PW : Handbook Data Management

By Team 14,

Ruchi Ravindra Manjalkar
Gaurav Sandeep Naik
Ashwini Bharat Nadekar
Darpen Jatin Bhandari

Dataset Selection and Description

Brainstorming and Initial Exploration

Our journey began with an extensive brainstorming session, exploring various data sources and platforms. We visited numerous websites and forums dedicated to datasets, seeking one that aligned with our project goals.

Collaboration with Professor Theo

To refine our search and gain expert insights, we scheduled a meeting with Professor Theo. Together, we discussed our objectives, challenges, and potential datasets. After careful consideration, we initially selected a dataset with 7 columns and around 100,000 rows.

Re-evaluation and Final Selection

Upon closer inspection, we realised that the initial dataset's limited columns hindered our ability to conduct a comprehensive analysis. We revisited our approach and, with Professor Theo's guidance, ultimately chose a music dataset containing 17 columns and 17,000 rows. This dataset encompassed diverse information about songs, artists, and various features.

Team Work

Dataset Discovery

All four team members actively participated in the dataset discovery process, contributing their perspectives and insights.

Further Tasks

- **Uncleaning of Data:** Ruchi, Darpen and Gaurav
- **Cleaning of Data:** Gaurav, Ashwini and Ruchi
- **User Stories & Data Profiling:** Ashwini and Ruchi
- **Presentation :** Ruchi and Darpen
- **Handbook Creation:** Gaurav and Darpen

Data Profiling Tools and Methods

Tool Selection

We opted for Tableau Prep as our primary data cleaning tool due to its user-friendly interface and robust data manipulation capabilities. This choice empowered us to efficiently profile and clean our dataset.

Data Uncleaning

Collectively, we injected artificial noise in various forms including:

- **Missing data randomly** - Some records have missing values.
- **Injecting incorrect data types** - class, mode, loudness & energy columns were provided with intentional data type errors
- **Incorrect column names** - class was provided with incorrect column name
- **Misleading Accurate Values** - danceability, time_signature, liveness, mode, class, popularity, speechiness and acousticness where intentionally given incorrect values using formulae and random selection.
- **Consistent Data Integrity** - Some numeric-only columns such as mode,class were injected with numeric data.

This manual noise addition prompted the need for data cleaning.

Data Cleaning and Quality Improvement

Collectively, we addressed various data quality issues, including:

- **Rectifying missing values**
- **Removing unnecessary columns**
- **Removing duplicates**
- **Accuracy**
- **Preserving Data Integrity**
- **Logical Column names**

These efforts laid a strong foundation for our subsequent analysis.

User Stories & Data Profiling

Our project's user stories guided our analysis, focusing on:

- To explore the relationship between track attributes and popularity, so user can understand what makes a song popular.
- To pinpoint lively tracks for an user’s event, so user can create an upbeat playlist.

Data was profiled according to the user stories and solved visually.

Progress Tracking and Documentation

Regular meetings and continuous documentation facilitated transparent communication and collaboration across the team. Each member provided updates, ensuring everyone stayed aligned with the project's direction.

Quality Assurance

Mutual review and verification were integral to maintaining high standards in our data analysis. Rigorous quality assurance processes ensured the robustness and reliability of our findings.

Risk Management

By identifying potential risks early in the project, such as data quality challenges and time constraints, we developed effective strategies to mitigate these risks. Proactive risk management contributed to the smooth execution of the project.

Conclusion

Our collaborative efforts resulted in significant insights into various music genres. Each team member's unique contributions played a crucial role in the project's success, underscoring the power of teamwork in tackling complex data analysis projects. The journey not only provided valuable insights into music data but also showcased the effectiveness of a well-coordinated team.