Project Assignment 5

Group Members: Amy Wang, Mickey Chen, Jyun-Hao Chen

Date: 05/10/2023

Background

Music has been such an impactful element in our life. There are everywhere; everything in our life is all about music. The music seamlessly integrates into commercials, movies, and various industries, establishing itself as one of the most lucrative enterprises worldwide. Thus, our team would like to research the transition of popular music. What kind of music that people love more nowadays? Are fast-paced or slow-paced songs more favored by individuals today? By analyzing the data, we also want to predict what kinds of music will be more prevalent in the future. If people want to make money through music, what type of music and genre will be their best chance?

Data, cleaning, and preparation

We found two data sets on GitHub: Billboard's weekly Top 100 record of songs and Spotify's song information data. Data cleaning involves several key steps to ensure data quality and integrity. Irrelevant columns, including URL, Track_id, Instance, Spotify_track_preview_url, Album, Spotify_track_album, and Spotify_track_popularity, were dropped. Data type conversions were performed, and the week_id column was converted to DateTime format. Missing values (NULL) were dropped from both datasets, resulting in a reduced dataset with 24,082 rows.

After cleaning the data and corresponding to our research question, we used song_id (song name+singer) as the key value. The Billboard and audio datasets were then merged using the song_id column, resulting in a new dataset with 282,371 rows.

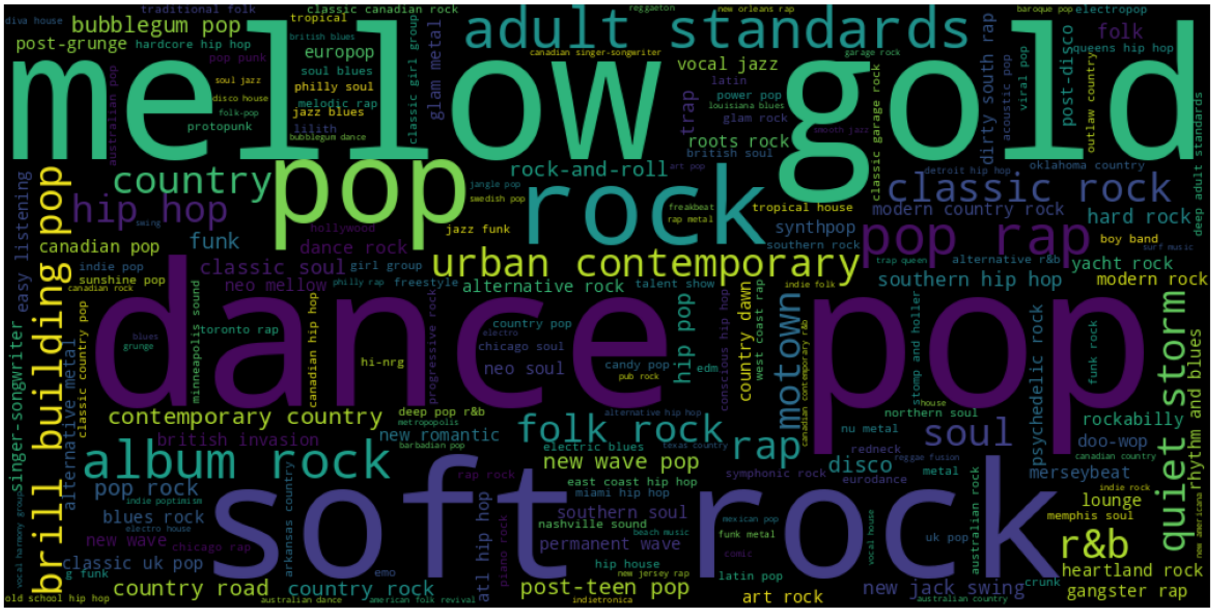
In the part of derivative data, we use the music attributes data of the latest two years as the analysis period to study the popular trend of music. We calculated the mean of the top 100 song's attributes for two years every week, so generating a line chart to view song trends later is convenient.

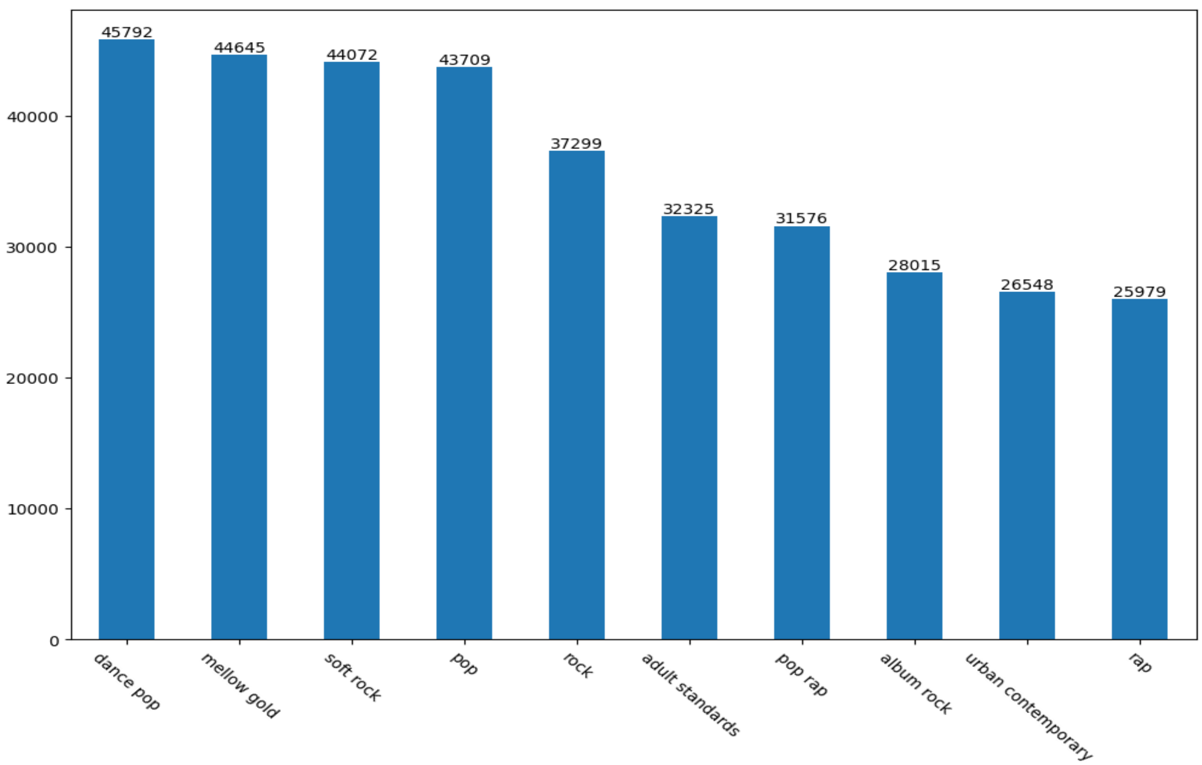
Research Question

Question 1: *What are the most frequently charted music genres in history, and how have music attributes changed in a period?*

<u>Genre</u>

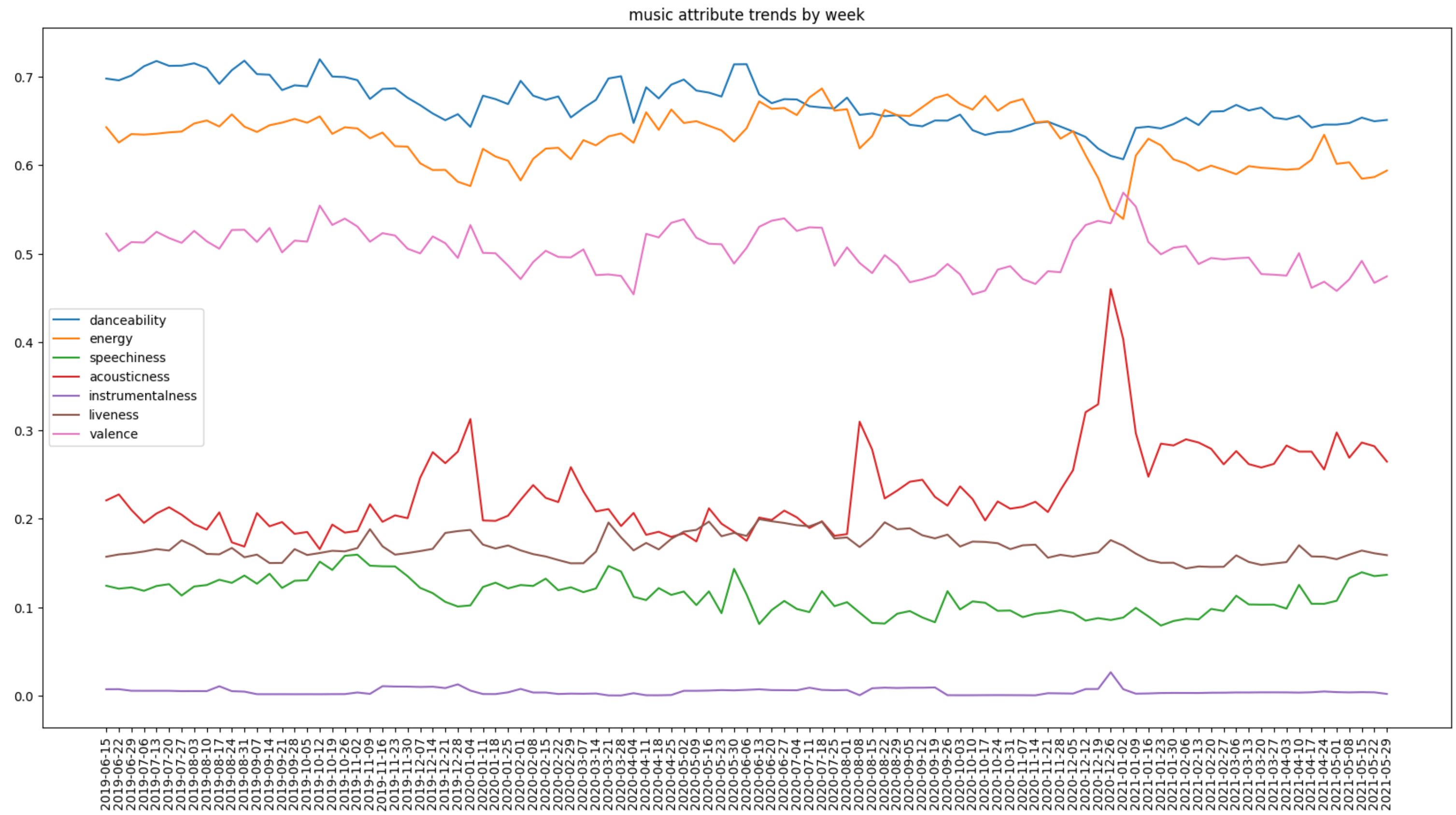
For the genre, dance, mellow gold, and soft rock are the most popular genres over time.





Music attributes

The danceability and energy are usually higher among the songs on the list in the past two years. And acoustics has a peak of rising in December.



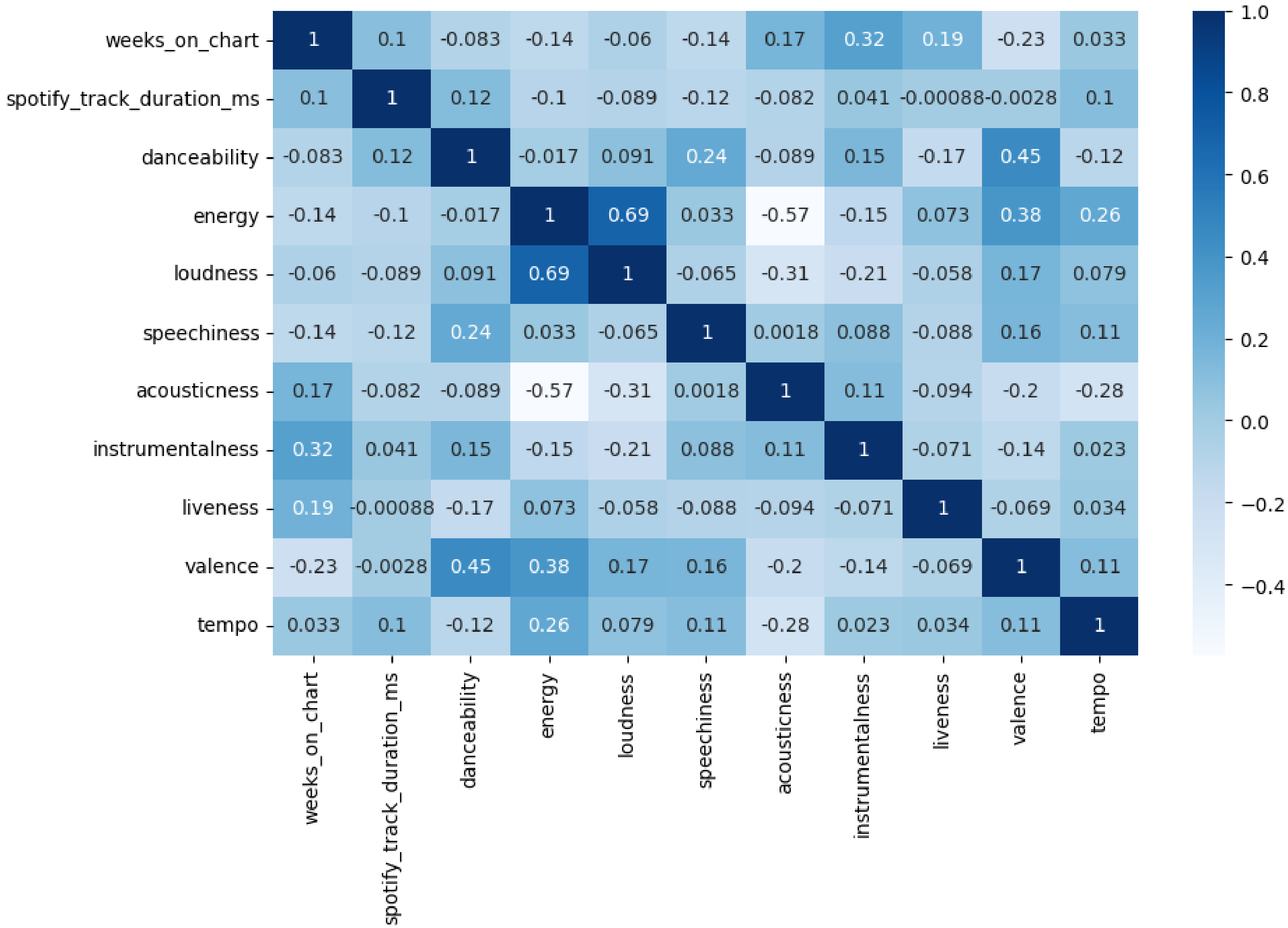
Question 2: *Is there a correlation between music attributes and popularity on Billboard? And the prediction of music trends for the 2020s.*

<u>Correlation of the musical attributes</u>

Simply put, a correlation test is a statistical technique used to determine if variables are related, considering both the strength and direction of their relationship. To analyze the correlation between musical attributes, we create a correlation matrix and visualize it as a heatmap.

The correlation matrix is based on the Pearson correlation coefficient, which shows a positive correlation in dark blue color and a negative in light blue color. Based on the heatmap, the following inferences can be made::

- There is a strong positive correlation between 'loudness' and 'energy'.
- 'Valence' has a strong positive correlation with 'energy' and 'danceability.'
- 'Acousticness' has a negative relationship with 'energy' and 'loudness'.
- Most of the correlations between music attributes are low, indicating the absence of multicollinearity.



<u>Multiple Linear Regression Model</u>

R-squared is a statistical measure that indicates the proportion of the variation in a dependent variable explained by the independent variables in a regression model.

We assume the null hypothesis H_0: the coefficient of independent variables are 0. The alternative hypothesis H_1 : the coefficient of independent variables are not 0.

According to the summary of the linear regression below, most music attributes (independent variables) with p-values > 0.05 indicate that they are not significantly different from 0. We failed to reject the null hypothesis H_0: the coefficient of independent variables are 0. The R-squared scores show the model's independent variables can only explain 23% of the variation in the dependent variable. (The ratio of our model's independent variables to sample sizes is 1:10.)

There are several reasons for failing to reject the null hypothesis:

- The effect truly doesn't exist in the population.

- The sample size was too small to detect the effect.
- The variability in the data was too high. The effect exists, but the noise in the data swamped the signal (effect).
- Collected a fluky sample. When dealing with random samples, chance always affects the results. The luck of the draw might have caused the sample not to reflect an effect that exists in the population.

OLS Regression Results

Dep. Variable:	weeks_on_chart	**R-squared:**	0.229
Model:	OLS	**Adj. R-squared:**	0.142
Method:	Least Squares	**F-statistic:**	2.637
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00734
Time:	20:11:23	**Log-Likelihood:**	-320.58
No. Observations:	100	**AIC:**	663.2
Df Residuals:	89	**BIC:**	691.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	53.0083	10.372	5.111	0.000	32.399	73.618
spotify_track_duration_ms	1.748e-05	2.01e-05	0.872	0.386	-2.24e-05	5.73e-05
danceability	-0.7473	6.425	-0.116	0.908	-13.514	12.020
energy	-3.5542	7.811	-0.455	0.650	-19.074	11.966
loudness	0.4504	0.474	0.950	0.345	-0.492	1.393
speechiness	-12.7439	11.371	-1.121	0.265	-35.338	9.850
acousticness	4.9271	3.940	1.250	0.214	-2.903	12.757
instrumentalness	34.6931	10.895	3.184	0.002	13.044	56.342
liveness	13.7441	6.083	2.259	0.026	1.657	25.831
valence	-3.8104	4.030	-0.945	0.347	-11.818	4.197
tempo	0.0217	0.026	0.833	0.407	-0.030	0.074

Omnibus:	48.226	**Durbin-Watson:**	0.475
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	127.572
Skew:	1.791	**Prob(JB):**	1.99e-28
Kurtosis:	7.218	**Cond. No.**	5.12e+06

Then, we remove those independent variables that are not significantly different from 0 (p>0.05). Consequently, the model's R-squared decreased to 0.146. This phenomenon occurs because R-squared tends to increase as more independent variables are joined in the model. Therefore, when we remove the independent variables with no significant difference, the R-squared declines as lesser independent variables are in the model. Hence, it is advisable to consider the Adjusted R-squared when using multiple linear regression.

According to the result, the Adjusted R-squared is only 0.129, indicating that this model fails to predict music trends for the 2020s. Since there are more factors that will affect the songs on Billboard, such as the advertisement, play counts, or social media, these missing data will provide a good future research aspect to dig into those potential avenues.

OLS Regression Results

Dep. Variable:	weeks_on_chart	**R-squared:**	0.146
Model:	OLS	**Adj. R-squared:**	0.129
Method:	Least Squares	**F-statistic:**	8.306
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.000468
Time:	20:11:24	**Log-Likelihood:**	-325.65
No. Observations:	100	**AIC:**	657.3
Df Residuals:	97	**BIC:**	665.1
Df Model:	2		
Covariance Type:	nonrobust		

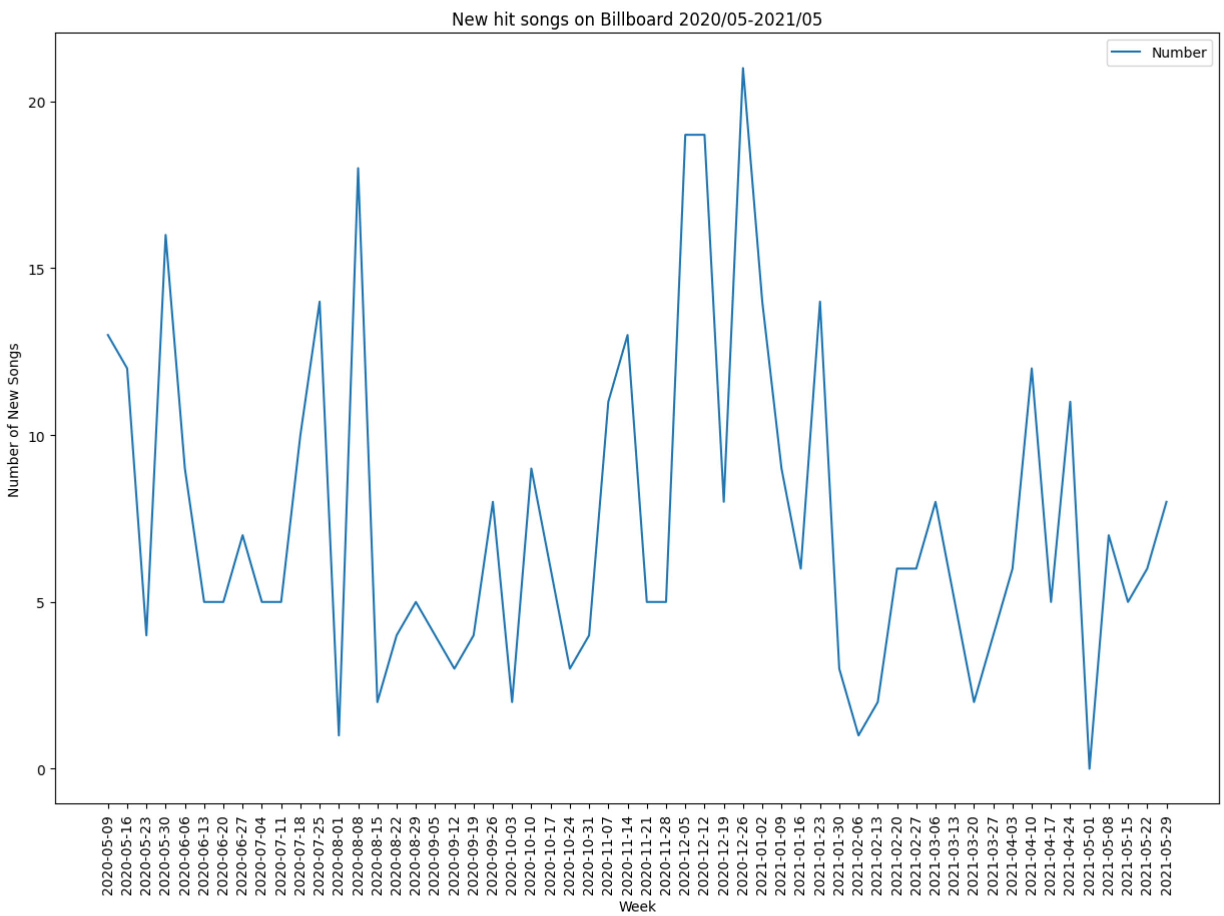
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	52.2948	1.137	46.007	0.000	50.039	54.551
instrumentalness	35.7923	10.117	3.538	0.001	15.713	55.872
liveness	13.4826	5.939	2.270	0.025	1.695	25.270

Omnibus:	50.779	**Durbin-Watson:**	0.326
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	137.679
Skew:	1.890	**Prob(JB):**	1.27e-30
Kurtosis:	7.330	**Cond. No.**	16.1

Question 3: *Do audiences prefer fixed songs, or are new songs introduced weekly?*

By calculating the number of new songs per week, it is possible to understand the audience's music-listening preferences. More specifically, Billboard has 100 songs a week. We record the song_id of the first week and compare the song_id of the second week, and the third week continues to compare. We can calculate quantitative results.

A line chart shows the number of new songs on the Billboard weekly. The results show that the number of new songs on Billboard changes between 5-15 songs weekly, meaning that listeners prefer to listen to fixed songs.



Conclusion

Based on the research experience, we have three key takeaways.

1. **Regarding the topic that could be related to strong personal preferences like art, the dataset probably cannot extrapolate outcomes accurately.**

As you can see, our analytics result is hard to explain the music trend and why specific music categories are popular. After discussion, we believe there can be some extra factors that fluctuate the music's popularity. For instance, if a fan is an ardent follower of Taylor Swift, regardless of music attributes or genre, they are likely to listen to her songs and vote for them. This situation could lead to opposite music attributes (fast and slow songs) intersecting on Billboard's rank, weakening the correlation between music attributes and popularity. Thus, we think topics containing too much personal preceptive could easily meet this situation and be challenging to analyze.

2. **A big dataset doesn't equal a good dataset.**

Initially, we were drawn towards this dataset due to our interests and the vast amount of data it contained. Unfortunately, we encountered a significant hindrance in our analysis as many of the 300,000 data points were impractical. Specifically, these two datasets were plagued with numerous null values and outliers, challenging our analysis. After data cleaning resulted in a loss of approximately 15% of the original dataset, which was a significant proportion and may have potentially distorted our final results.

3. **Storytelling is what makes datasets and analyses valuable.**

While many individuals can analyze and visualize data independently, their deliverables may not always be compelling to their intended audience. While the audience may be able to understand the presented results and visualizations, they may need to fully grasp the analytical process and rationale behind the approach used to arrive at those results. This is where storytelling can critically present a comprehensive view of the data analysis process and why specific decisions were made. By crafting a compelling narrative, data professionals can help their audience fully understand the data and its implications, leading to more informed and reasonable decision-making.

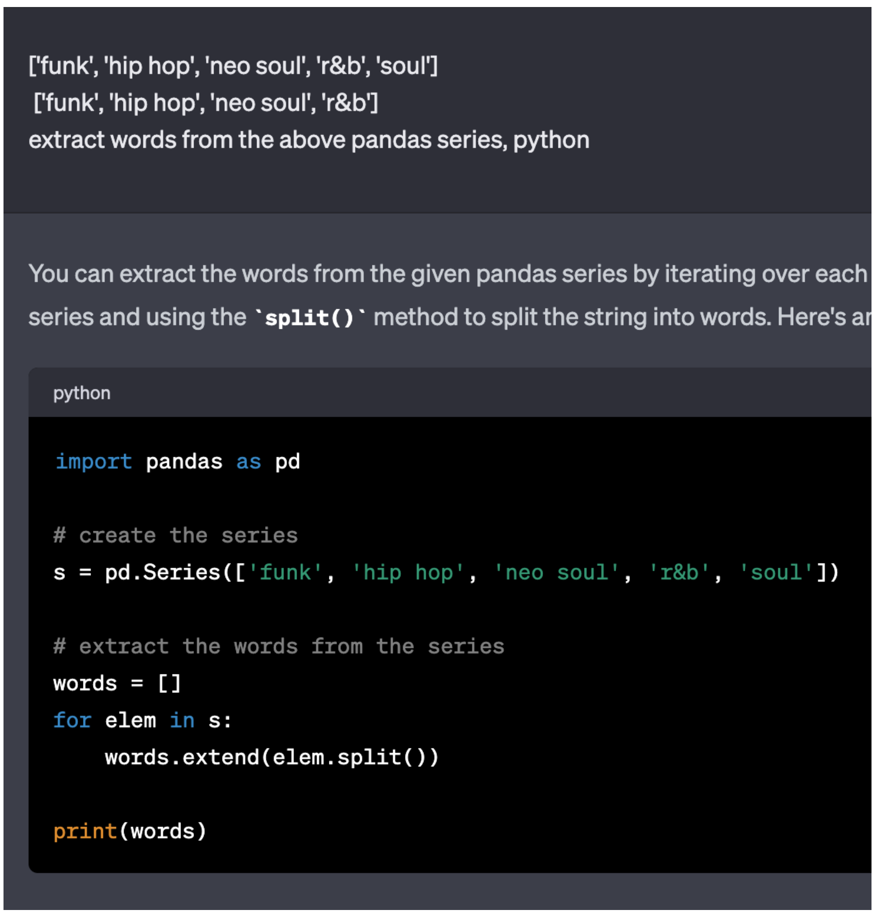
Appendix 1: Original Code

Colab:
https://colab.research.google.com/drive/1SueQeLtT8FnBzm_F0-3FCM00atWwARu-?usp=sharing

Appendix 2: ChatGPT Usages

1. Query Code

We often use chatgpt to query code. For example, we want it to generate code and examples of specific functions so that we can understand the usage and apply modifications to the project code. Below are functions of split and lineplot generated by chatgpt. This approach has greatly shortened the time we spent searching for the correct functions before, and the examples it provided are also relatively concise and easy to understand.

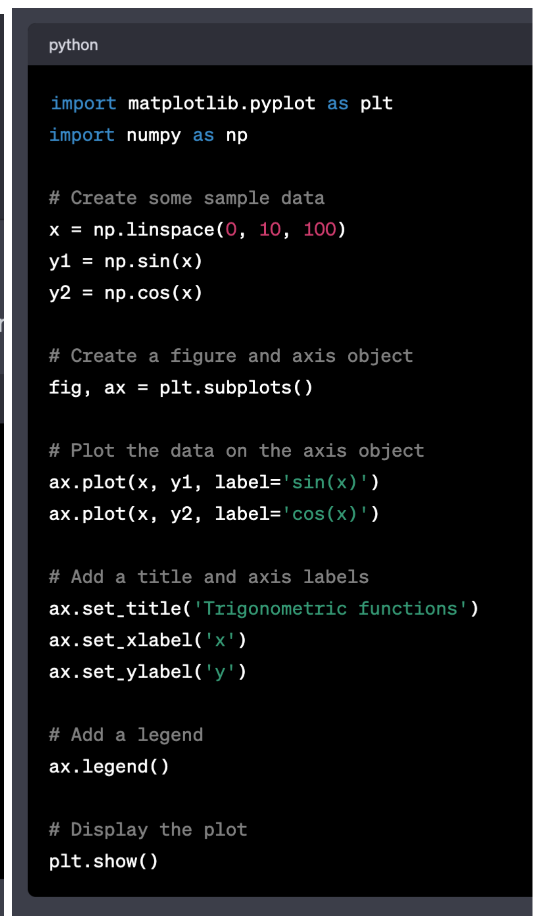


['funk', 'hip hop', 'neo soul', 'r&b', 'soul']
['funk', 'hip hop', 'neo soul', 'r&b']
extract words from the above pandas series, python

You can extract the words from the given pandas series by iterating over each
series and using the `split()` method to split the string into words. Here's ar

python

```python
import pandas as pd

# create the series
s = pd.Series(['funk', 'hip hop', 'neo soul', 'r&b', 'soul'])

# extract the words from the series
words = []
for elem in s:
    words.extend(elem.split())

print(words)
```

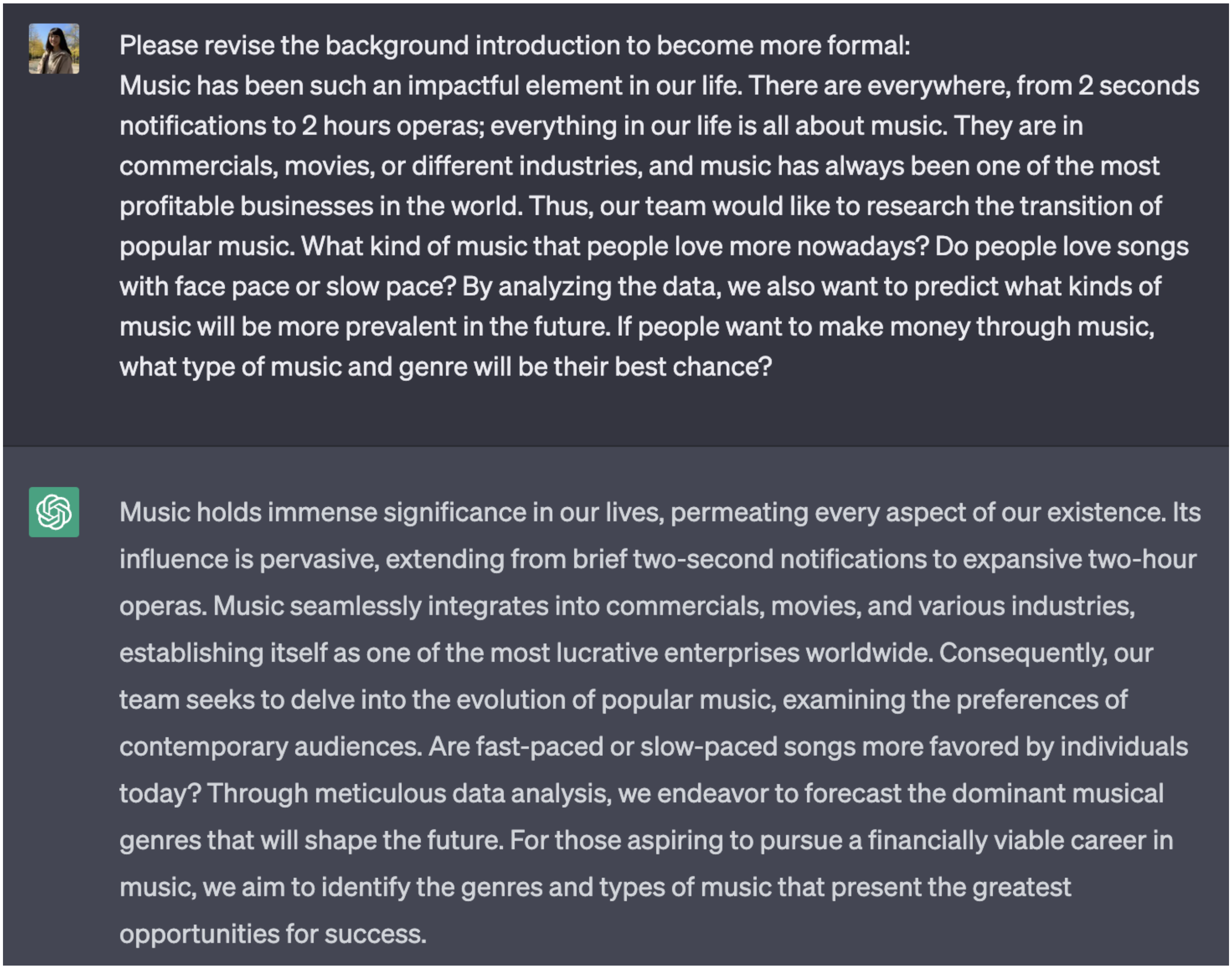


python

```python
import matplotlib.pyplot as plt
import numpy as np

# Create some sample data
x = np.linspace(0, 10, 100)
y1 = np.sin(x)
y2 = np.cos(x)

# Create a figure and axis object
fig, ax = plt.subplots()

# Plot the data on the axis object
ax.plot(x, y1, label='sin(x)')
ax.plot(x, y2, label='cos(x)')

# Add a title and axis labels
ax.set_title('Trigonometric functions')
ax.set_xlabel('x')
ax.set_ylabel('y')

# Add a legend
ax.legend()

# Display the plot
plt.show()
```

2. **Generate descriptive body text.**

We used project assignment 4 as the fundamental and revised the background text to make it more professional. ChatGPT often uses more formal and professional words that we seldom use. It makes the whole paragraph look more professional, but sometimes it takes effort to read due to its wordy content.



Please revise the background introduction to become more formal:
Music has been such an impactful element in our life. There are everywhere, from 2 seconds notifications to 2 hours operas; everything in our life is all about music. They are in commercials, movies, or different industries, and music has always been one of the most profitable businesses in the world. Thus, our team would like to research the transition of popular music. What kind of music that people love more nowadays? Do people love songs with face pace or slow pace? By analyzing the data, we also want to predict what kinds of music will be more prevalent in the future. If people want to make money through music, what type of music and genre will be their best chance?

Music holds immense significance in our lives, permeating every aspect of our existence. Its influence is pervasive, extending from brief two-second notifications to expansive two-hour operas. Music seamlessly integrates into commercials, movies, and various industries, establishing itself as one of the most lucrative enterprises worldwide. Consequently, our team seeks to delve into the evolution of popular music, examining the preferences of contemporary audiences. Are fast-paced or slow-paced songs more favored by individuals today? Through meticulous data analysis, we endeavor to forecast the dominant musical genres that will shape the future. For those aspiring to pursue a financially viable career in music, we aim to identify the genres and types of music that present the greatest opportunities for success.

3. Abstract for long paragraphs.

For the data cleaning and preparation paragraph, we already have a long and detailed description of the steps we have done to do the data cleaning. It is fast to ask ChatGPT to help us abstract and shorten the paragraph with only the main steps. One interesting thing is that, at first, ChatGPT may only shorten part of the text into 3 paragraphs. Therefore, after providing the first generated text, we asked ChatGPT to regenerate again with a more precise and shortened paragraph.

Abstract:

The data cleaning process involved several key steps to ensure the quality and integrity of the datasets used in the research. Firstly, seven columns that were deemed irrelevant to the research questions were dropped, including URL, Track_id, Instance, Spotify_track_preview_url, Album, Spotify_track_album, and Spotify_track_popularity. This step aimed to streamline the dataset and focus on the essential variables.

Next, data type conversions were performed, with the week_id column being converted to DateTime format to facilitate time-based calculations and analysis. Other columns remained unchanged as their original data types were appropriate for the analysis.

Duplicate rows were handled differently for the Billboard and audio datasets. In the Billboard dataset, duplicate rows were retained as each instance represented a unique occurrence of a song on the chart, contributing to its popularity. However, in the audio dataset, duplicate rows based on the song_id were removed to ensure only one unique row per song remained, facilitating subsequent merging with the Billboard dataset.

Missing values were addressed by dropping rows with incomplete data. The "Genre" column had a missing data rate of 4%, while the 11 musician attributes columns had a missing data rate of approximately 15%. Filling in the missing data was deemed inappropriate as the nature of the characters and numeric values made it impossible to accurately define them. After dropping the missing values, a reduced dataset with 24,082 rows remained.

Finally, the Billboard dataset and audio dataset were merged using the song_id column as a unique identifier. The merge operation combined the quantitative data from both datasets, resulting in a merged dataset with 282,371 rows and 23 columns.

Outliers were checked to ensure data integrity, and no outliers were found in the dataset. The range of the data was within the expected limits, with the earliest year being 1958 and the latest year being 2021.

Overall, the data cleaning process involved eliminating irrelevant columns, converting data types, handling duplicates, addressing missing values, merging datasets, and checking for outliers, resulting in a refined dataset ready for further analysis and research.

The 1st generated paragraph that ChatGPT generated is still too long and wordy.

May you shorten the abstract you generate into 3 paragraphs only?

Abstract:

The data cleaning process involved several key steps to ensure data quality and integrity. Irrelevant columns, including URL, Track_id, Instance, Spotify_track_preview_url, Album, Spotify_track_album, and Spotify_track_popularity, were dropped. Data type conversions were performed, and the week_id column was converted to DateTime format. Duplicate rows were handled differently for the Billboard and audio datasets, with duplicate song_id rows removed from the audio dataset. Missing values were dropped from both datasets, resulting in a reduced dataset with 24,082 rows. The Billboard and audio datasets were merged using the song_id column, resulting in a merged dataset with 282,371 rows. Outliers were checked and found to be within the expected range.

The data cleaning process involved eliminating irrelevant columns, converting data types, handling duplicates, addressing missing values, merging datasets, and checking for outliers. The resulting refined dataset is ready for further analysis and research.

The 2nd revised version is much better and shorter, and we just needed to make some revisions.

4. **Check the precision of the language used to describe the correlation and multiple linear regression conclusion.**

We also use the ChatGPT to help us check the description of the multiple regression and correlation model. When it comes to statistic analysis, language, and vocabulary usage should be very precise and correct. Therefore, we also used ChatGPT to help us check the usage to ensure we did not say something wrong, which was useful.

Please check the usage of words and grammer of this description of the multiple linear regression model:

R-squared is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model.

We assume the null hypothesis H0: the coefficient of independent variables are 0. The alternative hypothesis H1 : the coefficient of independent variables are not 0.

According to the summary of the linear regression below, most music attributes (independent variables) p-values > 0.05 indicate that they are not significantly different from 0. We failed to reject the null hypothesis H0: the coefficient of independent variables are 0. And the R-squared scores show the model's independent variables can only explain 23% of the dependent variable variation. (Our models' independent variables to sample sizes are 1: 10.)

The original command for ChatGPT.

The usage of words and grammar in the description of the multiple linear regression model appears to be correct. However, there are a few minor suggestions for improvement:

R-squared is a statistical measure that indicates the proportion of the variation in a dependent variable explained by the independent variables in a regression model.

We assume the null hypothesis H0: the coefficients of the independent variables are 0. The alternative hypothesis H1: the coefficients of the independent variables are not 0.

According to the summary of the linear regression below, most music attributes (independent variables) with p-values > 0.05 indicate that they are not significantly different from 0. We fail to reject the null hypothesis H0: the coefficients of the independent variables are 0. The R-squared score indicates that the model's independent variables can only explain 23% of the variation in the dependent variable. (The ratio of our model's independent variables to sample sizes is 1:10.)

There are several reasons for failing to reject the null hypothesis:

- The effect truly doesn't exist in the population.
- The sample size was too small to detect the effect.
- The variability in the data was too high, and the effect was swamped by noise.
- The sample collected was unrepresentative due to chance.

We then remove the independent variables that are not significantly different from 0 (p > 0.05). Consequently, the model's R-squared decreases to 0.146. This phenomenon occurs because R-squared tends to increase with the addition of more independent variables. Therefore, removing non-significant independent variables results in a decline in R-squared. Hence, it is advisable to consider the Adjusted R-squared when using multiple linear regression.

Based on the results, the Adjusted R-squared is only 0.129, indicating that this model fails to predict music trends for the 2020s. Other factors such as advertisements, play counts, and social media influence the songs on the Billboard charts, presenting potential avenues for future research.

The suggestion from ChatGPT.

The description of correlation is generally correct. However, here are some suggested revisions to improve grammar and clarity:

A correlation test is a statistical technique used to determine if variables are related, considering both the strength and direction of their relationship. To analyze the correlation between musical attributes, we create a correlation matrix and visualize it as a heatmap.

The correlation matrix is based on the Pearson correlation coefficient, which is represented by dark blue for positive correlation and light blue for negative correlation. Based on the heatmap, the following inferences can be made:

- There is a strong positive correlation between 'loudness' and 'energy'.
- 'Valence' shows a strong positive correlation with 'energy' and 'danceability'.
- 'Acousticness' exhibits a negative relationship with 'energy' and 'loudness'.
- Most of the correlations between music attributes are low, indicating the absence of multicollinearity.

Overall, the description accurately explains the concept of correlation and provides insights into the relationships between the musical attributes.

5. Help generate the conclusion of the whole report.

Since we have already made most of the report's conclusions, we also asked ChatGPT to help us generate some general takeaways and conclusions for the whole report, then finalize and modify by ourselves.

Based on the research experience, we have there key takeaways.

Based on the research experience, we can draw three key takeaways:

1. Music genres have changed over time, but dance, mellow gold, and soft rock have been the most popular genres throughout history. In terms of music attributes, danceability and energy have been increasing in the past two years, while acousticness has a peak of rising in December.
2. While there is a correlation between certain music attributes and Billboard popularity, our multiple linear regression model failed to predict music trends for the 2020s due to the limited number of independent variables and missing data on other important factors such as advertising, play counts, and social media.
3. Our analysis on the number of new songs on Billboard weekly indicates that listeners prefer fixed songs rather than new songs. This may have implications for the music industry and marketing strategies for musicians and record labels.

Overall, this research provides insights into the history and trends of popular music and highlights the need for further investigation into the factors affecting music popularity and trends.