Data Visualization Portfolio

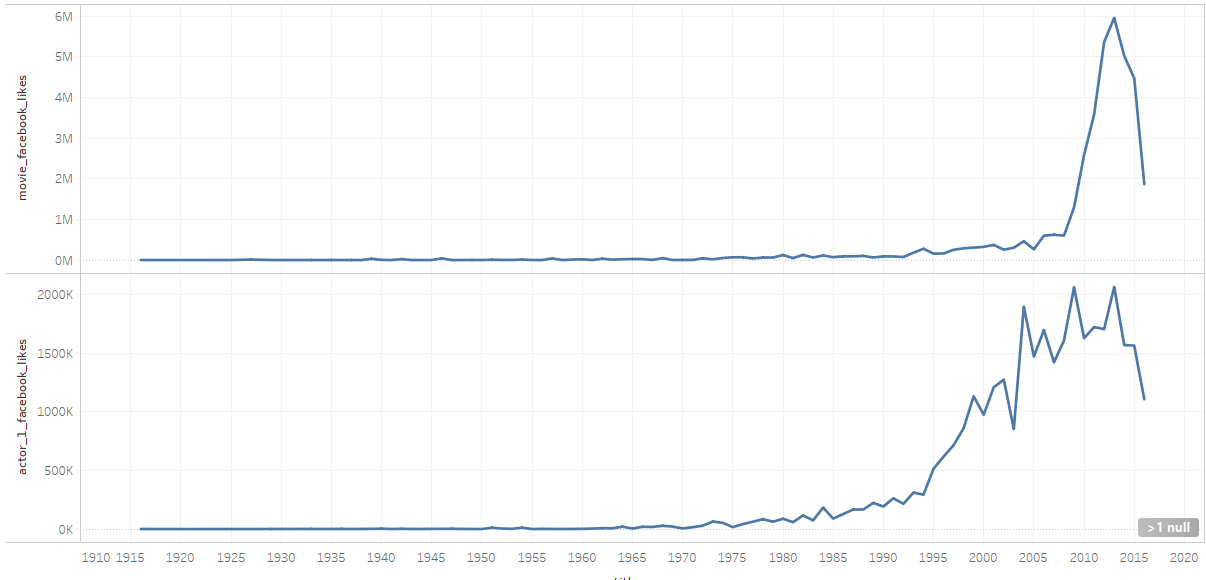
Jan Kappé

# DV – Dataset and Topic

Voor het DV project heb ik ervoor gekozen om met de IMDB 5000 movies dataset te werken. De dataset is te vinden op Kaggle: <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

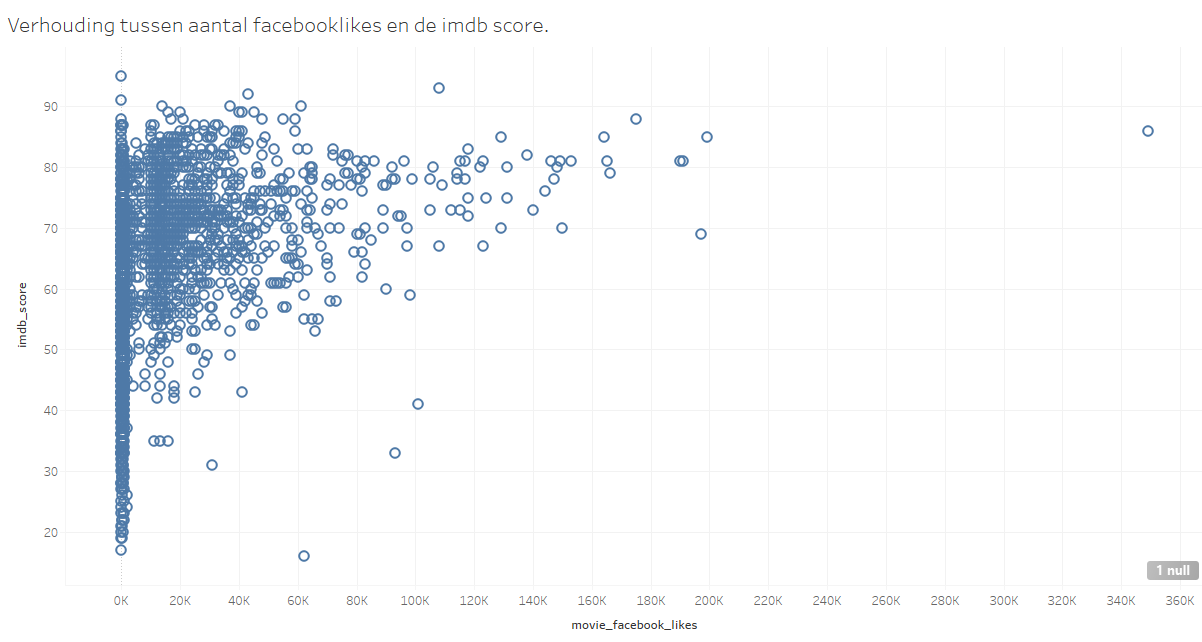
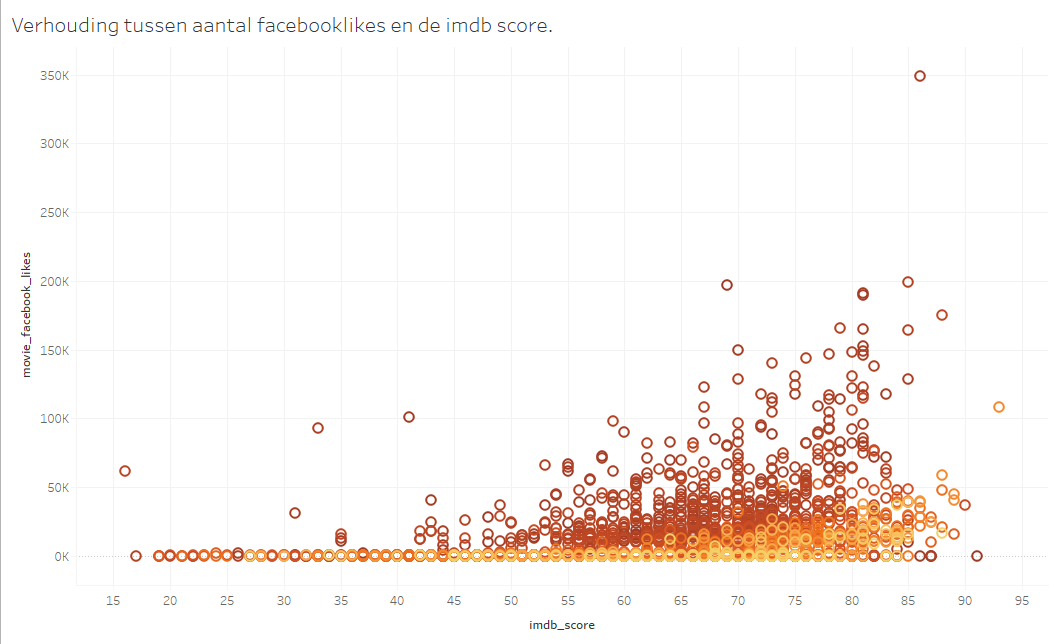
De dataset bevat algemene data over de film zoals genre, budget, director, taal, en land waar de film uit komt. Behalve deze algemene data is er ook data zoals het aantal Facebook-likes van de film, acteurs en director.

# DV Story- IMDB 5000 Movie dataset

Het belangrijkste onderdeel dat ik wil onderzoeken in deze dataset is of er een relatie is tussen het aantal Facebook-likes van een film en de IMDB-rating. Hierbij moet er rekening worden gehouden met het feit dat facebook pas vanaf een bepaald jaartal veel wordt gebruikt. Het is bijvoorbeeld ook interessant om te kijken of de leeftijd de hoofdrolspeler invloed heeft op het aantal likes dat hij krijgt op Facebook.

In de bovenste grafiek van het bovenstaande plaatje is het totaal aantal Facebook-likes voor films te zien voor elk jaar. Hier kun je zien dat het aantal likes van films uitgebracht vanaf 2008 erg veel omhoog is gegaan. Een ander interessant punt is dat er weer een daling in Facebook-likes is na 2013. Mijn eerste gok is dat dit komt omdat het een paar jaar duurt voordat de grote menigte de film heeft bekeken (wanneer de film ook op TV komt bijvoorbeeld).

In de scatter-plot hieronder is het verband te zien tussen het gemiddelde aantal facebook-likes per IMDB rating. Dit is een eerste visualisatie en ik ben van plan deze nog veel verder uit te breiden en te onderzoeken.

Wat als eerste opvalt, is de verticale lijn met films die geen Facebook-likes hebben. Hier gaat het om oudere films of onbekende films die geen Facebook-pagina hebben. Er is enigszins een verband te zien in de data. Je ziet ook duidelijke uitschieters; bijvoorbeeld Interstellar (helemaal rechts) met veruit het meeste Facebook-likes, of bijvoorbeeld de film “Justin Bieber: Never Say Never” (helemaal onderaan maar met 60 duizend Facebook-likes).

De bovenstaande chart is een uitbereiding op de op de vorige scatter-chart. De assen zijn omgedraaid en er is gefilterd op films tussen 1980 en 2016. Ook zijn de films weggehaald die geen Facebook pagina hebben (0 likes). Er is een kleur toegevoegd die te maken heeft met jaar waarin de film is uitgebracht. Hoe donkerder de kleur hoe nieuwer de film. Hiermee kun je goed zien dat het jaar veel invloed heeft op het aantal likes, wat logisch is gezien hoe nieuw Facebook is.

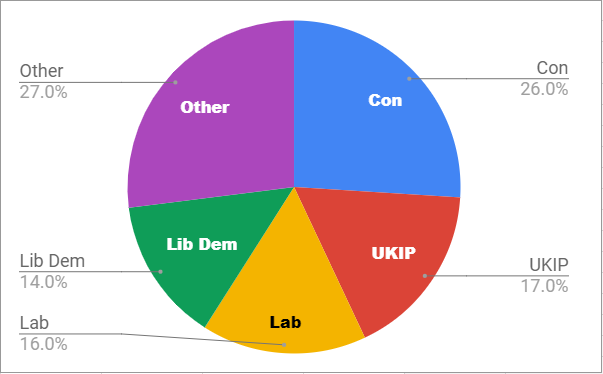
## Final Story

Met deze laatste exploratie fase heb ik besloten om me te gaan focussen op de Facebook-likes. Hierbij ga ik nog extra onderzoek doen naar de populariteit van facebook en de opkomst van Facebook-advertisement.

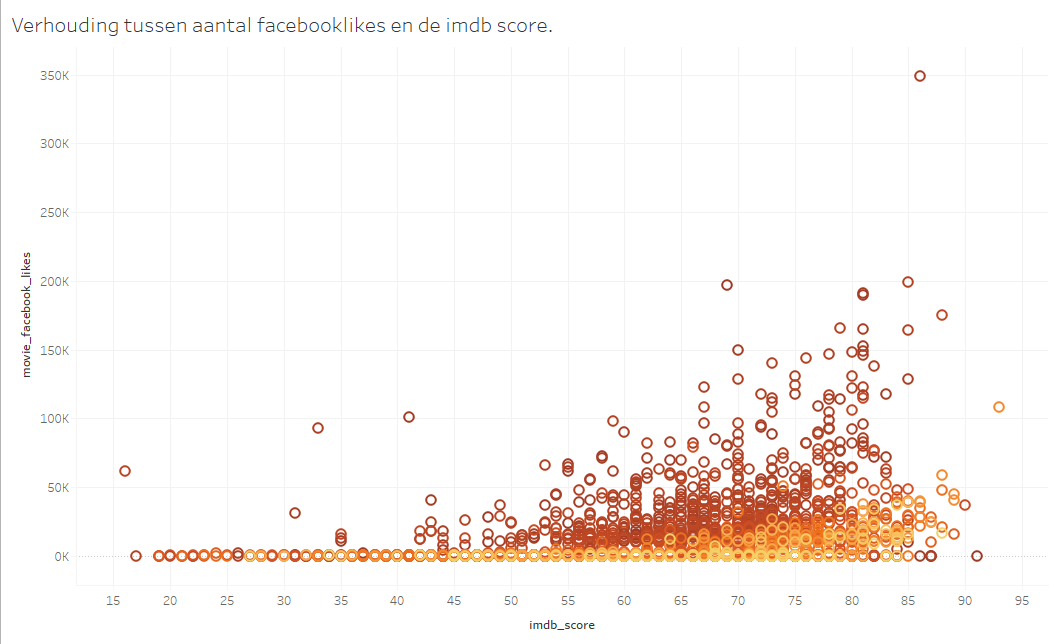
## Bad Charts

Hieronder een voorbeeld van een slechte grafiek. Ten eerste is het fout gevisualiseerd, de lengte van de charts komt niet overeen met het percentage. In dit geval zou het beter zijn om een pie-chart te gebruiken.



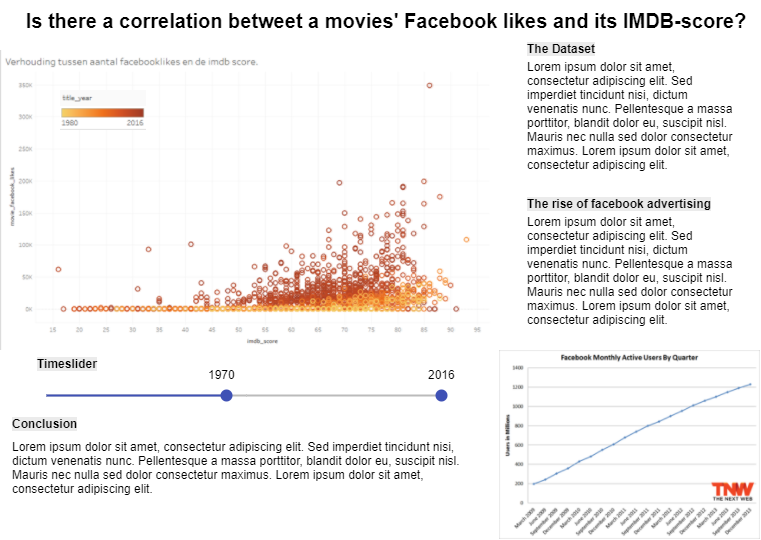


# DV Design

Als chart type heb ik gekozen voor een scatterchart. Hiervoor heb ik gekozen omdat ik wil gaan kijken naar 2 numerieke waardes van individuele films, het aantal facebook likes en IMDB -rating. Het jaar waarin de film is uitgekomen ga ik met de kleur van de punten aangeven. Welke kleuren ik hiervoor ga gebruiken ga ik nog onderzoeken door te kijken naar kleur standaarden of door rond te vragen wat het duidelijkste is.

Er zullen 2 soorten interactie plaatsvinden; de gebruiker kan door naar een punt te gaan met hun muis meer informatie krijgen over de film. Hiermee kan de gebruiker zelf naar de verschillende uitschieters kijken. Verder zal er een schuifbalk komen waarmee de gebruiker kan filteren op het jaartal waarin de film is uitgekomen. Hiermee kan de gebruiker kijken naar de groei van Facebook over de jaren.

### Layout

Voor de layout van de uiteindelijke site heb ik de volgende eerste schets gemaakt in de online tool draw.io. Na eerste feedback van klasgenoten zal de grafiek een stuk groter worden. Ook zal er duidelijk worden gemaakt dat de ‘timeslider’ bij de bovenstaande grafiek hoort. Verder ga ik nog onderzoeken of er een mogelijkheid is om de grafiek orgineler te maken, maar wel zonder dat dit ten koste gaan van de duidelijkheid.

Feedback op design van docent:   
Facebook likes en IMDB-score. Leuk idee om ook resultaten van ML mee te nemen. Kortere titel met boodschap, eventueel met een vraagteken. Je bent al bezig met opzetten van de site. Goed bezig. Je ligt goed op schema voor DV.

De titel van de webpagina mag korter en er daarin kan al een mogelijk uitkomst van het verhaal staan. Door misschien een vraagteken toe te voegen kun je de gebruiker aanmoedigen om zelf naar de grafiek te kijken of het ook echt waar is.

Er mag veel minder tekst, zo veel mogelijk moet door de grafiek zelf uitgelegd worden

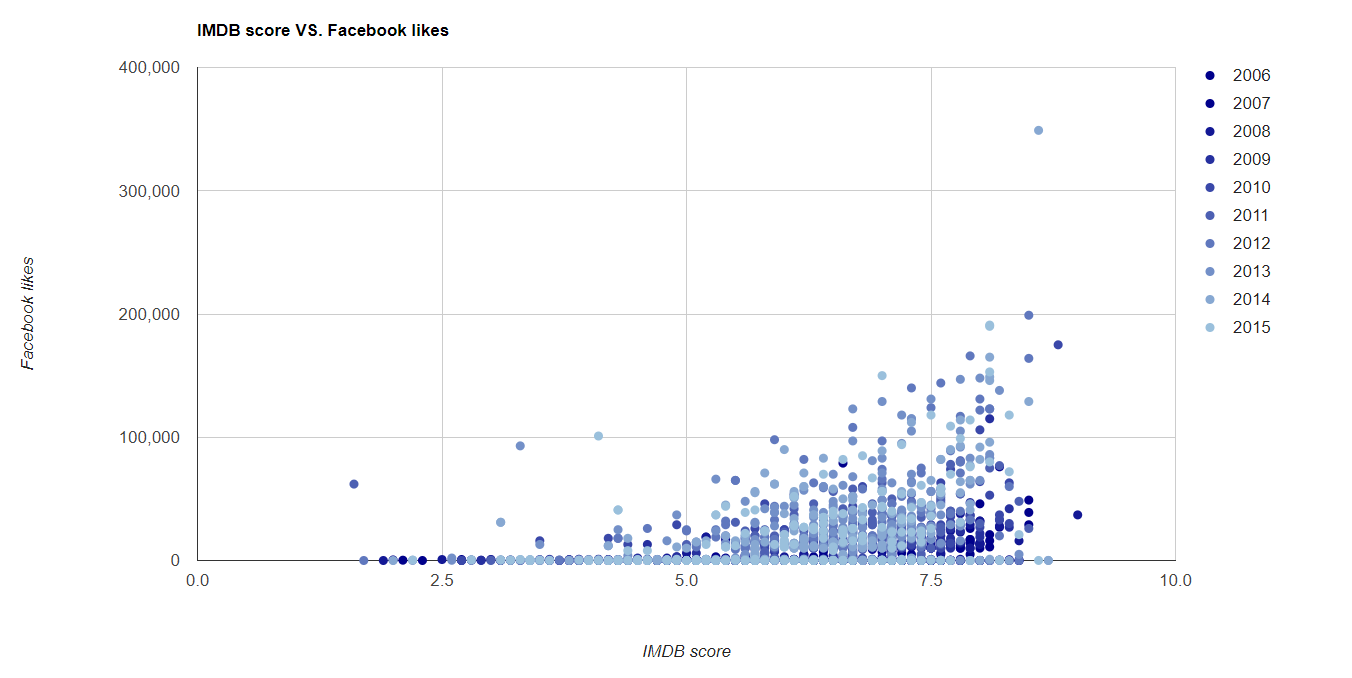
# Implementatie

### Google Visualization

In de implementatie fase ben ik begonnen met het onderzoeken van Dimple.js en D3.js. Allereerst heb ik geprobeerd een simpele scatter plot te maken met Dimple. Echter liep ik steeds tegen obstakels aan en ben ik na een dag niets verder gekomen. Hierna heb ik geprobeerd om direct met D3 te werken maar zonder succes.

Om toch verder te kunnen met mijn project heb ik ervoor gekozen om van Dimple en D3 af te stappen en te zoeken naar een andere tool zodat ik me kan focussen op de visualisatie in plaats van het uitzoeken van deze technieken. Al snel kwam ik tot Google Visualization, hiermee heb ik erg snel een simpele scatter plot kunnen maken.

Zo heb ik doorgewerkt naar het eerste prototype van mijn visualisatie die hieronder is te zien. De volgende belangrijke functie die ik heb toegevoegd is een filter waarmee de gebruiker een selectie kan maken op basis van het jaar dat de film is uitgekomen. Met deze versie heb ik tijdens het klassikaal feedback moment de volgende feedback ontvangen:

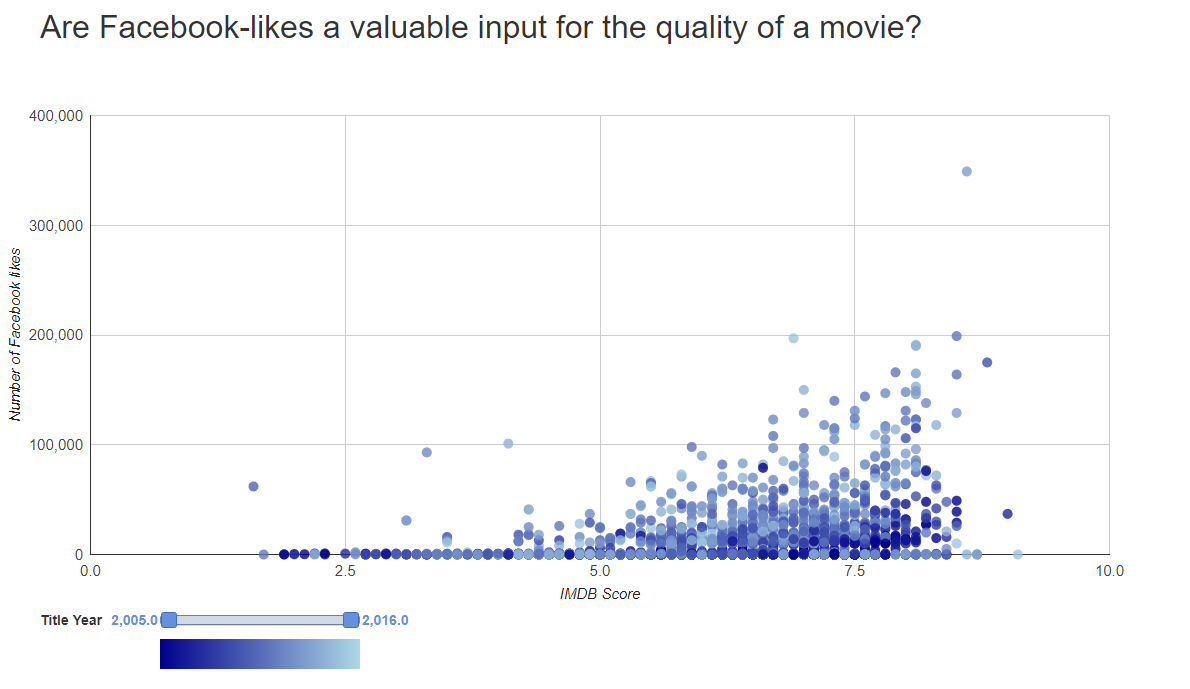
* Assen vast zetten om
* Titel toevoegen aan de website die de gebruiker uitdaagt om te onderzoeken.
* Titel van film toevoegen
* Naam van assen toevoegen
* Mogelijk poster van film toevoegen
* Kijken of het aantal mensen die hebben gestemd interessant is.

## Tooltips en filteren

De twee belangrijkste functionaliteiten van deze visualisatie is de mogelijkheid om te filteren en om te zien om welke film het gaat. Met filteren kan de gebruiker zelf onderzoeken welke invloed tijd heeft op de data, hiermee wordt de gebruiker uitgenodigd om met de visualisatie aan de gang te gaan.

Door het toevoegen van een tooltip op de verschillende data in de scatterchart kan de gebruiker zien over welke films het gaat. Dit is vooral interessant voor de uitschieters in de chart, de nieuwsgierige gebruiker zal willen weten welke film bijvoorbeeld veel likes heeft maar een lage beoordeling.

Het toevoegen van deze functionaliteiten ging een stuk minder gemakkelijk dan verwacht. Los van elkaar deden deze het erg goed en dit was erg simpel om te maken, echter wanneer ik twee functionaliteiten wilde combineren verdween een van de functionaliteiten (het laten zien van de film in de tooltip). Ik ben er erg lang mee bezig geweest om dit probleem op te lossen, ik ben uiteindelijk op de oplossing gekomen door een vraag te plaatsen op Stackoverflow.

Verder heb ik ervoor gekozen om de legende aan te passen/te verwijderen. In de vorige versie (hierboven te zien) staat de kleur aan de zijkant voor elk jaartal. De kleur is echter niet bedoeld om een precieze indicatie te geven van het jaartal maar meer een globaal overzicht. Bovendien zou dit erg onduidelijk worden als er veel jaartallen bij zouden komen. Daarvoor heb ik zelf een gradiënt gemaakt en die geplaats onder de filter. Ik heb aan mijn medestudenten gevraagd of dit dan duidelijk was dat het over het jaartal van de film gaan en die waren daar allen mee eens. Een klein puntje wat is deze versie nog niet is aangepast is het formaat van het jaartal bij de filter, hier wil ik dat er 2016 staat in plaats van 2,016.0.

## Extra filters en zoekfunctie

In het feedbackmoment in de klas werd er besproken dat het misschien interessant is om ook naar films te kijken voor Facebook bestond. Om de keuze te beperken voor de gebruiker heb ik een checkbox toegevoegd waar de gebruiker kan selecteren of ze films willen van het alleen het Facebook tijdperk willen zien. Wat ook besproken is tijdens dit feedback moment is het toevoegen van het aantal personen dat heeft gestemd. Deze informatie heb ik toegevoegd aan de tooltip en een drop-down menu toegevoegd waar de gebruiker eventueel kan filteren op een minimum aantal stemmen.

Verder heb ik van een klasgenoot feedback gehad dat het interessant zou zijn als de gebruiker kan zoeken naar films. Deze feedback heb ik verwerkt en de functie toegevoegd.

