SPOTIFY REWIND 2024

HAHA GROUP - 22KHDL1 - INTRO2DS

PRESENTED BY **CHAU ANH DANG**
KIM ANH NGUYEN
DIU HUYEN TRAN
MINH HUY DO

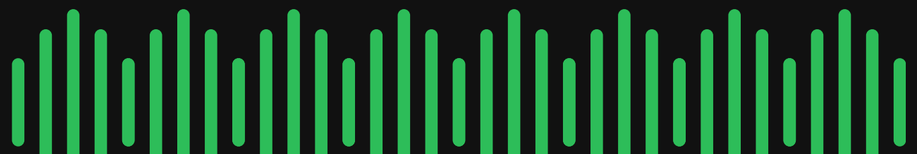


TABLE OF CONTENTS

TABLE OF CONTENTS

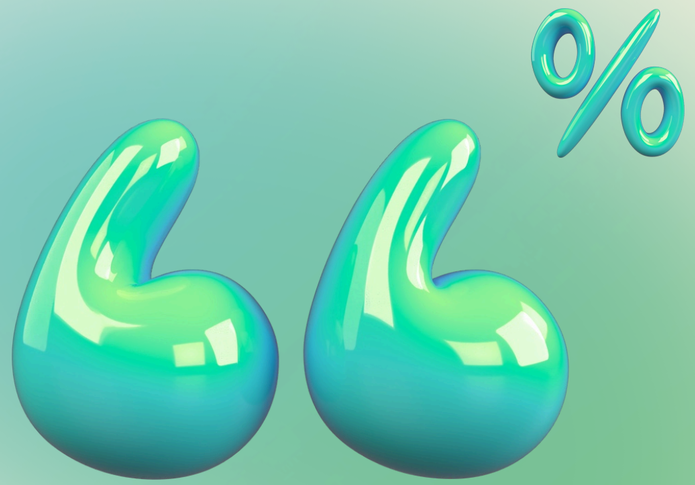
PROJECT MILESTONES

OF OUR PROGRESS

COMPLETED TASKS

ON-GOING TASKS

COMPLETED TASKS



DATA COLLECTION

Determined project purposes, objectives and data sources.

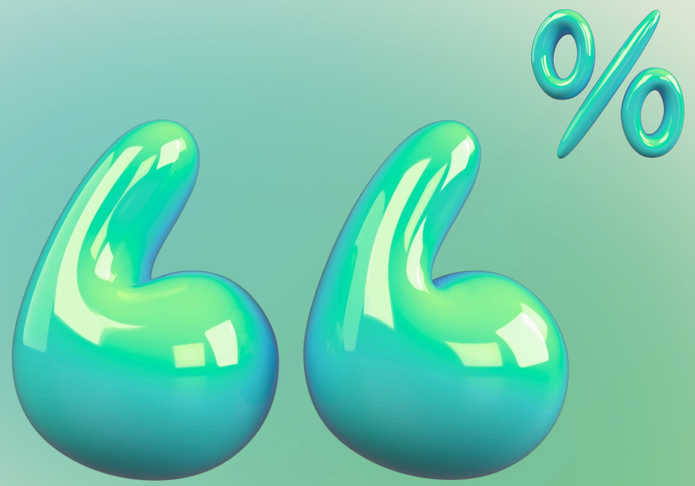
Scraped Spotify's weekly top 200 music charts for *Vietnam* (Jan–Oct 2024).

Utilized frameworks: *Selenium*, *BeautifulSoup*, *pandas*, and others.

Saved and **combined** data into unified CSV files.

Verified dataset accuracy to meet project requirements.

COMPLETED TASKS



DATA EXPLORATION

Reviewed dataset structure, row/column counts, and data types.

Checked for missing or duplicated values.

Generated summary statistics (mean, min, max).

Explored distributions of numerical and categorical data.

ON-GOING TASKS

DATA PREPARATION

Data Cleaning:

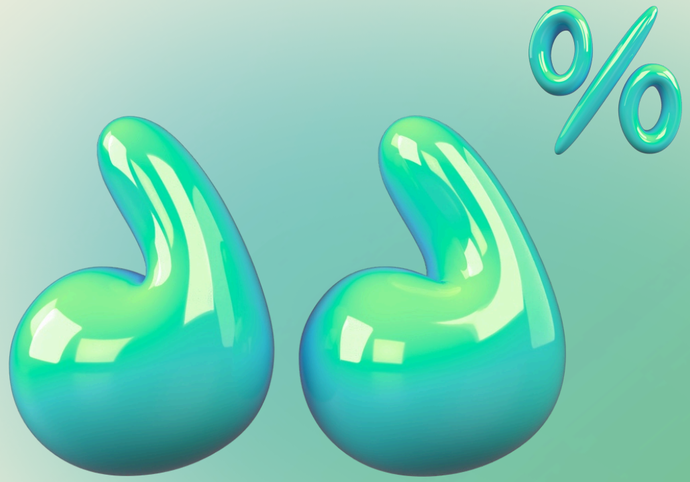
- Remove duplicates.
- Handle missing values.

Feature Engineering:

- Create new features such as:
  - Monthly streaming trends.
  - Streams+.
  - Number of peaks in each month.

Data Transformation:

- Encode categorical data.
- Normalize numerical data.
- Integrate various datasets.



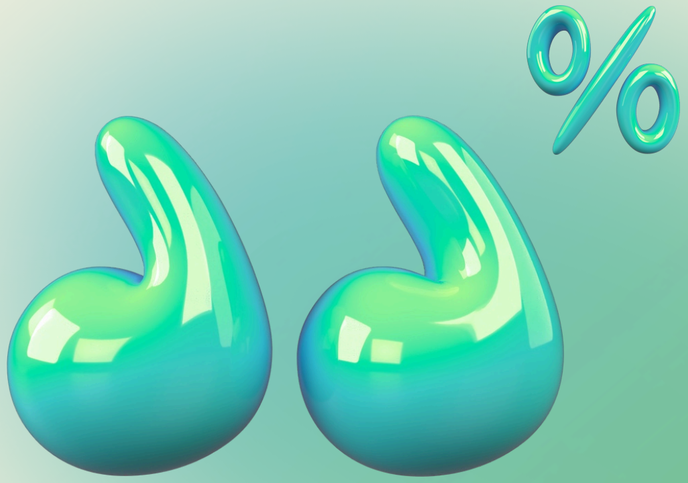
ON-GOING TASKS

VISUALIZATION

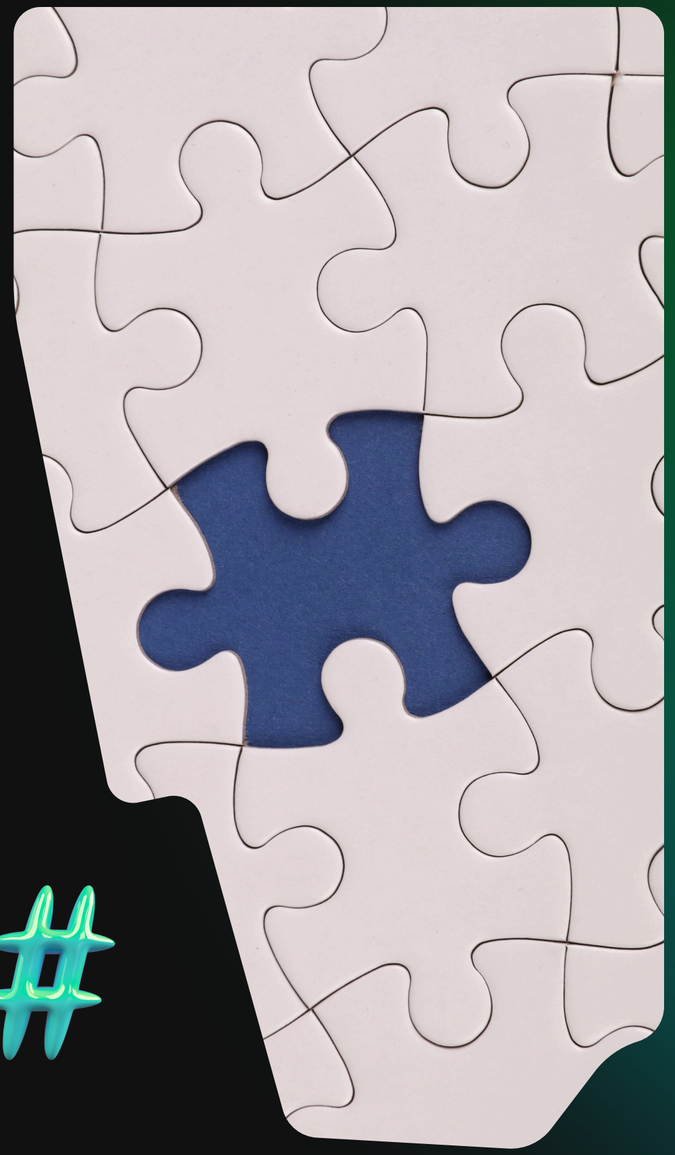
Calculating and **visualizing** correlations using *NumPy* and *Matplotlib*.

Analyzing relationships between different data attributes.

Extracting insights from visual plots.



CHALLENGES & SOLUTIONS



SCRAPING DATA

Scraping from a dynamic website using Selenium

Automating the data downloads

EXPLORING DATA

Overview of dataset

Exploratory Data Analysis (EDA)

Quality check

01

Dynamic Website Scraping

Spotify charts used dynamic loading and rendering.

Standard HTML parsing was insufficient.

02

Automating Data Downloads

Weekly chart data had unique URLs based on the date.

Required programmatic construction of URLs.

SCRAPING DATA (CHALLENGES)

01

Dynamic Website Scraping

Used *Selenium WebDriver* to automate:

- Browser interactions.
- Navigation to the charts page.
- Handling login and cookie pop-ups.

02

Automating Data Downloads

Built the *calculate_week_details()* function to:

- Compute start and end dates for each month.
- Construct URLs programmatically.

SCRAPING DATA (SOLUTIONS)

EXPLORING DATA

- Missing samples compared to expectations.
- Unclear column names or attributes.
- Missing or erroneous values (e.g., negative values for age).
- Duplicated records.
- Categorical attributes with multiple inconsistent values.

OVERVIEW (CHALLENGES)

EXPLORING DATA

01

Dataset Validation

- Use *shape* to check dataset size.
- Use *info()* to inspect column names, missing values, and data types.
- Handle missing values:
  - Mean/median for numerical data.
  - Mode for categorical data.

02

Duplicate Handling

- Use *duplicated()* with *sum()* to identify duplicates.
- Retain only unique records.

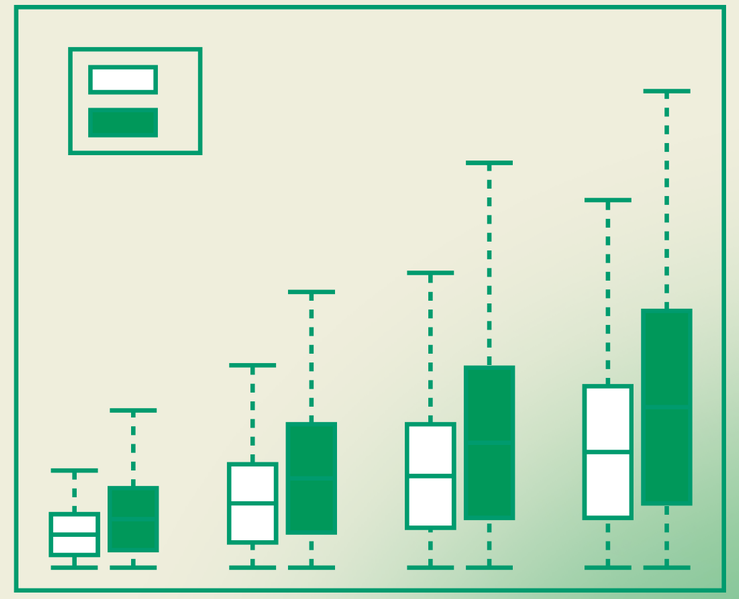
03

Categorical Data

- Use *nunique()* to find unique values.
- Apply encoding methods (one-hot for nominal, label encoding for ordinal).

OVERVIEW (SOLUTIONS)

EXPLORATORY DATA ANALYSIS (EDA)



Outliers in the data

CHALLENGES

SOLUTIONS

Use box plots (IQR) to identify outliers.

- Mitigate outliers with:
  - Z-scores.
  - IQR methods.
  - Log transformations to reduce their effect.

QUALITY CHECK

CHALLENGES

Out-of-range or **incorrect** values.

SOLUTIONS

Rank Validation:
- Ensure rank starts at 1, increments sequentially, and is positive.

Stream Validation:
- Check that stream values are non-negative.

URI Format Validation:
- Verify correct formatting of URI values.

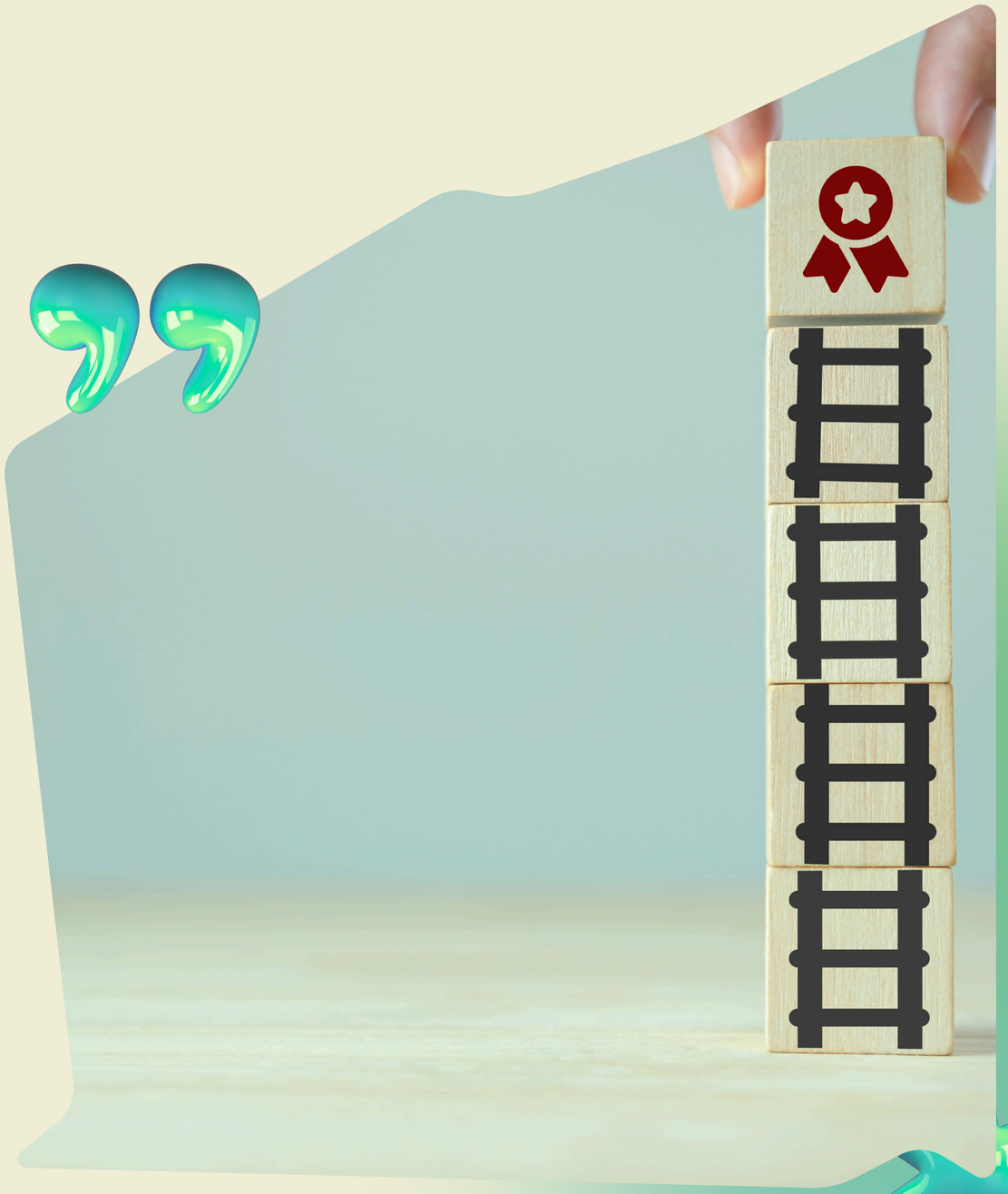
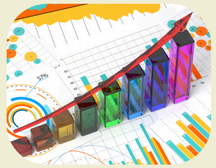
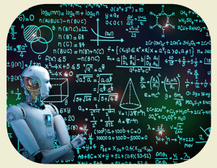


TABLE OF CONTENTS

NEXT STEPS


VISUALIZATION
17/11/2024 – 22/11/2024
MODELING
23/11/2024 – 30/11/2024
EVALUATION
1/12/2024 – 9/12/2024
DEPLOYMENT
10/12/2024 – 22/12/2024

DETAILED TASKS FOR NEXT STEPS

PHASE	TASKS	DELIVERABLES	
Visualization	• Calculate correlation matrix to uncover relationships. • Plot data and identify actionable insights. • Formulate key questions based on findings.	• Correlation matrix • Insights and relevant questions	
Modeling	• Model Selection: Time-series for trends, clustering for artist segmentation. • Build Models: Analyze genre trends, artist rankings, and peak release periods. • Model Assessment: Validate and interpret results.	• Model performance report • Preliminary insights	
Evaluation	• Review results to ensure alignment with project objectives. • Refine models for improved accuracy.	• Evaluation report • Key insights summary	
Deployment	• Summarize findings and recommendations in the final report. • Present insights through dashboards and presentations.	• Final report • Stakeholder presentation	

THANK YOU

FOR YOUR FOLLOWING