Music Analysis and Genre Classification

Team Members: Yiqi Rao, Haosheng Wang

Why this is important: Music Genre is an important point of consideration for music lovers to choose and filter out musics of their tastes. Moreover, well-constructed genre classification models are foundations for music recommendation system, which is crucial for music listening apps to improve user experience.

Goodness

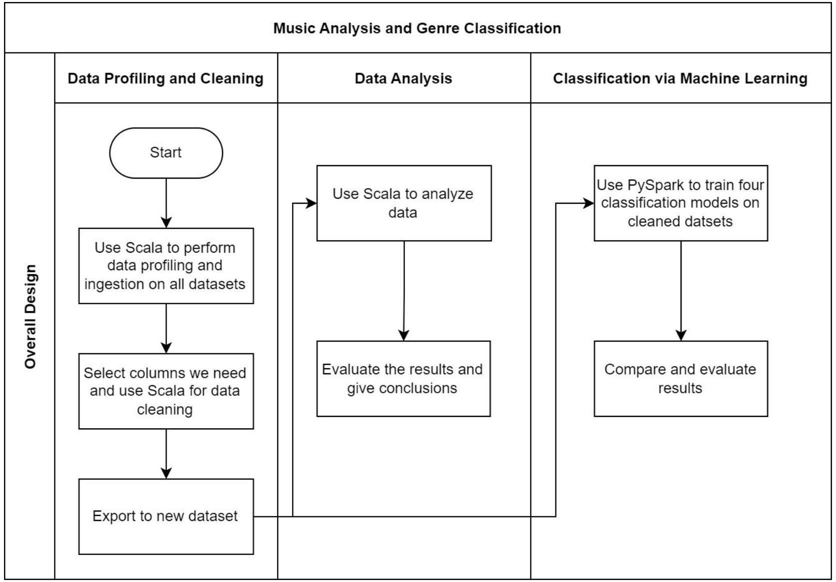
- In the beginning, we tried to analyze metadata of the tracks such as:
    - Artist preference (different artists could have different preferences over different genres)
    - Geographical effect (different location comes with different atmosphere could help form the trend of some genre more popular than other genres
- When building genre classification machine learning models:
    - We use cross validation for hyper-tuning and estimating performance of four different classification models, including decision trees, random forest, multilayer perceptron, and logistic regression.
    - We train our models respectively on two sets of music descriptors: Echo Nest and Mel-frequency Cepstral coefficients (MFCC). And MFCC is a commonly chosen and reliable music descriptor in scholarly papers which can yield rather high classification accuracy.

All Data Sources

1. tracks.csv
    a. Include tracks data and relevant information
    b. 137001 rows
2. gernes.csv
    a. Include all genres tracks may have
    b. 164 rows
3. echonest.csv
    a. Include Echo Nest music descriptors data for tracks
    b. 13128 rows
4. features.csv
    a. Include several sets of music descriptors for tracks
    b. 106575 rows

All the above data can be found on https://github.com/mdeff/fma.

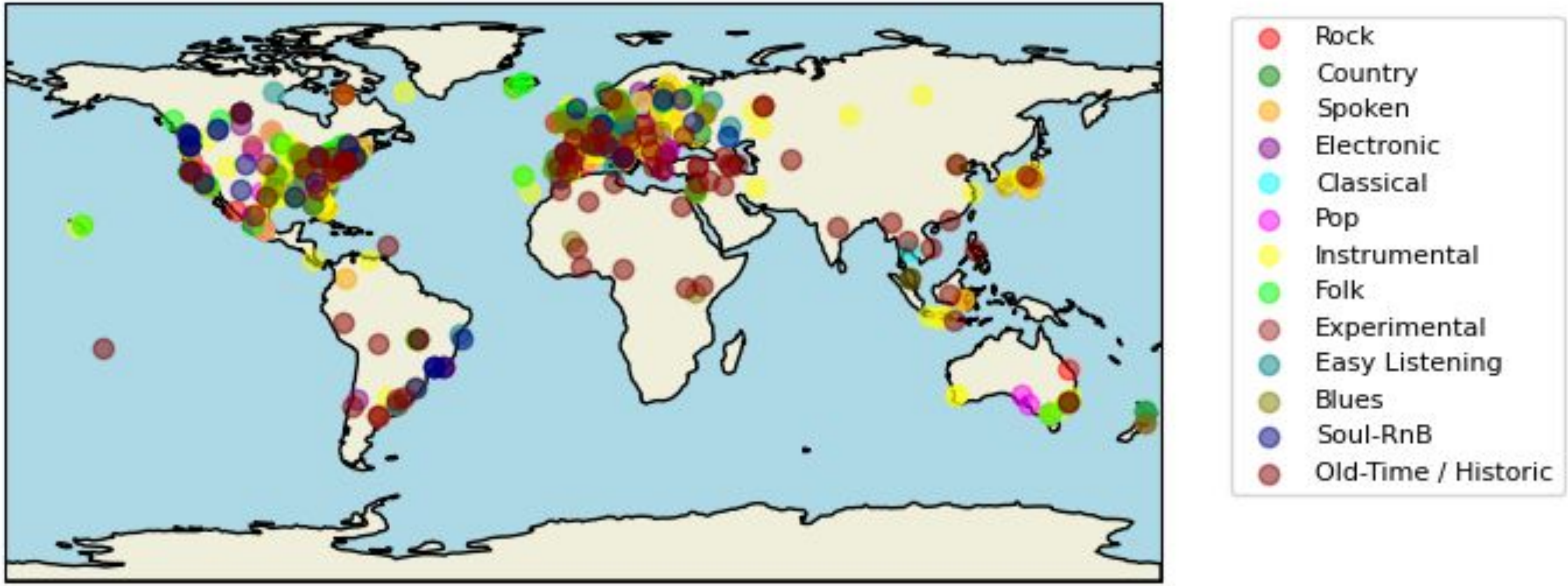
Design Diagram



Code Challenge

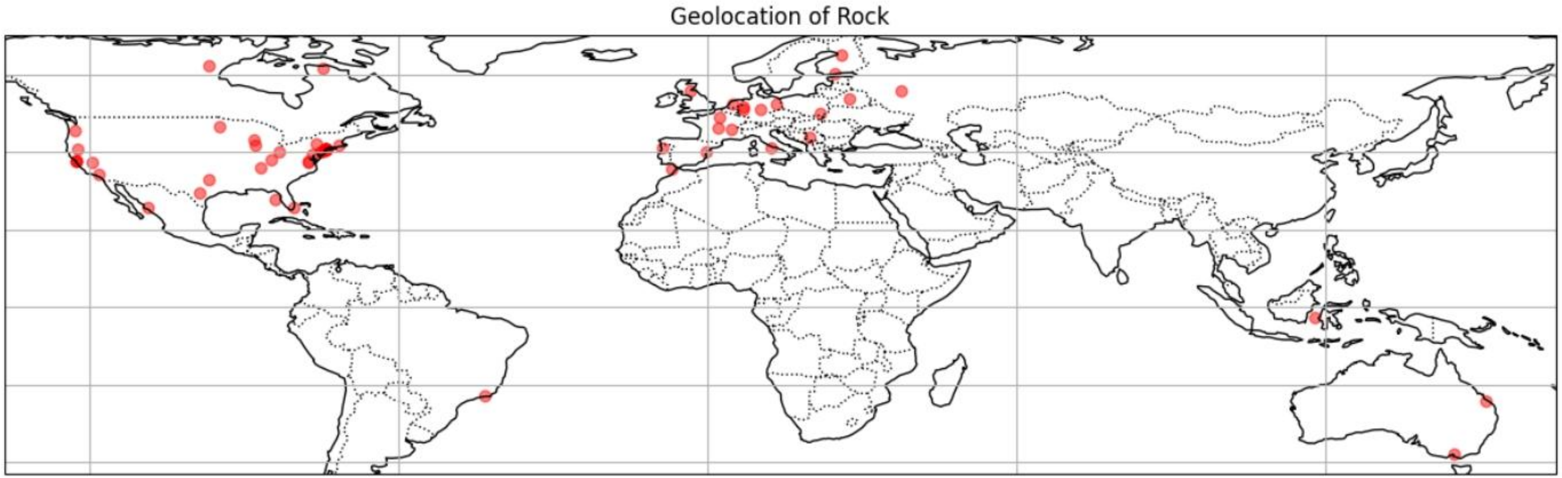
In order to analyze geographical distribution of music of different genres, we first use Apache Spark to clean the tracks.csv to get rid of those rows without valid numeric artist_longitude and artist_latitude. Then we use scala to group artist_longitude and artist_latitude into a tuple, and aggregate them into an array under the same genre using Spark SQL. And output a small txt file. We then use python to further organize the data getting around 900 points, and combined with genres.csv to get the top_genre of each genre (thus reduce the number of genres from 157 to less than 20). Then we use matplotlib and cartopy to visualize the geographical distribution of different genres. (The result in the following slides)

Geolocation of Music Genres



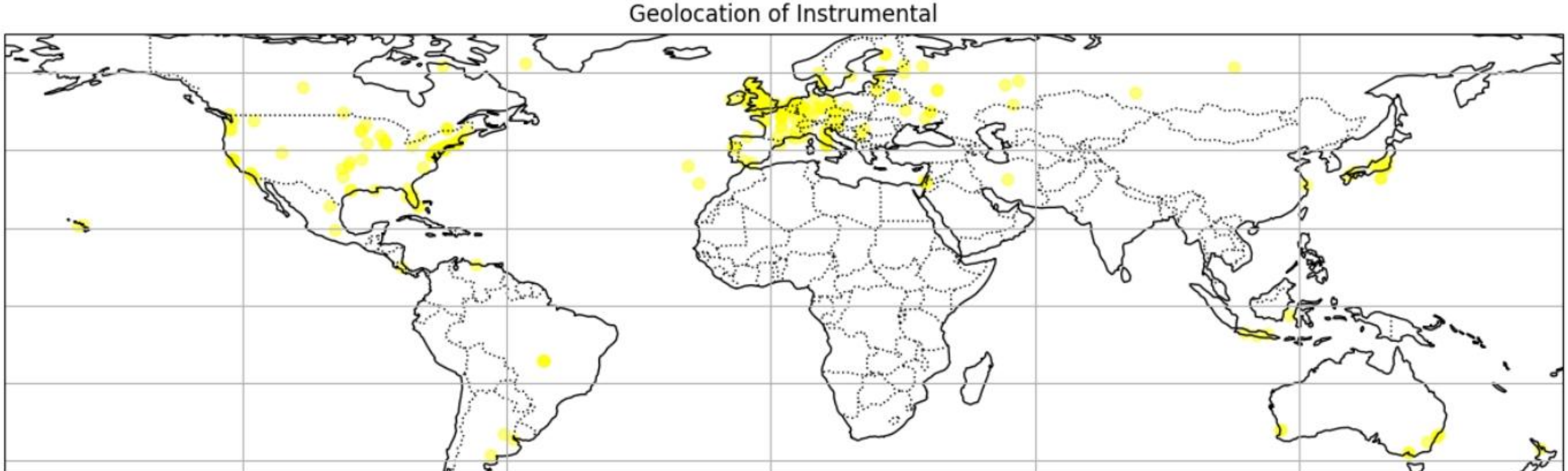
Insight 1

- Geographical distribution of Rock
    - E.g. Geographical distribution of Rock
    - We can see from the picture that the the majority of the artist of Rock music comes from Europe and North America



Insight 1 (cont.)

- Geographical distribution of Instrumental Music
    - E.g. Geographical distribution of Instrumental
    - For Instrumental music, the artists come from a range of worldwide locations—with most come from Europe, and less come from locations like America, Asia, Australia, etc.

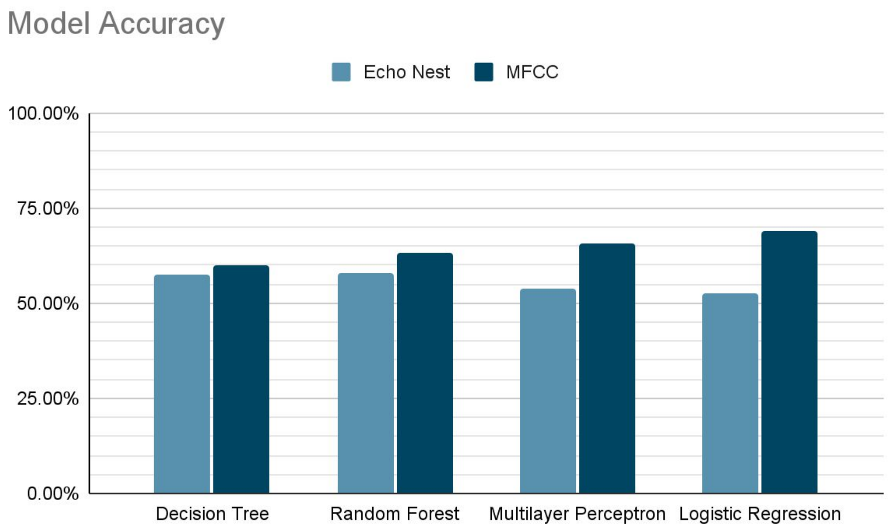


Insight 1 (cont.)

- From the previous two pictures (as well as other genres' geolocation pictures included in the article), we can see that the majority of the artists come from Europe and North America, with small differences among all the genres.
- This reflects that the majority of the music data in the dataset are collected from these areas, and thus not a solid evidence of geographical effect on music genres.
- It is interesting to see the distribution of different genres, but in order to really predict and classify music, we need to look into the music itself.

Insight 2

- MFCC music descriptor provides a rather reliable performance on music classification.
- Classification models based on 140 different MFCC music descriptor are obviously better than 8 different Echo Nest music features.
- Different sets of music descriptors may fit different classification models.



Acknowledgements

We would like to thank NYU HPC for providing us access to Apache Spark for data profiling, cleaning, and analysis. We are also grateful to Michaël Defferrard and other contributors for offering the open-source detailed datasets of music. Finally, we hope to extend our thanks to Google Colab for offering us the PySpark environment to train our models.

References

[1] Defferrard, Michaël, et al. "FMA: A dataset for music analysis." *arXiv preprint arXiv:1612.01840* (2016).

[2] Defferrard, Michaël. “MDEFF/FMA: FMA: A Dataset for Music Analysis.” *GitHub*, 2017, github.com/mdeff/fma.

[3] Li, Tao, Mitsunori Ogihara, and Qi Li. "A comparative study on content-based music genre classification." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. 2003.