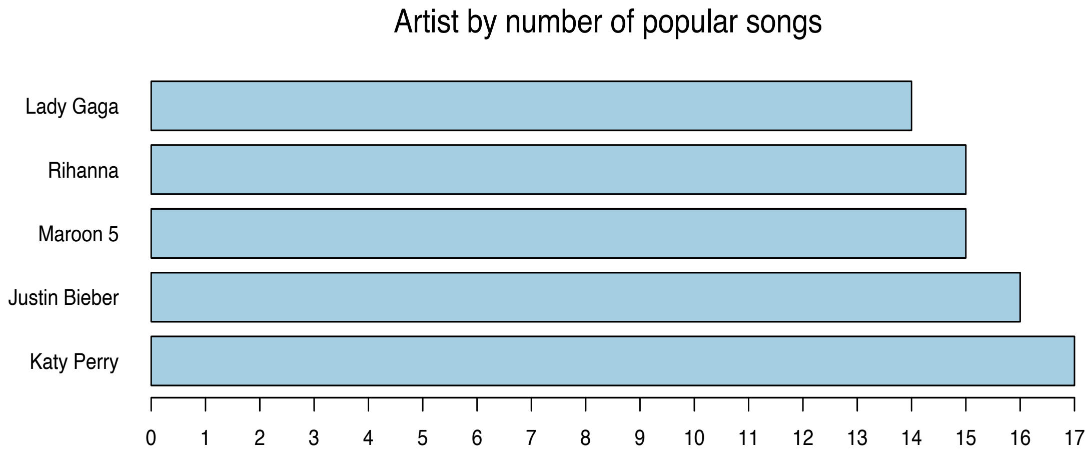
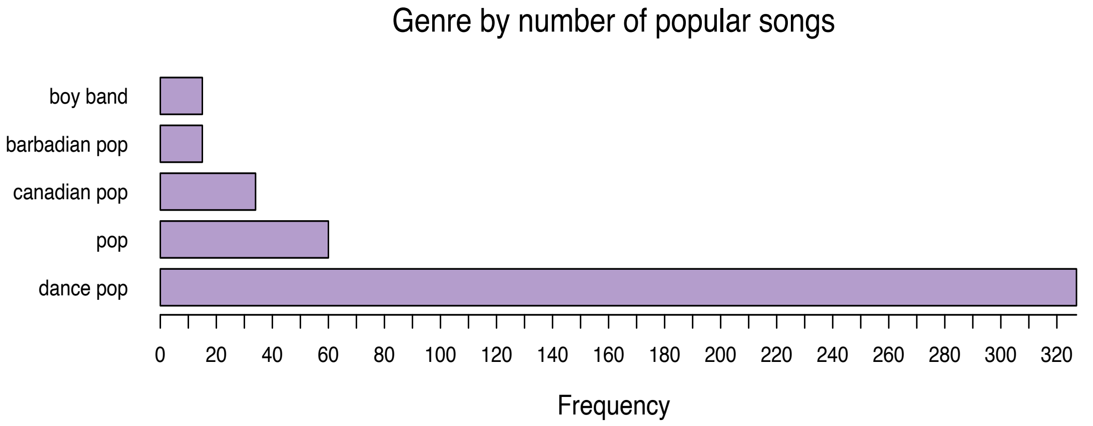
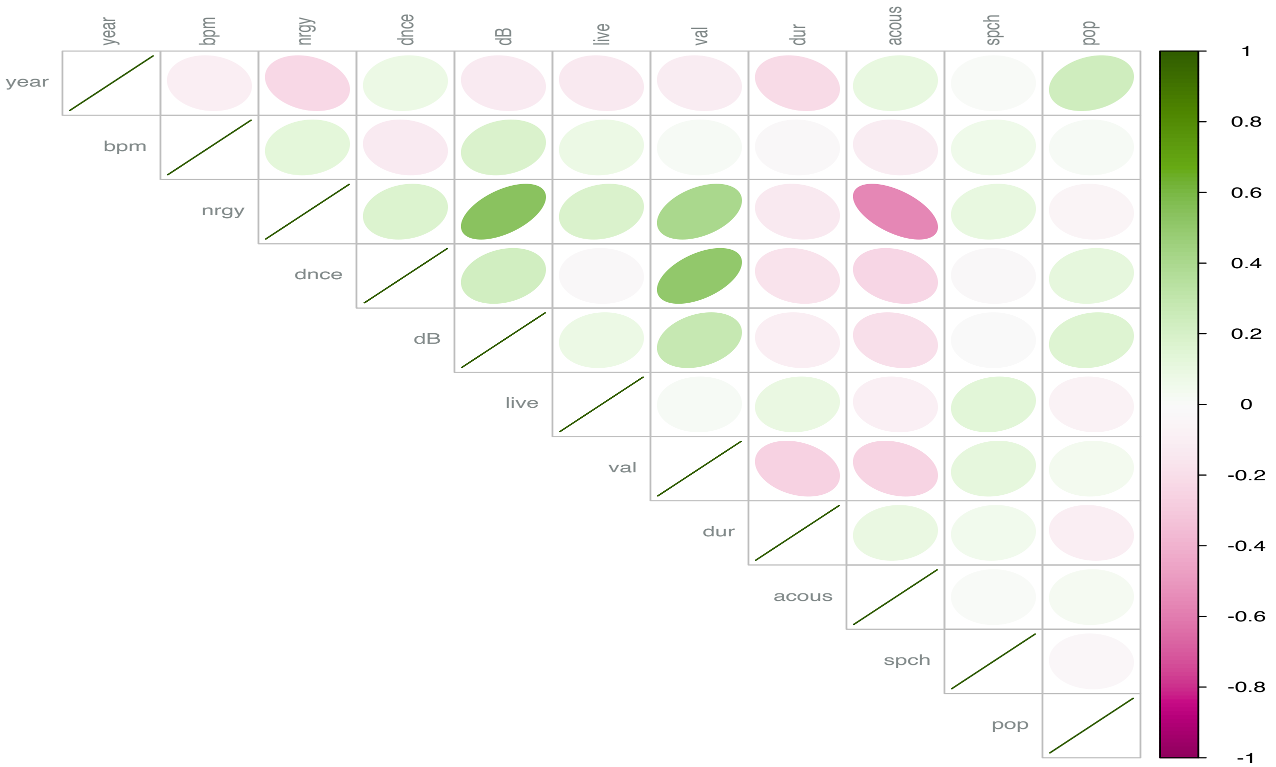
FDMTL Data Vis in R – Prof. Bruno Richard
Sasha Rider – 198008430

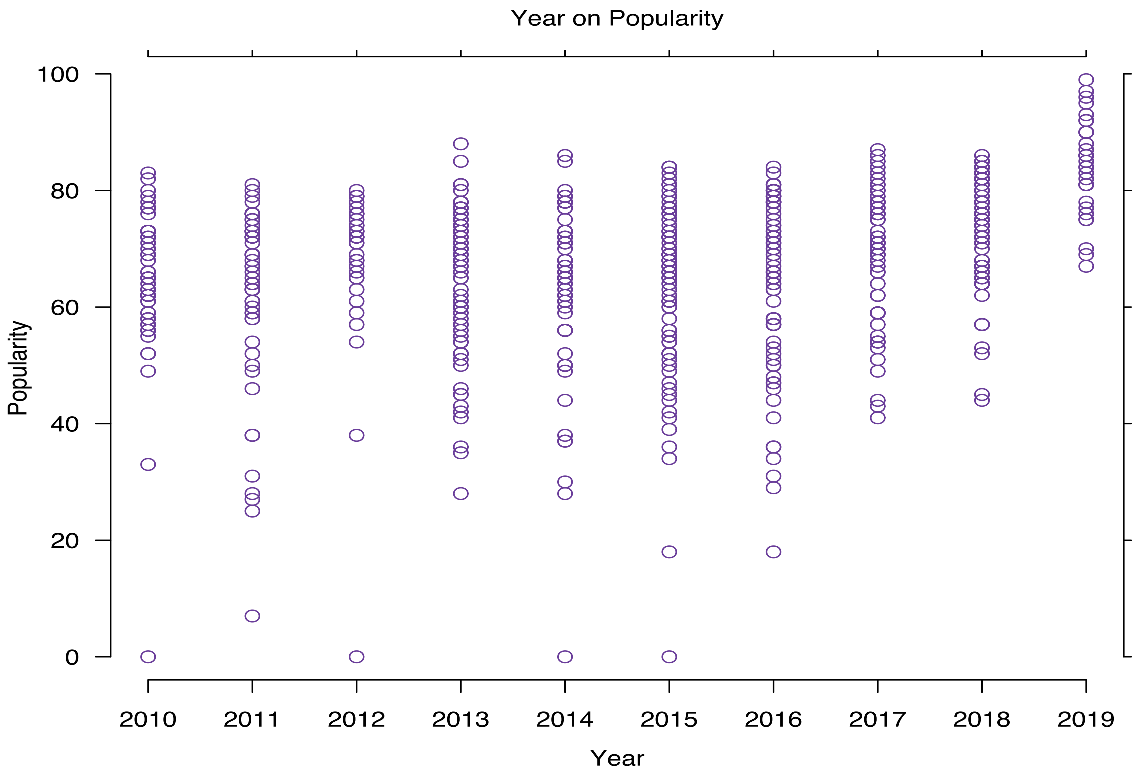
For the final project, I found a dataset on Kaggle that contains the most popular songs on Spotify from 2010-2019 with data such as duration, bpm, decibels, valence, etc. on each of the songs. I chose this dataset because I wanted to explore the effect that different factors had on the popularity of songs. Music has always been one of my passions and my goal was to learn about listening patterns among users as well as learning the effects that certain outliers, such as a lot of spoken words or high level of speechiness, has on how well liked a song is. I have always wondered why certain songs seem to hit home more than others, and through this analysis have learned a fair share of what makes these songs superior in comparison. Some notable information about the data is that there are 15 columns with 13 variables. Without an explanation needed we have song title, artist, genre, year, and pop for popularity. However, we also have columns that may not be obvious from the name alone. Notably the "bpm, "nrgy", "dnce", "dB", "live", "val", "acous", and "spch" columns which have strange names. Bpm stands for beats per minute and is basically how fast or slow the song is aka the tempo. Nrgy is for the energy of the song hence the higher the value the more energetic the song. Dnce stands for danceability, which is measured by musical elements including tempo, rhythm stability, beat strength, etc. to gauge how suitable a track is for dancing. dB is for decibels and the higher the value, the louder the song is however all of these values lie in the negative range. Live stands for liveliness and the higher the value is the more likely the song is a live recording. Furthermore, val stands for valence which describes the mood of the song, so a lower value will indicate a sad song whereas a higher value relates to a happier, upbeat song. Acous is for acousticness which means that the song primarily uses instruments, and these instruments create sounds without the help of electronics. Lastly spch stands for speechiness and represents how many spoken words are in a song. I found this dataset very interesting to work with because I know most of these songs, so I had a bit of nostalgia, and I am very interested in studying trends as it is the closest, I can get to predicting the future. While working with this dataset I came across many interesting questions, such as: why is pop the top genre? What makes a song popular? Who were the top artists from 2010-2019? How do certain musical elements combine? Is there a way to predict which song will be popular? As I came up with these questions I was intrigued and decided to put my R skills to good use so I could visualize the data. My hypothesis for this dataset is: What factors contribute to popularity in a song? As I started working with the dataset, I found a song by Adele called "Million Years Ago" which I had to remove as it had incorrect or missing information. This song contained a bpm of 0 which is impossible and would skew the scale of bpm when plotting. After cleaning the dataset, I wanted to create a barplot that would show the top 5 artists and genres. Then I wanted to create a correlation matrix to see how all of my predictor variables lined up. I would use the 11 numerical (integer) columns I had to plot the correlation matrix, and later to build linear models. A particular interest of mine is machine learning, so I found this to be challenging yet fascinating. Next, I would build scatterplots for each predictor variable to visualize how they are distributed. I thought it would be useful to see whether the predictors were evenly distributed or left/right skewed. After building the scatterplots, I continued building but this time it was for generalized linear models. After calculating the AIC scores, I chose which model was the best fit and singled that one out so that I could run it on the test dataset. It proved to be a good model and I was able to come to a conclusion. I will post all of my graphs/figures below and provide an explanation for each as well. Then I will provide a conclusion for my hypothesis.


Artist by number of popular songs

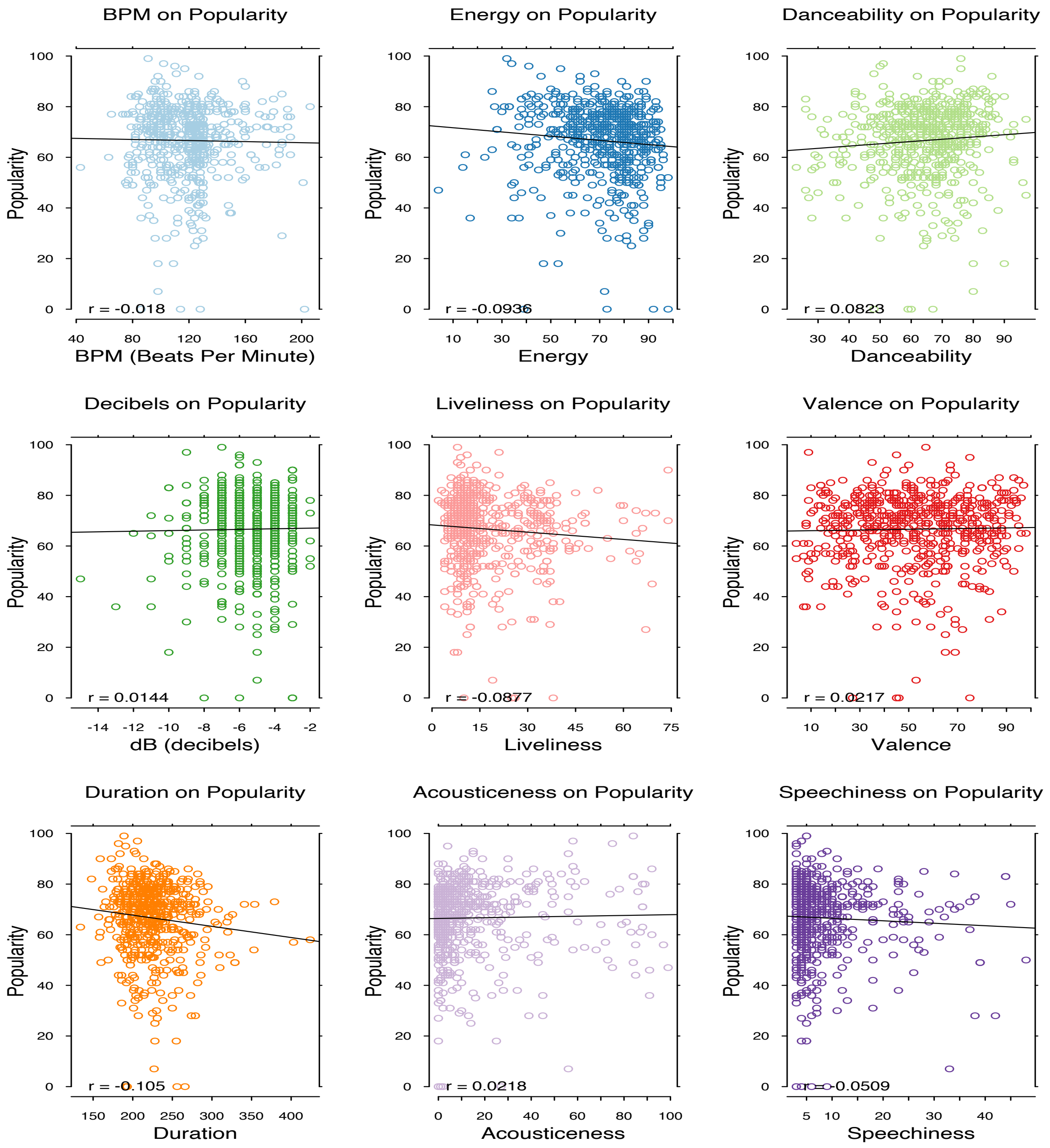

Genre by number of popular songs

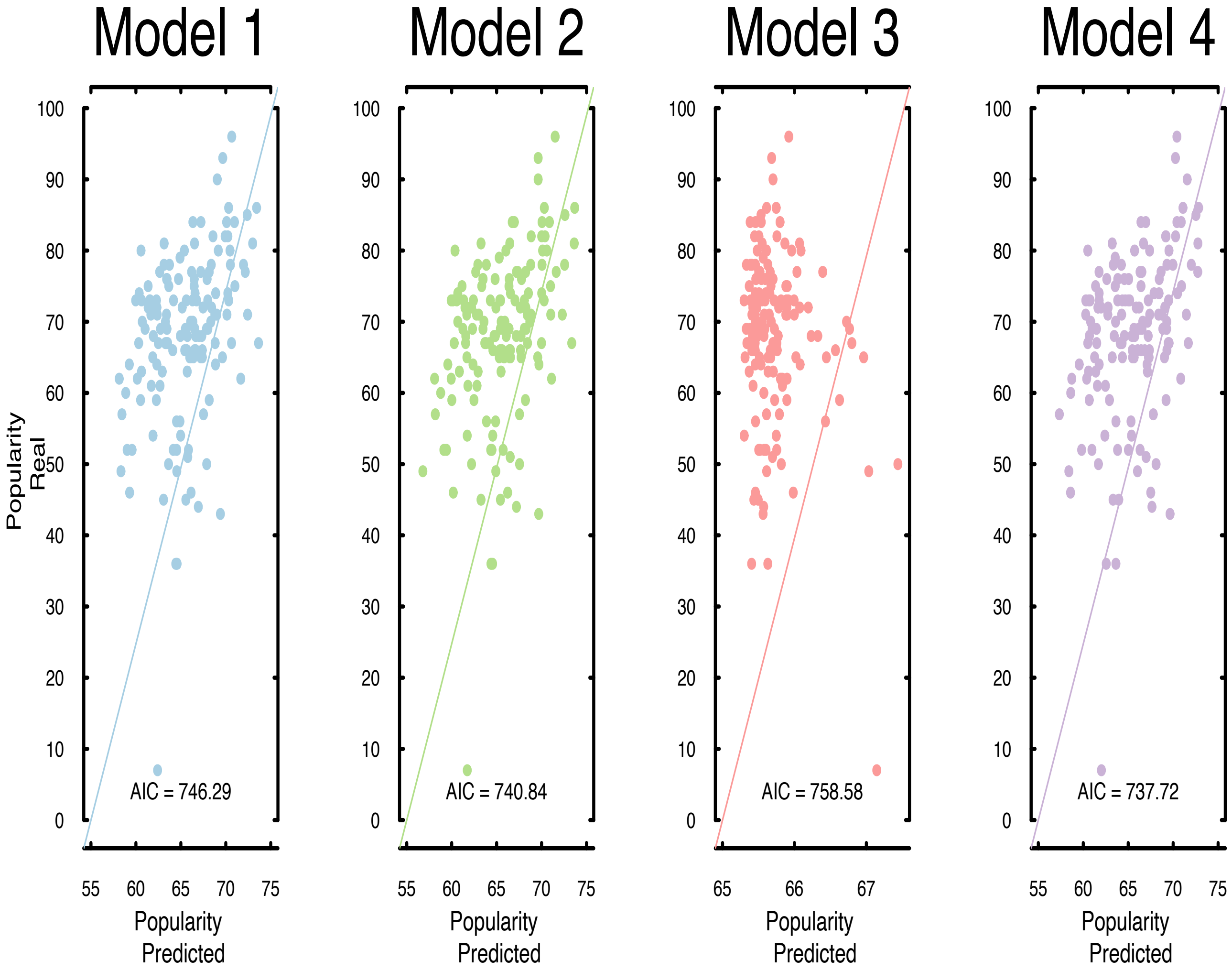
Katy Perry is the top artist from 2010-2019 with 17 popular songs. In the 2nd figure dance pop is the top genre with 330 songs on the top songs chart. It is clear to see that pop music dominates the "popular" music category, and this is because pop is a very general category that a lot of songs fall under. For example, the artists that fall under the pop category range from Alicia Keys to Nicki Minaj to Taylor Swift. Other genres like rock or country are not seen much on these charts as there is a smaller audience when it comes to listening to a more specific category of music. The correlation matrix on the next page was not significantly helpful towards figuring out what variables made a song popular, however it was helpful in other ways. If I had not made the correlation matrix, I would have not noticed how much of an impact year has on popularity. I decided to go on Spotify's website to see if I could find any more information about how popularity is calculated, and I found that year largely impacts it because a song is more popular depending on how many plays it has and how recent those plays are. I also noticed that danceability had a positive correlation with popularity and this explains why "dance pop" is the number one genre from 2010-2019. Some correlations just made sense such as energy and valence as most happy songs have high energy. It is interesting to see that energy has a negative correlation on popularity whereas valence has a positive one. Another one that caught my eye was danceability and valence having a strong positive correlation. Unlike energy and valence, danceability and valence both have a positive effect on popularity. On the next page are two graphs: the correlation matrix and the year by popularity scatterplot.


year
bpm
nrgy
dnce
dB
live
val
dur
acous
spch
pop
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

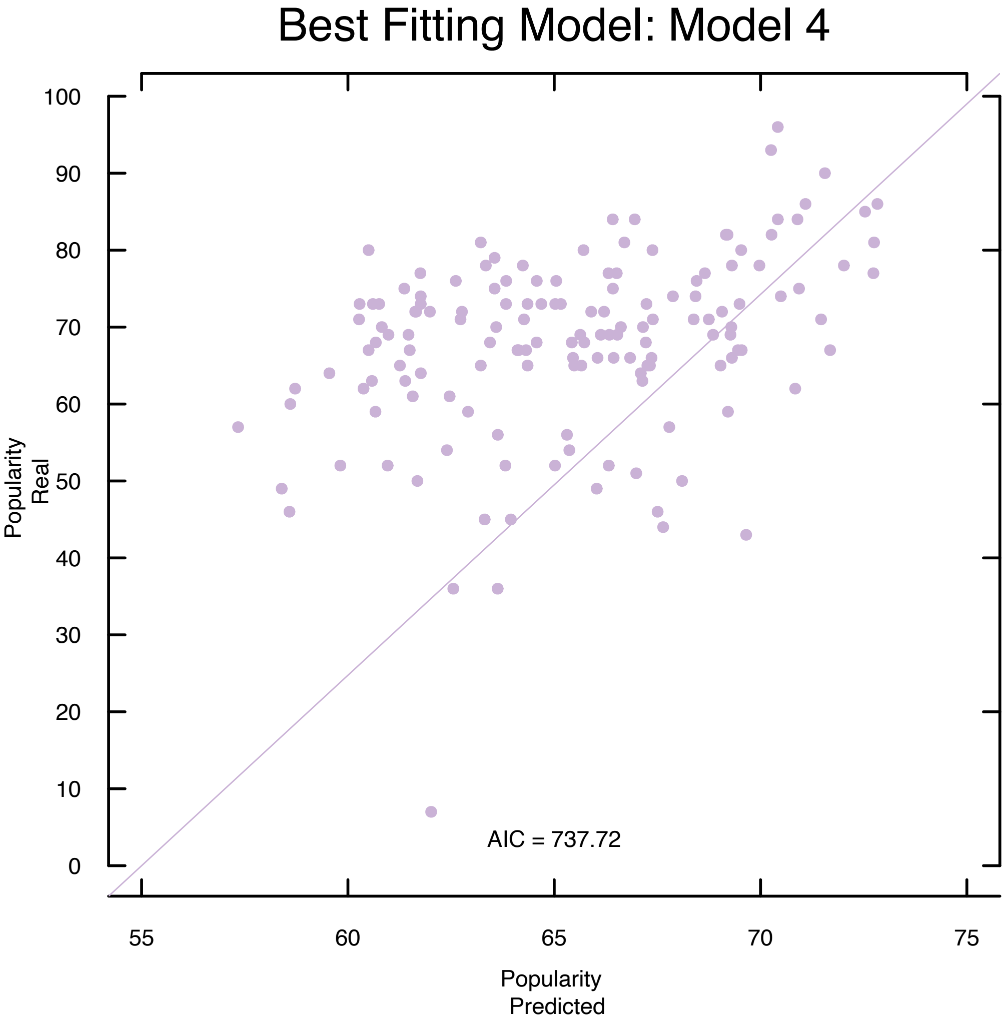

Year on Popularity
Popularity
Year
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

In this figure, as the year becomes more recent the popularity goes up. This is due to the fact that the algorithm measures song popularity by how many times the track was played with how recent those plays are. Although not all years follow this trend, such as from 2010-2011 and 2017-2018 the popularity goes down, it is easy to see that in 2019, the last year in the dataset, popularity increases to reach 100. None of the other years go to 100, the closest that it reaches is 85-90 in 2013. The most popular song in the whole dataset is “Memories” by Maroon 5, which came out in 2019. This song has 76 danceability out of 97, which is pretty high, a valence of 57 out of 74, which means it is more towards the upbeat side, and has an acoustic level of 84 out of 99. Looking at the top song gave me a good idea of what variables I needed to use to build a good model. I still wanted to plot them all though to see how they were distributed on a grand scale.





The first model contains all of the predictor variables I am using so: bpm, speechiness, acousticness, danceability, valence, energy, and year. The second model has bpm, danceability, energy, valence, and year and performs slightly better than model 1. Model 3 contains speechiness, acousticness, and valence and I put it together specifically to see the effect that speechiness and acousticness had on popularity. Acousticness increases popularity slightly, whereas more words spoken in a song decreases popularity. Model 3 is the worst model as the variables are pulling in opposite directions, and year is not included which makes a big impact. Model 4 is the model I chose, due to its low AIC score and it contains danceability valence and year. This leads me to my conclusion with the best model on the next page.



Model 4 is the best fitting model according to AIC. Model 4 contains danceability, valence, and year. The more danceable a song is, with a higher valence or more positive mood, and the more recent the year, the higher the popularity will be. In conclusion, in this dataset year has the greatest impact on popularity, with danceability and valence following. Variables that encourage positivity and happiness, specifically danceability and valence also contribute to popularity in a song. It is no wonder that people want to feel uplifted when listening to music, and it is on those rare occasions such as a bad day or going through a breakup that people want to listen to sad music. This is why valence is distributed very evenly, however there are more popular songs with higher valence. Another thing to note is that too much of something like energy will actually have a negative impact on popularity as there is a very delicate balance. These songs are widely played on the radio, and at venues like clubs which increases popularity. Lastly, I do think it’s possible to predict which songs will become hits but with more extensive research as music changes with time but repeats the same trends.