Technical perspective:

1. Cleanup dataset:

Some columns have few information or mostly null. Those columns were removed to simplify the dataset. Some rows as well were eliminated since they had no information. The rows with no information in their specific feature columns (track_listens, track_date…) were removed as well.

Drop tracks with very low value in column track_bit_rate (represents low track quality < 24000)

2. Feature Engineering:

- After analyzing the data, I realized some data such as track_interest have widespread range. Few values were above 300,000 and some have very low value. Therefore I decided to divide all these values into seven categories
- The dataset did not include price. I calculated each track's price based on its interest and quality factor.
- Year was extracted from track_date_created. In cases which track_date_created was null, album_date_created was used
- Dataset includes a column track_genres_all including list of genre ids. The descriptions were extracted to produce meaningful report (project/code/ecxploration/Extract Genre Description.ipynb).
- Track_duration was per seconds. Some tracks are very long since they are live performances. A new column, track_length, is defined dividing track_duration to four different ranges:
    1. less than three minutes
    2. between three and five minutes
    3. between five and seven minutes
    4. longer than 7 minutes

3. Modeling:

3.1. Model Training - Naive Bayes based on track_interest.ipynb

- Started modeling using track_interest as target.
- First set the following features as X columns: artist, track listens, quality, album listens
- Model runs very slow and accuracy is also very low
- Tried different combinations of features, none had good precision
- The highest result of 0.18% was achieved using track_year and track_listens as X_columns

3.2 Model Training- Track Interest and Track Listens.ipynb

- Tried running other models with track_interest
- Precision still low
- RandomForest would run out of memory

	model	precision	recall
2	DecisionTreeClassifier	0.002644	0.002644
1	KNeighborsClassifier	0.001841	0.001841
0	Naive Bayes	0.001275	0.001275

3.3 Model Training - Track Listens-Track Created Year-Artist.ipynb

- modeling using interest_factor as target.
- set the following features as X columns: artist, track listens, track year created
- Model runs very fast and accuracy is very high
- Random forest classifier with 100 estimators has highest precision

	model	precision	recall
2	RandomForestClassifier100	0.823513	0.823513
1	RandomForestClassifier10	0.821955	0.821955
4	DecisionTreeClassifier	0.810104	0.810104
3	KNeighborsClassifier	0.728234	0.728234
0	Naive Bayes	0.640085	0.640085

3.4 Model Training - Track Listens-Track Created Year and interest factor.ipynb

- modeling using interest_factor as target.
- set the following features as X columns: track listens, track year created
- Model runs very fast and accuracy is very high but slightly less than modeling including artist as well
- KNeighborClassifier has highest precision

	model	precision	recall
3	KNeighborsClassifier	0.822380	0.822380
4	DecisionTreeClassifier	0.812040	0.812040
2	RandomForestClassifier100	0.806091	0.806091
1	RandomForestClassifier10	0.802738	0.802738
0	Naive Bayes	0.653919	0.653919

3.5 Model Training - Track Listens,Created Year,-Artist,bit_rate_factor and Interest_factor

- Modeling using interest_factor as target.
- set the following features as X columns: track listens, track year created, artist, and bit_rate_factor(quality)
- Model runs very fast and accuracy is very high but 1% less than modeling including all factors but without bit_rate_factor
- Random forest classifier with 100 estimators has highest precision

	model	precision	recall
2	RandomForestClassifier100	0.816006	0.816006
4	DecisionTreeClassifier	0.802786	0.802786
1	RandomForestClassifier10	0.784466	0.784466
3	KNeighborsClassifier	0.728281	0.728281
0	Naive Bayes	0.640368	0.640368

3.6 Model Training - Track Listens and Interest_factor

- Modeling using interest_factor as target.
- set only track listens as X columns
- Model runs very fast and accuracy is high but less than modeling including all factors
- DecisionTreeClassifier has highest precision

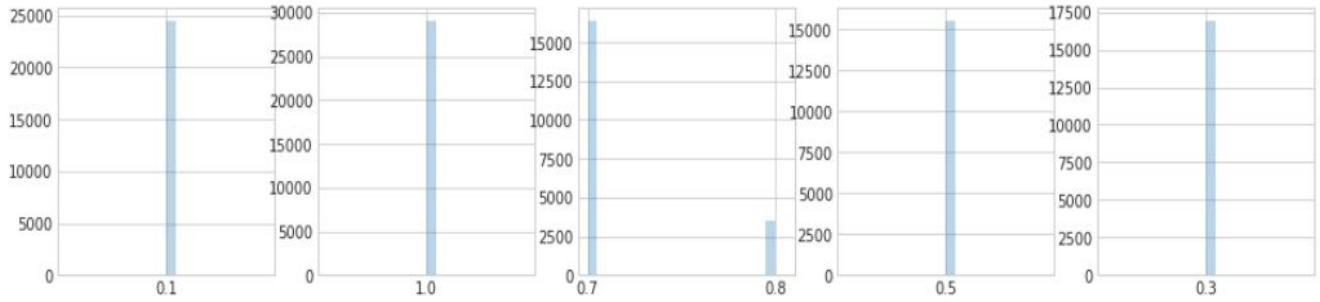
	model	precision	recall
4	DecisionTreeClassifier	0.745940	0.745940
2	RandomForestClassifier100	0.739849	0.739849
1	RandomForestClassifier10	0.731350	0.731350
3	KNeighborsClassifier	0.697545	0.697545
0	Naive Bayes	0.675968	0.675968

4 Clustering:

4.1 Model Training - Clustering Listen factor versus Ranking

- **K-Means with k=5**
- It shows the second cluster has the highest ranking. This cluster includes tracks with highest number of been listened.
- The first cluster has the lowest ranking which includes tracks being listened less than 1000 times

```
[('3', 11614), ('5', 6203), ('4', 3719), ('2', 2951), ('1', 5)]
[('1', 24720), ('2', 3176), ('3', 919), ('4', 169), ('5', 44)]
[('2', 12294), ('3', 3744), ('1', 2094), ('4', 1332), ('5', 469)]
[('4', 7745), ('3', 3911), ('5', 3865), ('2', 11)]
[('5', 12097), ('4', 4785), ('3', 30), ('2', 2)]
interest_factor
```



4.2 Model Training - Clustering-genres versus interest factor

- **K-Means with k=5**

  Rock, Folk, and electronic are always in top three of each cluster. The below graphs shows some have very high ranking and some low.

Counter({1: 28973, 0: 24284, 3: 16713, 8: 16325, 4: 15399, 5: 3545})

0: [('Rock', 1562), ('Folk', 1210), ('Electronic', 1100), ('Post-Punk,Rock,Punk', 837), ('Hip-Hop', 741)]

1: [('Rock', 1824), ('Folk', 1386), ('Electronic', 1304), ('Post-Punk,Rock,Punk', 1226), ('Hip-Hop', 799)]

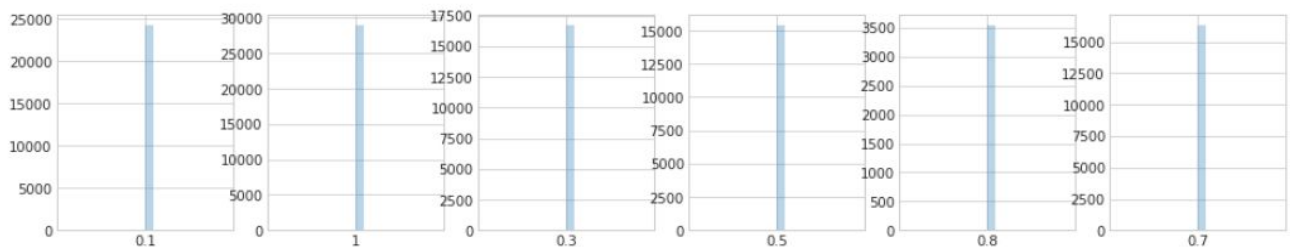
2: [('Rock', 1220), ('Electronic', 888), ('Folk', 752), ('Post-Punk,Rock,Punk', 583), ('Hip-Hop', 552)]

3: [('Rock', 1109), ('Folk', 817), ('Electronic', 702), ('Post-Punk,Rock,Punk', 581), ('Hip-Hop', 475)]

4: [('Rock', 198), ('Electronic', 179), ('Folk', 164), ('Post-Punk,Rock,Punk', 133), ('Hip-Hop', 83)]

5: [('Rock', 1048), ('Folk', 770), ('Electronic', 730), ('Post-Punk,Rock,Punk', 614), ('Hip-Hop', 470)]

```
interest_factor
```

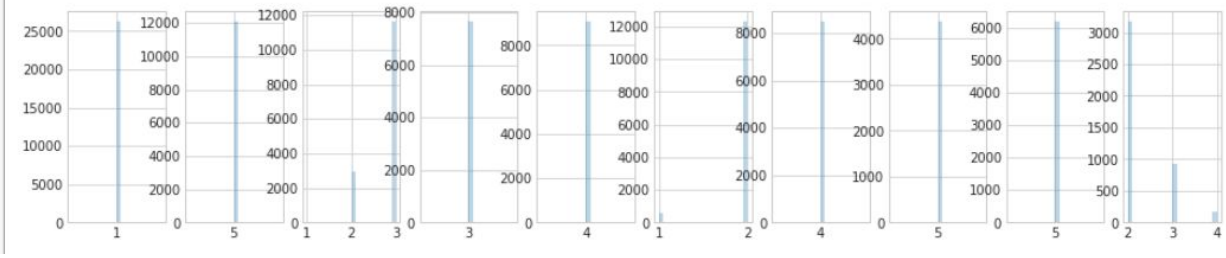


4.3 Model Training - Clustering-Track Year versus interest & listens

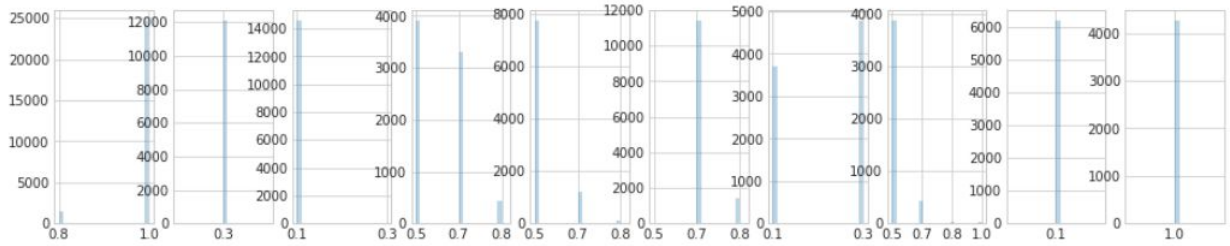
- **K-Means with k=10**

  First cluster has highest ranking and listen factor. Mostly includes tracks created after 2012.
  Second cluster has low ranking and listen factor. Mostly tracks released before 2012

```
[('2016', 5268), ('2015', 4833), ('2013', 3773), ('2014', 3189), ('2012', 3020)]
[('2010', 2146), ('2012', 1603), ('2013', 1561), ('2011', 1525), ('2009', 1192)]
[('2016', 2363), ('2015', 2221), ('2013', 1921), ('2014', 1623), ('2010', 1534)]
[('2009', 1221), ('2010', 1006), ('2012', 966), ('2013', 903), ('2016', 891)]
[('2010', 1480), ('2011', 1301), ('2013', 1198), ('2012', 1144), ('2009', 1098)]
[('2016', 2249), ('2015', 2143), ('2013', 1761), ('2014', 1564), ('2012', 1475)]
[('2009', 1335), ('2010', 1261), ('2013', 1032), ('2011', 1021), ('2012', 890)]
[('2010', 944), ('2009', 853), ('2011', 731), ('2013', 571), ('2012', 547)]
[('2009', 1049), ('2017', 1001), ('2016', 869), ('2010', 849), ('2011', 693)]
[('2009', 882), ('2011', 795), ('2010', 792), ('2012', 775), ('2013', 471)]
listens_factor
```



```
interest_factor
```



4.4 Model Training - Clustering-Track Year versus Genres

- **K-Means with k=10**

  Shows which genres have been more released each year

2010 the worst year (many electronics, hip-hop, experimental created)
After 2012 highest rank (mostly rock, folk, electronics)

Counter({4: 14471, 3: 13682, 1: 12919, 5: 12717, 6: 12346, 0: 11219, 2: 11030, 7: 9769, 8: 4233, 9: 2853})

0: [('Post-Punk,Rock,Punk', 969), ('Folk', 769), ('Rock', 635), ('Electronic', 436), ('Indie-Rock,Lo-Fi,Rock', 423), ('Avant-Garde,Experimental', 354), ('Punk,Rock', 350), ('Noise,Experimental', 308), ('Lo-Fi,Rock', 236), ('Folk,Pop,Experimental Pop,Singer-Songwriter', 216)]

1: [('Rock', 1042), ('Electronic', 626), ('Folk', 546), ('Hip-Hop', 377), ('Punk,Rock', 289), ('Noise,Experimental', 274), ('Experimental', 222), ('Indie-Rock,Rock', 219), ('Ambient Electronic,Electronic', 207), ('Post-Punk,Rock,Punk', 193)]

2: [('Hip-Hop', 441), ('Electronic', 429), ('Folk', 334), ('Rock', 311), ('Old-Time / Historic', 303), ('Punk,Rock', 276), ('Experimental', 221), ('Ambient Electronic,Electronic', 193), ('Indie-Rock,Rock', 181), ('Noise,Experimental,Electronic', 174)]

3: [('Rock', 1034), ('Electronic', 692), ('Folk', 563), ('Hip-Hop', 489), ('Punk,Rock', 374), ('Old-Time / Historic', 303), ('Avant-Garde,Experimental', 295), ('Indie-Rock,Rock', 250), ('Experimental', 232), ('Post-Punk,Rock,Punk', 227)]

4: [('Rock', 1656), ('Electronic', 901), ('Folk', 733), ('Hip-Hop', 541), ('Punk,Rock', 461), ('Indie-Rock,Rock', 368), ('Noise,Experimental', 350), ('Ambient Electronic,Electronic', 340), ('Rock,Garage', 285), ('Experimental,Radio,Spoken,Field Recordings', 274)]

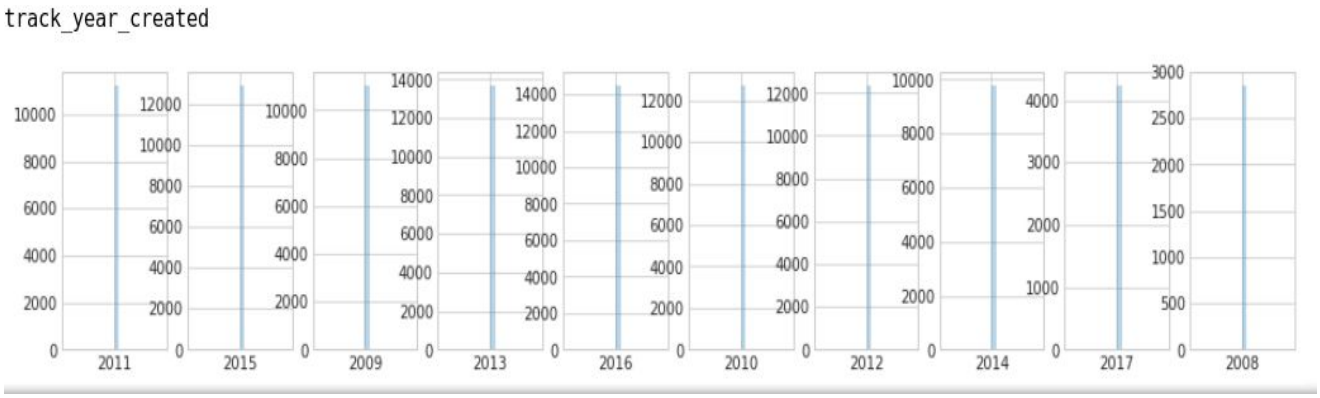
5: [('Electronic', 360), ('Hip-Hop', 313), ('Experimental', 292), ('Pop,Experimental Pop', 223), ('Chip Music,Electronic', 157), ('Chiptune,Chip Music,Electronic', 154), ('Rock', 138), ('Rock,Garage', 129), ('Indie-Rock,Rock', 117), ('Noise,Lo-Fi,Rock,Experimental', 108)]

6: [('Post-Punk,Rock,Punk', 2188), ('Folk', 1700), ('Rock', 1258), ('Indie-Rock,Lo-Fi,Rock', 928), ('Avant-Garde,Experimental', 806), ('Electronic', 800), ('Punk,Rock', 625), ('Lo-Fi,Rock', 516), ('Experimental,Field Recordings', 429), ('Noise,Experimental', 429)]

7: [('Electronic', 309), ('Hip-Hop', 281), ('Pop,Experimental Pop', 208), ('Experimental', 192), ('Chiptune,Chip Music,Electronic', 138), ('Chip Music,Electronic', 134), ('Rock,Garage', 128), ('Noise,Experimental,Electronic', 118), ('Rock', 93), ('Indie-Rock,Rock', 92)]

8: [('Old-Time / Historic', 229), ('Hip-Hop', 159), ('Avant-Garde,Experimental', 146), ('Electronic', 105), ('Punk,Rock', 104), ('Rock', 102), ('Dance,Pop,Electronic', 89), ('Experimental', 87), ('Techno,Electronic', 72), ('Folk', 72)]

9: [('Rock', 692), ('Electronic', 245), ('Folk', 206), ('Post-Punk,Rock,Punk', 108), ('Avant-Garde,Experimental', 103), ('Experimental,Field Recordings', 102), ('Punk,Rock', 96), ('Noise,Experimental', 96), ('Rock,Garage', 89), ('Spoken Weird,Spoken', 75)]



2008: variety of genres were released such as post-punk, avant-garde, experimental. Also including Rock, Electronic and Folk which had highest interest

2009: hip-hop, electronic, Folk, Rock

2010: variety of genres were released but not much interest. Few Rock, no Folk

2011 and 2012 toward punk, post-punk but still many Rock and Folk

2013: [('Rock', 1034), ('Electronic', 692), ('Folk', 563)

2014: 'Electronic', 309), ('Hip-Hop', 281), ('Pop,Experimental Pop'

2015: [('Rock', 1042), ('Electronic', 626), ('Folk'

2016: [('Rock', 1042), ('Electronic', 626), ('Folk'

2017: [('Old-Time / Historic', 229), ('Hip-Hop', 159), ('Avant-Garde,Experimental',