VNUHCM - University of Science
Faculty of Information of Technology

fit@hcmus

Progress

SPOTIFY REWINDS 2024

Introduction to Data Science

Lectures: *Lê Ngọc Thành*
Lê Nhựt Nam

Students: *Đặng Châu Anh*
Nguyễn Kim Anh
Trần Dịu Huyền
Đỗ Minh Huy

Class: *22KHDL1*

HO CHI MINH CITY, NOVEMBER 2024

Contents

I. Project milestones

1. List of completed tasks:

No.	Phase	Detail
1	Data Collection	Determined project purposes, objectives and data sources to crawl data.
		Successfully set up a process to scrape Spotify's weekly top 200 music charts for Vietnam by using these frameworks: ✓ `selenium` web scraping ✓ `BeautifulSoup` for HTML parsing ✓ `pandas` for data manipulation ✓ `time` for handling delays ✓ `os` for file system operations
		Crawled data of 10 months (January to October) in 2024 and saved the data into CSV files.
		Verified the dataset to ensure it meets the project requirements.
2	Data exploration	Read and combine the data from multiple CSV files.
		Explored the dataset's structure and quality: ✓ Checked the number of rows and columns to confirm the dataset's structure. ✓ Reviewed the data types ✓ Provided description for each column ✓ Identify missing values and duplicated values
		Viewed summary statistics to understand basic patterns such as mean, median, min, and max values.
		Performed initial exploratory data analysis to understand the basic patterns and characteristics of the data. ✓ Distribution of numerical data ✓ Distribution of categorical data

2. List of on-going tasks:

No.	Phase	Detail
1	Data Preparation	Data Cleaning: Remove duplicates, handle missing values.
		Feature Engineering: Create new features (monthly streaming trends, streams+, number of peaks in each month).
		Data Transformation: Encode, normalize, and integrate data.
2	Visualization	Calculating and visualizing correlations ✓ Using `numpy` and `matplotlib`
		Find out the correlation and relationship of data
		Gain insight from the plots.

II. Challenges and solutions

1. Scraping data:

a. Scraping data from a dynamic website using Selenium:

- **Challenge:** The Spotify charts website used dynamic loading and rendering, requiring a more advanced scraping approach beyond standard HTML parsing.
- **Solution:** Utilized `Selenium WebDriver` to automate browser interactions, including navigating to the charts page, handling the login process, and closing the cookies popup that blocked the download button.

b. Automating the data downloads:

- **Challenge:** The weekly chart data was available at different URLs based on the date, requiring programmatic construction of the correct URLs for each month.
- **Solution:** Implemented the `calculate_week_details()` function to compute the start and end dates for the last week of each month and then used these details to build the appropriate URL for downloading the data.

To summarize, the key challenges were dealing with the dynamic nature of the website and automating the data collection process across multiple months. The solutions involved leveraging `Selenium` for browser automation and programmatically constructing the download URLs based on the weekly chart details.

2. Exploring data:

a. Overview of dataset:

- **Challenge:**
  - + The dataset may be missing some samples compared to the initial expectation.
  - + Some column names or attributes might be unclear, making it hard to understand their real-world meaning.
  - + The dataset may contain missing or erroneous values (e.g., negative values for age).
  - + The dataset may contain duplicated values.

        + Categorical attributes can have multiple values.

    - **Solution:**

        + Check the number of rows and columns of the dataset (use property shape in pandas to see the size of the dataset).

        + Use function `info()` in pandas to see all the columns name, the total elements of each column (the number of missing values of each column), and the Dtype of each column. If there are missing values, we will use methods to fill the missing values such as mean, median for numerical attributes, and mode for categorical attributes.

        + We can use function `duplicated()` from `pandas` library combined with the `sum()` function to calculate the duplicate data in the dataset. If there is duplicate data, we will keep only one unique record.

        + Use function `nunique()` from `pandas` library to determine the number of unique values of each categorical attribute. We can use encoding methods like one – hot encoding for nominal data and label encoding for ordinal data.

b. **Exploratory Data Analysis (EDA):**

- **Challenge:** The data may have many outliers.
- **Solution:** Use box plot (IQR) to determine the outlier of the dataset. And can use some method, such as: Detecting outliers using the Z-scores, Detecting outliers using the Inter Quantile Range (IQR), or we can decrease the weight of the outliers by using log transformations

c. **Quality check:**

- **Challenge:** Some values can be out of range and can be in the wrong format.
- **Solution:** We validate the values in the dataset as follows:

    + **Rank Validation:**

    - The rank should start at 1 and increase by 1.
    - Ensure the rank values are positive integers starting from 1.
    - Check for invalid ranks — they should be between 1 and the maximum possible rank.

    + **Stream Validation:**

    - Ensure that stream values are non-negative.

    + **URI Format Validation:**

    - Verify that the URI values are in the correct format.

III. Next steps

Phase	Tasks	Timeline	Deliverables
Visualization	- Calculate correlation matrix and gain insigns about the relationship in data by plotting - Ask question to get actionable insigns	17/11/2024 – 22/11/2024	Calculate correlation matrix, ask questions, and gain insights
Modeling	- Model Selection: Choose time-series models for trend analysis and clustering for artist segmentation. - Build Models: Analyze genre trends, artist rankings, and peak release periods. - Model Assessment: Validate model performance and interpret results.	23/11/2024 – 30/11/2024	Model performance report and preliminary insights.
Evaluation	- Review Results: Ensure alignment with business objectives. - Refine Models: Adjust models as needed for accuracy.	1/12/2024 – 9/12/2024	Evaluation report and key insights summary.
Deployment	- Final Report: Summarize findings with actionable recommendations. - Stakeholder Presentation: Walk through key insights and dashboard.	10/12/2024 – 22/12/2024	Final report, presentation.

IV. Question

1. **What strategy should the artist use to launch their product on Spotify in 2024 to maximize the number of streams?**

   - Purpose: Find out the top months that the artist release new song and using these information to provide insights about what strategy the artist should use to launch their product on Spotify in 2024 to maximize the number of streams.

   - Steps to answer:

   + Calculate the total number of songs released each month in 2024.
   + Create a bar chart showing the number of songs released each month in 2024.
   + Identify the months with a significantly higher number of song releases in 2024.
   + Provide insights and propose a strategy for product releases.

2. **Should the artist choose to release a personal product or a collaboration in 2024? Compare the streams between the collaboration and the single to see which one is more effective?**

   - Purpose: Compare the streams between the collaboration and the single to see which one is more effective. From that, we can provide insights about the trend of collaboration and developed trends for the artist in 2024 (working alone or collaborating with other artists).

   - Steps to answer:

   + Identify the number of songs that are collaboration and single in 2024.
   + Calculate the average number of streams for solo and collaboration in 2024.
   + Compare the average number of streams between solo and collaboration in 2024.
   + Provide insights and determine which strategy is more effective for the artist in 2024.

3. **How the artist's market share based on audience's preference in 2024, Vietnamese or international artist?**

   - Purpose: Identify the artist with the most songs on the chart, the number of songs, and whether the artist is Vietnamese or international. From that, we can provide insights on the audience's preference for Vietnamese or international artists in 2024.

   - Steps to answer:

   + Calculate the number of songs each artist has on the chart in 2024.
   + Identify the artist with the most songs on the chart in 2024.
   + Determine whether the artist is Vietnamese or international.
   + Provide insights on the audience's preference for artists in 2024.

4. **What is the trend of removing a song from the chart top 20?**

   - Purpose: Identify the trend of removing a song from the chart top 20. From that, we can provide insights on the trend of turnover a song from the chart in 2024.

   - Steps to answer:

   + Filter the DataFrame for songs that have been in the top 20.
   + Calculate the duration each song stays in the top 20.
   + Compute the average duration.
   + Analyze the trend of song turnover in the top 20.

5. **What is the average time a song rises to the top 20 of the chart wheather calculated from the release date?**

   - Purpose: Gain the average time a song rises to the top 20 of the chart from the release date. From that, we can provide insights on the time it takes for a song to rise to the top 20 of the chart in 2024.

   - Steps to answer:

   + Calculate the time each song rises to the top 20 from the release date.
   + Compute the average time.
   + Analyze the trend of the time it takes for a song to rise to the top 20.