- How did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).

We first began by creating an index (python dictionary) for both table A and table B with the table ID as the key and the row of data as the value. This step took a bit of time for our larger, MSD table. After the indexing was complete, we iterated our table of matches (as predicted by our classifier) and accessed the appropriate table index with the appropriate IDs from the match table to retrieve the corresponding table data. We were then able to construct a new entry for our table E with all relevant attributes from both tables of a matching entity.

When merging these attributes, we constructed a few rules to provide consistency to data types in E.
For example, certain examples in both tables had missing values for the ‘year’ attribute. To combat this, when examining a ‘match’, if both years were present, we took the largest (latest) year. If neither were present, we set this value to 0. If only one was present, we used that value. Similarly for data attributes ‘Title’ and ‘Artist’, we took the shortest string.

- Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.

Schema: [“Title”,”Artist”,”Genre”,”Duration”, “Tempo”, “TimeSignature”, “KeySignature”, “Year”]
Number of Tuples: 3,751
Sample Tuples:

1. [“Leave (Get Out)”, ”JoJo”, pop, 230.55628, 122.647, 5, 9, 2004]
2. [“Beautiful Girls”, “Sean Kingston”, pop, 241.03138, 86.752, 3, 4, 2007]
3. [“Mrs. Robinson”, “Simon & Garfunkel”, rock, 243.30404, 183.557, 1, 6, 1968]
4. [“Reptilia”, “The Strokes”, rock, 219.81995, 157.862, 4, 9, 2003]

- What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.

Our original goal for this data science project was to be able to predict the genre value of a given song given other attributes it has in Table E. This is why we initially scraped the MusicBrainz site in order to obtain genre tags per song.

Classification Notes:

Before starting the development of the our classification models, the table E was split into a development (1500) and evaluation set (2251). Initially all numerical attributes of or table E were inputted as features into our models. This includes duration, tempo, time signature, key signature, and year.

Unfortunately, after testing 10 fold cross validation accuracy with a suite of different classifiers on these features, we were not able to produce an effective classifier. With every raw numerical data attribute used as a feature, our best model turned out to be a **Logistic Regression Classifier.**

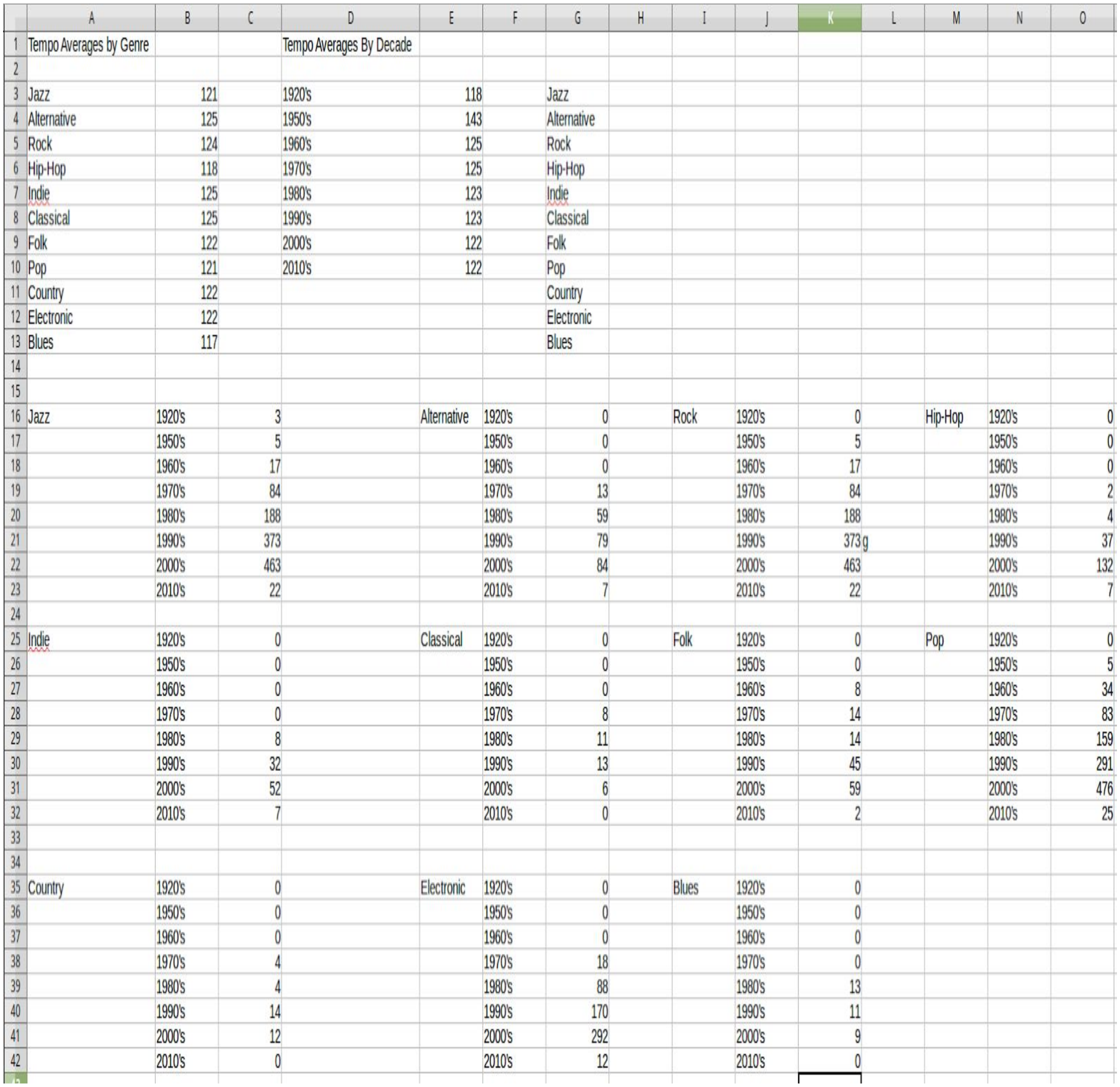
```
 Logistic Regression trained/scored on 2/3 sample. CVS: ---------  0.317443666178
 Metrics Accuracy Score: ---------------------------------------------------  0.317333333333
```

```
precision    recall  f1-score   support

alternative       0.00       0.00       0.00        93
     blues       0.00       0.00       0.00       15
  classical       0.00       0.00       0.00       14
    country       0.00       0.00       0.00       17
 electronic       0.00       0.00       0.00       246
      folk       0.00       0.00       0.00       53
    hip-hop       0.00       0.00       0.00       72
      indie       0.00       0.00       0.00       38
       jazz       0.00       0.00       0.00       52
        pop       0.20       0.01       0.01       420
       rock       0.32       0.99       0.48       480

avg / total       0.16       0.32       0.16       1500
```

Given this unacceptable accuracy, we performed some minor debugging by removing potentially problematic, unhelpful features. For example, ‘year’, which have a substantial amount of missing values was removed and increased the accuracy of our decision tree model by almost 10% (however it was still embarrassingly low). A handful of other subsets of the raw features were tried, however we were not able to beat the previously stated accuracy.

OLAP Notes:

For OLAP style analysis, we analyzed three different relationships: Tempo and Genres, Tempos by Decade, and number of songs in a Genre by Decade. Prior to actually running the code, we saw that OLAP data may be misleading. Our data is jaded in terms of number of songs collected per genre. For example, we had about a thousand instances of Jazz, Rock, and Pop, but in some genres like Country, or Indie for example, we had under a hundred data points. Because of this, we determined that outliers and insight from OLAP discovery may not be representative of the population of music.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Tempo Averages by Genre			Tempo Averages By Decade											
2															
3	Jazz	121		1920's	118		Jazz								
4	Alternative	125		1950's	143		Alternative								
5	Rock	124		1960's	125		Rock								
6	Hip-Hop	118		1970's	125		Hip-Hop								
7	Indie	125		1980's	123		Indie								
8	Classical	125		1990's	123		Classical								
9	Folk	122		2000's	122		Folk								
10	Pop	121		2010's	122		Pop								
11	Country	122					Country								
12	Electronic	122					Electronic								
13	Blues	117					Blues								
14															
15															
16	Jazz	1920's	3		Alternative	1920's	0		Rock	1920's	0		Hip-Hop	1920's	0
17		1950's	5			1950's	0			1950's	5			1950's	0
18		1960's	17			1960's	0			1960's	17			1960's	0
19		1970's	84			1970's	13			1970's	84			1970's	2
20		1980's	188			1980's	59			1980's	188			1980's	4
21		1990's	373			1990's	79			1990's	373	g		1990's	37
22		2000's	463			2000's	84			2000's	463			2000's	132
23		2010's	22			2010's	7			2010's	22			2010's	7
24															
25	Indie	1920's	0		Classical	1920's	0		Folk	1920's	0		Pop	1920's	0
26		1950's	0			1950's	0			1950's	0			1950's	5
27		1960's	0			1960's	0			1960's	8			1960's	34
28		1970's	0			1970's	8			1970's	14			1970's	83
29		1980's	8			1980's	11			1980's	14			1980's	159
30		1990's	32			1990's	13			1990's	45			1990's	291
31		2000's	52			2000's	6			2000's	59			2000's	476
32		2010's	7			2010's	0			2010's	2			2010's	25
33															
34															
35	Country	1920's	0		Electronic	1920's	0		Blues	1920's	0				
36		1950's	0			1950's	0			1950's	0				
37		1960's	0			1960's	0			1960's	0				
38		1970's	4			1970's	18			1970's	0				
39		1980's	4			1980's	88			1980's	13				
40		1990's	14			1990's	170			1990's	11				
41		2000's	12			2000's	292			2000's	9				
42		2010's	0			2010's	12			2010's	0				

- What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?

  Tempo of a song averaged over each Genre did not provide terribly interesting insights into the data. All of the genres had about that same average Tempo, of about 120 beats per minute, plus or minus about five beats per minute. An interesting insight gathered from the Tempo of music versus the decade it was created in: music created in the 1950's was about 20 beats per minute faster than music created from the 1960's to 2010's. If you look to the collection of Genres and the number of songs created during each decade, you will see that there is a massive variability between each Genre and each decade. Ideally, our data would be roughly equally distributed, which would give us the most accurate Tempo information for the Tempo by Genre and Tempo by Decade.

- If you have more time, what would you propose you can do next?

  If given more time to work on this project, the best way to enhance our overall exploration would be to collect data more equally. By this I mean that we would have roughly equal counts of each genre in each decade. This would limit the variability amongst counts, which would give a more representative view of the significant differences between Genres, Tempos, and Production date.

  Specifically, with classification, we would put more thought into classification feature generation for our set of data. For features like tempo and duration, we took their values and used them in a raw manner. However, pre-processing can be applied to the raw values to form more representative features. These new features can provide us with more flexible data analysis by generalizing the raw data into groups. For example, the tempo feature can be categorized into its string equivalence (Allegretto is categorized as tempos ranging from 112bpm to 120bpm). These new "looser" descriptions can provide us with the opportunity to get a better idea of the average tempo for a given genre (i.e. a good majority of "rock" songs have a tempo ranging from 100bpm - 160bpm and thus can be better categorized as Allegretto and Allegretto Moderato). Further classification optimizations could arise given these changes.