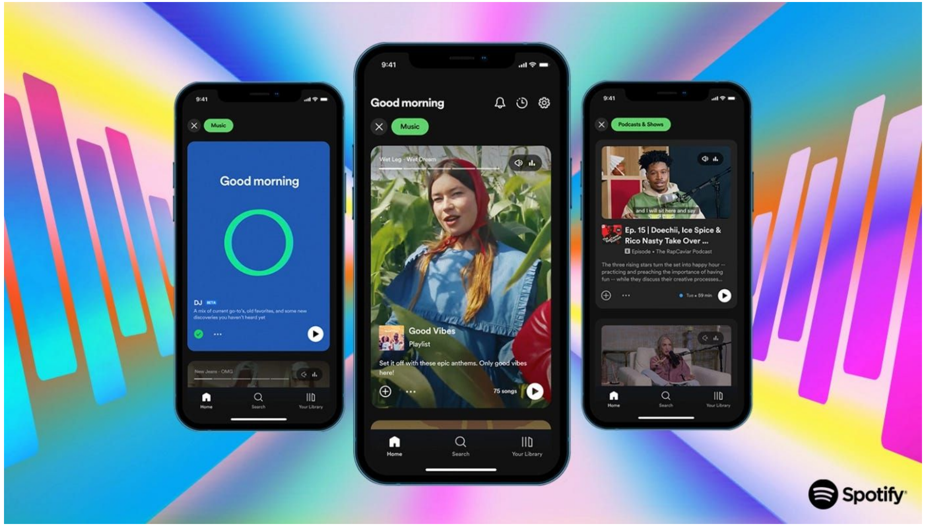
Capstone 2: Final Presentation

Ashley Kim

Background:

- Spotify is one of the top streaming platforms
- Main source of revenue = user subscription
  - Very important to increase user engagement and longevity
- This project aims to identify pre-released songs as “popular” or not by using the song characteristics

Spotify®

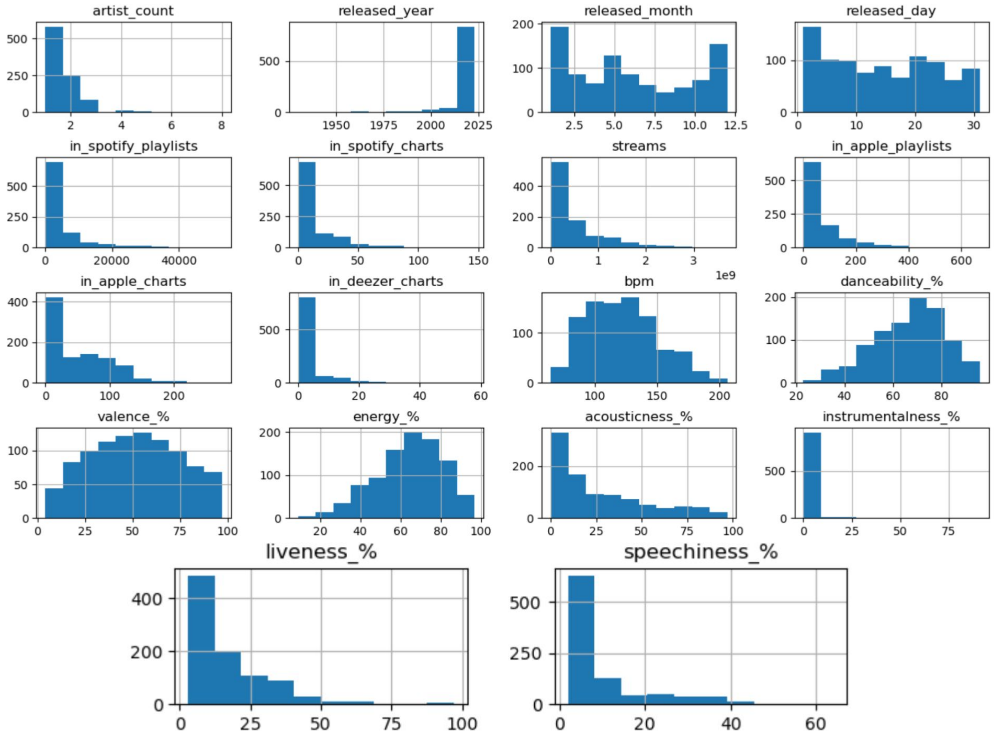


Good morning
Music
Good morning
DJ
Good Vibes
Playlist
Set it off with these epic anthems. Only good vibes here!
75 songs
Podcasts & Shows
Ep. 15 | Doechii, Ice Spice & Rico Nasty Take Over ...
Home
Search
Your Library
Spotify

Steps taken in the project:

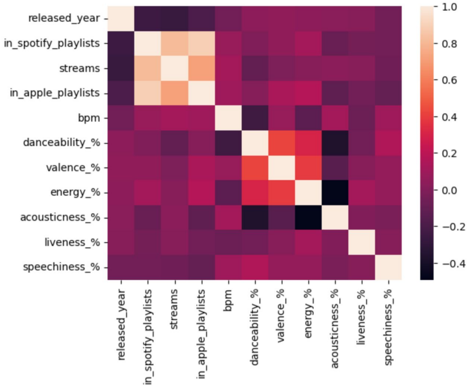
- Collect quality data
  - Used Kaggle.com
- Clean data through data wrangling
- Exploratory data analysis to analyze what the data is trying to show
- Develop machine learning model that predicts popularity of future songs

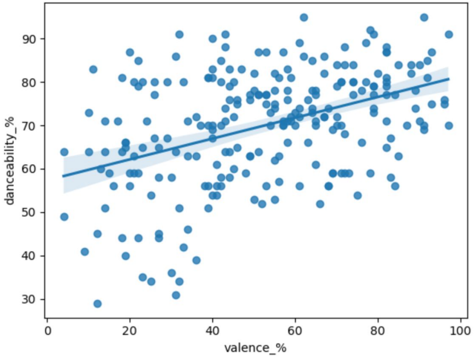
Data Wrangling:



- After cleaning up the dataset, I retrieved histograms of each numerical column
- Some important columns I further used include: streams, bpm, danceability, valence, energy, acousticness, liveness, and speechiness

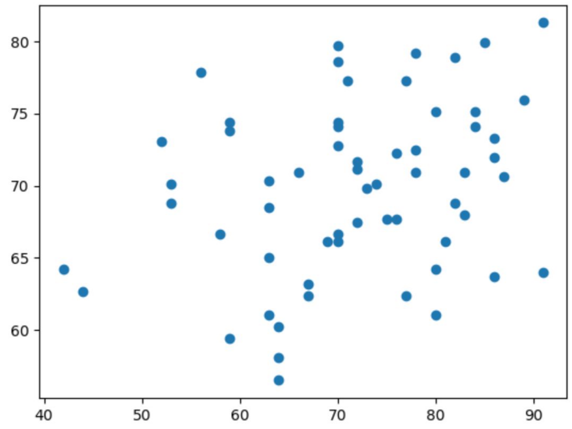
Exploratory Data Analysis:

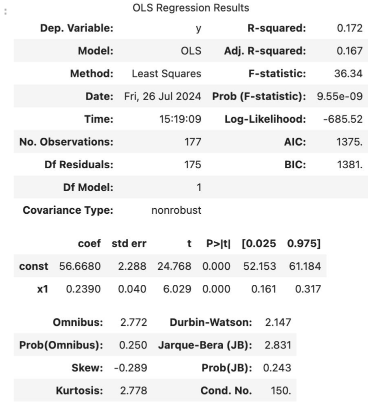




- Using a heat map I was able to identify features that correlate with one another
- As a result, I found that danceability and valence had the stronger correlation
  - Even though it was the strongest, it still was pretty weak
  - Overall, none of the features gave a strong correlation between each other

Models:

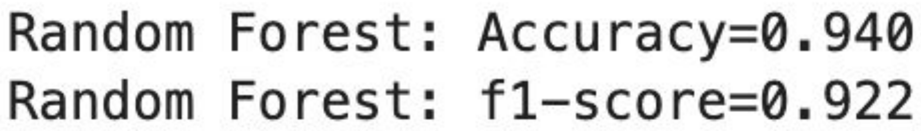




```
                         OLS Regression Results
       Dep. Variable:                 y       R-squared:      0.172
               Model:               OLS   Adj. R-squared:      0.167
              Method:     Least Squares      F-statistic:      36.34
                Date:  Fri, 26 Jul 2024 Prob (F-statistic):  9.55e-09
                Time:          15:19:09   Log-Likelihood:    -685.52
    No. Observations:               177              AIC:      1375.
        Df Residuals:               175              BIC:      1381.
            Df Model:                 1
     Covariance Type:         nonrobust

          coef   std err       t    P>|t|   [0.025   0.975]
const  56.6680     2.288  24.768    0.000   52.153   61.184
   x1   0.2390     0.040   6.029    0.000    0.161    0.317

       Omnibus:   2.772     Durbin-Watson:  2.147
Prob(Omnibus):    0.250  Jarque-Bera (JB):  2.831
          Skew:  -0.289          Prob(JB):  0.243
      Kurtosis:   2.778          Cond. No.   150.
```



```
Random Forest: Accuracy=0.940
Random Forest: f1-score=0.922
```

- Initially, I used a regression model using the two features, valence and danceability
- As a result, the R-squared value was 0.172, which is very weak
  - This model was not fit for this situation
- Next, I used a random forest model and as a result, accuracy was 0.940 and f1-score was 0.922
- The features used were: bpm, energy, and danceability

Future research:

- Try to use data from other streaming platforms
- Test out more models using additional features to see if I could get a stronger result