STATISTICAL DATA ANALYSIS

PROJECT - GEOGRAPHICAL ORIGIN OF MUSIC

Group 12 Team members:
1. Bhavani. T (S20180020251)
2. Navyasri. M (S20180020222)
3. Rohith. A (S20180020238)
4. Soumya.G (S20180020209)

1 Abstract

Geographical ethnomusicology gives the distribution of music from around the world. The research into the arts has been always based on the subjective judgment of human critics. The use of data mining tools to understand art has great promise as it is objective and operational. We performed exploratory data analysis and checked for outliers and performed normality check,factor analysis, PCA. We build different estimators SVM, KNN-classifier, Random Forest Classifier , Gradient boosting, logistic regression and also a multi layer neural network.

2 Introduction

The dataset consists of audio features that were extracted using MARSYAS. The features that we are using can be seen as the key audio features of a song. We need to predict the geographical region of a song based on these features. The data does not include geographical regions, but rather latitudanal and longitudinal coordinates.The task associated with the data is to predict the geographical origin of music.

3 Dataset description

- In 'default_features_1059_tracks.txt' file, out of 70 columns, the first 68 columns contain audio features and the last two columns describe the latitude and longitude of the place of origin of music.This latitude and longitude describes the capital city of the country.
- In 'default_plus_chromatic_features_1059_tracks.txt' file, the first 116 columns contain audio features, which includes chromatic features and the last two columns describe the latitude and longitude of the place of origin of music.
- The data is converted into a format by replacing latitude and longitude values with the respective country names. The corresponding country-wise hosts were plotted and observed that most of the tracks were taken from India.
- The distribution of data among different countries can be seen in below figure.

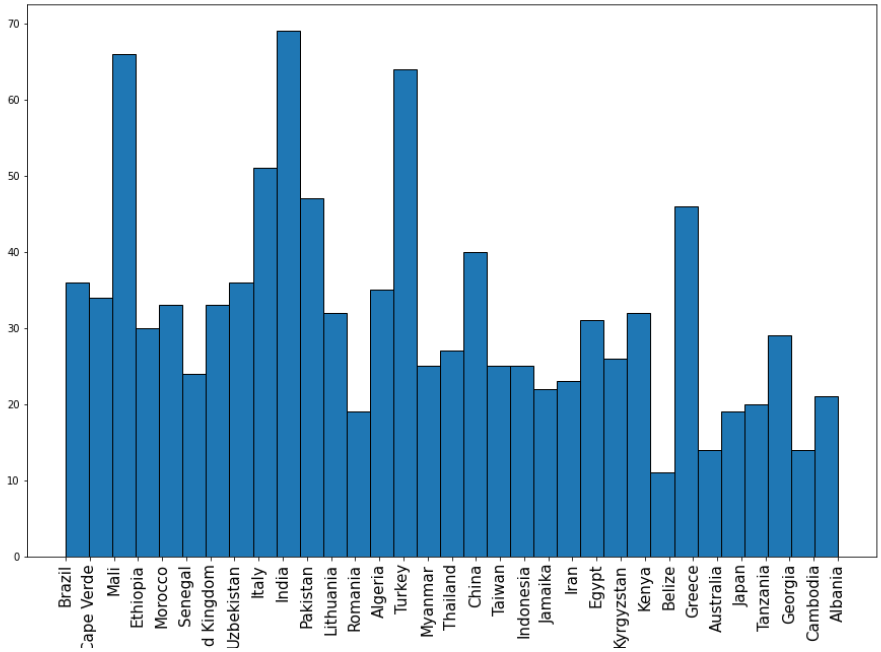


Figure 1: Distribution of music data among different countries

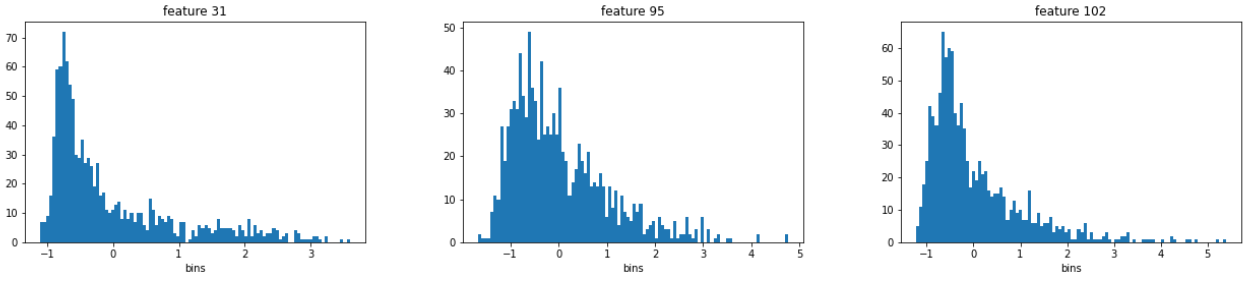
4 Analysis

4.1 data cleaning

- First of all the dataset shouldn't contain any misisng values, so we checked for it. But there are no null values in the entire datsets.
- The feature vectors in the dataset already have a mean zero and standard deviation 1.

4.2 data visulatisation

- Histogram of some of the features can be seen in below figure.



(a) feature 31 (b) feature 95 (c) feature 102

Figure 2: histograms of different features

- All the different locations were clustered using kmeans clustering and plotted latitude vs longitude as shown below.

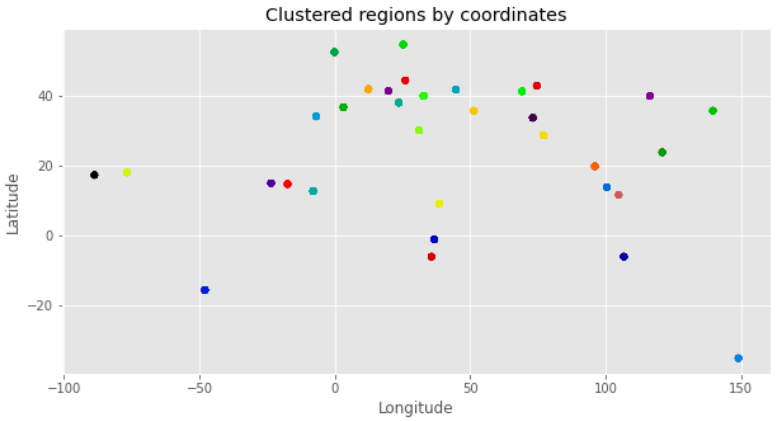


Figure 3: Latitude vs Longitude

- Now using DBSCAN which is Density-Based Spatial Clustering of Applications with Noise, a clustering method that is used to separate clusters of high density from clusters of low density. The clusters separated are plotted as shown in the below figure.
- Each colour represents a cluster. If we observe more denser clusters are at left bottom position in the plot.

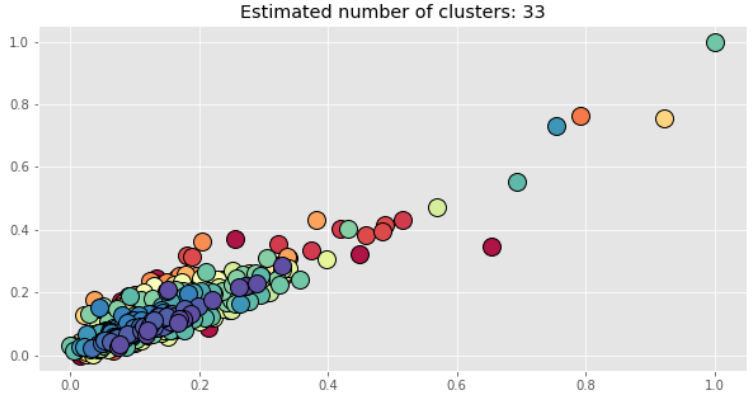
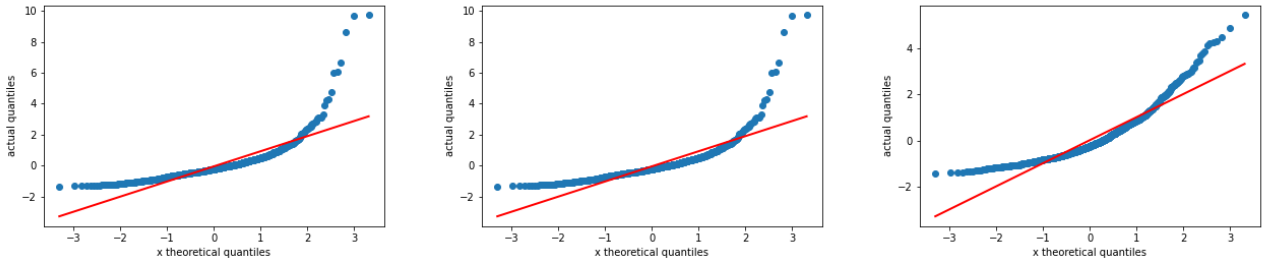


Figure 4: Different clusters separated by DBSCAN

- Univariate test using Q-Q plots is performed to check the normality condition. The feature vectors are all normally distributed with mean nearly zero and standard deviation as nearly one. Some of the plots can be seen in below figure.



(a) feature 1 q-q plot (b) feature 30 q-q plot (c) feature 100 q-q plot

Figure 5: q-q plots

4.3 Removing Outliers

- Outliers can be observed for some features in the below figure.

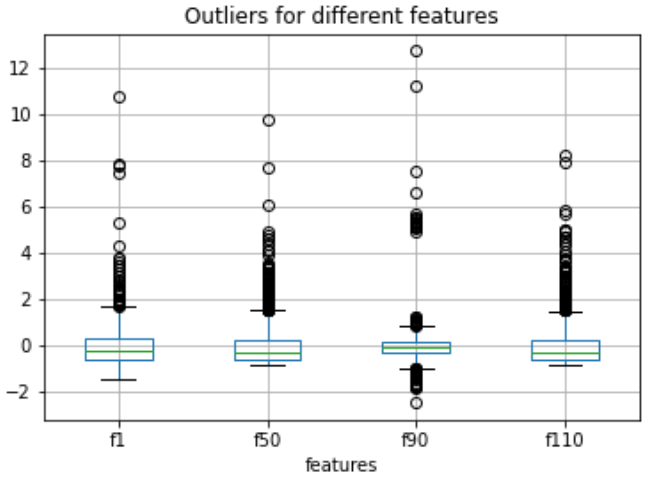


Figure 6: Boxplot representing outliers

  - Z- score: We considered if the Z- Score is >3 or < −3 then the observation in the data is said to be an outlier. An outlier is an observation that is abnormal in a random sample of the population data. So by calculating the z-scores we detect the outliers and remove it from the data.
    Z-scores were calculated for outliers detection
  - Quartiles: First Quartiles-Q1 (at 0.25) ,third-quartile- Q3 (at 0.75) were computed from which Inter quartile range is calculated. Any values greater than Q3+3*IQR or less than Q1-3*IQR are identified as outliers and are removed from the dataset.

- Later, the data sets after the removal of outliers using Z-scores and Quartiles are in for feature selection models (PCA and Chi-square analysis). The accuracies are compared in section-6 using the estimators.

4.4 Correlation

- Correlation heatmap is plotted for all the features. Generally a correlation matrix is a table showing correlation coefficients between variables.
- We can observe there is strong correlation between some features when chromatic features are included.

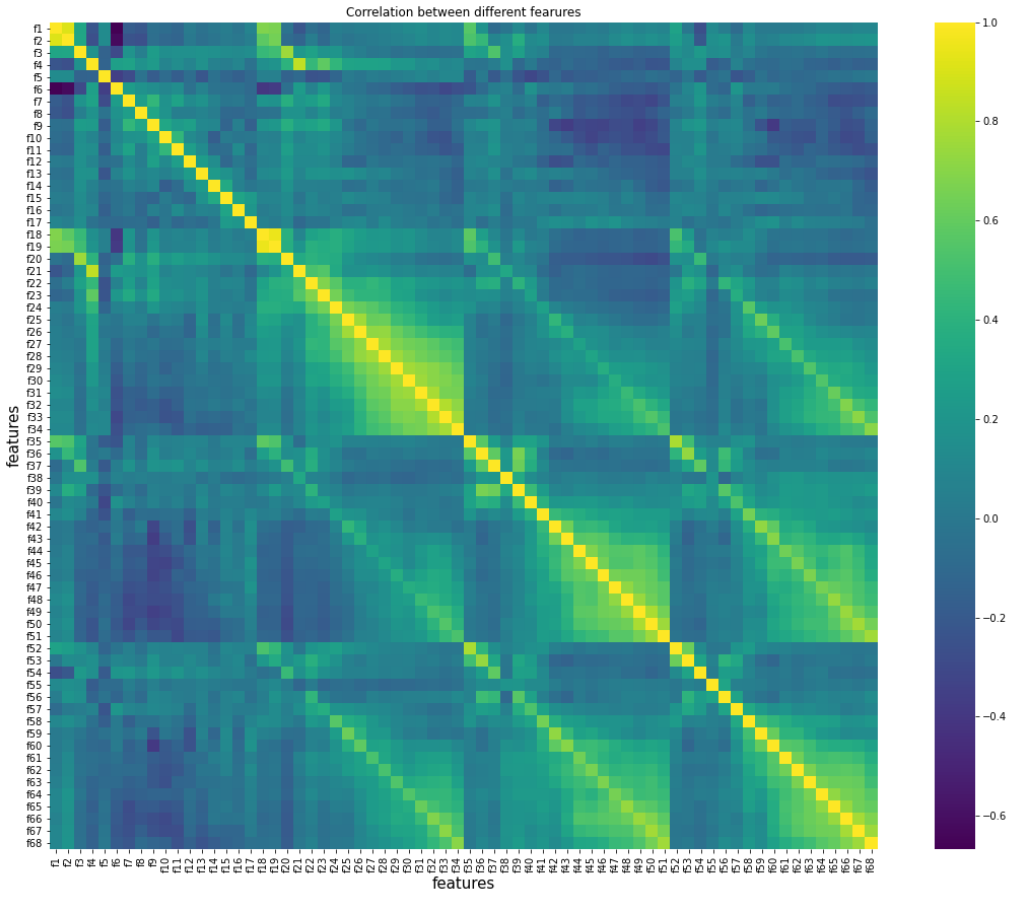


Figure 7: The correlation matrix for datset without chromatic features

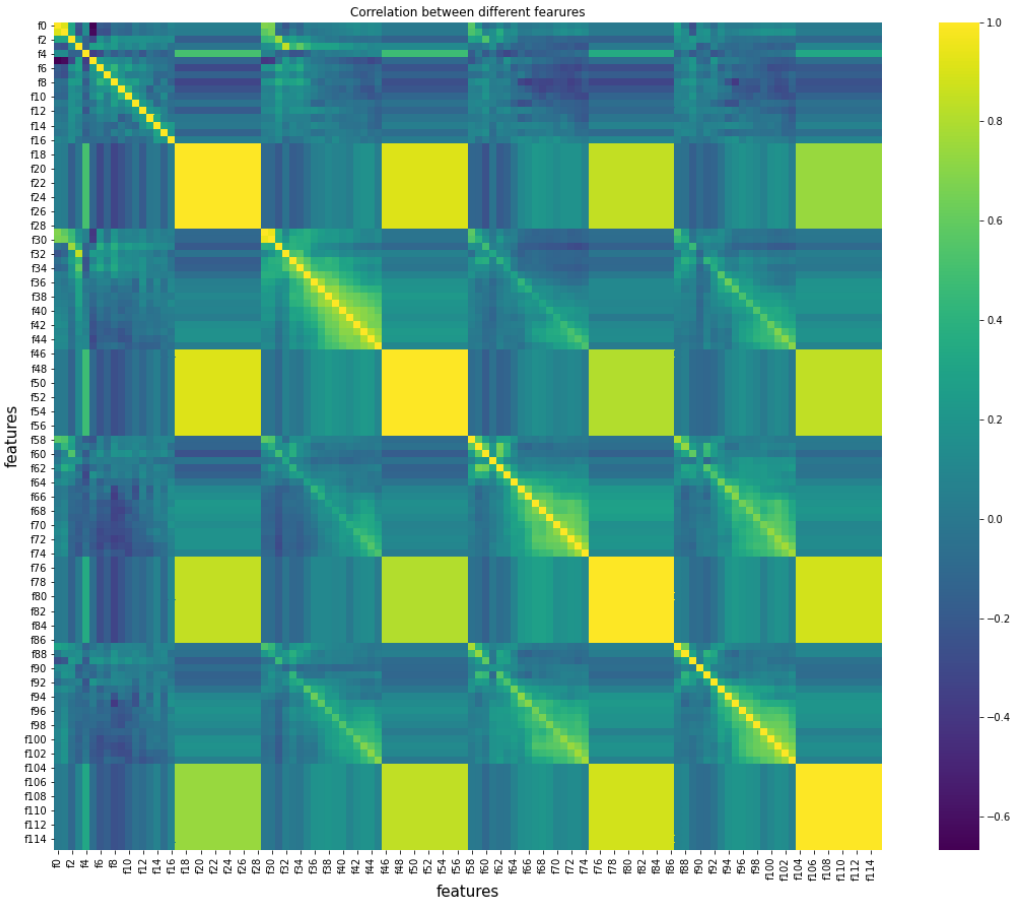


Figure 8: The correlation matrix for dataset with chromatic features

- Pearson's correlation coefficient test is performed on some of the features. It is the test statistics that measures the statistical relationship between two continuous variables. It is based on the method of covariance.

4.5 ANOVA

- Analysis of variance is performed between some features. For that we have split data set into two samples by randomly shuffling the data.
- If p is greater than 0.05, then it is considered to have same distribution.
- All the features seems to have same, because the main reason is, the dataset itself is normalised with variance nearer to 1.

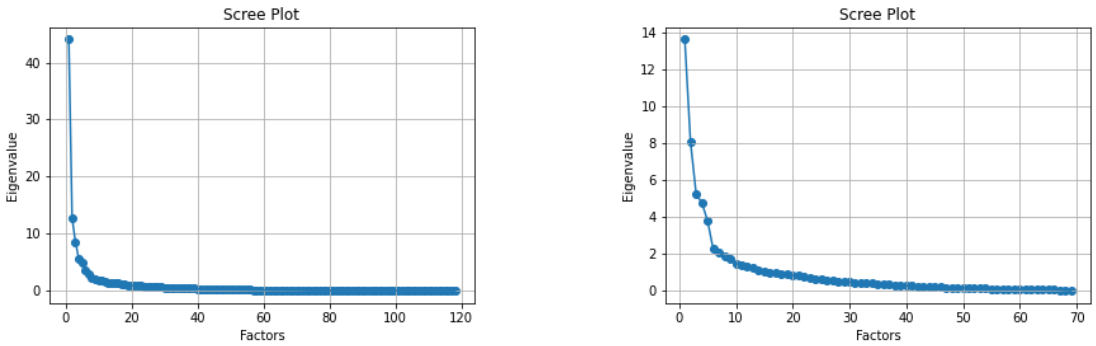
4.6 Feature selection

- One implemented process for feature selection is using chi-square scores. Chi-sqaure scores for each of the features are calculated and the scores ¿20 are considered as better features.

- On the other hand, the results (accuracies from the estimators) using these features are not as good as the values obtained when all the features are taken into account.
- PCA and factor analysis are statistical methods, used to reduce the dimensionality of the feature set.

4.7 Factor Anlaysis

- Factor Analysis is used to deal with data sets where there are large number of observed variables that are thought to reflect a smaller number of underlying variables. The key concept is that multiple observed variables have similar patterns of responses because they are all associated with the latent (not directly measured) variable.
- Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.
- The eigen value is the measure of how much of the variance of the observed variables a factor explains. Any factor with eigen value>=1 explains more variance than a single observed variable
- The factors that explain the least amount of variance are generally discarded. So using the eigen values of correlation, we have selected "some" features from the dataset which are assumed to be relevant based on this analysis.
- we can observe the below scree plot which shows the eigen values for different factors for both datasets with chromatic and without chromatic.



(a) with chromatic features

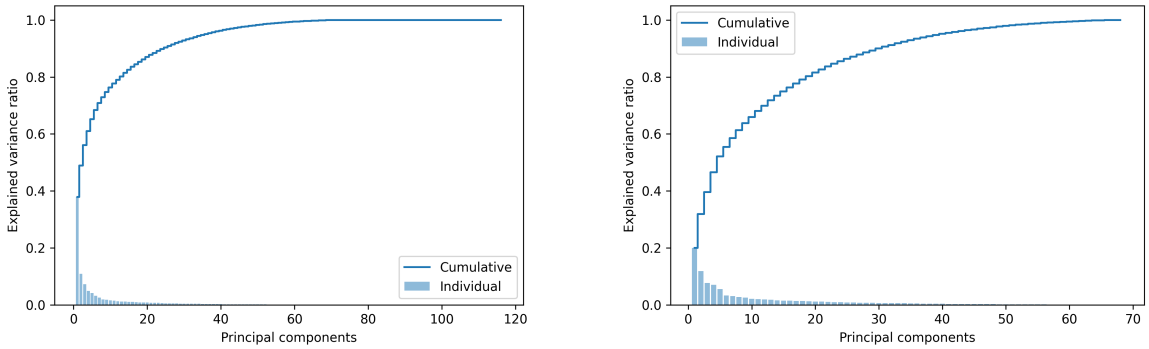
(b) without chromatic features

Figure 9: scree plots

- So we can see that the eigne values increased for the data with chromatic features, which tells it observed more variance.

4.8 PCA

- Principal component analysis: PCA uses an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables. These are called Principal Components.
- First covariance matrix is computed. Then eigen decomposition is performed on the covariance matrix and eigenvalues and vectors are computed. The PC's were sorted out based on the corresponding eigenvalues. And then the cumulative sum is computed and plotted as shown below.



(a) with chromatic features

(b) without chromatic features

Figure 10: ratio of variance explained

- From the plots, 95% of the variance is within the first 30 principal components for the data without chromatic features. Wherease 95% of the variance is within the first 40 principal components with chromatic data.

5 Models

- The datasets after removing the outliers are used for further analysis. This data set is divided into training and testing sets.
- Now for both the datasets, i.e with chromatic features and without chromatic features, the classification analysis is implemented separately using different estimators and the accuracies of both data sets for different estimators were compared.
- Also the obtained features from feature selection were used for different estimators.
- The estimators used are SVM, Logistic classifier, KNN, Random forest classifier, Decision tree classifier, multilayer Neural network and Gradient Booster.

S.NO	Model	Parameter	value
1	SVM-linear	kernel,C	'poly',1
2	SVM-Gauusian	kernel,C	'rbf',1
3	KNN	n beighbours	1
4	RFC	max depth	23
5	GB	max depth, learning rate	10,0.1
6	Decision tree	max depth	
7	Neural network	layers, activation, optimiser	3, 'ReLU','adam'
8	logistic regression	C	1

Table 1: parameters used for different estimators

6 Results

- Here we can see the performance of different estimators for both the datasets and using different feature selection approaches.
- The first two tables represent the model accuracies performed using features selected from PCA analysis. Both are without and with chromatic features respectively.

S.NO	Model	Accuracy
1	SVM-linear	34%
2	SVM-Gauusian	36%
3	KNN	35%
4	RFC	34.8%
5	GB	20.8%
6	Decision tree	15.4%
7	Neural network	41%
8	logistic regression	39.6%

Table 2: classification analysis without chromatic features for '30' features selected using PCA

S.NO	Model	Accuracy
1	SVM-linear	33%
2	SVM-Gauusian	32%
3	KNN	36%
4	RFC	47.8%
5	GB	19.6%
6	Decision tree	14.2%
7	Neural network	45%
8	logistic regression	39.9%

Table 3: classification analysis with chromatic features for '40' features selected using chi2 test

- These tables represent the model accuracies performed using features selected from chi2 test. In this test the first 30 features according to their scores are selected. Both are without and with chromatic features respectively.

S.NO	Model	Accuracy
1	SVM-linear	37%
2	SVM-Gauusian	38%
3	KNN	40%
4	RFC	45.3%
5	GB	23.7%
6	Decision tree	26.3%
7	Neural network	44%
8	logistic regression	36%

Table 4: classification analysis without chromatic features and features selected using chi-square

S.NO	Model	Accuracy
1	SVM-linear	21%
2	SVM-Gauusian	17%
3	KNN	26%
4	RFC	26.4%
5	GB	21.3%
6	Decision tree	19.7%
7	Neural network	28%
8	logistic regression	21.3%

Table 5: classification analysis with chromatic features and features selected using chi-square test

- From the computed results of estimators, feature selection based on PCA has shown better accuracy results when compared to Chi-square test for feature selection.
- When compared between the two datasets, PCA based analysis shown that chromatic features included data set has better results that the data whith no chromatic features. Whereas, Chi-square has better results for the data with no chromatic features.

7 Conclusion

So we are concluding that classifying the songs by geographical locations using the given features is unworthy. Even the instances for particular locations are very less. After identify the outliers and removing them, the dataset even becomes smaller. The dataset doesn't has the feature labels, so we dont know what data we are dealing with. Which suggests that the features extracted are not enough to differentiate the songs or the features are not able to capture the essential musical differences between the regions. Definitely another factors should be considered which explains the variance in song characteristics, so that we can classify the songs by their geographical locations.

References

[1] Zhou, Fang, Q. Claire, and Ross D. King. "Predicting the geographical origin of music." 2014 IEEE International Conference on Data Mining. IEEE, 2014.

[2] http://archive.ics.uci.edu/ml/datasets/geographical+original+of+music