NETFLIX

Big Data Management
2nd Assignment

Eleni Neti,
2022202204018

Petros-Fotis Kamberi,
2022202204012

1. Αποθήκευση δεδομένων στο MongoDB

Για την μετατροπή του αρχείου csv των δεδομένων σε αρχείο json, υλοποιήθηκε ένα python script το οποίο μετατρέπει τα πεδία show_id και release year κάθε εγγραφής σε ακέραιο και επιπλέον μετατρέπει τα πεδία που έχουν πολλαπλές τιμές σε λίστες με στοιχεία τις ίδιες τιμές. Συγκεκριμένα, τα πεδία που μετατράπηκαν σε arrays είναι τα listed_in, cast και director.

Στην συνέχεια, το αρχείο json που προκύπτει από την παραπάνω διαδικασία εισάγεται στην βάση με την χρήση του MongoDB Compass, το οποίο χρησιμοποιείται παρακάτω για την ανάλυση των δεδομένων.

Να σημειώσουμε ότι η ημερομηνία λανσαρίσματος μιας παραγωγής στην πλατφόρμα, το πεδίο date_added δηλαδή, επιλέξαμε να το αποθηκεύσουμε σαν string τύπο δεδομένων. Η επιλογή αυτή μας ωφέλησε να εκτελέσουμε το πρώτο ερώτημα χρησιμοποιώντας ένα απλό regular expression operation. Παράλληλα, η MongoDB δίνει τη δυνατότητα της μετατροπής από έναν τύπο δεδομένων σε έναν άλλον. Συνεπώς, ανά πάσα στιγμή μπορεί κανείς να μετατρέψει τα δεδομένα του στον επιθυμητό τύπο, κάτι που αξιοποιήσαμε για οπτικοποίηση με ανεξάρτητη μεταβλητή τον χρόνο, όπως θα δούμε παρακάτω. Σε κάθε περίπτωση, κρατάμε αυτή τη διευκόλυνση που μας παρέχει η MongoDB.

Δεύτερον, θα θέλαμε να επισημάνουμε πως το script μας δεν περιέχει κάποιου είδους καθαρισμό δεδομένων (π.χ. από nulls) που είθισται να γίνεται εξαρχής. Ωστόσο, αυτό δεν μας εμπόδισε κατά την υλοποίηση της άσκησης καθώς φαίνεται ότι η MongoDB διαχειρίζεται χωρίς ιδιαίτερο πρόβλημα τα nulls - τουλάχιστον για τα βασικά operations που επιτελέσαμε εμείς (filtering, aggregations κτλ.). Ακόμη και τον τελεστή $unwind, ο οποίος δημιουργεί νέα documents για κάθε στοιχείο μιας λίστας που του έχουμε ορίσει, διαχειρίστηκε τις κενές τιμές χωρίς πρόβλημα. Σε περιπτώσεις όπου εμφανίζονται ανεπιθύμητες null τιμές, όπως για παράδειγμα όταν δουλέψαμε με τα πεδία director και ratings στις οπτικοποιήσεις, “πετάξαμε” τα nulls χρησιμοποιώντας τον operator $ne στο query μας.

2. Ανάλυση των δεδομένων

Για την δημιουργία των ακόλουθων ερωτημάτων και την ανάκτηση των αντίστοιχων αποτελεσμάτων, εγκαταστήσαμε και χρησιμοποιήσαμε το εργαλείο MongoDB Compass. Ειδικότερα, τα αποτελέσματα κάθε ερωτήματος έχουν εξαχθεί το καθένα σε ξεχωριστό αρχείο με όνομα question*i*.csv, όπου *i* είναι ο αύξων αριθμός του ερωτήματος (από 1 μέχρι 5). Τα αρχεία αυτά βρίσκονται στον φάκελο του παραδοτέου. Επιπλέον, σε κάθε ένα από τα ερωτήματα που ακολουθούν, συμπεριλαμβάνονται το ερώτημα που γίνεται στην βάση, οι 20 πρώτες εγγραφές (ή όλες σε περίπτωση που είναι λιγότερες), καθώς και ο σχολιασμός των αποτελεσμάτων. Να σημειωθεί ότι στο παραδοτέο έχουμε επίσης συμπεριλάβει και τα ερωτήματα σε αρχεία .js.

Διαθέσιμο περιεχόμενο το 2019

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
    [
      {
        $match: {
          date_added: { $regex: '.*2019.*' }
        }
      },
      {
        $project: {
          _id: 0,
          show_id: 1,
          type: 1,
          title: 1
        }
      },
      { $limit: 20 }
    ],
    { maxTimeMS: 60000, allowDiskUse: true }
  );
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

show_id	type	title
81145628	Movie	Norm of the North: King Sized Adventure
80221550	TV Show	Archibald's Next Big Thing
81154455	Movie	Article 15
81113928	Movie	Care of Kancharapalem
81052275	Movie	Ee Nagaraniki Emaindi
81132437	Movie	Kill Me If You Dare
80178151	TV Show	The Spy
81176188	Movie	American Factory: A Conversation with the Obamas
81160036	Movie	Saawan
81173255	Movie	The Heretics
81078908	Movie	The World We Make
81155784	Movie	Watchman
81054495	Movie	Mo Gilligan: Momentum

81053893	Movie	Cultivating the Seas: History and Future of the Full-Cycle Cultured Kindai Tuna
80200087	Movie	Domino
81053892	Movie	TUNA GIRL
80225885	TV Show	Bard of Blood
81132443	Movie	Deliha 2
80218107	TV Show	Dragons: Rescue Riders
80231903	Movie	In the Shadow of the Moon

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Παρατηρούμε ότι η συντριπτική πλειοψηφία των παραγωγών της πλατφόρμας ανήκουν στην κατηγορία των ταινιών. Αυτό είναι αναμενόμενο, από την στιγμή που οι ταινίες συνήθως απαιτούν λιγότερο χρόνο παραγωγής καθώς και χρηματικούς πόρους από ότι μια σειρά πολλών επεισοδίων. Ενδεχομένως, αυτή η αναλογία να οφείλεται επίσης στο γεγονός ότι μια σειρά για να μπορέσει να συνεχιστεί και σε επόμενες σεζόν θα πρέπει να έχει σχετικά μεγάλη απήχηση στο συνδρομητικό κοινό, με άλλα λόγια αρκετές προβολές, που να δικαιολογεί την επένδυση χρημάτων για την δημιουργία νέων επεισοδίων. Αντιθέτως, οι ταινίες δεν διατρέχουν αυτό το ρίσκο, μιας και πολλές φορές δεν παράγονται με την προοπτική δημιουργίας sequel, ιδίως οι ταινίες που χρηματοδοτούνται για την παραγωγή τους στην συγκεκριμένη πλατφόρμα.

Χώρες παραγωγής σειρών

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
    [
      { $match: { type: 'TV Show' } },
      { $unwind: { path: '$country' } },
      {
        $group: {
          _id: '$country',
          count: { $sum: 1 }
        }
      },
      { $sort: { count: -1, _id: 1 } },
      {
        $project: {
          _id: 0,
          count: 1,
          country: '$_id'
```

```
        }
      },
      { $limit: 20 }
    ],
    { maxTimeMS: 60000, allowDiskUse: true }
  );
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

count	country
686	United States
223	United Kingdom
156	Japan
116	South Korea
107	Canada
70	France
65	Taiwan
55	India
50	Australia
45	Mexico
45	Spain
36	China
25	Germany
25	Turkey
23	Colombia
18	Brazil
18	Thailand
15	Italy
14	Argentina
14	Russia

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Όπως είναι φυσικό, οι χώρες οι οποίες εμφανίζονται στις κορυφαίες 20 όσον αφορά στις παραγωγές στην πλατφόρμα, είναι χώρες που είτε η επίσημή τους γλώσσα έχει μεγάλο αριθμό από ομιλητές παγκοσμίως (π.χ. αγγλικά, γαλλικά, ισπανικά) είτε έχουν μεγάλο πληθυσμό και άρα κοινό το οποίο θα παρακολουθήσει τις συγκεκριμένες παραγωγές (π.χ. Βραζιλία, Κίνα, Τουρκία) είτε διαθέτουν μία γενικότερη παράδοση στον χώρο του σινεμά και την παραγωγή σειρών τύπου Anime (Ν.Κορέα και Ιαπωνία) που έχουν ιδιαίτερη απήχηση.

Επιπλέον, με αρκετά μεγάλη διαφορά στην κορυφή των παραγωγών παγκοσμίων βρίσκονται οι Η.Π.Α που είναι η χώρα ίδρυσης της πλατφόρμας και όπως είναι γνωστό η αμερικανική βιομηχανία παραγωγής ταινιών και σειρών είναι διαχρονικά από τις μεγαλύτερες παγκοσμίως, αν όχι η μεγαλύτερη. Αξίζει να σημειωθεί ακόμη ότι στο τοπ 10 των χωρών-παραγωγών βρίσκονται 4 αγγλόφωνες χώρες (ΗΠΑ, Ηνωμένο Βασίλειο, Καναδάς και Αυστραλία), πράγμα το οποίο επίσης επιβεβαιώνει το γεγονός ότι χώρες με επίσημες γλώσσες που έχουν πολλούς ομιλητές παγκοσμίως έχουν γενικά στην πλατφόρμα περισσότερες παραγωγές σε σχέση με άλλες χώρες των οποίων η γλώσσα διαθέτει σχετικά περιορισμένο πληθυσμό ομιλητών.

Είδη διαθέσιμου περιεχομένου

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
    [
      { $unwind: { path: '$listed_in' } },
      {
        $group: {
          _id: '$listed_in',
          totalSum: { $sum: 1 },
          movies: {
            $sum: {
              $cond: [
                { $eq: ['$type', 'Movie'] },
                1,
                0
              ]
            }
          },
          TVShows: {
            $sum: {
              $cond: [
                { $eq: ['$type', 'TV Show'] },
                1,
                0
              ]
            }
          }
        }
      },
      { $sort: { totalSum: -1 } },
      {
        $project: {
          _id: 0,
          genre: '$_id',
```

```
          totalSum: {
            $add: ['$movies', '$TVShows']
          }
        }
      },
      { $limit: 20 }
    ],
    { maxTimeMS: 60000, allowDiskUse: true }
  );
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

genre	totalSum
International Movies	1927
Dramas	1623
Comedies	1113
International TV Shows	1001
Documentaries	668
TV Dramas	599
Action & Adventure	597
Independent Movies	552
TV Comedies	436
Thrillers	392
Children & Family Movies	378
Romantic Movies	376
Crime TV Shows	363
Kids' TV	328
Stand-Up Comedy	281
Docuseries	279
Romantic TV Shows	278
Horror Movies	262
Music & Musicals	243
British TV Shows	210

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Από τα αποτελέσματα φαίνεται ότι οι διεθνείς ταινίες (International Movies) είναι το πρώτο με σημαντική διαφορά είδος παραγωγής που απαντάται στην πλατφόρμα. Αυτό κατά πάσα πιθανότητα οφείλεται στην δυνατότητα αυτού του είδους ταινιών να απηχούν σε πολλά και ετερογενή πολλές φορές κοινά. Έπειτα ακολουθούν δύο διαχρονικά κλασικά είδη

παραγωγών που δεν είναι άλλα από το δράμα και την κωμωδία. Αυτό μπορεί ίσως να οφείλεται στο γεγονός ότι πολλές φορές μια παραγωγή μπορεί να μην εμπίπτει σε κάποιο άλλο είδος από τα διαθέσιμα στην πλατφόρμα ή μπορεί να είναι ένα κράμμα διαφορετικών ειδών και να μην είναι ξεκάθαρη η επιλογή κάποιου είδους. Έτσι, ενδεχομένως να επιλέγεται αντίστοιχα κάποιο από αυτά τα δύο ως το πιο κοντινό σε αυτά ως εναλλακτική ή επειδή υπερισχύει σε σχέση με τα άλλα είδη που μπορεί να εμπίπτει η εκάστοτε παραγωγή. Επίσης, είναι αξιοσημείωτο ότι σημαντικό πλήθος παραγωγών ανήκει στα Ντοκυμαντέρ, το οποίο υποδεικνύει ότι μεγάλο κομμάτι των συνδρομητών επιλέγει την πλατφόρμα και για λόγους που σχετίζονται με την καλλιέργεια και την επιμόρφωσή τους, πέρα από την ψυχαγωγία/διασκέδαση.

Εμφανιζόμενοι ηθοποιοί

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
  [
    { $unwind: '$cast' },
    {
      $group: { _id: '$cast', count: { $sum: 1 } }
    },
    { $sort: { count: -1, _id: 1 } },
    { $limit: 20 },
    {
      $project: { _id: 0, name: '$_id', count: 1 }
    }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

count	name
33	Anupam Kher
30	Shah Rukh Khan
27	Naseeruddin Shah
27	Om Puri
26	Akshay Kumar
26	Yuki Kaji
25	Paresh Rawal
25	Takahiro Sakurai
24	Amitabh Bachchan

23	Boman Irani
22	Andrea Libman
22	Ashleigh Ball
22	John Cleese
19	Kareena Kapoor
18	Daisuke Ono
18	David Attenborough
18	Erin Fitzgerald
18	Fred Tatasciore
18	Gulshan Grover
18	Kay Kay Menon

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Με βάση τα αποτελέσματα προκύπτει ότι οι πιο συχνά εμφανιζόμενοι ηθοποιοί προέρχονται κυρίως από ασιατικές χώρες και ιδίως από την Ινδία. Υποθέτουμε ότι αυτό οφείλεται στην μεγάλη παραγωγή ταινιών από το Bollywood, την ινδική βιομηχανία παραγωγής ταινιών, που ετησίως παράγει πολλές ταινίες και σειρές και επιπλέον δεν έχει τους περιορισμούς που ενδεχομένως να έχουν ταινίες αμερικανικής ή ευρωπαϊκής παραγωγής όσον αφορά στην διάθεση περιεχομένου εξαιτίας κείμενων νομοθεσιών που αφορούν τα πνευματικά δικαιώματα.

Κορυφαίες προτιμήσεις ηθοποιών

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
  [
    { $unwind: '$cast' },
    { $unwind: '$listed_in' },
    {
      $group: {
        _id: {
          cast: '$cast',
          genre: '$listed_in'
        },
        count: { $sum: 1 }
      }
    },
    { $sort: { '_id.cast': 1, count: -1 } },
    {
      $group: {
        _id: '$_id.cast',
```

```
        mostFrequentGenre: {
          $first: '$_id.genre'
        },
        count: { $first: '$count' }
      }
    },
    {
      $project: {
        _id: 0,
        name: '$_id',
        genre: '$mostFrequentGenre',
        count: 1
      }
    },
    { $sort: { name: 1, genre: 1, count: 1 } },
    { $limit: 20 }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

count	name	genre
1	Jr.	TV Dramas
1	2 Chainz	Docuseries
1	4Minute	International Movies
2	50 Cent	Action & Adventure
1	A Boogie Wit tha Hoodie	Docuseries
1	A-ra Go	Crime TV Shows
1	A. Murat Özgen	Horror Movies
1	A.C. Peterson	Dramas
2	A.D. Miles	TV Comedies
1	A.J. Cook	TV Mysteries
2	A.J. LoCascio	Kids' TV
3	A.K. Hangal	International Movies
1	A.R. Rahman	Documentaries
1	A.S. Sasi Kumar	International Movies
1	AFRA	International TV Shows
1	AJ Bowen	LGBTQ Movies

1	AJ Michalka	TV Action & Adventure
1	AJ Rivera	TV Dramas
1	Aabhas Yadav	Music & Musicals
1	Aachal Munjal	International Movies

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Αυτό που μπορούμε να εξάγουμε από τις 20 πρώτες εγγραφές των αποτελεσμάτων είναι ότι οι περισσότεροι ηθοποιοί εμφανίζονται σε σχετικά λίγες παραγωγές (π.χ. 1 ή 2). Αυτό μπορεί να οφείλεται σε αρκετούς παράγοντες όπως είναι η δημοτικότητα του ηθοποιού, αν δραστηριοποιείται κυρίως στο θέατρο, αν η κύρια ιδιότητά του δεν είναι εκείνη του ηθοποιού (π.χ. μπορεί να είναι μουσικός ή τραγουδιστής) κ.ά. Χαρακτηριστικό παράδειγμα της τελευταίας περίπτωσης αποτελεί η εμφάνιση δύο ράπερ (2 Chainz και 50 Cents) οι οποίοι δεν έχουν προφανώς ως κύρια δραστηριότητά τους την ηθοποιία και ενδεχομένως να χρειάστηκε να εμφανιστούν σε κάποια παραγωγή ως guest stars. Παρακάτω στην εργασία, θα δούμε ένα διαφορετικό εύρος των δεδομένων μας, έχοντας οπτικοποιήσει τα αποτελέσματά μας για τους ηθοποιούς με τις περισσότερες δουλειές.

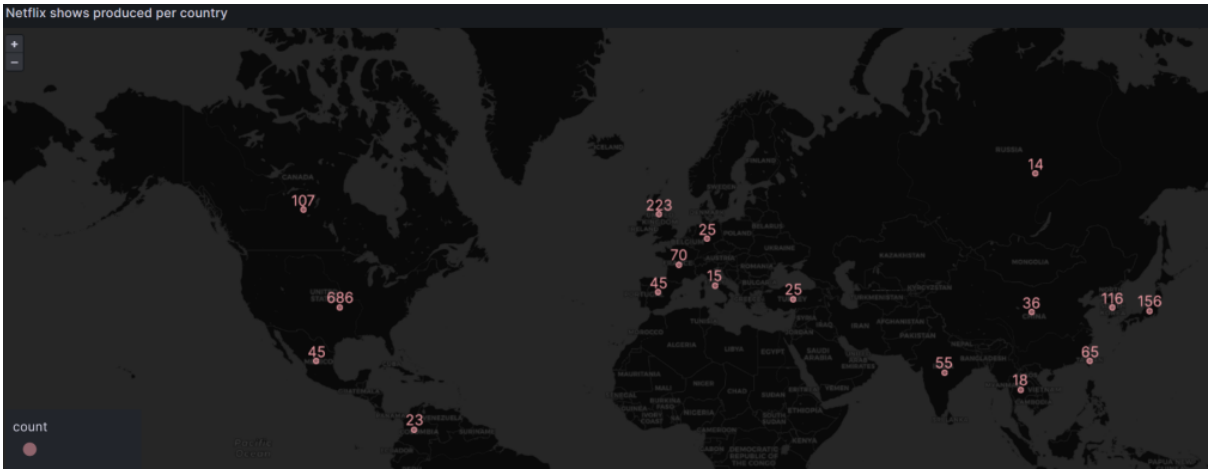
3. Bonus υλοποίηση

Για την δημιουργία των παρακάτω γραφικών απεικονίσεων χρησιμοποιήθηκαν τα εξής δύο εργαλεία οπτικοποίησης: **MongoDB Atlas** (ανοιχτόχρωμο background) και **Grafana** (σκουρόχρωμο background). Για τα επιπλέον ερωτήματα έχουμε συμπεριλάβει και τις 20 πρώτες εγγραφές (τα αποτελέσματα παρουσιάζονται κατά φθίνουσα σειρά στατιστικής μετρικής (count, average)) καθώς και έναν σύντομο σχολιασμό των αποτελεσμάτων, σε αντιστοιχία με το δεύτερο ερώτημα.

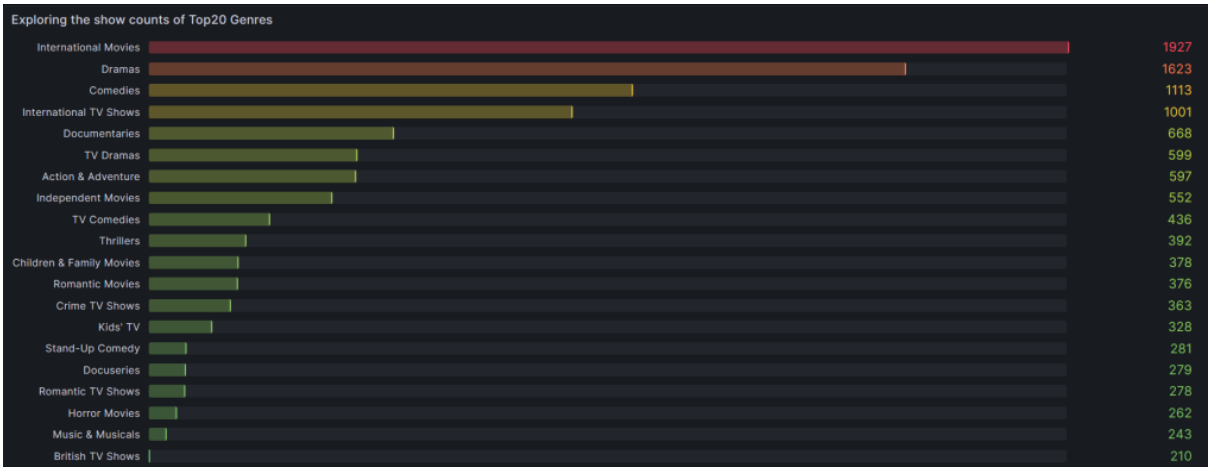
Διαθέσιμο περιεχόμενο το 2019

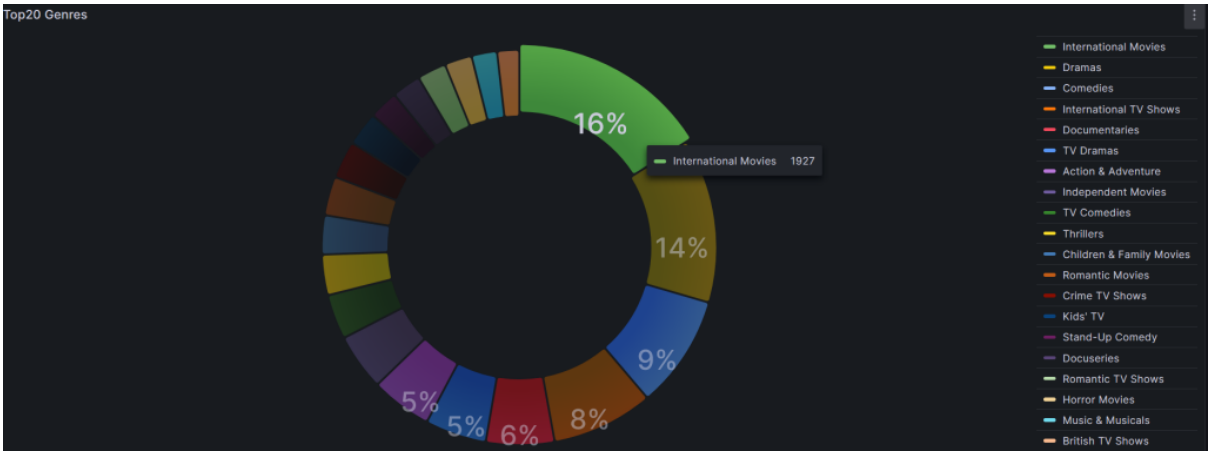
Καθότι το πρώτο ερώτημα επιστρέφει τα docs με τον μοναδικό κωδικό, τον τύπο και τον τίτλο των παραγωγών του 2019, για την οπτικοποίηση, προωθήσαμε εναλλακτικά ένα διαφορετικό query προκειμένου να φτιάξουμε ένα πάνελ καταμέτρησης του υλικού που έγινε διαθέσιμο το 2019 στην πλατφόρμα.

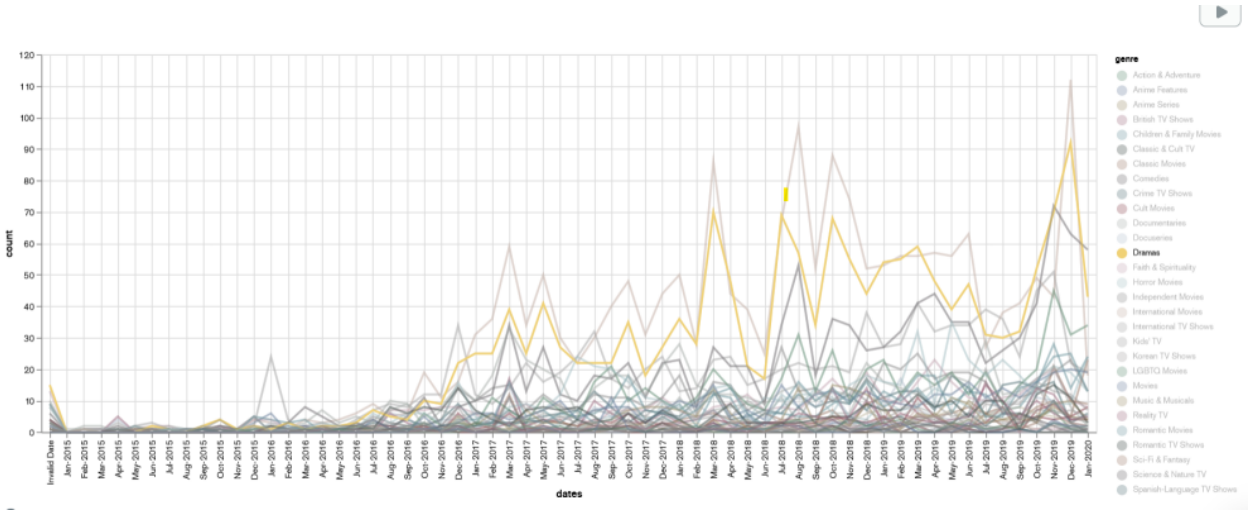
Χώρες παραγωγής σειρών (κορυφαίες 20)



Είδη διαθέσιμου περιεχομένου



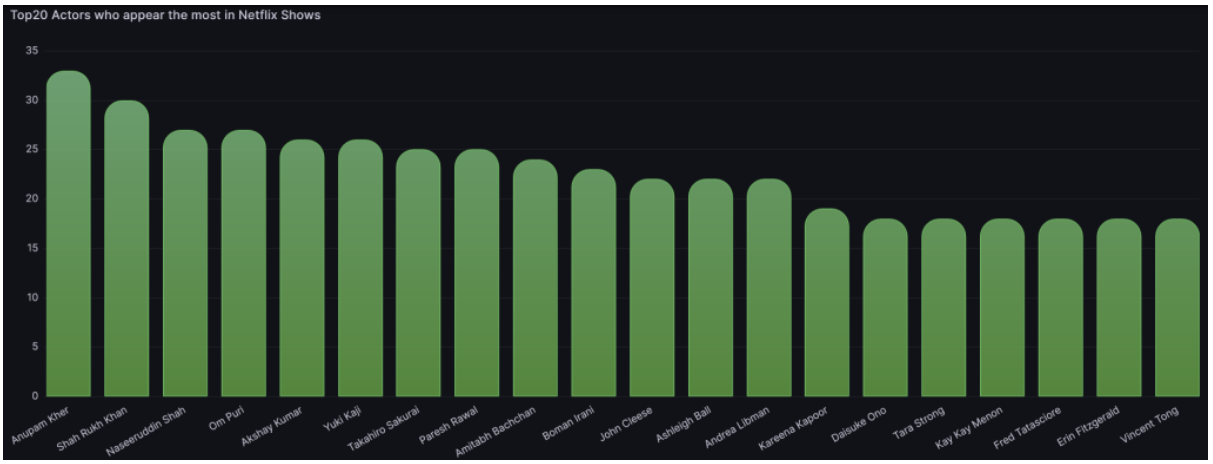




Το τελευταίο διάγραμμα αποτελεί μια οπτικοποίηση τύπου χρονοσειράς. Διαθέτει δύο ενεργά φίλτρα προκειμένου να φιλτράρει κανείς με βάση το είδος περιεχομένου (genre) αλλά και με βάση κάποιο χρονικό διάστημα ή στιγμιότυπο. Στην παραπάνω απεικόνιση έχουμε κάνει highlight το είδος Dramas (που αναφέρεται σε δραματικές ταινίες), ενώ παράλληλα έχουμε κλείσει το χρονικό παράθυρο από τον Ιανουάριο του 2015 μέχρι τον Ιανουάριο του 2020. Αφενός πρόκειται για ένα διάστημα ενδιαφέροντος όπου η Netflix “απογειώθηκε” ως πλατφόρμα, αφετέρου μπορεί κανείς να μελετήσει διακυμάνσεις ή τυχόν σημεία που να έχουν ενδιαφέρον να εξερευνήσει κάποιος, όπως λόγου χάρη την περίοδο της εμφάνισης της πανδημίας Covid-19, της κοινωνικής αποστασιοποίησης, και αν ή πώς επηρέασε την διάθεση περιεχομένου στην πλατφόρμα.

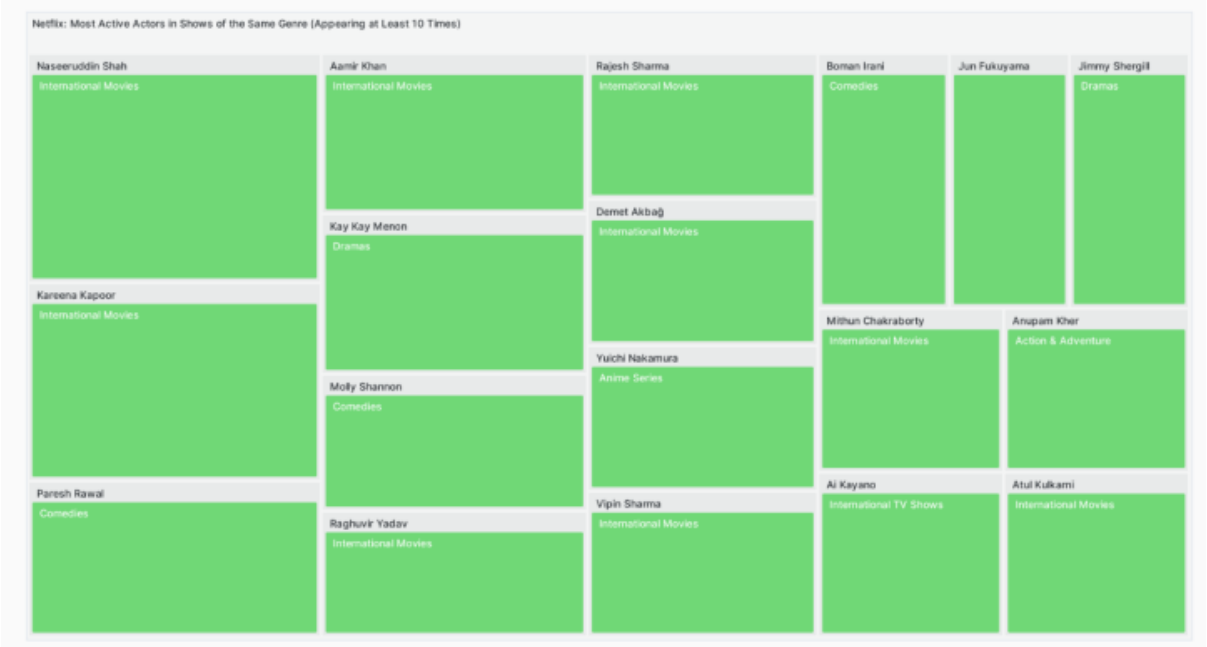
Τέλος, αξίζει να σημειωθεί ότι για την τελευταία οπτικοποίηση, πραγματοποιήσαμε μετασχηματισμό του τύπου δεδομένων του πεδίου “date_added” από string σε date. Το MongoDB Atlas έχει τη δυνατότητα να μετατρέπει τα δεδομένα στην default Date type μορφή της γλώσσας javascript. Έτσι λάβαμε τα δεδομένα στη μορφή που θέλαμε προκειμένου να χρησιμοποιήσουμε τον χρόνο σαν μεταβλητή.

Εμφανιζόμενοι ηθοποιοί



Η τελευταία οπτικοποίηση είναι της μορφής word cloud, όπου το μέγεθος του ονόματος σχετίζεται με τον αριθμό των φορών που εμφανίζεται στα δεδομένα, ενώ το χρώμα το συσχετίσαμε με το αν ο/η ηθοποιός εμφανίζεται σε ταινίες ή σειρές. Με αυτή την οπτικοποίηση, μπορεί κανείς να αντιληφθεί διαισθητικά τους συχνότερα εμφανιζόμενους ηθοποιούς και τον τύπο των παραγωγών (σειρά/ταινία) στις οποίες συμμετείχε. Για παράδειγμα, ο Anupam Kher, που είναι και ο κορυφαίος σε φορές εμφανίσεις στην πλατφόρμα, εμφανίζεται δύο φορές στο νέφος, μία με πράσινο και αυξημένη γραμματοσειρά (για 32 ταινίες) και μία φορά με μικρή γραμματοσειρά με μπλε (1 σειρά).

Κορυφαίες προτιμήσεις ηθοποιών (Most Active Actors in Netflix Shows of the same genre)



Η ιδέα του παραπάνω tree map είναι να συνοψίσει με οπτικό τρόπο πληροφορία τριών “διαστάσεων”. Τα δεδομένα όπως φαίνεται και παραπάνω, είναι ομαδοποιημένα ανά ηθοποιό, και με ένα απλό mouse over, μπορεί κανείς να δει το συχνότερο είδος περιεχομένου που αυτός/ή συμμετέχει, καθώς και το πλήθος των δουλειών σε αυτό το είδος. Για λόγους παρουσίασης, συμπεριλάβαμε μόνο τους ηθοποιούς με τουλάχιστον 10 εμφανίσεις στο εκάστοτε είδος.

Σημάνσεις Ηλικιακής Καταλληλότητας

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
  [
    {
      $group: {
        _id: '$rating',
        count: { $sum: 1 }
```

```
      }
    },
    { $sort: { count: -1 } },
    { $limit: 20 },
    {
      $project: {
        _id: 0,
        rating: '$_id',
        count: 1
      }
    }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

count	rating
2027	TV-MA
1698	TV-14
701	TV-PG
508	R
286	PG-13
218	NR
184	PG
169	TV-Y7
149	TV-G
143	TV-Y
95	TV-Y7-FV
37	G
7	UR
2	NC-17

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Από τα παραπάνω αποτελέσματα, προκύπτει ότι η πλειοψηφία των παραγωγών προορίζονται για ενήλικες ή εφήβους (TVMA, TV-14, R, NC-17). Αυτό είναι λογικό, αν αναλογιστούμε ότι για την δημιουργία λογαριασμού στην πλατφόρμα απαιτείται η χρήση πιστωτικής κάρτας και άρα η συμπλήρωση του 18ου συνήθως έτους της ηλικίας. Τα προγράμματα που προορίζονται για όλα τα κοινά (TV-G, G), για παιδιά (TV-Y7, TV-Y, TV-Y7-FV) ή απαιτούν γονική συναίνεση (PG-13, PG) είναι αισθητά λιγότερα, γεγονός το

οποίο συνάδει με την υπόθεσή μας ότι το κύριο κοινό της πλατφόρμας είναι έφηβοι και ενήλικες, και είναι έως έναν βαθμό αναμενόμενο, μιας και η παραγωγή τέτοιου είδους περιεχομένου υπόκεινται σε αρκετούς περιορισμούς καταλληλότητας που ενδεχομένως να δυσχεραίνουν ή ακόμη και να καθιστούν πρακτικά αδύνατη την παραγωγή περισσότερων τέτοιων προγραμμάτων.

Συχνότερα Εμφανιζόμενοι Σκηνοθέτες

Ερώτημα στην βάση

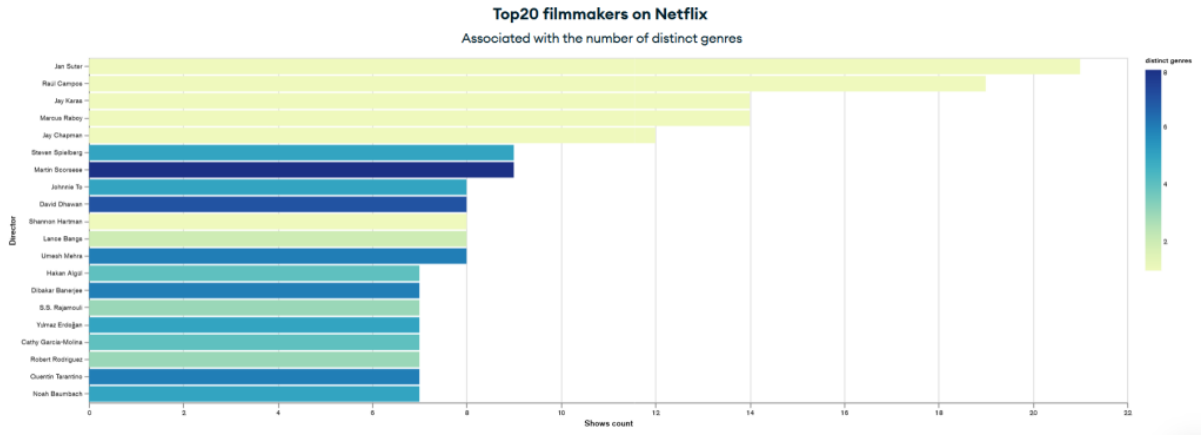
```
db.getCollection('Netflix').aggregate(
  [
    { $unwind: '$director' },
    {
      $group: {
        _id: '$director',
        count: { $sum: 1 }
      }
    },
    { $sort: { count: -1 } },
    { $limit: 20 },
    {
      $project: {
        _id: 0,
        director: '$_id',
        count: 1
      }
    }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Οι 20 πρώτες εγγραφές των αποτελεσμάτων

count	director
21	Jan Suter
19	Raúl Campos
14	Marcus Raboy
14	Jay Karas
12	Jay Chapman
9	Martin Scorsese
9	Steven Spielberg
8	Shannon Hartman
8	David Dhawan
8	Johnnie To
8	Lance Bangs
8	Umesh Mehra
7	Noah Baumbach
7	Ryan Polito
7	Quentin Tarantino
7	Dibakar Banerjee
7	Robert Rodriguez
7	Yılmaz Erdoğan
7	S.S. Rajamouli
7	Cathy Garcia-Molina

Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Αυτό που αξίζει να σημειωθεί στα παραπάνω αποτελέσματα, είναι το γεγονός ότι γνωστοί και πολυβραβευμένοι σκηνοθέτες, όπως οι Martin Scorsese, Steven Spielberg, Quentin Tarantino και Noah Baumbach βρίσκονται αισθητά χαμηλότερα όσον αφορά στην εμφάνιση τους σε περιεχόμενο της πλατφόρμας σε σχέση με λιγότερο “καταξιωμένους” σκηνοθέτες που δεσπόζουν στις κορυφαίες θέσεις της παραπάνω κατάταξης. Αυτό ίσως να οφείλεται στο γεγονός ότι οι τελευταίοι μπορεί να σκηνοθετούν παραγωγές που προορίζονται εξαρχής για την πλατφόρμα, ενώ τα έργα των πρώτων όχι.



Το παραπάνω διάγραμμα μας βοηθάει να ξεδιαλύνουμε λίγο παραπάνω αυτή την μη αναμενόμενη κατάταξη των σκηνοθετών. Στο παραπάνω barchart φαίνεται με χρώμα το πλήθος των μοναδικών ειδών στα οποία έχει αναφορά ένας σκηνοθέτης. Παρατηρούμε ότι στις πρώτες θέσεις δεσπόζουν σκηνοθέτες υψηλής “ειδίκευσης” ως προς το περιεχόμενο, σε αντίθεση με τις χαμηλότερες θέσεις όπου η ποικιλία αυξάνεται. Πράγματι, στις δύο πρώτες θέσεις αναγνωρίζουμε τους Jan Suter και Raul Campos, οι οποίοι είναι σκηνοθέτες και παραγωγοί με πολύ έντονη δραστηριότητα στον χώρο της “κινηματογραφικής” stand-up κωμωδίας και τεράστια απήχηση στη Λατινική Αμερική.

Μέση Διάρκεια σε λεπτά ανά Είδος Ταινίας

Ερώτημα στην βάση

```
db.getCollection('Netflix').aggregate(
  [
    {
      $project: {
        filteredListedIn: {
          $map: {
            input: '$listed_in',
            as: 'category',
            in: {
              $cond: [
                {
                  $regexMatch: {
                    input: '$duration',
                    regex: '^[0-9]+ min$'
                  }
                },
                '$$category',
                '$$REMOVE'
              ]
            }
          }
```

```
        },
        durationNumeric: {
          $toDouble: {
            $arrayElemAt: [
              { $split: ['$duration', ' '] },
              0
            ]
          }
        }
      }
    },
    { $unwind: '$filteredListedIn' },
    {
      $group: {
        _id: '$filteredListedIn',
        avgDuration: { $avg: '$durationNumeric' }
      }
    },
    { $sort: { avgDuration: -1, _id: 1 } },
    { $limit: 20 },
    {
      $project: {
        _id: 0,
        listed_in: '$_id',
        avgDuration: 1
      }
    }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

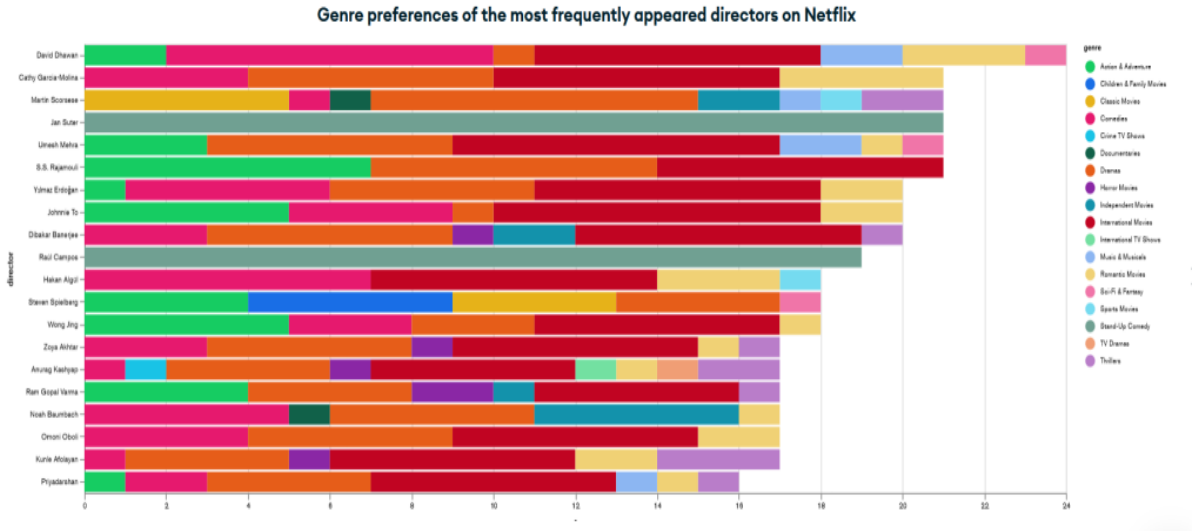
Οι 20 πρώτες εγγραφές των αποτελεσμάτων

avgDuration (min)	genre
113.98492462311557	Action & Adventure
113.86383240911891	Dramas
113.10714285714286	Classic Movies
111.57712765957447	Romantic Movies
110.97716658017644	International Movies
110.47736625514403	Music & Musicals
106.6839378238342	Sci-Fi & Fantasy
106.30102040816327	Thrillers

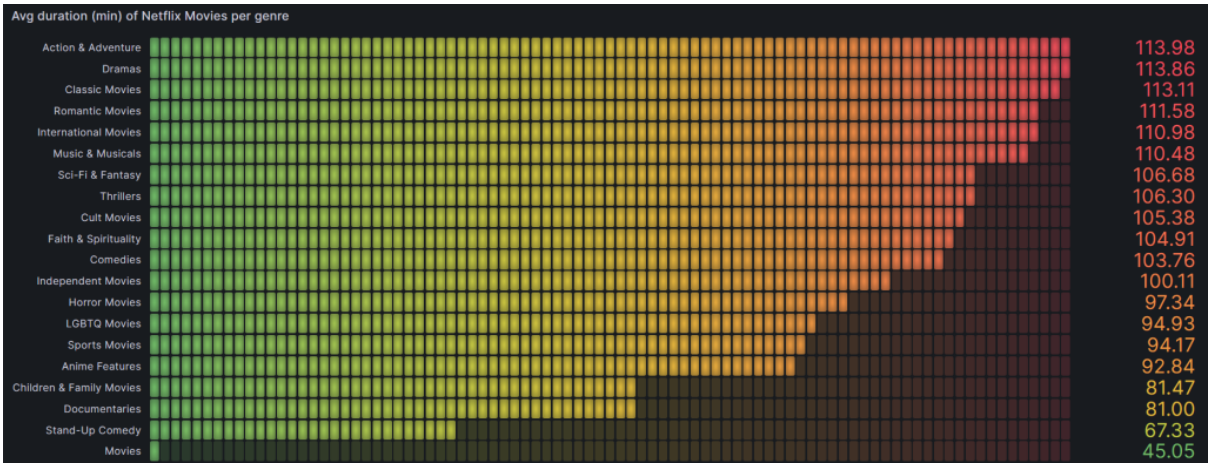
105.38181818181818	Cult Movies
104.91489361702128	Faith & Spirituality
103.76190476190476	Comedies
100.11413043478261	Independent Movies
97.33969465648855	Horror Movies
94.93333333333334	LGBTQ Movies
94.171974522293	Sports Movies
92.84444444444445	Anime Features
81.46825396825396	Children & Family Movies
81	Documentaries
67.33096085409252	Stand-Up Comedy
45.05357142857143	Movies

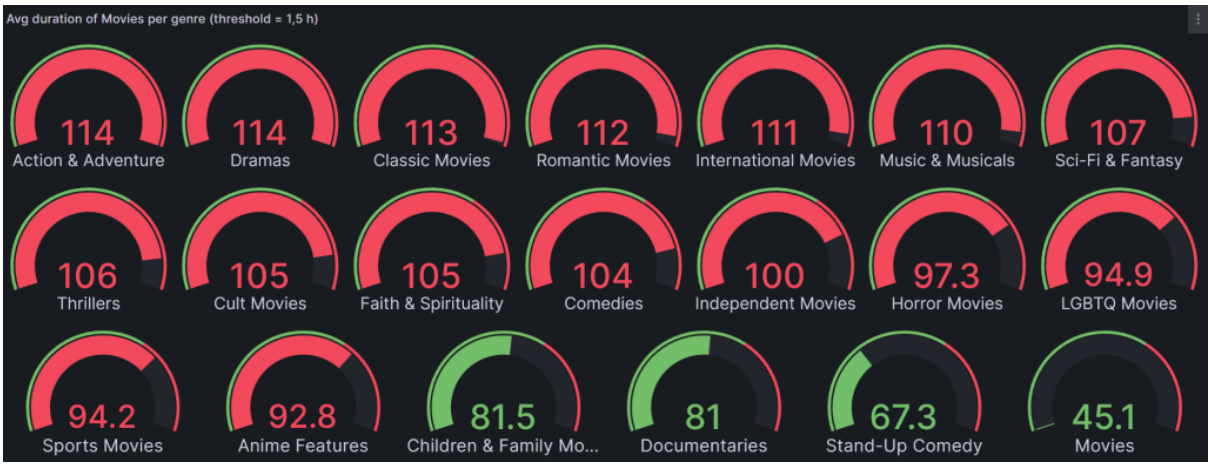
Συνοπτικός Σχολιασμός των Αποτελεσμάτων

Παρατηρούμε ότι για τα περισσότερα είδη ταινιών που αναγράφονται παραπάνω, η μέση διάρκεια τους είναι πάνω από μία ώρα και η πλειοψηφία των ταινιών αυτών ξεπερνούν τα 100 λεπτά, δηλαδή έχουν μέση διάρκεια πάνω από μιάμιση ώρα. Επίσης, βλέπουμε ότι είδη ταινιών με απλούστερη πλοκή έχουν σχετικά μικρότερη μέση διάρκεια (π.χ. Ντοκυμαντέρ, Αθλητικές και Παιδικές-Οικογενειακές ταινίες), ενώ ταινίες με περίπλοκη πλοκή (π.χ. Δράσης και Δράμα) έχουν κατά κανόνα μεγαλύτερη μέση διαρκεια.



Η παραπάνω οπτικοποίηση διαθέτει επίσης ενεργό φίλτρο, βάσει του οποίου μπορεί κανείς να εντοπίσει τους σκηνοθέτες που σχετίζονται με συγκεκριμένο είδος περιεχομένου, επιλέγοντας το επιθυμητό είδος από τη λίστα στα δεξιά του γραφήματος.





Στο παραπάνω πάνελ, έχουμε θέσει σαν threshold τη μιάμιση ώρα (90') , ως εκ τούτου αν η μέση διάρκεια των παραγωγών κάποιου είδους ξεπερνάει αυτό το κατώφλι, το αντίστοιχο gauge διάγραμμα κοκκινίζει.