SPOTIFY ANALYSIS

DATA 557 | WINTER 2022

GROUP MEMBERS

Aniket Fadia
Hannah Luebbering
Rebecca Klein
Teju Kandula

INTRODUCTION

DATA DESCRIPTION:

All of the data sourced for this project is from the **Spotify API**. There are two main sub categories of data from the Spotify API that we use: the Daily Top 200 Charts and Spotify Track Features.

1. Daily Top 200 Charts (SPOTIFY API)

The **Daily Top 200 Charts** shows the top 200 most streamed tracks each day from January 1, 2018 to December 31, 2021. For example, Spotify Daily Top Songs USA shows the daily update of the most played tracks across the US right now. The variables included in this dataset are `rank`, `uri`, `artist_names`, `track_name`, `source`, `peak_rank`, `previous_rank`, `days_on_chart`, `streams`.

#		TRACK	STREAMS
1	–	**We Don't Talk About Bruno** Carolina GaitÃ¡n - La Gaita, Mauro Castillo, Adassa, Rhenzy Feliz, Diane Guerr	998,966
2	–	**Heat Waves** Glass Animals	924,986
3	↓	**Super Gremlin**	787,867

5	↑	**Nail Tech** Jack Harlow	772,793
6	–	**STAY (with Justin Bieber)** The Kid LAROI, Justin Bieber	685,451

2. Spotify Track Features (SPOTIFY API)

The **Spotify Track Features** dataset shows audio features for each track streamed. A full list of these, along with their verbal definitions, can be found on Spotify's page for developers. There are 12 audio features for each track, including confidence measures `acousticness`, `instrumentalness`, `liveness`, `speechiness`; perceptual measures `danceability`, `energy`, `loudness`, `valence`; and descriptors `key`, `duration`, `mode`, `tempo`.

#	TRACK	DANCEAB...	ENERGY	KEY	LOUDN...	MODE	SPEECH...	ACOUST...
1	**abcdefu**	0.695	0.54	4	-5.692	1	0.0493	0.299
2	**We Don't Talk Abou**	0.577	0.45	0	-8.516	0	0.0834	0.357
3	**THATS WHAT I WAN**	0.737	0.846	1	-4.51	0	0.22	0.00614
4	**Heat Waves**	0.761	0.525	11	-6.9	1	0.0944	0.44

FINAL DATA

We merged the two data sources into a single data frame with the combination of being the unique identifier of an observation.

VARIABLE	TYPE	DESCRIPTION
date	Categorical, date	Date of the spotify chart
track_id	Categorical, str	Unique identifier for each track
track_name	Categorical, str	Title of the track

main_artist	Categorical, str	Name of the main artist
main_artist_id	Categorical, str	Unique identifier for each artist
rank	Quantitative, int	Rank from 1-200 (1 is the most streamed track that day)
streams	Quantitative, int	Total number of global streams that day
acousticness	Quantitative, float	Confidence measure of sound through acoustic (1.0 is the most acoustic)
danceability	Quantitative, float	Dance friendly measurement (1.0 is most danceable)
energy	Quantitative, float	Perceptual measure of intensity and activity
instrumentalness	Quantitative, float	Variety of instruments appeared
key	Categorical, int	Overall key of the track, sets of sharp or flat
liveness	Quantitative, float	Detection of whether a track was peformed live with an audience
loudness	Quantitative, float	Overall loudness of a track in decibels (dB)
mode	Categorical, int	Modality (major or minor) of a track, the type of scale
speechiness	Quantitative, float	Measures the number of spoken words
tempo	Categorical, int	Estimated tempo of a track in beats per minute (BPM)
valence	Quantitative, float	Measure from 0.0 to 1.0 describing the musical positiveness
duration	Quantitative, int	Duration of track in milliseconds
explicit	Categorical, boolean	True or false if contains explicit content
genre	Categorical, str	Name of the genre associated with that track

We do not have a genre for each song. So, we are utilizing the genres the artist belongs to and using those genres for each song. We want to discuss this with you in the meeting which we will be scheduling this week.

QUESTIONS/HYPOTHESES:

a Whether or not it is a weekday (Monday-Thursday) or weekend (Friday-Sunday)

b Whether or not it is in the holiday season (the day after Thanksgiving through December 31st)

c Meteorological season (summer, winter, spring, fall)

2 Did the popularity of happy songs (mean valence) in the top 200 Spotify US daily streams change during Covid?

3 What parameters are the most important in predicting the popularity on Spotify in the US?

METHODOLOGY:

1. Does time period change the popularity of genres in the Spotify US daily charts? (QUESTION 1)

Data Used:

We use the Spotify Daily Top Tracks as described in Data Source above. Specifically, we use the `track_id`, `Date`, and `genre` column variables from the Spotify Daily Top Tracks data. For sub-question (b), the date of Thanksgiving for each year is found using a holiday dataset from kaggle (https://www.kaggle.com/donnetew/us-holiday-dates-2004-2021/version/1).

Calculated Variables

Part a The calculated variable is the `day of the week` based on the Date column. Here, we examine weekday versus weekend, such that weekday is considered to be Monday-Thursday and weekend is considered to be Friday-Sunday.

Part b The calculated variable `holiday` is TRUE or FALSE depending on whether or not the Date column is within the holiday season. The holiday season includes the day after Thanksgiving through December 31st.

Part c The calculated variable is the meteorological `season` based on the Date column. For the seasons, we consider *Spring* to be March, April, and May; *Summer* to be June, July, and August; *Fall* to be September, October, and November; and *Winter* to be December, January, and February.

Statistical Method

Parts (a) and (b): large sample Z-test

For each sub-question, the two samples (weekday and weekend, holiday and not_holiday) are independent with independent observations, and the sample sizes are both large enough to use the large-sample Z-test.

We test the null hypothesis of no difference between proportions, i.e., we define the null hypotheses as $H_0: p_{\text{weekday}} = p_{\text{weekend}}$ and $H_0: p_{\text{holiday}} = p_{\text{not holiday}}$ versus the 2-sided alternative hypotheses $H_1: p_{\text{weekday}} \neq p_{\text{weekend}}$ and $H_1 : p_{\text{holiday}} \neq p_{\text{not holiday}}$, where p is the proportion of songs in the list of the top 200 daily streams on spotify that are of the genre being tested. This is calculated by taking the number of rows in the sample that contain the genre being tested divided by the total number of rows in the sample. A genre is to be considered if it includes that word in the genre (for example: 'pop' would include 'synthpop').

Part (c): Chi-square

The four samples (spring, summer, fall, and winter) are independent with independent observations, and the sample sizes are all large enough to can assume normality using the Central Limit Theorem.

We test the null hypothesis of no difference between proportions, i.e., we define the null hypotheses as $H_0: p_{\text{spring}} = p_{\text{summer}} = p_{\text{fall}} = p_{\text{winter}}$ versus the 2-sided alternative hypotheses that not all of the proportions are different (i.e., at least one season is not equal $H_1: p_{\text{season1}} \neq p_{\text{season2}}$)

The variable of interest is the proportion of songs in the list of the top 200 daily streams on spotify that are of the genre being tested. (See full description in sub-questions (a) and (b)).

How results will be reported and interpreted: p-value

Parts (a) and (b):

For each genre, we will calculate the proportion $\left(\hat{p}_{A:\text{ weekday/holiday}}, \hat{p}_{B:\text{ weekend/not holiday}}\right)$, length $\left(n_{A:\text{ weekday/holiday}}, n_{B:\text{ weekend/not holiday}}\right)$, and pooled sample proportion $\left(\hat{p} = \frac{n_A \hat{p}_A + n_B \hat{p}_B}{n_A + n_B}\right)$.

Using the above values, we can then calculate the test statistic:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}\,(1-\hat{p})\,(1/n_A + 1/n_B)}}$$

.

proportions
    - This would mean that the proportion of that genre for that subset is not equal (e.g., the proportion of pop songs in the spotify top 200 changes depending on whether or not it is the weekend)
- If the p-value is greater than α (i.e., $p > 0.01$), then we do not reject the null hypothesis of equal means
    - This would mean that the proportion of that genre for that subset is equal (e.g., the proportion of pop songs in the spotify top 200 does not change based on whether or not it is the weekend)

Part (c):

For each genre, we will create a dataframe with the count of songs in the top 200 of that genre and season, and a count of the total number of songs in the season. We will use that to run the chi-square test using `chisq.test(season_genre_df, correct = F)`, which will return a p-value.

- If the p-value is less than α (i.e., $p < 0.01$), then we reject the null hypothesis of equal proportions
    - This would mean that the proportion of that genre for that subset is not equal (e.g., the proportion of pop songs in the spotify top 200 changes depending on the season)
- If the p-value is greater than α (i.e., $p > 0.01$), then we do not reject the null hypothesis of equal means
    - This would mean that the proportion of that genre for that subset is equal (e.g., the proportion of pop songs in the spotify top 200 does not change based on the season)

---

2. Did the popularity of happy songs in the top 200 Spotify charts change during Covid? (QUESTION 2)

Data and Calculated Variables:

We use the Spotify Daily Top Tracks as described in the Data Description above. Particularly, question 2 makes use of the `track_name`, `valence`, and `Date` columns. Based on these columns, we created a `covid` variable to define whether a track entry was added to the top 200 playlist before Covid (`Date` < "03/13/2020") or after Covid (`Date` >= "03/13/2020"). Tracks added before Covid (03/13/2020) are labeled *before* whereas tracks added after Covid are labeled *after*.

Statistical Method:

two-sample large sample Z-test to compare the mean valences before and after Covid. This means that we compare the Z test statistic to the standard normal distribution.

We test the null hypothesis of no difference between valence means, i.e., we define the null hypothesis as $H_0 : \mu_{\text{before}} = \mu_{\text{after}}$ versus the 2-sided alternative hypothesis $H_1 : \mu_{\text{before}} \neq \mu_{\text{after}}$, where μ_{before} and μ_{after} are the mean valences per song for songs added before and after Covid, respectively. The test is conducted at a significance level of $\alpha = 0.05$.

How results will be reported and interpreted:

First, we calculate the mean (μ_{before}, μ_{after}), standard deviation (s_{before}, s_{after}), and size (n_{before}, n_{after}) for each of the two samples.

	MEAN	STD DEV	SIZE
before	0.4572265	0.2014658	1117863
after	0.4826522	0.2272466	592944

Using these values, we can then calculate the test statistic:

$$Z = \frac{\left|\bar{X}_{\text{before}} - \bar{X}_{\text{after}}\right|}{s^2_{\text{before}} / n_{\text{before}} + s^2_{\text{after}} / n_{\text{after}}} = \frac{|0.4572 - 0.4826|}{\sqrt{0.2015^2/1117863 + 0.2273^2/592944}} = 72.379$$

Next, we calculate the p-value using the standard normal distribution (since n_{before} and n_{after} are both large). The p-value is a probability about the test statistic, calculated under the assumption that the null hypothesis is true.

- If the p-value is less than α (i.e., $p < 0.05$), then we reject the null hypothesis of equal means.
    - This would mean that the mean valences per song before and after Covid are not equal (i.e., the popularity of happy songs changed during Covid).
- If the p-value is greater than α (i.e., $p > 0.05$), then we do not reject the null hypothesis of equal means.
    - This would mean that the mean valences per song before and after Covid are equal (i.e., the popularity of happy songs did not change during Covid).

The p-value for the test is $p < 0.001$. Based on the test, we reject the null hypothesis of equal valence means at the 0.05 level of significance.

We can also calculate the confidence interval for the difference between population means. A confidence interval provides additional information beyond the hypothesis test. In general, we can interpret a confidence interval as the set of all values of the population parameter that

The confidence interval for the difference between population means is $(0.0247, 0.0261)$, which is very similar to the result from the large-sample procedure. Hence, since the interval does not contain the value 0, we would reject the null hypothesis $H_0 : \mu_{\text{before}} - \mu_{\text{after}} = 0$ and conclude that the mean valences per song before and after Covid are different.

---

3. What parameters are the most important in predicting the popularity on Spotify in the US? (QUESTION 3)

Data Used:

Again, we use the Spotify Daily Top Tracks (described in Data Source). For question 3, we aggregate by song, so each song has its own row. Each song has the same values for each attribute so we take the mean of these attributes.

- The attributes include: `explicit`, `acousticness`, `danceability`, `duration`, `energy`, `instrumentalness`, `key`, `liveness`, `loudness`, `mode`, `speechiness`, `tempo` and `valence`.

We sum `streams` to find the total number of streams for each song. This does not mean the total number of streams since the song has come out. This value is only the total number of streams when the song is in the top 200 of spotify. After aggregation, the new dataframe has 7756 songs.

Calculated Variables: popularity

We create an `updated_rank` by doing 201 - `Rank` (i.e. rank 1 has score 200, rank 200 has score 1) so a higher updated_rank means the song is performing better in the ranks. We sum this `updated_rank` for each song to get a **popularity score**. This way the longer the song is in the Spotify top 200 the higher the popularity score.

Statistical Method: t-test

We will create a linear regression model (see next section) to calculate the estimated coefficients, β, for each predictor variable. We will use a **t-test** to determine which of these variables are significant:

$$H_0\colon\ \beta_X = 0 \qquad \text{versus} \qquad H_A\colon\ \beta_X \neq 0,$$

where β_X is each coefficient in the full linear regression model. We test at the significance level of $\alpha = 0.05$.

Statistical Model: Linear Regression Model

Dependent Variable: `Popularity` (see New Variables to be Calculated)

- `Acousticness` : A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- `Danceability` : Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- `Duration` : The duration of the track in milliseconds.
- `Energy` : Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- `Instrumentalness` : Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- `Key` : The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- `Liveness` : Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- `Loudness` : Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- `Mode` : Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- `Speechiness` : Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- `Tempo` : The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- `Valence` : The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

The sample size is big so we can assume normality using the Central Limit Theorem. We also know the data is independent.

How results will be reported and interpreted: p-value

models and see if they increased/decreased.