# Blank (Vector) Space



## Instructions

As a reminder, Taylor Swift is contemplating re-releasing three of her albums (Midnights, folklore, and 1989). She has also covered three different songs:

* Boyz II Men’s “End Of The Road”,
* Guns ‘N’ Roses’ “Sweet Child O’ Mine”, and
* Prince’s “Purple Rain”

She wants to include each cover song on the re-released album that the song is most similar to based on nine musical properties.

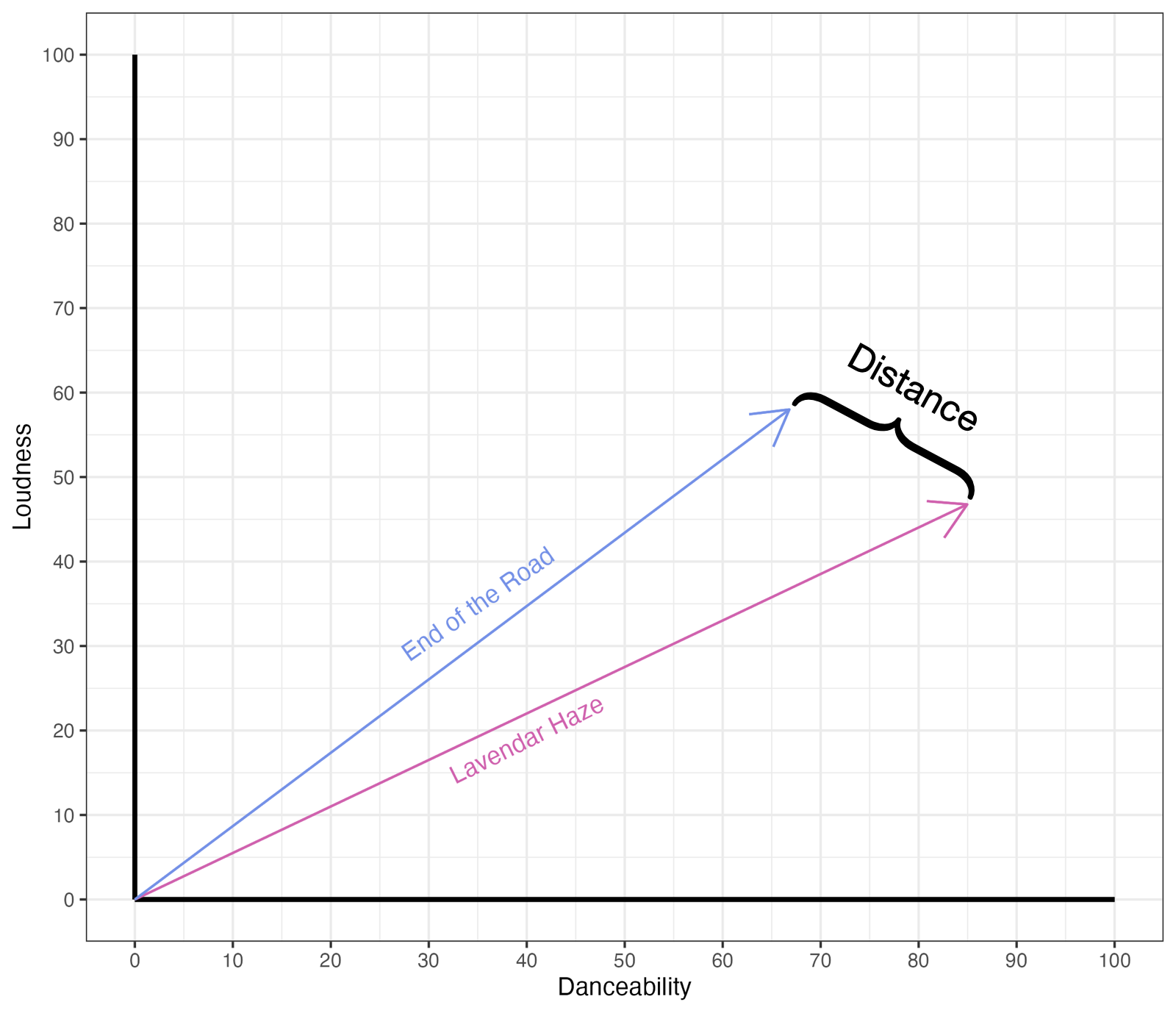
| Spotify collects data on the following nine musical properties:   * **Acousticness:** A confidence measure of whether the track is acoustic. Higher values indicate higher confidence that the track is acoustic. * **Danceability:** Measure of how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. Higher values indicate the song is more danceable. * **Energy:** Perceptual measure of a song’s intensity and activity, with higher values indicating more energy. Typically, energetic tracks feel fast, loud, and noisy. 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”. Higher instrumentalness values indicate a greater likelihood the track contains no vocal content. * **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. * **Loudness:** The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. * **Speechiness:** Measures the presence of spoken words in a track, with higher values indicating more words. (Because Spotify also includes audiobooks and podcasts as, most songs have low speechiness values.) * **Tempo:** The overall estimated speed/pace of a track. * **Valence:** Measure of whether a song is likely to make someone feel happy (higher valence) or sad (lower valence). |
| --- |

## Part I: The Mathematics of Euclidean Distance

In the previous set of activities, we computed the similarity between songs using Euclidean Distance. Mathematically, we define the distance between two observations *a* and *b* as:

where, *p* is the number of measured attributes for the observations. In higher levels of mathematics, ***ai*** and ***bi*** are vectors in *p*-space. To help you think about this, we will simplify this to two-space, something you graph in a lot in high school mathematics. Being in two-space means that there are two attributes measured for each song (e.g., danceability and loudness). You can visualize a vector as an arrow that has the butt-end of the arrow at the origin (0, 0), and the arrow-end at the (danceability, loudness)-value for the song in question.

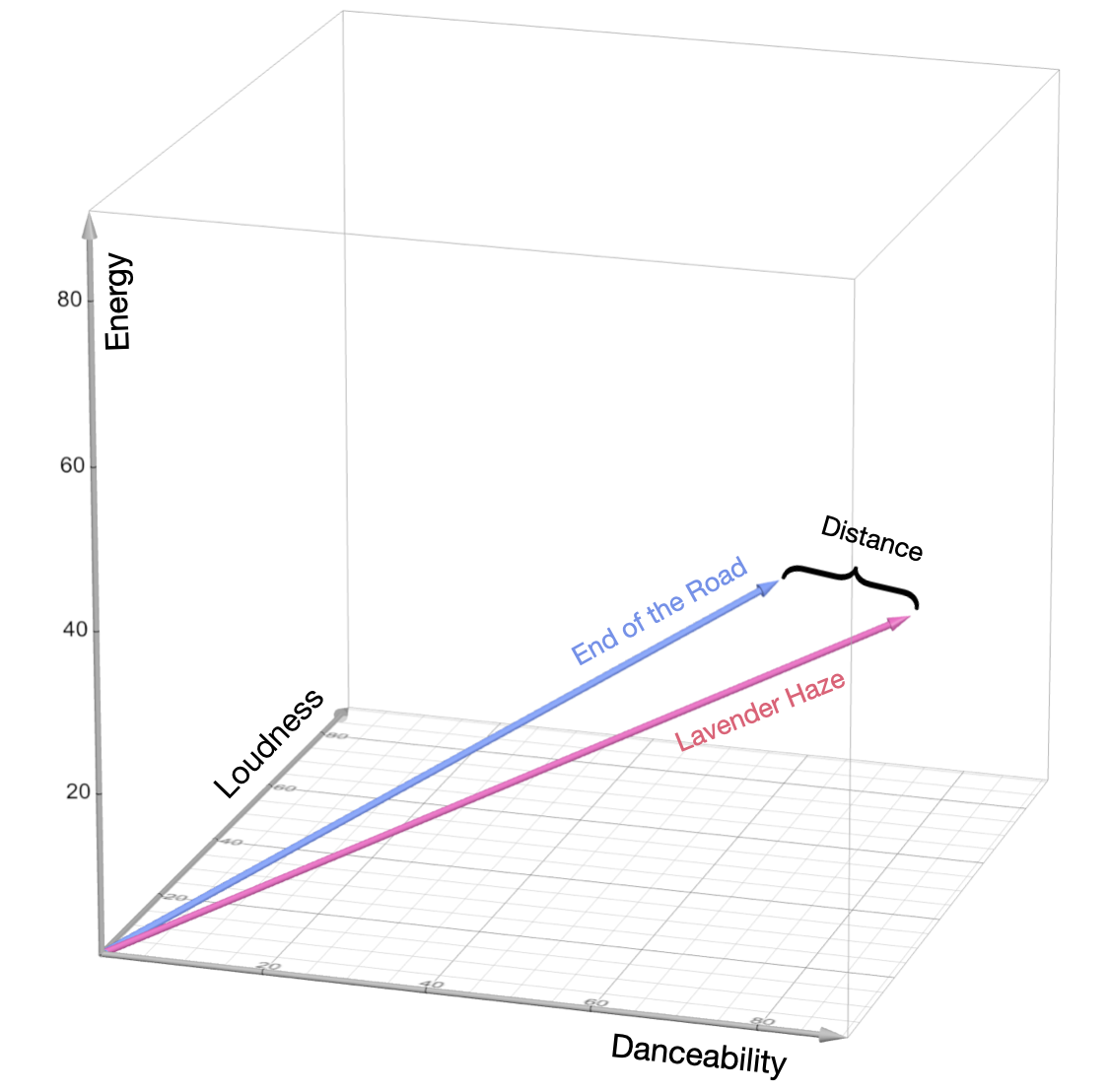
For example, consider the songs *Lavendar Haze* (danceability = 84.96, loudness = 46.76) and *End of the Road* (danceability = 66.80, loudness = 58.00). Each song would constitute a different vector (arrow) and the Euclidean Distance is visualized as the distance between the arrow tips.



Songs that were more similar to each other had smaller Euclidean Distance measures, while songs that were less similar to each other had larger Euclidean Distance measures.[[1]](#footnote-0)

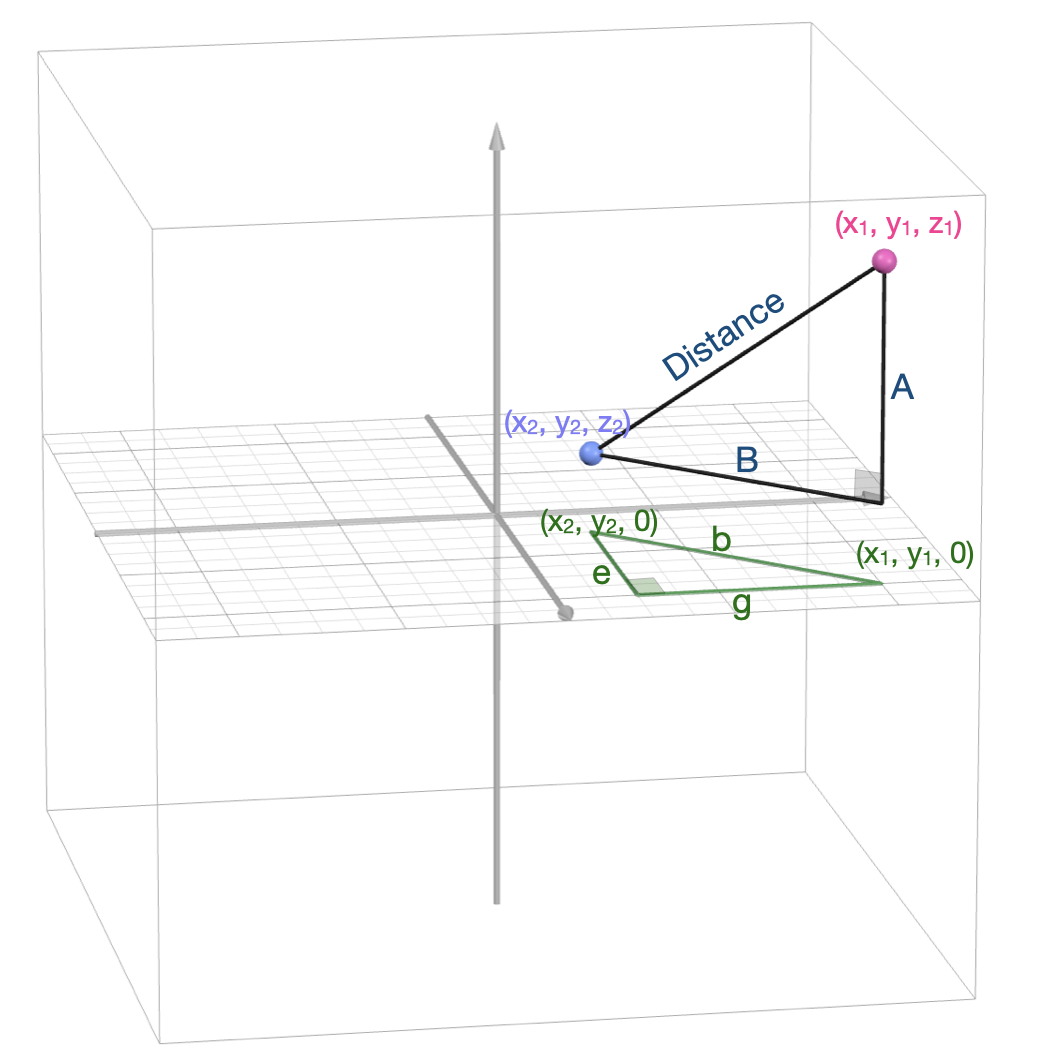
## Part II: Euclidean Distance in More Than Two Dimensions

We can also extend the idea of Euclidean Distance to measure similarity when we have more than two attributes. (For example, imagine we had measured three metrics for our songs: loudness, danceability, and energy.) Again, we are measuring the distance between two vectors, however this time they are plotted in 3-space; there would be a loudness axis, a danceability axis, and an energy axis.



Essentially we need to find the distance between two points in 3-space. Let’s say that these two points have the coordinates (x1, y1, z1) and (x2, y2, z2), respectively. The following plot shows two such points. To determine the distance between the two points, we can form a right triangle such that the hypotenuse of that triangle extends between the two points and one of the legs is parallel to the vertical axis (see black triangle in figure below).

We need to find the length of both legs of this right triangle, namely A and B, so that we can use the Pythagorean Theorem to find the distance between the points (x1, y1, z1) and (x2, y2, z2). Because we created this triangle so that A was parallel to the vertical axis, the length of A is simply (z1 – z2). To find the length of B, we project that leg onto the plane created by the non-vertical axes. (This projection creates the line segment b, which is now in two-space.) This segment is the hypotenuse of another right triangle (green triangle in figure below) that connects the points (x1, y1) and (x2, y2).



1. Find the length of b in the green triangle.
2. Since b is the projection of B, both b and B have the same length. Now that you know the length of B and A, use those values to determine the hypotenuse of the black triangle, which is the distance between (x1, y1, z1) and (x2, y2, z2).

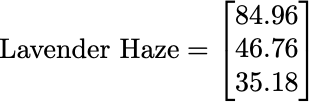
Euclidean Distance between two points in 3-space is just an extension of the Euclidean Distance between two points in 2-space! We are computing the squared differences between the corresponding coordinates, summing them, and finding the square root of that sum. In fact, if we move on to 4-, 5-, or even n-space, the Euclidean Distance is found similarly, we just have additional coordinates.

As an example, to find the Euclidean Distance between *Lavendar Haze* (danceability = 84.96, loudness = 46.76, energy = 35.18) and *End of the Road* (danceability = 66.80, loudness = 58.00, energy = 34.05), we compute:

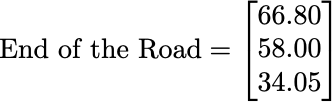
### **Math Extension: Matrix Algebra (Optional)**

Matrix algebra (also known as linear algebra) is the mathematics that focuses on matrices and vectors. Because we are essentially trying to find the distance between two vectors, matrix algebra offers a convenient notation and set of operations for expressing and working with these vectors.

The vectors we plotted earlier can also be expressed using the notation of matrix algebra. For example, *Lavendar Haze* (danceability = 84.96, loudness = 46.76, energy = 35.18) can be expressed as:



Each of the attribute values is indicated as a different element in the vector. Because there are three elements in the vector, we would refer to this as a 3-dimensional vector. (It would need to be plotted in three dimensions.) Similarly *End of the Road* (danceability = 66.80, loudness = 58.00, energy = 34.05) can be expressed as:

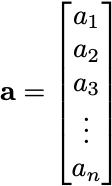


The Euclidean distance between two vectors can also be expressed as an operation in matrix algebra, namely the distance is equal to the norm of the difference between the two vectors. Notationally this looks like the following:



A norm is just a matrix algebra term that means length of a vector.

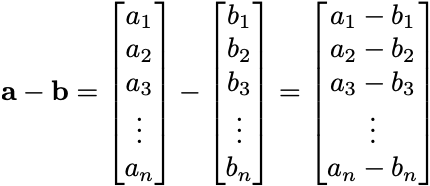
Consider an n-dimensional vector **a**. That is,



The norm of **a** (it’s length) is:



In our distance formula we are computing the norm of the difference between vector **a** and vector **b**. When we compute a difference between two vectors, we subtract the corresponding elements of the two vectors. For example,



Then to find the norm, we would square each of the elements in the resulting difference vector, add them together, and take the square root of that sum:



This is the same formula that we expressed earlier for Euclidean distance. The main advantage of using matrix algebra is that the notation allows us to express things very conveniently.

1. The nine Spotify metrics for *Lavender Haze* and *End of the Road* are given below. Find the Euclidean Distance between these two songs.

| **Property** | **Lavendar Haze** | **End Of The Road** |
| --- | --- | --- |
| Acousticness | 26.62 | 6.79 |
| Danceability | 84.96 | 65.59 |
| Energy | 35.17 | 34.05 |
| Instrumentalness | 0.12 | 0.00 |
| Liveness | 15.70 | 0.00 |
| Loudness | 46.76 | 58.00 |
| Speechiness | 15.28 | 0.00 |
| Tempo | 21.16 | 73.50 |
| Valence | 6.65 | 54.28 |

1. Open the ***taylor-swift-spotify-metrics*** Google Sheet. In the Activity 5 sheet, write a formula to compute the Euclidean distance between *End of the Road* and each of the tracks on 1989, folklore, and Midnights based on all nine Spotify metrics.
   1. Check that your formula is computing correctly by double-checking the Euclidean distance between *Lavender Haze* and *End of the Road*.
2. Based on your distance measures, which song is most similar to *End of the Road* on the nine metrics? Explain.
3. Use the optimal number of nearest neighbors to determine the album that *End of the Road* should be released on. (You computed the optimal number of nearest neighbors in the previous activity.) Also report the “vote” tally for each class (album).

## Part III: Classify the Other Two Songs

The Spotify metrics for the other two cover songs Taylor Swift wants to add to her re-released albums are given below.

| **Property** | **Purple Rain** | **Sweet Child O’ Mine** |
| --- | --- | --- |
| Acousticness | 3.56 | 9.25 |
| Danceability | 10.32 | 26.32 |
| Energy | 37.19 | 100.00 |
| Instrumentalness | 0.47 | 17.56 |
| Liveness | 100.00 | 8.57 |
| Loudness | 47.39 | 95.18 |
| Speechiness | 1.80 | 7.71 |
| Tempo | 37.48 | 49.51 |
| Valence | 16.74 | 64.77 |

1. Use the same process to determine the album that *Sweet Child O’ Mine* should be released on. Also report the “vote” tally for each class (album).
2. Use the same process to determine the album that *Purple Rain* should be released on. Also report the “vote” tally for each class (album).

1. Because larger distances mean that the songs are less similar, sometimes data scientists refer to Euclidean Distance as a *dissimilarity* measure. [↑](#footnote-ref-0)