Spec Category	Spec Details
Formatting	To reproduce this case study you will execute the following rubric chunks in order and provide a CSV with ● The top 10 most relevant lyrics for each of the seven cities ● All required scripts and data can be found in the following GitHub: https://github.com/madisonhgallagher/CS3/tree/main
Purpose	The purpose of this project is to demonstrate technical ability to reproduce TF-IDF analysis and then use critical thinking to analyze the results and make connections to real world events or trends. The student should be able to use the provided repo to execute TF-IDF analysis and then consider how different US cities' culture varies based on the results of the analysis.
Notes***	● A script called “lyric_scraper_script.py” is provided in the repo but is not necessary to be used for project reproduction as all needed data is already supplied to the user. ● The lyric_scraper_script.py may be useful if the user wants to scrape additional data outside of the scope of this case study reproduction
Data	● Locate the “DATA” folder in the provided Github Repo ● Open the “Appendix.pdf” to familiarize yourself with the data for the project ● Download all seven CSV files of lyrical data (one for each city) ● Store the CSVs in an easy to locate folder on your computer
EDA	● Locate the “SCRIPTS” folder in the provided Github Repo ● Open the file called “2.EDA.ipynb” ● Explore the exploratory plots that showcase some general trends for each city. ● Compare and contrast the most

	common genres and artists in each city
Analysis	*The analysis you will conduct is called TF-IDF which means Term Frequency-Inverse Document Frequency. This method of analysis computes the most frequently used words in a series of documents, and punishes for overall frequency. For example, words like "the" or "and" should not score highly because one would expect that they appear in all documents. The goal of using TF-IDF is to identify the words that are most* ***uniquely*** *relevant to a specific city. The TF-IDF function will assign a value between 0 and 1, the higher the value of that term for a specific document, the more relevant it is to that document.* To conduct the analysis: ● Locate the "SCRIPTS" folder in the provided Github Repo ● Open the file called "3.TF-IDF_Analysis.py" ● Download the file and open it in your preferred IDE (google Colab is recommended) ● Edit the file directory in line 17 to reflect the location of the seven CSV files on your computer. ● Run the script in its entirety
Output/Deliverable	*The output of the analysis file will be seven lists with ten words per list. Each list corresponds to one city's top 10 most relevant words.* ● Save the output of the analysis script (seven lists) as a CSV ● Compare how the lists are different and how they are similar ● Consider how many curse words are on the list, what cities have more or less curse words? ● What languages are on the list? Does it vary by city? ● How can you connect these findings to differences in populations within each city? ● Share and upload your results as desired.