Raven, Baldwin

Zhao, Bingxin

CS 242

4.19.2022

Final Project

Spotify Song Feature Analysis Project

My dataset is the track audio features of 154 songs from a playlist I made for the cafe I

work at. The problem I am examining is whether or not it meets the standards of the company as they grade each of the cafes on whether their music is upbeat enough to create the pleasant atmosphere they want the cafes to have. I am examining the overall statistics of the song features and determining whether or not the playlist meets the standards. My other problems and questions I will be looking at is that is there a relationship between the variables of the song and are they good for predicting other values for the variables of the same song. I will be also looking at whether or not the values are good for predicting the genre of the song. Overall, I have determined the playlist is fairly upbeat and filtered it to make it more upbeat and there is not much of a relationship between the variables and they are not good for prediction. The dataset is also not that useful for predicting genre.

Again, I am examining a dataset of audio features from a playlist of 154 songs I made to play in the cafe. The audio features I have in the dataset are danceability, acousticness, energy, speechiness, instrumentalness, and valence. I will be mainly focusing on danceability, valence, and energy as these features are what make an upbeat song. The problem of making an upbeat playlist for this cafe is important for all of the cafes as we are graded on how well our cafe functions, including atmosphere, as they look at these grades to determine if anything needs to be done to correct the cafes. So far, we have received fairly good grades. The other problems I have mentioned are things I am looking at out of interest as I want to work for Spotify so I wanted to see if the variables I mentioned are actually good predictors of each other as they seem related so I want to see if they are. I also wanted to know if the variables’ values are good predictors for the genre. I am using descriptive statistics to describe the data and then creating filters for the data by looking at them to see what kind of playlist results. I also used histograms and fitted linear models to visualize the data. I am using the variables Danceability, Energy, and Valence to create different linear models to see if any of the variables are good for predicting one another. I used KNN, logistic regression, and cross-validation to see if the variables are good predictors of the genre. Based on the linear models I discovered that the variables are not good predictors of each other. There is not much of a relationship between them. And the variables are only 50% useful for predicting genre.

As I stated before the variables I have in this dataset are danceability, valence, energy, aucousticness, speechiness, and instrumentalness. To obtain these values I used a Spotify web API that takes in songs id’s(which can be found at the end of sound links) and outputs the song's audio features. I inputted the songs IDs individually for all the songs on the playlist and inputted them into a CSV. I first created some code that took the playlist id and outputted all of the song titles, artists, and id which I took and used to input into their API tool. Below are what the features and their measurements [mean](https://medium.com/@boplantinga/what-do-spotifys-audio-features-tell-us-about-this-year-s-eurovision-song-contest-66ad188e112a):

Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

Acousticness: A measure from 0.0 to 1.0 of whether the track is acoustic.

Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.

Instrumentalness: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.

Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

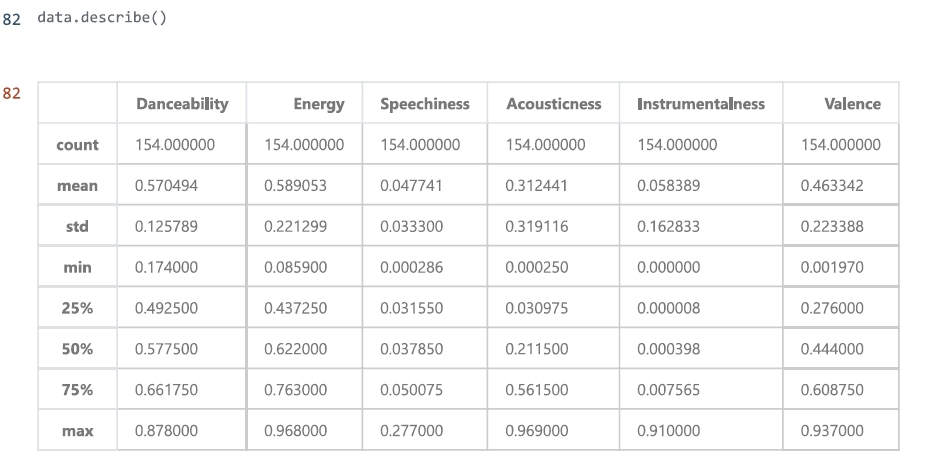
There are a total of 154 songs but after I applied my filters to the dataset I had a playlist of 39 songs.

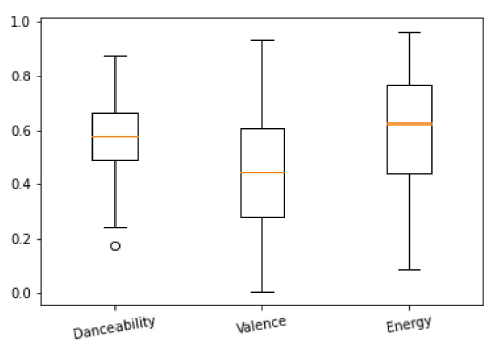
Although the problems do have many implications I really wanted to look at whether or not it met the standards of being upbeat in order for the cafe to receive good marks for atmosphere. This was the first problem I wanted to look at because it is a problem that directly relates to the data I am looking at and I can use the audio analysis values to determine whether or not it makes a good playlist. I relied on descriptive statistics as I could see overall what the composition of the playlist was in terms of audio features. I mostly looked at the mean and quartiles because those gave me a good idea about how much the playlist already fit the ideals I set. I wanted the danceability, energy, and valence to be above or close to .5. My next problem was to determine whether or not there was a relationship between danceability, energy, and valence by fitting linear models and looking at both plots and the model summaries, and looking at the R^2 and the p values to see if there is a relationship with the null hypothesis that the coefficients are good for prediction by using the alpha of .1. This doesn’t relate to my main problem very much but I am looking at it to see if the variables show any sign of a relationship. This problem does relate to my other problem which is whether or not danceability, energy, and valence are good predictors of the genre of the song. I am using KNN, cross-validation scores, and logistic regression to evaluate this. I feel my approaches will reveal what I want to know as they give me the exact answers I want and tell me whether or not the models I have are useful or whether or not a variable is a good predictor through scores and hypothesis tests. I did consider other methods but as I kept coding and looking at the results I found that they wouldn't be very helpful as I had already discovered that there was almost no relationship between the variables and they weren’t very good for prediction.

Process

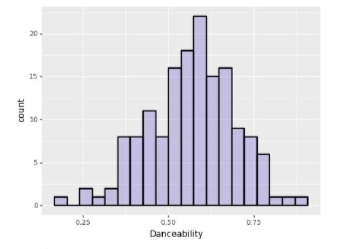
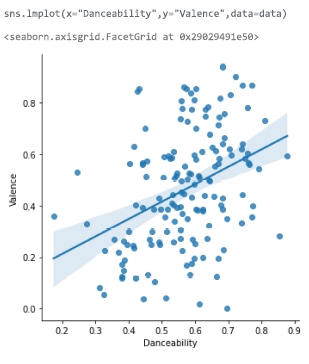
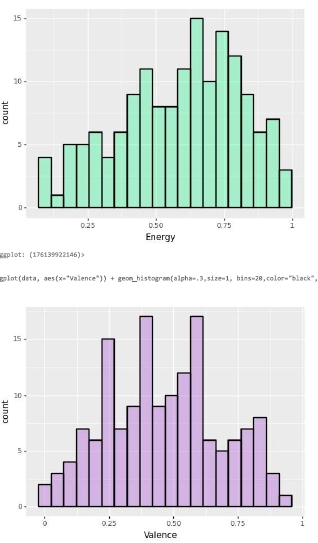
1. Examine the original dataset using summary statistics
2. Plot histograms and linear models with different variables to get an overall view of the data
3. Filter dataset and examine again using summary statistics
4. Repeat plotting to view the new dataset
5. Fit linear models: Danceability~Energy+Valence, Valence~Danceability+Energy, Energy~Danceability+Valence, examine summary statistics and plot to visualize
6. KNN, cross val scoring, logistic regression to determine if the 3 variables are good for prediction

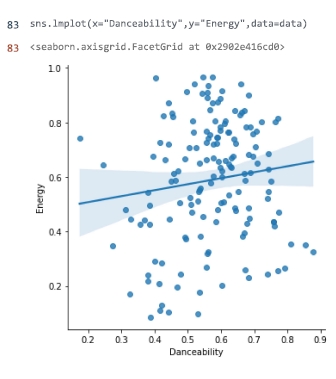
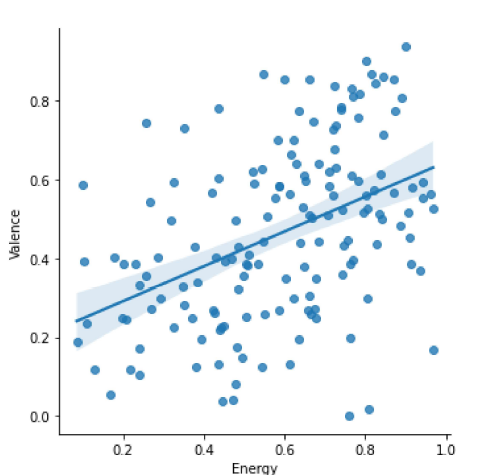
For my first question about how upbeat the songs were already, I visualized the data first to see how it looked overall. Like I have mentioned before I mostly focused on the 3 variables danceability, energy, and valence because based on their descriptions they would best convey how upbeat the song is. First I used descriptive statistics for the whole dataset:



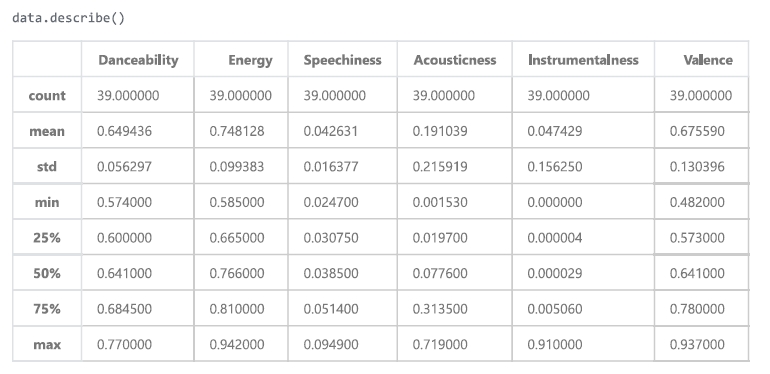


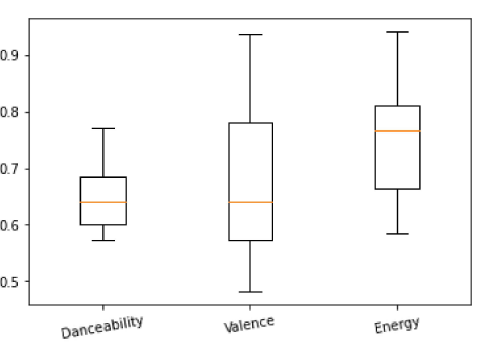
I mostly looked at the mean and quartiles. Overall it has a good distribution. The mean and median are both over the 50% goal I had for the songs for at least the danceability and energy. The valence was a little lower than what I was looking for. The three variables also have fairly high maximums and fairly high 75% quartiles and maximums. I then used these measurements to set my standards for how I should filter the playlist to make it more upbeat and I filtered it to be over the mean values cause I felt those averages better described what the dataset looked like. Before I did that I wanted to look more at the distributions and what the linear models plotted looked like before it was filtered:



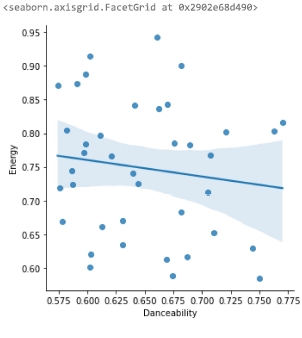
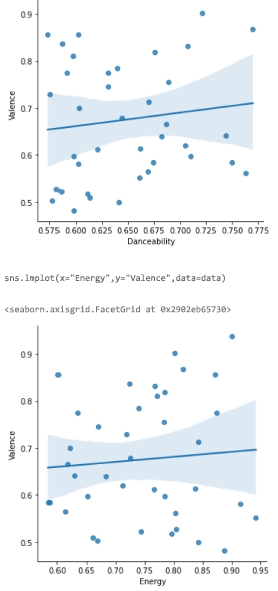
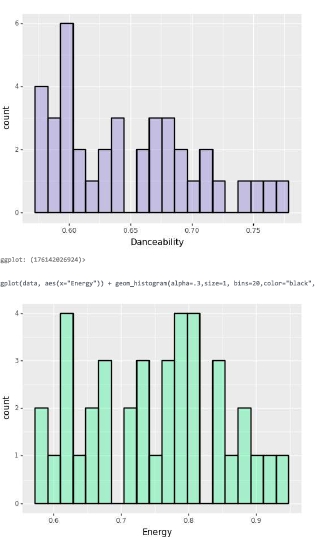


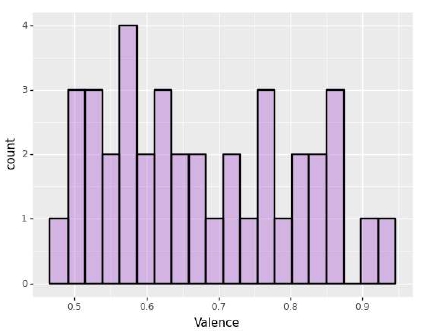
The linear plots get more into when I look at the relationships between the 3 target variables but I wanted to look at the distributions of the variables. Valence and danceability appear to have fairly normal distributions although danceability has a slight left skew while energy is skewed to the left. For the energy variable, this is good because it indicated that the songs are more energetic. This gets more into my second question but the linear models do not indicate much of a linear relationship between them and the points are fairly randomly distributed so I was interested to see if that would change after I filtered the dataset. After filtering:



As you can see the means have increased significantly, same with the medians. In fact, the medians are now fairly close to the means. 

The energy for the filtered dataset is a lot higher than it previously was and in general, this filtered dataset seems a lot more “upbeat” than before, but that danceability seems a bit lower than the other variables. The valence does have a lot more variance in it though I did set the limit for it to be about 10% less than the other limits for danceability and energy. I also created the same plots to compare to the ones from before it was filtered





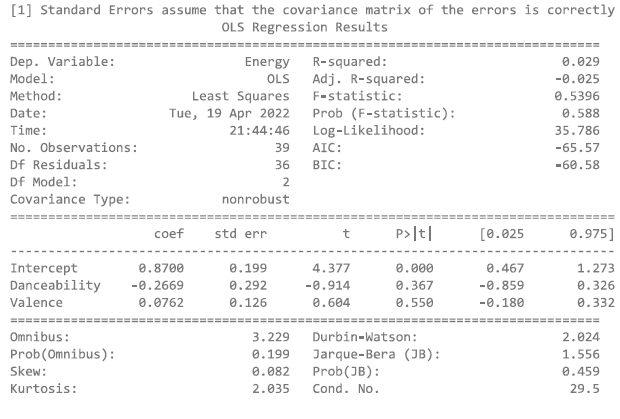
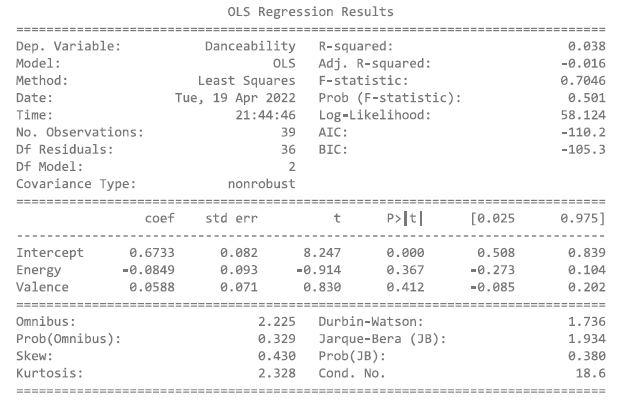
The linear plots have changed, especially danceability and energy as the slope is now negative. The distributions are also no longer normal. Overall based on the histograms and the summary statistics I would say that the filtered playlist is more upbeat now and would fit the standards.

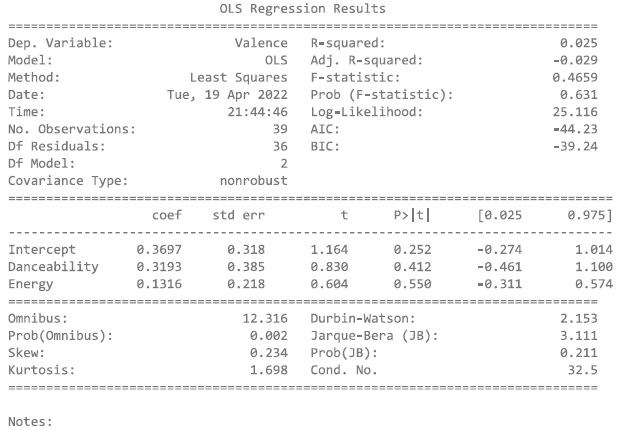
Next I will be further examining 3 linear models Danceability~Energy+Valence, Valence~Danceability+Energy, and Energy~Danceability+Valence to see if they are good for prediction. I will be looking at the summary statistics and setting my hypotheses:

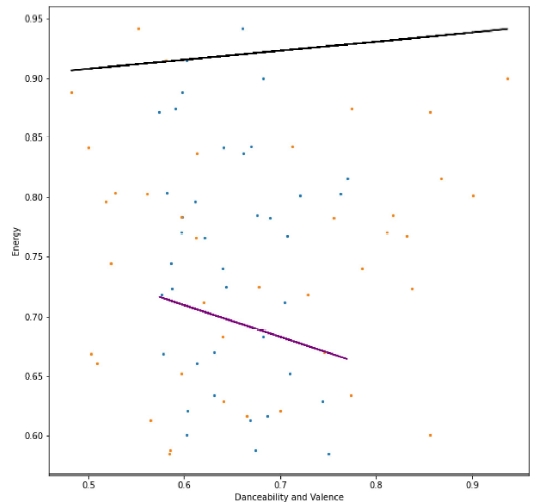
=The interaction between the X variables are useful for predicting the Y variable

=The interaction between the X variables are not useful for predicting the Y variable

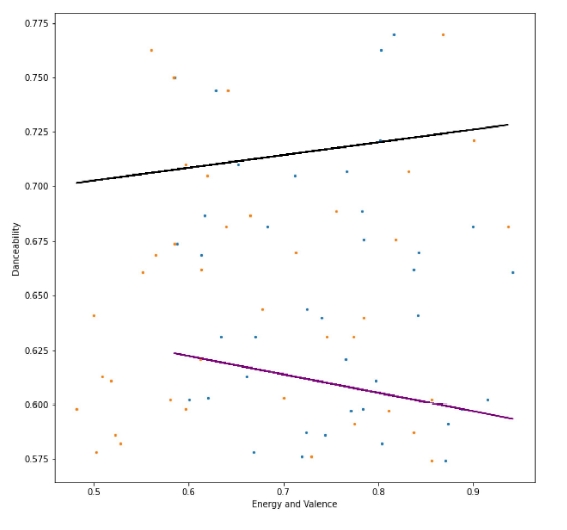
I will be using the alpha of .1. These are the summaries for each model:

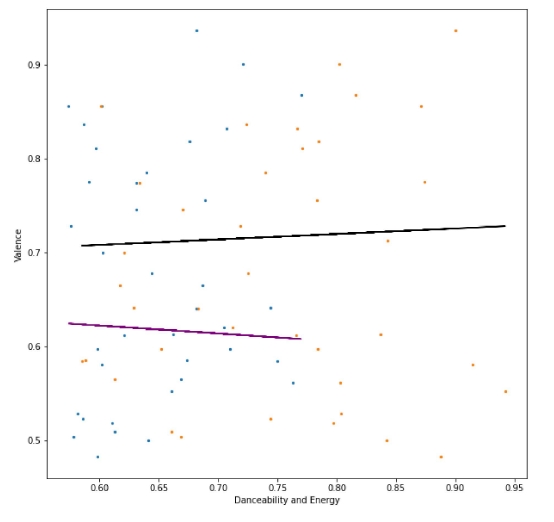


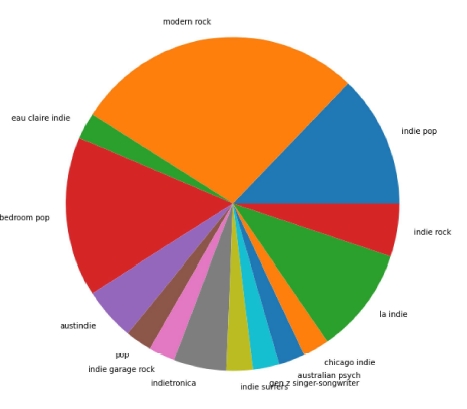
Based on the summary statistics all of these p values are over the alpha of .1 we fail to reject the null hypothesis that something within the models is useful for predicting the Y values as we have insufficient evidence. The R^2 values are also very low which also indicates a very small amount of the variation in the Y values is explained with its relationship with the X values. It would seem of any of the models that Danceability~Energy+Valence would be the best.I also plotted these linear models to visualize these results.



As we can see by these plots and the model summaries there does not seem to be much of a relationship between these variables and they might not be reliable for predicting each other at all.





Next, I will be answering my question about if the variables are useful for predicting the genre of the song based on their values. I used KNN, logistic regression, and cross-validation scoring. First I wanted to look at the proportions of the genres within the new playlist using a pie chart. This is on the next page. There are a total of 14 different genres in the filtered playlist containing 39 songs, a majority of these are modern rock. My KNN model used a number of 2 neighbors. The predicted result was modern rock and indie-pop. When scoring the model it resulted in a value of exactly 50% with a test size of .05. I adjusted the number of neighbors and the test size to see if I could get a better score but that was the best I could do. I did the same with logistic regression with the same test size and it outputted the same value as the KNN model except it predicted modern rock only and was the same when I adjusted it. The cross-validation scores across five groups were 0.125, 0.125, 0.25,0.25, and 0.14285714 with a mean of 0.1785. All of these values indicate the variables are not very useful for predicting the genre of the song. I believe this to be possible because of the fact that there is a wide variety of genres and there is a lot of variation in the values of danceability, valence, and energy.

In conclusion, I have learned a lot about this dataset, looking at the summary statistics let me know on average the playlist is already fairly upbeat yet it could be improved. I also found that the variables are not very useful for predicting each other and there is not much of a relationship between them. The variables are also not very useful for predicting the genre of the song. I have done a similar study with another playlist for another class and had produced similar results, but here I used some different methods and went more in-depth so it was not entirely surprising to see these results but I was still interested to see if it would change with a different dataset. A bit of a far reach but I think something interesting could be done with these statistics, since my first goal was to try and make a more upbeat playlist perhaps values for these variables could be used to create playlists based on what kind of mood or kind of song you want to be in your playlist. I myself would like to pursue this idea further and see what I could create!