1. **Data Description :**

I would like to present the dataset we have: it consists of two CSV files with the following description:

File1 : Mesures\_contenu\_volume\_audio\_à\_commander(in) **(Mesures Table)**

|  |  |
| --- | --- |
| Colonne | Description |
| Segment ID | **Unique identifier of the content segment** |
| Segment | **Title of the episode or topic** |
| Show ID | **Show identifier** |
| Show | **Name of the show** (e.g., Swiss Crimes, The Hourly News, The Good Speakers, etc.) |
| Publication Date | **Broadcast date** |
| App/Site Name | **Access platform** (rts.ch, rts-app-play, rts-app-info, etc.) |
| Device Class | **Device type** (Smartphone, PC/Laptop, etc.) |
| Segment Length | **Segment duration** (in seconds or minutes) |
| Media Views | **Number of views** |
| Avg Play Duration | **Average viewing duration** |
| Visitors | **Number of unique visitors** |
| New Visit Rate % | **Percentage of new visitors** |
| Entries | **Number of page entries** |
| Exits | **Number of exits** |
| Returning Visits | **Number of returning visits** |
| Bounces | **Number of bounces** |
| Total Play Duration | **Total play duration** (sum of viewed minutes/seconds) |

* This table contains the performance indicators associated with each segment.
* This file tells us how each content is performing.

File 2: Correspondance\_show\_segment\_tag(in) (Tags Table)

|  |  |
| --- | --- |
| **Colonne** | **Description** |
| Segment ID | **Unique identifier of the segment** (episode, excerpt, podcast, etc.) |
| Show | **Name of the show** to which the segment belongs |
| Show ID | **Internal show (program) identifier** |
| Assigned Tags | **Set of labels or “tags” describing the content** |

* This table is used to describe the metadata of each content segment (episode, topic, podcast, etc.)
* It represents *what* the content is.
* The link between the two tables is: SegmentID.
* The field Segment ID is the primary (unique) key in both files.
* It’s the connection point that allows us to merge the two data sources.
* This way, you can link the metadata (themes) to their performance metrics.

The Segment ID field is the primary (unique) key in both files.  
It’s the connection point that allows you to merge the two data sources.

Example :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **Segment ID** | **Show** | **Assigned Tags** | **Media Views** | **New Visit Rate %** | **Avg Play Duration** |
| 14572281 | Crimes suisses | media\_radio:podcasts-originaux | 7327 | 2.49 % | 00:24:41 |
| 14897825 | Le Journal horaire | rts\_info:la-1ere | 20762 | 84.56 % | 00:05:19 |

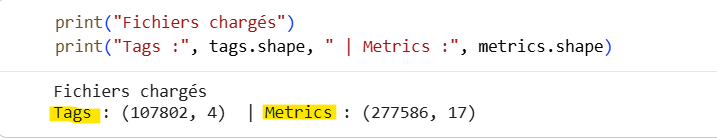
Before merging the two tables, we will first clean the data.

1. **Data Cleaning:**

For the data cleaning, I choose python (<https://colab.research.google.com/>) to manipulate the data.

So, at first, I load the CSV files that I called Measures and Tags into Pandas Data Frames.

Here is the initial data shape before cleaning:

****

**Cleaning data contains:**

1. Cleaning Column names: Removes leading/trailing spaces from column names.
2. Cleaning cell values: Removes leading/trailing spaces from column names.
3. Dropping fully empty rows: Deletes rows where all cells are empty.
4. Harmonizing identifiers: Converts Segment ID to string

Removes spaces to ensure proper matching during merge

1. Cleaning numeric and textual values: Convert "New Visit Rate %" to float

Convert time to seconds

1. Removing duplicates: Ensures each Segment ID appears only once

**Checking cleaning data:**

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

* At this stage, we can go further to understand why we have so much duplicated data and validate these changes with the business.

**Merging the two tables**:

Une image contenant texte, capture d’écran, reçu, Police

Le contenu généré par l’IA peut être incorrect.

Then the new shape of new table merged with a left join:

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

1. **Data Analysis :**

So, at first I want to use an LDA (Latent Dirichlet Allocation) model to automatically discover topics from the Assigned Tags, and then compare or align these topics with the existing categories :

* + media\_radio:societe
  + media\_radio:humour
  + media\_radio:info
  + media\_radio:musique
  + media\_radio:sport

The goal of using an LDA model is to:

* Clean all the tags (removing 'media\_radio', 'la-1ere', etc.).
* Transform the text into a word matrix using CountVectorizer.
* Train an LDA model to detect the dominant topics.
* Display the most representative words for each topic.
* Automatically map the documents to your main categories.

Here is the result :

Une image contenant texte, Police, papier, noir et blanc

Le contenu généré par l’IA peut être incorrect.

Now, we can construct our mapping:

theme\_keywords = {

    'Info': ['info', 'reportages', 'news', 'economie' ,'monde'],

    'Sport': ['sport', 'match', 'football', 'basket', 'tennis', 'rugby'],

    'Musique': ['musique', 'concert', 'chanson', 'album', 'pop', 'rock', 'classique'],

    'Société': ['societe', 'entretiens', 'social', 'documentaire social', 'culture'],

    'Humour': ['humour', 'comedy', 'blague', 'sketch', 'stand-up']

}

And we obtain the final tags:

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

**4.Metrics Définition :**

To recapitulate, tags are now cleaned, standardized and grouped by theme, duration are converted to seconds and unclassified segments remain marked as ‘other’ for the future review.

Now, we define the metrics aggregation. For each theme, we calculated:

* Number of segments **(Number)**
* Total views (Media Views) **(SUM)**
* Number of visitors (Visitors) **(SUM)**
* Returning visits (Returning Visits) **(SUM)**
* New visitor rate (New Visit Rate %) **(SUM)**
* Total and average playtime **(SUM, Average)**
* Bounces **(SUM)**

**5.Definition of Strategic Scores**

|  |  |  |
| --- | --- | --- |
| **Concept** | **Metrics** | **Purpose** |
| Acquisition | Visitors \* New Visit Rate % | Measures the ability to attract new visitors |
| Retention | Returning Visits | Measures the retention of existing visitors |
| Engagement | Total Play Duration | Measures overall interest in the content |

**6. Prioritization**

Scores are normalized for comparison:

* Acquisition (Visitors × New Visit Rate %) and Retention (Returning Visits) are measured in different units and can have very different ranges.
* Normalizing (Acquisition\_norm and Retention\_norm) scales both scores to a comparable range (typically 0–1).
* This ensures that one metric doesn’t dominate the other just because of its scale

Priority Score calculation: Priority Score = 0.6 \* Acquisition\_norm + 0.4 \* Retention\_norm :

* Acquisition is given a weight of 0.6, meaning attracting new visitors is slightly more important for priority decisions.
* Retention is weighted at 0.4, reflecting that keeping existing visitors is important but slightly less critical for this prioritization.
* The weights sum to 1 to make it a weighted average.

Themes are ranked by Priority Score to guide content production:

By combining both normalized metrics with specific weights, the Priority Score reflects the overall strategic value of a theme:

* Themes with high acquisition potential and reasonable retention rank higher.
* Themes that are strong in retention only may rank slightly lower unless they also have good acquisition.
* This allows teams to allocate production resources effectively, focusing on themes that will attract and retain the largest audience.

**7. Visualization**

Two visualizations are generated:

* Matplotlib / Seaborn line chart for visitor evolution
* Dashboard using Power BI for dynamic exploration

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

**8. Result interpretation:**

a. Regarding the graph of total audience by theme, we notice that visitors in 2024 are attracted by news more than music, society or sport and this maybe is explained by current political events (war, conflicts, crises…), also news is updated continuously and explain their behaviour to prioritize staying informed particularly with with significant global or local events.

b. Regarding the graph of the proportion of loyal visitors, I notice that returning visits equal or exceed visitors across all themes. This suggests visitors tend to come back multiple times within the measurement period, indicating strong loyalty or repeated engagement.

c. Regarding the graph of average engagement by theme, I notice that Music clearly drives the deepest engagement, with users spending the longest time on average per session or piece of content. Société and Humour also show solid engagement, keeping users interested for a significant amount of time. Info and Sport have lower average engagement times, possibly due to the typical format or consumption habits of their content.

d. Regarding the scatter plot of Acquisition Score vs Engagement Score by theme, with the size of each point representing the total number of visitors, is a great way to visualize how themes perform in attracting new users (acquisition) and keeping them engaged, while also reflecting their audience size. So, I notice that : Info: Likely a large-sized point (highest visitors), moderate acquisition and engagement. Musique: Moderate to high acquisition and very high engagement, medium audience size. Société: Smaller audience, moderate scores. Humour: Moderate audience with lower acquisition but high engagement. Sport: Smallest audience, likely lower scores overall

e. Regarding the graph of Priority score by theme, I notice that Info has the highest Priority Score, indicating it’s the top theme to focus on, as it combines strong acquisition, retention, and engagement with the largest audience. Musique is the second priority, showing solid engagement despite a smaller audience than Info. Société comes next, with moderate scores. Humour and Sport have the lowest Priority Scores, suggesting they are lower priorities in the current context, possibly due to smaller audiences or lower combined engagement and acquisition

**9. Strategic Recommendations**

* Info shows the higest priority score, large audience and strong retention so continue investing in timely and high quality new content
* Music also drives the best engagement and the decent acquisition so consider expanding music offerings
* Develop targeted marketing campaigns to increase new visitor especially for sport
* Explore ways to increase visitor loyalty and engagement through fresh content for sport and humour
* Use the Priority Score as a dynamic guide to reallocate resources as audience behavior evolves