Capstone Project 1 - Data Storytelling

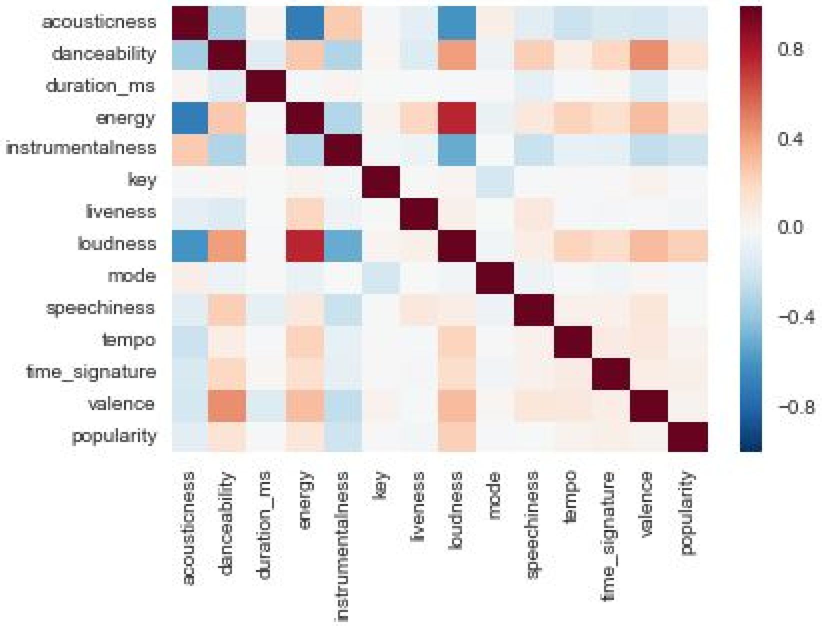
Spotify Audio Features

After going through a data wrangling/cleaning step, which was presented in the previous phase of the project, in this phase I attempt to answer the question below using visualizations only (i.e. without sophisticated statistical analyses). I will also look into the visualizations to see if there are any other interesting patterns I did not expect.

The Main Question: Is there a relationship between audio features of songs and their popularity in the dataset in hand?

To review the definitions of variables, please see the data wrangling ipython notebook for this project on GitHub:
https://github.com/hamidniki/Capstone-project1/blob/master/Capstone%201%20Data%20Wrangling.ipynb

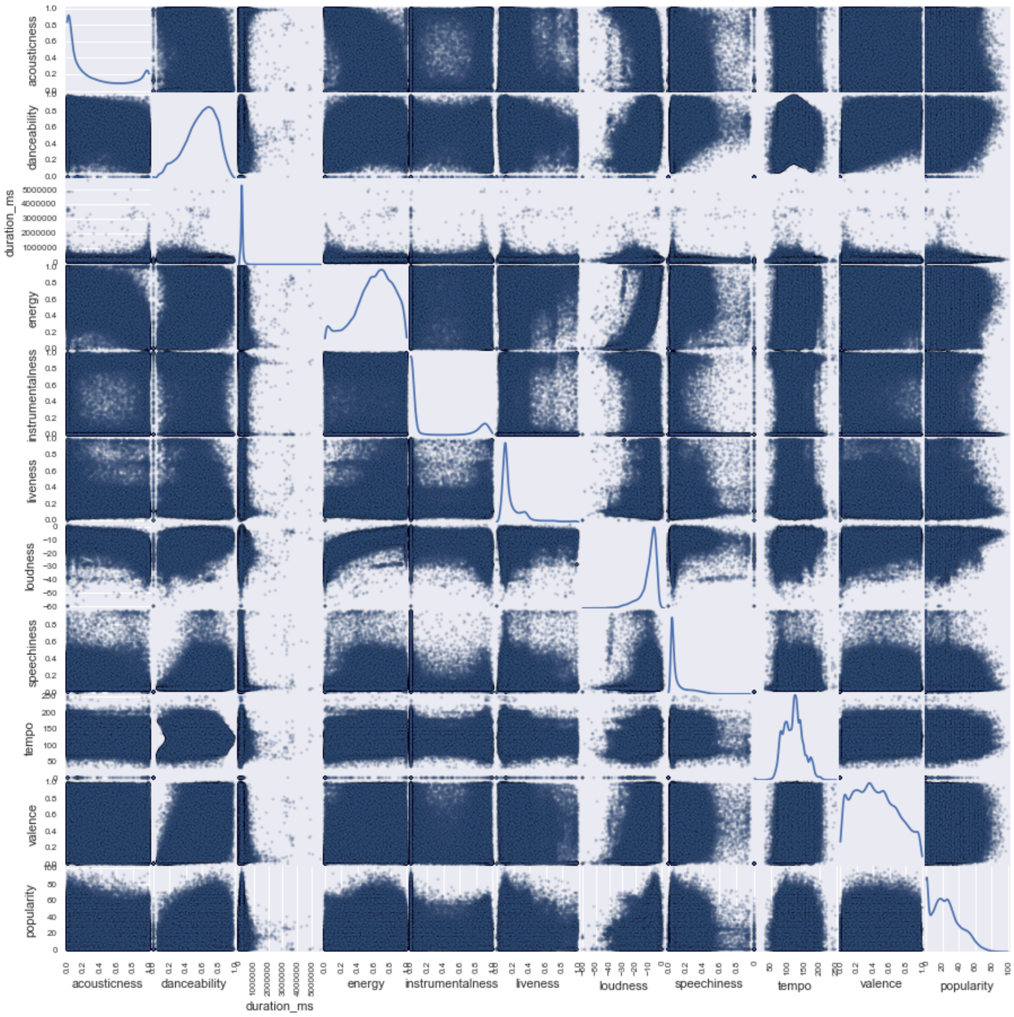
Looking at a visual presentation of the correlation matrix between the popularity variable and the variables that represent the audio features, the following relationships between the variables are apparent:



- Looking at the bottom row, popularity has the highest correlation with loudness. This is not surprising because louder musics with higher energy tend to be more danceable and played often at bars/parties
- Adding to the point above, in the plot we see loudness is highly correlated with energy and has, less but still noticeable, correlation with danceability and valence (how happy the music is), which are all elements that make a song more suitable for parties

- Valence and danceability are positively correlated and that is because happier songs tend to be more danceable
- Popularity also seems to be negatively correlated with instrumentalness(absence of vocal in the track). This indicates listeners are more inclined toward tracks that contain vocal. This makes sense as well because some people are less inclined by the music itslef and it is the lyrics that draws them to the music. So music tracks with vocal elements are at an advantage.
- Tempo has noticeable positive correlation with energya dnloudness
- Speechiness and instrumentalness are negatively correlated which makes sense given the definition of these two variables
- Loudness and energy are highly negatively correlated with acousticness. This is because musics produced using acoustic instruments tend to be softer than electronic music.

Then the scatter matrix of the numerical variables was examined to see if any additional points can be detected in the data. The additional observations were as follows:

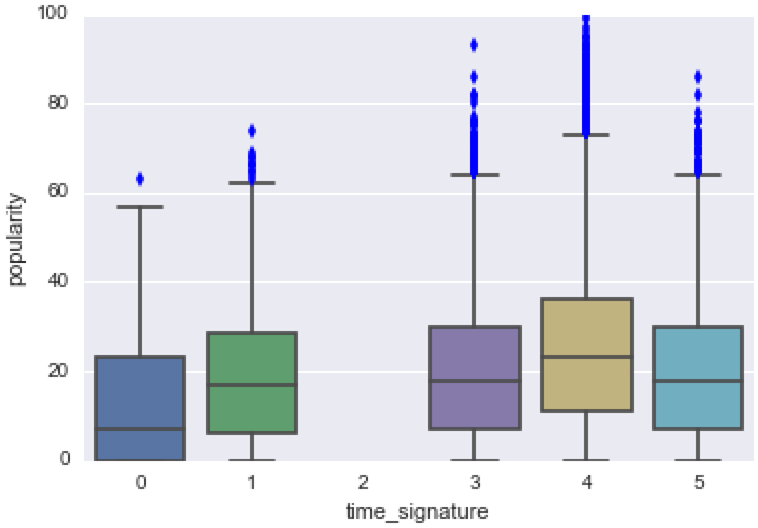


- Much of what we observed in correlation matrix can be seen here too as these two kinds of plots convey the same information
- Tracks with very long length (e.g. 80 or 90 minutes) tend to be mostly soft and instrumental and used less for dancing.
- Live tracks are less popular in spotify. This is because, eventhough listening to live music as it is being played is entertaining, because of the lower quality of sound, the recorded versions of live performances are not popular for listening in car or playing at home/parties
- Tracks with extreme tempo levels (< 50 BPM or > 250 BMP) are not popular

Visual Inspection of Categorical Variables (Mode, Key, Time_signature):

Variables mode, key and time_signature are represented as integer variables in the dataset. However, since these variables do not have ordinal nature, we will use them as categorical variables in future analyses. We visually inspected the relationship between these variables and popularity using boxplots and concluded that:

- There is no visible difference between minor vs major modes in terms of popularity
- There is no visible difference between different keys in terms of popularity
- Rhythms with 4 beats per measure seem to be more popular. This could be due to the following facts:
  - 4-beat rhythm is more common than any other kind of rhythm. This is the most familiar beat to human ear and is the most frequent time_signature in the dataset (~ 86%)
  - 4-beat rhythm is commonly used in popular genres of music which have fast tempos, high energy and are danceable.



Notice there are no 2-beat tracks in the dataset. This rhythm is not as common as other rhythms, but it exists. This might be because tracks with 2-beat rhythm are categorized as having had 3-beat rhythm, due to the similarity of these two kinds of rhythms (i.e. 2/4 signature vs 3/6 signature)

Checking for Interactions Between Numerical Variables and Levels of Time Signature:

Interaction between the time_signature variable and the other numerical variables was also inspected. For each level of time signature, the correlation matrix and plot was generated. However, the observed correlations were the same as those originally seen using all levels of time signature.

Conclusion:

From the visual analyses described above, it is apparent that loudness, energy, danceability and time_signature have the strongest positive correlations and acousticness and instrumentalness have a weak negative correlation with popularity. This hypothesis, however, needs to be formally tested using formal statistical analyses in future phases of the project. We visually inspected the interaction between time signature and the numerical variables and we did not detect any interaction. There could, however, be interaction between other variables which we did not check for. We will dig deeper into the data using statistical analyses in later phases of the project.