My final project in Data Science Programming was designed to create a Jupyter Notebook and answer a data science question. For my project I decided to focus on topics regarding music. I found a csv file with the top listened to songs from 2023 and many different attributes and variables that coincided with each song. I asked four questions during the project:

1. What is the average amount of streams in total?
2. What is the relationship between the artist and their stream count?
3. What is the relationship between length of song and amount of streams?
4. How many artists are in a certain range of streams?

Using the data I was able to find the answers to all of these questions. These questions were designed for the music companies use or for artists. These questions will show what type of songs do better on charts and what should be released for the utmost success.

Before answering the questions, I was required to clean and read the data. To do this, I imported many important python tools such as numpy, pandas, matplotlib.pyplot, seaborn, os, and nltk. Each of these tools played an important part in the process of cleaning the data. After importing these tools, I read my csv file. The file was then output so I could then further look into the data. After reading the file, I explored the data. I used pandas for things such as basic information, head, tail, pulling random samples, finding the names of the columns, the shape of the dataset, the types, and a statistical description of numeric columns.

Next, I worked to clean the data. Firstly, I noticed that the **streams** and **in\_shazam\_charts** columns were a string variable so I changed it to an integer. Then, I replaced any NaN value with ‘0’. I found the outliers in the data and output them to see if any needed removing. I declared all of them to be important which led me to leave them in the dataset. They were important to tell what is doing exceptionally well. After finding the outliers, I created boxplots for each of the rows and printed them in a clean way. I printed the values for the last time and noticed that they were still not as clean as I would prefer them to be, which prompted me to round the numbers. All of these steps led me to a clean dataset.

Subsequent to cleaning the data, I moved to answer my questions. I answered the first question using simple calculations found within the data. The second question I answered using a scatter plot to relate danceability to energy. Question 3 was answered to correlate the amount of streams to BPM using mean calculations and a line style graph. Graph 4 was created to represent the amount of artists within 100,000,000 stream increments.

Overall, this project cleaned and answered music industry related questions.