Japan and US EDA

Heavy_Rotation
skaragou - berdogdu - jkennan - ccataldo

Data

We collected daily 'Top 50 Viral Songs' that appeared on Spotify's charts from 1/1/17 to 3/1/20 via https://www.spotifycharts.com. The data was modified to include the Spotify song id, country name, and date for each entry within a viral chart. We then obtained song features (valence, acousticness, danceability, energy, instrumentalness, speechiness, and tempo), via the Spotify API for each unique song ID. We stored this information within two tables in a .db file. In total, we gathered 3,411,816 entries across 66 countries containing 109,658 unique songs. A small number of countries (ex: Vietnam) may have occasional gaps, which we attribute to errors within https://www.spotifycharts.com at the time of data collection. A total of seven countries do not have chart data beginning in January 2017 (two in the latter half of 2017, four in 2018, and one in 2019). We do not have a definitive answer for each country, though speculate that it is because Spotify was not deployed in these countries until their respective chart start dates.

Initial Findings and Trends

We initially hypothesized that the "song features associated with virality in the US are the same with those found in other countries." We chose to measure song virality as the number of days a song remains on the charts, which allowed us to perform a survival analysis (ie. multiple linear regression) for each country. While we ran models for every country we had data for, we found that many, but not all, countries had similar song features. Hence, we were curious about four countries (US, Japan, Greece and India) which had varying significant song features. The model coefficients, standard error and p-values can be seen below.

Model for US Chart Data

For the US it seemed that 'danceability', 'instrumentalness' and to a lesser degree 'speechiness' and 'tempo' were significant song features for virality in the US based on the chart data.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
valence	-0.0086483	0.9913890	0.0681119	-0.127	0.8990	
acousticness	-0.1186549	0.8881142	0.0647711	-1.832	0.0670	.
danceability	-0.8566995	0.4245610	0.0950568	-9.013	< 2e-16	***
energy	0.1558673	1.1686711	0.0941788	1.655	0.0979	.
instrumentalness	0.5146540	1.6730595	0.0870037	5.915	3.31e-09	***
speechiness	-0.2292220	0.7951520	0.1126999	-2.034	0.0420	*
tempo	-0.0010857	0.9989149	0.0004476	-2.425	0.0153	*

Model for Japan Chart Data

For Japan it seemed that 'valence', 'instrumentalness' and to a lesser degree 'energy', 'tempo' and 'acousticness' were significant song features for virality in Japan:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
valence	-0.1876069	0.8289405	0.0622361	-3.014	0.00257	**
acousticness	-0.1468174	0.8634516	0.0621528	-2.362	0.01817	*
danceability	0.0878457	1.0918196	0.0904056	0.972	0.33121	
energy	-0.1665210	0.8466051	0.0831973	-2.002	0.04534	*
instrumentalness	0.3333071	1.3955759	0.0584401	5.703	1.17e-08	***
speechiness	0.1133598	1.1200349	0.1388857	0.816	0.41438	
tempo	-0.0009362	0.9990643	0.0004412	-2.122	0.03384	*

Model for India Chart Data

For India interestingly seemed that only ‘instrumentalness’ was a significant song features for virality in India:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
valence	0.0484693	1.0496631	0.1398486	0.347	0.729	
acousticness	-0.1277168	0.8801026	0.1179994	-1.082	0.279	
danceability	-0.1615468	0.8508267	0.2020869	-0.799	0.424	
energy	0.0486389	1.0498412	0.1794214	0.271	0.786	
instrumentalness	0.5961428	1.8151040	0.1376177	4.332	1.48e-05	***
speechiness	-0.0073114	0.9927153	0.2745449	-0.027	0.979	
tempo	0.0003766	1.0003767	0.0008949	0.421	0.674	

Model for Greece Chart Data

For Greece it seemed that ‘danceability’ and ‘instrumentalness’ were significant song features for virality in Greece:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
valence	9.382e-02	1.098e+00	7.742e-02	1.212	0.226	
acousticness	-1.055e-01	8.999e-01	7.421e-02	-1.421	0.155	
danceability	-8.366e-01	4.332e-01	1.075e-01	-7.784	7.01e-15	***
energy	-9.909e-04	9.990e-01	1.073e-01	-0.009	0.993	
instrumentalness	5.739e-01	1.775e+00	9.287e-02	6.179	6.43e-10	***
speechiness	-1.541e-01	8.571e-01	1.370e-01	-1.125	0.261	
tempo	3.538e-05	1.000e+00	5.175e-04	0.068	0.945	

While the survival analysis model identified certain interesting trends about viral song traits in the data set, it could not conclusively confirm or deny our hypothesis. By comparing p-values between the US and other countries, we were making the assumption that virality is only dependent on the provided song features and external factors such as advertising and celebrity news does not play a role on how a song becomes viral. Additionally, the p-values can also be affected by song outliers which could push song feature significance without an actual correlation present.

To avoid coming to any incorrect conclusions, we instead chose to explore other interesting correlations and trends within our data, rather than attempt to test a specific hypothesis. Specifically, we chose to look at correlations between two features, danceability and valence, and the amount of time songs remained on their respective charts in Japan and the US. We chose to compare the US and Japan because their most significant survival features were different (danceability in the US and valence in Japan), and because they have a large number of songs recorded in their charts and hence we would not have to sample for a smaller population size.

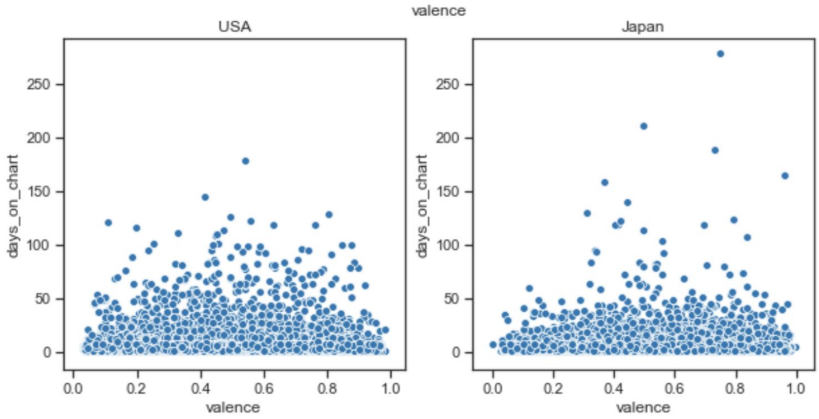
Exploratory Data Analysis

Finding #1: Songs seem to last longer on the US charts on average than on the Japanese Charts.

	days_on_chart	valence	acousticness	danceability	energy	instrumentalness	speechiness	tempo
mean (US)	9.427914	0.456383	0.231724	0.652117	0.623229	0.037300	0.121121	120.642003
mean 95% CI (US)	(9.101, 9.755)	(0.451, 0.462)	(0.225, 0.238)	(0.648, 0.656)	(0.619, 0.628)	(0.034, 0.041)	(0.118, 0.124)	(119.916, 121.368)
std (US)	13.000791	0.219050	0.255045	0.153134	0.182677	0.146241	0.119276	28.883970
mean (Japan)	7.814416	0.520991	0.206167	0.608906	0.709561	0.065295	0.086789	121.873689
mean 95% CI (Japan)	(7.575, 8.055)	(0.516, 0.526)	(0.2, 0.212)	(0.606, 0.612)	(0.705, 0.714)	(0.061, 0.07)	(0.085, 0.089)	(121.231, 122.516)
std (Japan)	10.492461	0.224410	0.257101	0.147929	0.203674	0.203300	0.088219	28.088414

Analysis for Finding #1: Above is a table that describes the mean, the 95% confidence interval for the mean and standard deviation of song features in the US and Japan. As seen, the mean of the amount of days a song spends on the charts is higher in the US (9.43) than in Japan (7.81). Given that there is no overlap in the confidence intervals of the country's respective means, the difference between the days on chart in the US and Japan are statistically significant. Since the boundaries of the confidence interval for days on the chart in the US are higher than the boundaries for Japan, on average, songs in the US stay longer on the charts than those in Japan.

Finding #2: There is a correlation between valence and time on charts in Japan and the US

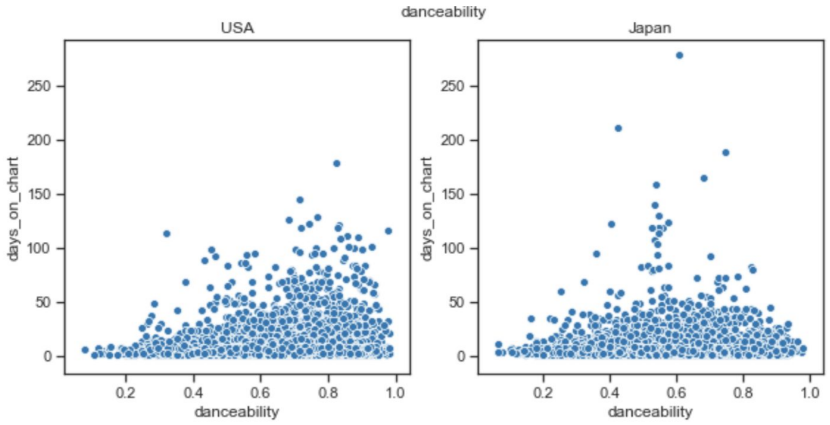


Country	ρ (Pearson Correlation Coef.)	P-value
US	0.0449	0.00045
Japan	0.0467	6.23410×10^{-5}

Analysis for Finding #2: The above scatter plot is that of the valence song feature and duration on charts in the two countries. From our survival analysis model we found that valence was a significant feature for predicting time on charts in Japan, while it was not for the US. However,

when we calculated the correlation coefficients for the US and Japan based on the scatterplots, we found that both countries have a positive correlation between valence and days on chart and the correlation is significant with a threshold of 0.05. While from the graph it seems that Japan might have a stronger correlation than the US, their correlation coefficients are not clearly different from one another and hence we can only come to the conclusion that a positive correlation exists for both countries. Additionally, we can make the observation that on average Japan's songs on charts have higher values of valence than those in the US. On our baseline table in the previous finding, the confidence intervals for valence do not overlap and both boundaries for Japan are higher than those of the US.

Finding #3: There is a stronger correlation between danceability and time on charts in US than in Japan



Country	ρ (Pearson Correlation Coef.)	P-value
US	0.1414	1.4309×10^{-28}
Japan	-0.0083	0.4721

Analysis for Finding #3: Similarly, this is a scatter plot of danceability vs. the number of days a song was on the chart in the two countries. The distribution of songs differs more in this variable than in valence. We chose to look into this correlation, as the model showed that danceability is a statistically significant variable in the US while it was not for Japan. We calculated the correlation coefficients and their corresponding two-tailed test for both the US and Japan. We can see that there is a positive correlation coefficient of 0.1414 for the US and since the p-value is below our threshold of 0.05, danceability in the US is likely correlated with days on chart. On the other hand, for Japan, we can see that there is not enough evidence to show that it is correlated.

We can also make the observation that on average the US's songs on charts have higher values of danceability on average than those in Japan. On our baseline table in the previous finding, the confidence intervals for danceability do not overlap and both boundaries for the US are higher than those of Japan.