Datavisualisatie-groep 6: logboek

Taakverdeling

Taakverdeling: 29 maart-22 April

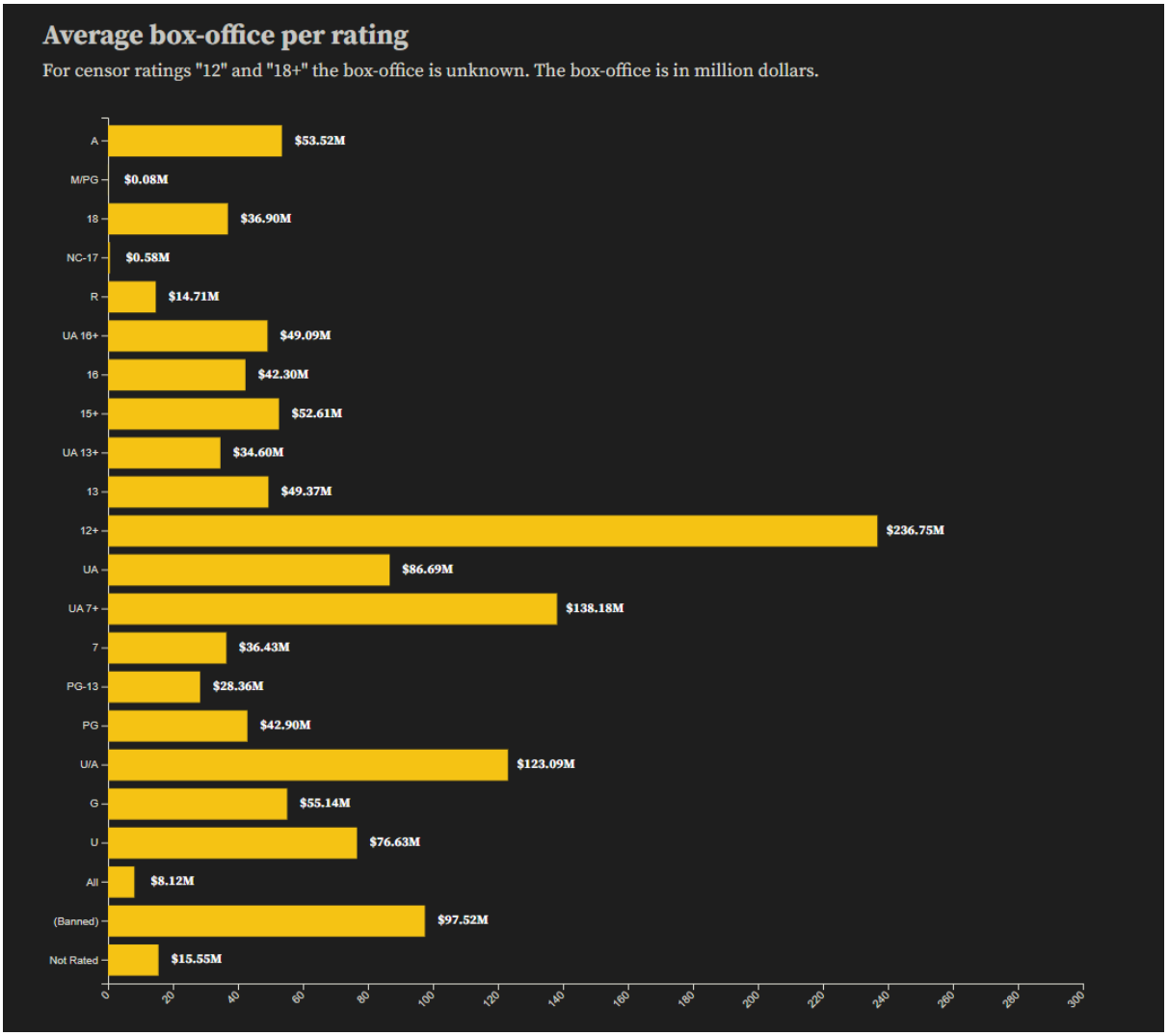
Groepslid	Taken
Bavo	• scatter plot: gross income vs score (do they match?)
Brent	• scatter plot. x: # movies, y: average movie score per actor/director. • bar chart: Welke rating heeft gemiddeld de grootste box-office
Pjotr	• wordcloud van woorden in filmtitels • words in title vs average gross income.
Kai	• bar plot: genre gemiddeld gross income per jaar (stacked)

Bar chart: Average box-office per rating

Het doel van deze toevoeging was om een eenvoudige staafdiagram te maken die de gemiddelde box-office-omzet weergeeft voor censuur classificaties. Dit zou gebruikers een helder overzicht geven van hoe de box-office-scores zich verhouden tot de verschillende filmratings.

Er was echter een probleem met deze benadering. Sommige van de censuur classificaties kwamen zelden voor, wat het gemiddelde vertekende. Als je naar de grafiek keek, zag je bijvoorbeeld dat de 12+-rating een zeer hoog gemiddelde had, maar dat was gebaseerd op slechts een handvol films in de dataset. Dit gaf een misleidend beeld.

Om dit te corrigeren en meer informatie te bieden, besloten we om in plaats van een staafdiagram een boxplot te gebruiken. Met de boxplot kun je niet alleen het gemiddelde zien, maar ook de spreiding, mediaan, en eventuele outliers. Dit biedt een duidelijker en completer beeld van de box-office-scores voor elke rating.

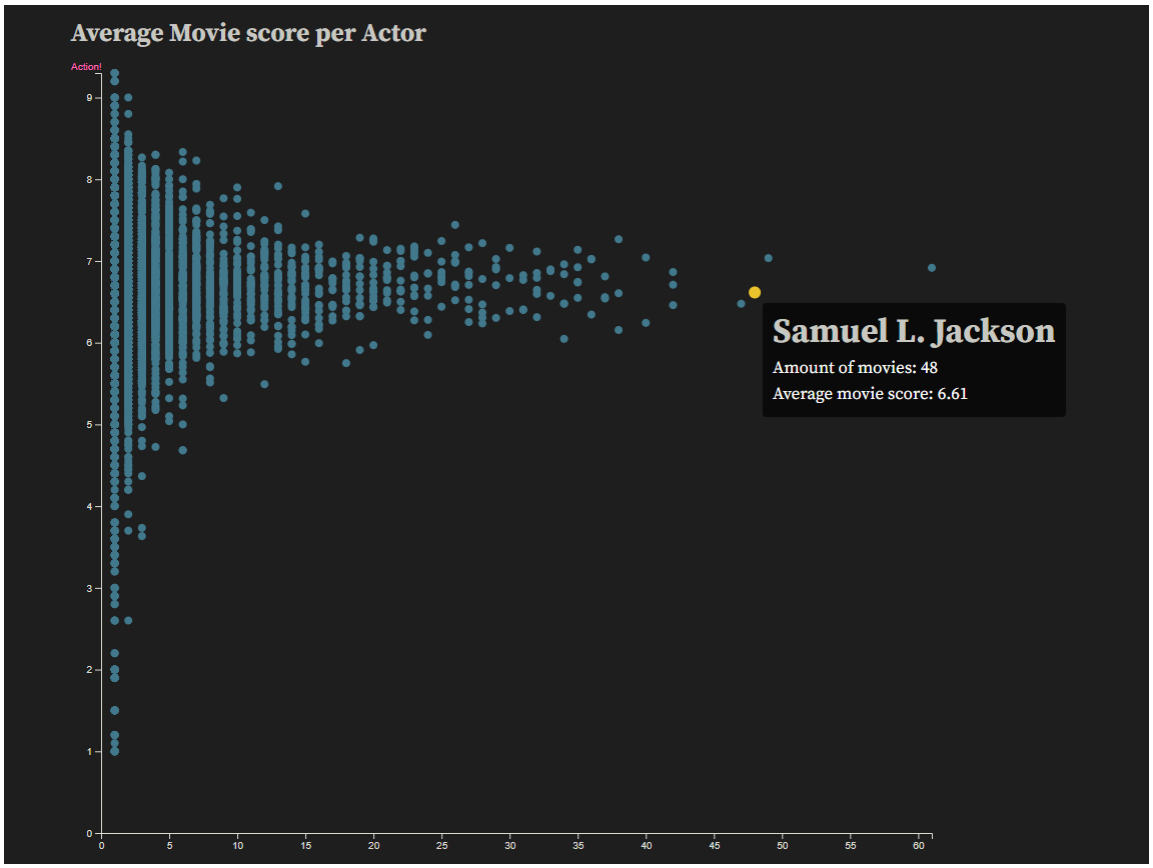


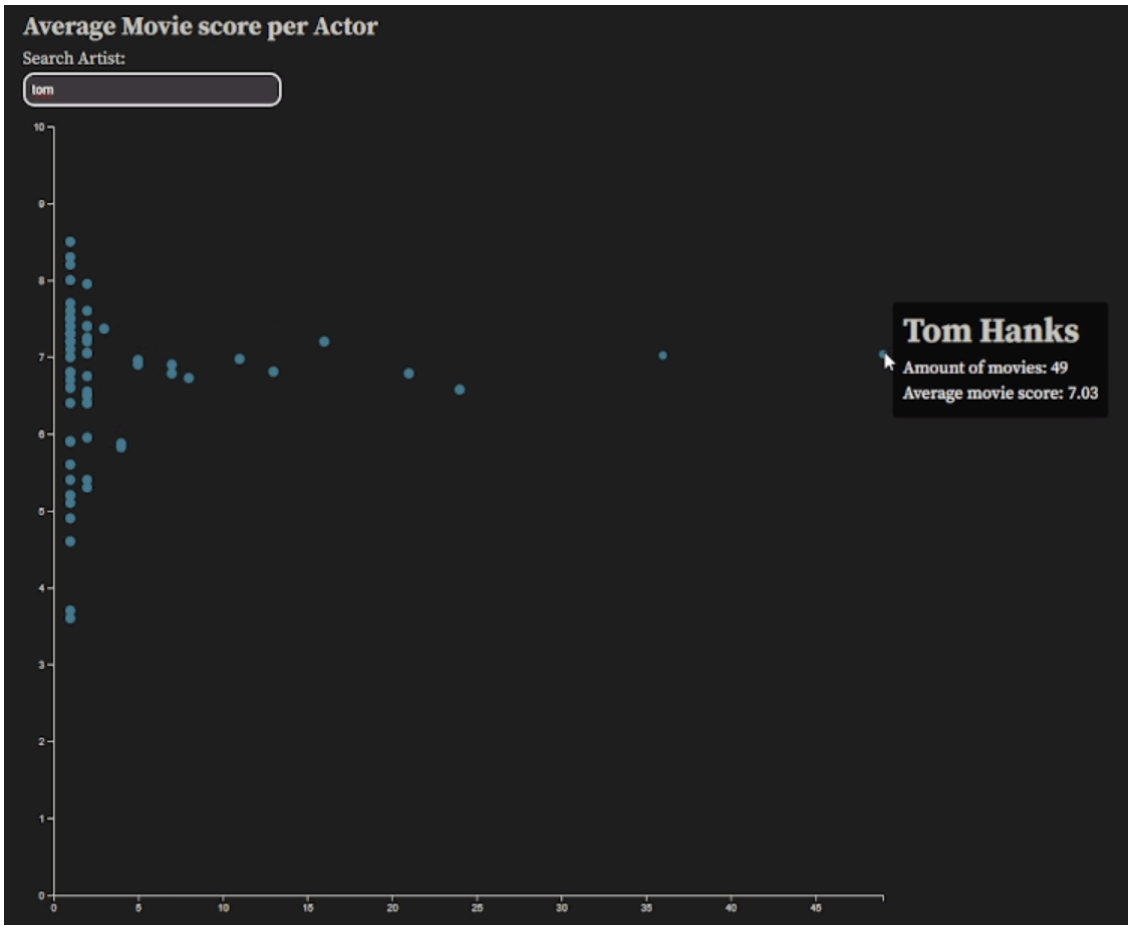
Scatter plot: Average movie score per actor/director

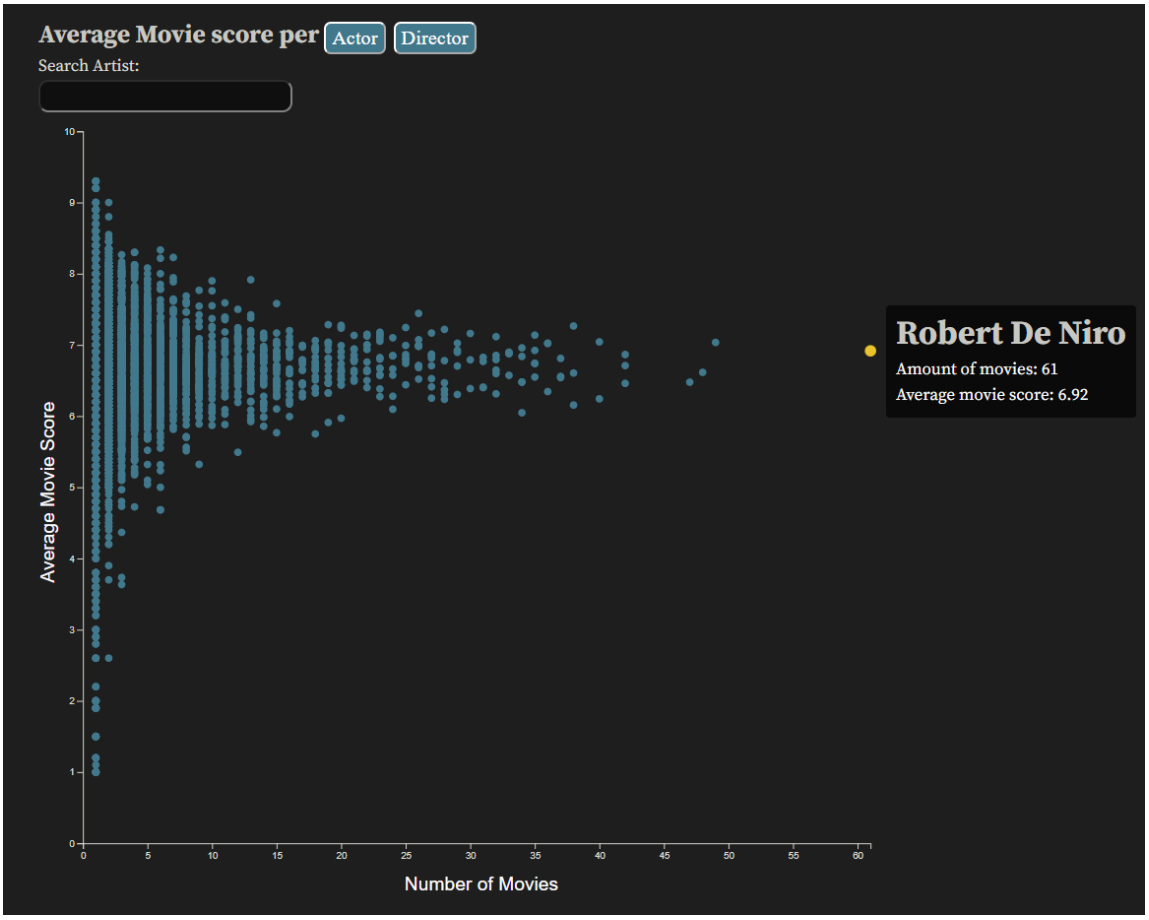
Deze scatterplot is gemaakt om elke stip te associëren met een acteur of regisseur. De x-as vertegenwoordigt het aantal films waarin ze hebben gespeeld of die ze hebben geregisseerd, en de y-as toont de gemiddelde score van al deze films. Het doel was om te kijken of er een verband bestaat tussen het aantal films dat iemand heeft gedaan en de gemiddelde beoordeling van die films. We wilden ook de mogelijkheid bieden om acteurs of regisseurs op te zoeken om hun prestaties in kaart te brengen.

We begonnen met het maken van deze scatterplot voor alleen de acteurs. We hebben ook tooltips toegevoegd, zodat je kunt zien welke acteur elk punt op de plot vertegenwoordigt. Zo kun je gemakkelijk de details bekijken van het aantal films en de gemiddelde score voor elke acteur.

Nadat dit goed werkte, hebben we een zoekfunctie toegevoegd en twee knoppen waarmee je kunt kiezen om alle regisseurs of alle acteurs te zien. Zo kun je snel schakelen tussen verschillende groepen en specifieke namen opzoeken binnen de scatterplot.


Average Movie score per Actor
Action!
Samuel L. Jackson
Amount of movies: 48
Average movie score: 6.61


Average Movie score per Actor
Search Artist:
tom
Tom Hanks
Amount of movies: 49
Average movie score: 7.03



Scatter plot: Movie score vs. box-office

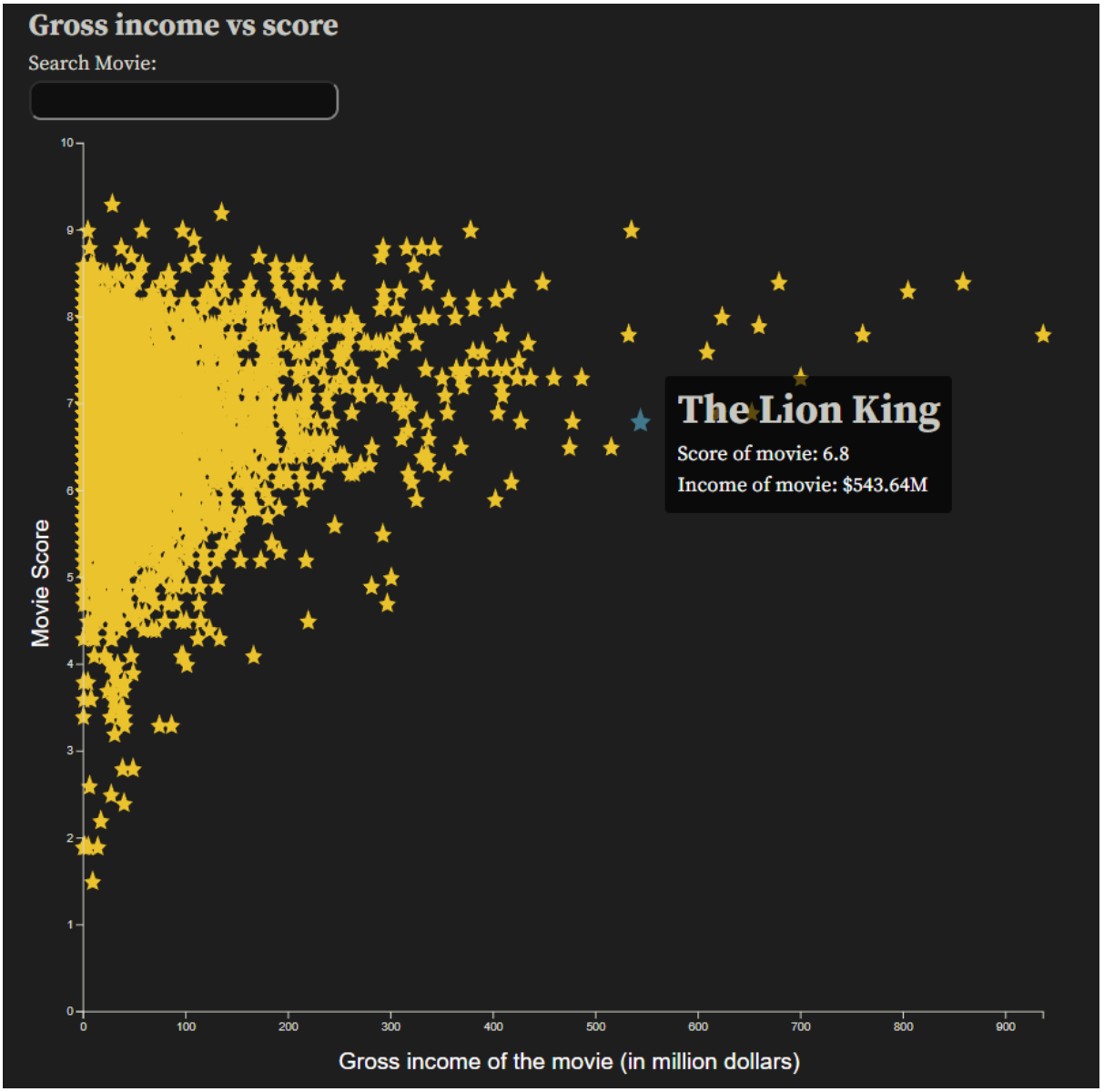
We hebben deze plot gemaakt met als doel om aan te tonen hoe belangrijk de kwaliteit van de film is, als het enige doel is om geld te verdienen. Het voordeel van een scatterplot is dat je de volledige correlatie kan waarnemen. Op deze manier kan je nagaan of een hoge score een noodzakelijke voorwaarde is, maar ook of een hoge score een voldoende voorwaarde is.
De x-as toont hoeveel miljoen dollar de film heeft opgebracht, en de y-as wat zijn score op IMDB was.

Als eerste iteratie was de plot heel gelijkaardig aan de andere scatter plot: dezelfde structuur, dezelfde kleuren, dezelfde zoekfunctie en dezelfde tooltips (allemaal natuurlijk op andere data). Om toch ervoor te zorgen dat de grafiek zijn eigen identiteit heeft, heb ik een aantal aanpassingen gedaan.

De grootste zulke aanpassing is de vorm van de datapunten. In de plaats van simpelweg cirkels te gebruiken, wou ik deze veranderen door sterren. Dit bleek echter een heel moeilijke en tijdrovende aanpassing.De enige manier om sterren als datapunten te tonen, is door datapunten voor te stellen als ‘paths’. Dit zijn objecten met maar 1 attribuut ‘d’: een lijst van coördinaten om te verbinden en dan in te kleuren. Ik dan een functie geschreven om deze sterren te genereren. Aangezien

datapunten moeten een beetje groter worden als je de muis erover houdt, moest deze functie ook een scale parameter meekrijgen.

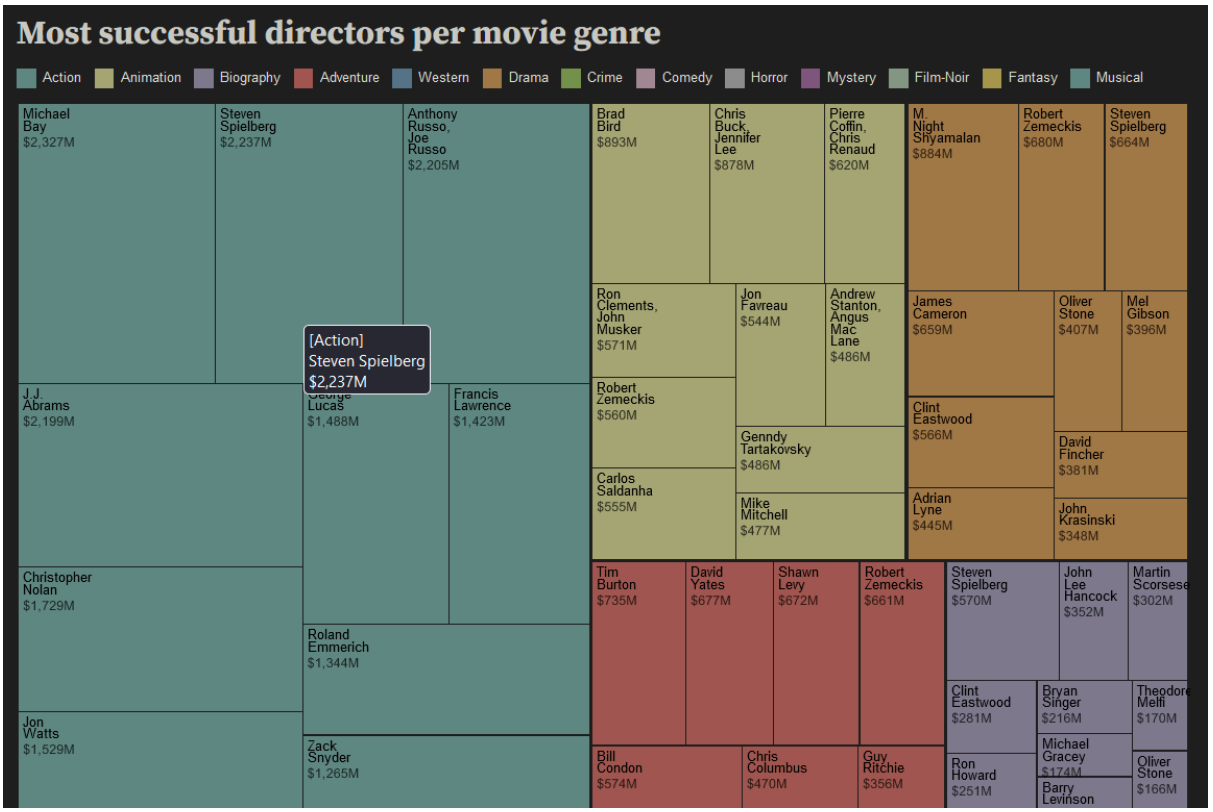
Er waren meer technische problemen dan deze (zoals de transitions, het feit da er meerdere paths zijn in een plot,...). Eens deze opgelost waren, heb ik nog als aanpassing de hover en standaard kleur omgewisseld: de sterren die standaard geel zijn oogt veel mooier dan een hoop blauwe sterren. Daarnaast zorgt dit weer voor een contrast met de andere scatter plot.



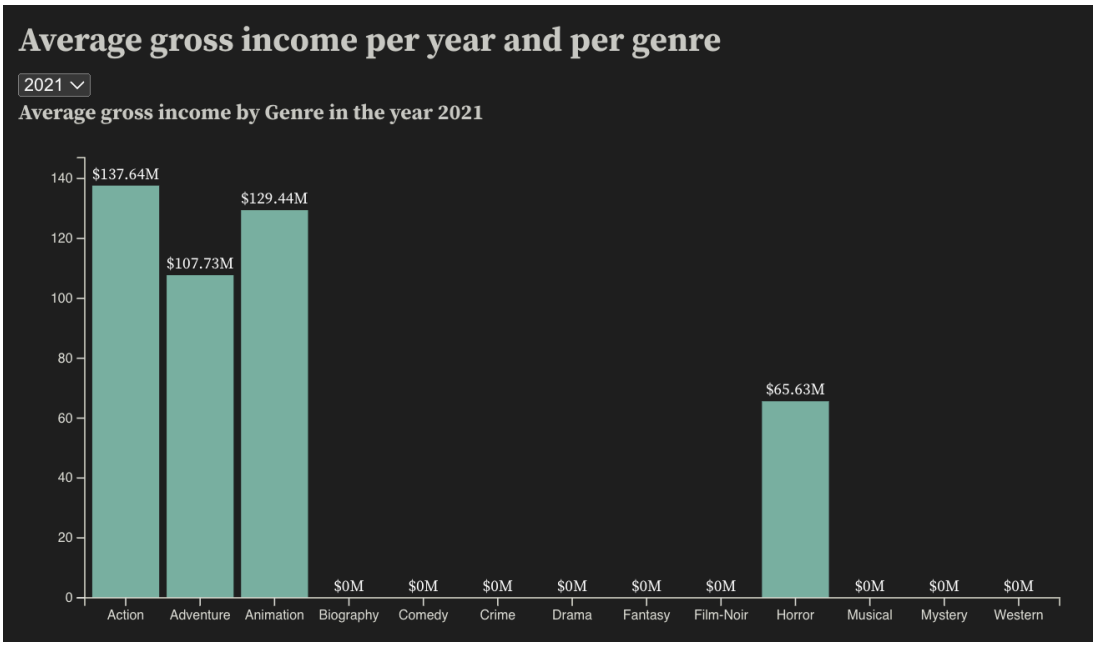
Wordcloud voor filmtitels

Het idee voor deze plots was om een idee te geven van welke woorden het meest gebruikt worden in filmtitels. Als uitbreiding hierop zou bekeken worden hoeveel de opbrengst zou zijn van films met bepaalde woorden in de titel en of er een verband te vinden zou zijn tussen de frequentie van deze woorden en de opbrengst.

Dit leek uiteindelijk toch niet zo'n goed idee en bestaande implementaties van wordclouds leken bovendien niet goed te werken.
Daarom is er gekeken naar alternatieve visualisaties waaruit een treemap een interessante keuze bleek. Hierop kunnen namelijk goed filmgenres en regisseurs gegroepeerd worden zodat per genre de best verdienende regisseurs te zien zijn; een blok van een regisseur is groter als deze meer verdiend heeft met zijn films.



Bar plot: Gemiddelde bruto-inkom per genre per jaar.



Deze plot stelt het gemiddelde bruto-inkom per genre per jaar voor. Met deze grafiek kunnen we bekijken welke genres populair zijn en dus het meest aantrekkelijk zijn voor de grote filmstudio's. Voor de eerste versie van de plot werd elk jaar geselecteerd aan de hand van een keuzebox. Sommige jaren hadden voor sommige genres geen film. Deze genres werden dan toegevoegd met een waarde van 0 miljoen om voor elk jaar dezelfde x-as te verkrijgen.

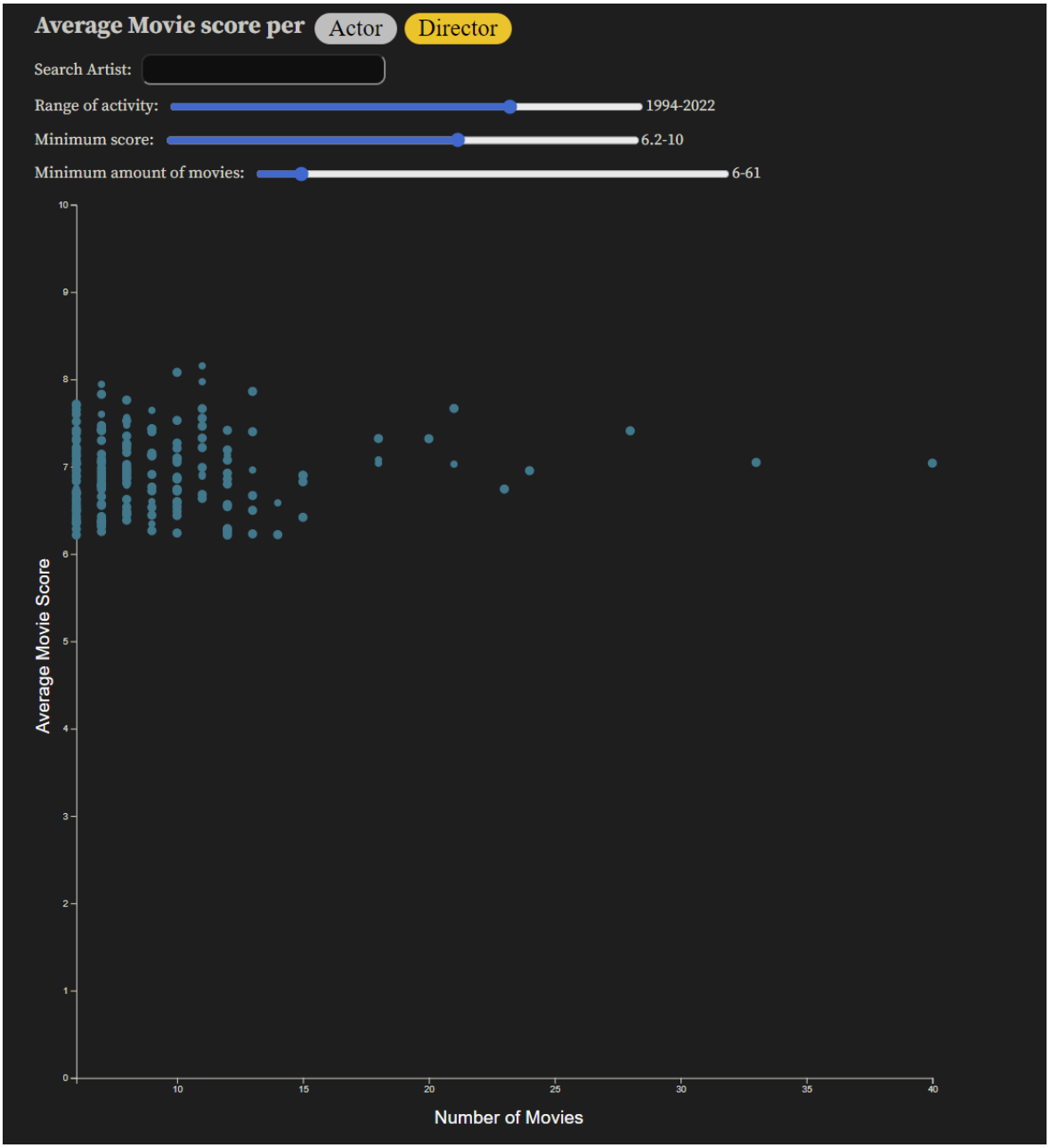
Taakverdeling: 22 April-6 Mei

Groepslid	Taken
Bavo	filter technieken toevoegen voor scatter plot
Brent	• tooltips fixen • filter technieken toevoegen voor scatter plot
Pjotr	• Treemap van meest succesvolle directors per filmgenre moet ook voor acteurs worden toegevoegd. • Maak de treemap ook meer interactief
Kai	• Maak een slider ipv van een dropdown voor de bar-chart voor genre gemiddeld gross income per jaar • Verander de bar chart voor “Welke rating heeft gemiddeld de grootste box-office”

Scatter plot: Average movie score per actor/director

Aanvankelijk ondervonden we problemen met de tooltips. Het probleem was dat de tooltip te ver van de muis af stond bij punten aan de rechterkant van de plot, maar juist veel te dicht bij punten aan de linkerkant. Dit veroorzaakte flikkeringen van de tooltips bij de punten aan de linkerkant. We hebben dit eenvoudig opgelost door de positie van de muis op een andere manier te bepalen, waardoor de tooltips nu soepeler en stabieler werken.

Een ander probleem met onze scatterplot was dat er enorm veel punten aanwezig waren op de plot. Hierdoor moesten we dus een manier bedenken om deze te filteren. We loste dit probleem op door te filteren op personen die nog actief waren in een bepaalde periode. We filteren ook op het aantal films dat iemand had gemaakt en wat zijn gemiddelde score is.



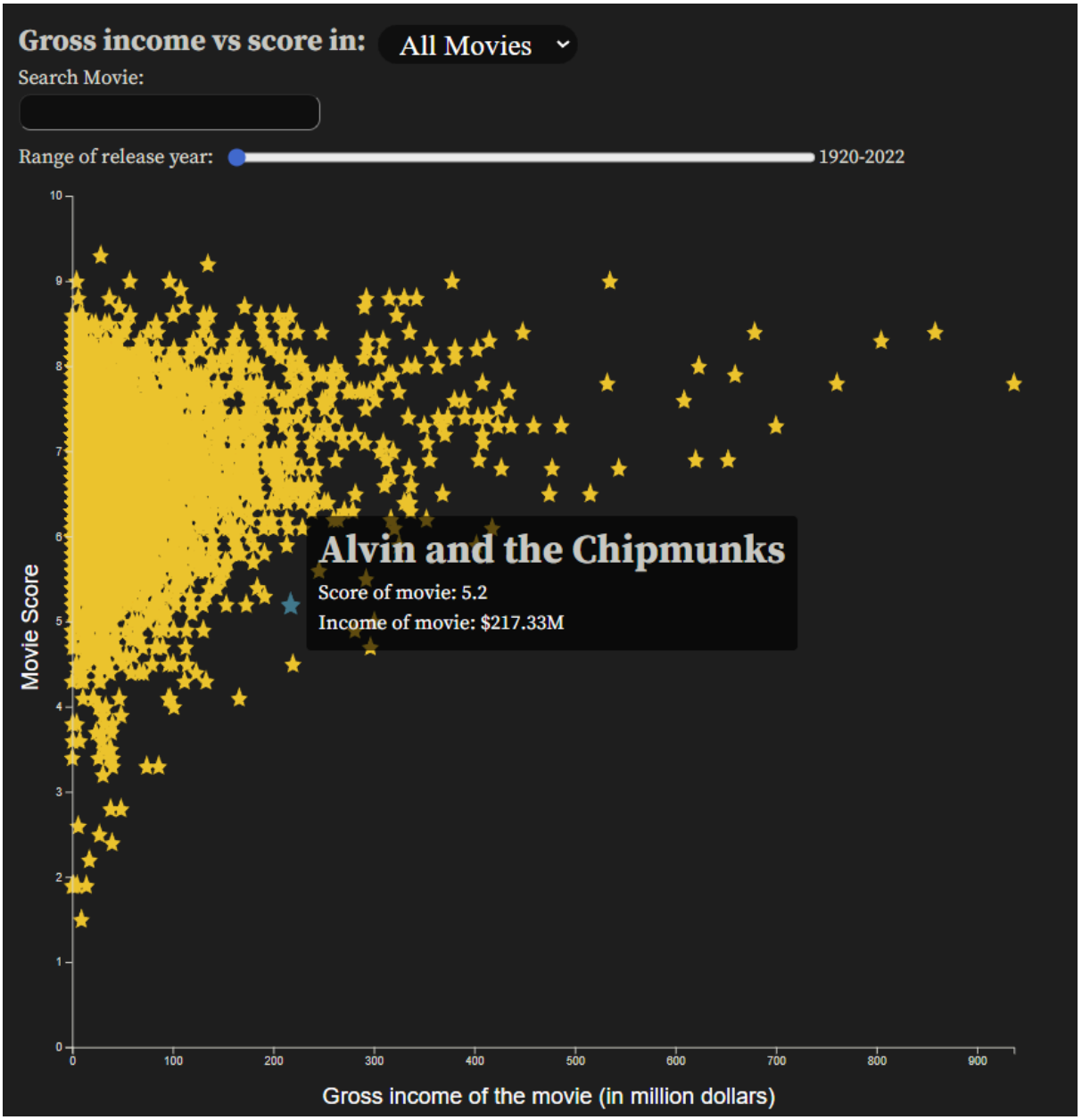
Scatter plot: Movie score vs. box-office

Deze scatter plot had exact dezelfde problemen als deze hierboven. De tooltips zijn op dezelfde manier opgelost.

Het te veel aan datapunten heb ik ook opgelost door het toevoegen van filters. Ten eerste heb ik, net zoals de andere scatter plot, een filter toegevoegd voor de recentheid van de films.

Een filter die ik ook heb toegevoegd, is om te kunnen filteren op genre. Dit is een logische filter: het valt te verwachten dat kijker bij verschillende genres verschillende verwachtingen hebben over kwaliteit. Deze genres staan gesorteerd. Natuurlijk is er

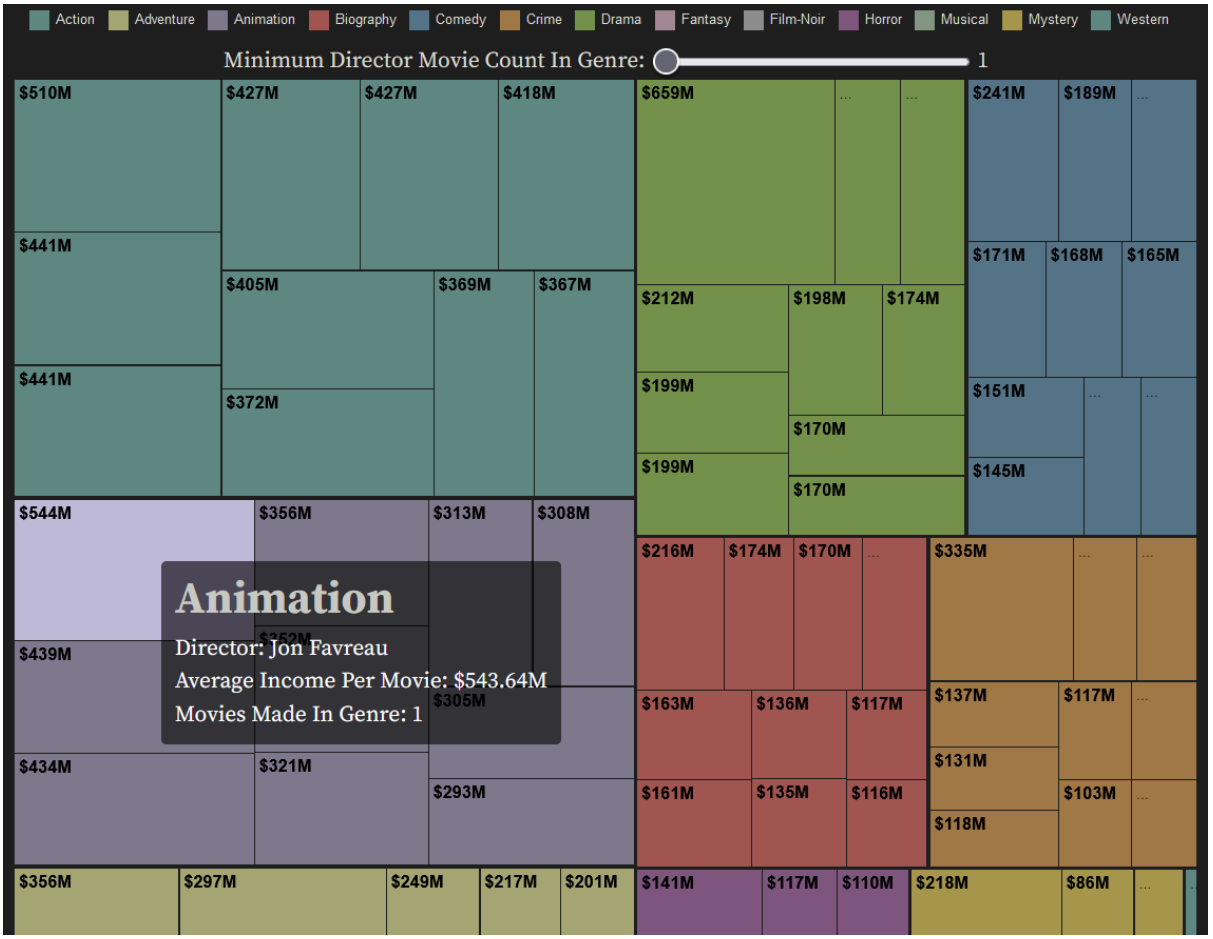
ook de optie gebleven om alle films tezamen in de plot te zetten, zodat ook het totaalplaatje kan bestudeerd worden.



Treemaps

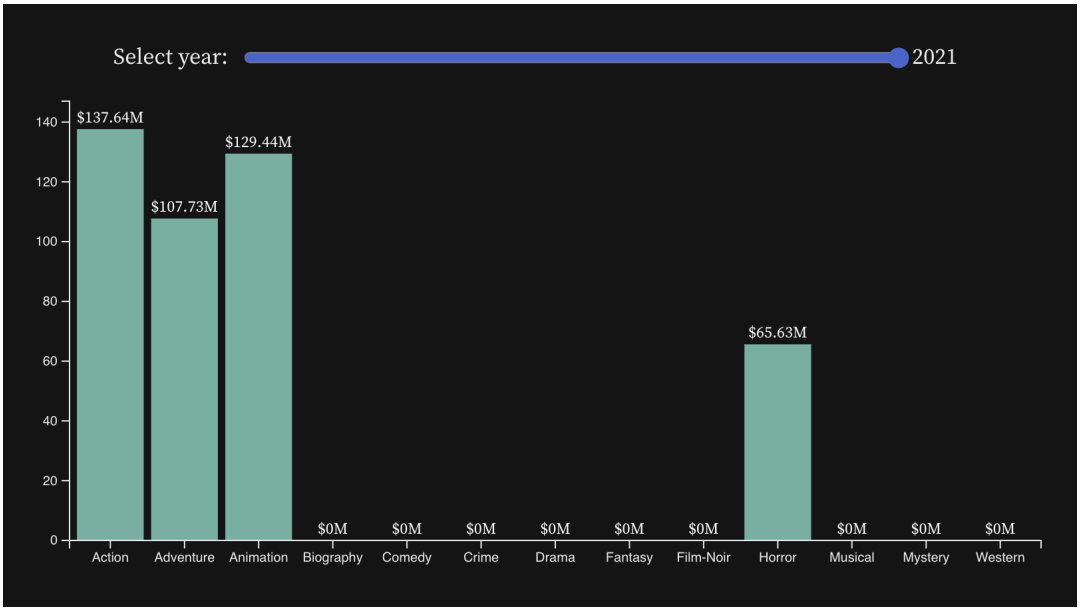
De styling van de treemap paste nog niet helemaal in het geheel van onze visualisaties en was ook niet interactief genoeg naar onze mening. Daarom is de styling aangepast zodat de tooltip er hetzelfde uitziet als bij de andere plots. Ook zijn de namen weggehaald van de rechthoeken om een ‘cleanere look’ te krijgen: als de muis over een rechthoek beweegt zal alle informatie getoond worden over de regisseur in kwestie. Verder wordt nu het gemiddelde inkomen per film per regisseur getoond in plaats van zijn totale inkomen, om een beter beeld te geven over hoe deze regisseur gemiddeld presteert.

Ten slotte is er nog een filter toegevoegd om het minimum aantal films te bepalen dat de regisseurs gemaakt hebben. Het gemiddelde inkomen zegt immers minder als een regisseur slechts één film heeft gemaakt die zeer goed gepresteerd heeft.



Line-plot: Gemiddelde bruto-inkom per genre per jaar.

Na het feedbackmoment is de keuzebox veranderd door een slider. Deze is makkelijker om te interpreteren en meer gebruiksvriendelijk dan een lange lijst van jaren om uit te kiezen. Voor het visueel effect is er ook een animatie toegevoegd bij het veranderen van jaartal. Origineel werd de grafiek gewoon verwijderd en vervangen door een andere grafiek. Met deze nieuwe animatie is de plot visueel aantrekkelijker en oogt het geheel professioneler dan voorheen.



Bar-chart average box rating per rating to rating boxplot

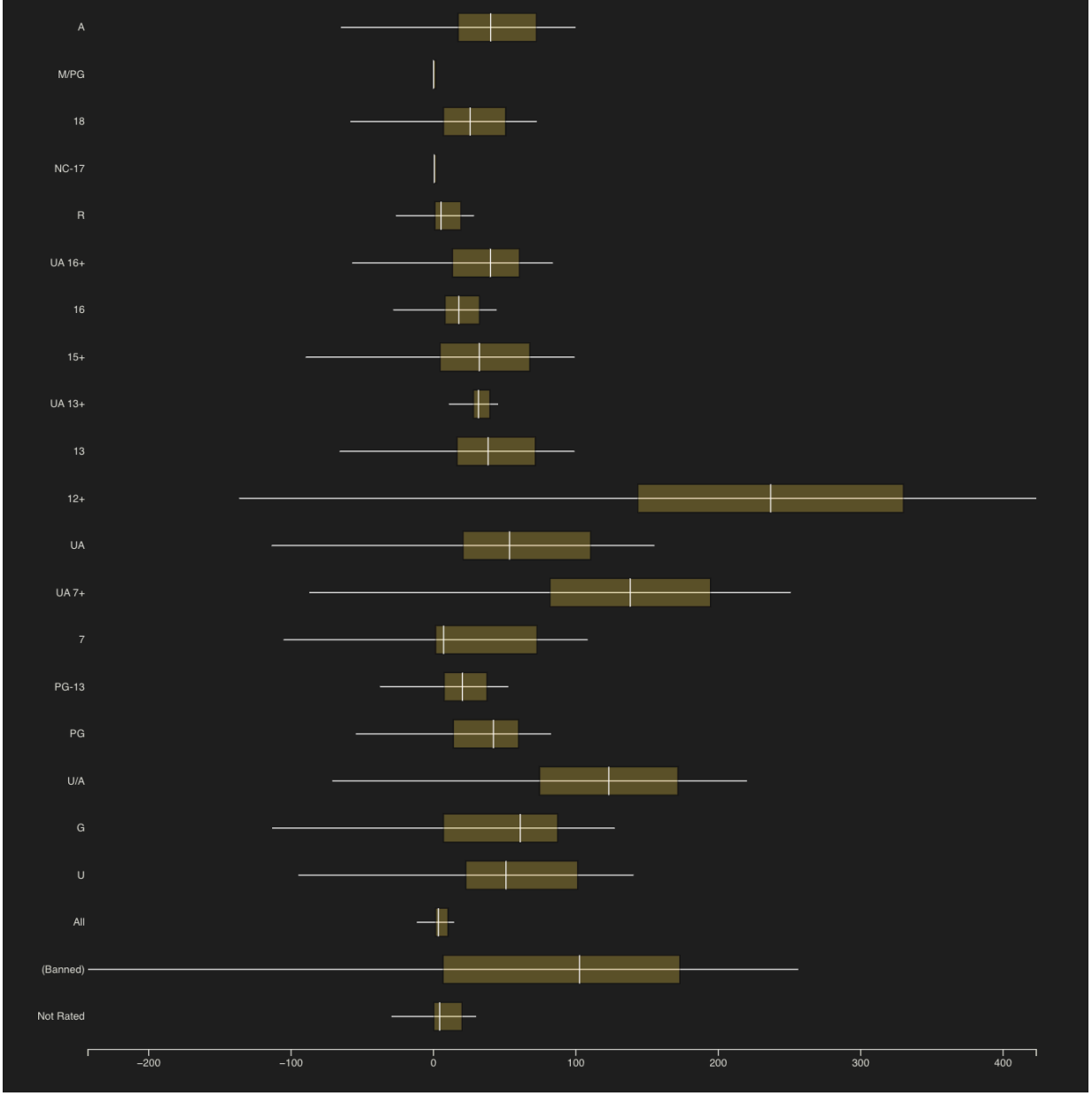
Origineel hadden we een bar chart die visueel aantoont welke filmrating het hoogste bruto inkomen had. Dit was een eerste goede richting maar gaf een verkeerd beeld. Elke rating had een verschillend aantal films, en er was een aanzienlijke variatie tussen de minst en meest winstgevende films binnen elke rating.

Daarom besloten we om over te gaan naar een boxplot. Deze weergeeft de informatie aan de hand van een gesorteerde lijst van alle films per rating. Deze lijst is van klein naar groot gesorteerd op het bruto inkomen van de film. Aan de hand van deze lijst bepalen we het eerste kwartiel, het derde kwartiel en de mediaan. Het minimum en maximum wordt dan bepaald aan de hand van de interkwartielafstand Q3 - Q1. Het minimum is gelijk aan Q1 - 1.5 * interkwartielafstand en het maximum is gelijk aan Q1 + 1.5 * interkwartielafstand. We trekken dan vervolgens een lijn van het minimum tot het maximum en maken een rechthoek aan de hand van Q1 en Q3. Met deze visualisatie krijgen we een duidelijke representatie van het bruto inkomen per genre.

Alleen merkte we al snel op dat filmratings die maar 1 film bevatten slecht gevisualiseerd worden. Bovendien, doordat het minimum en maximum aan de hand van de interkwartielafstand berekend wordt, hebben we negatieve waarden. Wat in feite niet mogelijk is, want we werken met geldsommen die altijd positief zijn.

Average box-office per rating

For censor ratings "12" and "18+" the box-office is unknown. The box-office is in million dollars.

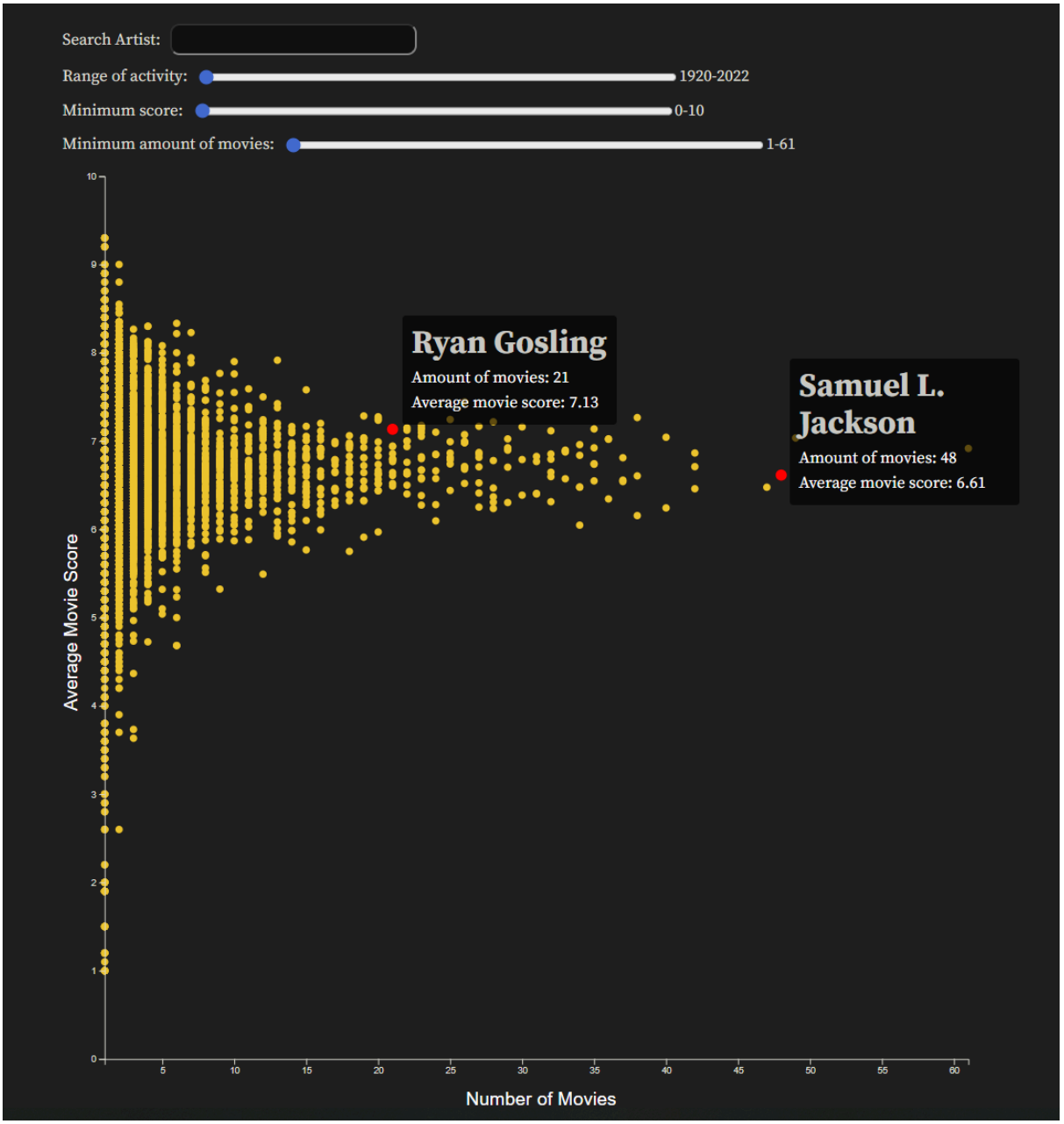


Taakverdeling: 6 Mei-13 Mei

Groepslid	Taken
Iedereen	• Tekst over grafiek schrijven
Bavo	• voorbeelden bij scatterplot toevoegen
Brent	• voorbeelden bij scatterplot toevoegen
Pjotr	• treemap splits de namen van directors & acteurs en hou ze dus niet in groep
Kai	• bug negatieve waardes oplossen bij boxplots • Voeg sommige ratings samen voor de boxplots

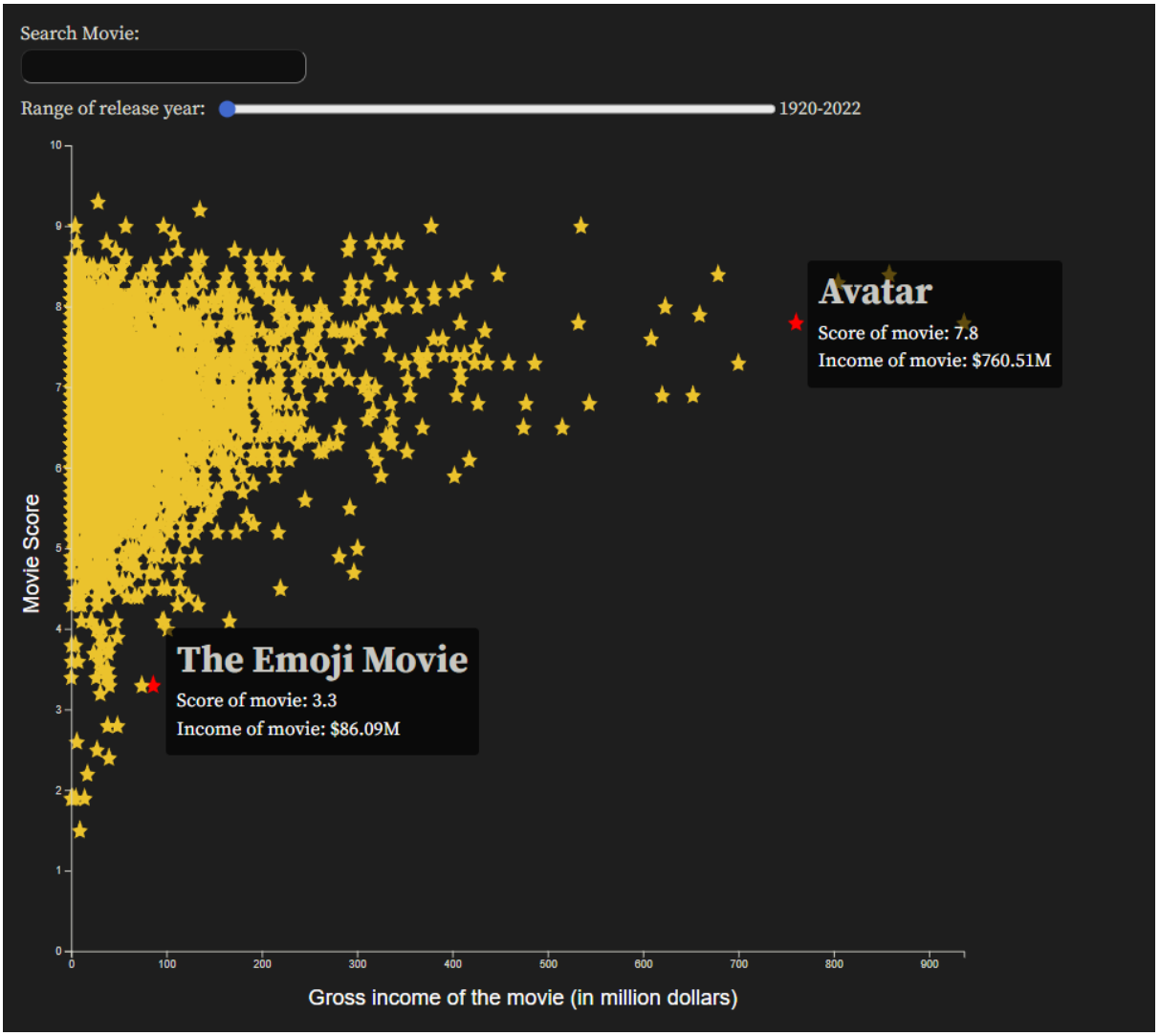
Scatter plot: Average movie score per actor/director

Als laatste toevoeging hebben we enkele voorbeelden ingebouwd om gebruikers te motiveren hun muis over de punten te bewegen. Door deze voorbeelden kunnen gebruikers snel zien wat er gebeurt als ze met hun muis over een punt bewegen, wat het aantrekkelijker maakt om met de scatterplot te interacteren.



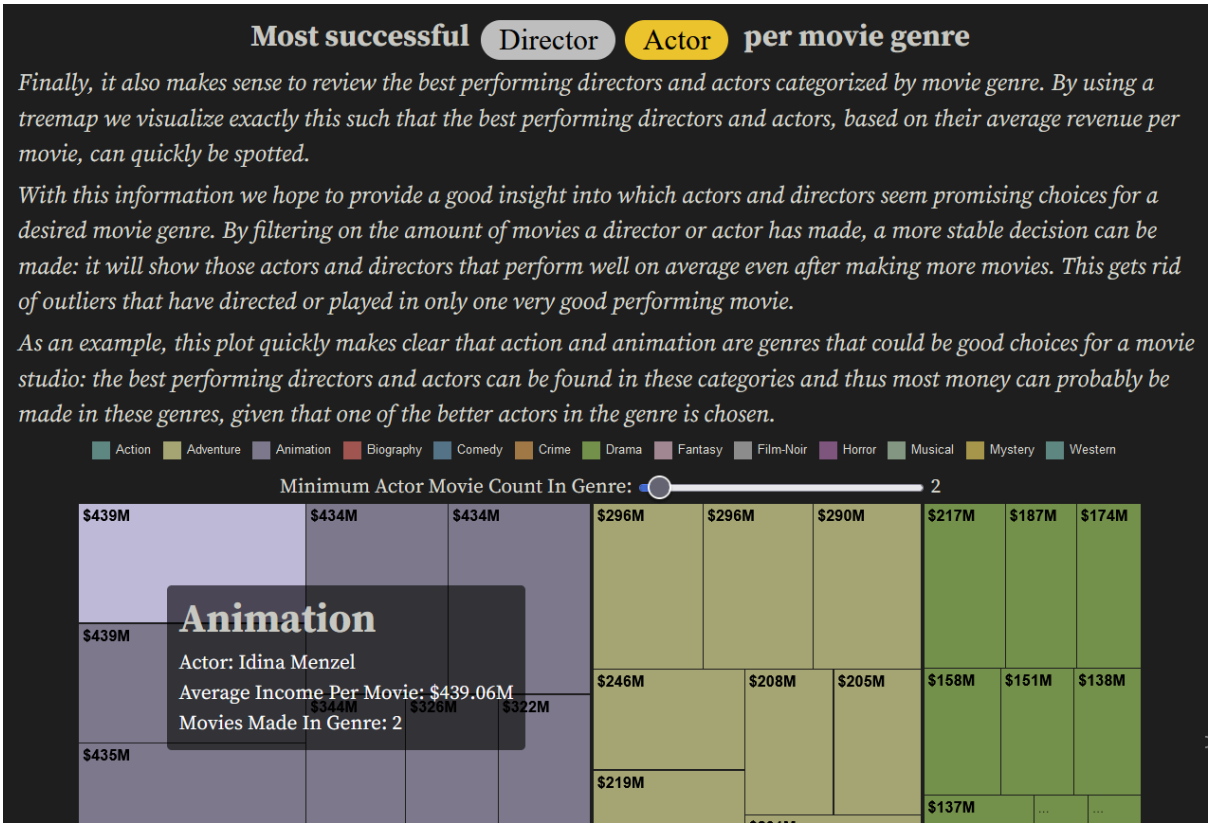
Scatter plot: Movie score vs. box-office

Exact dezelfde redenering als bij de vorige plot.



Treemaps

Voor de afwerking is er nog een kleine informatieve tekst geschreven bij de treemap. Bovendien is er eenzelfde treemap gemaakt die hetzelfde toont voor acteurs als voor regisseurs met een optie om hiertussen te wisselen. Tot slot is er opgesplitst tussen regisseurs en acteurs (eerst werden meerdere regisseurs/acteurs samengenomen).

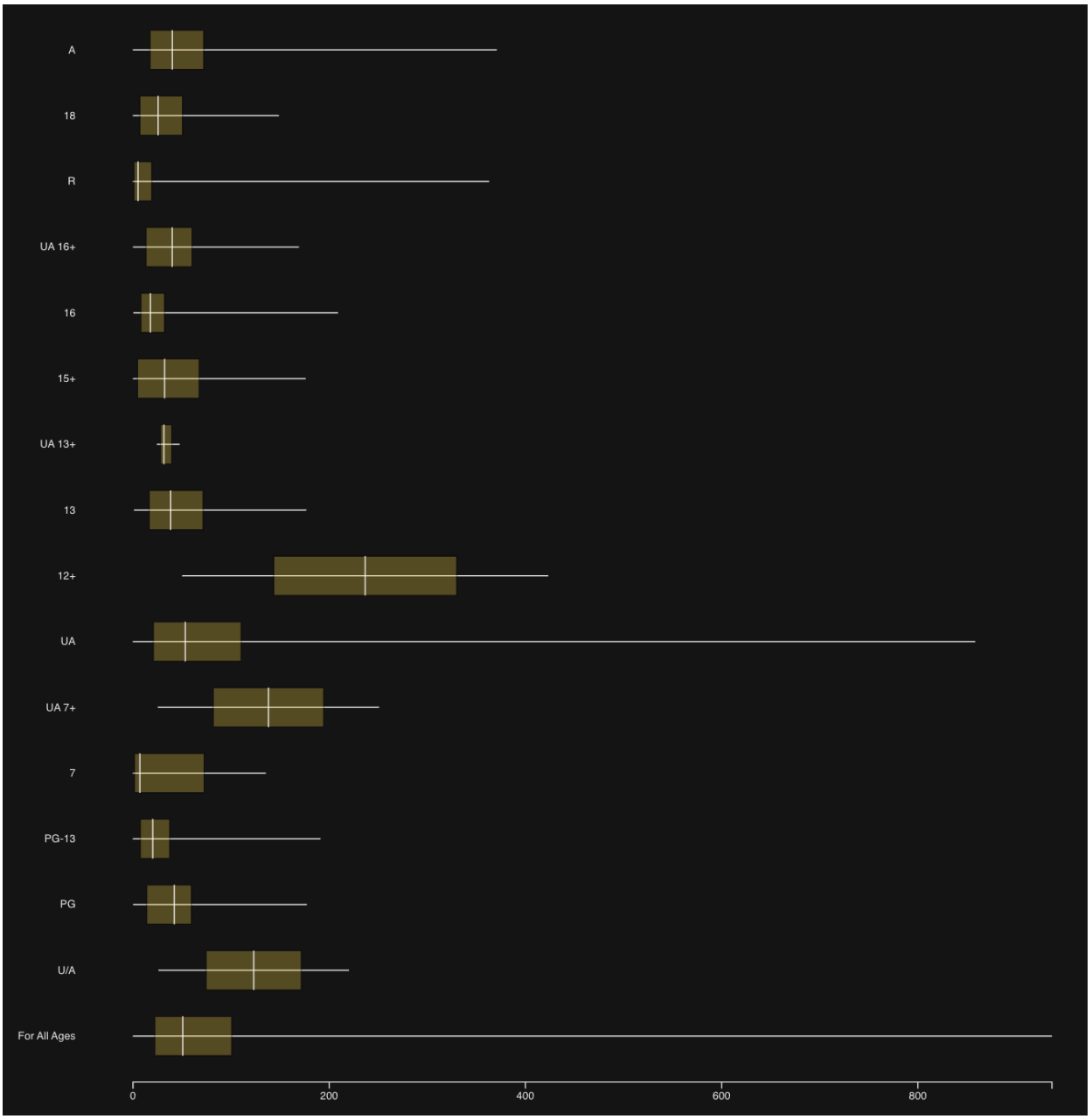


Box plot: box-office per rating: minimum fix and combining ratings.

Uit de vorige versie van de boxplot hadden we nog een paar problemen om op te lossen. De negatieve minimumwaarde voor een film rating moet weggewerkt worden. Sommige ratings hebben ook dezelfde betekenis, deze kunnen we samen nemen om een grotere populatie aan films te hebben. Bovendien hebben we een banned en unrated weggelaten omdat deze geen meerwaarde hebben in ons verhaal voor de grote filmstudio's.

Om de negatieve waarden weg te krijgen werken we niet meer met de interkwartielafstand. We nemen nu gewoon het minimum en maximum van de gesorteerde lijst als waarden. Dit betekent wel dat de grafiek meer naar links leunt omdat alle film ratings wel een film hebben van 0M bruto inkomen. Dit is visueel moeilijker te interpreteren, maar het is wel correcter dan een negatief bruto inkomen hebben.

Daarna hebben we de filmratings M/PG en NC-17 weggelaten omdat deze slechts één film bevatten. Hierdoor was de populatie te klein om een duidelijk beeld te geven van deze filmratings. Ten slotte hebben we de ratings U, G en ALL samengevoegd tot "For All Ages", omdat deze dezelfde filmrating vertegenwoordigen.


A
18
R
UA 16+
16
15+
UA 13+
13
12+
UA
UA 7+
7
PG-13
PG
U/A
For All Ages
0
200
400
600
800