Capstone Project 1 - Exploratory Data Analysis
Spotify Audio Features

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

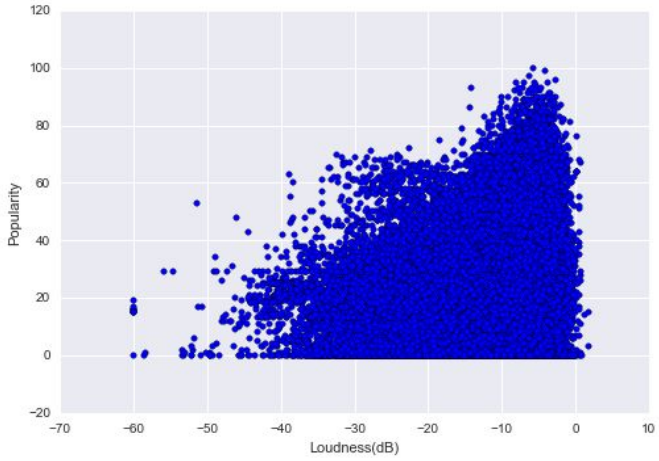
In this phase, we try to test whether there is significant correlation between some of the variables in the dataset and the outcome variable (popularity) using hypothesis testing and confidence intervals. The attempt to answer the following questions:

- Is there a significant correlation between popularity and Loudness(both numerical variables - Correlation test performed)
- Is there a significant correlation between popularity and Danceability(both numerical variables - Correlation test performed)
- Is there a significant correlation between Popularity and Instrumentalness(both numerical variables - Correlation test performed)
- Is there a significant correlation between popularity and track duration?(both numerical variables - Correlation test performed)
- Is there a significant correlation between popularity and Liveness?(both numerical variables - Correlation test performed)
- Does 4-beat time signature make a song more popular?(one categorical, the other numerical variable - two sample test performed)

Note 1: From our data storytelling (previous phase), we know that there is correlation between some of the independent variables (e.g. loudness and energy). But most of these correlations are obvious (i.e. given by definition) and are not of particular interest for testing here.

Note 2: All of the hypothesis tests listed above are done using the bootstrap method. Data are simulated under the null hypothesis, a test statistic is calculated, and finally the p_value is calculated by finding the probability of getting the observed statistic, or more extreme, under the null.

Popularity and Loudness:



H0: There is no correlation between loudness and popularity

Ha: There is positive correlation between loudness and popularity (Positive because having seen the scatter plot, testing for positive correlation seems appropriate)

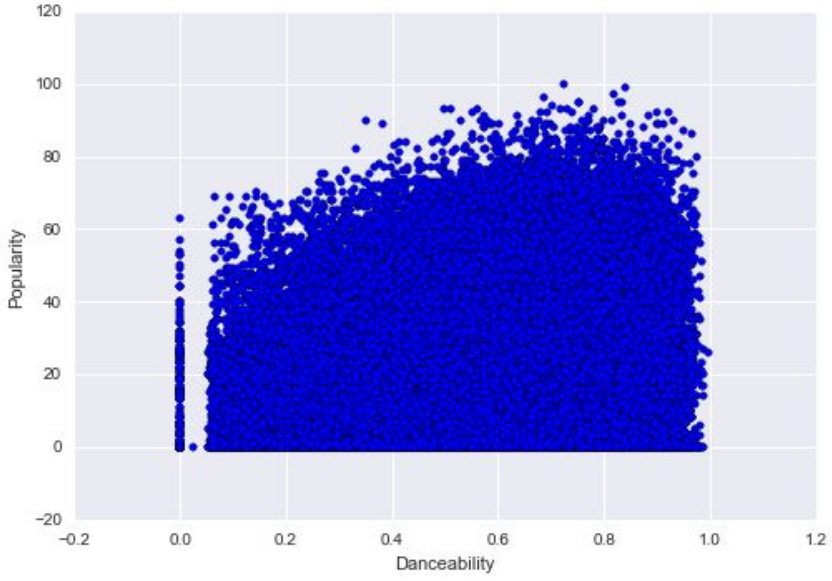
$\alpha = 0.05$

Observed correlation: 0.235

P_value ≈ 0.0

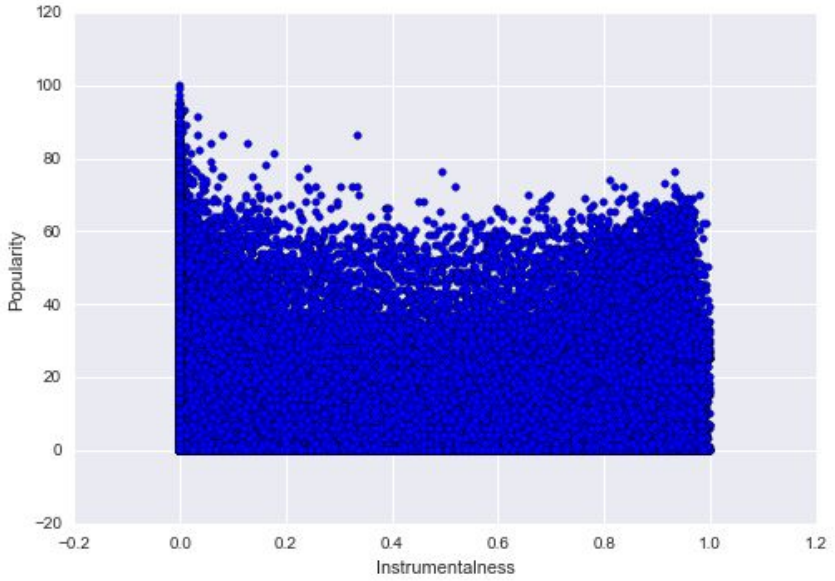
Conclusion: P_value is small. We conclude that there is a positive correlation between popularity and loudness.

Popularity and Danceability:



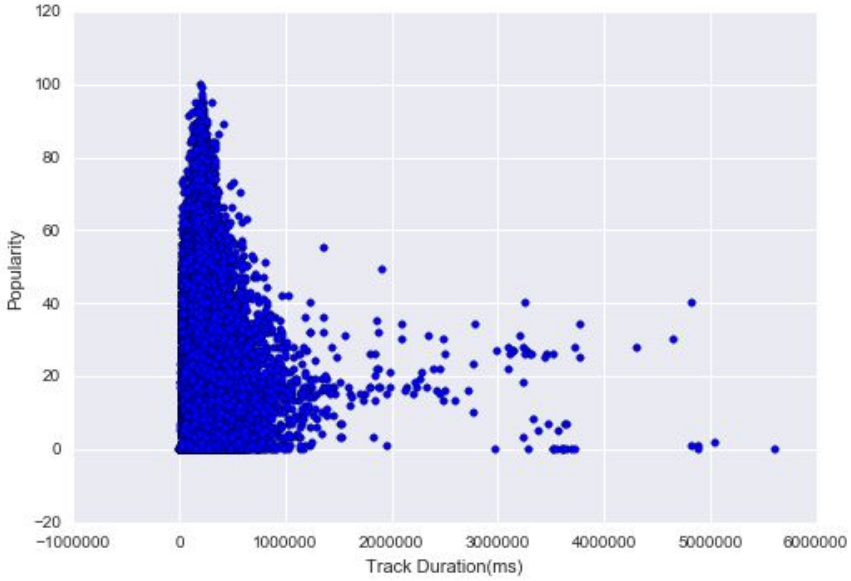
H0: There is no correlation between danceability and popularity
Ha: There is positive correlation between danceability and popularity (Positive because having seen the scatter plot, testing for positive correlation seems appropriate)
α= 0.05
Observed correlation: 0.134
P_value ≈0.0
Conclusion: P_value is small. We conclude that there is a positive correlation between popularity and danceability.

Popularity and Instrumentalness:



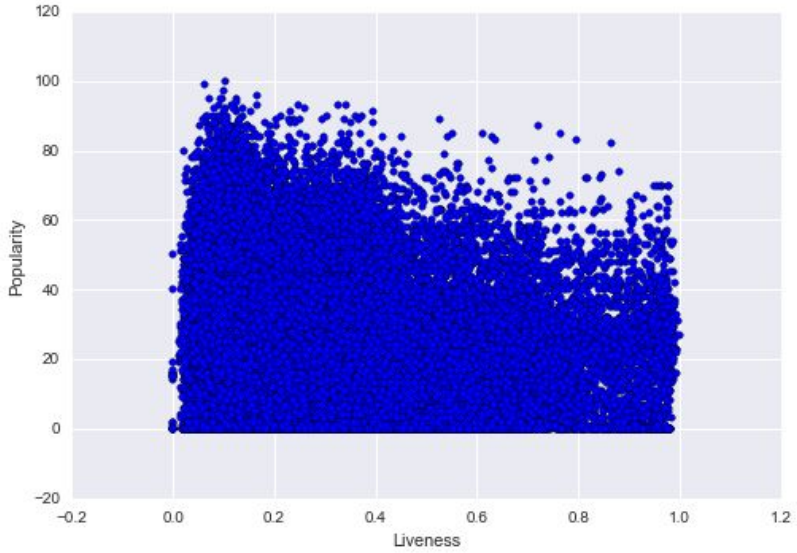
H0: There is no correlation between instrumentalness and popularity
Ha: There is negative correlation between instrumentalness and popularity (negative because having seen the scatter plot, testing for negative correlation seems appropriate)
α= 0.05
Observed correlation: -0.21
P_value ≈0.0
Conclusion: P_value is small. We conclude that there is a negative correlation between popularity and instrumentalness.

Popularity and Track Duration:



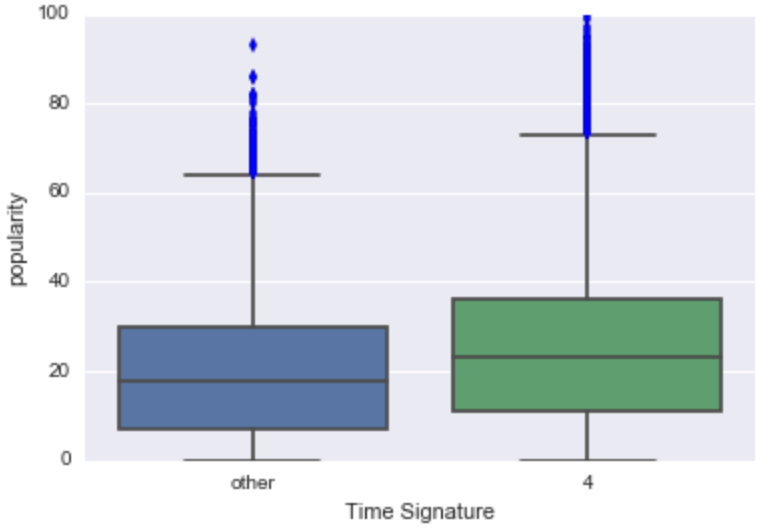
H0: There is no correlation between track duration and popularity
Ha: There is negative correlation between track duration and popularity (negative because having seen the scatter plot, testing for negative correlation seems appropriate)
α= 0.05
Observed correlation: -0.009
P_value $\approx$0.001
Conclusion: P_value is small. We conclude that there is a negative correlation between popularity and track duration.

Popularity and Liveness:



H0: There is no correlation between liveness and popularity
Ha: There is negative correlation between liveness and popularity (negative because having seen the scatter plot, testing for negative correlation seems appropriate)
$\alpha =$ 0.05
Observed correlation: -0.028
P_value $\approx$ 0.0
Conclusion: P_value is small. So we reject the null and conclude that there is a negative correlation between popularity and liveness.

Does 4-beat time signature make a song more popular?



H0: There is no difference in popularity bewteen tracks with 4-beat time signature and other time signatures
Ha: Tracks with 4-beat time signature are, on average, more popular than other time signatures
α = 0.05
Observed difference in means (mu_four - mu_other): 4.56
P_value $\approx$ 0.0
Conclusion: Given the small p_value, we reject the null and conclude that 4-beat time signatures, on average, are more popular than other time signatures