# Data Wrangling Report

## Project Summary

This project required data wrangling of two different datasets that are different formats and structures and combining those datasets into one file to discover insights about the data. The Spotify Charts dataset is a dataset of the songs that chart on the “Top 200” and “Viral 50” on Spotify which is in a CSV format. The Grammy Awards dataset is a dataset of Grammy award winners which is in a JSON format. The technologies I used to preform the data wrangling and combining the files were Python which is a programming language and Microsoft Excel. To combine the datasets, I had to audit, clean, and convert the data from the datasets before I combined the datasets. The datasets have the artist name and song title in common so this allowed me to combine the data based on those titles. A question I answered is “How many artists have made an appearance on the Spotify Charts and won at least one Grammy?”. The answer is 28 artists. This question gives the number of artists that have had a song chart on the Spotify Charts, and they have also won a Grammy. The number of artists that have their song chart on Spotify Charts and won a Grammy is very low. Another question I answered is “Has the artists of the top 10 most streamed songs on the Spotify Charts won a Grammy from those songs?”. This question required manipulating the data and shows the top 10 streamed songs and whether the artist of that song won a Grammy for that song. The answer is no, the top 10 streamed songs have not caused the artist of that song to win a Grammy. These are the top 10 streamed songs however they are not Grammy winning songs. Another question I answered was “Which Grammy Award category has the most artists that have made an appearance on the Spotify Charts and won a Grammy in that category?”. The answer to this question is “Song of the Year”. The question suggests that the “Song of the Year” Grammy Award winner category has the most artists that chart on the Spotify Charts in comparison to the other Grammy categories.

# Wrangling Details

Spotify Charts dataset:

Origin: https://www.kaggle.com/datasets/dhruvildave/spotify-charts

Characteristics: The format of this dataset is a CSV file. It contains 9 columns, title (the title of the song), rank (the rank of the song), date (the date that the Spotify chart was published), artist (the artist of the song), url, (the Spotify URL for the song), region (the region of the Spotify chart), trend (the trend of the song on Spotify charts), and streams (the number of streams of the song). The dataset has millions of rows of data.

Initial Audit: The dataset was unable to be opened by Excel and VSCode. The dataset was too large of size for the technologies used in the BUSAN 300 course therefore a subset of the dataset is required. There are a lot of rows of the same song if it is on multiple Spotify charts for different dates. There are blank values for the streams column when the "Chart" column is "Viral 50".

Grammy Awards dataset:

Origin: https://www.kaggle.com/datasets/theriley106/grammyawardsinnumbers

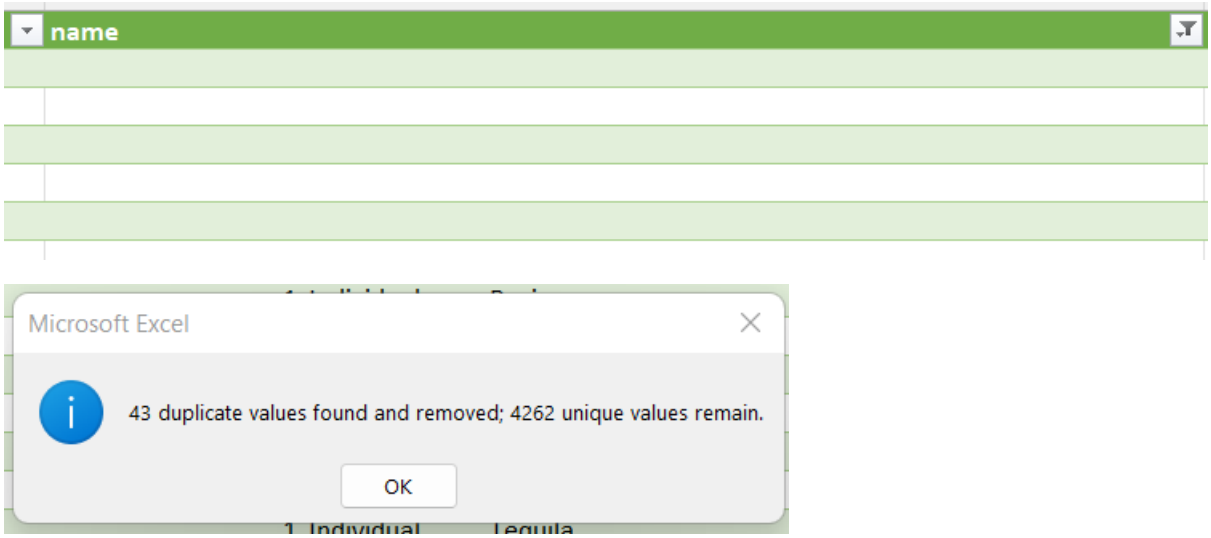
Characteristics: The format of the dataset is a JSON file. It contains 5 columns, name (the name of the recipient of the award), awardType (the type of the award), category (the category of the award), annualGrammy (the number of the Grammy award show), and awardFor (what the award is for). The dataset contains 4305 JSON objects.

Initial Audit: The dataset is in JSON format, so it must be converted to a CSV file so it would be easier to inspect. There are some NULL fields for the name and awardFor columns. The JSON objects in the dataset have a consistent format therefore it would be easy to convert it to a CSV using a Python program. There are some duplicate rows.

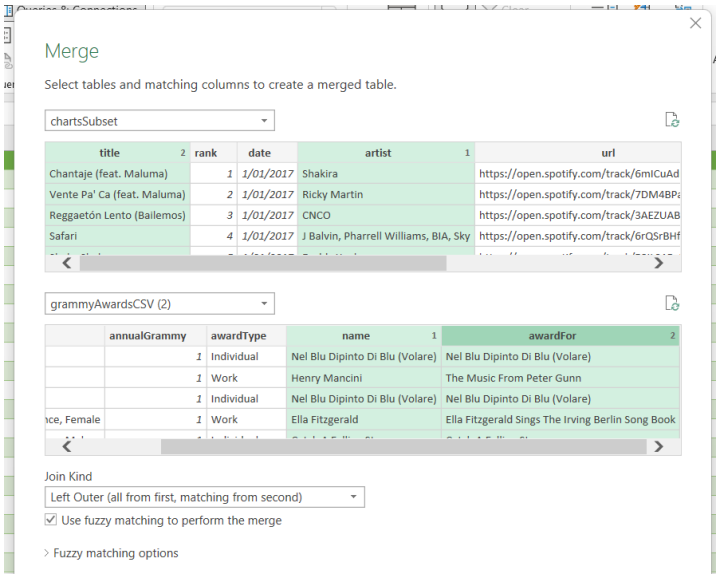
Steps performed to combine datasets:

The Spotify Charts dataset was too large for the tools used in BUSAN 300 therefore a sufficiently large subset of the dataset is required so I wrote a Python program to extract the header values and tens of thousands of rows of data from the Spotify Charts CSV file and write that data into a new CSV file. The program uses the CSV module that Python provides. The module allows the CSV file to be read and write the subset of the CSV file to another CSV file.

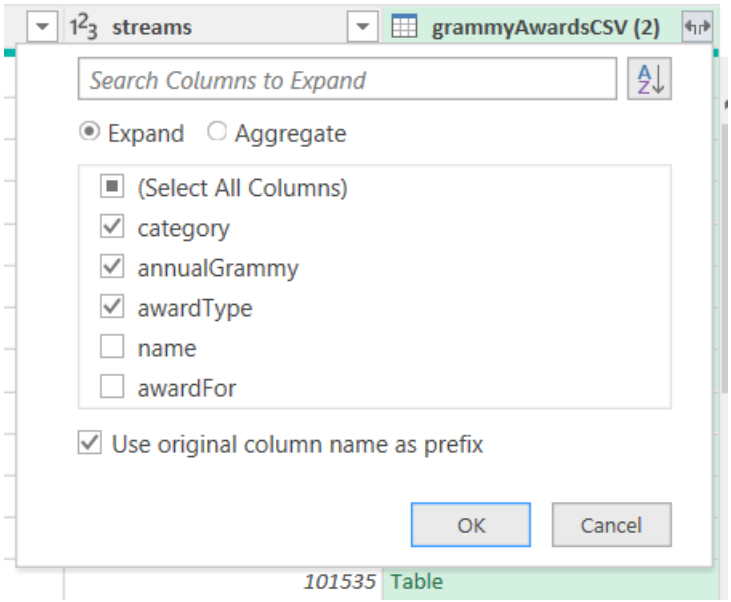
The Grammy Awards dataset had to be converted into a CSV file so that both datasets are in similar format and easier to inspect therefore I wrote a Python program to convert the JSON file to CSV. The program extracted the header values and the data and wrote the data into a new CSV file. The program uses the CSV module and the JSON module that Python provides. The module allows the JSON file to be read and write the contents of the JSON into a CSV file using the CSV module.




Using Excel, I imported the Spotify Charts dataset and the Grammy Awards dataset. The Grammy Awards dataset has duplicate rows, so I removed the duplicates. The Grammy Awards dataset has rows with no values, so I filtered the name and awardFor column to check which rows have “blanks” and deleted those rows. The other columns didn’t have any “blanks”. Rows should not have blank values for the name and awardFor column because the datasets will be combined using the name (artist) and song title (awardFor).




Using the Merge function on Excel, I combined both datasets based on the artist name column and the song title column. The Merge function performs a Left Outer Join on the datasets. I selected fuzzy matching so that the merge is not for exact match.




The merge function provides the Grammy Awards dataset as a table in a column of the Spotify Charts dataset so I expanded the table in PowerQuery and selected the columns that should be in the combined dataset. I unselected the name and awardFor columns because these columns are already in the combined dataset (artist and title).

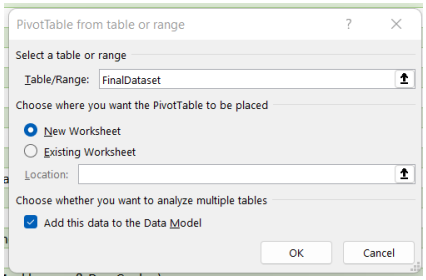
The two datasets are combined into a Excel workbook. I prefixed the columns that are from the Grammy Awards dataset with the grammyAwardCSV (2) to make it easier to recognise which columns are from the Grammy Award dataset.

The link to the combined Final dataset:

https://docs.google.com/spreadsheets/d/1T-B-zeNeBHDEBobzTBUeC7Lk2P0X1dcq/edit?usp=sharing&ouid=109419367565294155278&rtpof=true&sd=true

# Questions and Answers

How many artists have made an appearance on the Spotify Charts and won at least one Grammy?





| grammyAwardsCSV (2).category | (Multiple Items) |
|---|---|
| | |
| **Row Labels** | **Distinct Count of artist** |
| A Great Big World, Christina Aguilera | 1 |
| Adele | 1 |
| Alabama Shakes | 1 |
| Alicia Keys | 1 |
| Aretha Franklin | 1 |
| Beyoncé | 1 |
| Bruno Mars | 1 |
| Childish Gambino | 1 |
| Drake | 1 |
| Ed Sheeran | 1 |
| Eminem | 1 |
| Hank Solo | 1 |
| Imagine Dragons | 1 |
| John Mayer | 1 |
| Justin Hurwitz | 1 |
| Kanye West | 1 |
| Kanye West, Jamie Foxx | 1 |
| Kendrick Lamar | 1 |
| Kings of Leon | 1 |
| Maroon 5 | 1 |
| Michael Jackson | 1 |
| Portugal. The Man | 1 |
| Red Hot Chili Peppers | 1 |
| Rihanna, Calvin Harris | 1 |
| Sam Smith | 1 |
| The Chainsmokers | 1 |
| Twenty One Pilots | 1 |
| Whitney Houston | 1 |
| **Grand Total** | **28** |

28 artists.

Using Excel, I used a PivotTable and put the artist column in the Rows field, the Grammy Award category into the Filters field, and the distinct count of the artist column into the Values field of the PivotTable. I filtered the “blanks” out of the Grammy category column as there are “blanks” in the Grammy Awards category column to represent that the artist in that row has not won a Grammy.

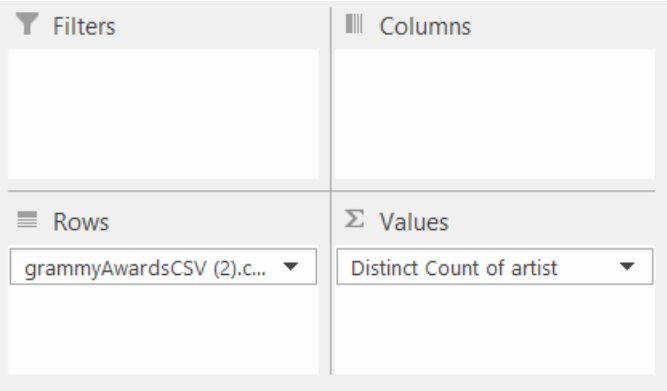
Have the artists of the top 10 most streamed songs on the Spotify Charts won a Grammy from those songs?

| | streams | grammyAwardsCSV (2).category |
|---|---|---|
| ON | 7572795 | |
| ON | 5999224 | |
| ON | 4293519 | |
| ON | 3342769 | |
| ON | 3334232 | |
| ON | 3284281 | |
| ON | 3135625 | |
| | 3046692 | |
| | 3042293 | |
| ON | 3019058 | |
| ON | 3015525 | |

No.

Using Excel, I filtered the streams column from largest to smallest and manually checked if the column of the Grammy category was blank for the top 10 rows as a blank value for the category column means that the artist of that song has not won a Grammy for that song.

Which Grammy Award category has the most artists that have made an appearance on the Spotify Charts and won a Grammy in that category?




| Row Labels | Distinct Count of artist |
|---|---|
| Song Of The Year | 4 |
| Best Short Form Music Video | 3 |
| Best Pop Duo/Group Performance | 3 |
| Record Of The Year | 3 |

Song of the Year.

Using Excel, I used a PivotTable and put the Grammy Award categories in the Rows field and the distinct count of artist in the Values field of the PivotTable. I filtered the “blanks” out of the Grammy Award category column. I sorted the distinct count of artist by largest to smallest.