قمنا **بتحميل مجموعتين من البيانات من موقع IMDb**

1. **ملف title.basics.tsv.gz**

هذا الملف يحتوي على معلومات أساسية عن جميع أنواع العناوين (أفلام، مسلسلات، حلقات، ...)، ويتضمن الأعمدة التالية:

- tconst :المعرف الفريد لكل عنوان (على شكل سلسلة حروف وأرقام).
- titleType: نوع العمل الفني) مثل.(... movie, short, tvSeries,
- primaryTitle: الاسم الشائع أو التجاري للعمل.
- originalTitle: الاسم الأصلي للعمل، بلغته الأصلية.
- isAdult :مؤشر يحدد إن كان العمل مخصصاً للبالغين فقط (0 = لا، 1 = نعم).
- startYear: سنة الإصدار أو بداية العرض (بالنسبة للمسلسلات).
- endYear: سنة نهاية العرض (فارغة لغير المسلسلات).
- runtimeMinutes: مدة العمل بالدقائق.
- genres: تصنيف العمل الفني (يمكن أن يتضمن حتى 3 أنواع).

2. **ملف title.ratings.tsv.gz**

هذا الملف يحتوي على تقييمات المستخدمين لكل عنوان، ويتضمن:

- tconst :المعرف الفريد للعمل (يستخدم للربط مع ملف الأساسيات).
- averageRating: متوسط التقييم (مرجّح) من قبل المستخدمين.
- numVotes :عدد الأصوات التي حصل عليها العمل.

كان حجم كل من مجموعات البيانات كالتالي:

بيانات عناوين الأفلام: 11,738,479 صف و 9 أعمدة تمثل كافة العناوين في IMDb من أفلام، مسلسلات، حلقات، إلخ.

بيانات التقييمات: 1,582,149 صف و 3 أعمدة تمثل فقط العناوين التي حصلت على تقييمات من المستخدمين.

تنظيف بيانات الأفلام واختيار النوع المطلوب:

حذف الأعمدة غير المفيدة : قمنا بحذف العمودين isAdult و endYear لأن isAdult يحتوي على نفس القيم تقريبا و endYear يحتوي على قيم NaN لمعظم العناوين. و قمنا بتصفية العناوين للاحتفاظ فقط بالعناوين من نوع "movie" ثم احتفظنا فقط بالأفلام التي لها تقييمات فعلية و تم التحقق من وجود تكرار في المعرف tconst وتبين عدم وجود تكرار.

أصبحت بيانات عناوين الأفلام تحتوي على 331,560 صف و 7أعمدة فقط.

أيضا قمنا بتنظيف البيانات لحذف أي فيلم لا يحمل عنوان إنجليزي. الهدف هو التركيز فقط على الأفلام ذات الأصل الإنجليزي .

الخطوات: كشف اختلاف الاسم بين primaryTitle و originalTitle

primaryTitle العنوان الشائع أو المستخدم في التوزيع.

originalTitle العنوان الأصلي بلغة الفيلم.

قمنا بإنشاء عمود title_diff لرصد الحالات التي يختلف فيها الاسمان و لاكتشاف العناوين غير الإنجليزية تم استخدام تعبيرات منتظمة regex للبحث عن الحروف الخاصة بلغات غير إنجليزية مثل (...é, à, ñ) أو كلمات شائعة من الفرنسية، الألمانية، الإسبانية... إلخ.

أيضا قمنا ب تصفية إضافية حسب الكلمات المفتاحية من عناوين أجنبية شهيرة للتأكد من إزالة أي بقايا لعناوين غير إنجليزية حتى لو لم تحتوي على حروف خاصة ثم قمنا بالاحتفاظ فقط بالعناوين التي تحقق الشرط التالي primaryTitle == originalTitle

أصبحت بيانات عناوين الأفلام تحتوي على 224,431 صف و 7 أعمدة فقط.

بعد تصفية جميع الأفلام غير الإنجليزية، كان من الضروري التحقق من وجود تكرارات لأفلام قد تكون مدخلة أكثر من مرة بنفس الاسم والسنة والنوع، خاصة أن بعض قواعد البيانات قد تحتوي على مدخلات مكررة لأسباب مثل إعادة إصدار أو أخطاء إدخال.

قمنا بفحص التكرار بناءً على أعمدة أساسية

primaryTitle الاسم الأساسي للفيلم.

originalTitle الاسم الأصلي.

startYear سنة الإصدار.

genres نوع الفيلم.

و كان عدد التكرارات المكتشفة **116** تم حذف التكرارات والاحتفاظ بسطر واحد فقط لكل فيلم .

بعد تصفية الأفلام غير الإنجليزية والتأكد من عدم وجود تكرارات، قمنا بخطوة إضافية تهدف إلى التخلص من أنواع الأفلام غير المرغوبة في التحليل، مثل العروض الواقعية أو الوثائقيات أو الأفلام القصيرة، وذلك للتركيز فقط على الأفلام السينمائية الدرامية أو الروائية التقليدية.

أصبحت بيانات عناوين الأفلام تحتوي على 175,993 صف و 7 أعمدة فقط.

بعد تصفية الأنواع غير المرغوبة، قمنا بمعالجة القيم المفقودة والقيم غير المنطقية في أعمدة السنوات ومدة عرض الأفلام.

و لتحقيق ذلك قمننا بالتالي: في عامود startyear تم حذف الصفوف التي تحتوي على سنة مفقودة وحولناها إلى int و حذفنا الأفلام التي سنة إنتاجها قبل 1900 أو بعد 2025 وفي عامود runtimeMinutes تم تحويل القيم إلى أرقام والتخلص من القيم غير الصالحة (مثل النصوص) أيضا تم ملاحظة قيم غير منطقية (مثل 59460 دقيقة) بالتالي اعتمدنا أن المدة المنطقية للفيلم تكون بين 30إلى 300 دقيقة و في عامود genres تم تعويض القيم المفقودة بكلمة 'Unknown'

أصبحت جميع القيم نظيفة ومنطقية.

إضافة تقييمات المستخدمين وإنشاء ميزات جديدة

بعد تنظيف البيانات، قمنا بدمج تقييمات المستخدمين من الملف الثاني title.ratings.tsv.gz، وإنشاء عدد من الميزات المشتقة لتسهل التحليل لاحقا. تم دمج عمود averageRating مع بيانات الأفلام باستخدام المفتاح المشترك tconst و إنشاء ميزات جديدة مثل :

تصنيف مدة الفيلم (runtimeCategory) , تم إنشاء تصنيف ثلاثي للأفلام بناءً على مدة الفيلم:

- **short** أقل من 75 دقيقة
- **standard** بين 75 و 120 دقيقة
- **Long** أكثر من 120 دقيقة

عمر الفيلم (movie_age) يحسب كم مضى من الزمن منذ عرض الفيلم لأول مرة حتى عام 2025

تصنيف التقييم (rating_category) لتسهيل الفهم البصري والتحليل، قمنا بتصنيف التقييم العددي إلى أربع فئات:

- **Excellent** التقييم من 8 فما فوق
- **Good** بين 6.5 و 7.9
- **Average** بين 5 و 6.4
- **Poor** أقل من 5

بالنهاية أصبح لدينا الآن 153,443 فيلم إنكليزي مرفق ببيانات شاملة:

- النوع
- المدة
- سنة الإنتاج
- التصنيف العام حسب تقييمات المستخدمين
- مدة الفيلم مصنفة
- عمر الفيلم

بيانات الممثلين المشهورين (Superstar Actors)

قمنا بتحليل ودمج بيانات الممثلين المشاهير من مصدرين مختلفين لتحديد قائمة"superstar actors" ، وهم الممثلون الأعلى شهرة والأعلى دخل. استخدمنا ملفين:

1. **Celebrity.csv**
يحتوي على بيانات عامة عن المشاهير (ممثلين، مغنيين... إلخ)، ويتضمن الأعمدة التالية:

- Unnamed: 0 رقم تسلسلي
- Name اسم الشهرة.
- original_name الاسم الحقيقي.
- Popularity مقياس للشعبية.
- Gender الجنس.
- Id معرّف المشهور.
- known_for_department المجال الذي يشتهر فيه (مثلاً التمثيل).
- Adult هل هو محتوى للبالغين فقط.

2. **forbes_celebrity_100.csv**
يحتوي على قائمة فوربس لأعلى المشاهير دخلاً، ويتضمن:

- Name اسم المشهور.
- Pay (USD millions) الدخل بملايين الدولارات.
- Year السنة التي تم فيها التقييم.
- Category المجال المهني (رياضة، موسيقى، أفلام...).

قاعدة بيانات المشاهيرتحتوي على **9980** صف و **8** أعمدة.

قاعدة Forbes تحتوي على**1647** صف و **4** أعمدة.

تنظيف بيانات المشاهير

حذف الأعمدة غير الضرورية : تم حذف الأعمدة التالية لعدم الحاجة إليها في التحليل الحالي 'Unnamed: 0', 'adult','gender','id'

و تم تصفية البيانات حسب مجال التمثيل تم الاحتفاظ فقط بالسجلات التي تنتمي إلى مجال التمثيل (Acting) ،

معالجة الأسماء غير الإنجليزية: نظراً لأن التحليل يركز فقط على الشخصيات المعروفة باللغة الإنجليزية، فقد تم استبعاد أي صف يحتوي على اسم (شهرة أو حقيقي) لا يتكون من أحرف إنجليزية فقط.
تم استخدام دالة isascii() للتحقق من أن الاسم لا يحتوي على رموز أجنبية، بالإضافة إلى استخدام تعبير منتظم (Regex) للتأكد من تطابق الأسماء مع نمط اللغة الإنجليزية و بعد التحقق من سلامة الأسماء، تم الاحتفاظ فقط بالسجلات التي يكون فيها name مطابقًا تمامًا لـ original_name.

لإزالة التكرارات تم التعرف على وجود أسماء مشاهير مكررة ضمن العمود nameكما تم فحص حالات وجود اختلاف بين nameو original_name.

بعد الانتهاء من تنظيف بيانات المشاهير، تم إجراء تحليل وصفي لمتغير popularityبهدف تحديد مدى انتشار وظهور الشخصيات في البيانات. و لتحليل أبرز الشخصيات، تم ترتيب البيانات ترتيبًا تنازليًا حسب عمود popularity، ثم اختيار أول 150 شخصية

أصبحت قاعدة بيانات المشاهيرتحتوي على 5946 صف و 4 أعمدة.

تنظيف قاعدة Forbes

تصفية فئة التمثيل : و نتج عن ذلك 176ممثلًا (Actors) و 107ممثلة (Actresses)

معالجة التكرارات: نظرًا لأن بعض المشاهير يظهرون في قائمة Forbes لعدة سنوات، تم استخراج الأسماء المكررة ثم تم اختيار أحدث ظهور فقط لكل شخصية بناءً على العمود Year، بهدف تحليل آخر دخل معروف ثم تم ترتيب النتائج تنازليًا حسب قيمة

Pay (USD millions) لتحديد الأعلى دخلًا.

أصبحت قاعدة Forbes تحتوي على **524** صف و **4** أعمدة.

مطابقة الأسماء بين الملفين: تم تحويل الأسماء في كلا الملفين إلى صيغة موحدة (أحرف صغيرة، بدون فراغات) بهدف تسهيل عملية المطابقة ثم تم استخراج التقاطع بين الأسماء المتوفرة في الملفين، وبلغ عدد الأسماء المتطابقة 73 اسم غيرر مكرر وتم الاحتفاظ فقط بالسجلات التي تحتوي على هذه الأسماء المتقاطعة ثم تم دمج الملفين باستخدام العمود الموحد name و بعد الدمج، تم الاحتفاظ فقط بأحدث سجل دخل لكل شخصية، ثم ترتيبهم حسب الدخل الأعلى

البيانات النهائية (merged_df) تحتوي على اسم المشهور و مستوى الشعبية popularity و الدخل السنوي Pay (USD millions) و نوع الفئة Category (ممثل أو ممثلة) هذا الجدول يمثل قائمة محدّثة من الممثلين والممثلات الأكثر شهرة والأعلى دخلًا، بناءً على تقاطع بيانات IMDbوForbes، وبعدد نهائي 73 شخصية.

ملاحظة: هناك العديد من أسماء الشخصيات الذي ظهرت اسمائها مكررة لان تم تصنيفها ضمن قائمة forbesعلى مدار اكثر من سنة و لكن تم الاعتماد على اجدد سنة تم تحديد دخل الشخصية فيها. لذلك ممكن ان يظهر عدد الأسماء المتطابقة بين الملفين 142 و لكن تم معالجة التكرار لتصبح 73 شخصية فقط

تم تعريف فئة"Superstar" نظرًا لأن عملية الدمج بين بيانات IMDb و Forbes أسفرت عن قائمة من الممثلين والممثلات الذين يتمتعون بشهرة واسعة popularity عالية ودخل سنوي مرتفع، فقد تم اعتبار هذه المجموعة نخبة المشاهير

و بعد تصنيف الشخصيات المدمجة بين IMDb و Forbesضمن فئة واحدة تعتبر "نجومًا كبارًا"(Superstars) ، تم إنشاء جدول عرض نهائي يضم المعلومات الأساسية التالية :

- Name اسم الشخصية.
- Popularityمؤشر الشهرة من.IMDb
- Pay (USD millions) الدخل السنوي كما ورد في قائمة.Forbes
- is_superstar مؤشر منطقي تم تعيينه إلى True لجميع السجلات، لتمييز الشخصيات المعتبرة نجوماً كباراً

يحوي علي 73 صف × 4 أعمدة أي أن عدد الشخصيات التي تتوفر لها معلومات كاملة من كلا المصدرين وتطابق الشروط بلغ 73 شخصية