ΔΙΑΧΕΙΡΙΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

2η Προγραμματιστική Εργασία

Ομάδα:
Γραμματικόπουλος Λάμπρος 2022201800038 dit18038@uop.gr
Κολοτούρος Κωνσταντίνος 2022201800090 dit18090@uop.gr

Περιεχόμενα:

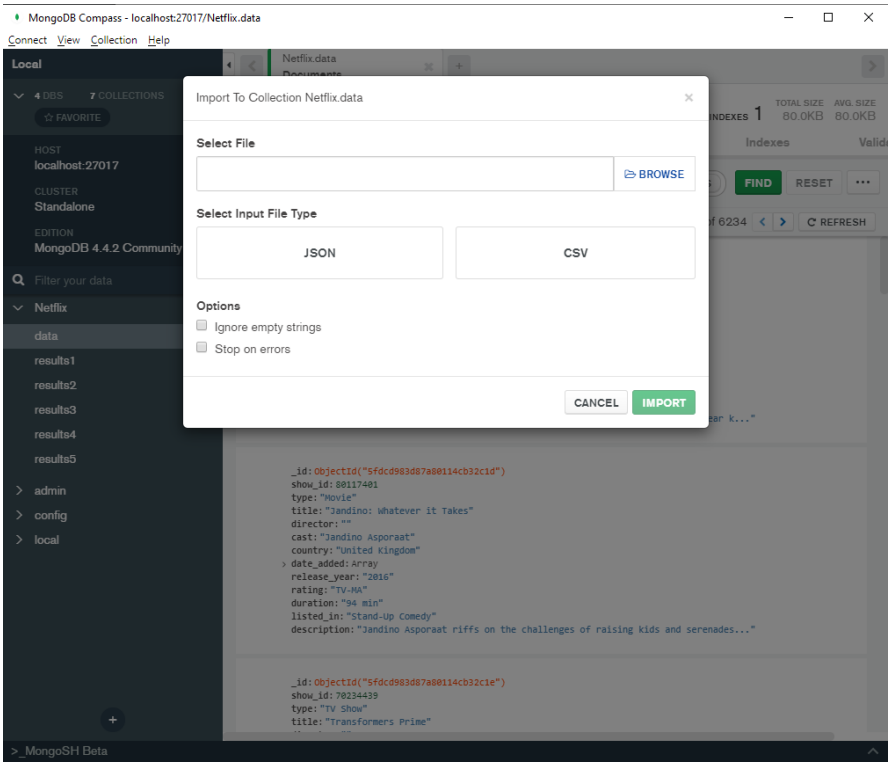
Απάντηση ερωτήματος 1:

Αρχικά, έγινε η εγκατάσταση του MongoDB V4.4.2 κατά την οποία εγκαταστήθηκε το MongoDB shell αλλά και το MongoDB Compass.

Έπειτα, πραγματοποιήθηκε η μετατροπή από το CSV αρχείο σε ένα αρχείο τύπου JSON με έναν online μετατροπέα (https://csvjson.com/csv2json).

Χρειάστηκε όμως περαιτέρω επεξεργασία του αρχείου που λήφθηκε από τον μετατροπέα διότι έπρεπε να άλλαξουν τα πεδία: director,cast,country,listed_in,date_added γιατί περιείχαν πολλαπλές τιμές σε ένα πεδίο και γι'αυτό έλαβαν μορφή πίνακα. Επίσης, άλλαξε το πεδίο release_year και από αριθμός έγινε αλφαριθμητικό για να είναι ευκολότερα διαχειρίσιμο από τυχόν queries σε περίπτωση που για παράδειγμα ήταν 19 αντί για 2019. Όλες αυτές οι αλλαγές στο αρχείο JSON έγιναν με χρήση ενός script (JsonAlternation) που υλοποιήθηκε σε γλώσσα JAVA, το οποίο περιέχεται στα υποβληθέντα αρχεία.

Ύστερα, επειδή η εισαγωγή όλων των δεδομένων από το MongoDB Shell ανά μία εγγραφή τη φορά ήταν πρακτικά αδύνατον, έγινε η μαζική εισαγωγή όλων των δεδομένων, από το τελικό JSON αρχείο που παράχθηκε, μέσα από το εργαλείο MongoDB Compass σε μια συλλογή δεδομένων (data) της βάσης δεδομένων (Netflix) που δημιουργήθηκε.



Απάντηση ερωτήματος 2:

Για την επίλυση του δεύτερου ερωτήματος χρησιμοποιήθηκε το εργαλείο MongoDB Compass και συγκεκριμένα η δομή Aggregation Pipeline Builder. Η δομή αυτή χρησιμοποιήθηκε διότι μέσω αυτής μπορούν αρκετά πιο εύκολα να γίνουν απλές αλλά και περίπλοκες ερωτήσεις για τα δεδομένα της βάσης. Επίσης, αυτή η δομή λειτουργεί σε στάδια, δηλαδή μετά από κάθε εντολή που εκτελείται, πραγματοποιείται μία εκτύπωση των αποτελεσμάτων, γεγονός που βοήθησε πολύ στην κατανόηση των ερωτήσεων που έγιναν για τα δεδομένα και στην διόρθωση λαθών που προέκυπταν. Έπειτα, σε κάθε ερώτημα μετά από όλα τα στάδιά του, προστέθηκε το στάδιο $out το οποίο έσωσε τα αποτελέσματα των ερωτημάτων σε ξεχωριστές συλλογές δεδομένων. Τέλος, για κάθε μια από τις τελικές συλλογές έγινε εκτύπωση των αποτελεσμάτων μέσω του MongoDB Compass σε αρχεία CSV τα οποία περιέχονται στα υποβληθέντα αρχεία. Οι κώδικες που υλοποιούν κάθε ερώτημα βρίσκονται στα υποβληθέντα αρχεία.

i. Διαθέσιμο περιεχόμενο το 2019

Χρησιμοποιήθηκε μια εντολή **$match** για να βρεθεί το περιεχόμενο που έγινε διαθέσιμο το 2019 και μια εντολή **$group** για την εμφάνιση στο αρχείο εξόδου τα απαραίτητα στοιχεία της εγγραφής.

	_id	type	title
2	70095135	Movie	Knowing
3	81031008	Movie	For the Birds
4	70264888	TV Show	Black Mirror
5	80216820	Movie	Never Heard
6	81001494	Movie	Our Godfather
7	80194956	TV Show	The Disappearance of Madeleine McCann
8	70018445	Movie	Mujhse Shaadi Karogi
9	81028895	TV Show	Arthdal Chronicles
10	80199689	Movie	Velvet Buzzsaw
11	81034012	Movie	Street Flow
12	81113887	TV Show	Inmates
13	80158800	Movie	The 24 Hour War
14	81035869	Movie	A Mission in an Old Movie
15	80153272	TV Show	Sin Senos sÃ- Hay ParaÃ-so
16	81094667	Movie	Hidden in Plain Sight
17	81172727	Movie	Merry Men: The Real Yoruba Demons
18	70109689	Movie	Paul Blart: Mall Cop
19	80228274	TV Show	Teasing Master Takagi-san
20	81029301	TV Show	Botched Up Bodies
21	80175934	TV Show	Pinky Malinky

Σχόλια:

Εμφανίστηκαν όπως περιμέναμε τα πεδία _id, type και title.

ii. Χώρες παραγωγής σειρών

Χρησιμοποιήθηκε μια εντολή **$match** για να βρεθούν όλες οι εγγραφές τύπου TV-Show και για να μην συμπεριληφθούν οι εγγραφές της συλλογής χωρίς καταχωρημένη χώρα. Έπειτα, χρησιμοποιήθηκε μια εντολή **$unwind** για να εφανιστούν σαν ξεχωριστές εγγραφές, οι εγγραφές με περισσότερες από μία χώρες. Ακόλουθα, έγινε χρήση μιας εντολής **$group** για την συγκέντρωση όλων των εγγραφών με κοινό όνομα χώρας και την καταμέτρηση των εγγραφών αυτών για την εύρεση του πλήθους των σειρών που έχει μια χώρα παράγει. Τέλος, χρησιμοποιήθηκε μια εντολή **$sort** για την ταξινόμιση σε φθίσουσα σειρά ως προς το πλήθος των σειρών.

1	_id	count
2	United States	616
3	United Kingdom	200
4	Japan	153
5	South Korea	111
6	Canada	84
7	United State	70
8	Taiwan	65
9	France	60
10	India	53
11	Australia	45
12	Spain	43
13	Mexico	41
14	China	34
15	Turkey	24
16	Canad	23
17	United Kingdo	23
18	Germany	21
19	Colombia	20
20	Brazil	18
21	Thailand	18

<u>Σχόλια:</u>

Εμφανίστηκαν όπως περιμέναμε τα πεδία _id,count ταξινομιμένα ως προς το count, δηλαδή το πλήθος των σειρών με φθίνουσα σειρά. Επίσης, οι εγγραφές χωρίς καταχωρημένη χώρα δεν έχουν εμφανιστεί όπως και θέλαμε.

iii. Είδη διαθέσιμου περιεχομένου

Χρησιμοποιήθηκε μια εντολή **$unwind** για να εφανιστούν σαν ξεχωριστές εγγραφές, οι εγγραφές με περισσότερες από μία κατηγορία περιεχομένου. Έπειτα, χρησιμοποιήθηκε μια εντολή **$match** για να μην συμπεριληφθούν οι εγγραφές της συλλογής χωρίς καταχωρημένη κατηγορία περιεχομένου. Ακόλουθα, έγινε χρήση μιας εντολής **$group** για την συγκέντρωση όλων των εγγραφών με κοινή κατηγορία περιεχομένου και την καταμέτρηση των εγγραφών αυτών για την εύρεση του πλήθους των σειρών και ταινιών (παραγωγών) που ανήκουν σε κάθε είδος. Τέλος, χρησιμοποιήθηκε μια εντολή **$sort** για την ταξινόμιση σε φθίσουσα σειρά ως προς το πλήθος των παραγωγών.

1	_id	count
2	International Movies	1844
3	Drama	988
4	Comedie	730
5	Dramas	635
6	TV Dramas	571
7	International TV Show	568
8	Independent Movies	535
9	Action & Adventur	529
10	International TV Shows	433
11	TV Comedies	399
12	Thrillers	392
13	Comedies	383
14	Romantic Movies	376
15	Documentarie	345
16	Documentaries	323
17	Crime TV Show	309
18	Stand-Up Comedy	281
19	Romantic TV Shows	257
20	Children & Family Movie	238
21	Music & Musicals	237

Σχόλια:

Εμφανίστηκαν όπως περιμέναμε τα πεδία _id,count ταξινομιμένα ως προς το count, δηλαδή το πλήθος των παραγωγών της κάθε κατηγορίας με φθίνουσα σειρά. Επίσης, οι εγγραφές χωρίς καταχωρημένη κατηγορία περιεχομένου δεν έχουν εμφανιστεί όπως και θέλαμε.

iv. Εμφανιζόμενοι ηθοποιοί

Χρησιμοποιήθηκε μια εντολή **$unwind** για να εφανιστούν σαν ξεχωριστές εγγραφές, οι εγγραφές με περισσότερους από έναν ηθοποιούς. Έπειτα, χρησιμοποιήθηκε μια εντολή **$match** για να μην συμπεριληφθούν οι εγγραφές της συλλογής χωρίς καταχωρημένο ηθοποιό. Ακόλουθα, έγινε χρήση μιας εντολής **$group** για την συγκέντρωση όλων των εγγραφών με κοινό ηθοποιό και την καταμέτρηση των εγγραφών αυτών για την εύρεση του πλήθους των σειρών και ταινιών (παραγωγών) στις οποίες έχει συμμετάσχει κάθε ηθοποιός. Επιπλέον, χρησιμοποιήθηκε μια εντολή **$sort** για την ταξινόμιση σε φθίσουσα σειρά ως προς το πλήθος των παραγωγών. Τέλος, χρησιμοποιήθηκε μια εντολή **$limit** για να εμφανιστούν οι 20 συχνότερα εμφανιζόμενοι ηθοποιοί σε παραγωγές.

1	id	count
2	Anupam Kher	30
3	Om Puri	26
4	Takahiro Sakurai	24
5	Shah Rukh Kha	24
6	Boman Irani	23
7	Paresh Rawal	22
8	Andrea Libman	22
9	Yuki Kaji	22
10	Naseeruddin Shah	19
11	Akshay Kuma	19
12	David Attenborough	18
13	John Cleese	18
14	Gulshan Grover	17
15	Ashleigh Ball	16
16	Erin Fitzgerald	16
17	Vincent Tong	16
18	Fred Tatasciore	16
19	Tomokazu Sugita	15
20	Asrani	15
21	Michael Palin	15

Σχόλια:

Εμφανίστηκαν όπως περιμέναμε τα πεδία _id,count ταξινομιμένα ως προς το count, δηλαδή το πλήθος των παραγωγών του κάθε ηθοποιού με φθίνουσα σειρά. Επίσης, οι εγγραφές χωρίς καταχωρημένο ηθοποιό δεν έχουν εμφανιστεί όπως και θέλαμε.

v. Κορυφαίες προτιμήσεις ηθοποιών

Χρησιμοποιήθηκε μια εντολή **$unwind** για να εφανιστούν σαν ξεχωριστές εγγραφές, οι εγγραφές με περισσότερους από έναν ηθοποιούς. Χρησιμοποιήθηκε μια εντολή **$unwind** για να εφανιστούν σαν ξεχωριστές εγγραφές, οι εγγραφές με περισσότερους μία κατηγορία περιεχομένου. Έπειτα, χρησιμοποιήθηκε μια εντολή **$match** για να μην συμπεριληφθούν οι εγγραφές της συλλογής χωρίς καταχωρημένο ηθοποιό και κατηγορία περιεχομένου. Ακόλουθα, έγινε χρήση μιας εντολής **$group** για την συγκέντρωση όλων των εγγραφών με κοινό ηθοποιό και ταυτόχρονα κοινή κατηγορία περιεχομένου και για την καταμέτρηση των εγγραφών αυτών για την εύρεση του πλήθους των σειρών και ταινιών (παραγωγών) κάθε μίας κατηγορίας περιεχομένου που έχει συμμετάσχει κάθε ηθοποιός. Επιπλέον, χρησιμοποιήθηκε μια εντολή **$sort** για την ταξινόμιση σε φθίσουσα σειρά ως προς το πλήθος των παραγωγών κάθε κατηγορίας περιεχομένου του κάθε ηθοποιού. Επιπρόσθετα, έγινε χρήση μιας εντολής **$group** για την επιλογή της πρώτης εγγραφής με το μεγαλύτερο πλήθος παραγωγών του κάθε ηθοποιού. Τέλος, χρησιμοποιήθηκε μια εντολή **$sort** για την ταξινόμιση σε αύξουσα σειρά ως προς τα ονόματα των ηθοποιών.

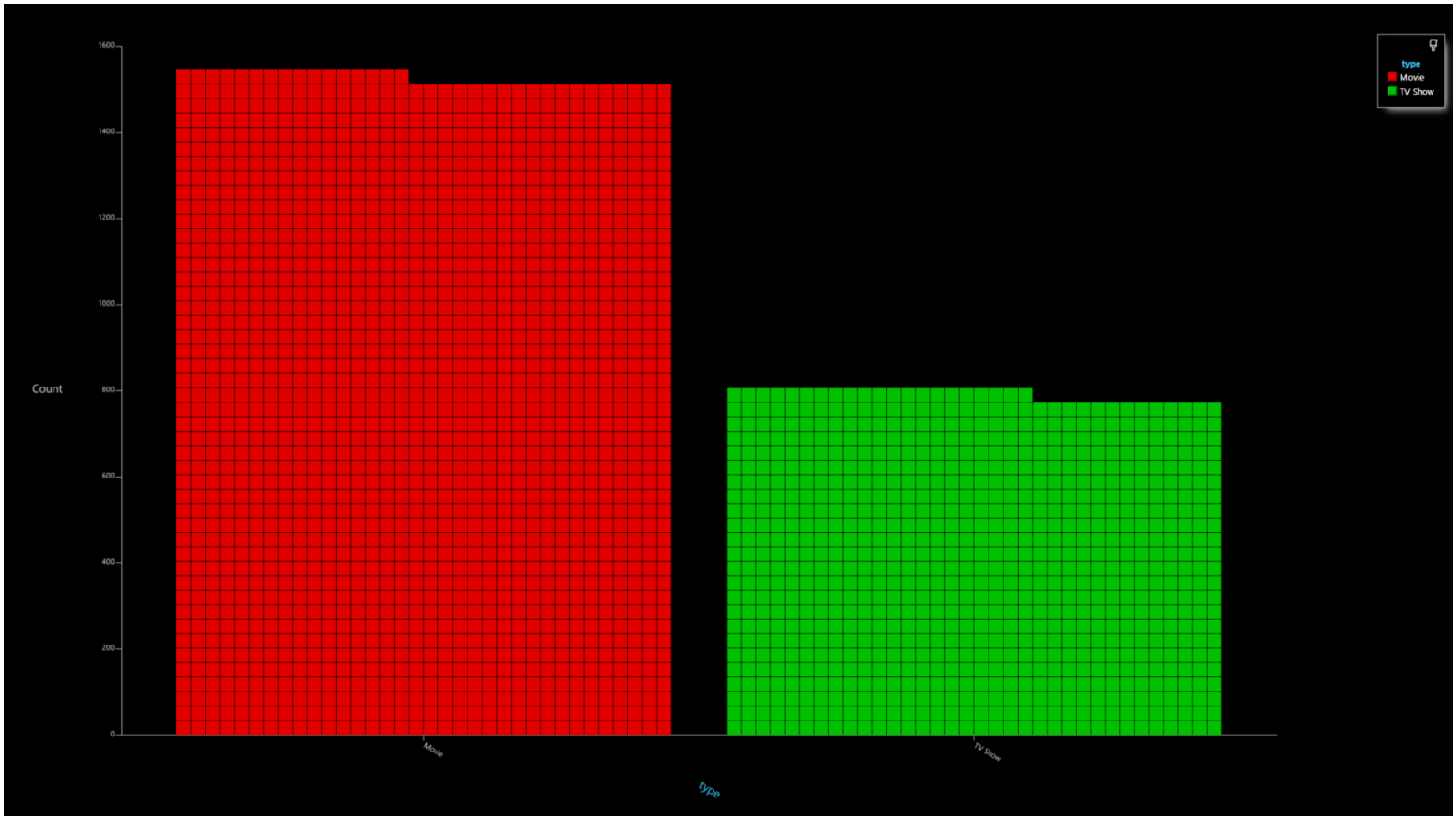
1	_id	listed_in	count
2	Jr.	TV Dramas	1
3	2 Chainz	Docuseries	1
4	4Minut	Music & Musicals	1
5	50 Cen	Action & Adventure	1
6	50 Cent	Thrillers	1
7	A Boogie Wit tha Hoodie	Docuseries	1
8	A-ra Go	International TV Shows	1
9	A. Murat Ã–zgen	Horror Movie	1
10	A.C. Peterson	Dramas	1
11	A.D. Miles	TV Comedies	2
12	A.J. Cook	Crime TV Show	1
13	A.J. LoCasci	Children & Family Movie	1
14	A.J. LoCascio	Kids' TV	2
15	A.K. Hangal	International Movies	3
16	A.R. Rahman	International Movies	1
17	A.S. Sasi Kumar	Independent Movies	1
18	AFRA	TV Horror	1
19	AJ Bowen	LGBTQ Movies	1
20	AJ Michalka	Kids' T	1

Σχόλια:

Εμφανίστηκαν όπως περιμέναμε τα πεδία _id, listed_in και count ταξινομιμένα ως προς το _ιδ, δηλαδή το όνομα των ηθοποιών με αύξουσα σειρά. Επίσης, οι εγγραφές χωρίς καταχωρημένο ηθοποιό ή κατηγορία περιεχομένου δεν έχουν εμφανιστεί όπως και θέλαμε.

Bonus υλοποίηση:
Για τις γραφικές απεικονίσεις χρησιμοποιήθηκε το εργαλείο Visual Studio Code και εγκαταστάθηκε η επέκταση SandDance.
Το SandDance χρησιμοποιεί οπτικοποιήσεις μονάδας, οι οποίες εφαρμόζουν αντιστοίχιση μεταξύ των σειρών στη βάση δεδομένων και εκτυπώνει σημάδια στην οθόνη.

i. Διαθέσιμο περιεχόμενο το 2019

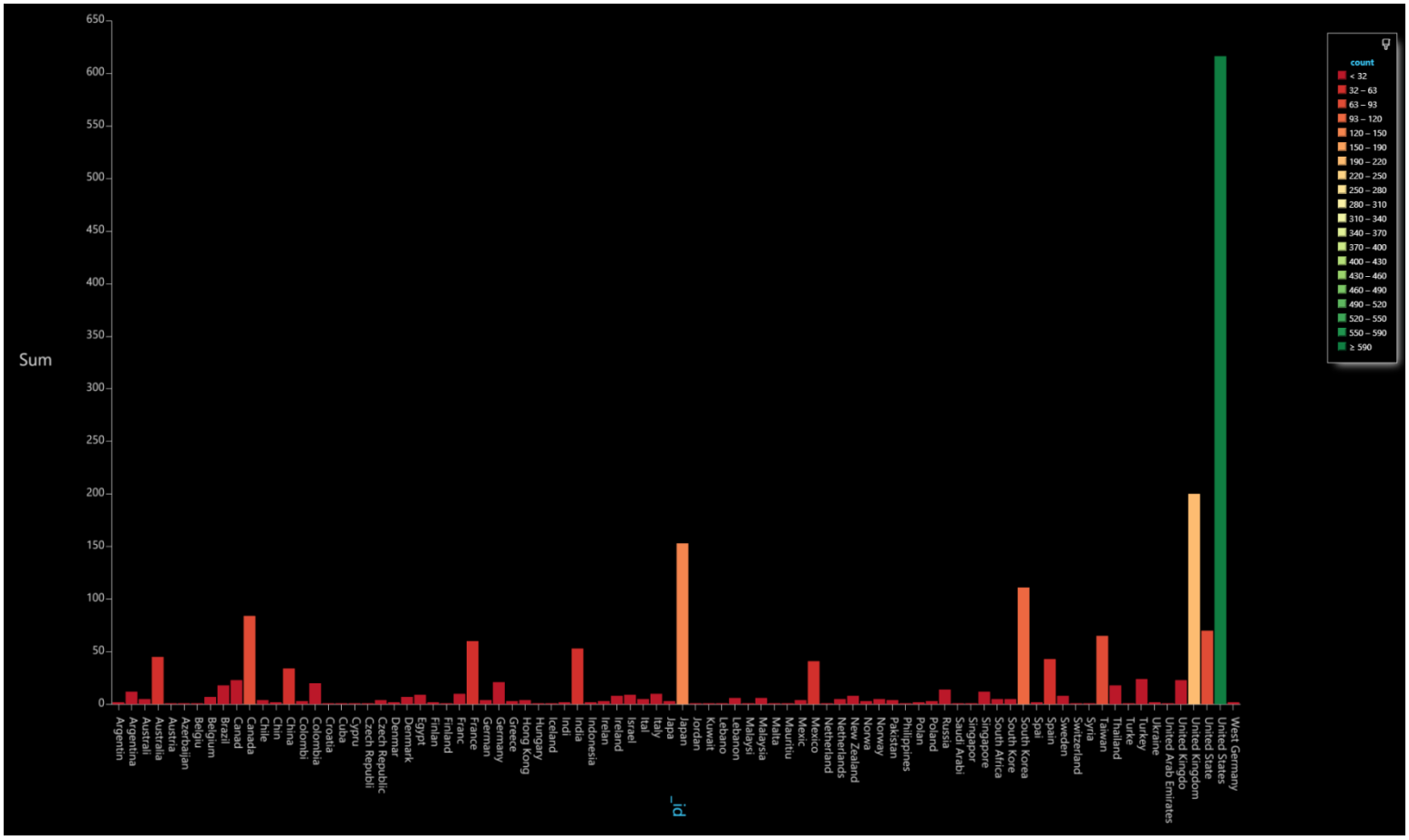


Στην παραπάνω εικόνα φαίνεται το πλήθος των ταινιών και το πλήθος των σειρών που ήταν διαθέσιμες το 2019 από την πλατφόρμα Netflix.

Για την αναπαράσταση των δεδομένων έγινε χρήση ενός διαγράμματος στηλών (Column Chart) σε δύο διαστάσεις, επειδή ένα χαρακτηριστικό προς εκτύπωση είναι ονομαστικό (nominal).

Σχόλια:
Παρατηρούμε ότι το 2019 η πλατφόρμα Netflix διέθετε σχεδόν διπλάσιες ταινίες απ'ότι σειρές κάτι το οποίο είναι αναμενόμενο σε μια πλατφόρμα.

ii. Χώρες παραγωγής σειρών

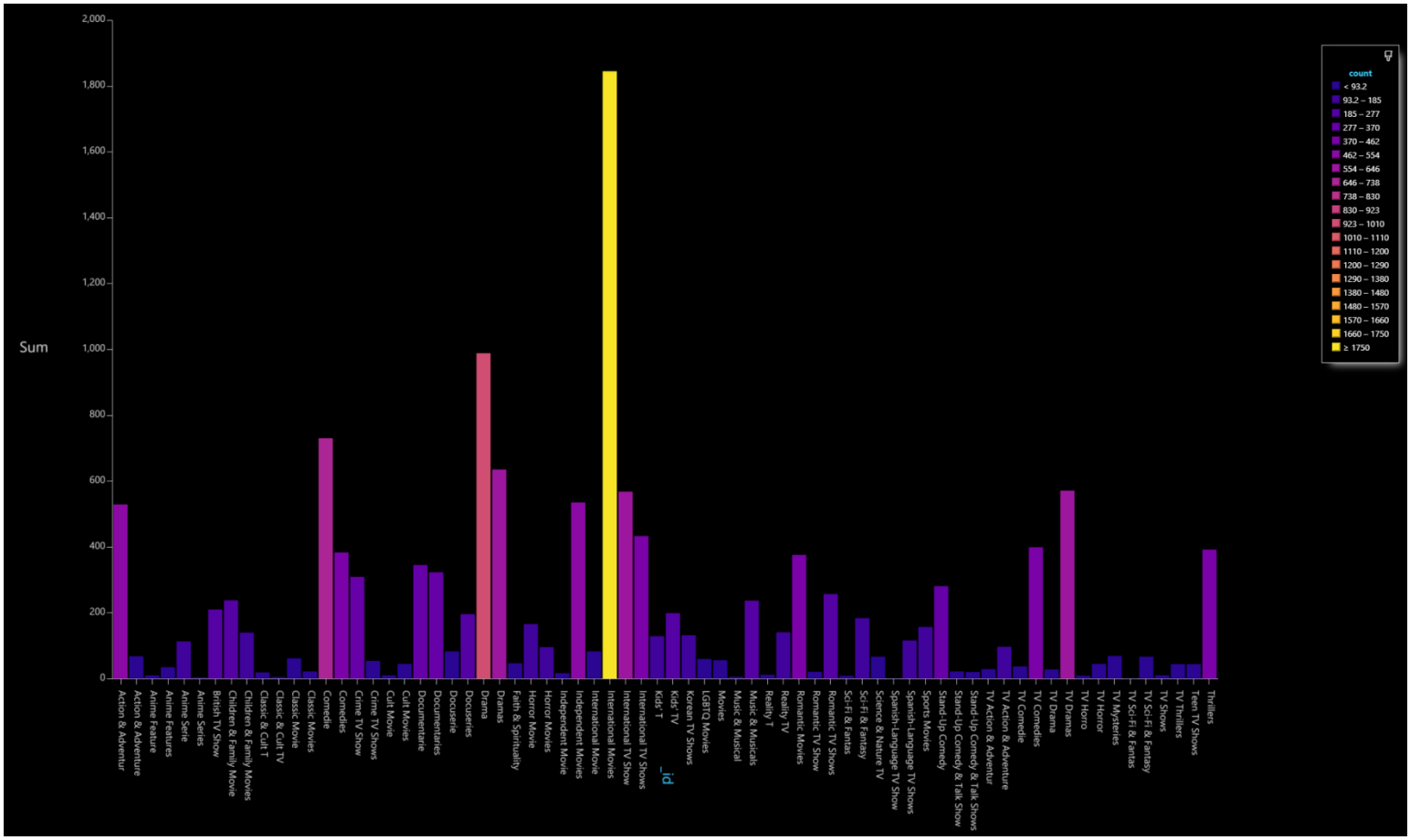


Στην παραπάνω εικόνα φαίνεται το πλήθος των σειρών ανά διαφορετική χώρα παραγωγής που είναι διαθέσιμες από την πλατφόρμα Netflix.

Για την αναπαράσταση των δεδομένων έγινε χρήση ενός διαγράμματος στηλών (Column Chart) σε δύο διαστάσεις, επειδή ένα χαρακτηριστικό προς εκτύπωση είναι ονομαστικό (nominal).

<u>Σχόλια:</u>
Παρατηρούμε ότι οι χώρες με τις περισσότερες παραγωγές είναι οι Ηνωμένες Πολιτίες, το Ηνωμένο Βασίλειο, η Ιαπωνία και η Νότια Κορέα κάτι το οποίο είναι αναμενόμενο διότι όλες αυτές οι χώρες, είναι χώρες που έχουν παγκοσμίως πολλές παραγωγές. Επίσης, οι Ηνωμένες Πολιτίες εμφανίζονται με τις περισσότερες παραγωγές από τις υπόλοιπες χώρες διότι η πλατφόρμα Netflix έχει την έδρα της εκεί.

iii. Είδη διαθέσιμου περιεχομένου

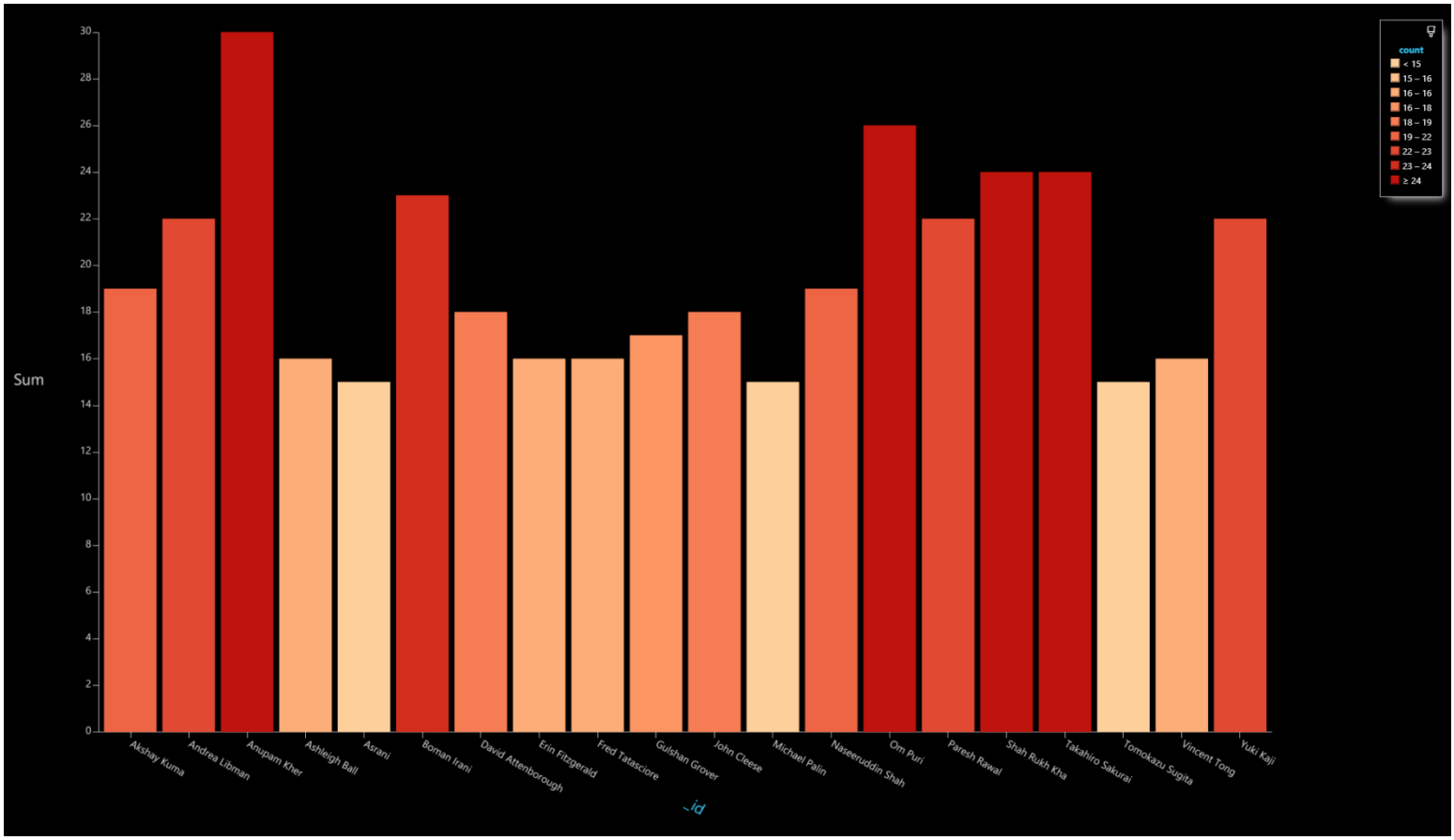


Στην παραπάνω εικόνα φαίνεται το σύνολο των ταινιών και σειρών που ανήκουν σε κάθε κατηγορία περιεχομένου που είναι διαθέσιμες από την πλατφόρμα Netflix.

Για την αναπαράσταση των δεδομένων έγινε χρήση ενός διαγράμματος στηλών (Column Chart) σε δύο διαστάσεις, επειδή ένα χαρακτηριστικό προς εκτύπωση είναι ονομαστικό (nominal).

<u>Σχόλια:</u>
Παρατηρούμε ότι η κατηγορία περιεχομένου με τις περισσότερες παραγωγές είναι οι διεθνείς τηλεοπτικές εκπομπές (International TV Shows), κάτι που είναι φυσιολογικό διότι συνεισφέρουν σε αυτήν την κατηγορία σχεδόν όλες οι χώρες.

iv. Εμφανιζόμενοι ηθοποιοί

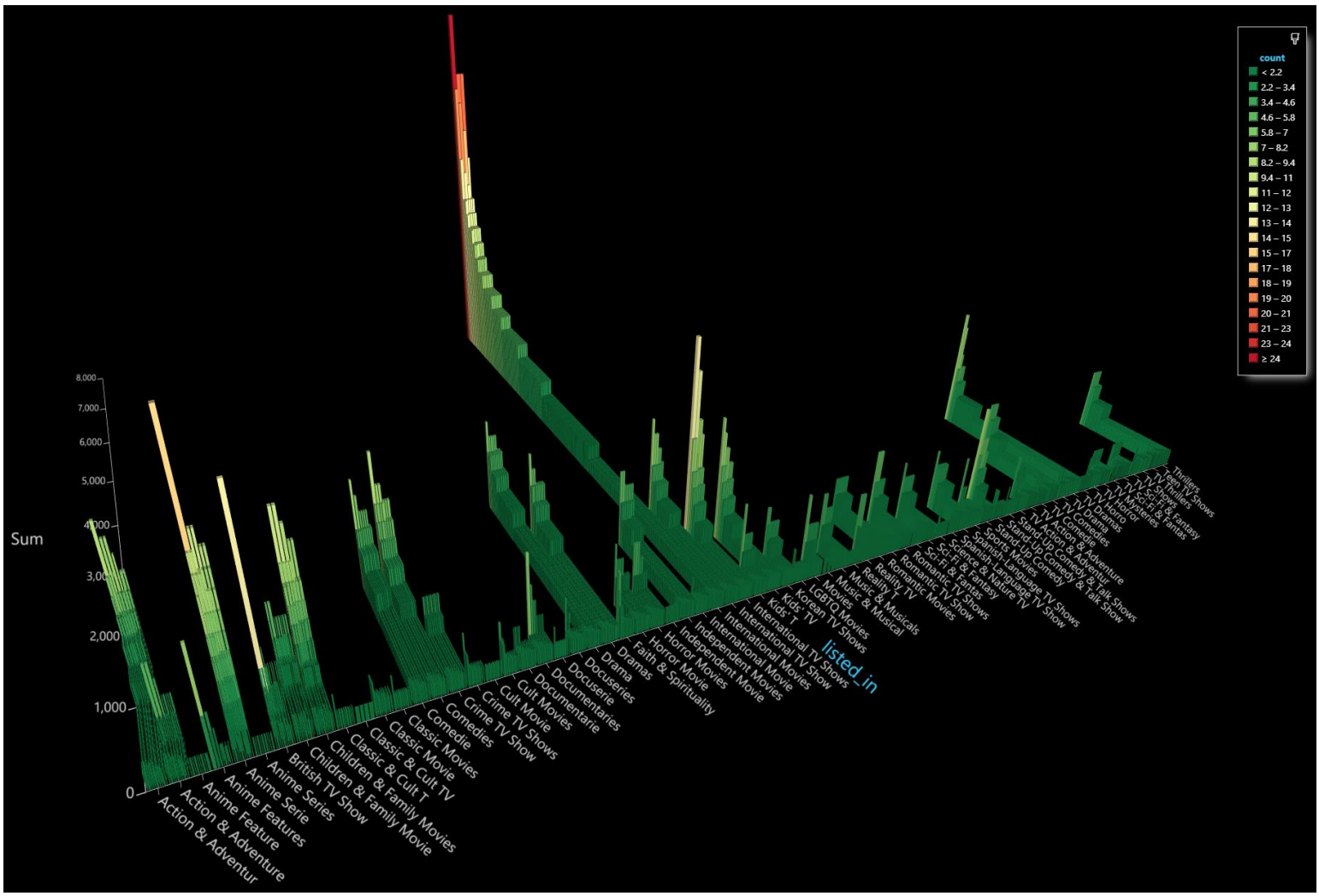


Στην παραπάνω εικόνα φαίνονται οι 20 συχνότερα εμφανιζόμενοι ηθοποιοί σε παραγωγές καθώς και το πλήθος των παραγωγών του κάθε ηθοποιού στην πλατφόρμα Netflix.

Για την αναπαράσταση των δεδομένων έγινε χρήση ενός διαγράμματος στηλών (Column Chart) σε δύο διαστάσεις, επειδή ένα χαρακτηριστικό προς εκτύπωση είναι ονομαστικό (nominal).

<u>Σχόλια:</u>
Παρατηρούμε ότι οι 20 συχνότερα εμφανιζόμενοι ηθοποιοί έχουν κατά μέσο όρο περίπου 20 παραγωγές με τις περισσότερες να είναι 30 και τις λιγότερες να είναι 15.

v. Κορυφαίες προτιμήσεις ηθοποιών



Στην παραπάνω εικόνα φαίνονται όλοι οι ηθοποιοί στην κατηγορία περιεχομένου στην οποία έχουν τις περισσότερες παραγωγές. Η κάθε στήλη αναπαριστά έναν ηθοποιό και το ύψος των στηλών αναπαριστά το πλήθος των παραγωγών του κάθε ηθοποιού.

Για την αναπαράσταση των δεδομένων έγινε χρήση ενός διαγράμματος στηλών (Column Chart) σε τρείς διαστάσεις, επειδή ένα χαρακτηριστικό προς εκτύπωση είναι ονομαστικό (nominal).

<u>Σχόλια:</u>
Παρατηρούμε ότι οι περισσότερες παραγώγες ανά κατηγορία περιεχομένου είναι στην κατηγορία διεθνείς τηλεοπτικές εκπομπές το οποίο και περιμέναμε, ένω στην κατηγορία αυτή βρίσκεται επίσης και ο ηθοποιός με τις περισσότερες παραγωγές.