Project Final

Project title8

- *Two Decades of Spotify: Analyzing Spotify's Top Hits (2000-2019)*

Team Members

- Erick Chacon
- Mariyah Shahmalak
- Brandon Sobremisana
- Ibrahim Kanu
- Lawrence Aryeh

Summary of Group Project, Accomplishments, and Learning Experience

- Our group project, "Two Decades of Spotify: Analyzing Spotify's Top Hits (2000-2019)," focused on exploring various aspects of popular music over 20 years using a dataset containing information on the top 2000 songs. Each group member took on specific questions and analyses, leveraging different Python tools to draw meaningful insights from the data.
- **Group Accomplishments:**
  - Throughout this project, our group utilized various Python tools, including Pandas, NumPy, Matplotlib, Seaborn, Plotly Express, and Jupyter Notebook, to manipulate, analyze, and visualize the data effectively. We tackled diverse questions, from identifying tempo ranges with the highest danceability to exploring correlations between song attributes like energy, loudness, and popularity. Each member brought their unique perspective and technical skills to the table, resulting in a comprehensive dataset analysis.
- **Challenges Encountered:**
  - As a group, we encountered several challenges, including sourcing a suitable dataset that provided enough detail for meaningful analysis and handling inconsistencies in the data, such as missing or incorrect values. We also needed help formatting data for specific visualizations and ensuring our analyses accurately reflected the insights we aimed to uncover. Each member overcame these challenges by collaborating, sharing ideas, and leveraging different Python tools to achieve our goals.
- **Learning Experiences:**
  - This project provided valuable learning experiences for each member. We all gained a deeper understanding of Python's data analysis and visualization capabilities, particularly in using Pandas, Matplotlib, Seaborn, and Plotly Express. The collaborative nature of the project allowed us to share knowledge, solve problems together, and learn from each other's approaches. We also improved our skills in data manipulation, graph formatting, and interpreting complex datasets.
- **Future Directions:**

- Moving forward, we would consider expanding our analyses to include more song attributes and exploring additional visualization tools like Bokeh or Dash for more interactive and dynamic presentations. We also recognized the potential of incorporating machine learning techniques to predict song popularity and track trends over time. Finally, enhancing our ability to work with larger and more complex datasets would allow us to conduct more in-depth analyses and uncover more nuanced insights into the music industry.

- In conclusion, this project was an enriching experience that allowed us to apply our Python skills to a real-world dataset, work collaboratively to solve challenges, and gain valuable insights into the factors contributing to a song's success on Spotify. We are grateful for the opportunity to work together as a team and look forward to applying what we've learned to future projects.

Summary of Individual Accomplishments and Learning Experience

- **Erick:**
  - I used Python to explore the relationships between various song attributes in this project, focusing on tempo, danceability, and popularity. The three questions I chose were:
    1. What tempo range has the highest danceability score?
    2. Who are the top 5 artists, and what is their most popular song?
    3. What is the relationship between tempo and popularity?

    Python Tools Used:

    1. **Pandas**: Utilized for data manipulation, cleaning, and analysis. It helped load the data, handle duplicates, group, and calculate averages.
    2. **NumPy**: Used for numerical operations, specifically to create the bins for the tempo ranges.
    3. **Matplotlib**: Employed for essential data visualization, creating bar and scatter plots to visualize the relationship between variables.
    4. **Seaborn**: An extension of matplotlib for more advanced and aesthetically pleasing visualizations like bar and regression plots.
    5. **Plotly Express**: A powerful tool for creating interactive plots, which was used to visualize the top 5 artists and their most popular songs and the relationship between tempo and popularity.
    6. **Jupyter Notebook**: Served as the platform for writing and executing code, documenting the analysis process, and displaying visualizations in an organized and interactive environment.

    Challenges Encountered:

    One of the primary challenges was ensuring that the data visualizations accurately reflected the analysis, mainly when working with the order of categories in the Plotly visualizations. Ensuring that the bar chart correctly ordered the artists from most popular to least popular required additional steps to categorize the data correctly. Additionally, handling and visualizing large datasets with multiple attributes posed a challenge in maintaining the clarity and interpretability of the graphs.

What Would I Have Done Differently?

If I were to approach this project again, I would consider preprocessing the data more thoroughly, particularly in handling missing values or outliers that could skew the analysis. I would also explore using other visualization libraries like Bokeh for more interactive plots. Additionally, incorporating machine learning techniques from Scikit-learn, such as clustering or classification, could provide deeper insights into the data.

What Would I Do Next?

I would expand the analysis to include more song attributes, such as acousticness, liveness, and speechiness, to see how they interact with popularity and other vital attributes. I would also consider conducting a time series analysis to explore how the popularity of specific song attributes has evolved over different decades. Additionally, implementing predictive modeling to forecast song popularity based on its attributes could be a valuable next step. Finally, I would explore deploying these visualizations to a web application using frameworks like Dash or Streamlit for broader accessibility and interaction.

- **Mariyah:**
  - For this project, my group and I examined a dataset sourced from a CSV file, which included columns such as artist, song title, duration, explicit content, year, popularity, danceability, energy, key, loudness, mode, speechiness, acoustics, instrumentals, liveness, valence, tempo, and genre. Using the data, we reached various conclusions.
  - I created a bar chart with Matplotlib to track the number of songs published each year between 2000 and 2019 that start with the letter "I." The results showed that 2002 had the highest number of such songs, followed by a noticeable decline in recent years, possibly reflecting shifts in music trends or songwriting practices. Second, I used Matplotlib and Seaborn to investigate whether a longer track duration correlates with popularity. Still, the scatter plot revealed no clear relationship, indicating that other factors might be more influential in determining a song's popularity. Finally, I examined the correlation between loudness and energy levels using a hexbin plot, demonstrating a positive correlation, especially within the loudness range of -7.5 to -5.0 dB, where the highest energy levels were observed. This suggests that louder tracks are often perceived as more energetic due to production techniques designed to enhance their impact.
  - Python tools:

    In my project, I utilized several Python tools to analyze and visualize song data effectively. Pandas were employed for data manipulation, filtering songs, and counting occurrences by year, allowing for detailed dataset exploration. Matplotlib was used to create various visualizations, including bar charts and hexbin plots, which helped uncover patterns and relationships within the data. Seaborn enhanced these visualizations, mainly through aesthetically pleasing scatter plots. Although not explicitly mentioned, SciPy is often involved in statistical analysis, such as calculating correlation coefficients and adding depth to the insights. Additionally, Jupyter Notebook served as a platform for writing and running the code, organizing the workflow, and interactively displaying the visualizations. Together, these tools facilitated a comprehensive analysis of the relationships between different song attributes.

  - Challenges Encountered:
    One of the primary challenges encountered in this project was finding a suitable CSV file with enough detailed information to conduct a thorough analysis. The dataset needed various attributes, such as song titles, publication years, and musical features, to allow for meaningful exploration of trends and correlations. Sourcing a CSV that met these criteria took considerable effort, as many datasets needed more depth of information or the specific variables required for the analyses. Finding a suitable dataset was crucial to ensuring the project's success.

- **Brandon:**
  - The three questions I chose to answer were: what years had the most top songs, what ratio of top songs contained explicit content vs. non-explicit, and what was the average duration of top songs? I created a bar chart representing the top 10 years by popularity and a scatter plot representing the average duration of the top songs. I analyzed explicit content by creating the pie chart visualization. Through this project, I gained experience using Seaborn and Matplotlib for data visualization, enhanced my skills in using Plotly to create interactive visualizations, and improved my understanding of Python functions and data manipulation using Pandas. I used six Python tools: pandas for data manipulation, numpy for numerical operations, plotly for creating visualizations, seaborn for creating attractive statistical graphics, matplotlib for plotting data, and converted milliseconds to minutes and seconds with a Python function. Challenges encountered included managing missing or incorrect values and variable-related inconsistencies in my code and creating visualizations that effectively communicated insights. In the future, I aim to incorporate additional APIs or data sources to enhance the analysis and apply machine-learning techniques to predict future trends and understand the factors driving song popularity. My next steps involve formulating more complex and specific questions to answer, aiming to quantify the song precisely attributes that correlate with popularity. Ideally, I would develop algorithms to track cultural or other relevant influences and extract

patterns and trends from the collected data. This analysis would help create projections to predict which types of songs might become popular.

- **Ibrahim:**
    - In completing this project, I was tasked with answering three questions: which artist has the most danceable songs, the most energetic songs that score high on acoustics (how it is labeled in the program), and the lowest valence tracks by year? The challenge in answering these questions came from formatting the graphs and determining how to choose the data. I chose data by taking the average of a statistic and then setting a lower limit to a statistic based on that. An example of this would be setting the lower limit of energy to be .79 and the lower limit of acoustincess to be .4. The average energy level of songs in the dataset was .72. In contrast, the average acousticness level was approximately .13. Given this I decided to set the lower limit for energy not much higher than the average. Then I set the lower limit for acousticness, which was much higher than the average. Doing this would allow a broader range of energetic songs to pull from before filtering them based on acousticness. This was done in two lines to ensure these results. As for formatting, when using the .plot method, most of it is done automatically, but the data frame itself must be in a particular form, which is the main formatting problem. I formatted the energy and acoustics graph by forcing the index to be the song numbers and then setting the index to the song titles. The .plt method does not require such formatting, so I used it on the valence chart. It was nice to work on a project with other people. It made solving problems and finding new inspirations way easier than solo. I am genuinely thankful for working with such a great team.
- **Lawrence:**
    - For my part of the group project, I answer three questions using data visualization. The first question is, what are the top 5 artists with the most hit songs? I used a bar graph to answer this question to help visualize and appreciate the extent of disparity of the number of hit songs among the top 5 artists. The second question is, what is the percentage of hit songs per genre in relation to each other? The best graph visualization would be a pie chart to help us visualize and appreciate the popularity of each genre from 1999 through 2019. The third question is the same as the second one, focusing on 2000. I wanted to see how it compares to the data results from the second question. The best graph visualization type would be a pie chart for the aforementioned reason.
    - What challenges did you encounter?:

        Finding a suitable dataset to work on was very challenging. We had to find a dataset that is simple enough and has some feasible complexity. Working Dataframe can be confusing, especially when trying to analyze data based on some specified criteria. For example, I wanted to collect value count for each genre, but there were too many genres, so I cut down Labels in a pie chart plot using a list didn't work. I realized that I could work around it by filtering all data through a list of the specified genres that I decided on.

- Which 6 Python tools did you use?:

  The tools I used to do my part were VS studio code, anaconda python environment, pyplot of matplotlib python package, and pandas python package.

- What would you have done differently?:

  I would have tried to look for a more extensive dataset that would have allowed me to conduct more in-depth data analysis, which would have given me more flexibility and range when it comes to choosing different questions.

- What would you do next?:

  I would continue to improve my use of Python to analyze datasets and use data visualization to help achieve that. I would also try to understand how to utilize other graph types besides pie and bar.