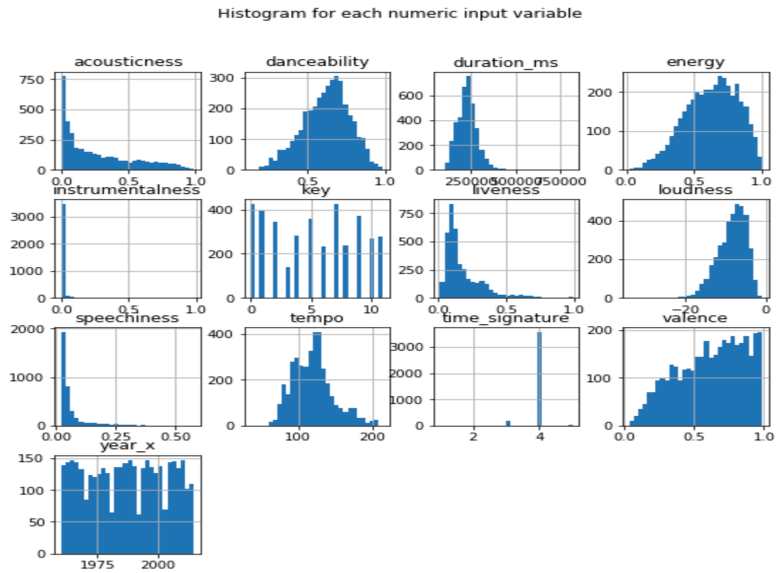
Mid Project Review

MSiA Yi Feng

Highlights

- Finished the Data Preparation for the modeling
  - After merging data sources, there are 5027 rows of observations, 14 potential predictors and 2 potential response variables
- Initial EDA and modeling
  - Histogram of Numeric Predictors



Highlights

- Modeling
  - Build several regression models based on the response variable “popularity” which is computed as 1/(Billboard rank)
  - Build several classification models based on the binary response variable “rank” which is computed as
    - 1 if rank is less or equal to 25
    - 0 if rank is greater than 25

Review Progress

- Finish up to Epic 2 Story 1
  - **Backlog**
    - Story 1: Merge databases (4 point)
      - Online datasets searching
      - Merge several datasets to include more features of songs that will be needed for modeling
    - Story 2: EDA (2 point)
      - Explore the potential variables that could be used to better predict the songs' popularities
      - Perform necessary variables transformation
  - **Epic 2**: Modeling Build the predicting models such as linear regression, neural networks and etc. Choose the optimal model by the ML metrics.
  - **Backlog**
    - Story 1: Build initial models (4 points)
      - Build several predicting models with features from the first epic.
      - Use common ML metrics and test dataset to choose the final model.

Demo/Analysis

- Models for Regression
  - Random Forest has cross validation training R square as 0.16 and testing MSE of 141.91
  - Neural Network has cross validation training R square as 0.015 and testing MSE 140.23

- Models for Classification
  - Logistic Regression has cross validation training R square as 0.80 and accuracy of testing data as 0.784
  - Random Forest has cross validation training R square as 0.80 and accuracy of testing data as 0.784

Lessons Learned

- The predictive models do not work well with the dataset
  - Most of the features about the song do not influence the ranking much
  - For classification model, the machine learning metrics were satisfying because of unbalanced data
  - Considering changing the project direction to clustering and recommendation

Recommendation/Next Step

- Build clustering models with the dataset
- Build recommendation model such as kNN with the clusters
- Construct the front end of the web app