Analyse de l’industrie de la K-pop à travers les données

Module Big Data encadré par BERGÈRE, Alexandre

Rendu Final – Analyse de la K-pop à travers les données

1. Introduction 2
   1.1 Objectif principal 2
   1.2 Problématique 2
   1.3 Méthodologie 2
2. Présentation du projet 2
   2.1 Description générale 2
   2.2 Outils utilisés 3
3. Sources et Préparation des Données 9
   3.1 Source de données 9
   3.2 Préparation des Données 11
4. Analyse et Visualisation des Données 25
   4.1 Présentation des Onglets et Visualisations 26
      4.1.1 Analyse Temporelle des Performances 26
      4.1.2 Comparaison des Groupes et Genres 27
      4.1.3 Analyse de la Longévité des Groupes K-pop 29
      4.1.4 Analyse Géographique des Idols 30
5. Interprétation des Résultats 31
   5.1 Facteurs de Popularité 31
   5.2 Comparaison des Groupes et Genres 33
   5.3 Analyse de la Longévité des Groupes K-pop 34
   5.4 Analyse Géographique des Idols 35
   5.5 Conclusion Globale 35
6. Difficultés rencontrées 36
   6.1 Problèmes rencontrés 36
   6.2 Solution temporaire 37
7. Prise de recul et axes d'amélioration 38

7.1 Compétences acquises et technologies apprises 39
Conclusion 41

1. Introduction

1.1 Objectif principal

Réaliser une analyse approfondie de l'industrie de la K-pop en exploitant les données relatives aux groupes, aux idols et aux vidéos musicales (MV - Music Videos).

1.2 Problématique

Quels sont les facteurs clés qui influencent la **popularité** et la **longévité** des groupes et idols dans l'industrie de la K-pop ?

1.3 Méthodologie

La démarche s'appuie sur l'utilisation de plusieurs datasets, transformés selon une architecture ***Bronze*** → ***Silver*** → ***Gold***. Cette structuration permet de nettoyer, préparer et enrichir les données afin de faciliter leur analyse et leur visualisation à l'aide d'outils dédiés.

2. Présentation du projet

2.1 Description générale

Ce projet s'appuie sur diverses données collectées au sujet de l'industrie de la K-pop, couvrant plusieurs années. Ces données concernent les groupes, les idols et les MV (Music Videos), permettant ainsi une analyse approfondie de cet univers dynamique. L'objectif est de mieux comprendre les mécanismes qui structurent cette industrie, ses facteurs d'équilibre et ses tendances d'évolution.

2.2 Outils utilisés

- **Langage Python :**

Utilisation :
Langage principal pour la préparation et la transformation des données.

Où ? :
Étapes **Bronze** → **Silver** → **Gold**.

Pourquoi ? :
Automatisation des processus d'extraction, de nettoyage et de transformation des données pour garantir leur qualité.

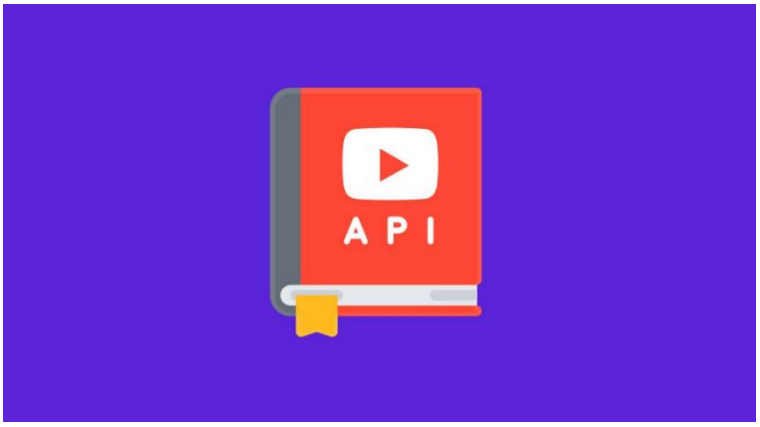
- **Pandas** (Bibliothèque Python) :

Utilisation :
Manipulation des données sous forme de DataFrames.

Où ? :
Nettoyage des fichiers CSV, détection des valeurs manquantes, remplacement des séparateurs décimaux (ex: `.` → `,`) et création de nouvelles colonnes calculées (ex: ratio_engagement).

Pourquoi ? :
Facilite l'organisation et la transformation des données complexes.

- **API YouTube Data v3**



Utilisation :
Enrichissement des données des MV (Music Videos) avec des informations telles que le nombre de vues, likes et commentaires.

Où ? :
Récupération des métriques des vidéos à partir des liens présents dans les datasets.

Pourquoi ? :
Obtenir des données dynamiques et à jour directement depuis YouTube.

- **Power BI**

Power BI

Utilisation :
Visualisation des données enrichies sous forme de graphiques interactifs.

Où ? :
Création de rapports détaillés pour analyser les vues, les ratios d'engagement, les comparaisons entre genres (Boy Group vs Girl Group), ainsi que la longévité des groupes/idols.

Pourquoi ? :
Permet une interprétation claire et visuelle des tendances et des résultats issus de l'analyse.

- **Kaggle**

kaggle

Utilisation :
Plateforme pour la recherche et la récupération des datasets.

Où ? : Source principale pour collecter les fichiers CSV sur les groupes de K-pop, leurs idols, et les vidéos.

Pourquoi ? :
Fournit des datasets bien structurés et accessibles pour démarrer rapidement un projet d'analyse de données.

- **CSV (Comma-Separated Values)**

Utilisation :
Format de stockage des données brutes et transformées.

Où ? :
Les fichiers ***Bronze*** contiennent les données brutes, les fichiers ***Silver*** les données nettoyées, et les fichiers ***Gold*** les données enrichies prêtes à l'analyse.

Pourquoi ? :
Format simple, flexible et facilement manipulable.

- **Azure Blob Storage**

Utilisation :
Solution de stockage pour héberger les fichiers de données brutes, nettoyées et enrichies dans les différents conteneurs (***Bronze***, ***Silver***, ***Gold***).

Où ? :

- ***Bronze*** : Données brutes dans des fichiers CSV.
- ***Silver*** : Données nettoyées et enrichies en format Delta.
- ***Gold*** : Dimensions, métriques, et table des faits en format Delta.

Pourquoi ? :
Stockage sécurisé et évolutif. Intégration avec d'autres services Azure (comme Azure Synapse, Databricks, ou Power BI). Compatibilité avec divers formats de fichiers et protocoles, notamment abfss:// pour Azure Data Lake Storage Gen2.

- **Delta (Delta Lake)**

Utilisation :
Les données en format Delta sont utilisées dans les conteneurs ***Silver*** (données nettoyées et transformées) et ***Gold*** (données enrichies et prêtes pour l'analyse).

Où ? :
Les fichiers ***Bronze*** contiennent les données brutes, les fichiers ***Silver*** les données nettoyées, et les fichiers ***Gold*** les données enrichies prêtes à l'analyse.

Pourquoi ? :
- Supporte les transactions ACID, assurant la cohérence des données même en cas d'échecs.
- Permet de gérer les versions des données (time travel) pour auditer ou restaurer des états passés.
- Optimise les performances pour les charges de travail analytiques grâce au format Parquet sous-jacent et aux indices générés.
- Idéal pour des pipelines de données scalables et fiables.

- **GitHub**

Utilisation :
Gestion des versions du projet et collaboration.

Où ? :
Suivi de l'évolution du code Python et des étapes du projet.

Pourquoi ? :
Assure la sauvegarde et la traçabilité des modifications apportées au projet.

- **Visual Studio Code (VSCode)**

Utilisation :
Environnement de développement intégré (IDE) pour écrire et exécuter les scripts Python.

Où ? :
Étapes de codage, tests et débogage des scripts de transformation et d'enrichissement des données.

Pourquoi ? :
Facilite le développement avec des extensions adaptées (Python, Git, etc.).

3. Sources et Préparation des Données

3.1 Source de données

Les datasets bruts collectés constituent la base de ce projet et couvrent plusieurs aspects de l'industrie K-pop. Voici une description de chaque fichier utilisé :

- ***kpop_idols.csv***

  Contenu :

Informations générales sur les idols, incluant leur nom de scène, nom complet, date de naissance, pays d'origine, et leur appartenance à un groupe.

Colonnes principales :
Stage Name, Full Name, Date of Birth, Group, Country, Gender.

- ***kpop_idols_boy_groups.csv***

  Contenu :
  Liste des boy groups, avec des détails sur leur nom, l'année de début, le nombre de membres, et leur fanclub.

  Colonnes principales :
  Name, Debut, Members, Fanclub Name, Active.

- ***kpop_idols_girl_groups.csv***

  Contenu :
  Similaire au dataset des boy groups, mais pour les girl groups.

  Colonnes principales :
  Name, Debut, Members, Fanclub Name, Active.

- ***kpop_music_videos.csv***

  Contenu :
  Informations sur les vidéos musicales, incluant l'artiste, le nom de la chanson, la date de sortie, ainsi que des liens YouTube.

  Colonnes principales :
  Artist, Song Name, Video, Release, Director.

3.2 Préparation des Données

Transition de Bronze à Silver

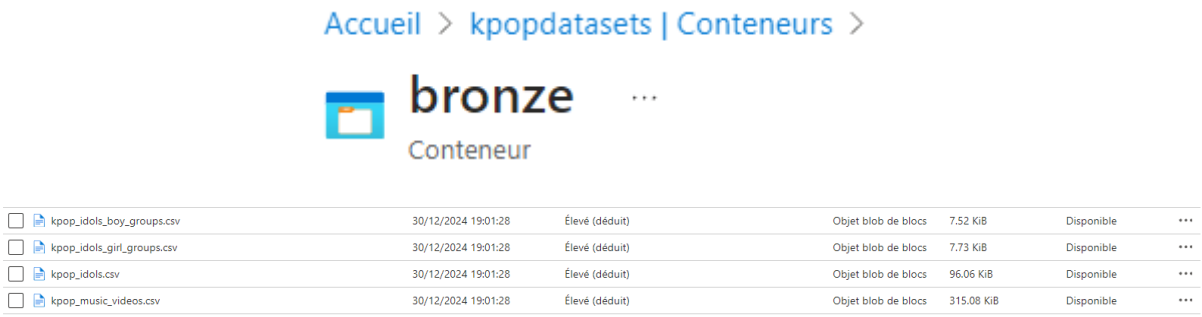
Objectif de cette étape :

Cette étape vise à charger, vérifier et nettoyer les données brutes stockées dans la couche ***Bronze*** pour garantir leur qualité avant de les transformer dans la couche ***Silver***. L'accent est mis sur la suppression des doublons et l'identification des problèmes potentiels dans les fichiers sources.

Étapes clés effectuées dans le script :

A. Chargement des fichiers depuis la couche Bronze :

- Les fichiers bruts sont situés dans le conteneur ***bronze*** au sein de notre compte de stockage Azure (kpopdatasets).



- Une boucle parcourt tous les fichiers listés (kpop_idols.csv, kpop_idols_boy_groups.csv, kpop_idols_girl_groups.csv, kpop_music_videos.csv), elle va les charger, nettoyer les doublons et charger les fichiers propres dans le conteneur ***silver*** dans Azure.

```python
# Itérer sur tous les fichiers
for file_name in FILES:
    try:
        # Charger et vérifier le fichier
        df_cleaned = load_and_check_duplicates(file_name)

        # Enregistrer les données nettoyées dans le conteneur Silver
        save_csv_to_blob(SILVER_CONTAINER, file_name, df_cleaned)
        print(f"\nDonnées nettoyées enregistrées dans le conteneur Silver : {file_name}")
    except Exception as e:
        print(f"Erreur lors du traitement de {file_name} : {e}")
```

Vérification et exploration des données :

- Affiche un aperçu des données (les premières lignes).
- Génère un résumé des colonnes pour chaque fichier (df.info()), incluant :
  - Les types de données (texte, numérique, etc.).
  - Les colonnes avec des valeurs manquantes.
- Affiche les statistiques descriptives pour analyser les distributions.

Exemple de réponse à la suite du traitement, montre les types de colonnes et les valeurs statistiques.

```
--- Chargement du fichier : ../data/bronze\kpop_idols.csv ---

Aperçu des données de kpop_idols.csv :
  Stage Name      Full Name Korean Name K. Stage Name Date of Birth      Group      Country Birthplace Other Group Gender
0        A.M  Seong Hyunwoo         성현우           에이엠    1996-12-31  Limitless  South Korea        NaN         NaN      M
1        Ace  Jang Wooyoung         장우영           에이스    1992-08-28        VAV  South Korea        NaN         NaN      M
2      Aeji      Kwon Aeji         권애지            애지    1999-10-25    Hashtag  South Korea      Daegu         NaN      F
3      AhIn       Lee Ahin         이아인            아인    1999-09-27   MOMOLAND  South Korea      Wonju         NaN      F
4      Ahra        Go Ahra         고아라            아라    2001-02-21   Favorite  South Korea      Yeosu         NaN      F

Résumé des données :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1310 entries, 0 to 1309
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Stage Name     1310 non-null   object
 1   Full Name      1304 non-null   object
 2   Korean Name    1304 non-null   object
 3   K. Stage Name  1309 non-null   object
 4   Date of Birth  1310 non-null   object
 5   Group          1219 non-null   object
 6   Country        1310 non-null   object
 7   Birthplace     689 non-null    object
 8   Other Group    122 non-null    object
 9   Gender         1310 non-null   object
dtypes: object(10)
memory usage: 102.5+ KB
None

Statistiques descriptives :
       Stage Name    Full Name Korean Name K. Stage Name Date of Birth Group      Country Birthplace     Other Group Gender
count        1310         1304        1304          1309          1310  1219         1310        689            122   1310
unique       1135         1251        1241          1082          1181   208           12        135             55      2
top         Jisoo  Lee Minhyuk         이서영            유진    2001-11-13   NCT  South Korea      Seoul  Super Junior-M      M
freq            4            3           3             5             3    18         1204        188              6    676

Nombre total de doublons dans kpop_idols.csv : 0

Taille après suppression des doublons : (1310, 10)

Données nettoyées enregistrées dans : ../data/silver\kpop_idols.csv
```

B. Détection et suppression des doublons :

- Identifie le nombre total de doublons (df.duplicated().sum()).
- Supprime les doublons en conservant uniquement les lignes uniques (df.drop_duplicates()).
- Affiche un exemple des doublons détectés.

Extrait du script détectant et supprimant les doublons

```python
# Vérification des doublons
total_duplicates = df.duplicated().sum()
print(f"\nNombre total de doublons dans {blob_name} : {total_duplicates}")

if total_duplicates > 0:
    print("\nExemples de doublons :")
    print(df[df.duplicated()].head())

# Nettoyage des doublons
df_cleaned = df.drop_duplicates()
print(f"\nTaille après suppression des doublons : {df_cleaned.shape}")
```

Affichage en console en cas de doublons

```
Nombre total de doublons dans kpop_music_videos.csv : 2

Exemples de doublons :
            Date  Artist   Song Name Korean Name Director                        Video Type   Release
504   2019-04-21  Target   Beautiful        아름다워      NaN  https://youtu.be/CdlqPQY7LIM  Boy     Major
1610  2017-01-17    100%  How To Cry         NaN      NaN  https://youtu.be/cjKtjg9Ia68  Boy  Japanese
```

kpop_music_videos.csv* -> *Bronze

```
1611    2017-01-17,100%,How To Cry,,,https://youtu.be/cjKtjg9Ia68,Boy,Japanese
1612    2017-01-17,100%,How To Cry,,,https://youtu.be/cjKtjg9Ia68,Boy,Japanese
```

```
505    2019-04-21,Target,Beautiful,아름다워,,https://youtu.be/CdlqPQY7LIM,Boy,Major
506    2019-04-21,Target,Beautiful,아름다워,,https://youtu.be/CdlqPQY7LIM,Boy,Major
```

kpop_music_videos.csv* -> *Silver

```
1609    2017-01-18,Kasper,Lean On Me,린온미,,https://youtu.be/-TqPIirTV-s,Girl Solo,Major
1610    2017-01-17,100%,How To Cry,,,https://youtu.be/cjKtjg9Ia68,Boy,Japanese
1611    2017-01-17,CLC,Hobgoblin,도깨비,Vikings League,https://youtu.be/u9OxRFab6o4,Girl,Major
```

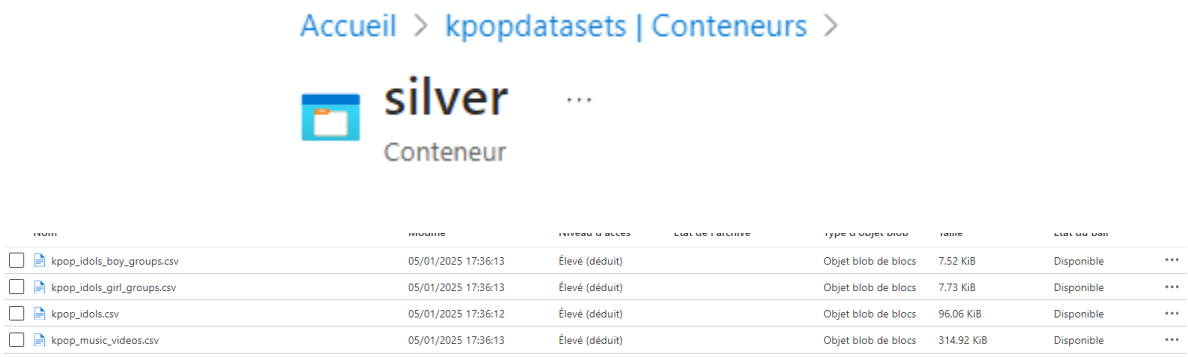
```
504    2019-04-22,Pentagon,Genius (feat. Pentagon's Fathers),,,https://youtu.be/BCeeJAxUoos,Boy,Minor
505    2019-04-21,Target,Beautiful,아름다워,,https://youtu.be/CdlqPQY7LIM,Boy,Major
506    2019-04-19,ENOi,Bloom,,,https://youtu.be/OyAIt76cx1o,Boy,Major
```

C. Export des données nettoyées vers la couche Silver :

- Chaque fichier nettoyé est sauvegardé dans le conteneur ***silver***.

- La structure des fichiers est préservée, et les données sont prêtes pour les étapes suivantes.

*Répertoire **Silver** après exportation des fichiers nettoyés.*



Résumé des transformations entre *Bronze* et *Silver* :

Action	Détail
Chargement	Lecture des fichiers CSV bruts dans le conteneur ***Bronze*** du compte de stockage Azure.
Vérification	Identification des doublons, valeurs manquantes, et types de colonnes.
Nettoyage	Suppression des doublons.
Export vers Silver	Sauvegarde des fichiers nettoyés dans le conteneur ***Silver***.

Transition de *kpop_music_videos.csv* à *kpop_music_videos_enriched.csv*

Objectif de cette étape :

Cette phase vise à enrichir les données de vidéos musicales avec des métriques essentielles (vues, likes, commentaires) récupérées dynamiquement via l'API YouTube, et à préparer ces données pour une analyse approfondie dans la couche ***Gold***.

Étapes clés effectuées dans le script :

A. Chargement des vidéos depuis la couche *Silver*

- Le fichier ***kpop_music_videos.csv*** (précédemment nettoyé dans la couche ***Silver***) est chargé pour être enrichi avec des métriques supplémentaires.

B. Extraction des identifiants des vidéos YouTube

- L'URL de chaque vidéo dans la colonne Video est analysée pour extraire l'ID unique YouTube, essentiel pour interroger l'API.

- La fonction extract_video_id gère les deux formats d'URL (standard et raccourci `youtu.be`).

Exemple de traitement :

✓ URL : `https://www.youtube.com/watch?v=dQw4w9WgXcQ` → ID : `dQw4w9WgXcQ`.
✓ URL : `https://youtu.be/dQw4w9WgXcQ` → ID : `dQw4w9WgXcQ`.

C. Récupération des métriques via l'API YouTube

- La fonction get_video_metrics interroge l'API YouTube Data v3 pour chaque ID de vidéo.

- Métriques collectées :
  - **`views`** : Nombre de vues de la vidéo.
  - **`likes`** : Nombre de likes.
  - **`comments`** : Nombre de commentaires.

- En cas d'erreur (si la vidéo a été supprimée ou que l'API est hors quota), des valeurs par défaut (`None`) sont attribuées pour éviter l'interruption du script.

D. Intégration des métriques dans le DataFrame

Les métriques récupérées sont intégrées dans le DataFrame existant (kpop_music_videos.csv) sous forme de nouvelles colonnes : views, likes, et comments.

- Validation des données :
  - Les vidéos sans ID valide ou sans données YouTube sont identifiées et peuvent être filtrées ultérieurement.

E. Sauvegarde des données enrichies dans la couche *Silver* (intermédiaire)

- Les données enrichies sont sauvegardées sous le nom ***kpop_music_videos_enriched.csv*** dans le conteneur ***Silver***.

- Ce fichier sera ensuite structuré dans la couche ***Gold***.

*Fichier initial avant enrichissement (**kpop_music_video.csv**)*

kpop_music_videos.csv

data > silver > kpop_music_videos.csv > data

```
Date,Artist,Song Name,Korean Name,Director,Video,Type,Release
2020-05-22,Agust D,Daechwita,대취타,,https://youtu.be/qGjAWJ2zWWI,Boy Solo,Major
2020-05-21,Yubin,yaya (Me Time),냅냅,,https://youtu.be/BJD0arHF_5c,Girl Solo,Major
2020-05-21,OnlyOneOf,Angel (prod. GRAY),,,https://youtu.be/vVTo8p72FxQ,Boy,Major
2020-05-20,Ryu Sujeong,Tiger Eyes,,,https://youtu.be/aE6curPGQRY,Girl Solo,Major
2020-05-20,Crush,Mayday feat. Joy (Red Velvet),자나깨나,,https://youtu.be/29ycT6fA-Rs,Boy Solo,Major
2020-05-19,NCT 127,Punch,,,https://youtu.be/U08OSl3V4po,Boy,Major
2020-05-19,SECRET NUMBER,Who Dis?,,,https://youtu.be/mkYwq_CKpyw,Girl,Major
2020-05-18,REDSQUARE,ColorFull,,,https://youtu.be/we1wqjcsYBU,Girl,Major
2020-05-18,N.Flying,Amnesia,,,https://youtu.be/UcxNeKzdDN4,Boy,Japanese
```

*Fichier final après enrichissement (**kpop_music_video_enriched.csv**)*

kpop_music_videos_enriched.csv

data > silver > kpop_music_videos_enriched.csv > data

```
Date,Artist,Song Name,Korean Name,Director,Video,Type,Release,views,likes,comments
2020-05-22,Agust D,Daechwita,대취타,,https://youtu.be/qGjAWJ2zWWI,Boy Solo,Major,463211676.0,15026379.0,1377719.0
2020-05-21,Yubin,yaya (Me Time),냅냅,,https://youtu.be/BJD0arHF_5c,Girl Solo,Major,1196539.0,53236.0,2697.0
2020-05-21,OnlyOneOf,Angel (prod. GRAY),,,https://youtu.be/vVTo8p72FxQ,Boy,Major,7630194.0,85263.0,3992.0
2020-05-20,Ryu Sujeong,Tiger Eyes,,,https://youtu.be/aE6curPGQRY,Girl Solo,Major,2675044.0,43217.0,3486.0
2020-05-20,Crush,Mayday feat. Joy (Red Velvet),자나깨나,,https://youtu.be/29ycT6fA-Rs,Boy Solo,Major,18690673.0,665181.0,31744.0
2020-05-19,NCT 127,Punch,,,https://youtu.be/U08OSl3V4po,Boy,Major,46555941.0,1828052.0,456340.0
2020-05-19,SECRET NUMBER,Who Dis?,,,https://youtu.be/mkYwq_CKpyw,Girl,Major,48191871.0,1638851.0,452695.0
2020-05-18,REDSQUARE,ColorFull,,,https://youtu.be/we1wqjcsYBU,Girl,Major,,,
2020-05-18,N.Flying,Amnesia,,,https://youtu.be/UcxNeKzdDN4,Boy,Japanese,485902.0,29920.0,1603.0
2020-05-18,TXT,Can't You See Me?,"세계가 불타버린 밤, 우린...",,https://youtu.be/cMFHUTJ13Ys,Boy,Major,78528314.0,3638426.0,590468.0
```

*État du conteneur **Silver***

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bail
kpop_idols_boy_groups.csv	05/01/2025 17:36:13	Élevé (déduit)		Objet blob de blocs	7.52 KiB	Disponible
kpop_idols_girl_groups.csv	05/01/2025 17:36:13	Élevé (déduit)		Objet blob de blocs	7.73 KiB	Disponible
kpop_idols.csv	05/01/2025 17:36:12	Élevé (déduit)		Objet blob de blocs	96.06 KiB	Disponible
kpop_music_videos_enriched.csv	05/01/2025 18:09:19	Élevé (déduit)		Objet blob de blocs	400.56 KiB	Disponible
kpop_music_videos.csv	05/01/2025 17:36:13	Élevé (déduit)		Objet blob de blocs	314.92 KiB	Disponible

Résumé des transformations entre *kpop_music_videos.csv* et *kpop_music_videos_enriched.csv*

Action	Détail
Chargement des données	Lecture du fichier ***Silver*** ***kpop_music_videos.csv***.
Extraction des ID vidéo	Extraction des identifiants YouTube depuis les URL des vidéos.
Récupération des métriques	Collecte des données `views`, `likes`, `comments` via l'API YouTube Data v3.
Enrichissement des colonnes	Ajout des métriques comme colonnes supplémentaires dans le DataFrame.
Sauvegarde des données enrichies	Export des vidéos enrichies dans ***kpop_music_videos_enriched.csv***, prêtes pour l'étape ***Gold***.

Ajout : Script de test de l'API YouTube

Pour garantir le bon fonctionnement de l'interrogation de l'API YouTube avant l'enrichissement des données, un script de test test_api_call.py a été mis en place

Objectif du script :

- Vérifier que la connexion à l'API YouTube Data v3 fonctionne correctement

- Tester la récupération des données pour une vidéo spécifique

Fonctionnement :

1. **Chargement de la clé API** :

   La clé API est chargée depuis un fichier `.env` pour garantir la sécurité des identifiants

2. **Requête de test :**

   Une vidéo YouTube spécifique est interrogée via son ID (EZntLk9bTUw), en utilisant le service youtube.videos().list avec le paramètre part="statistics"

3. **Affichage des résultats :**

   Les métriques de la vidéo sont affichées sous forme de réponse JSON, incluant des données telles que les *viewCount*, *likeCount*, et *commentCount*

Exemple de sortie JSON :

```
C:\Users\mouss\Documents\Efrei\Cours\Big data\Projet\Big-Data-K-pop-Analytics\scripts>python test_api_call.py
{'kind': 'youtube#videoListResponse', 'etag': 'jRuWphiikFh0Pc-Plw0VSI_wnSQ', 'items': [{'kind': 'youtube#video'
, 'etag': 'McQiBkaJ-RrU4MJfoIh1ntr-uak', 'id': 'EZntLk9bTUw', 'statistics': {'viewCount': '16849575', 'favorite
Count': '0', 'commentCount': '13359'}}], 'pageInfo': {'totalResults': 1, 'resultsPerPage': 1}}
```

Rôle dans le projet :

- **Validation technique :**
  - La clé API est correctement configurée
  - L'API répond avec les informations attendues

- **Réduction des erreurs :**
  Il a servi à isoler les problèmes liés à l'authentification ou aux restrictions de l'API avant d'intégrer les appels dans le script principal **`get_metrics_music_video.py`**

Transition de Silver à Gold

Objectif de cette étape :

Structurer les données nettoyées et enrichies (couche ***Silver***) en tables relationnelles prêtes pour l'analyse et la visualisation dans Power BI. Cette étape comprend la création de dimensions et de faits, le calcul de métriques spécifiques (comme le ratio d'engagement) et l'export des données sous un format adapté.

Étapes clés effectuées dans le script

A. Chargement des données *Silver*

- Les fichiers enrichis de la couche ***Silver*** sont importés
  - ***kpop_idols.csv***
  - ***kpop_idols_boy_groups.csv***
  - ***kpop_idols_girl_groups.csv***
  - ***kpop_music_videos_enriched.csv***

- Chargement des fichiers dans des DataFrames Pandas pour les manipulations ultérieures

B. Création des clés primaires

- Ajout d'identifiants uniques (colonnes id) pour chaque table
  - **Idols** : id_idol
  - **Boy Groups :** id_group
  - **Girl Groups :** id_group
  - **Vidéos :** id_video

- Garantir des relations uniques entre les tables et éviter les conflits lors des jointures

Ajout des colonnes id nouvellement créées

```
# Ajouter une colonne 'id' unique pour chaque DataFrame
idols['id_idol'] = idols.index + 1
boy_groups['id_group'] = boy_groups.index + 1
girl_groups['id_group'] = girl_groups.index + 1
videos['id_video'] = videos.index + 1
```

dimension_videos.csv

```
Date;Artist;Song Name;Korean Name;Director;Lien_video;Type;Release;vues;Likes;comments;id_video;ratio_engagement
```

dimension_idols.csv

```
nom_idol;Full Name;Korean Name;K. Stage Name;Date of Birth;Group;Country;Birthplace;Other Group;Gender;id_idol
```

dimension_groupes.csv

```
nom_du_groupe;Short;Korean Name;Debut;Company;Members;Orig. Memb.;Fanclub Name;Active;id_group;genre
```

A. Transformation des données en dimensions

Dimension Groupes

- Fusion des fichiers ***kpop_idols_boy_groups.csv*** et ***kpop_idols_girl_groups.csv***
  - Une colonne genre est ajoutée pour différencier les Boy Groups et Girl Groups
  - Les noms des groupes sont standardisés (lowercase et suppression des espaces)

- Renommage de la colonne Name en nom_du_groupe

Aperçu de la table dimension_groupes

dimension_groupes.csv

data > gold > dimension_groupes.csv > data

```
nom_du_groupe;Short;Korean Name;Debut;Company;Members;Orig. Memb.;Fanclub Name;Active;id_group;genre
100%;;백퍼센트;2012-09-18;TOP Media;4;7;Perfection;Yes;1;Boy Group
14u;;원포유;2017-04-17;BG;14;14;;Yes;2;Boy Group
1the9;;원더나인;2019-02-09;MBK;9;9;;Yes;3;Boy Group
24k;;투포케이;2012-09-06;Choeun;8;6;24U;Yes;4;Boy Group
2am;;투에이엠;2008-06-21;JYP, Big Hit;4;4;I Am;No;5;Boy Group
2pm;;투피엠;2008-07-04;JYP;6;7;Hottest;Yes;6;Boy Group
8eight;;에이트;2007-09-06;Big Hit;3;3;Sweet Voice;No;7;Boy Group
a-jax;;에이젝스;2012-06-01;DSP;5;7;A-LIGHT;No;8;Boy Group
a.c.e;ACE;에이스;2017-05-23;Beat;5;5;Choice;Yes;9;Boy Group
```

Dimension Idols

- Renommage de la colonne *Stage* Name en nom_idol
- Ajout d'un identifiant unique (id_idol)

Aperçu de la table dimension_idols

dimension_idols.csv

data > gold > dimension_idols.csv > data

```
nom_idol;Full Name;Korean Name;K. Stage Name;Date of Birth;Group;Country;Birthplace;Other Group;Gender;id_idol
A.M;Seong Hyunwoo;성현우;에이엠;1996-12-31;Limitless;South Korea;;;M;1
Ace;Jang Wooyoung;장우영;에이스;1992-08-28;VAV;South Korea;;;M;2
Aeji;Kwon Aeji;권애지;애지;1999-10-25;Hashtag;South Korea;Daegu;;F;3
AhIn;Lee Ahin;이아인;아인;1999-09-27;MOMOLAND;South Korea;Wonju;;F;4
Ahra;Go Ahra;고아라;아라;2001-02-21;Favorite;South Korea;Yeosu;;F;5
Ahyoung;Cho Jayoung;조자영;아영;1991-05-26;Dal Shabet;South Korea;Seoul;;F;6
Ahyung;Kang Ahyoung;강아형;아형;1996-08-27;P.O.P;South Korea;Pohang;;F;7
Ailee;Lee Yejin;이예진;에일리;1989-05-30;;South Korea;Denver;;F;8
Aini;Kim Heejung;김희정;아이니;1991-07-13;Pink Fantasy;South Korea;;;F;9
```

Dimension Vidéos

- Renommage des colonnes
  - views → vues
  - likes → likes
  - Video → lien_video

- **Calculs spécifiques**
  - Calcul du ratio d'engagement : ratio_engagement = likes / vues
  - Gestion des cas où vues = 0 pour éviter les valeurs infinies

Aperçu de la table dimension_videos enrichie

dimension_videos.csv

data > gold > dimension_videos.csv > data

```
Date;Artist;Song Name;Korean Name;Director;lien_video;Type;Release;vues;likes;comments;id_video;ratio_engagement
2020-05-22;agust d;Daechwita;대취타;;https://youtu.be/qGjAWJzzWWI;Boy Solo;Major;463211676,0;15026379,0;1377719,0;1;0,03243955145897488
2020-05-21;yubin;yaya (Me Time);냅냅;;https://youtu.be/BJD0arHF_5c;Girl Solo;Major;1196539,0;53236,0;2697,0;2;0,04449165468070827
2020-05-21;onlyoneof;Angel (prod. GRAY);;;https://youtu.be/vVTo8p72FxQ;Boy;Major;7630194,0;85263,0;3992,0;3;0,011174420991130763
2020-05-20;ryu sujeong;Tiger Eyes;;;https://youtu.be/aE6curPGQRY;Girl Solo;Major;2675044,0;43217,0;3486,0;4;0,016155622113131596
2020-05-20;crush;Mayday feat. Joy (Red Velvet);자나깨나;;https://youtu.be/29ycT6fA-Rs;Boy Solo;Major;18690673,0;665181,0;31744,0;5;0,035588927161691826
2020-05-19;nct 127;Punch;;;https://youtu.be/U08OSl3V4po;Boy;Major;46555941,0;1828052,0;456340,0;6;0,039265708322811045
2020-05-19;secret number;Who Dis?;;;https://youtu.be/mkYwq_CKpyw;Girl;Major;48191871,0;1638851,0;452695,0;7;0,03400679338637838
2020-05-18;redsquare;ColorFull;;;https://youtu.be/welwqjcsYBU;Girl;Major;nan;nan;nan;8;nan
2020-05-18;n.flying;Amnesia;;;https://youtu.be/UcxNeKzdDN4;Boy;Japanese;485902,0;29920,0;1603,0;9;0,0615762026087565
```

B. Création de la table des faits

- Fusion entre *dimension_videos* et *dimension_groupes* sur la base des colonnes Artist (dans *dimension_videos*) et nom_du_groupe (dans *dimension_groupes*)

- Suppression des lignes contenant des valeurs manquantes dans des colonnes critiques (vues, likes, ratio_engagement)

- Colonnes finales dans la table des faits :
  - id_video, nom_du_groupe, vues, likes, ratio_engagement, lien_video

C. Formatage et export des données Gold

- Conversion des formats
  - Remplacement des séparateurs décimaux (. → ,) dans les colonnes numériques (vues, likes, ratio_engagement)
  - Sauvegarde des fichiers en CSV avec le séparateur “;” pour assurer leur compatibilité avec Power BI

- Export final
  - ***dimension_idols.csv***
  - ***dimension_groupes.csv***
  - ***dimension_videos.csv***
  - ***table_faits.csv***

*Aperçu du dossier **Gold** contenant les fichiers prêts à être exporté.*

Accueil > kpopdatasets | Conteneurs >

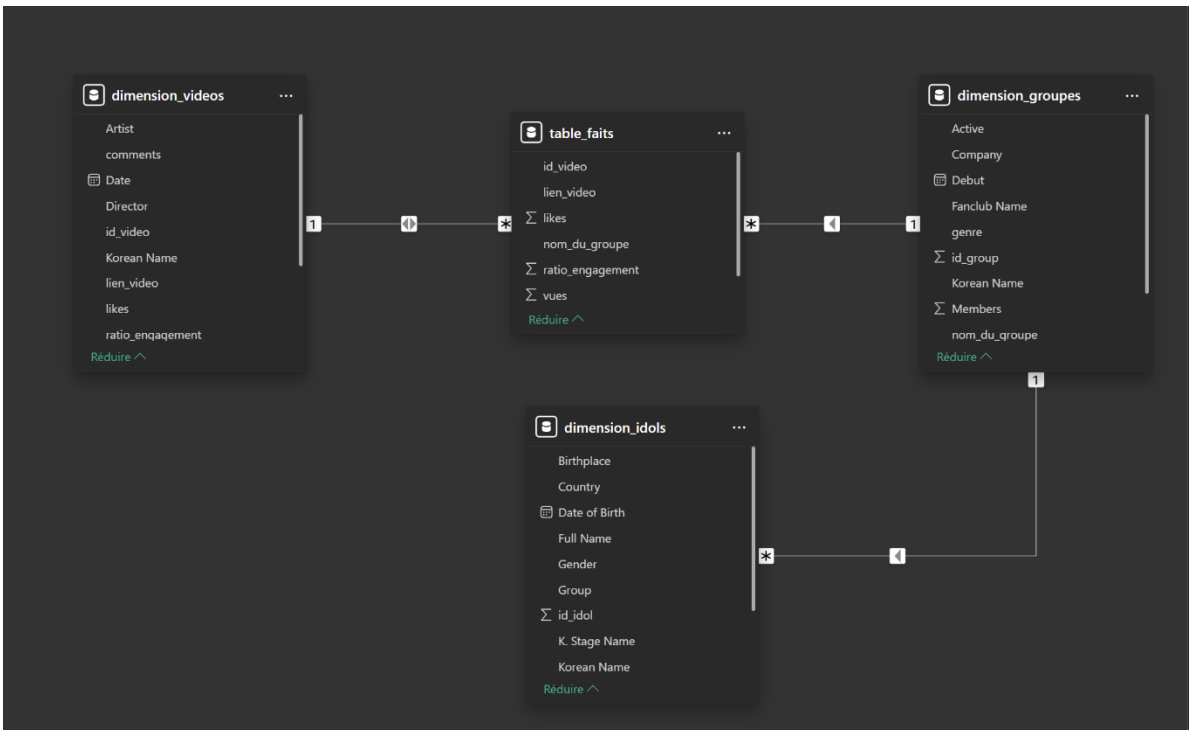
gold

Conteneur

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bail
dimension_groupes.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	19.2 KiB	Disponible
dimension_idols.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	101.35 KiB	Disponible
dimension_videos.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	490.53 KiB	Disponible
table_faits.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	181.73 KiB	Disponible

- Modèle et relations entre les tables

Aperçu de la vue du modèle dans Power BI



Résumé des transformations entre *Silver* et *Gold*

Action	Détail

Chargement des données	Lecture des fichiers enrichis depuis la couche ***Silver***.
Création des clés primaires	Ajout de colonnes ID (id_idol, id_group, id_video) pour les relations entre tables.
Transformation des données	Renommage des colonnes, calcul du ratio d'engagement, nettoyage des valeurs manquantes/infinies.
Jointure des tables	Fusion des dimensions et création de la table des faits.
Export *Gold*	Conversion des données en format CSV avec un séparateur adapté, prêtes pour Power BI.
Mise en place des relations et du modèle	Ajout des relations entre les différentes tables.

4. Analyse et Visualisation des Données

Pour la phase d'analyse et de visualisation, nous avons utilisé Power BI. Cet outil a permis de structurer les données issues de la couche **Gold** et de générer des visualisations interactives pour répondre à la problématique posée. L'utilisateur peut explorer les graphiques et interagir directement pour analyser des points spécifiques.

Rappel de l'outil :

- Power BI est un outil de Business Intelligence qui permet de transformer des données en visualisations interactives.

- Les données issues des tables de faits et de dimensions (couche **Gold**) ont été intégrées dans Power BI pour créer un modèle relationnel robuste et dynamique.

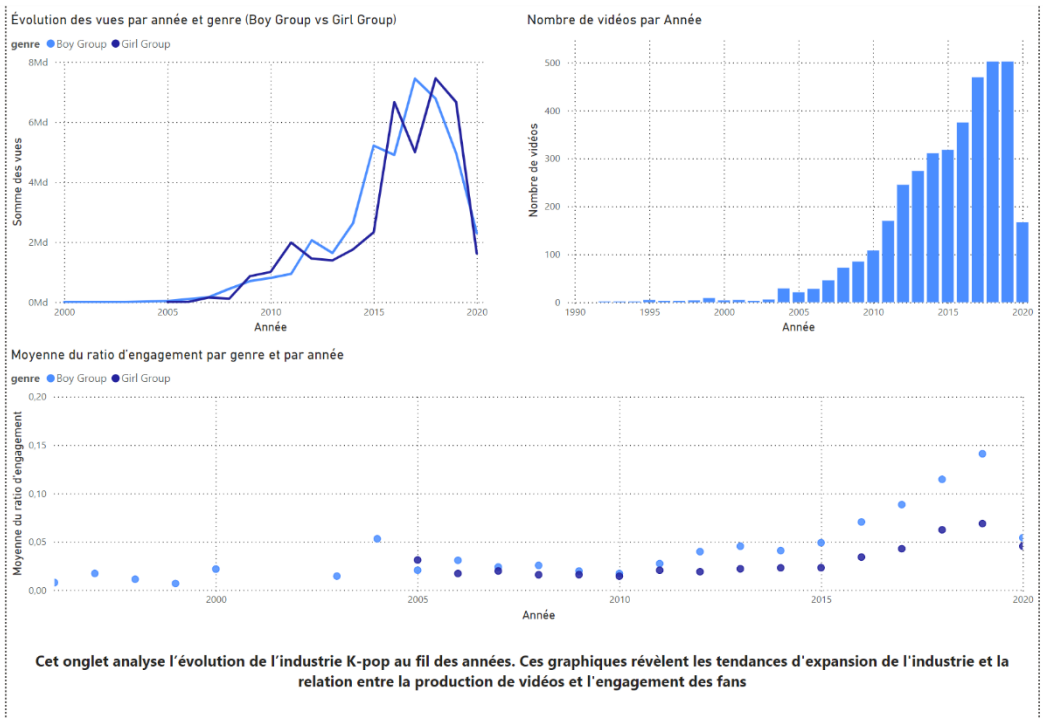
4.1 Présentation des Onglets et Visualisations

4.1.1 Analyse Temporelle des Performances

Cet onglet met en avant l'évolution de l'industrie K-pop au fil des années avec un focus sur la production de vidéos et l'engagement des fans.

- **Graphique 1 : Évolution des vues par année et genre**
  - Description : Ce graphique en ligne présente la somme des vues pour les Boy Groups et les Girl Groups sur une échelle temporelle.
  - **Analyse** : On observe une augmentation constante des vues depuis les années 2000, avec une accélération notable à partir de 2013.
  - **Intérêt** : Il permet de voir l'évolution des préférences du public entre les deux genres.
- **Graphique 2 : Nombre de vidéos par année**
  - **Description** : Un graphique en barres qui montre le volume de vidéos publiées chaque année.
  - **Analyse** : L'augmentation régulière du nombre de vidéos reflète la croissance rapide de l'industrie K-pop.
- **Graphique 3 : Moyenne du ratio d'engagement par genre et par année**
  - **Description** : Un nuage de points qui montre la moyenne du ratio d'engagement (*likes*/*vues*) pour chaque genre.
  - **Analyse** : Une tendance croissante du ratio d'engagement indique une interaction plus élevée des fans au fil des années.

Interaction suggérée : Filtrer les années spécifiques ou se concentrer sur un genre en particulier.



4.1.2 Comparaison des Groupes et Genres

Cet onglet compare les performances des **Boy Groups** et des **Girl Groups** en termes de popularité.

- **Graphique 1 : Moyenne des vues par groupe et genre**

  - **Description** : Un graphique en barres horizontales qui présente les groupes avec la moyenne de vues la plus élevée.

  - **Analyse** : Des groupes comme Blackpink et BTS dominent largement, soulignant leur popularité mondiale.

- **Graphique 2 : Popularité des groupes (vues et likes)**

◊ **Description** : Un tableau récapitulatif montrant la somme des vues et des likes par groupe et par genre.

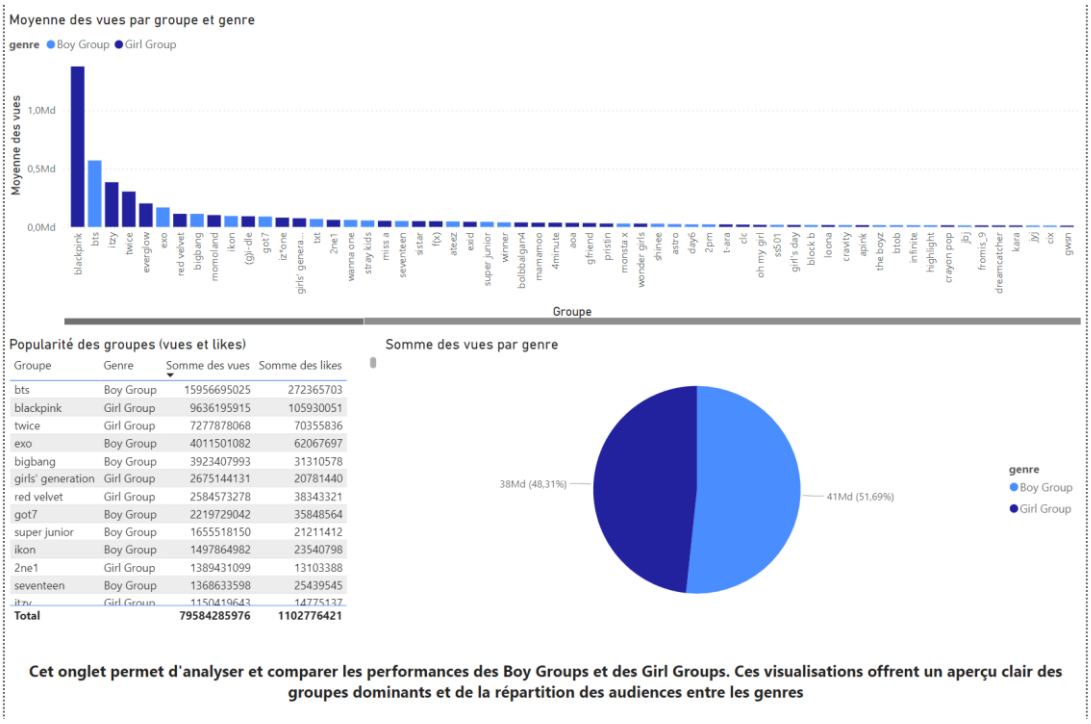
◊ **Analyse** : Ce tableau offre un aperçu détaillé des performances des groupes dominants.

♦ **Graphique 3 : Somme des vues par genre**

◊ **Description** : Un diagramme circulaire qui montre la répartition globale des vues entre **Boy Groups** et **Girl Groups**.

◊ **Analyse** : La répartition est équilibrée, avec une légère avance pour les Boy Groups.

Interaction suggérée : Trier les tableaux pour observer les groupes émergents et dominants.



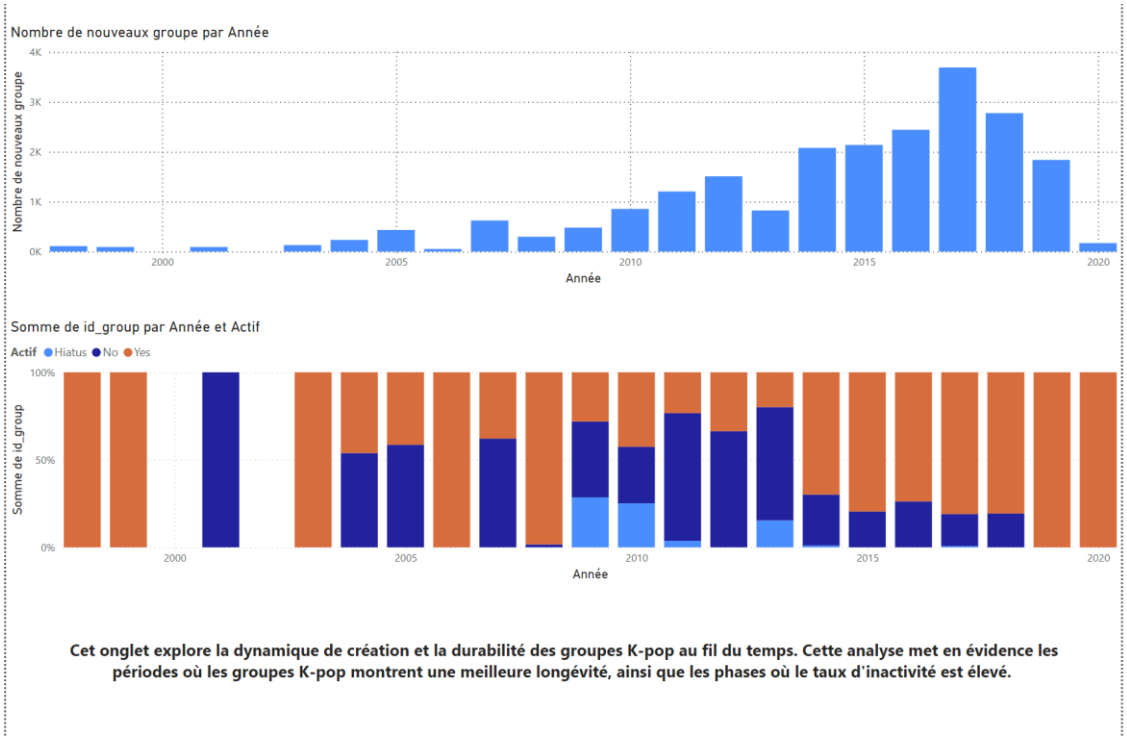
Groupe	Genre	Somme des vues	Somme des likes
bts	Boy Group	15956695025	272365703
blackpink	Girl Group	9636195915	105930051
twice	Girl Group	7277878068	70355836
exo	Boy Group	4011501082	62067697
bigbang	Boy Group	3923407993	31310578
girls' generation	Girl Group	2675144131	20781440
red velvet	Girl Group	2584573278	38343321
got7	Boy Group	2219729042	35848564
super junior	Boy Group	1655518150	21211412
ikon	Boy Group	1497864982	23540798
2ne1	Girl Group	1389431099	13103388
seventeen	Boy Group	1368633598	25439545
itzy	Girl Group	1150419643	14775137
Total		**79584285976**	**1102776421**

4.1.3 Analyse de la Longévité des Groupes K-pop

Cet onglet examine les dynamiques de création et de durabilité des groupes K-pop.

- **Graphique 1 : Nombre de nouveaux groupes par année**
  - **Description** : Un graphique en barres verticales qui montre la création de nouveaux groupes chaque année.
  - **Analyse** : Une croissance importante est observée entre 2010 et 2017.
- **Graphique 2 : Somme des groupes actifs/inactifs par année**
  - **Description** : Un graphique en barres empilées qui présente la proportion de groupes actifs, inactifs et en hiatus.
  - **Analyse** : Une tendance d'inactivité est visible pour les groupes plus anciens, ce qui souligne les défis de longévité dans l'industrie K-pop.

Interaction suggérée : Utiliser des filtres pour comparer les groupes actifs par année.



4.1.4 Analyse Géographique des Idols

Cet onglet se concentre sur la répartition géographique des idols de K-pop dans le monde.

- **Graphique : Répartition des idols par pays**
  - **Description** : Une carte interactive qui montre la provenance géographique des idols à travers des cercles proportionnels.
  - **Analyse** : La majorité des idols sont originaires de **Corée du Sud**, mais on observe également des contributions de pays voisins comme le Japon et la Chine.

Interaction suggérée : Cliquer sur les régions spécifiques pour voir le nombre exact d'idols par pays.

Invitation à Interagir

Pour une analyse approfondie, nous vous invitons à explorer ces graphiques directement dans Power BI. Vous pourrez filtrer par période, par groupe, ou par région pour obtenir des insights plus détaillés sur l'industrie K-pop.

5. Interprétation des Résultats

Dans cette section, nous allons interpréter les résultats des visualisations précédemment présentées pour répondre à notre problématique : Quels sont les facteurs de popularité et de longévité des groupes/idols de K-pop ?

L'analyse se concentre sur les tendances temporelles, la comparaison entre genres, les dynamiques de longévité et la répartition géographique.

5.1 Facteurs de Popularité

Analyse Temporelle des Performances

- **Graphique : Évolution des vues par année et genre**

  - Les vues des vidéos K-pop ont connu une croissance exponentielle à partir de 2010, avec un pic autour de 2016-2018. Cette tendance coïncide avec l'essor international des groupes comme BTS et Blackpink.

  - Les Boy Groups et Girl Groups affichent une évolution similaire, bien que les Boy Groups montrent une légère domination en volume global de vues.

- **Graphique : Nombre de vidéos par année**

  - La production de vidéos a fortement augmenté depuis les années 2010, atteignant son apogée entre 2015 et 2019.

  - Ce phénomène suggère une forte corrélation entre la quantité de contenu produit et l'engagement des fans. Une production régulière et abondante semble être un levier de popularité.

- **Graphique : Moyenne du ratio d'engagement par année et genre**

  - Le ratio d'engagement reste globalement faible mais montre une tendance croissante après 2015.

  - Les Boy Groups ont un engagement légèrement supérieur, probablement en raison d'une fanbase plus vaste et active.

Conclusion partielle :

La popularité d'un groupe est influencée par :

- La production continue de contenu (vidéos et performances).

- L'exposition médiatique accrue après 2010 grâce aux plateformes comme YouTube.

- L’engagement des fans, qui augmente avec la régularité du contenu.

5.2 Comparaison des Groupes et Genres

- **Graphique : Moyenne des vues par groupe et genre**

  - Des groupes comme Blackpink, BTS et Twice dominent clairement le classement avec des moyennes de vues très élevées.

  - Les Girl Groups rivalisent de plus en plus avec les Boy Groups, brisant le monopole masculin observée au début des années 2000.

- **Graphique : Popularité des groupes (vues et likes)**

  - Les Boy Groups affichent une somme totale de vues légèrement supérieure, mais les Girl Groups montrent des performances impressionnantes sur des périodes plus courtes.

  - Les likes, souvent considérés comme un indicateur d’appréciation directe, suivent une tendance similaire.

- **Graphique : Répartition des vues par genre**

  - La répartition des vues est relativement équilibrée : 51,69% pour les Boy Groups et 48,31% pour les Girl Groups.

  - Cette parité témoigne de l’évolution des préférences du public et de la montée en puissance des Girl Groups.

Conclusion partielle :

La popularité est aujourd'hui partagée de manière équilibrée entre les Boy Groups et les Girl Groups, bien que les Boy Groups conservent une légère avance grâce à leur historique plus long dans l'industrie.

5.3 Analyse de la Longévité des Groupes K-pop

- **Graphique : Nombre de nouveaux groupes par année**
  - L'industrie K-pop connaît des cycles de création massive de nouveaux groupes, notamment après 2010.
  - Le pic de création en 2016-2018 coïncide avec l'intérêt mondial pour la K-pop.

- **Graphique : Taux d'activité des groupes par année**
  - Une proportion élevée de groupes actifs dans les années récentes montre une meilleure longévité des groupes récents.
  - En revanche, les groupes plus anciens affichent un taux d'inactivité élevé, soulignant les défis de durabilité dans l'industrie.
  - Les périodes de "hiatus" deviennent de plus en plus fréquentes pour les groupes actifs après 5 à 10 ans d'existence.

Conclusion partielle :

- La longévité d'un groupe dépend de sa capacité à rester actif et pertinent sur le marché.

- Les groupes qui survivent au-delà de 5 ans sont rares, mais ceux qui le font ont généralement une fanbase solide et une stratégie médiatique efficace.

5.4 Analyse Géographique des Idols

- **Graphique : Répartition des idols par pays**

  - La Corée du Sud domine naturellement la scène K-pop, avec une majorité des idols originaires du pays.

  - Toutefois, des pays voisins comme le Japon, la Chine et la Thaïlande commencent à émerger, indiquant une influence régionale croissante.

  - Quelques idols provenant des États-Unis et d'Europe montrent l'impact global de la K-pop.

Conclusion partielle :

La répartition géographique des idols souligne l'influence mondiale de la K-pop tout en restant fortement centrée sur l'Asie.

5.5 Conclusion Globale

À travers l'analyse des visualisations interactives, nous pouvons identifier les principaux facteurs de popularité et de longévité dans l'industrie de la K-pop :

1. Production continue de contenu : La régularité et la quantité de vidéos influencent directement la visibilité et l'engagement.

2. Équilibre des genres : Les Boy Groups et Girl Groups se partagent désormais la scène, avec une montée en puissance des Girl Groups.

3. Longévité limitée : La majorité des groupes ne dépassent pas 5 ans d’activité, sauf exception grâce à une base de fans solide.

4. Influence mondiale : La K-pop reste dominée par la Corée du Sud, mais son influence s’étend désormais à d’autres régions.

Ces résultats mettent en lumière les dynamiques complexes de la K-pop et répondent à notre problématique initiale en identifiant les stratégies clés pour assurer la popularité et la durabilité des groupes/idols dans cette industrie compétitive.

6. Difficultés rencontrées

6.1 Problèmes rencontrés

Dans le cadre de ce projet, nous avons pris la décision de ne pas utiliser Databricks et de travailler directement en local sur nos postes. Ce choix nous semblait pertinent, car il offrait plusieurs avantages, notamment une plus grande liberté dans les traitements (par exemple, l'appel à l'API YouTube) et la possibilité de travailler dans un environnement qui nous était plus familier.

Cependant, cette décision a engendré plusieurs difficultés qui, malheureusement, restent sans solution viable à ce jour.

Ces problèmes concernent principalement les deux objectifs majeurs du projet : réaliser les traitements en utilisant le format Delta et produire un rapport de visualisation basé sur ce format.

Traitements au format Delta :
Nous avons tenté d'effectuer les traitements en local via des scripts Python, mais nous avons rencontré de nombreux obstacles liés à l'environnement, à la compatibilité des outils, ainsi qu'aux accès aux comptes de stockage et aux conteneurs. Ces difficultés nous ont empêchés de réaliser les objectifs dans les délais impartis.
En conséquence, nous avons uniquement pu convertir les fichiers CSV déjà traités au format Delta et les uploader directement dans leurs conteneurs respectifs. Ce résultat est loin d’être satisfaisant et nous

regrettons sincèrement de ne pas avoir pu aboutir à une solution plus complète.

Rapport de visualisation basé sur le format Delta :
Nous avons tenté d'intégrer les données au format Delta dans Power BI, mais le logiciel ne prend pas en charge ce type de données via un import direct. Nous avons également exploré l'option d'utiliser Azure Synapse Analytics en créant des pools SQL serverless, mais le format Delta n'était pas pris en charge. Enfin, l'utilisation de pools Spark a été envisagée, mais cette méthode s'est révélée payante et coûteuse.

En résumé, le choix de travailler en local, bien qu'initialement prometteur, s'est avéré devenir un piège face à la complexité des traitements Delta que nous avions sous-estimée. Cela a conduit à un rendu incomplet qui ne répond pleinement ni à vos attentes ni aux nôtres.

Nous avons néanmoins fait de notre mieux pour proposer un projet aussi abouti que possible, en essayant de pallier les lacunes du rendu. Nous sommes conscients que ce résultat reste en deçà des objectifs visés, et nous exprimons nos regrets à ce sujet.

6.2 Solution temporaire

Afin de nous rapprocher au maximum des attentes du projet, nous avons tout de même développé un script permettant de convertir les fichiers CSV traités présents dans nos conteneurs ***Silver*** et ***Gold*** au format Delta. Ce script, intitulé **« `convert_csv_to_delta.py` »**, réalise cette conversion de manière autonome.

Aperçu de la console suite à l'exécution du script

```
Traitement des fichiers dans le conteneur : bronze
Traitement du fichier : kpop_idols.csv
25/01/05 19:54:05 WARN SparkStringUtils: Truncated the string represe
sql.debug.maxToStringFields'.
Conversion et upload terminés pour : kpop_idols.csv
Traitement du fichier : kpop_idols_boy_groups.csv
Conversion et upload terminés pour : kpop_idols_boy_groups.csv
Traitement du fichier : kpop_idols_girl_groups.csv
Conversion et upload terminés pour : kpop_idols_girl_groups.csv
Traitement du fichier : kpop_music_videos.csv
Conversion et upload terminés pour : kpop_music_videos.csv
Traitement des fichiers dans le conteneur : silver
Traitement du fichier : kpop_idols.csv
Conversion et upload terminés pour : kpop_idols.csv
Traitement du fichier : kpop_idols_boy_groups.csv
Conversion et upload terminés pour : kpop_idols_boy_groups.csv
Traitement du fichier : kpop_idols_girl_groups.csv
Conversion et upload terminés pour : kpop_idols_girl_groups.csv
Traitement du fichier : kpop_music_videos.csv
Conversion et upload terminés pour : kpop_music_videos.csv
Traitement du fichier : kpop_music_videos_enriched.csv
Conversion et upload terminés pour : kpop_music_videos_enriched.csv
Traitement des fichiers dans le conteneur : gold
Traitement du fichier : dimension_groupes.csv
Conversion et upload terminés pour : dimension_groupes.csv
Traitement du fichier : dimension_idols.csv
Conversion et upload terminés pour : dimension_idols.csv
Traitement du fichier : dimension_videos.csv
Conversion et upload terminés pour : dimension_videos.csv
Traitement du fichier : table_faits.csv
Conversion et upload terminés pour : table_faits.csv
```

*Aperçu du contenu final du conteneur **silver***

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bail
kpop_idols						-
kpop_idols_boy_groups						-
kpop_idols_girl_groups						-
kpop_music_videos						-
kpop_music_videos_enriched						-
kpop_idols_boy_groups.csv	05/01/2025 19:10:16	Élevé (déduit)		Objet blob de blocs	7.52 KiB	Disponible
kpop_idols_girl_groups.csv	05/01/2025 19:10:23	Élevé (déduit)		Objet blob de blocs	7.73 KiB	Disponible
kpop_idols.csv	05/01/2025 19:10:03	Élevé (déduit)		Objet blob de blocs	96.06 KiB	Disponible
kpop_music_videos_enriched.csv	05/01/2025 18:09:19	Élevé (déduit)		Objet blob de blocs	400.56 KiB	Disponible
kpop_music_videos.csv	05/01/2025 19:10:24	Élevé (déduit)		Objet blob de blocs	314.92 KiB	Disponible

*Aperçu du contenu final du conteneur **gold***

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bail
dimension_groupes						-
dimension_idols						-
dimension_videos						-
table_faits						-
dimension_groupes.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	19.2 KiB	Disponible
dimension_idols.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	101.35 KiB	Disponible
dimension_videos.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	490.53 KiB	Disponible
table_faits.csv	05/01/2025 17:57:59	Élevé (déduit)		Objet blob de blocs	181.73 KiB	Disponible

7. Prise de recul et axes d'amélioration

Ce projet d'analyse de l'industrie de la K-pop a été l'occasion d'appliquer de manière concrète des compétences techniques, méthodologiques et

analytiques. Il a également permis d'identifier des axes d'amélioration pour la gestion future de données et de visualisations.

7.1 Compétences acquises et technologies apprises

- **Gestion et transformation des données**
  - **Compétence acquise :** Maîtrise du processus ETL (Extract, Transform, Load) avec une structure ***Bronze*** → ***Silver*** → ***Gold***.
  - **Technologie :** Utilisation de Pandas pour le traitement, la déduplication et le nettoyage des données.
    - Cela a permis de standardiser les données issues de multiples sources et d'assurer leur qualité avant analyse.
  - **Apprentissage :** Optimisation des scripts Python pour le traitement volumineux des datasets, gestion des erreurs et validations des données.
- **Connexion à des API externes**
  - **Compétence acquise :** Extraction de données via des APIs tierces, notamment l'API YouTube.
  - **Technologie :** Utilisation de Google API Client pour récupérer des statistiques de vidéos (vues, likes, commentaires).
    - Cela m'a permis de comprendre la structure d'une API REST, son authentification et son utilisation efficace pour enrichir un dataset.
  - **Apprentissage :** Gestion des limites d'appels API, manipulation des réponses JSON et intégration avec des fichiers de données existants.

- **Visualisation des données avec Power BI**

  - **Compétence acquise :** Création de tableaux de bord interactifs pour explorer et analyser des données.

  - **Technologie :** Utilisation de Power BI pour construire des graphiques variés (lignes, barres, secteurs, cartes géographiques).

    - Cela m'a permis de comprendre comment transmettre des insights complexes de manière visuelle et intuitive.

  - **Apprentissage :** Configuration de relations entre tables dans un modèle de données, utilisation de mesures et calculs DAX pour des indicateurs comme le ratio d'engagement.

- **Organisation et structuration d'un projet de données**

  - **Compétence acquise :** Structuration rigoureuse du projet en couches ***Bronze*** → ***Silver*** → ***Gold*** pour assurer la traçabilité et la qualité des données.

  - **Technologie :** Structuration des répertoires et automatisation des pipelines avec des scripts Python.

    - L'organisation des couches a facilité le suivi des données depuis leur état brut jusqu'à leur exploitation finale.

  - **Apprentissage :** Importance d'un workflow structuré pour la reproductibilité et l'évolutivité du projet.

- **Analyse des résultats et storytelling**

  - **Compétence acquise :** Interprétation des visualisations pour répondre à une problématique métier.

- ◊ **Technologie :** Rédaction d’une analyse détaillée basée sur les graphiques créés dans Power BI.
    - ◊ Cela a permis d’extraire des insights clés et de formuler des conclusions exploitables.
- ◊ **Apprentissage :** Traduction des données techniques en un récit compréhensible pour un public non technique.

Conclusion

Ce projet nous a offert une opportunité précieuse de consolider nos compétences en data engineering et data analysis tout en explorant une industrie aussi dynamique que la K-pop. Bien que nous n'ayons pas pleinement atteint les objectifs fixés, notamment en ce qui concerne l'intégration et l'exploitation complète du format Delta, nous avons néanmoins acquis des connaissances techniques et identifié des pistes d'amélioration pour le futur.

Ce projet nous a permis de mieux comprendre les limites de certains choix techniques et d’affiner notre approche méthodologique. Les compétences techniques développées, ainsi que les enseignements tirés des difficultés rencontrées, nous serons d'une grande utilité pour aborder de futurs projets.