winner in what makes a song popular interestingly enough. Third was speechiness which had the most interesting graph. With speechiness, while most good tracks were with low speechiness a decently fair amount of tracks with moderate speechiness were also popular. And finally with instrumentalism, people definitely don't seem to like it a lot.

Something interesting to research further would be: did a track's speechiness contribute to the song's popularity less and less over time? Furthermore, does artist popularity influence the popularity of the music even though it has the key features that make a song popular. Lastly, an additionally interesting question to work with could be that whether for a musical artist to be successful, do they have to release music consistently so they remain relevant and popular even if it lowers the quality of the music or can they release music when they really feel something personally and make quality music like how they want. In other words, can musical artists make albums and release them whenever they want or do they need to publish from time to time so they don't lose their popularity? With all these questions in hand, I’m still interested in exploring further and trying to see if there are potentially better problem statements that I could do Data Analysis with.

The next dataset that I had picked to work with was a Housing market price. This dataset had a few features but is still useful in exploring its potential and figuring out if it's worth searching for a more sophisticated dataset later on to do a more in depth analysis.

The features we have in this one are bedroom_count, net_sqm, metro_distance, floor, age, and price. With this dataset, I was interested in finding what factors played a major role in deciding the price for the house. So first I checked the correlations between metro distance and price. And found that there wasn’t any particular pattern. Next, I checked with bedroom count and as one can expect the prices were indeed showing a pattern. But the most interesting one was the floor in which the house is located. Usually, one might think that housing in higher floors would be preferred because of the view you get at such heights but the graphs tell a different story. People seem to prefer lower floors and they are priced much higher. And finally with age, there seems to be no pattern and correlation with price. New and old houses are priced very differently. So of the given attributes, what one would generally expect to play a role in the pricing of the house like its location or proximity to the metro area and age don’t have as much of an impact as does floor. Some interesting questions to work with in this case might be, especially in a housing price dataset with more attributes, does the pricing of the houses have a correlation with cost of land or demand for housing and so on.

With some analysis on both these datasets, I find the spotify one to have more potential, although, it may not be the most relevant dataset for this subject. For that reason the housing price dataset might be a better choice. But even with housing prices, it isn’t the most interesting problem to work with. There is more potential in exploring bigger and more sophisticated insights that are not readily apparent to an individual, insights that can only be deduced with more exploratory data analysis, graphical interpretations and studies. Further research in the types of empirical questions and datasets available would prove to be much more fruitful.