Energy of Electronic Music

Data 8: Final Project

Tarsus Lam & MaryEllen Miyashita

Introduction:

Free Music Archive

- 559 tracks, a random subset of FMA introduced at 2017 Music Information Retrieval Conference
- Three tables with categorical and numerical variables
    1. tracks
    2. genres
    3. features

Questions

1. Hypothesis Testing - ***Are there track features that uniquely determine the Electronic genre?***

- **Hypothesis**
  Null: There is no relationship between any of the features and whether the track is Electronic
  Alternative: There is a relationship between at least one of the features and whether the track is Electronic

- **A/B Testing**
  Difference of means, p-value of 5%

2. Prediction - ***Can these features be used to classify a track as Electronic or other?***

- **K-Nearest Neighbor**, classification method

Exploration

Joined Table

track_genre	acousticness	danceability	energy	instrumentalness	liveness	speechiness	tempo	valence
Jazz	0.235506	0.438672	0.487752	0.716122	0.0703593	0.0472978	120.79	0.650452
Rock	0.981657	0.142249	0.912122	0.967294	0.36351	0.087527	91.912	0.0343253
Folk	0.991813	0.461855	0.543751	0.964922	0.137006	0.0256877	93.945	0.758632
Rock	0.77377	0.552026	0.251328	0.568976	0.110743	0.0506326	117.247	0.356984
Rock	0.335481	0.390263	0.0210674	0.937508	0.0890457	0.0414906	60.382	0.0399321

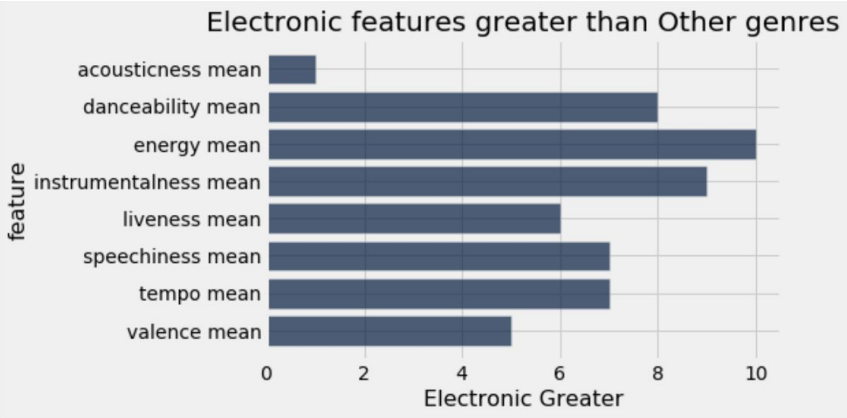
... (454 rows omitted)

Aggregated Table

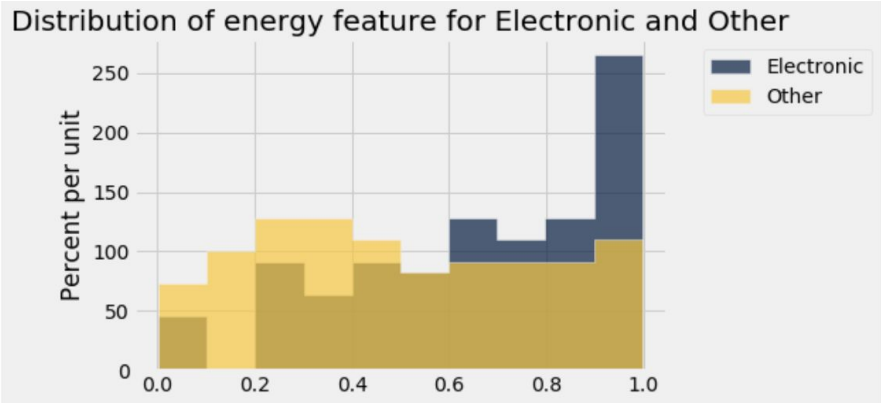
track_genre	count	proportion
Blues	5	0.0108932
Classical	5	0.0108932
Electronic	110	0.239651
Folk	33	0.0718954
Hip-Hop	46	0.100218
Instrumental	3	0.00653595
International	5	0.0108932
Jazz	16	0.0348584
Old-Time / Historic	8	0.0174292
Pop	13	0.0283224
Rock	215	0.46841

Exploration

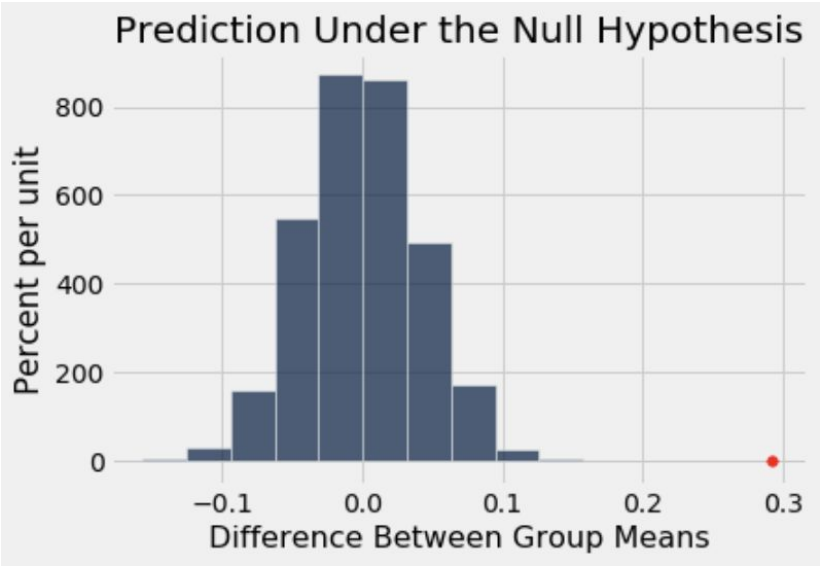
Qualitative plot:



Quantitative plot:



Hypothesis Testing - A/B test



1. **Observed test statistic = 0.2921**
   - Found by taking the mean of all *other* tracks' energy feature
2. **Find the simulated test statistic**
   - 5000 differences of means using shuffled tables
3. **Compare to p-value of 5%**
   - **0.0 < 5%,** statistically significant → **reject null hypothesis**

Prediction - Classification

K-NEAREST NEIGHBORS

1. **Separate shuffled data into train and test set (80:20)**
2. **Find Euclidean distance between the train and test rows**
   - First classifier based on each tracks' only using energy
   - Second on acousticness, danceability, energy, and instrumentalness
3. **Use distances, in ascending order, to find k-nearest neighbors**
   - Classifier uses first K neighbors to produce a guess
   - Smaller k of 3 → higher accuracy because high variance of dataset
4. **Compare number predicted correct using classifier to total number of tracks for accuracy**
   - Accuracy of 55% using just energy
   - Accuracy of 83% using the four features

Conclusion

- Electronic genre has a lower energy compared to the other genres, reject the null hypothesis
- Predict a track is Electronic based on its acousticness, danceability, energy and instrumentalness with 83% accuracy
- Limitations:
  - A/B test found an association, not precise prediction
  - Number of tracks varies for each genre
  - Method of collection is unknown