Netflix Movies and TV Shows

Hinweis: Bei der Erstellung dieser Arbeit wurde ChatGPT als unterstützendes Werkzeug zur sprachlichen Überarbeitung eingesetzt. Dabei wurden vollständige Textpassagen vorgegeben, die anschließend von ChatGPT in wissenschaftlichere Formulierungen überführt wurden. Zudem diente das System punktuell als Hilfsmittel zur Ergänzung einzelner Cypher-Abfragen, Visualisierungen und zur Schließung technischer Verständnislücken.

1. Datenquelle und Kontext

Der gewählte Datensatz stammt von der Plattform Kaggle und trägt den Titel **“Netflix Movies and TV Shows”**. Die Daten wurden über die Netflix-Plattform zusammengestellt und enthalten Informationen zu Filmen und Serien, die dort weltweit verfügbar sind. Die Daten wurden vermutlich direkt von der Netflix-Website durch **Web Scraping** erhoben. Sie erfassen verschiedene Eigenschaften von Titeln, wie Erscheinungsjahr, Genre, Besetzung und Produktionsland.

Wir haben diesen Datensatz gewählt, weil er eine gute Mischung aus strukturierten Informationen und potenziellen Beziehung Netzwerken bietet (z. B. Schauspieler spielen in mehreren Filmen, Filme gehören zu mehreren Genres usw.). Dadurch eignet sich der Datensatz sehr gut für eine spätere Modellierung mit **Neo4j**.

2. Dateninhalt

Der Datensatz beschreibt konkrete Filme und Serien auf Netflix. Enthalten sind über **8.000 Einträge**, aufgeteilt in die folgenden Spalten:

column	Datentyp	optimales Datentyp	description
show_id	String	–	Eindeutige ID
type	String	–	“Movie” oder "Tv Show"
title	String	–	Titel
director	String	–	Regisseur/-in
cast	String	–	Hauptdarsteller/-innen
country	String	–	Produktionsland
date_added	String	–	Veröffentlichungsdatum auf Netflix
release_year	Integer	–	Ursprüngliches Erscheinungsjahr
rating	String	Integer	Altersfreigabe
duration	String	Integer	Laufzeit oder Anzahl Staffeln
listed_in	String	–	Kategorie
description	String	–	Kurze Inhaltsangabe

3. Beschreibung der Datenlücken

Bei der Analyse des Datensatzes zeigen sich mehrere Fehlstellen, die berücksichtigt werden müssen:

- **Regisseur:in (`director`) – fehlende Einträge**
  Ein erheblicher Teil der Titel enthält keine Angabe zur Regie. Dies betrifft vor allem Serien, bei denen die Regie häufig nicht zentral kommuniziert wird.
- **Land (`country`) – fehlende Einträge**
  Die Herkunftsländer der Produktionen fehlen bei einem signifikanten Teil der Daten. Dies erschwert länderspezifische oder kulturvergleichende Analysen.
- **Besetzung (`cast`) – fehlende Einträge**
  Auch die Schauspieler:innen sind bei zahlreichen Titeln nicht angegeben. Dies kann die Durchführung netzwerkanalytischer oder genderbezogener Untersuchungen behindern.
- **Hinzugefügt am (`date_added`) – fehlende Einträge**
  Bei wenigen Titeln fehlt das Veröffentlichungsdatum auf der Plattform. Diese Lücken sind quantitativ gering, könnten aber bei zeitlichen Analysen dennoch Verzerrungen verursachen.
- **Altersfreigabe (`age_restriction`) – fehlende Einträge**
  Die fehlenden Angaben zur Altersfreigabe sind vernachlässigbar gering, sollten bei Sicherheits- oder Jugendschutz bezogene Fragestellungen dennoch berücksichtigt werden. Außerdem sind Altersfreigaben in unterschiedlichen Regionen unterschiedlich zugeordnet. Die Daten wurden aber auf den USA Standard vereinheitlicht und damit könnten leichte Verfälschungen vorliegen.
- **Erscheinungsjahr (`release_year`) – falsche Einträge**
  An einzelnen Stellen steht die Duration anstatt dem Jahr, was dazu führt, dass es mehr Labels für die Jahreszahlen gibt, die das Ergebnis verfälschen.

4. Aufbereitung der Daten

- **Altersfreigabe geändert**

Wir haben ein Mapping erstellt, um die Spalte `age_restriction` Altersfreigaben als Strings zu haben. Folgendes Mapping wurde hierbei genutzt:

Gegebene Beschreibung	Skala Umwandlung	Alter Umwandlung
G	0	0
TV-Y	1	2
TV-G	2	4
TV-Y7; TV-Y7-FV	3	7
PG	4	8
TV-PG	5	10
PG-13	6	13
TV-14	7	14
R	8	16

TV-MA	9	17
NC-17	10	18
Not Rated; Unrated	11	–

- **Umwandlung der Spalte „duration“**

Im ursprünglichen Datensatz war die Variable `Duration` als String codiert, wobei die Einträge typischerweise in der Form *„90 min“* oder *„1 Season“* vorlagen. Diese Darstellung vereint numerische Werte mit Einheiten Texten, was eine direkte quantitative Analyse erschwert. Zur Vereinfachung der Weiterverarbeitung wurde `Duration` in ein Integer-Variable überführt. Dabei wurden die numerischen Bestandteile extrahiert, während die Einheiten (z. B. *„min“* für Filme bzw. *„Season(s)“* für Serien) entfernt wurden. Diese Transformation ermöglicht eine statistische Auswertung der Laufzeiten, z. B. zur Berechnung von Mittelwerten oder zur Klassifizierung nach Dauer.

- **Ergänzung fehlender Länderangaben mittels externer Datenquelle**

Im Ausgangsdaten Satz wiesen insgesamt 831 Einträge fehlende Angaben zum Produktionsland auf. Um diese Lücken zu minimieren und die Datenbasis für geografisch orientierte Analysen zu verbessern, wurde eine externe Datenquelle herangezogen. Die Vervollständigung der fehlenden `Country-Angaben` erfolgte durch einen Abgleich mit Wikidata, einer offenen Wissensdatenbank, die strukturierte Informationen zu Filmen und Serien enthält. Mittels Title Abgleich konnten die fehlenden Länderangaben für einen Großteil der betroffenen Datensätze rekonstruiert und ergänzt werden.

- **Ergänzung fehlender Kontinent Angaben**

Im Datensatz lagen keine Kontinentdaten vor. Um eine Aggregation zu ermöglichen, wurden den vorhandenen Länderangaben ihre jeweiligen Kontinente zugeordnet. Die Zuordnung basierte auf einer CSV, erstellt durch ChatGPT, in der jedem Land genau einen Kontinent (z. B. Europa, Asien, Afrika) zugewiesen wurde.

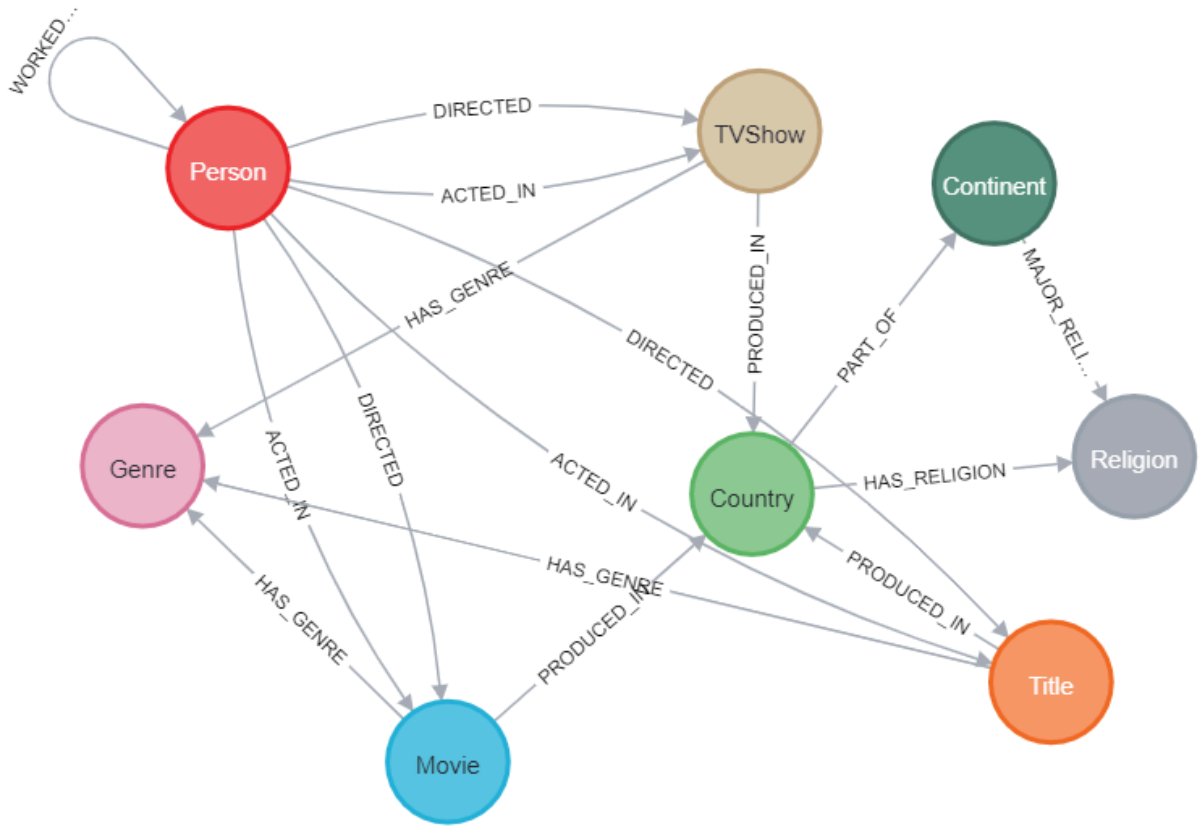
Diese Ergänzung ermöglicht es, Analysen nicht nur auf Länderebene, sondern auch aggregiert nach Kontinent durchzuführen – etwa zur Untersuchung kultureller, politischer oder wirtschaftlicher Unterschiede zwischen Weltregionen.

- **Ergänzung fehlender Religionsangaben**

Im Datensatz fehlten Informationen zur dominanten Religion einzelner Länder. Um kulturelle Vergleichsanalysen zu ermöglichen, wurden den Ländern Hauptreligionen zugewiesen. Die Zuordnung erfolgte mithilfe einer manuell erstellten CSV-Datei, in der jedem Land eine dominante Religion (z. B. Christentum, Islam, Hinduismus) zugeordnet ist.

Diese Ergänzung schafft die Grundlage für Analysen auf Basis religiöser Zugehörigkeiten – beispielsweise zur Untersuchung von Unterschieden in der Medienproduktion oder -darstellung zwischen Kulturräumen.

5. Graphmodell



6. Erste kritische Reflexion

1. **Unvollständige Felder (director, cast)**

→ Fehlende Werte in zentralen Feldern erschweren Analysen zur kreativen Beteiligung (z. B. Regie-Netzwerke oder Darstellervergleiche).

2. **Mehrfacheinträge in einem Feld (cast, listed_in)**

→ Kommagetrennte Einträge verhindern eine relationale Modellierung und erschweren die Zuordnung einzelner Akteur:innen oder Genres.

3. **Unvollständige oder mehrdeutige Länderdaten (country)**

→ Bei einigen Titeln fehlt das Herkunftsland ganz, bei anderen sind mehrere Länder angegeben. Das erschwert geografisch differenzierte Analysen oder Aggregationen nach Ländern.

4. **Fehlende Veröffentlichungsdaten (date_added)**

→ Ohne genaue Angabe des Veröffentlichungszeitpunkts lassen sich Zeitreihenanalysen oder saisonale Auswertungen nicht zuverlässig durchführen.

5. **Lücken in der Altersfreigabe (rating)**

→ Fehlt die Klassifikation, ist eine Analyse nach Zielgruppen oder Jugendschutzkriterien eingeschränkt.

6. **Fehlzuweisungen durch Titelähnlichkeit bei externen Quellen**

→ Ergänzungen fehlender Daten (z. B. country) durch Abgleich mit externen Quellen wie Wikidata können zu Fehlzuweisungen führen, insbesondere bei nicht eindeutigen oder mehrfach vorhandenen Titeln

Die beschriebenen Lücken, Formatierungsprobleme und potenziellen Fehlzuweisungen stellen zentrale Herausforderungen für die Datenqualität dar. Sie beeinträchtigen insbesondere Analysen, die auf vollständigen Beziehungen beruhen, wie etwa die Abbildung von *„Schauspieler spielt in Film X“*, oder Vergleiche zwischen Ländern und Zeiträumen. Ohne sorgfältige Vorverarbeitung besteht das Risiko verfälschter Visualisierungen, irreführender Trends oder fehlerhafter Schlussfolgerungen. Eine valide und belastbare Recherche erfordert daher einen methodisch fundierten Umgang mit diesen Datenmängeln.

Wie lässt sich das weltweite Inhaltsangebot von Netflix differenziert nach Ländern/ Regionen beschreiben und bewerten?

Frage 1: Bevorzugt Netflix das Serienformat gegenüber dem Filmformat – und wie unterscheiden sich diese Formate je nach Herkunftsland der Inhalte?

Formalisierung:

Untersucht wird der Einfluss des Produktions Landes bzw. der Region auf das bevorzugte Inhaltsformat (Film oder Serie). Zur Analyse werden das Produktionsland und das Content-Format (Serie/Film) herangezogen. Ziel ist es, den durchschnittlichen Anteil von Serien und Filmen pro Land zu berechnen, um regionale Unterschiede in der inhaltlichen Ausrichtung aufzuzeigen.

Operationalisierung:

- *Format* → Kategorische Variable → Produktionsformat
- *Herkunftsland* → Kategorische Variable → Produktionsland
- Schritte:
    - Zähle Anzahl Serien und Filme pro Land und berechne den prozentualen Anteil von Serien und Filmen je Land
    - Visualisiere im Koordinatensystem als Balkendiagramm
        - **X-Achse:** Länder
        - **Y-Achse:** Anteil Serien/Film
        - Die Darstellung wird nach dem Serienanteil sortiert, nicht alphabetisch nach Ländern

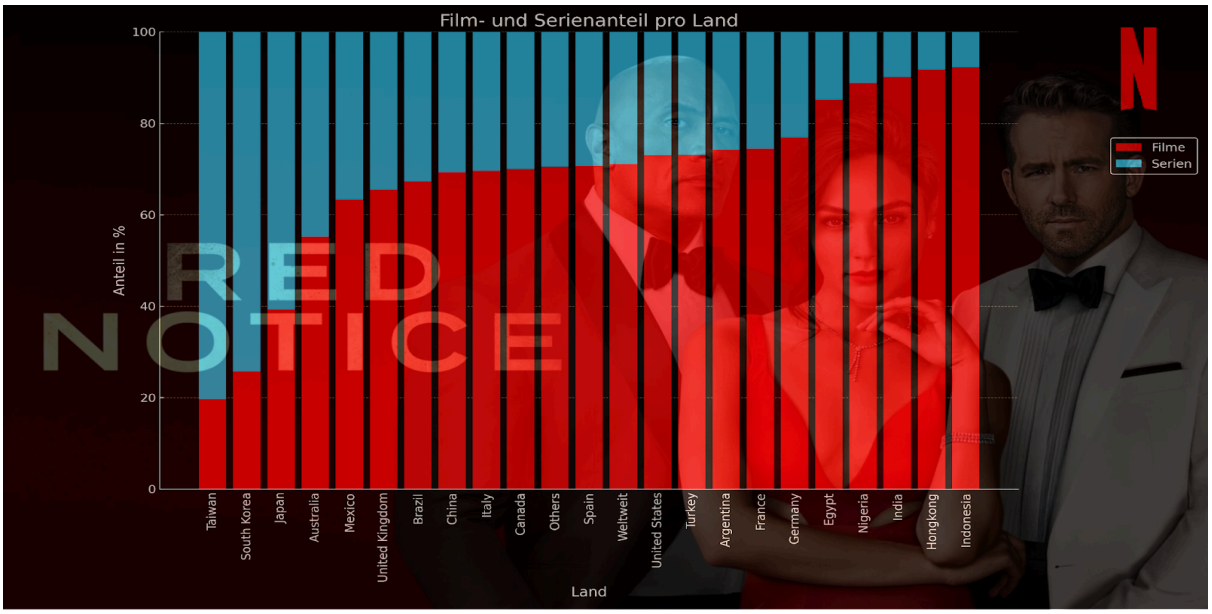
Kritische Reflexion:

Ein methodisches Problem ergibt sich aus der Tatsache, dass viele Inhalte unter **Mehrfachnennungen von Herkunftsländern** geführt werden. Eine eindeutige Zuordnung zu einem einzelnen Produktionsland ist daher nicht immer möglich und kann die Formatverteilung pro Land verzerren.

Zudem ist zwischen **nationalen Vorlieben** und **globalen Produktionsstrategien** zu unterscheiden: Netflix produziert zunehmend lokal ausgerichtete Inhalte, die jedoch international vertrieben werden. Die Analyse zeigt daher weniger reale Nutzerpräferenzen einzelner Länder, sondern vielmehr den Veröffentlichungsschwerpunkt innerhalb des Katalogs. Eine abschließende Bewertung, ob ein Land tatsächlich ein bestimmtes Format „bevorzugt“, bleibt dadurch eingeschränkt.

Cypher Code:

```
// Länder mit Film- und Serienproduktionen zählen
MATCH (c:Country)
OPTIONAL MATCH (c)<-[:PRODUCED_IN]-(m:Movie)
OPTIONAL MATCH (c)<-[:PRODUCED_IN]-(s:TVShow)
WITH c.name AS Country, count(DISTINCT m) AS Moviecount, count(DISTINCT s) AS Showcount

// Gruppieren kleiner Länder unter "Others"
WITH
  CASE
    WHEN (Moviecount + Showcount) <= 100 THEN "Others"
    ELSE Country
  END AS GroupedCountry,
  sum(Moviecount) AS Moviecount,
  sum(Showcount) AS Showcount

// Länderwerte zwischenspeichern und später "Weltweit" ergänzen
WITH collect({Country: GroupedCountry, Moviecount: Moviecount, Showcount: Showcount}) AS
Daten

// "Weltweit" Werte hinzufügen
WITH Daten + [
  {
    Country: "Weltweit",
    Moviecount: reduce(m = 0, row IN Daten | m + row.Moviecount),
    Showcount: reduce(s = 0, row IN Daten | s + row.Showcount)
  }
] AS FullData

// Prozentuale Verteilung berechnen
UNWIND FullData AS row
WITH
  row.Country AS Country,
  row.Moviecount AS Moviecount,
  row.Showcount AS Showcount,
  toFloat(row.Moviecount + row.Showcount) AS Total

// Ergebnis ausgeben
WITH
  Country,
  Moviecount,
  Showcount,
  round((toFloat(Moviecount) / Total) * 100, 2) AS MoviePercent,
  round((toFloat(Showcount) / Total) * 100, 2) AS ShowPercent

RETURN Country, Moviecount, Showcount, MoviePercent, ShowPercent
ORDER BY ShowPercent DESC;
```



Auswertung:

Zur Beantwortung der Fragestellung wurde der prozentuale Anteil von Serien und Filmen auf Netflix nach Produktionsland analysiert. Die Ergebnisse zeigen, dass der weltweite Katalog ein weitgehend ausgewogenes Verhältnis beider Inhaltsformate aufweist. Global betrachtet lässt sich daher keine klare Bevorzugung des Serien- oder Filmformats feststellen.

Auf regionaler Ebene zeigen sich jedoch deutliche Unterschiede. In ostasiatischen Ländern wie Taiwan, Südkorea und Japan dominieren Serien signifikant. Der Anteil von Serien liegt dort zum Teil deutlich über dem globalen Durchschnitt und markiert die höchsten Werte im Vergleich aller Länder. Diese Länder weisen eine ausgeprägte serielle Erzählkultur auf, was sich in der Formatverteilung auf Netflix deutlich widerspiegelt.

In Afrika sowie in Teilen Südasiens, insbesondere in Nigeria, Indien und Indonesien, zeigt sich hingegen ein gegenteiliger Trend: Hier liegt der Fokus stark auf dem Filmformat. Der Anteil von Serien ist in diesen Ländern deutlich geringer, während Filme einen Großteil des Angebots ausmachen. Diese Ausrichtung lässt sich in vielen Fällen mit historisch gewachsenen, nationalen Filmindustrien erklären, die das Medienangebot maßgeblich prägen.

Europa nimmt eine Mittelstellung ein. Länder wie Deutschland, Frankreich und Spanien zeigen ein relativ ausgeglichenes Verhältnis zwischen Serien und Filmen, wobei eine leichte Tendenz zur Filmorientierung feststellbar ist. Im Vergleich zu den Extremen anderer Regionen erscheint Europa damit eher balanciert.

Die aggregierte Kategorie „Weltweit“ spiegelt das globale Verhältnis wieder und dient als Referenzpunkt. Insgesamt bestätigen die Ergebnisse, dass sich das Verhältnis von Serien und Filmen auf Netflix deutlich nach Herkunftsland unterscheidet. Gleichzeitig bleibt auf globaler Ebene ein strukturelles Gleichgewicht erhalten, sodass keine eindeutige, weltweite Bevorzugung eines bestimmten Formats erkennbar ist.

Darüber hinaus lassen sich verschiedene mögliche Gründe identifizieren, die zur beobachteten Formatverteilung beitragen. Mögliche Gründe für dies sind erstens die strategische Anpassung von Netflix an lokale Produktionsbedingungen und

Konsumgewohnheiten. Zweitens werden Produktionsentscheidungen maßgeblich durch regionale Unterschiede in Kostenstrukturen und Produktionslogiken beeinflusst. Drittens variiert die Verfügbarkeit von Inhalten im Netflix-Katalog in Abhängigkeit vom geografischen Standort, was die wahrgenommene Formatverteilung zusätzlich verzerren kann. Diese Aspekte verdeutlichen, dass die Unterschiede zwischen Ländern nicht ausschließlich auf inhaltliche Präferenzen der Nutzer:innen zurückzuführen sind, sondern in erheblichem Maße durch ökonomische, strukturelle und distributionsbezogene Rahmenbedingungen bedingt werden.

Frage 2: Unterscheidet sich die durchschnittliche Filmlänge pro Land auf Netflix?

Formalisierung:

Es wird untersucht, ob das Produktionsland einen Einfluss auf die durchschnittliche Dauer von Filmen hat. Grundlage der Analyse sind die Filmlänge (in Minuten) sowie das jeweilige Produktionsland. Ziel ist es, die durchschnittliche Filmlänge pro Land zu berechnen und miteinander zu vergleichen.

Operationalisierung:

- *Länge* → Numerische Variable → Filmdauer in Minuten
- *Land* → Kategorische Variable → Produktionsland
- Schritte:
  - Filterung: Nur Filme (keine Serien) berücksichtigen
  - Gruppierung der Daten nach Produktionsland
  - Berechnung der durchschnittlichen Filmlänge pro Land
  - Visualisierung als Balkendiagramm
    - **X-Achse:** Länder
    - **Y-Achse:** Durchschnittliche Filmlänge (in Minuten)
    - Die Balken werden absteigend nach Filmlänge sortiert, nicht alphabetisch nach Ländern

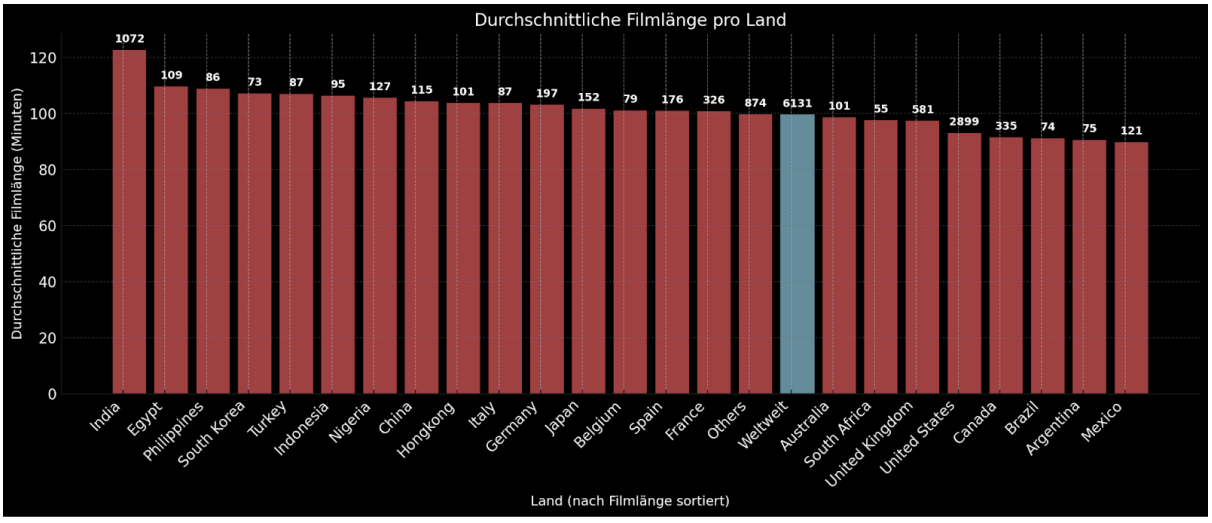
Kritische Reflexion:

Ein methodisches Problem ergibt sich aus **fehlenden oder fehlerhaften Laufzeitangaben**. Insbesondere extreme Ausreißer (z. B. sehr kurze oder sehr lange Angaben) können den Mittelwert verzerren.
Darüber hinaus können **kulturelle Erzähltraditionen** die durchschnittliche Filmlänge beeinflussen, was nicht durch eine feste zeitliche Referenz (z. B. 100 Minuten) adäquat abgebildet werden kann.

Cypher Code:

```
CALL {
 // Länder mit durchschnittlicher Filmlänge und Filmanzahl
 MATCH (m:Movie)-[:PRODUCED_IN]->(c:Country)
 WHERE m.duration IS NOT NULL
 WITH c.name AS Country, avg(toInteger(m.duration)) AS AvgDuration, count(m) AS MovieCount

 // Kleine Länder zu "Others" gruppieren
 WITH
   CASE WHEN MovieCount <= 50 THEN "Others" ELSE Country END AS GroupedCountry,
   collect({dur: AvgDuration, count: MovieCount}) AS Infos
```

```
// Durchschnitt neu berechnen
WITH GroupedCountry,
    reduce(totalDur = 0.0, entry IN Infos | totalDur + (entry.dur * entry.count)) AS TotalDuration,
    reduce(totalCount = 0, entry IN Infos | totalCount + entry.count) AS TotalCount
RETURN GroupedCountry AS Country,
      round(TotalDuration / TotalCount, 2) AS AvgDuration,
      TotalCount AS MovieCount

UNION

// Weltweiter-Durchschnitt berechnen
MATCH (m:Movie)
WHERE m.duration IS NOT NULL
RETURN "Weltweit" AS Country,
      round(avg(toInteger(m.duration)), 2) AS AvgDuration,
      count(m) AS MovieCount
}
RETURN Country, AvgDuration, MovieCount
ORDER BY AvgDuration DESC;
```



Auswertung:

Zur Überprüfung der Hypothese wurde die durchschnittliche Filmlänge von Netflix-Filmen in Abhängigkeit vom Produktionsland untersucht. Grundlage der Analyse bildeten ausschließlich als „Movie“ klassifizierte Titel. Serien sowie Datensätze mit fehlender oder unplausibler Laufzeit (z. B. extrem kurze oder ungewöhnlich lange Werte) wurden vorab aus dem Datensatz ausgeschlossen.

Die Ergebnisse wurden in Form eines Balkendiagramms visualisiert und nach absteigender durchschnittlicher Filmlänge sortiert. Auffällig ist, dass Filme aus Indien mit einer durchschnittlichen Laufzeit von über 120 Minuten die höchsten Werte aufweisen. Damit liegen sie signifikant über dem Niveau aller anderen Länder. Ebenfalls überdurchschnittlich lange Filme (mehr als 110 Minuten) stammen aus Ägypten, den Philippinen, Südkorea und der Türkei.

Im mittleren Bereich (etwa 100 bis 105 Minuten) befinden sich Länder wie Deutschland, Japan, Frankreich und Belgien, die im internationalen Vergleich eine durchschnittliche Filmlänge aufweisen. Diese Länder repräsentieren einen stabilen Mittelwert im Sample.

Am unteren Ende der Skala finden sich Kanada, Brasilien, Argentinien und insbesondere Mexiko, deren Filme durchschnittlich unter 95 Minuten lang sind. Diese Länder weisen somit eine tendenziell kürzere Filmdauer auf.

Darüber hinaus wurden auch aggregierte Kategorien wie „Weltweit“ und „Others“ ausgewertet. Diese dienen als Referenzgrößen, erlauben jedoch keine eindeutige geografische Zuordnung und wurden daher nicht in die Länderwertung einbezogen.

Die Analyse bestätigt die Hypothese: Die durchschnittliche Filmlänge unterscheidet sich signifikant in Abhängigkeit vom Produktionsland. Damit lässt sich ein klarer Einfluss der geografischen Herkunft auf die Laufzeit von Netflix-Filmen feststellen.

Mögliche Gründe für diese Unterschiede liegen unter anderem in variierenden Erzählkulturen. So zeichnen sich beispielsweise indische Filme traditionell durch lange Handlungsbögen und komplexe narrative Strukturen aus, die längere Laufzeiten begünstigen. Ebenso spielen historisch gewachsene Produktionsstandards eine Rolle, die je nach Land unterschiedliche Vorstellungen davon prägen, wie viel Erzählzeit ein Film benötigt. Schließlich können auch plattformspezifische Faktoren wie strategische Streamingentscheidungen eine Rolle spielen. In einigen Märkten setzt Netflix gezielt auf kürzere Produktionen, etwa um die Produktionskosten an lokale Gegebenheiten anzupassen.

Frage 3: Haben die Regionen/ Kontinente einen Einfluss auf dem Inhalt mit einer hohen Altersfreigabe?

Formalisierung:

Untersucht wird, ob ein Zusammenhang zwischen der geografischen Herkunft (Region bzw. Kontinent) und der durchschnittlichen Altersfreigabe von Inhalten auf Netflix besteht. Hierzu werden die Inhalte nach Kontinenten bzw. Ländergruppen geclustert und die Altersfreigaben in numerisch vergleichbare Werte überführt.

Operationalisierung:

- *Altersfreigabe* → Numerische Variable → In Jahre umwandeln (z. B. PG-13=13, R=16, TV-MA=17)
- *Ländergruppen* → Kategorische Variable → Gruppierung der Länder nach Kontinenten oder Ländergruppen (z. B. Europa, Asien, Nordamerika usw.)
- Schritte:
    - Berechnung der durchschnittlichen Altersfreigabe pro Region
    - Vergleich der Durchschnittswerte

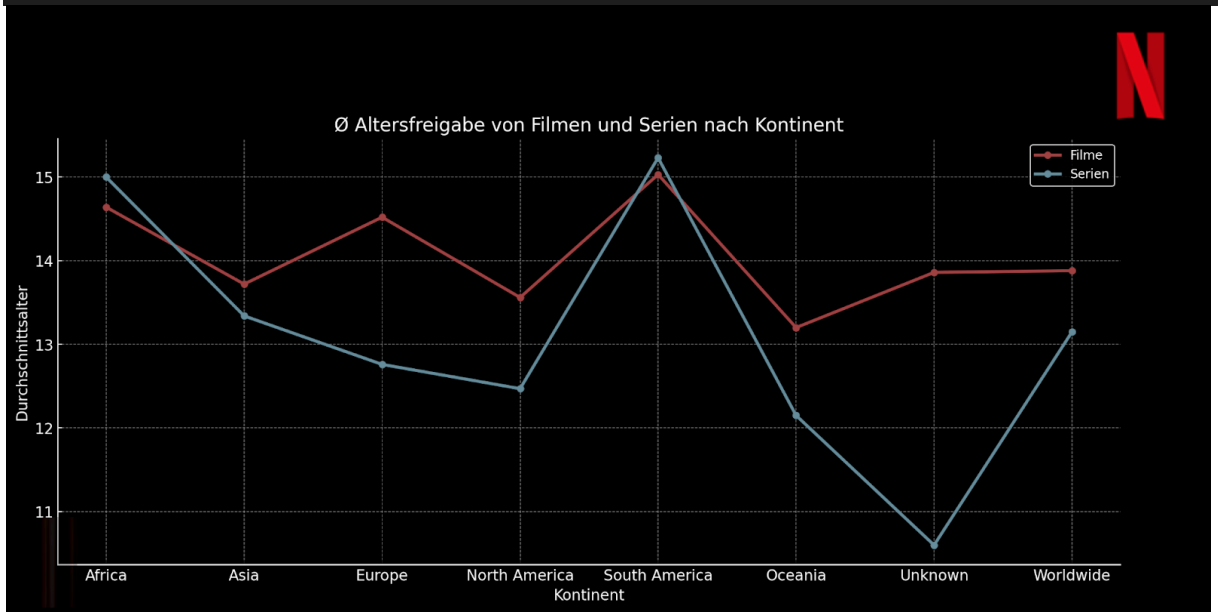
Kritische Reflexion:

Die numerische Standardisierung der Altersfreigaben ist methodisch notwendig, birgt jedoch Risiken. Altersfreigaben unterliegen **länderspezifischen Regelungen**, sodass z. B. eine US-amerikanische „TV-MA“-Einstufung nicht direkt mit einer indischen oder südkoreanischen Altersfreigabe vergleichbar ist.

Zudem kann die Gruppierung nach Kontinenten **kulturelle Unterschiede** innerhalb dieser Regionen **überdecken**. Besonders in heterogenen Regionen wie Asien oder Afrika ist die Verallgemeinerung mit Vorsicht zu interpretieren.

Cypher Code:

```
// Altersfreigabe von Filmen pro Kontinent
MATCH (m:Movie)-[:PRODUCED_IN]->(:Country)-[:PART_OF]->(co:Continent)
WHERE m.age IS NOT NULL
WITH co.name AS Continent, avg(toFloat(m.age)) AS AvgMovie

// Altersfreigabe von Serien im gleichen Kontinent
MATCH (s:TVShow)-[:PRODUCED_IN]->(:Country)-[:PART_OF]->(co2:Continent)
WHERE s.age IS NOT NULL AND co2.name = Continent
WITH Continent,
    round(AvgMovie, 2) AS Avg_Movie_Age,
    round(avg(toFloat(s.age)), 2) AS Avg_Series_Age

RETURN Continent, Avg_Movie_Age, Avg_Series_Age

UNION

// Weltweite Altersfreigabe
MATCH (mAll:Movie)
WHERE mAll.age IS NOT NULL
WITH avg(toFloat(mAll.age)) AS AvgMovie
MATCH (sAll:TVShow)
WHERE sAll.age IS NOT NULL
WITH AvgMovie, avg(toFloat(sAll.age)) AS AvgSeries

RETURN
 "Worldwide" AS Continent,
 round(AvgMovie, 2) AS Avg_Movie_Age,
 round(AvgSeries, 2) AS Avg_Series_Age
ORDER BY Continent;
```



Auswertung:

Zur Beantwortung der Hypothese wurde die durchschnittliche Altersfreigabe von Netflix-Inhalten in Abhängigkeit von ihrer geografischen Herkunft untersucht. Die Inhalte wurden nach Kontinenten gruppiert und ihre Altersfreigaben numerisch standardisiert, um Vergleichbarkeit herzustellen.

Die Ergebnisse zeigen deutliche Unterschiede zwischen den Regionen. Inhalte aus Afrika und Südamerika weisen mit durchschnittlich über 14 Jahren die höchsten Altersfreigaben auf. Dies lässt auf eine verstärkte Präsenz expliziter Inhalte schließen, etwa in den Bereichen Gewalt, Sexualität oder sozialkritische Themen. Demgegenüber zeigen Asien, Nordamerika und Ozeanien deutlich niedrigere Durchschnittswerte. Besonders bei Serien liegen die Altersfreigaben dort häufig unter 13 Jahren, was auf eine stärkere Orientierung an jugendfreundlichen Formaten oder auf restriktivere nationale Klassifizierungsstandards hinweisen könnte.

Europa liegt im mittleren Bereich. Die durchschnittlichen Freigaben sind niedriger als in Afrika und Südamerika, aber höher als in Asien und Nordamerika. Der Kontinent zeigt ein relativ ausgewogenes Verhältnis zwischen jugend- und erwachsenenorientierten Inhalten.

Darüber hinaus ist auffällig, dass Serien mit unbekannter Herkunft („Unknown“) im Durchschnitt die niedrigsten Altersfreigaben aller Serien aufweisen. Filme aus dieser Kategorie sind hingegen vergleichsweise hoch eingestuft. Inhalte mit der Herkunftsangabe „Worldwide“ liegen im globalen Mittelwert.

Ein weiterer Befund betrifft den Unterschied zwischen den Formaten: Filme erhalten weltweit tendenziell höhere Altersfreigaben als Serien, was auf komplexere, dramatischere oder inhaltlich anspruchsvollere Darstellungen in Filmformaten hinweist.

Mögliche Gründe für diese Unterschiede liegen unter anderem in den länderspezifischen Bewertungssystemen, die sich hinsichtlich ihrer Normen, Toleranzgrenzen und kulturellen Maßstäbe stark unterscheiden können. Darüber hinaus kann die technische Vereinheitlichung auf ein US-amerikanisches Bewertungssystem zu Verzerrungen führen, da dabei lokale Unterschiede nivelliert oder falsch abgebildet werden. Zudem sind kulturelle Vorstellungen darüber, was als jugendgerecht gilt, stark kontextabhängig und können erheblich variieren. Schließlich ist auch zu berücksichtigen, dass Serien oftmals komplexe, fortlaufende Erzählstrukturen mit sensiblen Thematiken behandeln, was ebenfalls zu höheren Altersfreigaben führen kann.

Die Ergebnisse stützen somit die Hypothese: Die durchschnittliche Altersfreigabe unterscheidet sich signifikant je nach geografischer Herkunft der Inhalte. Ein Einfluss der Region auf die inhaltliche Einstufung ist klar erkennbar.

Frage 4: Sind Filme auf Netflix, die in den letzten Jahren veröffentlicht wurden, im Durchschnitt kürzer als ältere Filme?

Formalisierung:

Untersucht wird, ob sich die durchschnittliche Filmlänge im zeitlichen Verlauf verändert hat. Im Fokus steht die Frage, ob Filme, die in den letzten Jahren erschienen sind, im Durchschnitt kürzer sind als ältere Produktionen. Ergänzend könnte perspektivisch auch die Entwicklung von Zuschauerpräferenzen hinsichtlich der Filmlänge analysiert werden, beispielsweise anhand von IMDb-Bewertungen.

Grundlage der Untersuchung sind die Variablen Veröffentlichungsjahr des Films sowie die Filmdauer in Minuten.

Operationalisierung:

- *Länge*: Numerisch → Numerische Variable → Filmdauer in Minuten
- *Jahr*: Numerisch → Numerische Variable → Veröffentlichungsjahr
- Schritte
  - Durchschnittslänge je Zeitraum berechnen
  - Vergleich: Neuere Filme < Ältere Filme?
  - Einteilung in Zeiträume (z. B. Jahrzehnte: 1942–2022 in 10-Jahres-Schritten)
  - Berechnung der **durchschnittlichen Filmlänge pro Jahrzehnt**
  - Visualisierung in einem Liniendiagramm oder Balkendiagramm
    - X-Achse: Jahrzehnte
    - Y-Achse: Durchschnittliche Filmlänge (in Minuten)

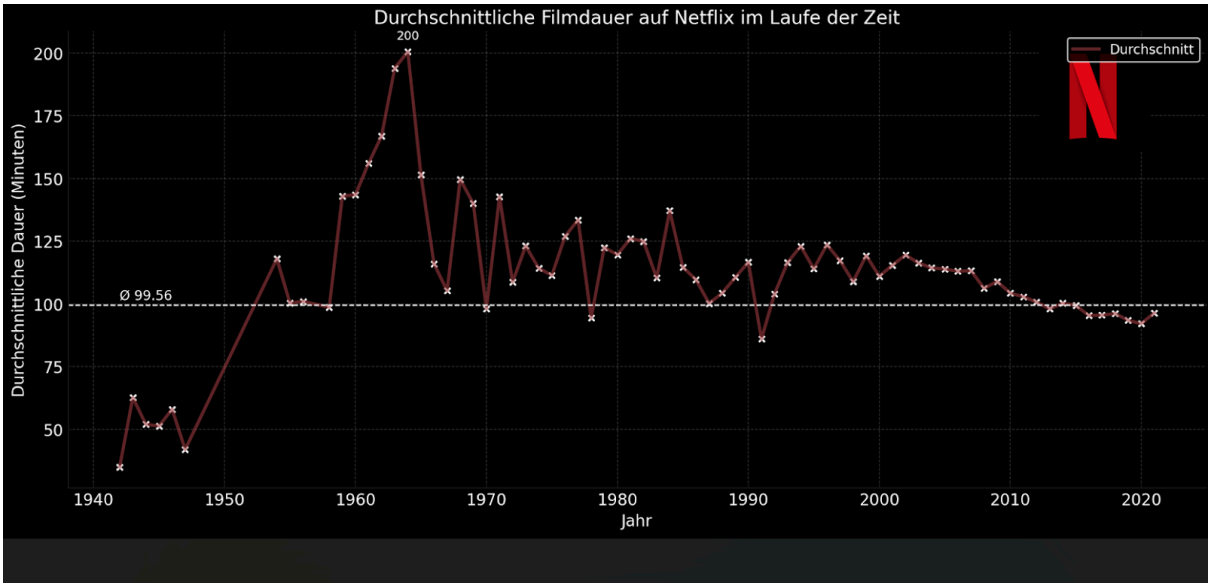
Kritische Reflexion:

Ein methodischer Unsicherheitsfaktor besteht in der **mangelnden Datenqualität bei älteren Filmen**. Ältere Produktionen weisen häufig unvollständige oder inkonsistente Laufzeitangaben auf. Dadurch kann der Eindruck entstehen, dass neuere Filme kürzer seien, obwohl diese Verzerrung lediglich auf Datenlücken beruht.

Ein weiterer Aspekt betrifft die **ungleiche Repräsentation der Zeiträume**: Filme, die vor dem Jahr 2000 erschienen sind, sind im Netflix-Katalog nur in geringer Anzahl vertreten. Dadurch entsteht eine statistische Verzerrung, die den Vergleich der Filmlängen über mehrere Jahrzehnte hinweg einschränkt. Zudem kann die algorithmische Kuratierung von Inhalten auf Netflix dazu führen, dass vorwiegend kürzere oder kommerziell verwertbare Filme prominent platziert und in die Analyse einbezogen werden.

Cypher Code:

```
CALL {
  // Teil 1: Durchschnitt und Anzahl pro Jahr
  MATCH (m:Movie)
  WHERE m.duration_value IS NOT NULL AND m.release_year IS NOT NULL
  WITH m.release_year AS Year, avg(toFloat(m.duration_value)) AS Average, count(*) AS
FilmCount
  RETURN Year, round(Average, 2) AS AverageDuration, FilmCount

  UNION

  // Teil 2: Gesamt-Durchschnitt und Gesamtanzahl (Jahr = 0)
  MATCH (m:Movie)
  WHERE m.duration_value IS NOT NULL
  WITH 0 AS Year, avg(toFloat(m.duration_value)) AS Average, count(*) AS FilmCount
  RETURN Year, round(Average, 2) AS AverageDuration, FilmCount
}

// sortiert ausgeben
RETURN Year, AverageDuration, FilmCount
ORDER BY Year ASC;
```



Auswertung:

Zur Beantwortung der Fragestellung wurde analysiert, ob sich die durchschnittliche Filmlänge auf Netflix im Zeitverlauf verändert hat. Berücksichtigt wurden ausschließlich als „Movie“ klassifizierte Inhalte. Als Analysevariablen dienten das Veröffentlichungsjahr sowie die Filmlänge in Minuten. Ziel war es, die durchschnittliche Filmlänge über mehrere Jahrzehnte hinweg zu ermitteln und etwaige Entwicklungen sichtbar zu machen.

Die Daten umfassen den Zeitraum von etwa 1940 bis 2022. Die durchschnittliche Filmlänge pro Jahr wurde in einem Liniendiagramm visualisiert, wobei der Gesamtdurchschnitt aller Filme bei rund 111,7 Minuten liegt und als Referenzpunkt markiert wurde.

Die Visualisierung zeigt einen deutlichen zeitlichen Verlauf. Besonders in den 1950er- und 1960er-Jahren ist ein signifikanter Anstieg der durchschnittlichen Filmlänge zu beobachten, mit einem Höhepunkt im Jahr 1960, in dem die Laufzeit im Durchschnitt bei etwa 200 Minuten lag. In den darauffolgenden Jahrzehnten sinkt die durchschnittliche Filmlänge allmählich, wobei die 1970er- bis 1990er-Jahre noch durch stärkere Schwankungen geprägt sind.

Ab 2000 lässt sich ein klarer Abwärtstrend feststellen, der sich in den Folgejahren fortsetzt. In den letzten Jahren liegt die durchschnittliche Filmlänge deutlich unter dem historischen Mittelwert, was auf eine veränderte Ausrichtung der Filmproduktion im Streaming-Zeitalter hinweist. Die Ergebnisse stützen somit die Hypothese, dass neuere Filme auf Netflix im Durchschnitt kürzer sind als ältere Produktionen.

Es könnten verschiedene mögliche Ursachen sein. Einerseits führen veränderte Rezeptionsgewohnheiten sowie zunehmend kürzere Aufmerksamkeitsspannen dazu, dass kompaktere Formate bei einem breiteren Publikum besser anschlussfähig sind. Auch die mobile Nutzung von Streamingdiensten begünstigt Inhalte mit kürzerer Laufzeit. Andererseits spielen plattformspezifische Strategien eine zentrale Rolle: Streaming-Anbieter wie Netflix bevorzugen häufig kürzere Formate, um beispielsweise das Binge-Watching zu fördern oder algorithmische Empfehlungen zu optimieren. Hinzu kommen veränderte Erzählstrukturen, die auf eine verdichtete Dramaturgie und eine schnellere Handlungsentwicklung setzen. Kommerzielle Überlegungen verstärken diesen

Wandel zusätzlich: Kürzere Filme verursachen geringere Produktionskosten und bieten größere Flexibilität im Katalogmanagement.

Frage 5: In welchem Genre sind Inhalte mit der Altersfreigabe „16“ auf Netflix häufiger zugeordnet?

Formalisierung:

Untersucht wird, welche Genres auf Netflix besonders häufig mit der Altersfreigabe „ab 16 Jahren“ versehen sind. Ziel ist es, festzustellen, ob bestimmte Genretypen systematisch mit einer höheren Altersfreigabe verbunden sind. Zur Beantwortung der Frage werden die Variablen Altersfreigabe, Genre und Produktionsland herangezogen.

Operationalisierung:

- *Altersfreigabe* → Filterkriterium → Filter auf Inhalte mit der Freigabe ab 16
- *Genre* → Kategorische Variable → Hauptgenre
- Schritte:

- Zähle Anzahl „16+“-Inhalte pro Genre
- Berechne die Zahl für alle Genre
- Sortiere Genres nach Anteil oder Häufigkeit
- Visualisiere in einem Balkendiagramm:
  - **X-Achse:** Genres
  - **Y-Achse:** Anteil bzw. absolute Anzahl der „TV-MA“-Inhalte

Kritische Reflexion:

Ein zentrales methodisches Problem stellt die **Mehrfachzuordnung von Genres** dar. Viele Titel sind mehreren Genres gleichzeitig zugeordnet. Wird nur das erste (Haupt-)Genre berücksichtigt, kann dies zu Verzerrungen führen. Eine alternative Vorgehensweise wäre eine proportionale Gewichtung oder Mehrfachzählung, was jedoch die Vergleichbarkeit erschwert.

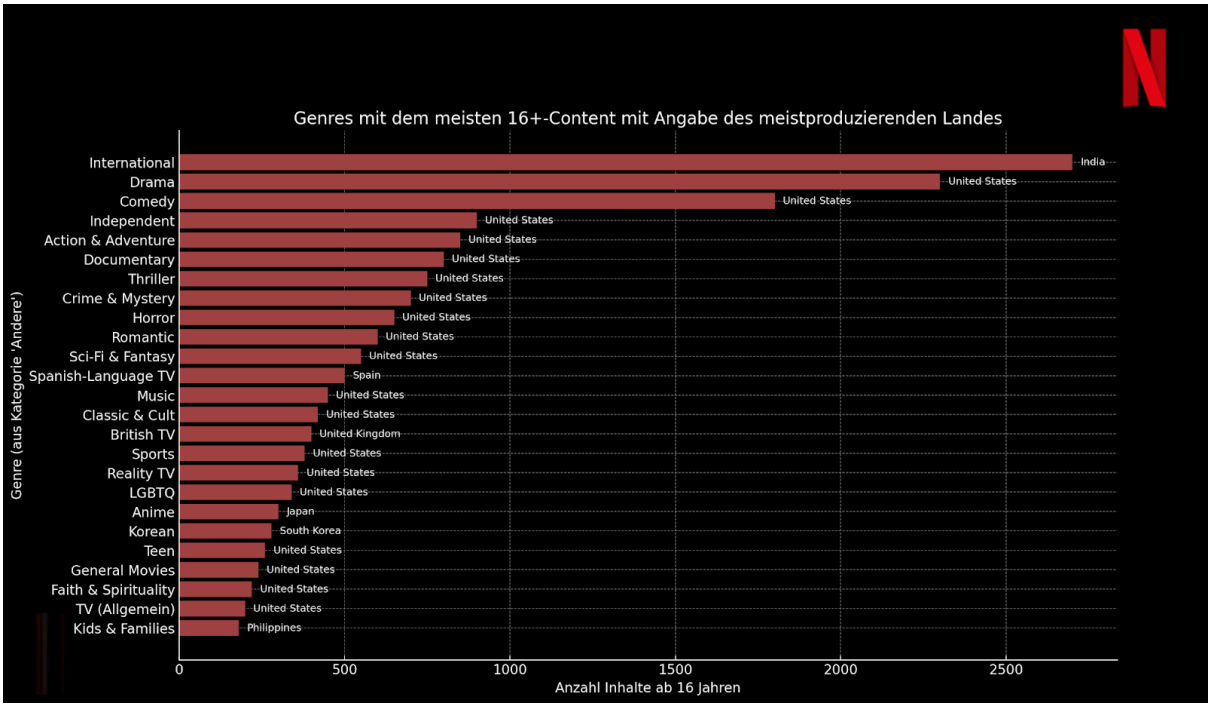
Zudem ist die **Genre-Terminologie uneinheitlich**. Es existieren keine verbindlichen Standards (z. B. „Crime“ vs. „Crime Thriller“), wodurch inhaltlich ähnliche Kategorien möglicherweise getrennt gezählt werden. Eine Vorverarbeitung zur Normalisierung der Genrebezeichnungen wäre daher sinnvoll.

Cypher Code:

```
// Inhalte mit Genre, Altersfreigabe ab 16 und Land
MATCH (m)-[:HAS_GENRE]->(g:Genre),
      (m)-[:PRODUCED_IN]->(c:Country)
WHERE (m:Movie OR m:TVShow)
  AND m.age IS NOT NULL
  AND m.age >= 16  // Nur Inhalte ab 16 Jahren

WITH toLower(g.name) AS genreName, c.name AS countryName

// Genres zu Kategorien zusammenfassen
WITH
 CASE
  WHEN genreName CONTAINS "comedy" THEN "Comedy"
  WHEN genreName CONTAINS "tv comed" THEN "TV Comedy"
```

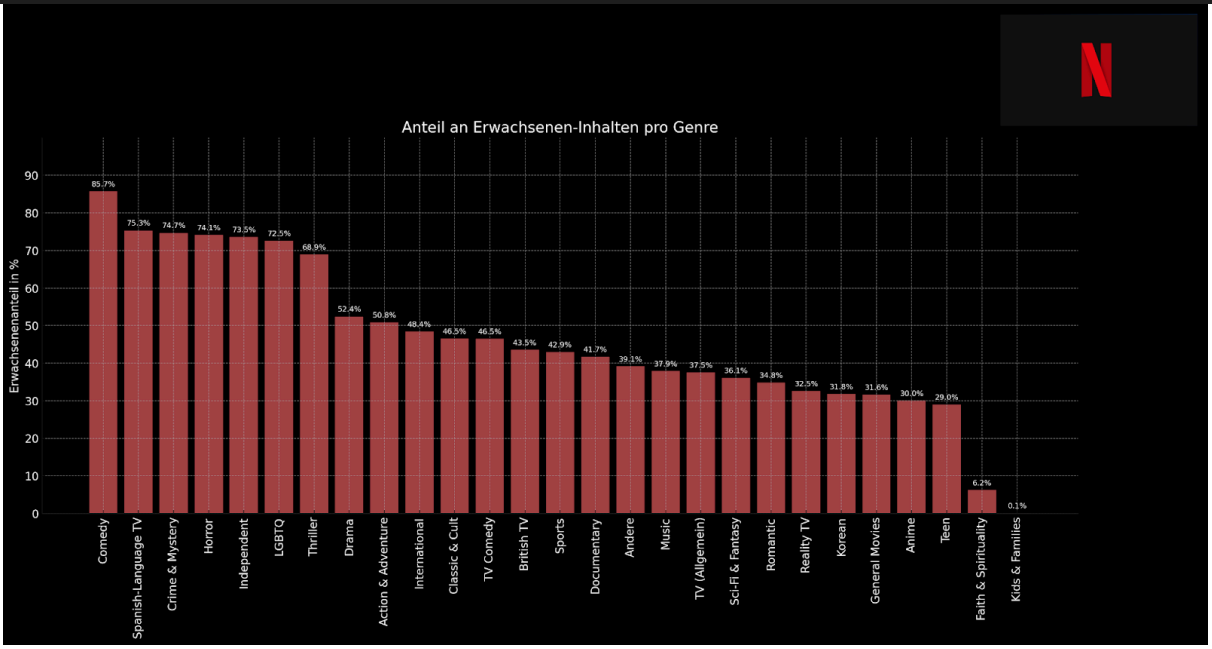
```
  WHEN genreName CONTAINS "horror" THEN "Horror"
  WHEN genreName CONTAINS "crime" OR genreName CONTAINS "mystery" THEN "Crime &
Mystery"
  WHEN genreName CONTAINS "thriller" THEN "Thriller"
  WHEN genreName CONTAINS "drama" THEN "Drama"
  WHEN genreName CONTAINS "action" OR genreName CONTAINS "adventure" THEN "Action &
Adventure"
  WHEN genreName CONTAINS "romance" OR genreName CONTAINS "romantic" THEN
"Romantic"
  WHEN genreName CONTAINS "documentary" OR genreName CONTAINS "docu" OR
genreName CONTAINS "science & nature" THEN "Documentary"
  WHEN genreName CONTAINS "teen" THEN "Teen"
  WHEN genreName CONTAINS "sport" THEN "Sports"
  WHEN genreName CONTAINS "music" THEN "Music"
  WHEN genreName CONTAINS "korean" THEN "Korean"
  WHEN genreName CONTAINS("kids") OR genreName CONTAINS("families") OR genreName
CONTAINS("children") THEN "Kids & Families"
  WHEN genreName CONTAINS("anime") THEN "Anime"
  WHEN genreName CONTAINS("classic") OR genreName CONTAINS("cult") THEN "Classic &
Cult"
  WHEN genreName CONTAINS("british") THEN "British TV"
  WHEN genreName CONTAINS("faith") OR genreName CONTAINS("spiritual") THEN "Faith &
Spirituality"
  WHEN genreName CONTAINS("independent") THEN "Independent"
  WHEN genreName CONTAINS("international") THEN "International"
  WHEN genreName CONTAINS("reality") THEN "Reality TV"
  WHEN genreName CONTAINS("sci-fi") OR genreName CONTAINS("fantasy") THEN "Sci-Fi &
Fantasy"
  WHEN genreName CONTAINS("lgbtq") THEN "LGBTQ"
  WHEN genreName CONTAINS("spanish") THEN "Spanish-Language TV"
  WHEN genreName CONTAINS("tv shows") OR genreName = "tv" THEN "TV (Allgemein)"
  WHEN genreName CONTAINS("movie") THEN "General Movies"
  ELSE "Andere"
END AS Genre, countryName

WHERE Genre IS NOT NULL

// Zähle pro Genre und Land
WITH Genre, countryName, count(*) AS countryCount

// Finde das häufigste Land pro Genre
WITH Genre, collect({country: countryName, count: countryCount}) AS countryStats
WITH Genre,
     reduce(s = {country: null, count: 0}, x IN countryStats |
       CASE WHEN x.count > s.count THEN x ELSE s END) AS topCountry

// Zähle gesamt pro Genre
MATCH (m)-[:HAS_GENRE]->(g:Genre)
WHERE (m:Movie OR m:TVShow)
  AND m.age IS NOT NULL
```

```
  AND m.age >= 16
  AND toLower(g.name) CONTAINS toLower(Genre)

WITH Genre, topCountry.country AS Top_Country, count(*) AS Count_Mature

RETURN Genre, Count_Mature, Top_Country
ORDER BY Count_Mature DESC;
```



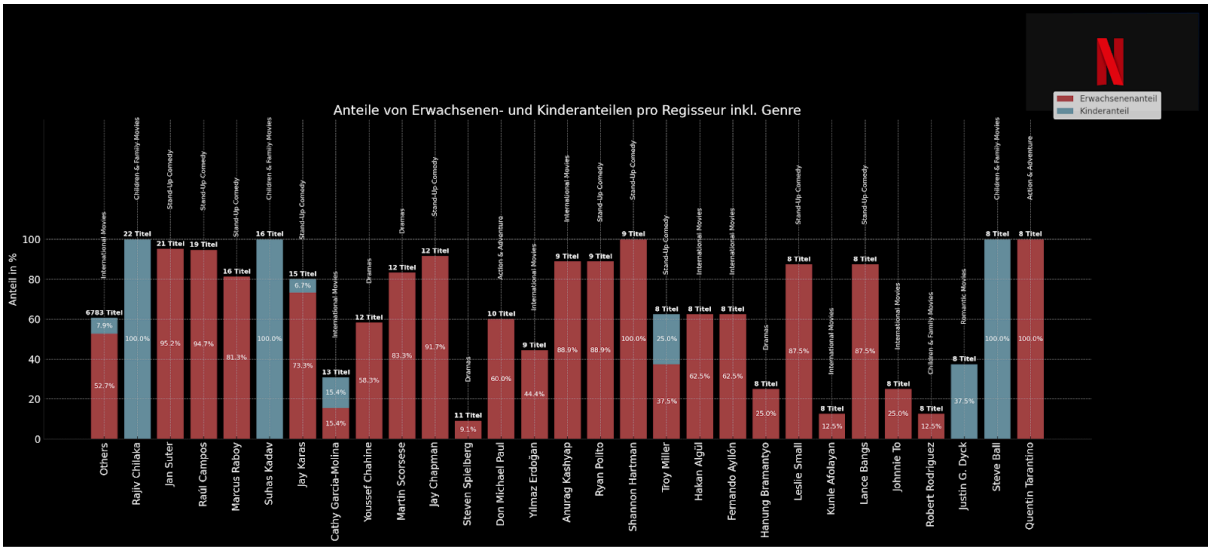

```
// Alle Inhalte mit Genre
MATCH (m)-[:HAS_GENRE]->(g:Genre)
WHERE (m:Movie OR m:TVShow)
WITH m, toLower(g.name) AS genreName

// Genre-Kategorien zusammenfassen
WITH m,
  CASE
    WHEN genreName CONTAINS "comedy" THEN "Comedy"
    WHEN genreName CONTAINS "tv comed" THEN "TV Comedy"
    WHEN genreName CONTAINS "horror" THEN "Horror"
    WHEN genreName CONTAINS "crime" OR genreName CONTAINS "mystery" THEN
"Crime & Mystery"
    WHEN genreName CONTAINS "thriller" THEN "Thriller"
    WHEN genreName CONTAINS "drama" THEN "Drama"
    WHEN genreName CONTAINS "action" OR genreName CONTAINS "adventure" THEN
"Action & Adventure"
    WHEN genreName CONTAINS "romance" OR genreName CONTAINS "romantic"
THEN "Romantic"
    WHEN genreName CONTAINS "documentary" OR genreName CONTAINS "docu" OR
genreName CONTAINS "science & nature" THEN "Documentary"
    WHEN genreName CONTAINS "teen" THEN "Teen"
    WHEN genreName CONTAINS "sport" THEN "Sports"
    WHEN genreName CONTAINS "music" THEN "Music"
```

```
    WHEN genreName CONTAINS "korean" THEN "Korean"
    WHEN genreName CONTAINS("kids") OR genreName CONTAINS("families") OR
genreName CONTAINS("children") THEN "Kids & Families"
    WHEN genreName CONTAINS("anime") THEN "Anime"
    WHEN genreName CONTAINS("classic") OR genreName CONTAINS("cult") THEN
"Classic & Cult"
    WHEN genreName CONTAINS("british") THEN "British TV"
    WHEN genreName CONTAINS("faith") OR genreName CONTAINS("spiritual") THEN
"Faith & Spirituality"
    WHEN genreName CONTAINS("independent") THEN "Independent"
    WHEN genreName CONTAINS("international") THEN "International"
    WHEN genreName CONTAINS("reality") THEN "Reality TV"
    WHEN genreName CONTAINS("sci-fi") OR genreName CONTAINS("fantasy") THEN
"Sci-Fi & Fantasy"
    WHEN genreName CONTAINS("lgbtq") THEN "LGBTQ"
    WHEN genreName CONTAINS("spanish") THEN "Spanish-Language TV"
    WHEN genreName CONTAINS("tv shows") OR genreName = "tv" THEN "TV
(Allgemein)"
    WHEN genreName CONTAINS("movie") THEN "General Movies"
    ELSE "Andere"
  END AS Genre, m.age AS age

// Gruppieren: Gesamtzahl + davon ab 16
WITH Genre,
     count(*) AS TotalInGenre,
     count(CASE WHEN age IS NOT NULL AND age >= 16 THEN 1 END) AS
Count_Mature

// Prozent berechnen
RETURN
  Genre,
  TotalInGenre,
  Count_Mature,
  round(toFloat(Count_Mature) * 100.0 / TotalInGenre, 2) AS PercentMature
ORDER BY PercentMature DESC;
```



```
// Schritt 1: hole alle relevanten Daten
MATCH (p:Person)-[:DIRECTED]->(t)
WHERE (t:Movie OR t:TVShow) AND t.age IS NOT NULL
OPTIONAL MATCH (t)-[:HAS_GENRE]->(g:Genre)

// Schritt 2: Vorbereitung
WITH
  p.name AS Director,
  t,
  toInteger(t.age) AS Rating,
  collect(DISTINCT g.name) AS Genres

// Schritt 3: Titel, Altersfreigaben und Genres pro Regisseur gruppieren
WITH
  Director,
  collect(DISTINCT t) AS Titles,
  collect(Rating) AS Ratings,
  reduce(allGenres = [], gList IN collect(Genres) | allGenres + gList) AS AllGenres

// Schritt 4: Gruppierung in "Others" wenn < 8 Titel
WITH
  CASE WHEN size(Titles) < 8 THEN "Others" ELSE Director END AS GroupedDirector,
  Titles,
  Ratings,
  AllGenres

// Schritt 5: Gruppiere alle Daten unter dem selben Director-Namen
WITH
  GroupedDirector AS Director,
  reduce(tFlat = [], tList IN collect(Titles) | tFlat + tList) AS AllTitles,
  reduce(rFlat = [], rList IN collect(Ratings) | rFlat + rList) AS AllRatings,
  reduce(gFlat = [], gList IN collect(AllGenres) | gFlat + gList) AS AllGenres

// Schritt 6: Auswertung
WITH
  Director,
  size(AllTitles) AS TotalTitles,
  size([r IN AllRatings WHERE r >= 16]) AS AdultCount,
  size([r IN AllRatings WHERE r <= 7]) AS ChildCount,
  AllGenres

WITH
  Director,
  TotalTitles,
  round(toFloat(AdultCount) * 100 / TotalTitles, 1) + " %" AS AdultShare,
  round(toFloat(ChildCount) * 100 / TotalTitles, 1) + " %" AS ChildShare,
  AllGenres

UNWIND AllGenres AS genre
WITH Director, TotalTitles, AdultShare, ChildShare, genre
WITH Director, TotalTitles, AdultShare, ChildShare, genre, count(*) AS GenreCount
```

```
WITH Director, TotalTitles, AdultShare, ChildShare,
    collect({genre: genre, freq: GenreCount}) AS GenreCounts

WITH Director, TotalTitles, AdultShare, ChildShare,
    reduce(best = {genre: "", freq: 0}, entry IN GenreCounts |
        CASE WHEN entry.freq > best.freq THEN entry ELSE best END
    ).genre AS MostFrequentGenre

RETURN Director, TotalTitles, AdultShare, ChildShare, MostFrequentGenre
ORDER BY TotalTitles DESC;
```



Auswertung:

Zur Beantwortung der Fragestellung wurde analysiert, in welchen Genres Inhalte mit der Altersfreigabe ab 16 Jahren auf Netflix besonders häufig vertreten sind. Grundlage der Analyse war ein gefilterter Datensatz, der ausschließlich Inhalte mit einer Altersfreigabe von mindestens 16 Jahren umfasste. Für jedes Genre wurde die absolute Anzahl dieser Inhalte ermittelt und visualisiert.

Die Ergebnisse zeigen, dass Inhalte mit der Altersfreigabe ab 16 Jahren am häufigsten in den Genres „International“, „Drama“ und „Comedy“ vorkommen. Besonders das Genre „International“ sticht dabei hervor: Es weist mit Abstand die höchste Anzahl an 16+-Inhalten auf. Die Analyse zeigt, dass diese Inhalte überwiegend aus Indien stammen, was auf eine regionale Schwerpunktbildung innerhalb des Genres hinweist.

Auch die Genres „Drama“ und „Comedy“ sind stark mit 16+-Inhalten belegt. In beiden Fällen dominiert die US-amerikanische Produktion, die insgesamt auch in anderen Genre-Kategorien einen Großteil der Inhalte mit dieser Freigabe liefert. Dies gilt insbesondere für genrespezifisch anspruchsvollere oder emotional intensivere Formate.

Weitere Genres mit auffällig vielen Inhalten ab 16 Jahren sind „Independent“, „Action & Adventure“, „Documentary“, „Thriller“, „Crime & Mystery“ sowie „Horror“. Diese Genres zeichnen sich durch inhaltliche Komplexität, dramatische Zuspitzungen oder explizite Darstellungen aus und weisen insgesamt eine erhöhte Einstufung hinsichtlich Jugendschutz auf.

Am unteren Ende der Skala stehen familien- und kinderorientierte Kategorien wie „Kids & Families“, „Faith & Spirituality“, „TV (Allgemein)“ und „General Movies“, die nur sehr wenige Inhalte mit Altersfreigabe ab 16 Jahren enthalten. Diese Genres richten sich vorrangig an ein jüngeres oder allgemeinpublikumstaugliches Zielpublikum und weisen entsprechend niedrigere Freigaben auf.

Auffällig ist zudem, dass auch in spezifischen Nischengenres wie „LGBTQ“, „Reality TV“ und „Anime“ eine nennenswerte Zahl an 16+-Inhalten vorhanden ist. Auch in diesen Fällen dominieren häufig die USA als Produktionsland, was auf eine starke Rolle des US-amerikanischen Marktes bei der Bereitstellung jugendschutzrelevanter Inhalte hinweist.

Die Ergebnisse belegen somit, dass Inhalte mit Altersfreigabe „16+“ auf Netflix besonders häufig in genrespezifisch ernsten, dramatischen oder spannungsbasierten Formaten vertreten sind, wobei der geografische Schwerpunkt überwiegend in den Vereinigten Staaten liegt – mit Ausnahme des internationalen Genres, das stark von indischen Produktionen geprägt ist.

Mögliche Ursachen für die Häufung von 16+-Inhalten in bestimmten Genres lassen sich in mehreren strukturellen und inhaltlichen Faktoren verorten. Zum einen gehen viele dieser Genres mit sensiblen Themen einher, etwa Gewalt, Sexualität, Drogenmissbrauch oder gesellschaftliche Konflikte, die eine höhere Alterseinstufung erforderlich machen. Zum anderen sind insbesondere indische Produktionen im internationalen Genre häufig länger und intensiver gestaltet, was mit narrativen und stilistischen Besonderheiten einhergeht, die eine höhere Freigabe rechtfertigen. Die Dominanz US-amerikanischer Produktionen in vielen dieser Genres verstärkt diesen Trend zusätzlich, da sie weltweit eine bedeutende Rolle bei der Verbreitung emotional aufgeladener, spannungsreicher oder kontroverser Inhalte spielen. Hinzu kommt, dass jugendschutzrechtliche Regelungen in bestimmten Genrebereichen systematisch zu höheren Einstufungen führen, unabhängig vom konkreten Inhalt. Schließlich sind viele dieser Genres besonders gut auf die Logik von Streaming-Plattformen zugeschnitten: Sie sprechen erwachsene Zielgruppen gezielt an, lassen sich gut in Empfehlungsalgorithmen integrieren und bieten Raum für kreative und gesellschaftlich relevante Inhalte abseits des Mainstreams.

Erläuterung:

Die aufgestellten Hypothesen ermöglichen eine differenzierte Analyse der strategischen Inhaltsgestaltung von Netflix. In den Fokus rücken dabei zentrale Dimensionen wie der **zeitliche Kontext** (Veröffentlichungsjahr, Filmlänge), die **geografische Herkunft** (Produktionsland bzw. -region), das **Format** (Film oder Serie), die **adressierte Zielgruppe** (Altersfreigabe) sowie **thematische Aspekte** (Genre).

Daraus ergibt sich folgende übergeordnete Forschungsfrage:
Wie lässt sich das weltweite Inhaltsangebot von Netflix differenziert nach Ländern/ Regionen beschreiben und bewerten?

Die Analyseergebnisse geben Aufschluss darüber, inwieweit Netflix sein Programmangebot an **regionale Gegebenheiten** anpasst und welche Inhalte in spezifischen Märkten besonders stark vertreten sind. Daraus lassen sich Rückschlüsse

auf strategische Ausrichtungen, kulturelle Positionierungen und marktorientierte Programmplanung ziehen.

Gesamtauswertung:
Das weltweite Inhaltsangebot von Netflix erweist sich als mehrdimensional und differenziert. Es lässt sich entlang verschiedener Analyseebenen untersuchen und zeigt eine klare Anpassung an regionale Kontexte.

Format (Serie/Film):
Die Verteilung zwischen Serien und Filmen variiert deutlich je nach Region. In ostasiatischen Ländern wie Südkorea, Japan und Taiwan dominieren Serien, was auf eine ausgeprägte Dramakultur und serielle Erzähltradition hinweist. In Afrika sowie in Teilen Südasiens (z. B. Indien, Nigeria, Indonesien) hingegen überwiegen Filme – ein Effekt historisch gewachsener Filmindustrien. Netflix steuert sein Formatangebot gezielt nach lokalen Produktionsbedingungen und Sehgewohnheiten aus.

Filmlänge:
Die durchschnittliche Länge von Filmen unterscheidet sich deutlich zwischen einzelnen Produktionsländern. Länder wie Indien und Ägypten produzieren im Schnitt längere Filme (über 110 Minuten), während Länder wie Kanada, Mexiko und Brasilien kürzere Formate bevorzugen (teils unter 95 Minuten). Diese Unterschiede spiegeln kulturelle Erzählstrukturen und nationale Produktionsstandards wider.

Altersfreigaben:
Die durchschnittliche Altersfreigabe variiert zwischen den Kontinenten. Afrika und Südamerika weisen die höchsten Freigaben auf, während Asien, Nordamerika und Ozeanien deutlich niedrigere Werte zeigen – insbesondere bei Serien. Europa liegt im mittleren Bereich. Diese Unterschiede lassen sich sowohl auf gesetzliche Regulierungen als auch auf gesellschaftliche Normen und Inhaltspräferenzen zurückführen.

Temporale Entwicklung:
Ein langfristiger Abwärtstrend in der durchschnittlichen Filmlänge ist erkennbar. Während Filme in den 1950er- und 1960er-Jahren teils deutlich über 120 Minuten lagen, zeigt sich ab den 2000er-Jahren eine kontinuierliche Verkürzung, mit Durchschnittslängen unter 100 Minuten in den letzten Jahren. Diese Entwicklung lässt sich mit veränderten Konsumgewohnheiten erklären – etwa durch mobile Nutzung, kürzere Aufmerksamkeitsspannen sowie die zunehmende Nachfrage nach kompakten, schnell konsumierbaren Formaten.

Genrezuordnung bei Altersfreigaben:
Inhalte mit einer Altersfreigabe ab 16 Jahren sind besonders häufig in genrespezifisch ernsten, komplexen oder spannungsreichen Kategorien vertreten. Zu den am stärksten betroffenen Genres zählen Drama, Thriller, Crime, Action & Adventure sowie das Genre „International“. In letzterem dominieren indische Produktionen, während in den meisten anderen Genres die USA als führendes Produktionsland auftreten. Auch spezifischere Nischengenres wie LGBTQ, Reality TV und Anime tragen zur Vielfalt innerhalb dieser Kategorie bei.

Fazit:
Inhalte mit einer Altersfreigabe ab 16 Jahren sind besonders häufig in genrespezifisch ernsten, komplexen oder spannungsreichen Kategorien vertreten. Zu den am stärksten betroffenen Genres zählen Drama, Thriller, Crime, Action & Adventure sowie das Genre

„International“. In letzterem dominieren indische Produktionen, während in den meisten anderen Genres die USA als führendes Produktionsland auftreten. Auch spezifischere Nischengenres wie LGBTQ, Reality TV und Anime tragen zur Vielfalt innerhalb dieser Kategorie bei.

CYPHER IMPORT CODES

Alle **rot und fett hervorgehobenen Abfragen** stehen in direktem Zusammenhang mit den formulierten Hypothesen. Die übrigen rot markierten Abfragen sind ausschließlich aus Gründen der Vollständigkeit aufgeführt. **Hell rot (und fett hervorgehobene) Abfragen** wurden im ersten Teil erstellt, jedoch als unwichtig empfunden und entfernt/ umgeändert. Die folgenden CRUD Operationen sind wie gefolgt gekennzeichnet:

Create
Read
Create oder Read
Update
Delete

```
// 1. Zuerst Filme und TV-Shows importieren (mit entsprechenden Labels)
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.show_id IS NOT NULL
CREATE (t:Title {
  show_id: line.show_id,
  title: line.title,
  release_year: toInteger(line.releaseyear),
  rating: toInteger(line.rating),
  duration: line.duration,
  description: line.description
})
WITH t, line
FOREACH (_ IN CASE WHEN line.type = 'Movie' THEN [1] ELSE [] END | SET t:Movie)
FOREACH (_ IN CASE WHEN line.type = 'TV Show' THEN [1] ELSE [] END | SET t:TVShow)
RETURN count(*);
```

Zunächst erfolgt der Import sämtlicher Filme und Serien aus der bereitgestellten CSV-Datei mittels des Befehls LOAD CSV. Jede Zeile, die eine gültige show_id enthält, wird in einen Knoten mit dem Label Title überführt. In diesem Schritt werden zentrale Attribute wie Titel, Erscheinungsjahr, Altersfreigabe, Laufzeit und Beschreibung ausgelesen und als Eigenschaften gespeichert. Dabei wird sichergestellt, dass das Erscheinungsjahr und die Altersfreigabe in ganzzahlige Werte umgewandelt werden, um eine korrekte Datentypisierung zu gewährleisten.

Anschließend erfolgt eine Differenzierung nach Typ: Basierend auf dem Attribut type wird jeder Titel zusätzlich entweder mit dem Label Movie oder TV Show versehen. Diese Unterscheidung ist essentiell für spätere inhalts- oder Format Bezogene Analysen und bildet die Grundlage für eine strukturierte semantische Erschließung des Netflix-Datenbestands.

```
// 2. Länder importieren und PRODUCED_IN-Beziehung erstellen
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.country IS NOT NULL AND line.country <> ''
MATCH (t:Title {show_id: line.show_id})
WITH line, t, split(line.country, ',') AS countries
UNWIND countries AS country_name
MERGE (c:Country {name: trim(country_name)})
MERGE (t)-[:PRODUCED_IN]->(c)
RETURN count(*);
```

Im zweiten Schritt werden Informationen zu den Produktionsländern der Titel aus der ursprünglichen CSV-Datei extrahiert und in das Datenmodell integriert. Hierzu wird zunächst überprüft, ob für einen Datensatz ein gültiger Eintrag im Feld Country vorliegt. Für alle entsprechenden Einträge wird der zugehörige Titel-Knoten anhand der Show_id identifiziert und mit einem oder mehreren Country-Knoten verknüpft.

Da einzelne Titel mehreren Ländern zugeordnet sein können, wird das Feld country mit Hilfe der split-Funktion anhand des Kommas separiert und anschließend über UNWIND in Einzeleinträge zerlegt. Für jedes Land wird (unter Vermeidung von Duplikaten mittels MERGE) ein eigener Country-Knoten erzeugt, dessen Name vor der Speicherung getrimmt wird, um etwaige führende oder nachgestellte Leerzeichen zu eliminieren. Anschließend wird zwischen dem jeweiligen Titel und dem zugehörigen Land eine PRODUCED IN-Beziehung hergestellt.

Diese Modellierung ermöglicht eine differenzierte Analyse der geographischen Herkunft der Inhalte im Katalog. Es wurde jedoch festgestellt, dass rund 830 Titel keine Angabe zum Produktionsland enthielten. Diese Lücke wurde durch eine ergänzende Abfrage identifiziert und stellt ein potenzielles Defizit in der Ausgangsdatengrundlage dar.

```
//2.1 Abfrage nach Anzahl fehlender Länder
MATCH (t)
WHERE (t:Movie OR t:TVShow) AND NOT (t)-[:PRODUCED_IN]->(:Country)
RETURN count(t) AS missing_country_count
```

Durch eine gezielte Abfrage wurde festgestellt, welche Titel – trotz des initialen Import – keine Verknüpfung zu einem Country-Knoten besitzen. Dies betraf zunächst rund 830 Titel. Zur Verbesserung der Datenqualität wurde ein ergänzendes Python-Skript erstellt, das mit Unterstützung von ChatGPT entwickelt wurde. Dieses nutzt WikiData als externe Datenquelle, um fehlende Länderinformationen anhand der Titel zu recherchieren und dem Datensatz hinzuzufügen. Trotz dieser Maßnahme verblieben jedoch noch etwa 480 Titel ohne Herkunftsangabe.

```
//2.2 India nach Titeln auffüllen, um Lücken zu reduzieren
MATCH (t:Movie|TVShow)
WHERE NOT (t)-[:PRODUCED_IN]->(:Country)
  AND (
    toLower(t.title) CONTAINS "singham" OR
```

```
    toLower(t.title) CONTAINS "telugu" OR
    toLower(t.title) CONTAINS "tamil" OR
    toLower(t.title) CONTAINS "bheem" OR
    toLower(t.title) CONTAINS "kushh" OR
    toLower(t.title) CONTAINS "boomika" OR
    toLower(t.title) CONTAINS "bhole" OR
    toLower(t.title) CONTAINS "malayalam" OR
    toLower(t.title) CONTAINS "ishq" OR
    toLower(t.title) CONTAINS "patlu" OR
    toLower(t.title) CONTAINS "rudra" OR
    toLower(t.title) CONTAINS "shiva" OR
    toLower(t.title) CONTAINS "maharaja" OR
    toLower(t.title) CONTAINS "baahubali" OR
    toLower(t.title) CONTAINS "rakhwale" OR
    toLower(t.title) CONTAINS "bheemayan" OR
    toLower(t.title) CONTAINS "bure kaam" OR
    toLower(t.title) CONTAINS "jugaad" OR
    toLower(t.title) CONTAINS "krish" OR
    toLower(t.title) CONTAINS "krishna" OR
    toLower(t.title) CONTAINS "jai aur" OR
    toLower(t.title) CONTAINS "satrangi" OR
    toLower(t.title) CONTAINS "swami baba"
  )
MERGE (c:Country {name: "India"})
MERGE (t)-[:PRODUCED_IN]->(c)
```

Zur weiteren Reduktion der verbleibenden Titel ohne Länderzuordnung wurde ein ergänzendes Verfahren zur inhaltlichen Mustererkennung angewendet. Dabei wurden charakteristische Begriffe identifiziert, die mit hoher Wahrscheinlichkeit auf eine indische Herkunft des jeweiligen Titels schließen lassen (beispielsweise „telugu“, „tamil“, „krishna“). Titel, deren Bezeichnung solche Merkmale enthielt, wurden systematisch dem Land „India“ zugeordnet. Durch diese inhaltlich motivierte Klassifikation konnte die Anzahl der Titel ohne Herkunftsangabe um weitere rund 110 Einträge reduziert werden.

//2.3 Country-Import durch eine weitere hilfsdatei

```
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/blob/master/Titel_nach_Produktionsland.csv" AS row
WITH trim(row.Titel) AS title, trim(row.Produktionsland) AS country
MATCH (t)
WHERE (t:Movie OR t:TVShow) AND t.title = title
MERGE (c:Country {name: country})
MERGE (t)-[:PRODUCED_IN]->(c)
```

Zur abschließenden Vervollständigung der Länderinformationen wurde eine separate CSV-Datei mit manueller Zuweisung von Produktionsländern erstellt. Diese Datei enthielt Titel, die trotz vorheriger Verfahren keine gültige Länderinformation besaßen. Durch die Verwendung dieser Datei konnten die verbleibenden Lücken geschlossen werden, sodass am Ende alle Titel im Datensatz mindestens einem Herkunftsland zugeordnet werden konnten.

```
//2.4 Herausfiltern von Ländern mit Problemen
MATCH (c:Country)
WHERE trim(c.name) = "" OR c.name CONTAINS ","
RETURN c.name AS ProblematischerName
ORDER BY c.name
```

Im Anschluss an die vollständige Integration der Länderinformationen wurde eine Validierung der Country-Knoten durchgeführt. Dabei wurden zwei Fehlerkategorien identifiziert: Knoten mit leerem Namen sowie solche, die mehrere Länderbezeichnungen in einem Feld enthielten (z. B. „France, Germany“). Diese problematischen Einträge wurden für eine nachfolgende Korrektur extrahiert.

```
//2.5 Eine Trennung all dieser Länder, die vorher nicht richtig getrennt wurden
MATCH (c:Country)
WHERE c.name CONTAINS ","
WITH c, split(c.name, ",") AS names
UNWIND names AS rawName
WITH c, trim(rawName) AS name
MERGE (newC:Country {name: name})
WITH c, newC
MATCH (n)-[r:PRODUCED_IN]->(c)
MERGE (n)-[:PRODUCED_IN]->(newC)
DELETE r
DELETE c
```

Knoten mit mehrfachen Ländereinträgen wurden im Rahmen eines Bereinigungsschritts getrennt. Dabei wurden die Einträge auf Basis des Kommas gesplittet und jeder Teilbegriff in einen neuen Knoten überführt. Anschließend wurden die bestehenden Beziehungen des Titels auf die neu erzeugten Länderknoten übertragen. Der ursprüngliche fehlerhafte Knoten sowie seine Beziehungen wurden gelöscht. Dadurch konnte die Datenbankstruktur semantisch konsistent gehalten werden.

```
//2.6 Löschen alle Country Knoten mit leeren Namen
MATCH (t)-[r:PRODUCED_IN]->(c:Country)
WHERE trim(c.name) = ""
DELETE r
WITH c
DETACH DELETE c
```

Als abschließender Schritt wurden alle Länder-Knoten mit leerem Namen inklusive ihrer Beziehungen entfernt. Dieser Schritt sichert die formale Integrität des Graphenmodells und verhindert fehlerhafte oder unbrauchbare Abfragen auf Basis leerer Entitäten.

```
// 3. Genres importieren und LISTED_IN-Beziehung erstellen
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.listed_in IS NOT NULL AND line.listed_in <> ''
MATCH (t:Title {show_id: line.show_id})
WITH line, t, split(line.listed_in, ',') AS genres
UNWIND genres AS genre_name
```

```
MERGE (g:Genre {name: trim(genre_name)})
MERGE (t)-[:CATEGORIZED_AS]->(g)
RETURN count(*);
```

Zur semantischen Erschließung der inhaltlichen Kategorisierung jedes Titels wurde eine separate Genre-Struktur aufgebaut. Die Informationen aus dem Feld „listed_in“ der CSV-Datei wurden dabei in einzelne Einträge aufgeteilt, da mehrere Genres häufig in einem Eintrag gemeinsam gespeichert waren. Durch eine Trennung an den Kommas konnte eine exakte Erfassung und Zuordnung gewährleistet werden. Für jedes Genre wurde ein eigener Knoten erstellt, der anschließend über eine Beziehung mit dem jeweiligen Titel verknüpft wurde.

```
//4. Auf Netflix geaddet
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.date_added IS NOT NULL AND line.date_added <> ''
MATCH (t:Title {show_id: line.show_id})
SET t.date_added = line.date_added
RETURN count(*);
```

In einem weiteren Schritt wurde für alle Inhalte das Datum ergänzt, an dem sie dem Netflix-Katalog hinzugefügt wurden. Diese Informationen wurden ebenfalls aus der ursprünglichen CSV-Datei entnommen und direkt als Attribut in die bereits bestehenden Titelknoten eingefügt. Obwohl diese Information im Rahmen der vorliegenden Hypothesen keine zentrale Rolle spielt, wurde sie der Vollständigkeit halber in das Modell integriert.

```
//5. Age_restriction mit der Beziehung HAS_AGERATING
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.age_restriction IS NOT NULL AND trim(line.age_restriction) <> ''
MATCH (t:Title)
WHERE trim(t.show_id) = trim(line.show_id)
MERGE (r:AgeRating {rating: trim(line.age_restriction)})
MERGE (t)-[:HAS_AGERATING]->(r)
RETURN count(*);
```

Zur Abbildung der Altersfreigaben wurde für jeden eindeutigen Eintrag im Feld „age_restriction“ ein eigener Knoten angelegt und mit dem jeweiligen Titel durch eine entsprechende Beziehung verknüpft. Da der ursprüngliche Datensatz in diesem Bereich unvollständig war, wurde ein ergänzendes Python-Skript entwickelt, welches mithilfe von WikiData zusätzliche Altersfreigaben bereitstellte. Diese Daten wurden in einer zweiten Datei verarbeitet und in das bestehende Modell integriert.

```
//5.1 Dataset updaten mit Hilfe von imdb Datei
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/imbd.csv" AS line
MATCH (m:Movie)
WHERE toLower(trim(m.title)) = toLower(trim(line.title))
  AND (m.age_restriction IS NULL OR m.age_restriction = '')
```

```
SET m.age_restriction = line.certificate
```

Um weitere Lücken bei den Altersfreigaben zu schließen, wurde zusätzlich eine externe IMDb-Datei herangezogen. Für alle Filme, die keine Altersfreigabe aufwiesen, jedoch in der IMDb-Datei identifiziert werden konnten, wurde das Feld mithilfe des Zertifikatswertes aus dieser Quelle ergänzt.

Ein ergänzendes Python-Import-Skript, das mit Unterstützung von Chat GPT erstellt wurde, dient der Vervollständigung des Datensatzes durch die Anreicherung fehlender Alters Informationen auf Basis von WikiData.

```
//5.2. Agerating in Jahre umwandeln
MATCH (a:AgeRating {rating: 0}) SET a.age = 0;
MATCH (a:AgeRating {rating: 1}) SET a.age = 2;
MATCH (a:AgeRating {rating: 2}) SET a.age = 4;
MATCH (a:AgeRating {rating: 3}) SET a.age = 7;
MATCH (a:AgeRating {rating: 4}) SET a.age = 8;
MATCH (a:AgeRating {rating: 5}) SET a.age = 10;
MATCH (a:AgeRating {rating: 6}) SET a.age = 13;
MATCH (a:AgeRating {rating: 7}) SET a.age = 14;
MATCH (a:AgeRating {rating: 8}) SET a.age = 16;
MATCH (a:AgeRating {rating: 9}) SET a.age = 17;
MATCH (a:AgeRating {rating: 10}) SET a.age = 18;
```

Die ursprünglichen numerischen Ratingwerte (0 bis 10) wurden im Zuge der Visualisierung und Analyse in konkrete Altersangaben in Jahren umgerechnet. Diese Transformation dient der besseren Interpretierbarkeit innerhalb der nachfolgenden Auswertungen.

```
//5.3 Age_Rating Knoten als Properties setzen
// Für Movie
MATCH (m:Movie)-[:HAS_AGERATING]->(r:AgeRating)
SET m.age_rating = r.rating;

// Für TVShow
MATCH (s:TVShow)-[:HAS_AGERATING]->(r:AgeRating)
SET s.age_rating = r.rating;

//5.4 Age in Jahren als Properties setzen
MATCH (m:Movie)
WHERE m.age_rating IS NOT NULL
SET m.age =
  CASE m.age_rating
    WHEN 0 THEN 0
    WHEN 1 THEN 2
    WHEN 2 THEN 4
    WHEN 3 THEN 7
    WHEN 4 THEN 8
    WHEN 5 THEN 10
    WHEN 6 THEN 13
    WHEN 7 THEN 14
```

```
    WHEN 8 THEN 16
    WHEN 9 THEN 17
    WHEN 10 THEN 18
  END;

MATCH (t:TVShow)
WHERE t.age_rating IS NOT NULL
SET t.age =
  CASE t.age_rating
    WHEN 0 THEN 0
    WHEN 1 THEN 2
    WHEN 2 THEN 4
    WHEN 3 THEN 7
    WHEN 4 THEN 8
    WHEN 5 THEN 10
    WHEN 6 THEN 13
    WHEN 7 THEN 14
    WHEN 8 THEN 16
    WHEN 9 THEN 17
    WHEN 10 THEN 18
  END;
```

Die Altersfreigabe wurde zuvor als ein Node behandelt. In der neuen Version wird sie direkt als Property gespeichert zur besseren Lesbarkeit. Dadurch bleibt der Graph übersichtlicher und die Informationen sind direkt als properties verfügbar

```
//6 Duration updaten mit der Beziehung HAS_DURATION
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.duration IS NOT NULL AND trim(line.duration) <> ''
MATCH (t:Title {show_id: trim(line.show_id)})
WITH t, trim(line.duration) AS raw_duration
// Filme (Minuten)
FOREACH (_ IN CASE WHEN raw_duration CONTAINS 'min' THEN [1] ELSE [] END |
  MERGE (d:Duration {type: "minutes", value: toInteger(split(raw_duration, ' ')[0])})
  MERGE (t)-[:HAS_DURATION]->(d)
)

// Serien (Staffeln)
FOREACH (_ IN CASE WHEN raw_duration CONTAINS 'Season' THEN [1] ELSE [] END |
  MERGE (d:Duration {type: "seasons", value: toInteger(split(raw_duration, ' ')[0])})
  MERGE (t)-[:HAS_DURATION]->(d)
);

6.1 //Type und value verknüpfen um beides stehen zu haben
MATCH (d:Duration)
SET d.label = d.value + " " + d.type;
```

Die im Feld „duration“ enthaltenen Angaben wurden anhand ihrer Struktur in zwei Kategorien unterteilt: Filme mit einer Angabe in Minuten und Serien mit einer Angabe in Staffeln. Für jede dieser Kategorien wurde ein Knoten erzeugt, der sowohl den numerischen Wert (etwa 90 für Minuten oder 2 für Staffeln) als auch die Einheit als Typ (z. B. „minutes“ oder „seasons“) beinhaltet. Anschließend wurde eine zusätzliche Beschriftung in Form eines kombinierten Labels (z. B. „90 minutes“) eingefügt, um eine konsistente Darstellung in Visualisierungen zu ermöglichen.

```
// 6.2 Duration als Properties speichern
// Für Movie
MATCH (m:Movie)-[:HAS_DURATION]->(d:Duration)
SET m.duration_value = d.value,
    m.duration_type = d.type,
    m.duration_label = d.label;

// Für TVShow
MATCH (s:TVShow)-[:HAS_DURATION]->(d:Duration)
SET s.duration_value = d.value,
    s.duration_type = d.type,
    s.duration_label = d.label;
```

Die Dauer wurde zuvor als ein Node behandelt. In der neuen Version wird sie direkt als Property gespeichert – aufgeteilt in duration_value, duration_type (z. B. Minuten oder Staffeln) und duration_label zur besseren Lesbarkeit. Dadurch bleibt der Graph übersichtlicher und die Dauer Informationen sind direkt als properties verfügbar.

```
//7. Release Year mit Beziehung RELEASED_IN
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.release_year IS NOT NULL AND trim(line.release_year) <> ''
MATCH (t:Title {show_id: line.show_id})
MERGE (y:ReleaseYear {year: toInteger(trim(line.release_year))})
MERGE (t)-[:RELEASED_IN]->(y)
RETURN count(*);
```

Für analytische Zwecke, insbesondere zur Durchführung zeitbezogener Auswertungen, wurde das Veröffentlichungsjahr jedes Titels als separater Knoten gespeichert und mit dem Titel über eine entsprechende Beziehung verknüpft. Die Jahrgänge wurden vorab in Integer-Werte umgewandelt, um die numerische Verarbeitung zu gewährleisten. Diese Struktur erlaubt eine gezielte Aggregation und Filterung nach dem Erscheinungsjahr.

```
// 7.1 Erscheinungsjahr als Properties direkt speichern
// Für Movie
MATCH (m:Movie)-[:RELEASED_IN]->(y:ReleaseYear)
SET m.release_year = y.year;

// Für TVShow
MATCH (s:TVShow)-[:RELEASED_IN]->(y:ReleaseYear)
SET s.release_year = y.year;
```

Der ursprüngliche Code hat für jedes Veröffentlichungsjahr (release year) eine eigene Node erstellt und diese mit den Titeln verknüpft. In der überarbeiteten Version wird das Veröffentlichungsjahr stattdessen direkt als Property zur bestehenden Nodes hinzugefügt. Dadurch wird der Graph vereinfacht und die Information direkt gespeichert.

//5.4, 6.3, 7.2 Löschen von Beziehung und Knoten von AgeRating, Duration und ReleaseYear

```
MATCH (:Movie)-[r:HAS_AGERATING|HAS_DURATION|RELEASED_IN]->()
DELETE r;

MATCH (:TVShow)-[r:HAS_AGERATING|HAS_DURATION|RELEASED_IN]->()
DELETE r;

MATCH (r:AgeRating) DETACH DELETE r;
MATCH (d:Duration) DETACH DELETE d;
MATCH (y:ReleaseYear) DETACH DELETE y;
```

Die Knoten und Beziehungen zu Erscheinungsjahr, Altersfreigabe und Dauer werden gelöscht.

// 8. Direktoren importieren und DIRECTED-Beziehung erstellen

```
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line WHERE line.director IS NOT NULL AND line.director <> ''
MATCH (t:Title {show_id: line.show_id})
WITH line, t, split(line.director, ',') AS directors
UNWIND directors AS director_name
MERGE (d:Person {name: trim(director_name)})
MERGE (d)-[:DIRECTED]->(t)
RETURN count(*);
```

Zur Vervollständigung der Datenbank wurden Informationen zu Regisseuren verarbeitet, obwohl diese für die zentrale Fragestellung keine unmittelbare Relevanz besitzen. Die im Feld „director“ der CSV-Datei enthaltenen Einträge wurden anhand von Kommata aufgesplittet und für jede Person ein Knoten erstellt. Anschließend wurde eine Beziehung zum jeweiligen Titel etabliert, wodurch auch eine spätere netzwerkanalytische Auswertung möglich ist.

//8.1 More Director hinzufügen

```
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles_filled_with_imdb_final.csv?ref_type=heads" AS row
WITH row
WHERE row.director IS NOT NULL AND row.director <> ''
MERGE (p:Person {name: row.director})
WITH row, p
MATCH (m:Movie {title: row.title})
WHERE NOT ( (:Person)-[:DIRECTED]->(m) )
MERGE (p)-[:DIRECTED]->(m)
```

Durch eine gezielte Cypher-Abfrage wurden Titel identifiziert, denen trotz vorheriger Datenanreicherung weiterhin keine Regieperson zugeordnet war. Zur Schließung dieser Lücken wurde eine ergänzte CSV-Datei importiert, die durch externe API-Recherchen mit TMDb und OMDb angereicherte Regieangaben enthält. Mithilfe eines Skripts wurde daraufhin ein neuer *Person*-Knoten erstellt und – sofern noch keine Verbindung bestand – entsprechende *DIRECTED*-Beziehungen zu den zugehörigen Filmen angelegt.

```
// 9. Schauspieler importieren und ACTED_IN-Beziehung erstellen
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/netflix_titles.csv" AS line
FIELDTERMINATOR ','
WITH line
WHERE line.cast IS NOT NULL AND line.cast <> ''
MATCH (t:Title {show_id: line.show_id})
WITH t, split(line.cast, ',') AS castList
UNWIND castList AS rawName
WITH t, trim(rawName) AS actorName
MERGE (a:Person {name: actorName})
MERGE (a)-[:ACTED_IN]->(t)
```

Analog zu den Regisseuren wurden auch die im Feld „cast“ gelisteten Schauspielerinnen und Schauspieler in die Datenbank aufgenommen. Die Einträge wurden getrennt, bereinigt und jeweils als eigener Personenknoten modelliert. Für jede mitwirkende Person wurde eine Beziehung zum entsprechenden Titel hergestellt. Auch diese Information dient primär der Vollständigkeit und wird nicht direkt in den Hypothesen verwendet.

```
//10. Personen finden, die zusammengearbeitet haben (Co-Working-Beziehung)
MATCH (p1:Person)-[:ACTED_IN]->(t)<-[:ACTED_IN]-(p2:Person)
WHERE p1.name < p2.name  // Vermeidet Duplikate
MERGE (p1)-[r:WORKED_WITH]->(p2)
ON CREATE SET r.count = 1
ON MATCH SET r.count = r.count + 1
RETURN count(*);
```

Basierend auf den Schauspiel-Informationen wurde eine Netzwerkanalyse der Zusammenarbeit zwischen Personen ermöglicht. Alle Personen, die gemeinsam in einem Titel mitgewirkt haben, wurden über eine Beziehung verknüpft. Um doppelte Verbindungen zu vermeiden, wurde die Bedingung eingeführt, dass die alphabetisch erste Person immer am linken Rand der Beziehung steht. Die Beziehung wurde zudem mit einem Zählwert versehen, der angibt, wie häufig diese Zusammenarbeit aufgetreten ist.

//11. Kontinente hinzufügen; da hierzu kein Dataset auffindbar war, ließen wir ChatGPT dies machen

```
UNWIND [
  ["Afghanistan", "Asia"],
  ["Albania", "Europe"],
  ["Algeria", "Africa"],
  ["Angola", "Africa"],
  ["Argentina", "South America"],
  ["Armenia", "Asia"],
```

["Australia", "Oceania"],
["Austria", "Europe"],
["Azerbaijan", "Asia"],
["Bahamas", "North America"],
["Bangladesh", "Asia"],
["Belarus", "Europe"],
["Belgium", "Europe"],
["Bermuda", "North America"],
["Botswana", "Africa"],
["Brazil", "South America"],
["Bulgaria", "Europe"],
["Burkina Faso", "Africa"],
["Cambodia", "Asia"],
["Cameroon", "Africa"],
["Canada", "North America"],
["Cayman Islands", "North America"],
["Chile", "South America"],
["China", "Asia"],
["Colombia", "South America"],
["Congo", "Africa"],
["Croatia", "Europe"],
["Cuba", "North America"],
["Cyprus", "Asia"],
["Czech Republic", "Europe"],
["Denmark", "Europe"],
["Dominican Republic", "North America"],
["Ecuador", "South America"],
["Egypt", "Africa"],
["Estonia", "Europe"],
["Ethiopia", "Africa"],
["Finland", "Europe"],
["France", "Europe"],
["Georgia", "Asia"],
["Ghana", "Africa"],
["Greece", "Europe"],
["Guatemala", "North America"],
["Hongkong", "Asia"],
["Hungary", "Europe"],
["Iceland", "Europe"],
["India", "Asia"],
["Indonesia", "Asia"],
["Iran", "Asia"],
["Iraq", "Asia"],
["Ireland", "Europe"],
["Israel", "Asia"],
["Italy", "Europe"],
["Jamaica", "North America"],
["Japan", "Asia"],
["Jordan", "Asia"],
["Kazakhstan", "Asia"],

["Kenya", "Africa"],
["Kuwait", "Asia"],
["Latvia", "Europe"],
["Lebanon", "Asia"],
["Liechtenstein", "Europe"],
["Lithuania", "Europe"],
["Luxembourg", "Europe"],
["Malawi", "Africa"],
["Malaysia", "Asia"],
["Malta", "Europe"],
["Mauritius", "Africa"],
["Mexico", "North America"],
["Mongolia", "Asia"],
["Montenegro", "Europe"],
["Morocco", "Africa"],
["Mozambique", "Africa"],
["Namibia", "Africa"],
["Nepal", "Asia"],
["Netherlands", "Europe"],
["New Zealand", "Oceania"],
["Nicaragua", "North America"],
["Nigeria", "Africa"],
["Norway", "Europe"],
["Pakistan", "Asia"],
["Pala Empire", "Asia"],
["Palestine", "Asia"],
["Panama", "North America"],
["Paraguay", "South America"],
["Peru", "South America"],
["Philippines", "Asia"],
["Poland", "Europe"],
["Portugal", "Europe"],
["Puerto Rico", "North America"],
["Qatar", "Asia"],
["Romania", "Europe"],
["Russia", "Europe"],
["Samoa", "Oceania"],
["Saudi Arabia", "Asia"],
["Senegal", "Africa"],
["Serbia", "Europe"],
["Singapore", "Asia"],
["Slovakia", "Europe"],
["Slovenia", "Europe"],
["Somalia", "Africa"],
["South Africa", "Africa"],
["South Korea", "Asia"],
["Spain", "Europe"],
["Sri Lanka", "Asia"],
["Sudan", "Africa"],
["Sweden", "Europe"],

```
  ["Switzerland", "Europe"],
  ["Syria", "Asia"],
  ["Taiwan", "Asia"],
  ["Thailand", "Asia"],
  ["Turkey", "Asia"],
  ["Uganda", "Africa"],
  ["Ukraine", "Europe"],
  ["Ungarn", "Europe"],
  ["United Arab Emirates", "Asia"],
  ["United Kingdom", "Europe"],
  ["United States", "North America"],
  ["Unlisted", "Unknown"],
  ["Uruguay", "South America"],
  ["Vatican City", "Europe"],
  ["Venezuela", "South America"],
  ["Vietnam", "Asia"],
  ["Yugoslavia", "Europe"],
  ["Zimbabwe", "Africa"]
] AS pair
MATCH (c:Country {name: pair[0]}), (k:Continent {name: pair[1]})
MERGE (c)-[:PART_OF]->(k)
```

Da bestimmte Hypothesen eine Aggregation auf Kontinentebene voraussetzen, wurde jedes Land einem Kontinent zugewiesen. Da kein geeignetes offenes Datenset verfügbar war, wurde diese Kategorisierung mithilfe von ChatGPT erstellt. Für jedes Land wurde die zugehörige Kontinenteinheit manuell definiert, und eine entsprechende Beziehung zwischen Land und Kontinent eingefügt.

```
//12. Religions adden, falls eine Abfrage demnach überprüft wird
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/religion.csv" AS line
WITH line, split(trim(line.religion), ',') AS religions
UNWIND religions AS religion
WITH trim(line.name) AS country, trim(religion) AS religion
MERGE (c:Country {country: country})
MERGE (r:Religion {name: religion})
MERGE (c)-[:HAS_RELIGION]->(r);
```

Eine der ursprünglich betrachteten Hypothesen erforderte die Kenntnis der dominanten Religionen in den einzelnen Ländern. Zu diesem Zweck wurde eine eigene Datei auf Basis öffentlich verfügbarer Daten (Wikipedia) erstellt. Jedem Land wurde seine jeweils meistverbreitete Religion zugeordnet, und diese Information als Beziehung zwischen Land und Religion in das Modell integriert. Für die korrekte Trennung mehrerer Einträge innerhalb eines Feldes wurde auch hier eine Vorverarbeitung angewendet.

```
//13. Meistverbreitete Religion je Kontinent (nach Länderanzahl) ermitteln
MATCH (cont:Continent)<-[:PART_OF]-(c:Country)-[:HAS_RELIGION]->(r:Religion)
WITH cont.name AS Kontinent, r.name AS Religion, count(*) AS Anzahl
ORDER BY Kontinent, Anzahl DESC
WITH Kontinent, collect({Religion: Religion, Anzahl: Anzahl})[0] AS meist
```

```
RETURN Kontinent, meist.Religion AS Meistgefolgte_Religion, meist.Anzahl AS Länderanzahl
ORDER BY Kontinent
```

Um weiterführende Auswertungen auf Kontinentebene zu ermöglichen, wurde für jeden Kontinent die Religion ermittelt, die unter den zugeordneten Ländern am häufigsten vertreten ist. Dies erfolgte durch eine Aggregation der Beziehungen zwischen Ländern und Religionen, gefolgt von einer Sortierung nach Anzahl. Das Ergebnis liefert für jeden Kontinent die am weitesten verbreitete Religion sowie die Anzahl der Länder, in denen sie dominiert.

```
//13.1 Meistgefolgte Religion per Kontinent
LOAD CSV WITH HEADERS FROM
"https://git.thm.de/aygr61/dataliteracyteam3/-/raw/master/religion_continent.csv" AS line
WITH trim(line.Kontinent) AS continentName,
    trim(line.Meistgefolgte_Religion) AS religionName
WHERE continentName IS NOT NULL AND religionName IS NOT NULL
MATCH (c:Continent {name: continentName})
MERGE (r:Religion {name: religionName})
MERGE (c)-[:MAJOR_RELIGION]->(r);
```

Die Cypher-Abfrage verarbeitet eine externe CSV-Datei, um für jeden Kontinent die am häufigsten vertretene Religion zu ermitteln und semantisch im Graphenmodell zu verankern. Dabei werden bestehende Kontinentknoten identifiziert, passende Religionsknoten bei Bedarf erzeugt und über eine strukturierte Beziehung dauerhaft miteinander verknüpft.