Name:Dingyu Fu
Student ID: N12718276
Instructor: Pascal Wallisch
Introduction to Data Science (DS-UA-112)
16 Dec 2023

Capstone Project

Introduction

A set of 52,000 songs is given. Methods like Principal Component Analysis, Linear Regression, Statistical Test are used to clean and analyze the data, figure out the relationship between different features and make predictions.

Data handling

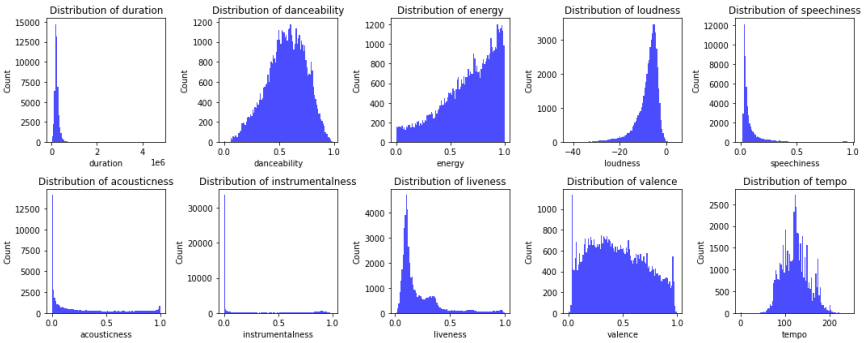
a. At the beginning of the code, I removed the Null value by using file = file.dropna().
b. For data that I need to calculate correlation coefficient from (PCA included), I did Z-Score standardization.
c. For String/boolean categorical data that need to be transformed into numerical label, I either ran a for loop and create a new data frame to replace the old data or use
   le = preprocessing.LabelEncoder()
   genre = le.fit_transform(genre)
   to change the label into numerical label
d. Subsets of data are named based on which question they are in plus simple letter description.
e. Seeding process is included inside test set,training set split function and any function that requires random generator.
f. Mean is the test statistic of all permutation test

Analysis

1. Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Is any of these features reasonably distributed normally? If so, which one?

Data Analysis: Since the none of shapes of distribution is similar to bell shape, it seems that none of these features are normally distributed. Even for danceability whose distribution looks like normal distribution, it is not normally distributed. I performed a Kolmogorov-Smirnov (KS) test to test whether these features are normally distributed, but all result p values are smaller than 0.05, which enables me to drop the null hypothesis that danceability was drawn from a normal distribution.

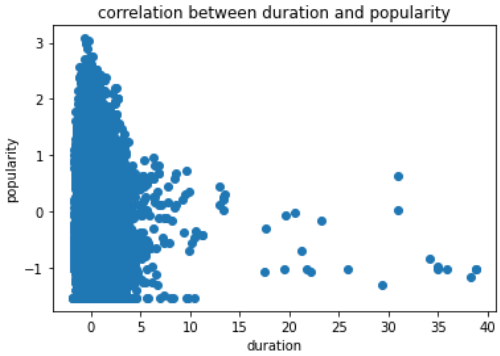
Index	p value
duration	0
danceability	1.66265e-45
energy	0
loudness	0
speechiness	0
acousticness	0
instrumentalness	0
liveness	0
valence	1.4725e-125
tempo	1.33311e-80



2. Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative?

Procedure: According to the scatter plot of duration and popularity, a linear relationship is not clearly shown. Therefore I calculated both Pearson and Spearman's correlation coefficient between song length and popularity.

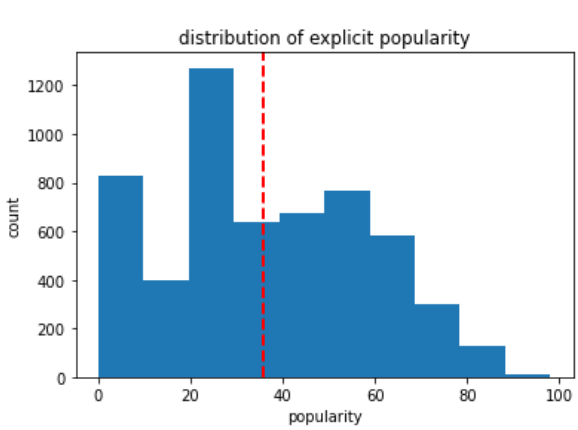
Data Analysis: The Pearon correlation coefficient is **-0.0546511959363764** and the Spearman's correlation coefficient is **-0.03728567620648788**. Both coefficients suggest that the relationship between song length and popularity is weak because two coefficients are close to 0.

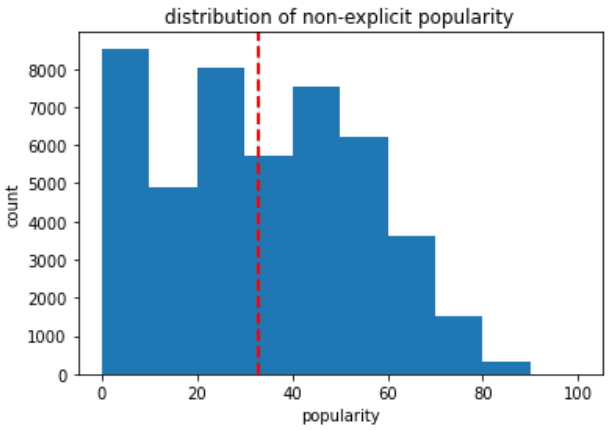


3. Are explicitly rated songs more popular than songs that are not explicit?

Procedure: My null hypothesis: that explicitly rated songs are more popular than non_explicit songs is due to chance.
To test this hypothesis, I first split the data into two groups, explicity and non-explicit and calculate mean of each group, which are **35.81**, and **32.79**. The median of each group is 34 (explicit)and 33 (non explicit). Then I feed these two data group into both the Mann-Whitney U test and permutation test, since the feature explicity is not normally distributed. I took the mean of popularity as the test statistic of permutation test.

Data Analysis: The p_value I get from these two tests are **3.0679199339114678e-19(U test)**, and **0.001998001998001998(permutation test)**. Both of them are below alpha value(0.05). Therefore I can drop the null hypothesis and conclude that explicit song is popular than non explicit song is not due to chance.



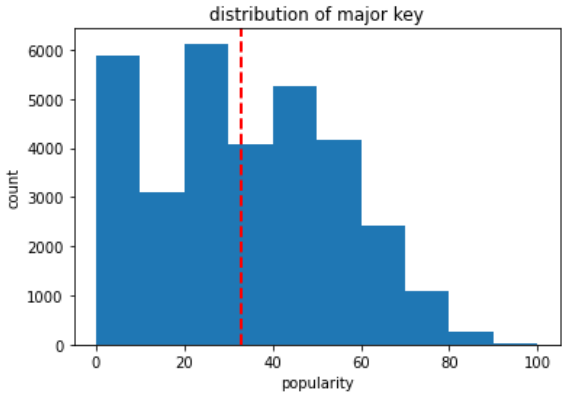


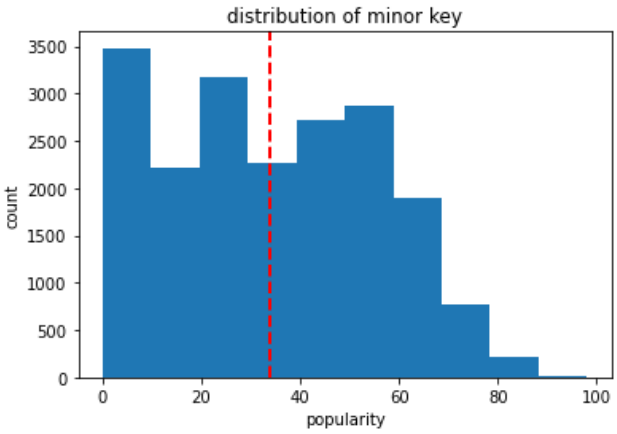
4. Are songs in major key more popular than songs in minor key?

Procedure: The way I did this question is same as the last question. Since both distributions are not normally distributed, Mann Whitney U and permutation test were used.

Data Analysis and Explanation: Mean of major key is **32.76** and the mean of minor key is **33.71**. Median of major key and minor are 32 and 34 respectively. Based on the mean and median of major key and minor key, we can know that songs with minor key is slightly popular than songs with major key.

The p value from Mann Whitney U test is **2.0175287554899416e-06** while the p value from permutation test is **0.001998001998001998**; Both of them are lower than alpha level(0.05). Therefore the differences in mean are not due to chance. The result that minor key is more popular than major key is significant.

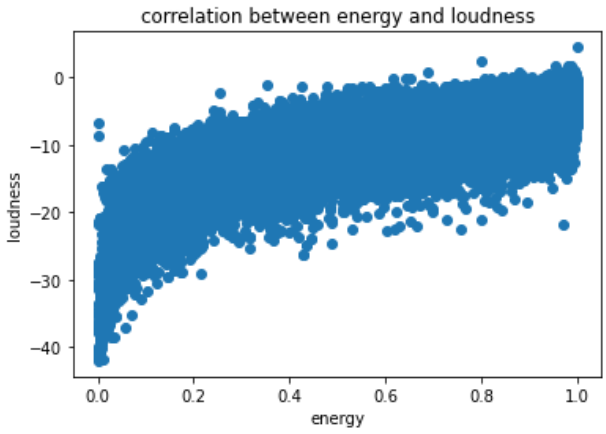




5.Energy is believed to largely reflect the “loudness” of a song. Can you substantiate (or refute) that this is the case?

Procedure: I first plot the Energy and Loudness scatterplot and found out a clear linear relationship. Then I performed the Pearson correlation calculation.

Data Analysis: The Pearson correlation coefficient calculated is 0.77, which is close to 1. Therefore there is a strong positive correlation between the energy and loudness of a song. This correlation coefficient does demonstrate energy is believedto largely reflect the loudness of a song..



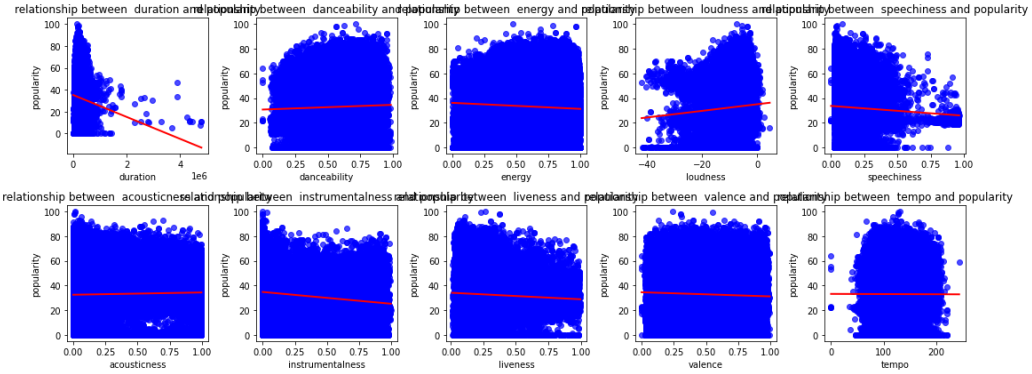
6. Which of the 10 song features in question 1 predicts popularity best? How good is this model?

Procedure: I used R^2 to assess if models can predict popularity best. R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent

variables in a regression model. The closer R^2 to 1 is, the more variance the model can account, and the better the model can predict.

Data Analysis and Explanation: Based on the R^2 I calculated, instrumentalness has highest $R^2 = 0.02186$ among all features.Therefore instrumentalness predict popularity best among 10 features. Also, this value suggests that instrumentalness can explain 2.1% of variance of popularity. However, although it has highest R^2, 0.021 is still too low to predict popularity well. Low value R^2 for all 10 features indicates that the 10 features individually only explain a little amount of variance of popularity so a better model/feature is needed to predict popularity.

(The figure below shows the relationship between each one of 10 features in Q1 and popularity)



Index	slope	intercept	Rsq
duration	[-1.00726444e-05]	35.306	0.00178847
danceability	[3.84571434]	30.8146	0.00185842
energy	[-4.89135063]	36.2499	0.00206895
loudness	[0.26731556]	35.1131	0.00266517
speechiness	[-7.92230276]	33.7444	0.00145993
acousticness	[1.85665518]	32.4766	-0.000539979
instrumentalness	[-9.54731966]	34.7352	0.0218602
liveness	[-5.29377414]	34.1101	0.000242218
valence	[-3.35289952]	34.499	-0.000792088
tempo	[-0.00142007]	33.1583	-0.000901095

7. Building a model that uses *all* of the song features in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question 7). How do you account for this?

Procedure: After doing general linear model using all of the song features, I used R^2 to assess if this model can predict popularity. I also used cross validation method and got its mean score.

Data Analysis: The R^2 I got is **0.046**, larger than the highest value in the question 6. Therefore, this model does do a better job than using simple linear regression of each feature to predict popularity.The mean cross validation score is **0.04732591847666727.**

Explanation: However,$R^2 \approx 0.046$ means this model can only explain **4.6%** of variance of popularity of the song. Therefore, this model is not too helpful in predicting popularity.
Since each of original 10 features only account for a little variance of the popularity(from last question), the multi-regression can't account for it well either. However, since each feature more or less account for the variance of the popularity of songs, the multi-regression model which take all of the predicators into account can make a better prediction about popularity.

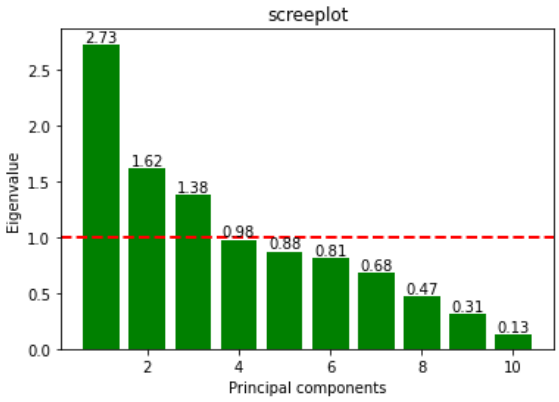
Index	R Square	RMSE	intercept	coef
duration	0.0464243	21.2196	52.2536	-8.15233e-06
danceability	0.0464243	21.2196	52.2536	4.41815
energy	0.0464243	21.2196	52.2536	-13.8073
loudness	0.0464243	21.2196	52.2536	0.681601
speechiness	0.0464243	21.2196	52.2536	-7.03205
acousticness	0.0464243	21.2196	52.2536	1.09933
instrumentalness	0.0464243	21.2196	52.2536	-8.60598
liveness	0.0464243	21.2196	52.2536	-2.70334
valence	0.0464243	21.2196	52.2536	-8.13751
tempo	0.0464243	21.2196	52.2536	0.00796462

8.When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using the principal components, how many clusters can you identify?

a. **Procedure:** After Z scoring of data,I made correlation matrix of the 10 features. I fed the matrix into PCA and got 3 principal components whose eigenvalues are greater than 1 based on Kaiser's rule. The Table below shows proportion of each principal components account for. I also used 90% rule to find out important principal components.The first 7 principal components together are able to account for over 90% of the variance.

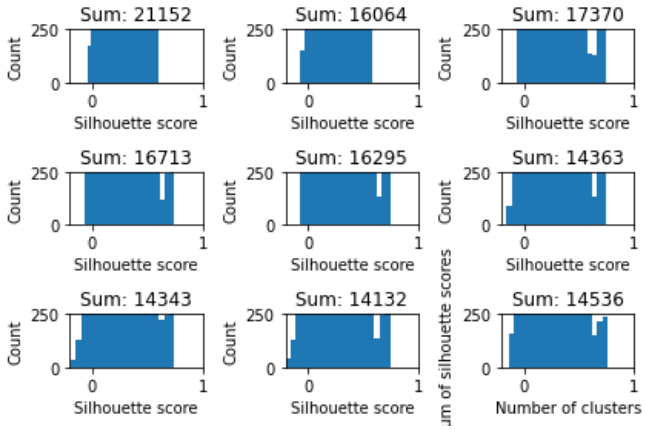
b. **Data Analysis**: The 3 principal components account for $37.3388\% + 16.1738\% + 13.8458\% = 57.3582\%$ of total variance in the data.

c. **Explanation:** The 7 principal components that I got from 90% rule suggest that 10 features of original data are to some extent independent of each other so there are still 7 significant important components.
The first 3 principal components coming from Kaiser rule can only explain 57% of the variance. This also means that more than 3 principal components from different dimension/independent features are needed to account for the variance.



Index	eigenvalue	ınce account fı
0	2.73393	27.3388
1	1.61739	16.1736
2	1.38461	13.8458
3	0.979607	9.79588
4	0.875226	8.75209
5	0.814846	8.14831
6	0.678282	6.78269
7	0.471581	4.71572
8	0.31314	3.13134
9	0.131581	1.31578

d. **Procedure:** Then I fed my original data into new coordinate system and use K means to classify my data. After that, I used Silhouette method to determine which K to use: I calculate silhouette coefficient of each K from K=2 to K =10;
e. **Data Analysis:** I used K=2 since when K=2, sum of Silhouette score is largest. (the closer to 1 each silhouette coefficient is, the more ideal the classification is; therefore the larger the sum of Silhouette score is, the better the classification).

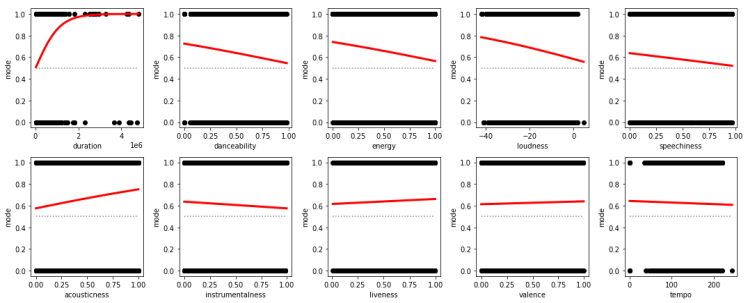


9. Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?

Procedure:I first split data of 10 features and mode into training set and test. I fed training set to Logistic Regression model and calculate AUC value of each feature and get tables as shown.

Data Analysis: Valence is not a good predictor of major or minor key since its AUC value is 0.505. This is suggesting the modelis not much better than randomly guessing the class labels. It is also suggesting that the model is not able to differentiate between minor key and major key. And among 10 features, the speechiness has highest AUC value, but it is also around 0.5.

Explanation: Since AUC value of 10 features from question 1 are around 0.5, they are not able to differentiate between major key and minor key.



Index	AUC
duration	0.476207
danceability	0.558439
energy	0.549762
loudness	0.535355
speechiness	0.563183
acousticness	0.559206
instrumentalness	0.533181
liveness	0.512221
valence	0.505073
tempo	0.512569

10.Can you predict the genre, either from the 10 song features from question 1 directly or the principal components you extracted in question 8?

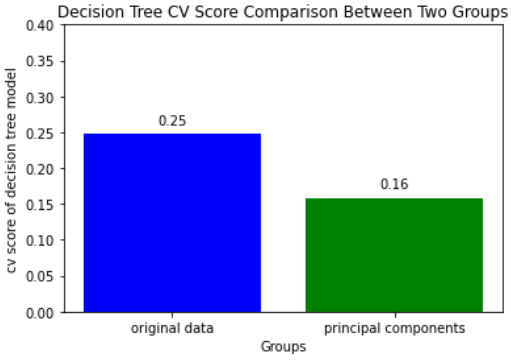
Procedure:I first converted string label of genre to numerical label by using le = preprocessing.LabelEncoder(). Then I fed 10 song features from question 1 and genre to DecisionTreeClassifier and Random Forest Classifier. Then I calculated cross validation scores of 2 classifiers.

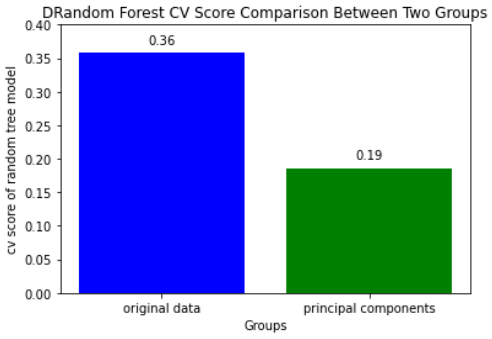
I also performed two classifier on 3 principal components (and 10 principal components, but not shown) after PCA.

Data Analysis:

Using Decision Tree:
For original data fed into DecisionTree model I got cross validation score of **0.25**, which means the classifier only get **25%** classification corretly. I got cross-validation score of **0.1578** for model constructed using 3 principal components I extracted from PCA**.** Both assessments indicate the data doesn't fit into the DecisionTree model well. The higher score suggests that random forest is a better classifier than DecisionTreeClassifier.

Using Random Forest:
The random forest classifier gave me cross validation score of original data of **0.358**,(0.36 in the graph) while cross validation of PCA data gave me score of **0.1882**(0.32 on the graph). This means the Random Forest Model can only predict genre correctly around 35% of the time.





Explanation:
Based on the cross validation score, both models have score/accuracy lower than 50%, no matter using original data or 3 significant principal components. This classification is even worse than randomly guessing. Therefore these 10 features are actually not good determining factors for genre, neither the new principal components from PCA are.

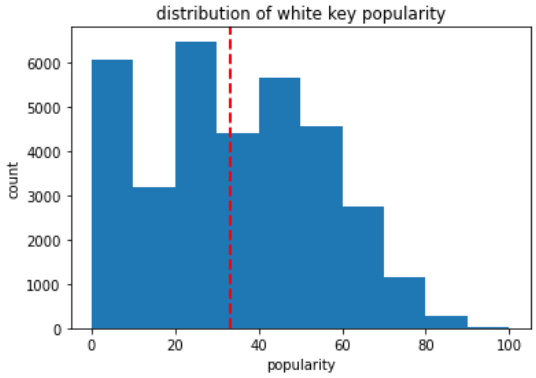
Based on the cross validation score, Random forest model is doing a better job in accuracy of predicting genre. Besides, models using original data have a higher accuracy than PCA do. That is because I only used 3 principal components which only explain 57% of variance. Therefore some information is missing. Unexplained variance in the rest components are needed for predicting genre more correctly.

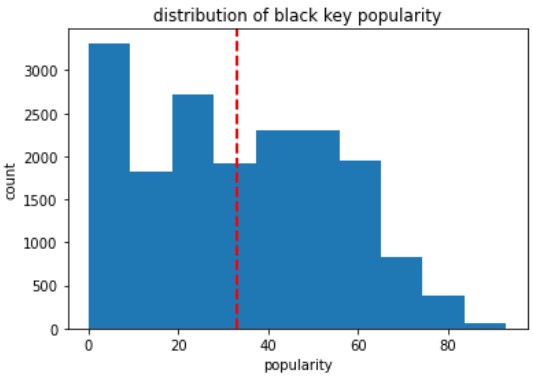
(PS: When I fed the 10 principal components into models, they still performed worse than the model using original data)

Extra credit

Procedure: I considered the relationship between key(not the mode) and popularity. I split the key [0 , 2 , 4 , 5 , 7 , 9 , 11]into white key and [1 , 3 , 6 , 8 , 10] into black key and tried to find out if the popularity is different in these groups.I calculated the mean of popularity of white keys and black keys.And then I performed welch t test and permutation test to verify the possibility of the result. I also used Mann Whitney test to test difference between these 2 groups.

Data Analysis: White key's popularity is **33.24** and black key's popularity is **32.88.** White key is more popular than black key. However, after performing welch t test and permutation test, p-values for both tests are 0.07605 and 0.08591. Therefore I fail to reject the null hypothesis that the difference in popularity between white keys and black keys are due to chance.
However, the Mann Whitney U test gave me a p value of **0.026,** which enables me to drop the null hypothesis that their difference in median is due to chance.Given median of white key is 33, median of black key is 32, popularity are very close. Even though p value suggests the significant difference, in practical term, we can still consider them similar in popularity.





Explanation:The popularity of a song isn't affected by the use of white key and black key. Therefore the use of white key of black key isn't a good predictor of popularity.