Music and Its Effects on Personality

Abraham Do

DSC 240 Final

Fall 2017

Introduction

1. What is the purpose of this project?

Music has had a high level of influence throughout the entirety of human history. Today, there are a plethora of musical genres that are readily available for consumers to listen to, and each genre entails with it a classification of the type of individual that would listen to its songs. For example, only the rich and intellectual listen to classical music while only those that are struggling and unintelligent listen to rap. These are universally known opinions regarding such musical genres, but is it possible to take these opinions and validate their categorizations of personalities and individualizations through the analyzation of a dataset?

2. How will I achieve this goal?

Because this project requires me to gather human data, I decided that the best method of approach was to find a survey already completed online. As much as I wanted to gather my own data, I came to the realization that even if I were to ask students at the University, I would only get a few hundred results at most. This was insufficient for me because I wanted to see results that better suited a larger sample of individuals, so I instead discovered a dataset on the website Kaggle. This dataset was much more ideal for me as it was a survey completed in 2013 by college students in the United Kingdom with over a thousand total responses with approximately 150 different types of questions asked. Furthermore, the dataset itself was very clean as it was done with results on a scale of 1-5, which made analyzing the data much more convenient without much alterations needed.

After much deliberation as to which algorithms I wanted to attempt to use on my dataset, I decided to implement the Logistic Regression, PCA, and Random Forest algorithms to determine the correlations between music and its effects on an individual's personality. I had originally planned to implement a XGBoost algorithm as well on my project but I quickly realized that the algorithm was extremely complex to install and get working properly; therefore, I unfortunately had to drop this approach as I did not have the proper amount of time to figure out the proper way to use the algorithm. The main reason that I decided to use multiple algorithms was to make sure that I was getting similar results with different approaches. This was important to me because I did not want to create results that were incorrect, as accuracy is imperative in a project such as this one.

3. Algorithms

Each algorithm has a different purpose, and I decided that it was important to make a clear distinction between all 3 before I go onto the research that I found.

Logistic Regression: This algorithm is a method in which the correlations in a dataset are found through the analyzation of one or more variables to output the best fitting number of value between each variable. In a sense, the algorithm is finding the line of best fit between all of the data points to see how each one relates to one another in a linear model. This is an extremely useful method for me as it should take all of the survey results that I have in my possession and output a result that will explain to me how each individual feels about a certain genre of music with a linear relationship between the two variables that I input.

PCA: This algorithm is simply a technique that individuals use to emphasize variations and to showcase the patterns in any given dataset. This particular dataset is more visual based, and it will just make it easier to have an idea of what every individual prefers because it will show every result and its correlations on a visual plot.

Random Forest: This algorithm works by creating multiple decision trees depending on the dataset given and outputting the mean predictions of each individual tree made through regression. I felt that this was a great algorithm to implement because of its relative similarity to Logistic Regression. Although the approaches are different, the actual results should be similar to one another which is exactly what I am trying to achieve through my coding. I could have stuck with just using Logistic Regression, but I wanted to ensure that the results that I was obtaining were all accurate and I decided that using this particular algorithm would ensure that I was indeed doing the right steps to achieve success in my project.

4. Research

In my research, I discovered that Campaña, Arroyo, and Yoo implemented the SEMMA methodology (sampling, exploration, modification, modeling, assessment) along with the OCEAN model ((O)penness to experience, (C)onscientiousness, (E)xtraversion, (A)greeableness, and (N)euroticism) to determine the connections between musical preferences and its associations with personalities. To accomplish this, they implemented K-means clustering along with ROC curves to view any potential similarities between the two topics at hand and determine if music can truly influence a person's life on a personal, social, and cultural level.

In the second paper that I read, Pandey attempted to implement the Myers-Briggs Type Indicator (MBTI) to see if an individual's personality traits can effectively predict what kind of book or other media content a person may be drawn to. Pandey tested this theory by hosting an online survey through several platforms and gathering data to be analyzed through the use of Principal Component Analysis (PCA) and K-means clustering. This test ended inconclusively because Pandey came to the realization that more variables had to be tested to avoid any skewed data, which is something that I could improve on with the dataset that I acquire.

In the final paper that I had looked at, Ferwerda, Tkalcic, and Schedl managed to mine data from the online streaming service Last.fm and cross reference the found data with personality tests that Facebook users had completed on an application known as “myPersonality”. The OCEAN model was once again used to map out any correlations between the different personalities, and unfortunately, there is no explicit description of which algorithm they had used. However, it is clear that they had used a clustering method to efficiently analyze the data, so we will take this into consideration for the project as well.

5. **Data Organization**

For my particular project, I did not need to use all 150 variables asked to the participants of the survey. To fix this, I decided to use the .iloc method to combine several groups of variables together depending on their relations to one another, and compared all of this data to an independent variable of music which also consisted of 17 different genres of music. The listing goes as follows:

Music: The constant variable which I am using to find correlations within the dataset

Math: Math, Physics

Science: Psychology, Biology, Chemistry

Spending: Shopping Centers, Branded Clothing, Partying/Socializing, Appearance, Gadgets

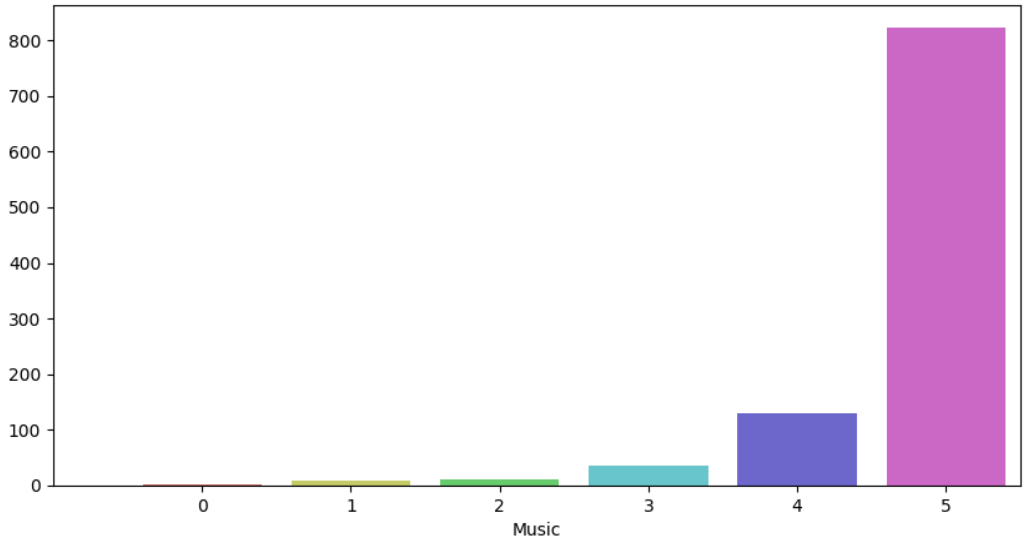
Active: Outdoor Activities, Passive Sports, Active Sports, Adrenaline Sports

Religion: Religious, Belief in God

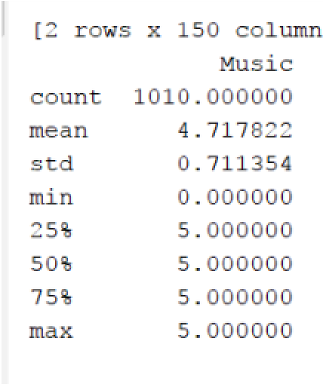
The reason that I set the data up as such was to make the comparisons simple, and so that only two variables would be involved even though in reality, there could be dozens of variables actually being tested at once.

Data Collection

1. Introduction

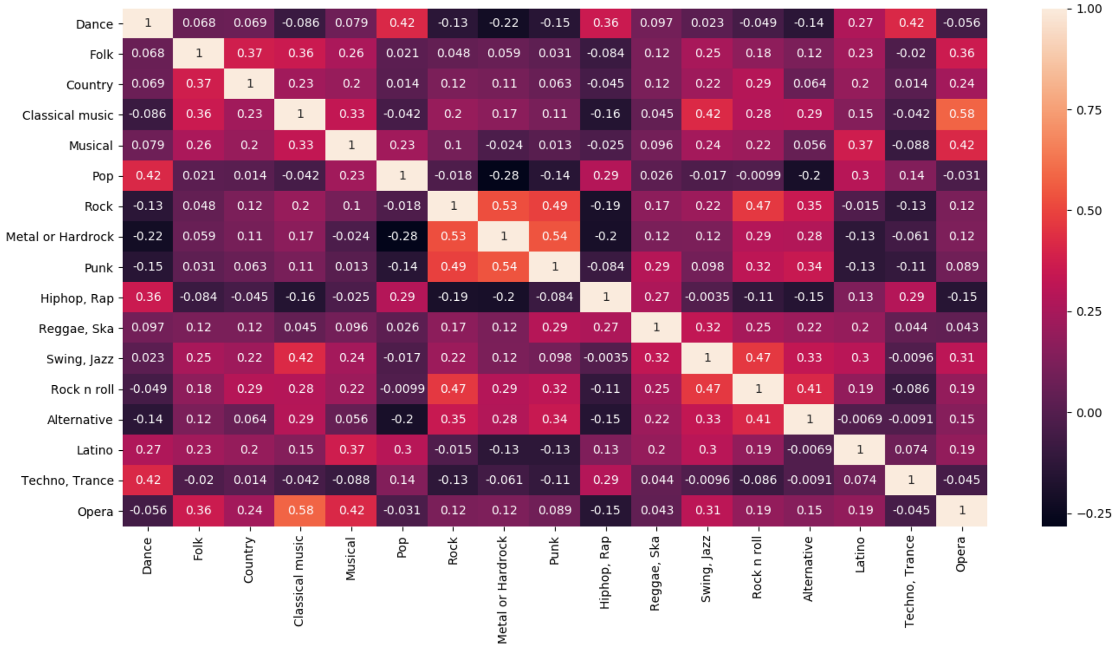


To start, I wanted to ensure that the dataset that I was using was the right choice by checking to see if the individuals involved really did like listening to music. It would have been useless to choose this dataset if it was visibly noticeable that the participants had little to no interest in actually listening to music, and thankfully this dataset had over 80 percent of its users answering that they loved music with a 5 result, and more than 15% of the users either giving the survey a 3 or 4 which is a great result for a project of this sort.



```
[2 rows x 150 column
               Music
count   1010.000000
mean       4.717822
std        0.711354
min        0.000000
25%        5.000000
50%        5.000000
75%        5.000000
max        5.000000
```

Furthermore, when printing the .describe() function of the survey, you can see that the results are extremely positive with a mean value of 4.7178 over all 1,010 responses for the dataset. This is more than what I could have asked for, and coming across a dataset with this magnitude of positive results was extremely beneficial to me in terms of my confidence of how accurate this project would be for music listeners.

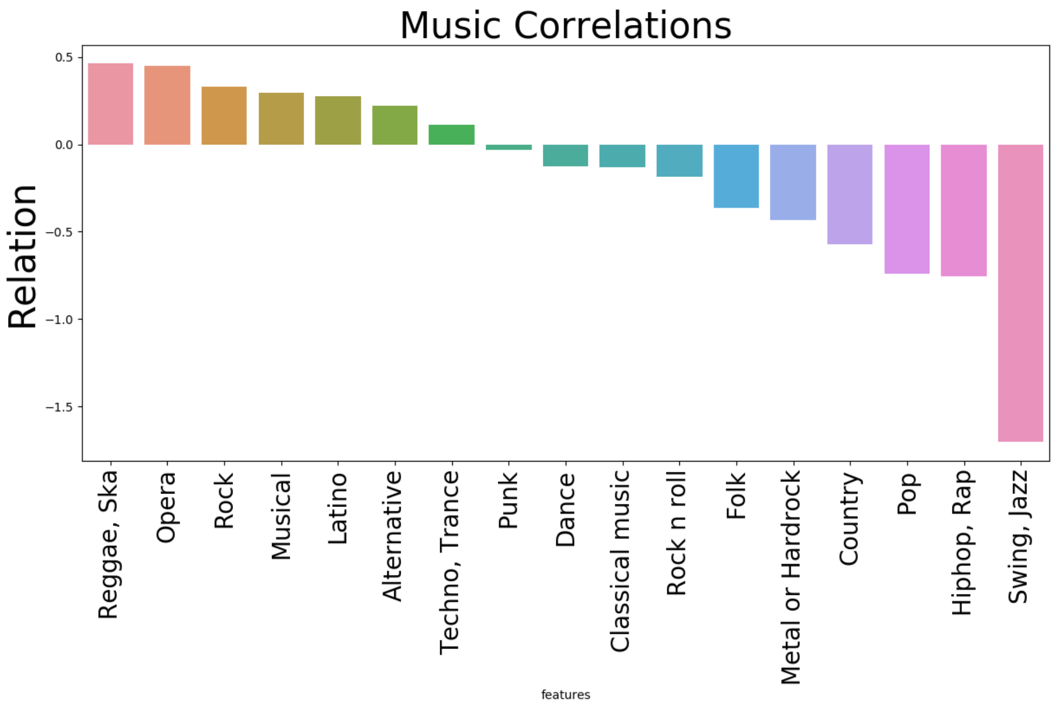


Along with discovering the correlations between an individual's personality and their preferences of music, I wanted to see how accurate the relationships were in terms of the music itself. To elaborate, according to this graph people who enjoy classical music have a fairly significant chance of also enjoying opera music with a significance level of 0.58. However, does this mean that a person who enjoys Math will also enjoy both styles of music as well, or is this correlation actually different depending on the individuals involved? Vice versa, people who listen to Metal have a -0.28 significance level in comparison with Pop music. Will this be the case with all variables involved, or will there actually be a group that enjoys both genres of music equally?

I thought that this would be an interesting subject to cover on the side, and I wanted to explain it on my report as well because the results that I had obtained were actually extremely interesting in certain cases. Of course, I will not be comparing every single correlation as that would take far too long, but I decided to take 2 or 3 of the strongest and weakest relationships and compare the results to see if this correlation between musical genres really stayed consistent with every variable type.

Strongest Correlations	**Weakest Correlations**
Classical/Opera	Metal/Pop
Punk/Metal	Metal/Dance
Rock/Swing,Jazz	Alternative/Pop

2. **Logistic Regression**
   a) Music vs Math



With the implementation of Logistic Regression, I was able to come up with five different graphs that I will explain in detail. There are five graphs because I am comparing the independent variable (music) with the five different dependent variables (Math, Science, Spending, Active, Religion) that I have discussed before.

In this particular graph, you can see that individuals who enjoy mathematics enjoy listening to Reggae, and very much dislike Swing. Something that I thought was strange about this particular graph was the musical preference of Reggae. Personally, I don't really know anybody who listens to Reggae and I was worried that the data might have been off, but I came to the realization that it would not be right for me to completely judge this dataset with my own preferences as this is a UK based survey and not my own. I can't judge as to what kind of music individuals from the UK enjoy, but I know enough to know that their culture is different from ours here in the US so that realization helped me accept the results with more confidence.

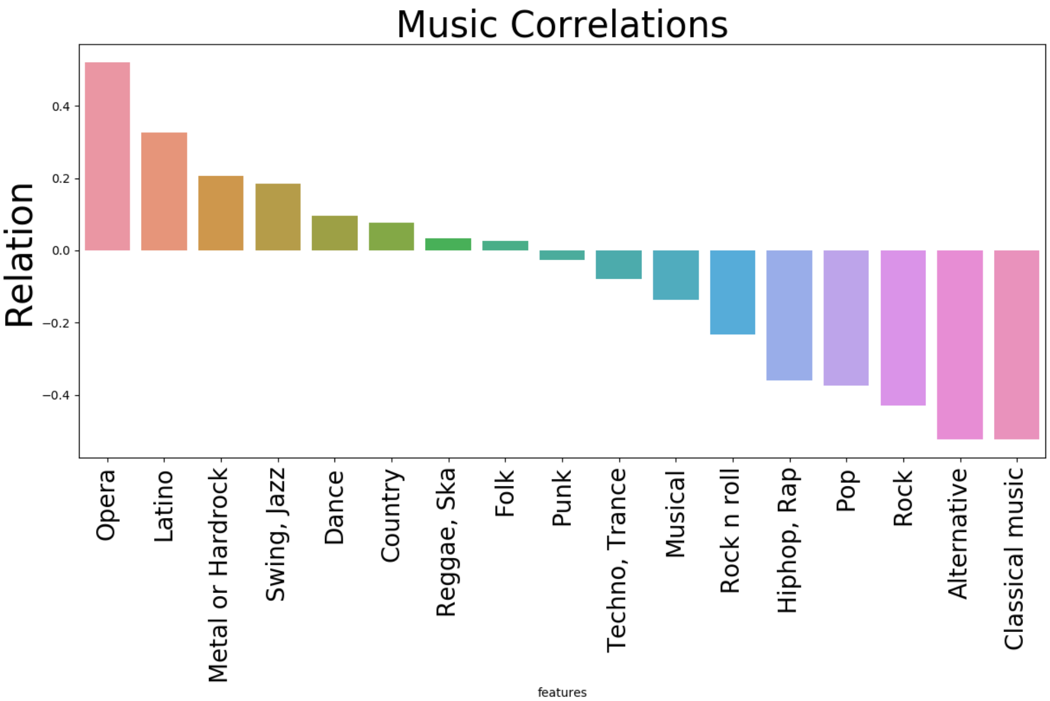
Correlations

Rock and Swing are very different but they should be similar according to the heatmap.

Opera and Classical are also very different from one another and they should be similar.

Metal is similar to both Dance and Pop, but they should actually be very different.

b) Music vs. Science



In this particular scenario, individuals who enjoy Science tend to also enjoy listening to Opera music and dislike Classical music. If there's something that sounds strange about that, it's because this completely contradicts the heatmap results from earlier. How could the strongest correlation in the entire heatmap have a completely different relation value in this particular graph? This is a question I will answer later on, but I felt that it was extremely important to bring this question up now so that we can see how recurring this particular anomaly is through all five graphs. Other than this, the only other outlier for me is the fact that these individuals really enjoy Latino music, but I believe that this also has to do with the fact that the dataset is based on the UK so we will simply have to accept the fact that this is indeed the case in this part of the world.

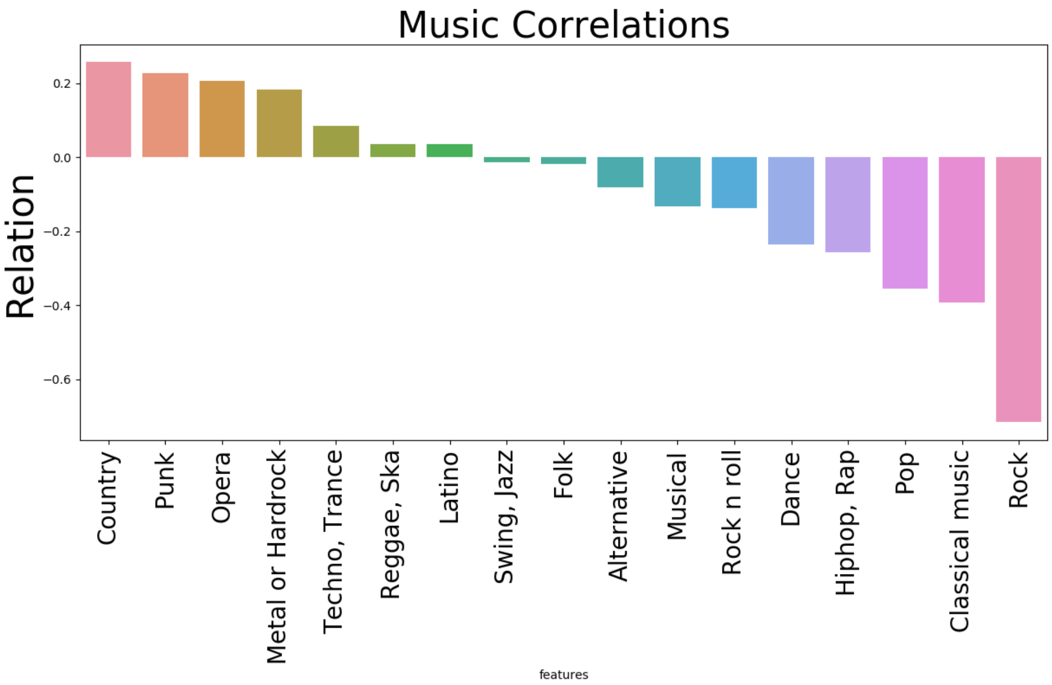
Correlations

Classical and opera are polar opposites, but they should be one of the closest related variables.

Swing and Rock are also very different, which should not be the case.

Finally, both Metal and Jazz are positively correlated, but they should instead be different.

c) Music vs Spending



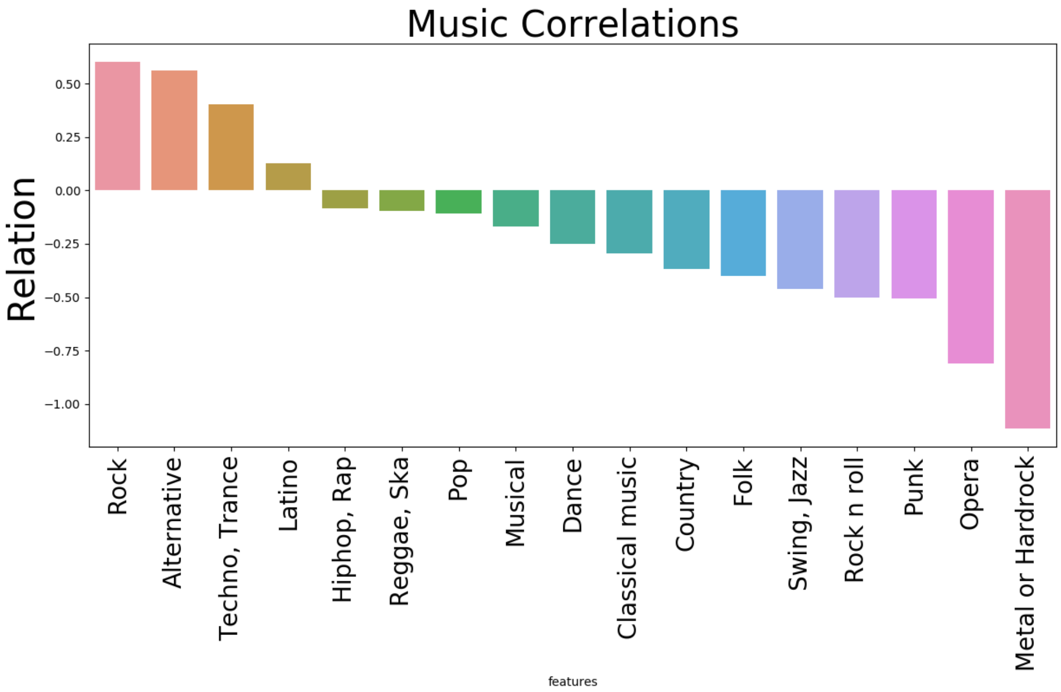
In this comparison, people who like to spending money very much enjoy Country music and hate listening to Rock. Of every test, this one had the most unexpected results as I had always believed that people with money would mostly enjoy classy music such as Opera and Classical. Furthermore, I always believed that Country and Punk music was listened to individuals with less money, so to see these results were very eye opening and made me extremely happy to see because it shows me that perhaps the stereotypes involved with music aren't always as valid as they seem. Again however, it was strange to see Opera and Classical on almost opposite ends of the graph's spectrum, and we can see that there is a recurring pattern with the initial heatmap's failure to repeat its findings when compared to the actual personalities of the grouped individuals in our tests.

Correlations

Classical and Opera are once again very different from each other when they should be similar.

Swing and Rock are both negatively correlated, but they should be different instead.

d) Music vs Active



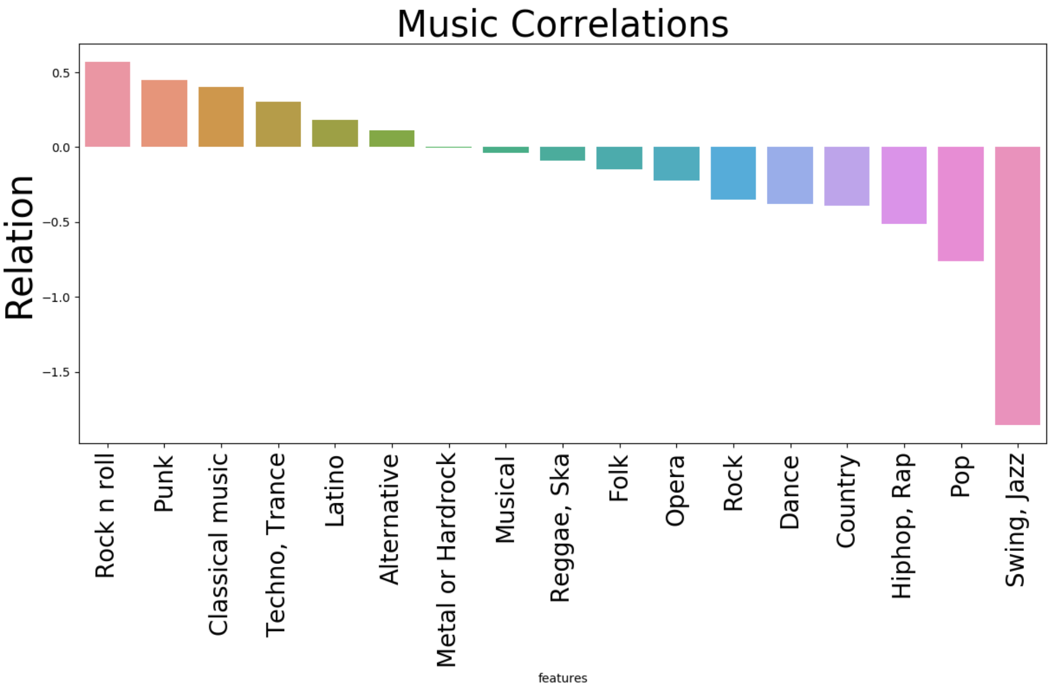
In this particular scenario, individuals that live an active life enjoy Rock but also dislike Metal. Of all the scenarios that I had tested, I believed that this was the most “normal” result based on my own preferences. It was definitely a bit strange to see that people who enjoy Rock for whatever reason cannot stand Hardrock, but this just goes to show I really do not understand the differences in genres because they are on completely opposite ends of the Logistic Regression graph. It is interesting that it took 4 tests to see a result that I can finally somewhat relate with in terms of the expectations that I had for every variable, and this is something that I will discuss further on in my conclusion.

Correlations

Rock and Swing are very different but they should be similar.

Metal and Pop and both negatively related but they should actually be different.

e) Music vs Religion

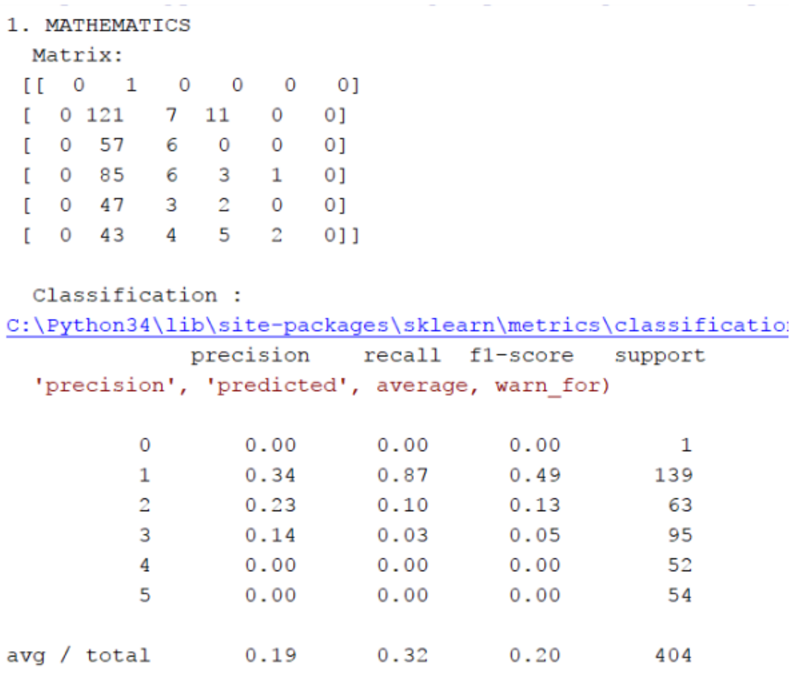


Religious individuals in the UK thoroughly enjoy Rock n Roll, but hate Swing music. For me, I couldn't believe that Rock n Roll and Punk were the two top options in this particular run of the test, and it was a complete break of expectations as I had always believed that religious individuals would much rather prefer slow and peaceful music. My only explanation for this would be that being college students, many of them may be listening to these genres of music as a form of defiance against their parents or their religion in general, but this is something that I will not go into detail without any proof or validation. Regardless, it was really interesting to see these results and I really believe that this goes to show that it may not be analytically correct to suggest a person's personality with the type of music that they like to enjoy on a daily basis.

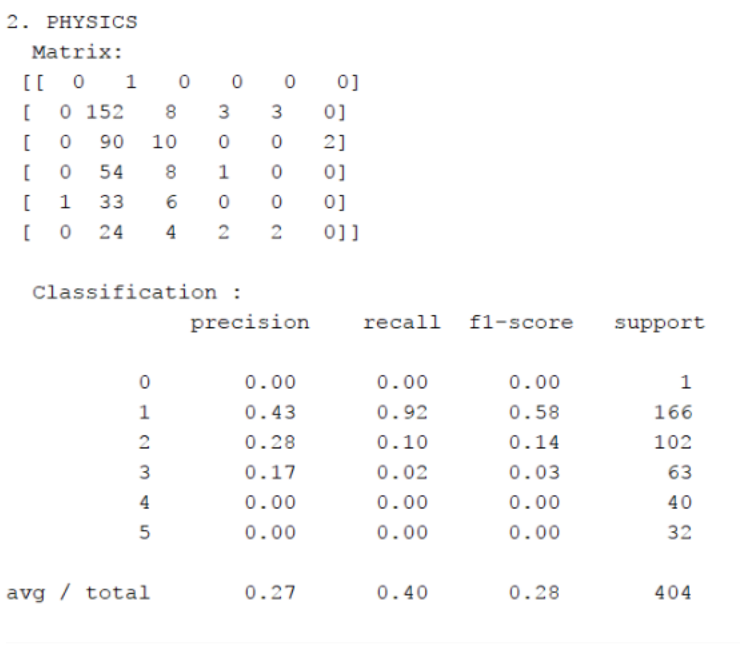
Correlations

Classical and Opera are once again far from each other when they should be closely related.

2.1. Logistic Regression Math Comparison



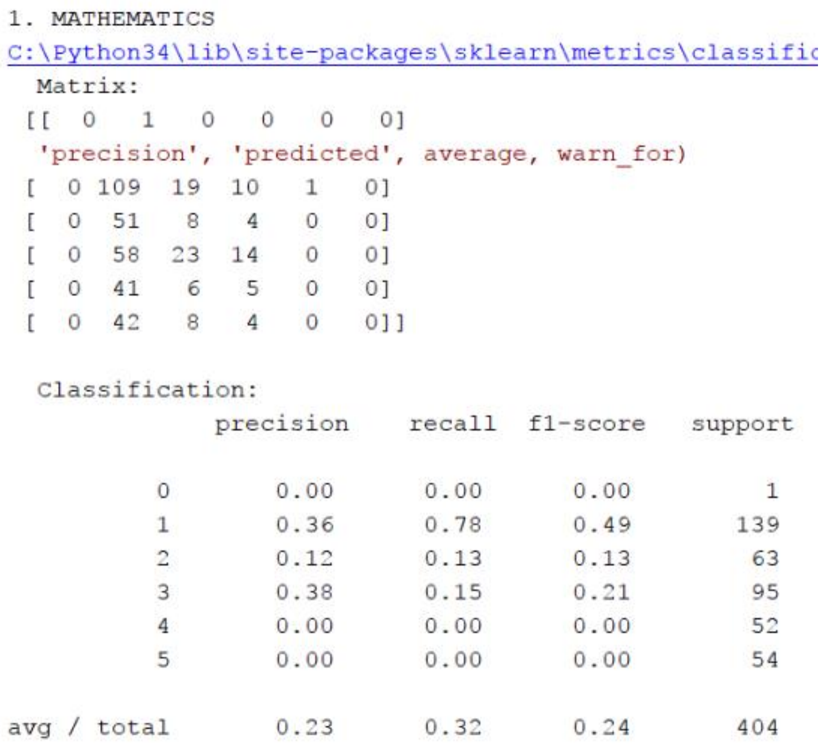
```
1. MATHEMATICS
  Matrix:
 [[  0   1   0   0   0   0]
 [  0 121   7  11   0   0]
 [  0  57   6   0   0   0]
 [  0  85   6   3   1   0]
 [  0  47   3   2   0   0]
 [  0  43   4   5   2   0]]

  Classification :
C:\Python34\lib\site-packages\sklearn\metrics\classificatio
              precision    recall  f1-score   support
  'precision', 'predicted', average, warn_for)

          0       0.00      0.00      0.00         1
          1       0.34      0.87      0.49       139
          2       0.23      0.10      0.13        63
          3       0.14      0.03      0.05        95
          4       0.00      0.00      0.00        52
          5       0.00      0.00      0.00        54

avg / total       0.19      0.32      0.20       404
```



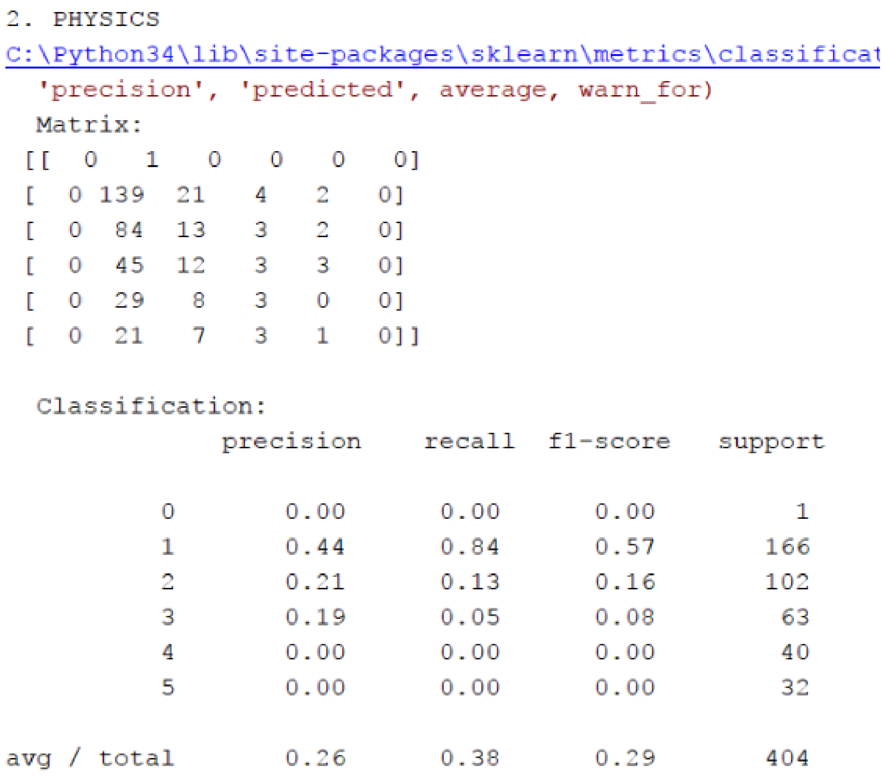
```
2. PHYSICS
  Matrix:
 [[  0   1   0   0   0   0]
 [  0 152   8   3   3   0]
 [  0  90  10   0   0   2]
 [  0  54   8   1   0   0]
 [  1  33   6   0   0   0]
 [  0  24   4   2   2   0]]

  Classification :
              precision    recall  f1-score   support

          0       0.00      0.00      0.00         1
          1       0.43      0.92      0.58       166
          2       0.28      0.10      0.14       102
          3       0.17      0.02      0.03        63
          4       0.00      0.00      0.00        40
          5       0.00      0.00      0.00        32

avg / total       0.27      0.40      0.28       404
```

The reason that I have included this data is because I did not make graphs for the Random Forest algorithm, so I instead decided to use the precision, recall, f1-score and support values instead to determine the relationships between the variables and music.

3. Random Forest



```
 1. MATHEMATICS
C:\Python34\lib\site-packages\sklearn\metrics\classific
  Matrix:
 [[  0   1   0   0   0   0]
  'precision', 'predicted', average, warn_for)
 [   0 109  19  10   1   0]
 [   0  51   8   4   0   0]
 [   0  58  23  14   0   0]
 [   0  41   6   5   0   0]
 [   0  42   8   4   0   0]]

  Classification:
              precision    recall  f1-score   support

          0       0.00      0.00      0.00         1
          1       0.36      0.78      0.49       139
          2       0.12      0.13      0.13        63
          3       0.38      0.15      0.21        95
          4       0.00      0.00      0.00        52
          5       0.00      0.00      0.00        54

avg / total       0.23      0.32      0.24       404
```



```
2. PHYSICS
C:\Python34\lib\site-packages\sklearn\metrics\classificat
  'precision', 'predicted', average, warn_for)
 Matrix:
[[   0   1   0   0   0   0]
[   0 139  21   4   2   0]
[   0  84  13   3   2   0]
[   0  45  12   3   3   0]
[   0  29   8   3   0   0]
[   0  21   7   3   1   0]]

 Classification:
              precision    recall  f1-score   support

          0       0.00      0.00      0.00         1
          1       0.44      0.84      0.57       166
          2       0.21      0.13      0.16       102
          3       0.19      0.05      0.08        63
          4       0.00      0.00      0.00        40
          5       0.00      0.00      0.00        32

avg / total       0.26      0.38      0.29       404
```

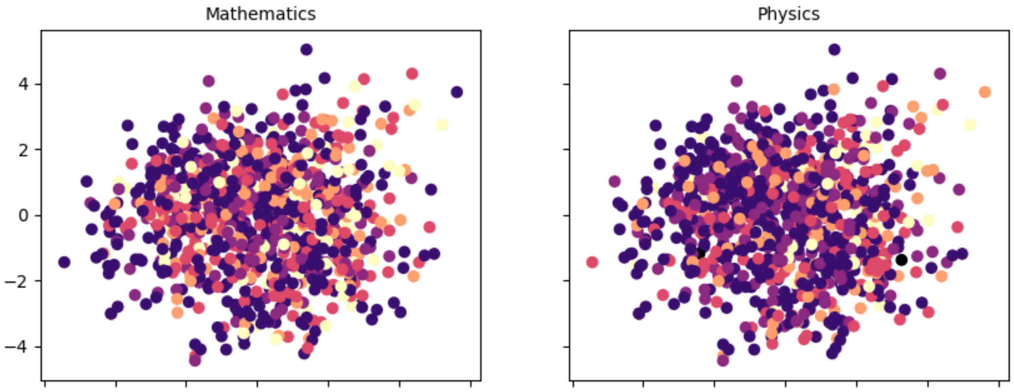
For the Random Forest algorithm, I decided to not include graphs and to rely on the classification scores as no matter what, the scores should all be relatively similar to the numbers produced through the Logistic Regression method.

Furthermore, I decided to only use the Math variable in this demonstration as it was the smallest sample set with only two sections being Math and Physics, whereas the other groups have many more sections which would have taken up too much space in this report. I did however check the results separately because I wanted to ensure that they were consistent, and thankfully every single variable had the same results as the Math group in terms of their own respective tests.

As you can see from the tests above, every single result was very close to one another, if not identical on the different categorical scores. This was precisely the result that I needed because this tells me that my Logistic Regression was also correct as well. If the numbers had been different, I would have had to go back and try to find where I went wrong in my coding which would have been extremely painful and time-consuming, and unfortunately this actually did happen to me a few times before I finally got the similar results that I was looking for.

The reason that I felt that no graphs were necessary for the Random Forest algorithm was because I felt that this method relates very closely with Logistic Regression, and it would have been repetitive and unnecessary to reshow every single result in my report when it would look very close to what I have already shown above. Needless to say, I was very happy to see that the highest margin of difference amongst the tests was only .04 on the precision and f1-scores of Math between the Logistic Regression and Random Forest algorithms. This guaranteed to me that no error was made, and it helped me move on to my next algorithm which was the implementation of the PCA graphs.

4. PCA



Finally, I decided to implement the PCA algorithm to describe the correlations between the Music and Math variables once again. Realistically, this particular algorithm isn't the most ideal solution to determining the correlations between the two variables but I just thought that this would be interesting to try and attempt for my project. As you can see, this algorithm takes every response in the Mathematics and Physics variables and relates it to the survey results of the Music group, creating this monstrosity of a graph which might be pretty to look at, but very difficult to understand.

I could definitely see this algorithm being extremely helpful for smaller datasets or even a differently organized assignment from mine, but for my project the PCA algorithm is particularly difficult to truly understand. However, being a data mining project, I wanted to challenge myself to see if I could get this algorithm up and running so that later on in the future, I can attempt to use the PCA method again in a better scenario as this method would definitely be useful to visually describe any variations and patterns in a dataset.

If you were to try and decipher the graphs created by the algorithm, you would see that the points are indeed accurately relaying the information that was seen in the Logistic Regression and Random Forest tests. Although my particular case is a bit of a visual nightmare, it still does aid me in helping me come to the conclusion that my project was finished without any errors, and that the results that I had obtained above still stand true which was the main reason that I attempted to use other algorithms along with the Logistic Regression method.

Data Analysis

1. **What Does It All Mean?**

What exactly can we conclude from the plethora of numbers and graphs shown above? To start, I believe that it is safe to say that although there are many stereotypes involved with different personalities, not all of these stereotypes hold true. The two biggest examples that I can give from my data collection are the correlations between Music vs. Spending and Music vs. Religion. Originally, I believed that people who loved to spend money were individuals who also enjoyed Classical music and anything that might assume an upscale lifestyle. Instead, these individuals actually enjoyed Country and Punk music while completely detesting Classical Music, which is a far deviation from what I had initially assumed. In the case of Religion, the top music of choice for individuals were Rock n Roll and Punk. I can't really describe as to why this may be without another study, but the results were definitely not I would have ever assumed as I had always thought that religious individuals would stray away from these kinds of music. However, it is also important to realize that this analytical approach is opinion based, so this may differ from person to person.

I am extremely glad that I performed my second analysis between music correlations using the data with the heatmap because it also helps me explain the conclusion as a whole. There were several cases where a group of individuals had a stark contrast in expected musical appreciation correlations as originally shown in the heatmaps. The most notable differences were the repetition of people who were supposed to like Classical Music and Opera, but ended up liking one and disliking the other. After a lot of thinking as to why this may have happened, I came to the realization that we are still using human subjects in our surveys, and no matter how complex an algorithm may be, it will always be impossible to completely understand the human mind through a computer. This would explain why even though the subjects were theoretically supposed to enjoy both genres of music almost equally, there were cases in which it was actually the complete opposite result. I think that it would be simple to dismiss these cases as being "noisy" data and move on, but it is too important to forgo the fact that we are trying to analyze the human mind, and it would be foolish to try and undermine this situation.

So how does this contribute to the overall conclusion? I believe that this fact helps us come to realize that we cannot generalize stereotypes in music and personality. Just because a group of individuals like something does not mean that another group will appreciate the same things, and the cross analyzation of the heatmaps and the algorithms help us see this picture. I came into this project fully expecting that the general stereotypes would hold true, but after discovering my results I was pleasantly surprised to see that I was wrong. The human mind is an unpredictable enigma, and it is foolish to think that a stereotype would always hold true for every single individual when there are simply so many other factors at play. The algorithms that

I used go to show that although there are some things that can be predicted, there are also many other variables that may surprise anyone at any given moment. Music and personalities do not always go hand in hand in every scenario, and this project proves that this is true.

2. Issues and Onwards

Figuring out the best methods of actually combining and arranging the variables were extremely difficult, and attempting to narrow down my selections from the 150 original questions asked was a task that took a long time to complete. The coding itself was also troublesome at times, and it was unfortunate that I didn’t get to try the XGBoost algorithm because of my issues with getting the package successfully installed onto my laptop.

If I were to ever repeat this task, I would attempt to get a more local dataset so that the results that I find are more relatable to me personally. I felt that because of the clash of cultures, I found myself doubting my results a lot more than I should have because they simply didn’t make much sense to me until I started to realize that it may have been because of the difference in regions. Furthermore, I would try and collect a much larger data sample if at all possible so that the results that I got were much more accurate. It never hurts to have a larger sample size as it ultimately paints a better picture of what individuals like, and there is always a chance that my particular project is inaccurate in a broader sense because of my smaller dataset.

Bibliography

X. Campaña, R. Arroyo, and S. Yoo. 2017. Experience in Applying Data Mining Techniques to Musical Content Database to Identify Personality Traits . International Journal of Applied Engineering Research 12 (2017), 3298–3304.

Animesh Pandey. 2013. Idea of a New Personality-Type Based Recommendation Engine. (November 2013).

Ferwerda. 2017. Personality Traits and Music Genres: What Do People Prefer to Listen To? UMAP 2017 Short Paper (July 2017).

Miroslav Sabo. 2016. Young People Survey. (December 2016). Retrieved November 14, 2017 from https://www.kaggle.com/miroslavsabo/young-people-survey