**CS703 6.0 Data Deployment**

*Madeline Lin*

**Task 6.1: Planning Deployment**

**Deliverable: Deployment Plan**

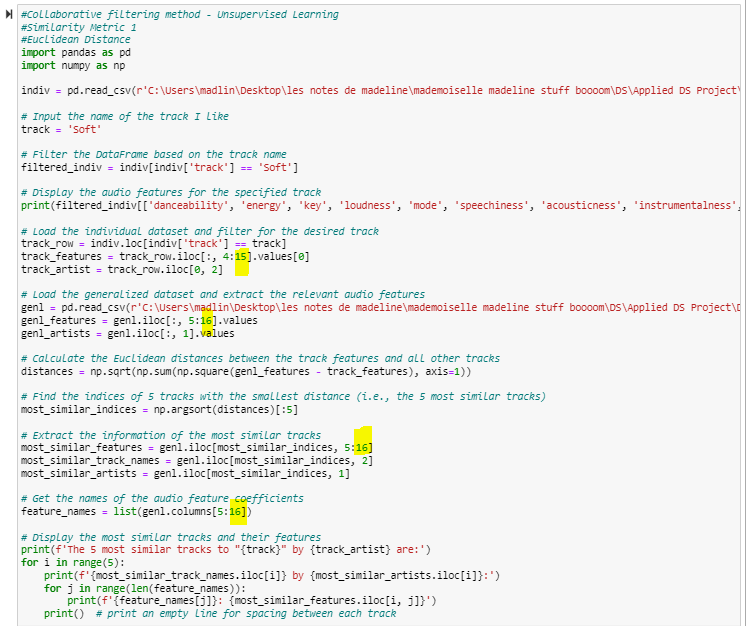
Before coming up with a strategy for putting the models to work in my business, my first step is to conduct a thorough inspection of my models and generalized dataset to eliminate any possible errors or mistakes that may have been made during the modeling/evaluation processes.

Below is a recap of my outcome.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Test Dataset | Target Song | Outcome | Accuracy |
| *Content-based filtering method* | | | | |
| *Original* | | | | |
| Simple Linear Regression Model | Global Weekly Top 100 Songs Dataset | N/A | Six out of 20 songs were added to my library. | 30% |
| *Revised* |  |  |  |  |
| Simple Linear Regression Model | Adjusted Generalized Dataset  (Songs from year 2011 to 2020) | N/A | One out of 20 songs was added to my library. | 5% |
| *Collaborative filtering method* | | | | |
| *Original* | | | | |
| Euclidean Distance | Generalized Dataset | One Song | None of five songs were added to my library. | 0% |
| Cosine Similarity | Generalized Dataset | One Song | One of five songs was selected. This song was already in my library. | 20% |
| *Revised* | | | | |
| Euclidean Distance | Generalized Dataset | Five Songs | One of five songs were added to my library. | 20% |
| Cosine Similarity | Generalized Dataset | Five Songs | One of five songs was selected. This song was already in my library. | 20% |

* Content-based filtering method
* Based on my review of the codes, I did not find any issues.
* Collaborative filtering method
  + Based on my review of the codes, in terms of the collaborative filtering method, I found there was an overlook in terms of the audio features across from individual dataset and generalized dataset. I accidently included the duration column when I retrieved the audio features. I revised my codes accordingly. There should be 11 columns of audio features included instead of 12 columns.
* One Target Song
* Euclidean Distance

*Old Code*

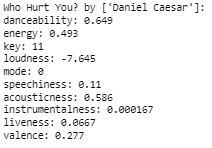
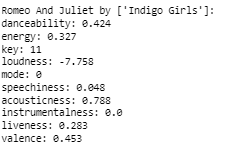
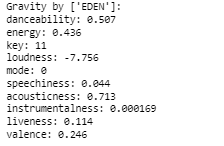


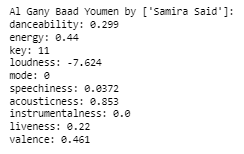
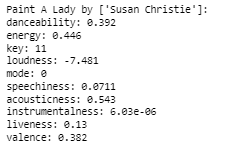
*Revised Code*



*Revised Output*

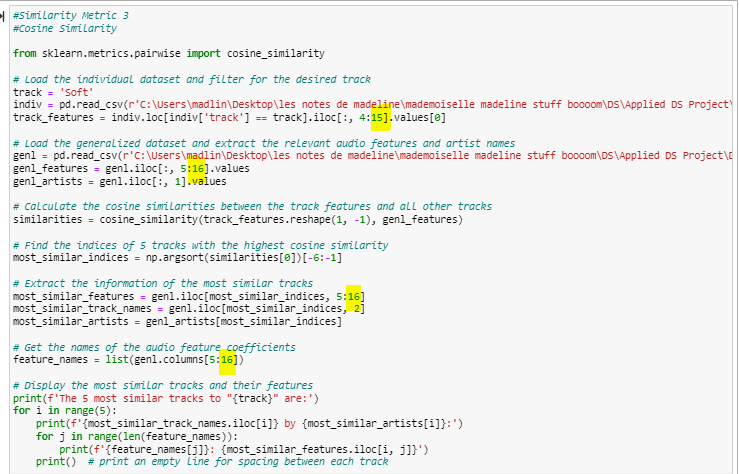






* Cosine Similarity

*Old Code*

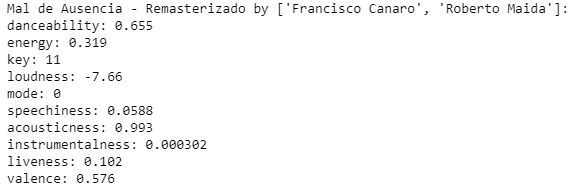
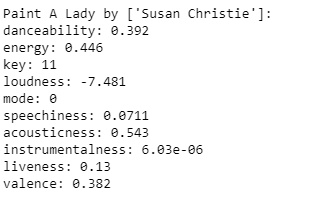


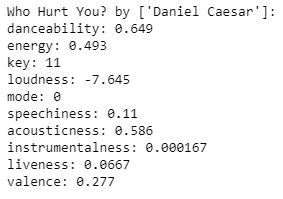
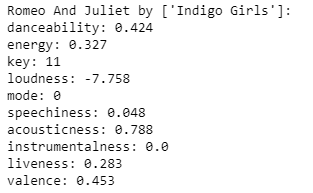
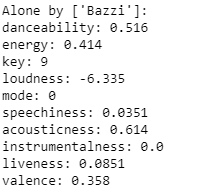
*Revised Code*



*Revised Output*

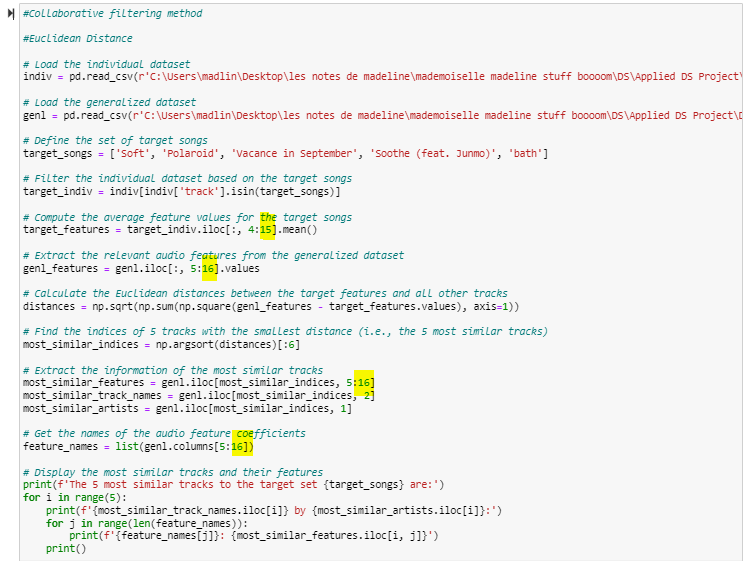


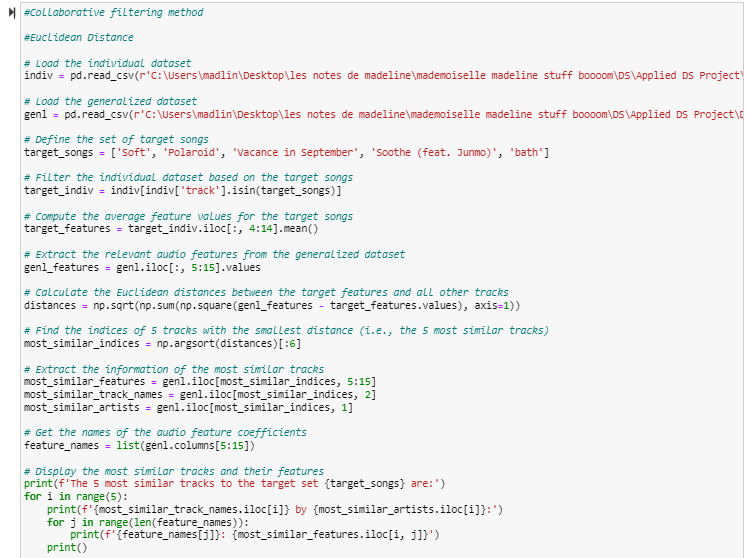
  

* Five Target Songs
  + Euclidean Distance

*Old Code*

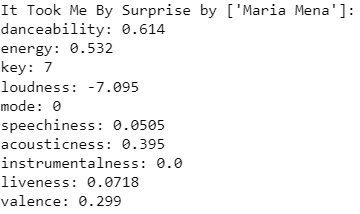
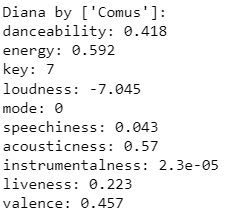
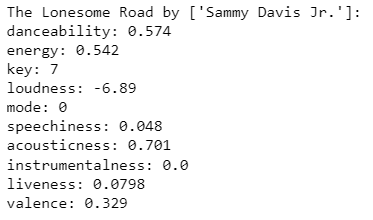


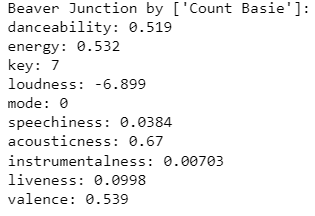
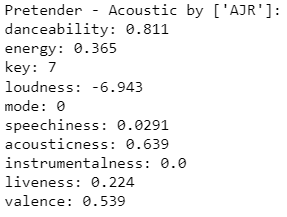
*Revised Code*



*Revised Output*



*Old Code*

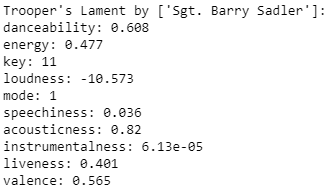
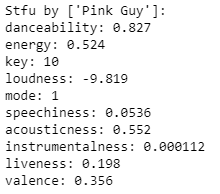
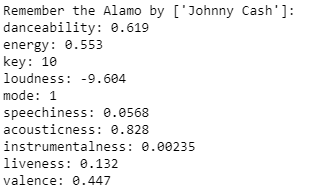


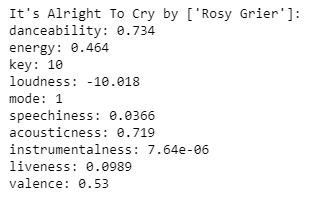
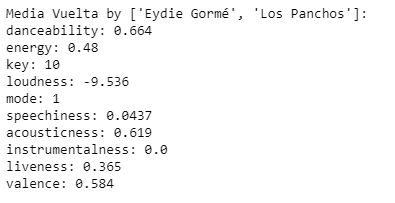
*Revised Code*



*Revised Output*



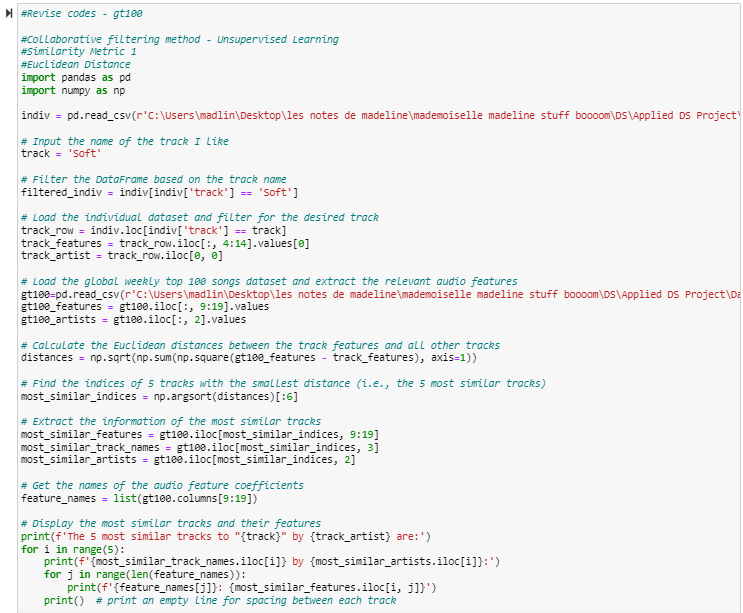


* Summary of the outcomes based on the revised codes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| *One Target Song* | | | | |
| *Euclidean Distance* | | | | |
| 1 | Gravity | EDEN | 3 | No |
| 2 | Romeo And Juliet | Indigo Girls | 3 | No |
| 3 | Who Hurt You? | Daniel Caesar | 5 | No |
| 4 | Paint A Lady | Susan Christie | 1 | No |
| 5 | Al Gany Baad Youmen | Samira Said | 2 | No |
| *Cosine Similarity* | | | | |
| 1 | Mal de Ausencia – Remasterizado | Francisco Canaro Roberto Maida | 1 | No |
| 2 | Paint A Lady | Susan Christie | 1 | No |
| 3 | Who Hurt You? | Daniel Caesar | 5 | No |
| 4 | Romeo And Juliet | Indigo Girls | 3 | No |
| 5 | Alone | Bazzi | 4 | No |
| *Five Target Songs* | | | | |
| *Euclidean Distance* | | | | |
| 1 | It Took Me By Surprise | Maria Mena | 6 | No |
| 2 | The Lonesome Road | Sammy Davis Jr. | 1 | No |
| 3 | Diana | Comus | 1 | No |
| 4 | Beaver Junction | Count Basie | 4 | No |
| 5 | Pretender – Acoustic | AJR | 10 | Yes (already in my library) |
| *Cosine Similarity* | | | | |
| 1 | Trooper’s Lament | Barry Sadler | 2 | No |
| 2 | Stfu | Pink Guy | 6 | No |
| 3 | Remember the Alamo | Johnny Cash | 3 | No |
| 4 | Media Vuelta | Eydie Gorme Los Panchos | 3 | No |
| 5 | It’s Alright To Cry | Rosy Grier | 4 | No |

* According to the table shown above, it appears that the revisions made to the code did not result in any improvement. When using a single target song, none of the recommended songs were added to my library. When using five target songs, only one of the recommended songs, *Pretender – Acoustic*, which I gave a 10 rating, was already in my library based on Euclidean Distance metric. It is interesting to note that the five target songs I picked are all K-Indie songs, which typically have a R&B or Soul vibe. However, the *Pretender – Acoustic* is actually an EDM track. I am unsure how the algorithm was able to suggest such a seemingly ironic result. One good thing I observed that, two songs *–* *Who Hurt You?* and *Paint A Lady*, were both selected according to two different similarity metrics, which somewhat proved that the revised codes were accurate.
* I decided to have one last try by using the Global Weekly Top 100 Songs dataset as my selection pool.
* One Target Song
* Euclidean Distance

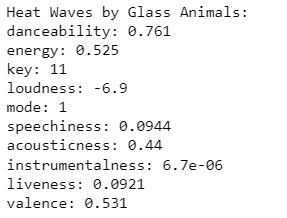
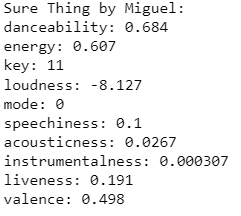
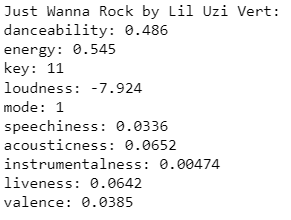
*Revised Code – gt100*

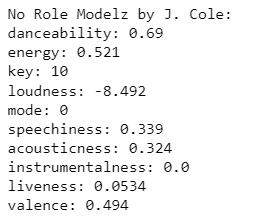
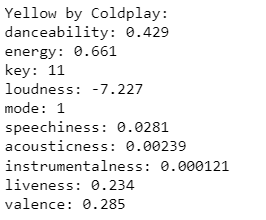


­

*Revised Output – gt100*





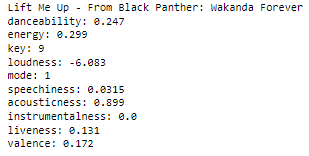
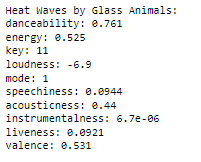
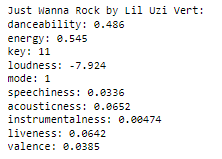
* Cosine Similarity

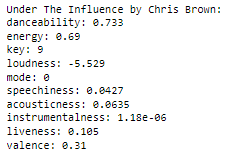
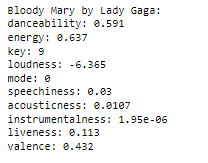
*Revised Code – gt100*



*Revised Output – gt100*



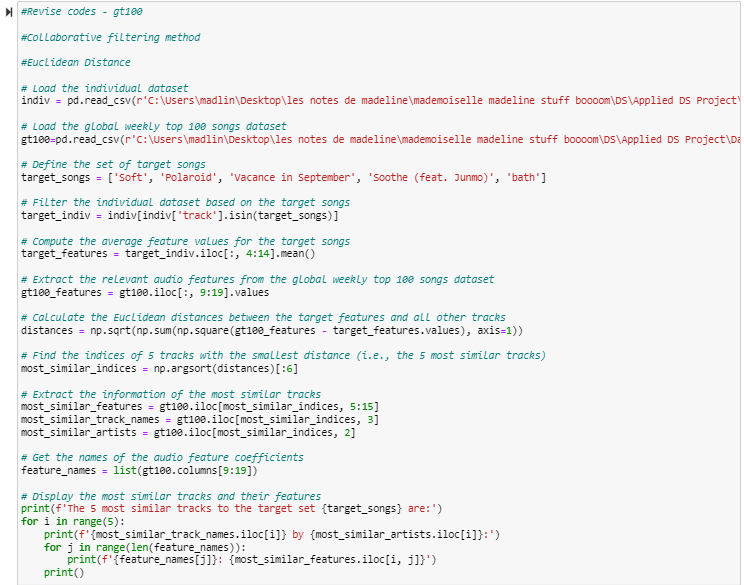
 

* Five Target Songs
* Euclidean Distance

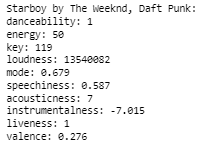
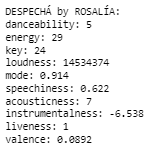
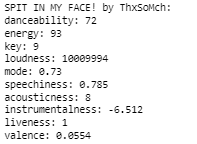
*Revised Code – gt100*

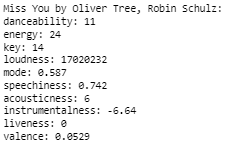
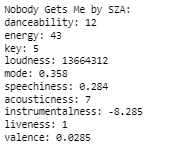
­



*Revised Output – gt100*



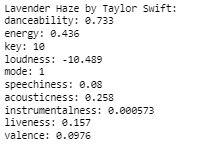
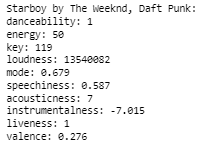
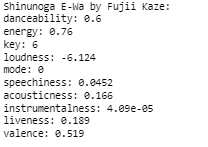
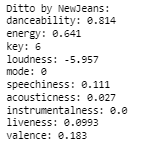
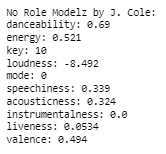
* Cosine Similarity

*Revised Code – gt100*



*Revised Output – gt100*



* Summary of the outcomes tested on the Global Weekly Top 100 Songs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Track | Artist | Scale | Added to my library |
| *One Target Song* | | | | |
| *Euclidean Distance* | | | | |
| 1 | Sure Thing | Miguel | 7 | Yes |
| 2 | Heat Waves | Glass Animals | 6 | No |
| 3 | Just Wanna Rock | Lil Uzi Vert | 4 | No |
| 4 | Yellow | Coldplay | 3 | No |
| 5 | No Role Modelz | J. Cole | 6 | No |
| *Cosine Similarity* | | | | |
| 1 | Lift Me Up | Rihanna | 5 | No |
| 2 | Just Wanna Rock | Lil Uzi Vert | 4 | No |
| 3 | Heat Waves | Glass Animals | 6 | No |
| 4 | Under The Influence | Chris Brown | 5 | No |
| 5 | Bloody Mary | Lady Gaga | 4 | No |
| *Five Target Songs* | | | | |
| *Euclidean Distance* | | | | |
| 1 | Starboy | The Weeknd Daft Punk | 9 | Yes (already in my library) |
| 2 | DESPECHA | ROSALIA | 6 | No (Latin Pop) |
| 3 | SPIT IN MY FACE! | ThxSoMch | 3 | No |
| 4 | Miss You | Oliver Tree Robin Schulz | 8 | Yes (already in my library) |
| 5 | Nobody Gets Me | SZA | 3 | No |
| *Cosine Similarity* | | | | |
| 1 | Lavender Haze | Taylor Swift | 5 | No |
| 2 | Shinunoga E-Wa | Fujii Kaze | 4 | No (J-pop) |
| 3 | Starboy | The Weekend Daft Punk | 9 | Yes (already in my library) |
| 4 | Ditto | NewJeans | 8 | Yes (K-pop) |
| 5 | No Role Modelz | J.Cole | 6 | No |

* According to the table shown above, when using a single target song, one song was added to my library based on the Euclidean Distance metric while no song was added to my library based on the Cosine Similarity metric. When using five target songs, two out of five songs were selected and they were already in my library based on the Euclidean Distance metric; two out of five songs were selected and one song was already in my library. What is more, the Euclidean Distance and Cosine Similarity metrics enabled me to discover Latin Pop, K-pop, and J-pop songs, in addition to American pop songs with English lyrics. This brought me joy as it has added diversity to the recommended songs. Additionally, it is pretty satisfactory that Euclidean Distance metric recommends me a song by ROSALIA, who is one of my favorite artists.
* Below is a summarized table outlining the improvements after I changed the test dataset to Global Weekly Top 100 Songs Dataset. The accuracy of two similarity metrics improved from 20% to 40% if I used five target songs. The accuracy of Euclidean distance’s accuracy increased from 0% to 20% while the Cosine similarity drops from 20% to 0% if I used one target song.

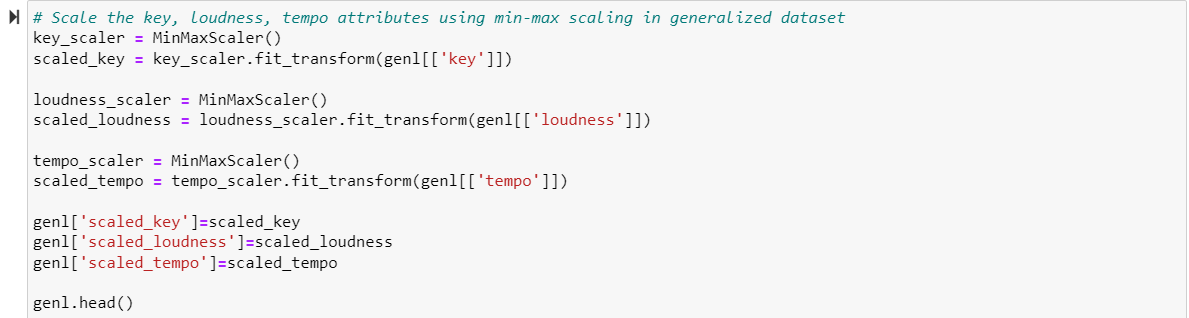
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Test Dataset | Target Song | Outcome | Accuracy |
| *Content-based filtering method* | | | | |
| *Original* |  |  |  |  |
| Simple Linear Regression Model | Global Weekly Top 100 Songs Dataset | N/A | Six out of 20 songs were added to my library. | 30% |
| *Revised* |  |  |  |  |
| Simple Linear Regression Model | Adjusted Generalized Dataset  (Songs from year 2011 to 2020) | N/A | One out of 20 songs was added to my library. | 5% |
| *Collaborative filtering method* | | | | |
| *Original* |  |  |  |  |
| Euclidean Distance | Generalized Dataset  Global Weekly Top 100 Songs Dataset | One Song | None of five songs were added to my library.  One of five songs was added to my library. | 0%  20% |
| Cosine Similarity | Generalized Dataset  Global Weekly Top 100 Songs Dataset | One Song | One of five songs was selected. This song was already in my library.  None of five songs were added to my library. | 20%  0% |
| *Revised* |  |  |  |  |
| Euclidean Distance | Generalized Dataset  Global Weekly Top 100 Songs Dataset | Five Songs | One of five songs were added to my library.  Two of five songs were selected. These two songs were already in my library. | 20%  40% |
| Cosine Similarity | Generalized Dataset  Global Weekly Top 100 Songs Dataset | Five Songs | One of five songs was selected. This song was already in my library.  Two of five songs were selected. One song was already in my library. | 20%  40% |

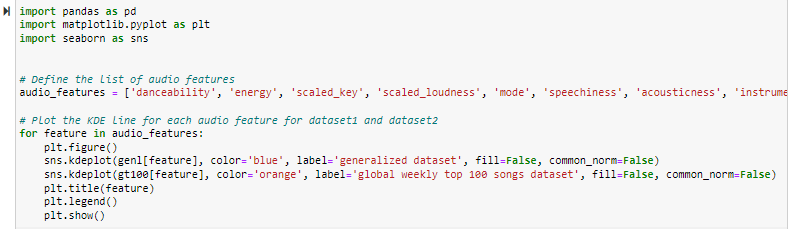
* At this point, it really triggers my curiosity to analyze the generalized dataset further. Here are the steps I followed.
  + Data Quality Recheck
    - Range
    - Missing Values
    - Outliers

I have performed this step in the data preparation assignment, but it is worthwhile for me to double check. Upon re-examining the generalized dataset, I have confirmed that there are no missing values or outliners among the 11 audio features I used to predict the streaming time. The ranges of key, loudness, and tempo were scaled between 0 to 1 using min-max scaling method. Therefore, the data quality is good for modeling.

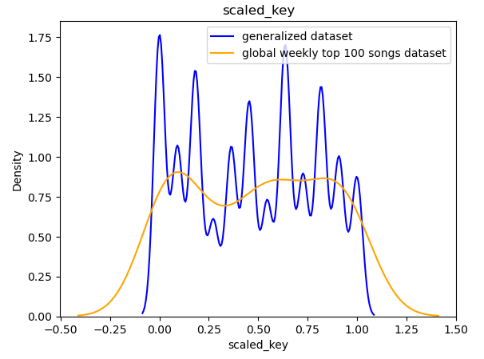
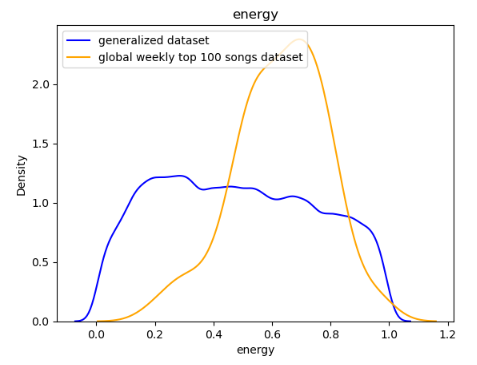
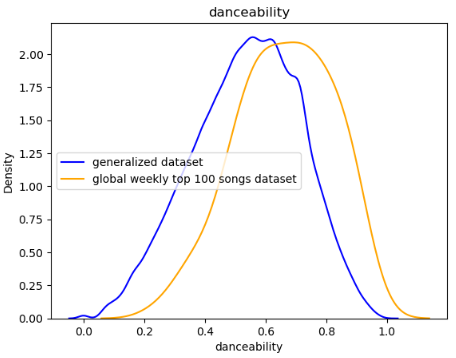
* + Visualizations
    - I decided to visualize the range and distribution of the audio features between the generalized dataset and global weekly top 100 songs dataset to dig into further.
    - *Code*

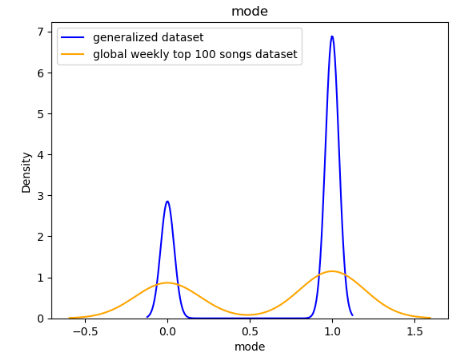
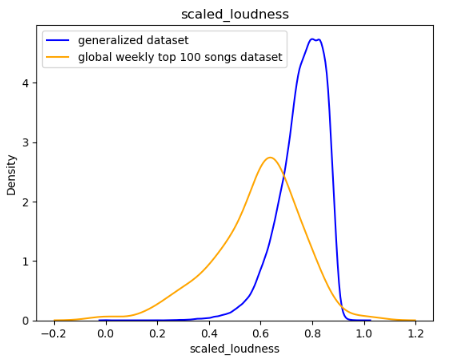
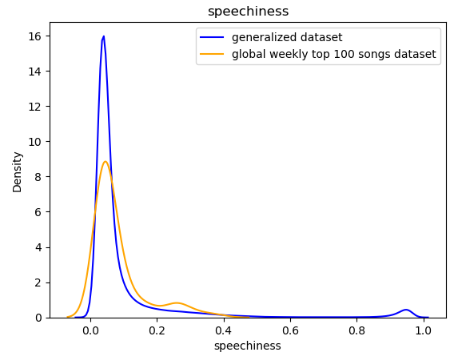


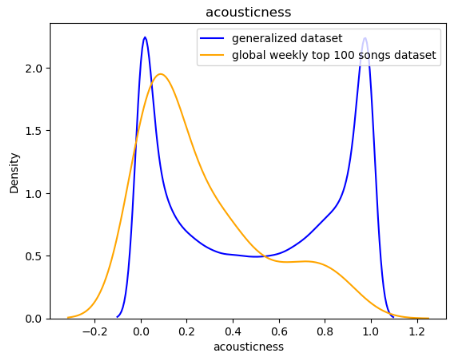
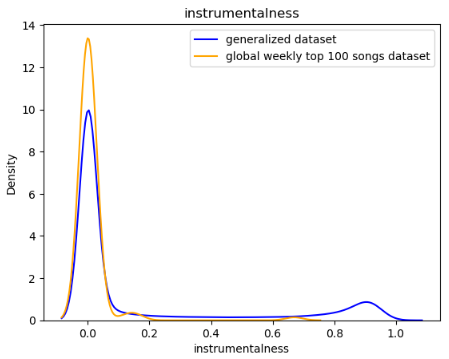
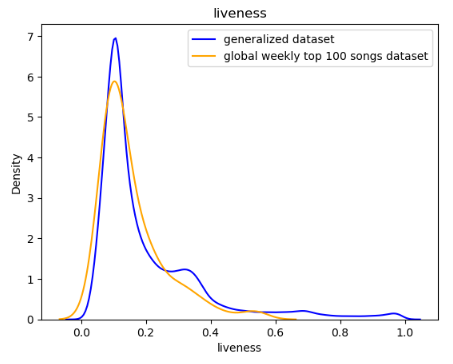


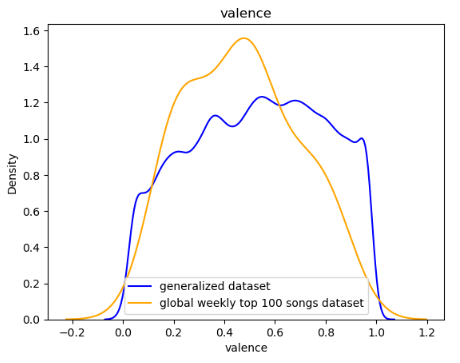
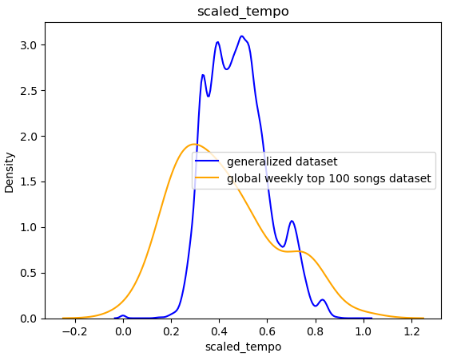


*Output*



* + - I used density (relative value) as the y-axis instead of count (absolute value) because the there was a stark contrast in the number of observations between two datasets. The x-axis is the range of the audio features. Key, loudness and tempo had been scaled. Then I visualized each audio feature between two datasets and found some interesting patterns.
      * Mode, speechiness, instrumentalness, and liveness had similar distributions in both datasets, while the distributions of the other seven audio features differed significantly from one another[[1]](#footnote-1).
      * On average, songs in global weekly top 100 songs dataset have higher danceability, energy but lower loudness, valence and tempo compared to songs in generalized dataset. My guess is that the current trend is leaning towards a genre like techno, which is a type of electronic dance music (EDM) but has a lower loudness and slight lower tempo.
      * The most significant difference between two datasets is key. Songs in generalized dataset have more variances in key compared to songs in global weekly top 100 songs dataset. This makes sense to me because as there are more genres of songs in the generalized dataset while songs in the global weekly top 100 songs dataset focus on some mainstreams with less diversity.
      * Another interesting point is that in terms of acousticness, songs in the generalized dataset are either low (close to 0.05) or very high (close to 1) while songs in the global weekly top 100 songs dataset are generally low (close to 0.1). I am not exactly sure about this difference.
* Based on the above explorations, there are variances of distribution between certain audio features including the danceability, energy, loudness, valence, tempo and acousticness between two test datasets. These variances potentially impact the performances of the models. Hence, while building the model is crucial, the selection of the appropriate test datasets is equally important. In certain cases, even with a powerful model, incorrect selection of the test dataset can result in the failure of the model. For instance, if someone is not a fan of J-pop, using a test dataset comprising solely of J-pop won’t give favorable results no matter how good the model is. In my case, which is not as extreme as the J-pop example, I would prefer to use recent global songs as the test data pool because the generalized dataset consisting of all songs from 1921 to 2020 really is outdated and does not match my preferences.

Now, the models are ready to use. As this is not a company-wise project but a personal project, I do not have the resources to plan for deployment from a company level, such as building the platform for deployment and continuously monitoring. However, the models can still be utilized for recommendation purposes and provide benefits to the public.

Below is a summary of my strategy with the detailed steps.

First, once I finalize my paper, I will share it in my github page. I hope people who are interested in Spotify recommendation system can offer their comments or ask questions of the work I did. By exchanging ideas and thoughts, we can work together to polish the work and improve the algorithms.

Second, I will consolidate the codes (revise if needed) and give instructions on how to apply my codes to user’s personal music data.

Third, I will collect each user’s results, including 1) which supervised learning model has the best performance in each user’s case; 2) which test dataset has a better result in each user’s case; 3) the accuracy of the models.

Fourth, I will conduct an analysis to determine whether the results based on my personal music data overall are consistent with those of other users. If consistent, it would indicate that the algorithms are fairly accurate. If it is not consistent, it would indicate further improvements still need to be made.

At the same time, I will also follow the latest news and research related to this topic and keep myself up to date with the latest developments in this field.

**Task 6.2: Planning Monitoring and Maintenance**

**Deliverable: Monitoring and Maintenance Plan**

This section pertains to the monitoring and maintenance, including reviewing of the model’s performance, updating models based on model performance or changes in available data. I have already completed these steps at the beginning of this assignment to ensure that the models are correct from a code perspective before my strategic planning deployment.

I agree that data-mining work, like application development, is a cycle. As I plan to publicize my recommendation system algorithms, more people will be involved in testing the performance of the models. If there is any discussion arise regarding the model maintenance or change in available dataset, I am open to them.

**Task 6.3: Reporting Final Results**

**Deliverable 1: Final Report**

By now, I haven’t consolidated all the assignments into a single document. I will summarize the entire project by assembling all the reports created up to this point, updating those reports where needed, and adding an overview summarizing the entire project and its results. I probably will use the track changes mode so that you can trace back the changes that I’ve made. The final report will be submitted separately via Blackboard before April 8.

**Deliverable 2: Final Presentation**

By now, I haven’t prepared a slide for summary of the final report to present. I will submit the slide in pdf format before the presentation date, April 10, preferably April 8. I will discuss about any open questions or concerns if necessary during the presentation.

**Task 6.4: Review Project**

**Deliverable: Team Experience Assessment**

Since this is an individual project, there is no collaboration with a team, so I will skip this part. However, it is always good to have some teaming experience. As I plan to share my project paper and codes on my github page, there might be some collaboration with others in terms of feedbacks or suggestions, which will be a good opportunity to gain some teamwork experience.

Note I also attached the codes for this assignment in the submission. The three datasets used in the codes are the same as the ones I had previously sent to you.

1. The density is estimated based on the kernel function and bandwidth used for Kernel Density Estimation (KDE). The density values between two datasets are not directly comparable. The shape and scale of the KDE curves are what most informative in comparing the distributions of the audio features in the two datasets. [↑](#footnote-ref-1)