X^i Records

Can Popularity Be Predicted?

Introduction

- With Spotify being so well used is it possible to predict the popularity of song?

- Are there certain characteristics of a song that make it more popular?

- Can we predict whether a new artist has released a song to rival that of Ed Sheeran or Taylor Swift?

Let’s See What the Data Says!

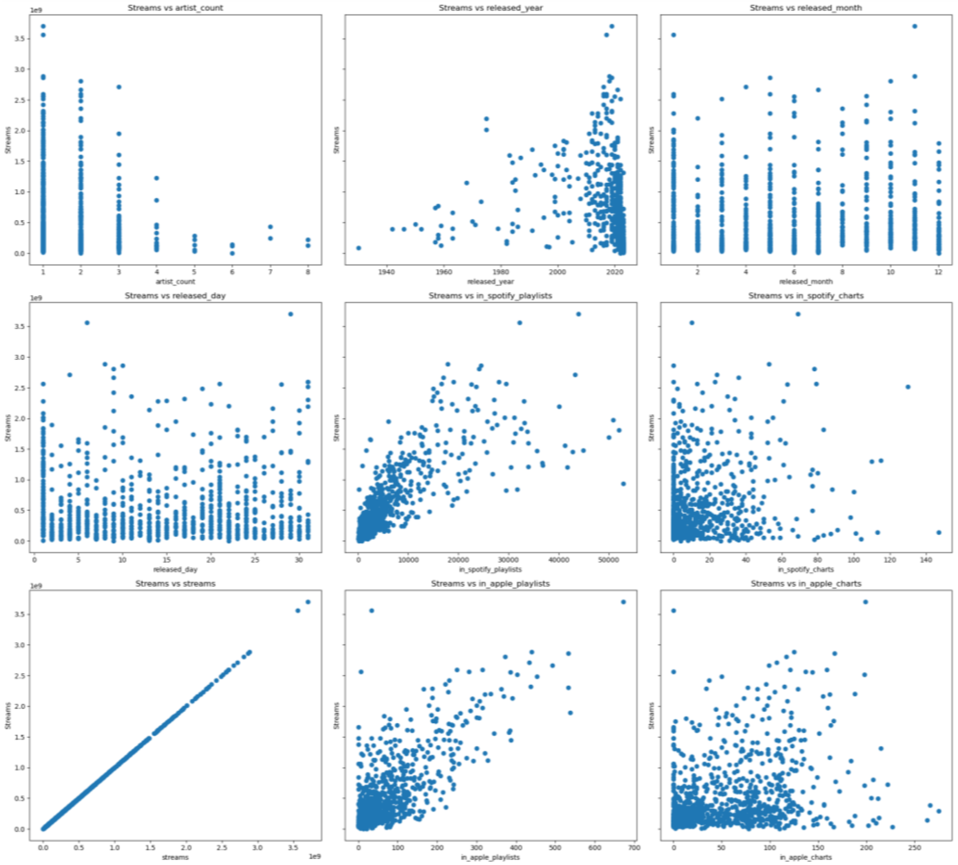
The Data

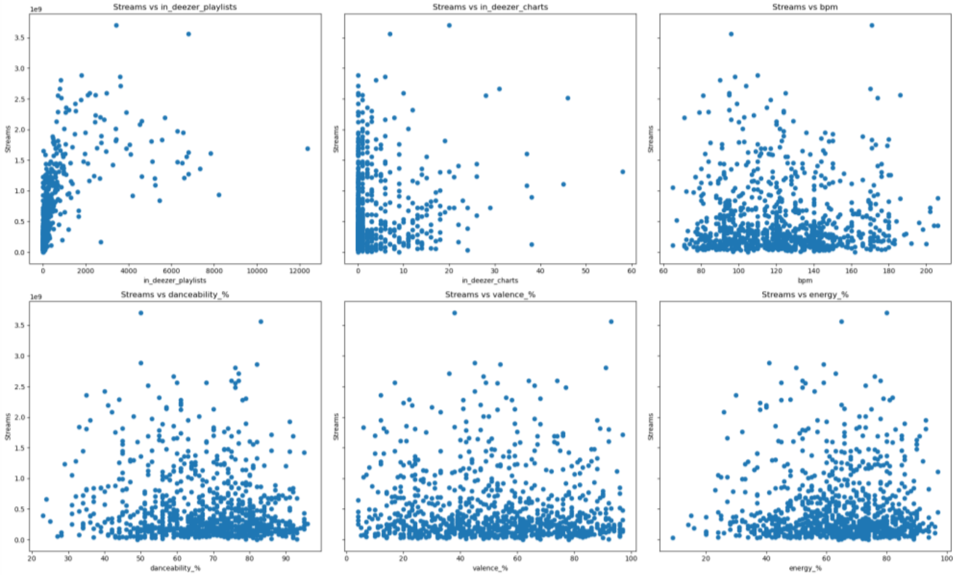
The dataset I have sourced from Kaggle and is entitled “Most Streamed Spotify Songs 2023”. It has 943 rows and 24 columns with the titles:

- **track_name:** Name of the song
- **artist(s)_name:** Name of the artist(s) of the song
- **artist_count:** Number of artists contributing to the song
- **released_year:** Year when the song was released
- **released_month:** Month when the song was released
- **released_day:** Day of the month when the song was released
- **in_spotify_playlists:** Number of Spotify playlists the song is included in
- **in_spotify_charts:** Presence and rank of the song on Spotify charts
- **streams:** Total number of streams on Spotify
- **in_apple_playlists:** Number of Apple Music playlists the song is included in
- **in_apple_charts:** Presence and rank of the song on Apple Music charts
- **in_deezer_playlists:** Number of Deezer playlists the song is included in
- **in_deezer_charts:** Presence and rank of the song on Deezer charts
- **in_shazam_charts:** Presence and rank of the song on Shazam charts
- **bpm:** Beats per minute, a measure of song tempo
- **key:** Key of the song
- **mode:** Mode of the song (major or minor)
- **danceability_%:** Percentage indicating how suitable the song is for dancing
- **valence_%:** Positivity of the song's musical content
- **energy_%:** Perceived energy level of the song
- **acousticness_%:** Amount of acoustic sound in the song
- **instrumentalness_%:** Amount of instrumental content in the song
- **liveness_%:** Presence of live performance elements
- **speechiness_%:** Amount of spoken words in the song

Initial Modelling

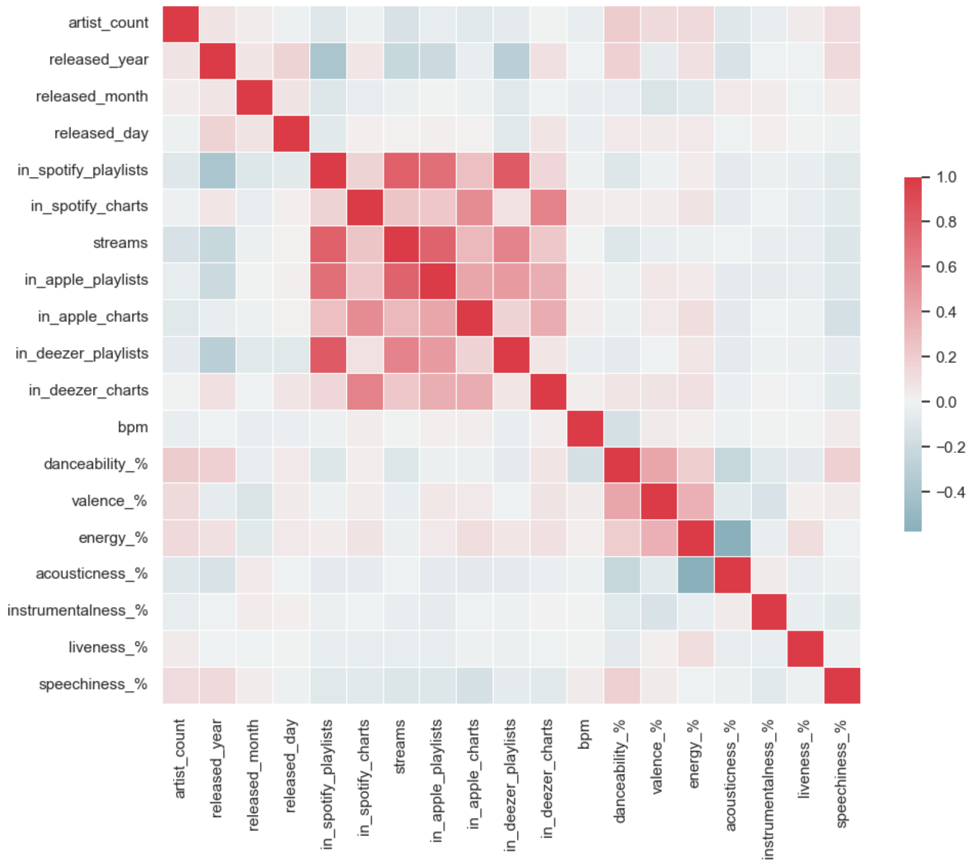
To begin with a created a simple plot of all variables against streams to get a feel for the data.





Modelling

This was followed by collinearity tests:



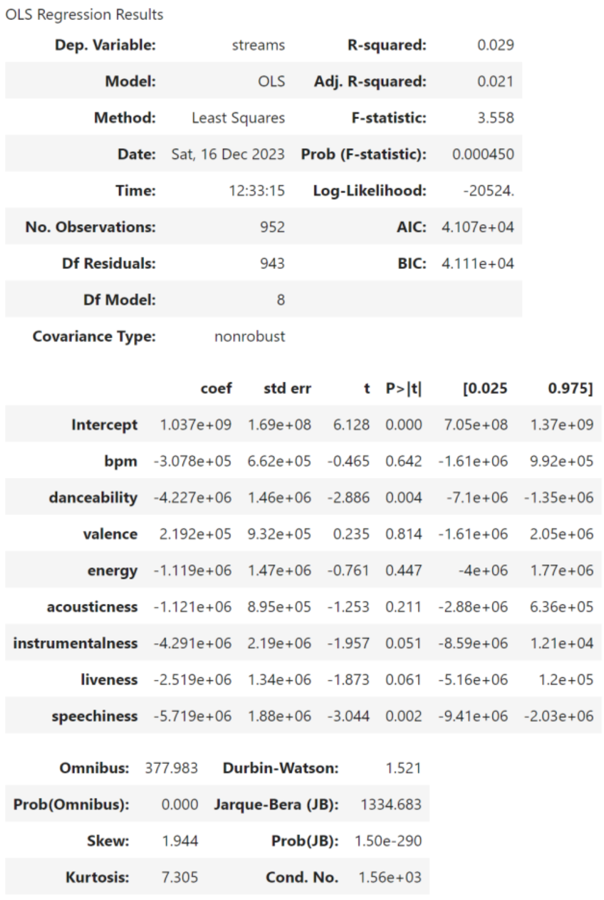
This showed that the variables with the highest collinearity with no. of streams were:

- in_spotify_playlists
- in_spotify_charts

Modelling

To answer the brief I then reduced the dataset to just necessary variables.

Unfortunately the initial OLES results weren’t favourable:

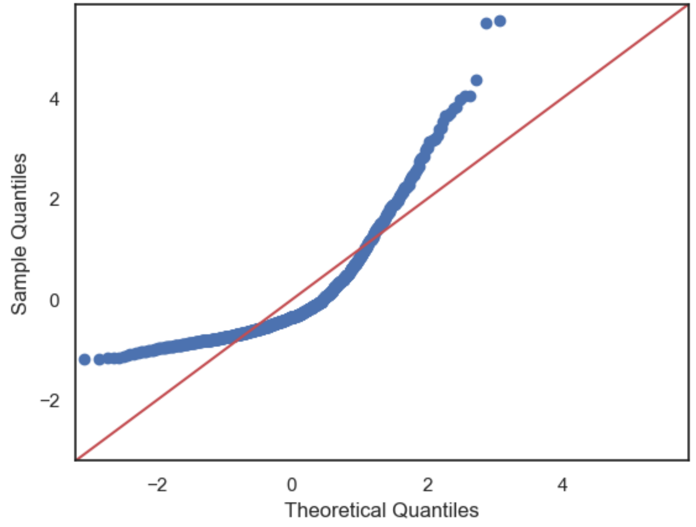


```
OLS Regression Results
       Dep. Variable:          streams        R-squared:      0.029
               Model:              OLS    Adj. R-squared:      0.021
              Method:    Least Squares       F-statistic:      3.558
                Date: Sat, 16 Dec 2023 Prob (F-statistic): 0.000450
                Time:         12:33:15   Log-Likelihood:     -20524.
    No. Observations:              952              AIC:  4.107e+04
        Df Residuals:              943              BIC:  4.111e+04
            Df Model:                8
     Covariance Type:        nonrobust

                         coef   std err       t  P>|t|     [0.025     0.975]
       Intercept    1.037e+09  1.69e+08   6.128  0.000   7.05e+08   1.37e+09
             bpm   -3.078e+05  6.62e+05  -0.465  0.642  -1.61e+06   9.92e+05
    danceability   -4.227e+06  1.46e+06  -2.886  0.004   -7.1e+06  -1.35e+06
         valence    2.192e+05  9.32e+05   0.235  0.814  -1.61e+06   2.05e+06
          energy   -1.119e+06  1.47e+06  -0.761  0.447     -4e+06   1.77e+06
    acousticness   -1.121e+06  8.95e+05  -1.253  0.211  -2.88e+06   6.36e+05
instrumentalness   -4.291e+06  2.19e+06  -1.957  0.051  -8.59e+06   1.21e+04
        liveness   -2.519e+06  1.34e+06  -1.873  0.061  -5.16e+06    1.2e+05
     speechiness   -5.719e+06  1.88e+06  -3.044  0.002  -9.41e+06  -2.03e+06

        Omnibus: 377.983     Durbin-Watson:     1.521
 Prob(Omnibus):   0.000  Jarque-Bera (JB):  1334.683
           Skew:   1.944          Prob(JB): 1.50e-290
       Kurtosis:   7.305         Cond. No.  1.56e+03
```

And the resulting Q-Q plot was not ideal.

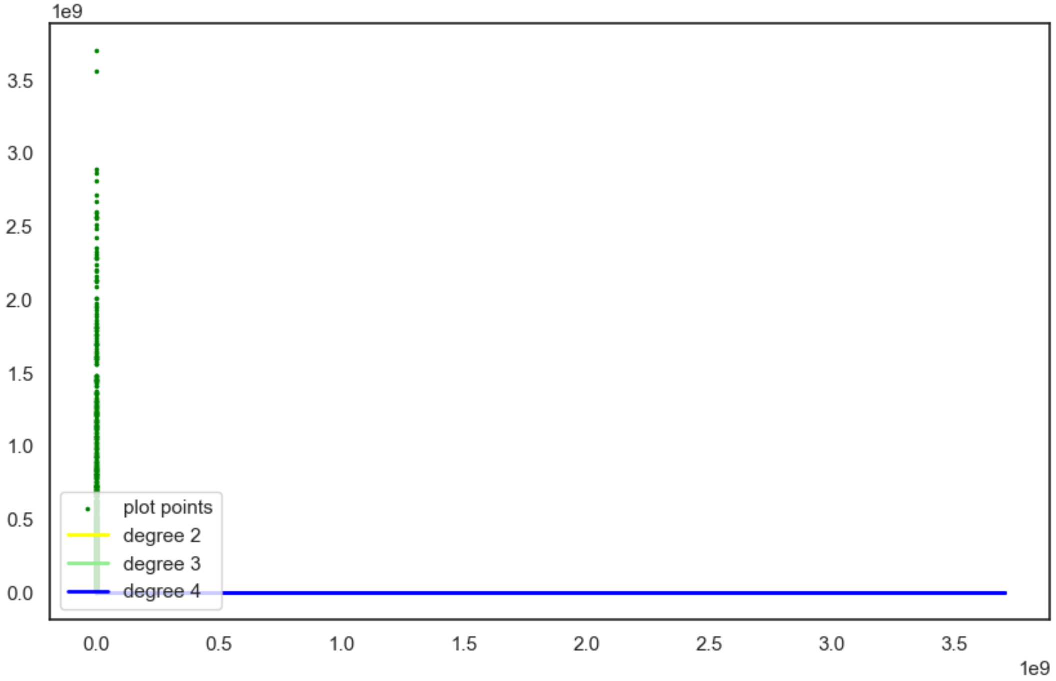


Modelling

Needing some direction I performed a stepwise selection that showed the best result would come from the two variables: speechiness and danceability.

I then proceeded create another model with only those two variables. The results did not improve. I even tried to apply polynomial regression to see if the plot could be more linear.

It was not conclusive.



Insights

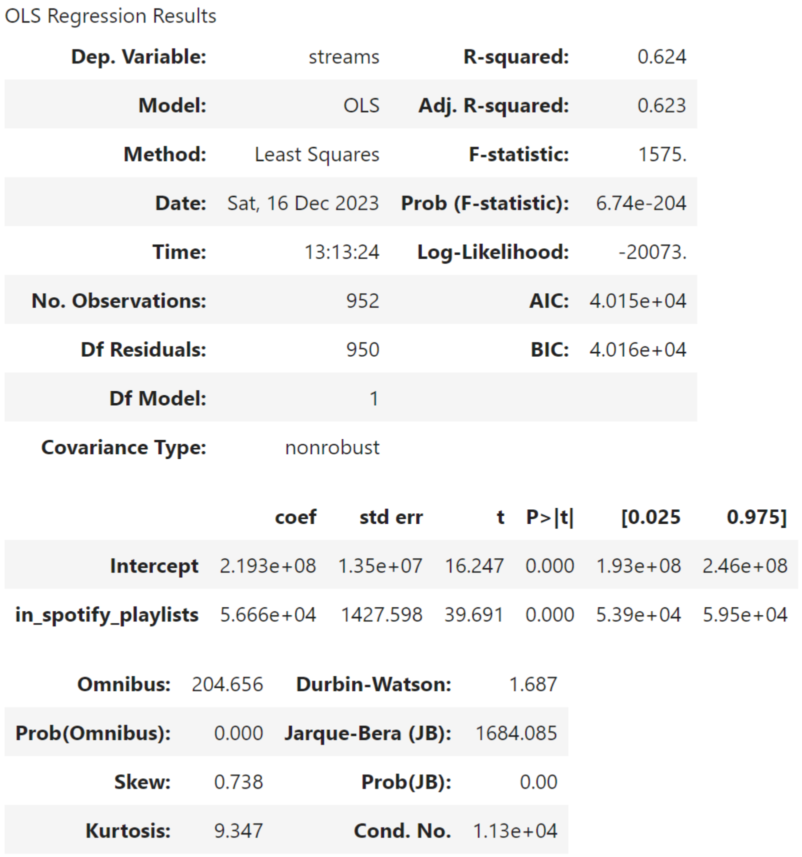
From these models I was able to conclude:

- That with this data set it is unable to be predicted whether a song would be popular from it’s characteristics using a linear regression model.

From an early collinearity test I did see a high collinearity between streams and in_spotify_playlists. I decided to create a further model around this.

Further Modelling.

This new model between streams and in_spotify_playlists had promising results:

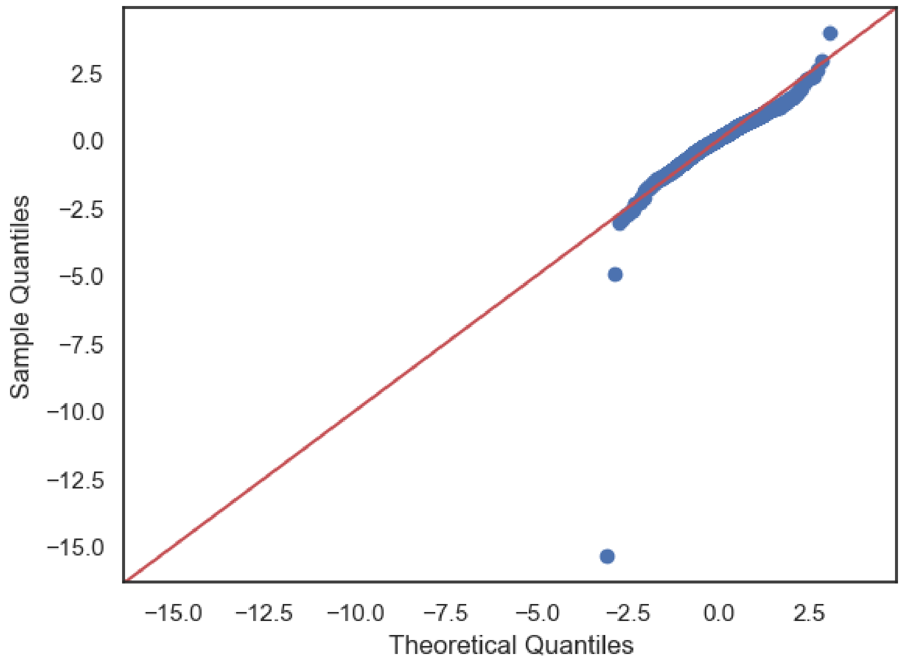


OLS Regression Results

Dep. Variable:	streams	**R-squared:**	0.624
Model:	OLS	**Adj. R-squared:**	0.623
Method:	Least Squares	**F-statistic:**	1575.
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	6.74e-204
Time:	13:13:24	**Log-Likelihood:**	-20073.
No. Observations:	952	**AIC:**	4.015e+04
Df Residuals:	950	**BIC:**	4.016e+04
Df Model:	1		
Covariance Type:	nonrobust		

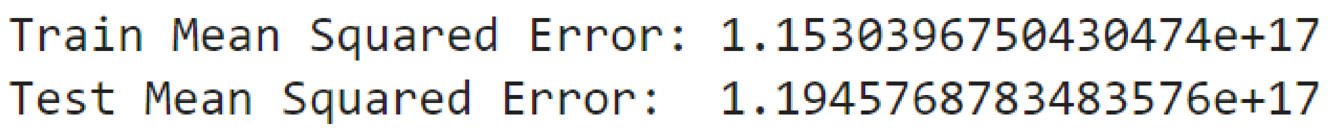
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.193e+08	1.35e+07	16.247	0.000	1.93e+08	2.46e+08
in_spotify_playlists	5.666e+04	1427.598	39.691	0.000	5.39e+04	5.95e+04

Omnibus:	204.656	**Durbin-Watson:**	1.687
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1684.085
Skew:	0.738	**Prob(JB):**	0.00
Kurtosis:	9.347	**Cond. No.**	1.13e+04



Testing

Testing was simple I split the data at 75 % to train the model and test it which resulted in:



```
Train Mean Squared Error: 1.1530396750430474e+17
Test Mean Squared Error:  1.1945768783483576e+17
```

Conclusion

Although we may not have been able to use a linear regression model to predict the popularity of a song based on its characteristics it is possible to predict the popularity of a song based on the number of playlists it appears in, using this dataset.

Thank you

Nyssa Mitchell

Data Scientist

nysmitch@gmail.com

Any Questions?

Questions?