

# Dataflix, a Visual Exploration of Netflix

Artur Romão (nMec 98470), Eva Bartolomeu (nMec 98513)

Information Visualization, 2022 (MSc Informatics Engineering, University of Aveiro)

## Abstract

As we know, *Netflix* is the most known and used streaming video *on-demand* platform in the world. Therefore, *Netflix* has thousands of contents and a few options to search and filter it, which led us to think on how it would be to visually represent those contents. That's how the idea of *Dataflix* came about: using a dataset with information related to *Netflix* movies and TV shows, allowing users to have a global perspective about those contents, ensuring, at the same time, a data visualization experience.

## Introduction

This work consists in the conceptualization and implementation of a simple application prototype for data visualization and exploration.

*Dataflix* aims to give an insight and a visual representation of data related to *Netflix*'s content, such as country of production, year added to *Netflix*, age restriction groups, and so on, allowing the users to understand the characteristics of the data and the content distribution among those same characteristics while providing a dynamic way to explore and filter the content in order to find patterns and interesting phenomenons.

In this report, you will find the most important topics of our implementation, such as:

- the dataset used;
- the personas that are the target audience of our application;
- a set of relevant questions that our application should help to answer;
- a low fidelity prototype and associated user tests;
- the final high fidelity prototype and associated user tests;
- a conclusion and ideas for future work.

## Motivation and objectives

When it comes to motivation, we must say that this is a topic that we particularly enjoy since we're *Netflix* almost daily users, that made us enjoy even more developing this project.

About the objectives of our application, the main goals are:

- Organize *Netflix*'s content in many different ways (ex: age restriction, country, etc.);
- Provide a dynamic and visual way of displaying *Netflix*'s content using d3.js charts;

Besides this, we hope that our application is useful for curious users that want to visualize *Netflix*'s content in many different ways.

As a last ambition, we also want to provide our service as an open source solution so that everyone can have a complete data visualization experience of *Netflix*'s content.

## Personas and the Questions

### Characterization of the personas and their context

In this application there is only one type of user. For this user we created a persona, Inês Amorim who is a *Netflix*'s customer, as we can see in the table below.

Name	Job Title	Demographics
Inês Amorim	Student	Aveiro 16 years old Cinephile <i>Netflix</i> 's Customer

Goals/Tasks	Context
She wants to check the number of adequate movies for her age	She wants to know where most of Netflix contents come from

Table 1 - Persona: Netflix's Customer

Inês Amorim is a cinephile, and as she uses *Netflix* everyday, she intends to explore its content in different ways, and it is in this context that Inês will use *Dataflix*. Her objectives, in terms of visualization/data analysis, are:

- Observe the distribution of Movies/TV Shows by country of creation;
- Observe the distribution of Movies/TV Shows by different age restriction groups;
- Observe the distribution of movies/TV Shows by year added to *Netflix*.

## Questions to Answer

The following questions are examples of some that our system should help to answer:

- Does *Netflix* have more focus on TV shows than movies in the past few years?
- What are the countries that produce more content for *Netflix*?
- How many adequate movies does *Netflix* provide for children? And how many TV shows for teenagers under 16 years old?
- What are the most common genres among *Netflix* contents?
- How many movies were added to *Netflix* between 2010 and 2014? And what about TV shows?

## Dataset

The dataset that we used consists of listings of all movies and TV Shows available on *Netflix*. We

retrieved the dataset from the Kaggle website, the link can be found in the references [1]. From all the attributes available, we used the following ones:

- Type (Movie or TV Show)
- Country
- Date added to Netflix
- Age group rating
- Genres

## Visualization Solution

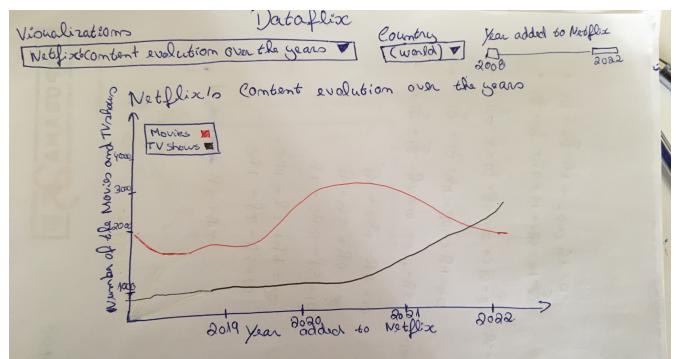
Firstly, a Low Fidelity Prototype (LFP) was developed, in paper, and some user tests were carried out to gather information and redefine some mistakes in the prototype, so that the final High Fidelity Prototype (HFP) could be as good as possible.

### Low fidelity prototype and user feedback

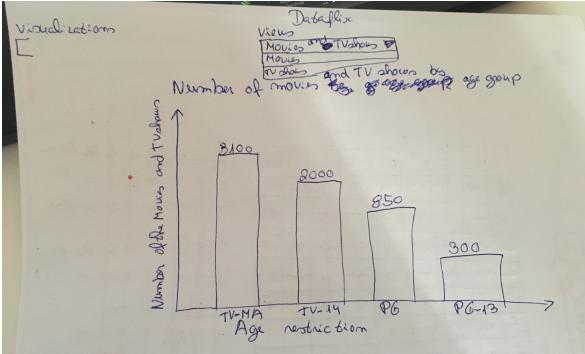
A low fidelity prototype was made so that we could try out which type of charts were better for our data representation. In the following images, we can see our first conception for the application pages, with different charts for each representation.



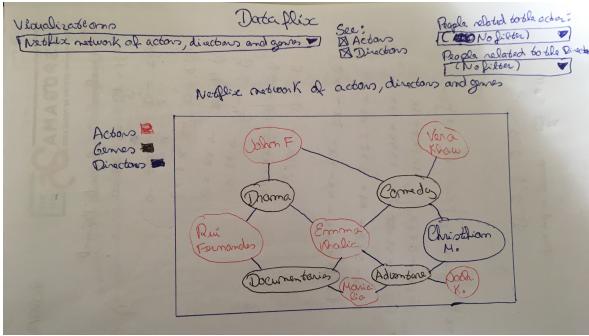
Figure 1 - Example of a Bubble Map to represent *Netflix*'s content throughout the countries



**Figure 2** - Example of a Line Chart to represent the content added to *Netflix* throughout the years



**Figure 3** - Example of a Bar Chart to represent *Netflix's* content by each age group restriction



**Figure 4** - Example of a Network Diagram that represents relations between actors, directors and genres

After finishing the LFP, it was time to conduct the user tests, which were done again for the HFP and will be explained further on.

From those tests, the users reported the following problems:

- the Bubble Map is a bit confusing to represent the number of contents per country, because the bubbles are small and poorly representative (usually used for cities);
  - there is lack of information and documentation throughout the application;
  - the Network Diagram is not that useful and does not contribute that much for a better understanding of the data.

After analyzing this feedback, we decided to change some aspects of our application, which will be explained in the section *Evaluation and changes in the prototype*.

## *Functional prototype*

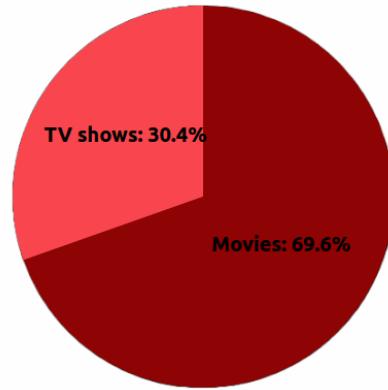
The functional prototype was developed using the React tool [2] and using the D3.js library [3]. Its main functionalities are:

- View the amount of content produced in different countries.
  - Analyze the evolution of *Netflix* content over the years.
  - Insight about the *Netflix* content distribution when it comes to age restrictions.

Let's now analyze the charts developed in **Dataflex**.

## *Pie Chart*

To give an insight about the distribution of *Netflix*'s content by type (Movies and TV Shows), we developed this simple Pie Chart (Figure 5) [11].



**Figure 5** - Pie Chart: Netflix's content evolution over the years

## *Word Cloud*

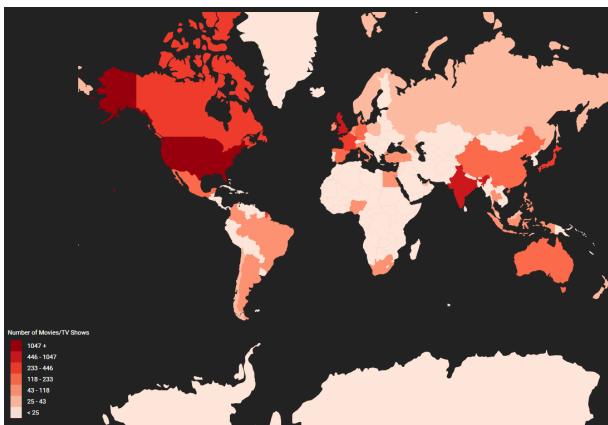
In order to slightly infer about the most common genres among *Netflix*'s content, we decided to add this Word Cloud [12]. The size of the words varies according to how trendy the genre is (a bigger word means a more common genre).



**Figure 6** - Word Cloud of the Netflix's genres

### Choropleth Map

This Choropleth Map represents Netflix's contents throughout the world [9] [10]. The darker the color is, the more number of Movies and TV Shows it has on *Netflix*. By analyzing this map, we can conclude that North America dominates the origin of *Netflix*'s video productions, followed by India, The United Kingdom, Japan and France.

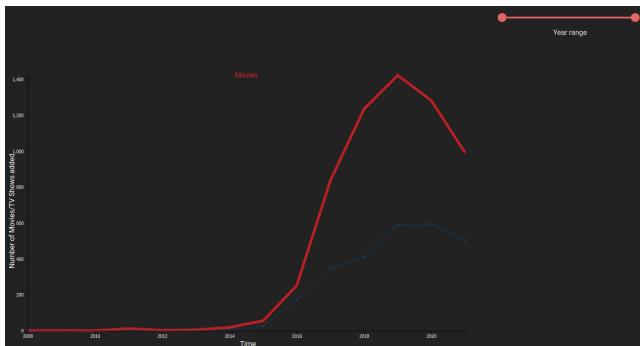


**Figure 7** - Choropleth Map: Netflix's content throughout the world

### Line Chart

Line Chart was the technique chosen to show the evolution of *Netflix* content over the years (Figure 8). We chose this chart to understand *Netflix*'s focus over the years, whether it was Movies or TV Shows.

In this chart we can see the number of movies and TV Shows added to *Netflix* per year [4]. Furthermore, we can select the interval of years we want to analyze.



**Figure 8** - Line Chart: Netflix's content evolution over the years

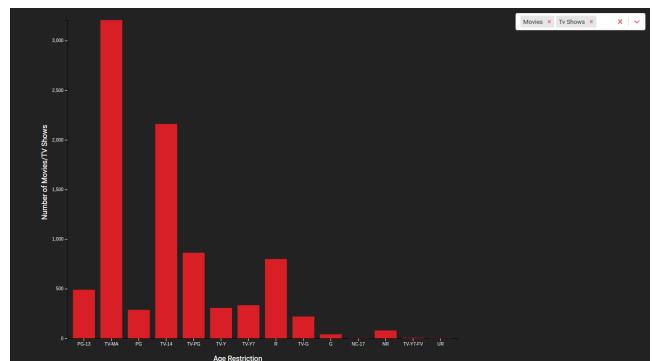
From this chart we can conclude:

- Every year, there have always been more movies added than TV Shows;
- From 2009 to 2012, only movies were added;
- Over the years, *Netflix* has added more and more content per year, highlighting the years 2016 to 2018;
- From the year 2020, a drop in the number of contents added is visible, which could be explained by the impact of the Covid-19 pandemic.

### Bar Chart

The Bar Chart was selected to view the number of movies and series by age group (Figure 9). It was chosen to realize the amount of content in each age group.

It is possible to filter by Movies, TV Shows or both at the same time [5] [6].



**Figure 9** - Bar chart: Number of Movies and TV Shows by Age Group

TV-MA is by far the tendency in both Netflix's Movies and TV Shows.

### Implementation challenges

As main implementation challenges, we would like to highlight the following:

- developing the Choropleth Map took a bit longer than we expected and, although we tried, we didn't manage to implement the zoom functionality;
- the year range filter in Line Chart was also a bit hard to develop since we had to

- update the chart every time the years interval changed;
- a demo in the deployed project [15] can be done with no problems, except if the page is reloaded, that returns a 404 error, although that doesn't happen with the localhost project.

### ***Evaluation and changes in the prototype***

Due to the tests carried out on the low fidelity prototype, we came to the conclusion that the prototype would have to undergo some changes:

- change from a Bubble Map to a Choropleth Map because that would be more appropriate to represent the number of *Netflix* contents per country;
- remove the Network Diagram;
- put more help and documentation throughout the application.

### ***Questions of the usability tests***

During the usability tests, we asked some questions to the 3 participants.

We tested 3 tasks for each participant, at the end of each task we asked 3 questions about the developed task (Table 2) [7].

<b>1</b>	Overall, I am satisfied with the ease of completing the tasks in this scenario.
<b>2</b>	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.
<b>3</b>	Overall, I am satisfied with the support information (online help, messages, documentation) when completing the tasks.

**Table 2 - Questions: After-Scenario Questionnaire**

After each participant took the test, we also asked 10 more general questions about the system (Table 3) [8].

<b>1</b>	I think that I would like to use this system frequently.
----------	--

<b>2</b>	I found the system unnecessarily complex.
<b>3</b>	I thought the system was easy to use.
<b>4</b>	I think that I would need the support of a technical person to be able to use this system.
<b>5</b>	I found the various functions in this system were well integrated.
<b>6</b>	I thought there was too much inconsistency in this system.
<b>7</b>	I would imagine that most people would learn to use this system very quickly.
<b>8</b>	I found the system very cumbersome to use.
<b>9</b>	I felt very confident using this system.
<b>10</b>	I needed to learn a lot of things before I could get going with this system.

**Table 3 - Questions: After Test**

For each question, the participant had to respond according to the following scale:

**1 - Strongly disagree 2 - Disagree 3 - Neutral  
4 - Agree 5 - Strongly agree**

### ***Usability Tests***

After conducting the Usability Tests, it is now time to analyze the results. We considered the arithmetic mean of the participants' ratings for each question.

Task	Q1	Q2	Q3
Visualize the number of Movies and TV Shows of the United States	4.3	4.6	4.3
Analyze the number of Movies and TV Shows added to Netflix between 2017 and 2019	5	5	4.6

Visualize the number of Movies classified with the restriction age of TV-MA	5	5	4.6
---	---	---	-----

**Table 4** - Results of the After-Scenario Questionnaire

Questions	Average Score
Q1	4.6
Q2	1
Q3	4.6
Q4	1
Q5	5
Q6	1
Q7	5
Q8	1
Q9	5
Q10	1

**Table 5** - Results of the After Test Questionnaire

If we analyze the tables above carefully, we can infer that the final application was a success, considering the evaluations of the users.

For the **After-Scenario Questionnaire**, the ratings attributed by the users are well satisfactory, with rates equal or better than 4.3/5 for every question of each task.

And, for the **After Test Questionnaire**, the questions that refer to quality aspects like: "I think that I would like to use this system frequently." and "I thought the system was easy to use." received average ratings equal or better than 4.6/5, while questions referring to downsides of the application like: "I found the system unnecessarily complex." and "I thought there was too much inconsistency in this system." received low ratings (1 for every question).

### **Heuristic evaluation**

We also asked participants to perform a heuristic evaluation of the application, following Nielsen's Heuristics [13] (Table 6).

The severity classification of the usability problem found was on a scale of 0-4, 0 meaning "I don't agree that this is a usability problem at all" and 4 "Usability catastrophe" [14].

Heuristic	Problem	André	Filipe	Marta
User control and freedom	Zoom in Choropleth Map is not possible	-	3	2
Help users recognize, diagnose, and recover from errors	Bar Chart filters for both type of contents when none is selected	1	-	-

**Table 6** - Heuristic evaluation results.

### **Feedback from users**

During the tests we got some feedback from users.

One of the reviews obtained was the fact that we had few filters in some charts, it was not possible to manipulate the way in which they were displayed. For example, filtering by country of origin in the Line Chart, to see the evolution of content coming from that country, and filtering by type of content (movie / TV Show) in the Choropleth Map, in order to be able to see which country has the most series, or more movies separately.

Another negative point was the fact that the Choropleth Map didn't have the zoom functionality. If the user wants to select smaller countries it is necessary to be careful and go slowly to be able to see the label.

## **Conclusion and Future Work**

In conclusion, we can say that we achieved all the proposed objectives. In addition, the questions

that the system must answer, related to the user, indicated above, were also fulfilled.

We also concluded that information visualization is an important practice to be able to better observe extensive data and see it in different ways. It greatly facilitates the interpretation of the data, which we can take from the charts.

By analyzing the implemented data visualization, we took several conclusions from within the charts, which we wouldn't just from simply analyzing the dataset.

For future work the most important feature to implement would be the zoom in the Choropleth Map. Adding more filters to the charts would also be a good idea, for example: filtering by type of content in the Choropleth Map, or filtering by country of origin in the Line Chart.

## References

- [1] <https://www.kaggle.com/datasets/shivamb/netflix-shows>
- [2] <https://reactjs.org/>
- [3] <https://d3js.org/>
- [4] <https://codepen.io/zakariachowdhury/pen/JEmjwq>
- [5] [https://d3-graph-gallery.com/graph/barplot\\_button\\_data\\_simple.html](https://d3-graph-gallery.com/graph/barplot_button_data_simple.html)
- [6] <https://perials.github.io/responsive-bar-chart-with-d3/>
- [7] <https://www.digital-management.at/asq-after-scenario-questionnaire/>
- [8] <https://www.playbookux.com/usability-testing-questions/>
- [9] <https://codepen.io/bwiacek/pen/pXZEvK>
- [10] [https://d3-graph-gallery.com/graph/choropleth\\_hover\\_effect.html](https://d3-graph-gallery.com/graph/choropleth_hover_effect.html)
- [11] [https://d3-graph-gallery.com/graph/pie\\_annotation.html](https://d3-graph-gallery.com/graph/pie_annotation.html)
- [12] <https://codesandbox.io/s/50wzl?file=/src/index.js>
- [13] <https://www.nngroup.com/articles/ten-usability-heuristics/>
- [14] <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>
- [15] <https://artur-romao.github.io/dataflix-vi-project/>