eda

November 19, 2021

1 Exploratory Data Analysis of Spotify tracks!

1.0.1 Hope these tracks can be popular ...

[3]:
```
import pandas as pd
from pandas_profiling import ProfileReport
import altair as alt
```

1. Formulate our question

We want to predict the popularity of a song, given various features such as genre, duration, energy, tempo and acousticness. Can our raw data do this for us?

2. Read in the data

[85]:
```
audio = pd.read_csv('../data/audio_features.csv')
```

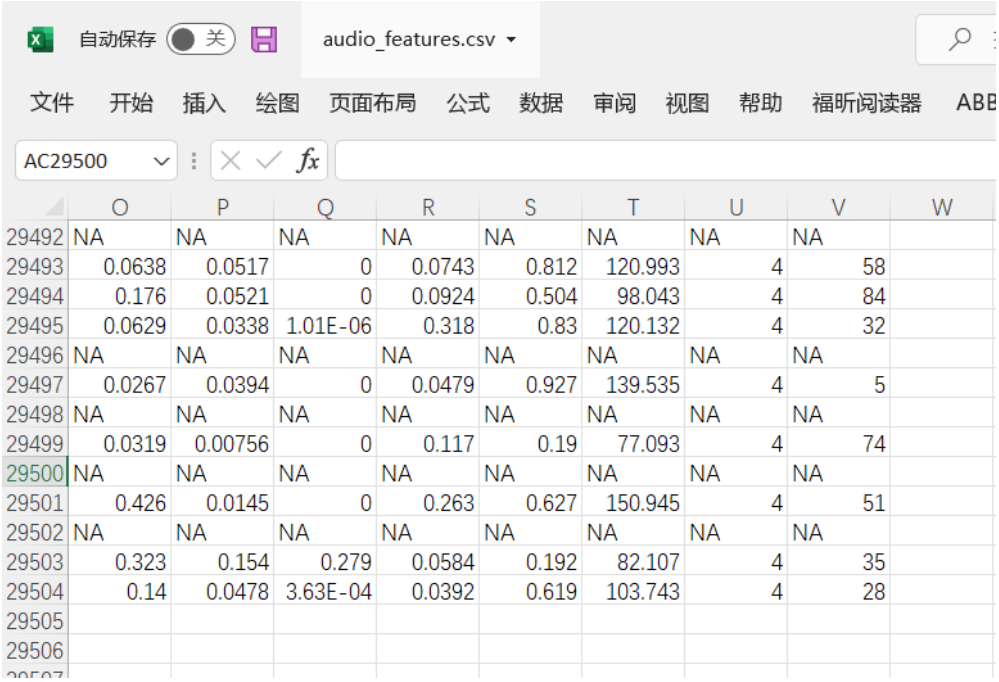
3. Check the packaging

[19]:
```
audio.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29503 entries, 0 to 29502
Data columns (total 22 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   song_id                    29503 non-null  object
 1   performer                  29503 non-null  object
 2   song                       29503 non-null  object
 3   spotify_genre              27903 non-null  object
 4   spotify_track_id           24397 non-null  object
 5   spotify_track_preview_url  14491 non-null  object
 6   spotify_track_duration_ms  24397 non-null  float64
 7   spotify_track_explicit     24397 non-null  object
 8   spotify_track_album        24391 non-null  object
 9   danceability               24334 non-null  float64
 10  energy                     24334 non-null  float64
 11  key                        24334 non-null  float64
 12  loudness                   24334 non-null  float64
 13  mode                       24334 non-null  float64
```

```
 14  speechiness                 24334 non-null  float64
 15  acousticness                24334 non-null  float64
 16  instrumentalness            24334 non-null  float64
 17  liveness                    24334 non-null  float64
 18  valence                     24334 non-null  float64
 19  tempo                       24334 non-null  float64
 20  time_signature              24334 non-null  float64
 21  spotify_track_popularity    24397 non-null  float64
dtypes: float64(14), object(8)
memory usage: 5.0+ MB
```

We have 22 columns and 29503 rows in the data. Though there are some missing data, this dataset can support us to build a predictive model. We can check with the original data, they match!



	O	P	Q	R	S	T	U	V	W
29492	NA	NA	NA	NA	NA	NA	NA	NA	
29493	0.0638	0.0517	0	0.0743	0.812	120.993	4	58	
29494	0.176	0.0521	0	0.0924	0.504	98.043	4	84	
29495	0.0629	0.0338	1.01E-06	0.318	0.83	120.132	4	32	
29496	NA	NA	NA	NA	NA	NA	NA	NA	
29497	0.0267	0.0394	0	0.0479	0.927	139.535	4	5	
29498	NA	NA	NA	NA	NA	NA	NA	NA	
29499	0.0319	0.00756	0	0.117	0.19	77.093	4	74	
29500	NA	NA	NA	NA	NA	NA	NA	NA	
29501	0.426	0.0145	0	0.263	0.627	150.945	4	51	
29502	NA	NA	NA	NA	NA	NA	NA	NA	
29503	0.323	0.154	0.279	0.0584	0.192	82.107	4	35	
29504	0.14	0.0478	3.63E-04	0.0392	0.619	103.743	4	28	
29505									
29506									

4. Look at the Top and the Bottom of our Data

[20]:
```
audio.head()
```

[20]:
```
                                              song_id           performer  \
0     -twistin'-White Silver SandsBill Black's Combo  Bill Black's Combo
1  ¿Dònde Està Santa Claus? (Where Is Santa Claus…          Augie Rios
2             …And Roses And RosesAndy Williams       Andy Williams
3          …And Then There Were DrumsSandy Nelson        Sandy Nelson
4               …Baby One More TimeBritney Spears      Britney Spears

                                          song  \
0                    -twistin'-White Silver Sands
1  ¿Dònde Està Santa Claus? (Where Is Santa Claus?)
2                         …And Roses And Roses
```

```
3                      …And Then There Were Drums
4                             …Baby One More Time

                                      spotify_genre        spotify_track_id  \
0                                                []                     NaN
1                                      ['novelty']                     NaN
2  ['adult standards', 'brill building pop', 'eas…  3tvqPPpXyIgKrm4PR9HCf0
3   ['rock-and-roll', 'space age pop', 'surf music']  1fHHq3qHU8wpRKHzhojZ4a
4              ['dance pop', 'pop', 'post-teen pop']  3MjUtNVVq3C8Fn0MP3zhXa

                          spotify_track_preview_url  \
0                                               NaN
1                                               NaN
2  https://p.scdn.co/mp3-preview/cef4883cfd1e0e53…
3                                               NaN
4  https://p.scdn.co/mp3-preview/da2134a161f1cb34…

   spotify_track_duration_ms spotify_track_explicit  \
0                        NaN                    NaN
1                        NaN                    NaN
2                   166106.0                  False
3                   172066.0                  False
4                   211066.0                  False

                             spotify_track_album  danceability  …  \
0                                            NaN           NaN  …
1                                            NaN           NaN  …
2                     The Essential Andy Williams         0.154  …
3                            Compelling Percussion         0.588  …
4  …Baby One More Time (Digital Deluxe Version)         0.759  …

   loudness  mode  speechiness  acousticness  instrumentalness  liveness  \
0       NaN   NaN          NaN           NaN               NaN       NaN
1       NaN   NaN          NaN           NaN               NaN       NaN
2   -14.063   1.0       0.0315       0.91100          0.000267     0.112
3   -17.278   0.0       0.0361       0.00256          0.745000     0.145
4    -5.745   0.0       0.0307       0.20200          0.000131     0.443

   valence    tempo  time_signature  spotify_track_popularity
0      NaN      NaN             NaN                       NaN
1      NaN      NaN             NaN                       NaN
2    0.150   83.969             4.0                      38.0
3    0.801  121.962             4.0                      11.0
4    0.907   92.960             4.0                      77.0

[5 rows x 22 columns]
```

[21]:
```
audio.tail()
```

[21]:
```
                                                song_id  \
29498  Zoo YorkLil Tjay Featuring Fivio Foreign & Pop…
29499                                        ZoomFuture
29500                 ZoomLil' Boosie Featuring Yung Joc
29501    Zorba The GreekHerb Alpert & The Tijuana Brass
29502                                     Zunga ZengK7

                                        performer            song  \
29498  Lil Tjay Featuring Fivio Foreign & Pop Smoke        Zoo York
29499                                        Future            Zoom
29500               Lil' Boosie Featuring Yung Joc            Zoom
29501              Herb Alpert & The Tijuana Brass  Zorba The Greek
29502                                          K7       Zunga Zeng

                                        spotify_genre  \
29498                                             NaN
29499  ['atl hip hop', 'hip hop', 'pop rap', 'rap', '…
29500         ['baton rouge rap', 'deep southern trap']
29501   ['adult standards', 'easy listening', 'lounge']
29502                                     ['freestyle']

           spotify_track_id  \
29498                   NaN
29499  2IG6Te7JyvrtqhFeOF7le4
29500                   NaN
29501  3WLEVNohakzZmMpN5W7mHK
29502  0XevPPcCBPovknaBw3lFvh

                           spotify_track_preview_url  \
29498                                            NaN
29499  https://p.scdn.co/mp3-preview/cb8fde6edc08e70a…
29500                                            NaN
29501  https://p.scdn.co/mp3-preview/1841a4034ba42fc0…
29502  https://p.scdn.co/mp3-preview/8d5174aeb7d6b740…

       spotify_track_duration_ms spotify_track_explicit spotify_track_album  \
29498                        NaN                    NaN                NaN
29499                   278429.0                   True              FUTURE
29500                        NaN                    NaN                NaN
29501                   264853.0                  False  !!!Going Places!!!
29502                   273000.0                  False  Swing Batta Swing!

       danceability  …  loudness  mode  speechiness  acousticness  \
29498           NaN  …       NaN   NaN          NaN           NaN
29499         0.852  …    -7.673   1.0        0.426        0.0145
```

```
29500           NaN  …       NaN   NaN          NaN           NaN
29501         0.531  …   -12.702   1.0        0.323        0.1540
29502         0.846  …    -9.642   1.0        0.140        0.0478

       instrumentalness  liveness  valence    tempo  time_signature  \
29498               NaN       NaN      NaN      NaN             NaN
29499          0.000000    0.2630    0.627  150.945             4.0
29500               NaN       NaN      NaN      NaN             NaN
29501          0.279000    0.0584    0.192   82.107             4.0
29502          0.000363    0.0392    0.619  103.743             4.0

       spotify_track_popularity
29498                       NaN
29499                      51.0
29500                       NaN
29501                      35.0
29502                      28.0

[5 rows x 22 columns]
```

5. Check our "n"s

To have a general understanding of our data, we'll use pandas profiling from here. The script that generate the eda report of a `html` version is in `src`. The `html` version eda report is here.

[26]:
```
profile = ProfileReport(audio, title="Pandas Profiling Report")  #,␣
↪minimal=True)
profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|          | 0/36 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

We can find there are some duplicate rows in the dataset, which suggests that we need to remove them later in the data analysis process.

6. Validate with at least one external data source

The website musicstax provides us all the data from spotify. However, it takes some time to crawl all the data to match ours in the database so let's just try one observation. We pick `bad bad bad` to check. Since the popularity may change time by time, we can validate other features.

[33]:
```
audio.loc[audio['song']=='Bad Bad Bad', ['song', 'danceability', 'energy',␣
↪'valence','speechiness', 'liveness', 'instrumentalness']].iloc[0, ]
```

[33]:
```
song                Bad Bad Bad
danceability              0.974
energy                    0.596
valence                   0.892
speechiness               0.184
liveness                  0.151
instrumentalness            0.0
Name: 2124, dtype: object
```

Seems not that bad :)

7. Make a plot

The distribution and the correlation of the data are in the eda report. We'll explore the relationship between `spotify_track_popularity` and the features which have at least weak correlation (pearson's r $>$ 0.1).

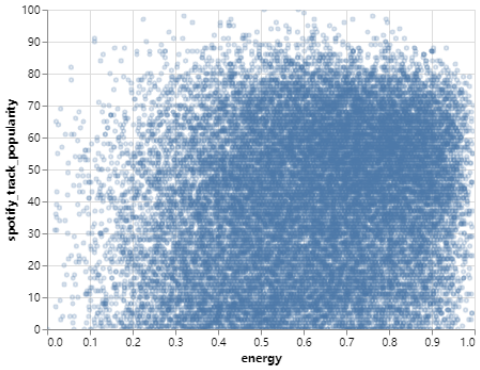
[60]:
```
# List out the correlated features
imp_features = list(audio.corr().query('spotify_track_popularity > 0.1').
 ↪sort_values(by=['spotify_track_popularity']).index)
imp_features.remove('spotify_track_popularity')
imp_features.remove('time_signature')
imp_features
```

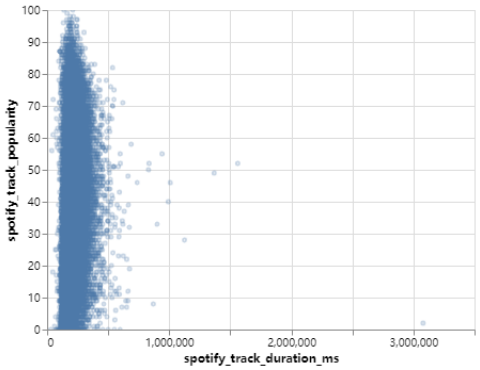
[60]:
```
['energy',
 'spotify_track_duration_ms',
 'danceability',
 'speechiness',
 'loudness']
```

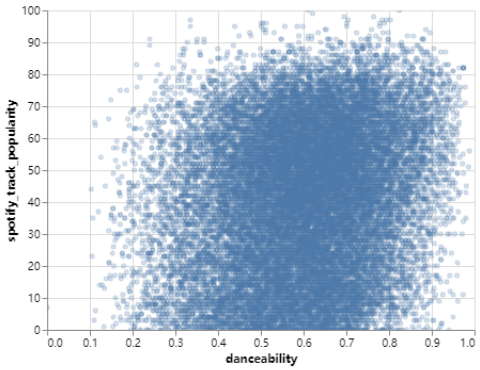
Notice that we remove `time_signature` since it is actually a categorical feature. We may have another plot for this one.

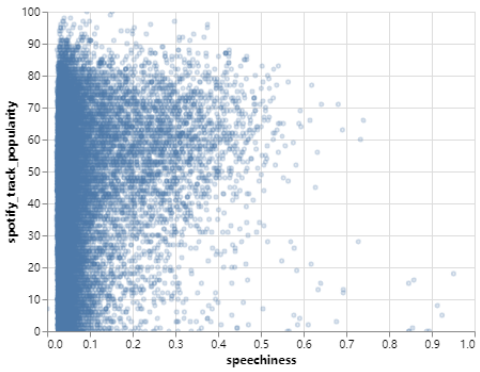
[69]:
```
# Draw them ( ´ )
# The scatter plot for numeric features
alt.renderers.enable('mimetype')
alt.data_transformers.disable_max_rows()

rela = alt.Chart(audio).mark_point(opacity=0.3, size=10).encode(
     alt.X(alt.repeat('row'), type='quantitative',
          scale=alt.Scale(zero=False)),
     alt.Y('spotify_track_popularity', type='quantitative',
          scale=alt.Scale(zero=False)),
     tooltip='song'
).repeat(
    row=imp_features
).interactive()

rela
```

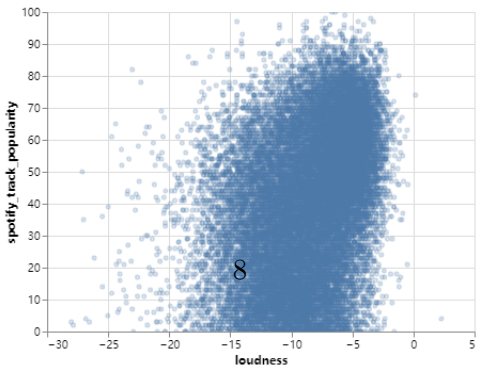
[69]:


spotify_track_popularity
energy


spotify_track_popularity
spotify_track_duration_ms

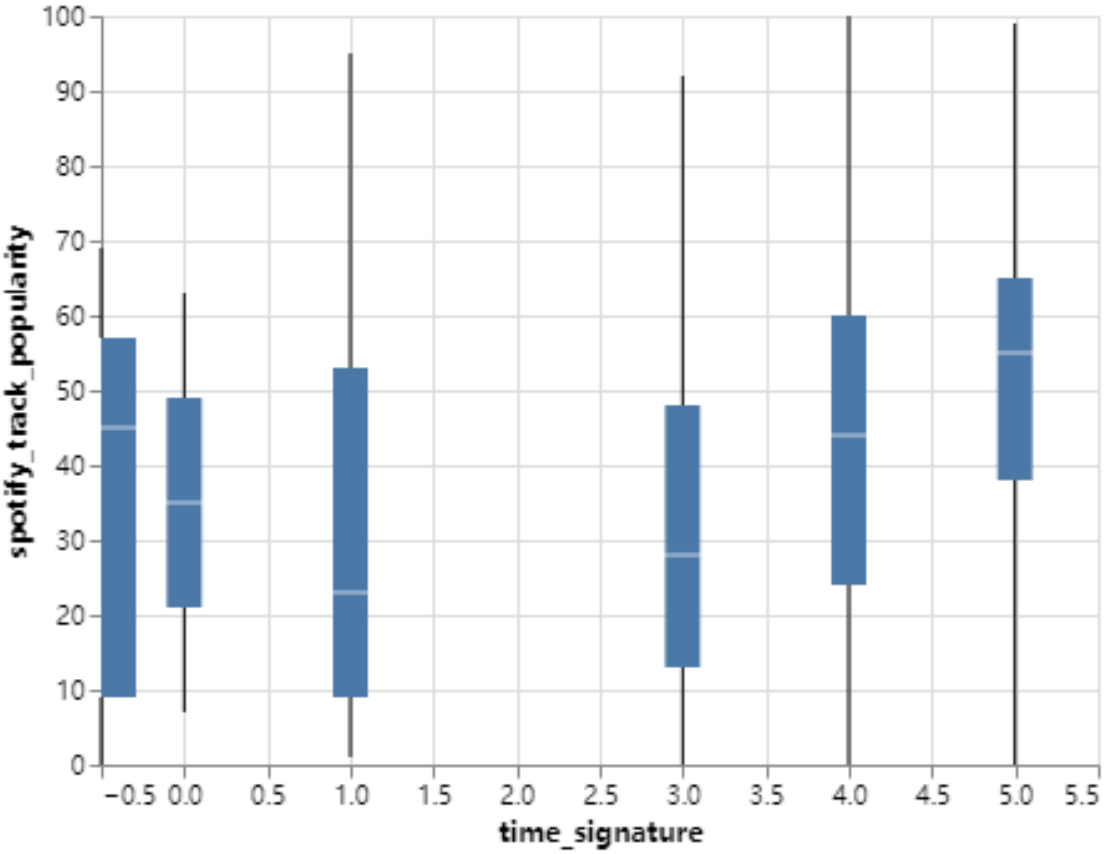

spotify_track_popularity
danceability


spotify_track_popularity
speechiness


spotify_track_popularity
loudness

[73]:
```
# The box plot for categorical features

box = alt.Chart(audio).mark_boxplot().encode(
     alt.X('time_signature', scale=alt.Scale(zero=False)),
     alt.Y('spotify_track_popularity',
          scale=alt.Scale(zero=False))
).interactive()

box
```

[73]:



8. Try the Easy Solution First

Before the machine learning models, let's build up a simple linear model using these features first.

[96]:
```
import numpy as np
from sklearn.linear_model import LinearRegression
```

```
reg_col = list(audio.corr().query('spotify_track_popularity > 0.1').
 ↪sort_values(by=['spotify_track_popularity']).index)
reg_df = audio[reg_col]
reg_df = reg_df.dropna()

reg_X = reg_df[imp_features]
reg_y = reg_df['spotify_track_popularity']
reg = LinearRegression().fit(reg_X, reg_y)
reg.score(reg_X, reg_y)
```

[96]:
```
0.19835091018875328
```

[97]:
```
reg.coef_
```

[97]:
```
array([-2.05478281e+01,  5.80294169e-05,  1.64532456e+01,  3.45413859e+01,
        2.69579132e+00])
```

9.Follow-up Questions

- Do we have the right data? > The data can be matched with external data, and it has enough features and observations for us to answer the question.
- Do we need other data? > This data may be enough.
- Do we have the right question? > We found that linear regression did not perform well on our selected features, that can be caused by distribution of the data as well as the method of the model. We can fix this by transforming the columns and changing the model e.g. ridge.