Report Visual Analytics: TV Movies Guide Analysis

Simone Gennenzi 1848670 - Giorgia Ristich 1839919 - Florin Cuconasu 1835605

January 28, 2023

1 Introduction

Nowadays, our TV provides a really busy schedule. For this reason, our project takes a look at the programming of different Italian networks such as Sky and Mediaset. However, since we do not have enough data to make a concrete and realistic analysis of purely Italian programs, we decided to focus on the movies broadcast on each day for one year (2022). In this way, we were able to take the information for a particular movie from IMDB (see more details in the **Dataset** section 3.4). The idea behind our project started from our curiosity in understanding how Italian TV programming is managed and according to what characteristics it is made (e.g., daytime or advertising). So, the final tool is an interactive visualization where different parts work together to give the user an accurate overview of the whole dataset, which can be used to highlight patterns, find outliers and more. This tool is directed to a specialized user that is a TV manager in charge of the movies scheduling of TV channel/s. Indeed, the answers to these questions may be helpful to his investigation:

- Is there a channel that broadcasts fewer movies but with a high sharing audience?
- What are the most screened genres in a network and in a channel in a particular month?
- What is the mean duration time of broadcast movies by different channels during a particular month?

In particular, we use:

- Stacked Bubble chart: to see the number of movies broadcast on a channel in relation to the sharing;
- Bubble plot: to see different information (duration, daytime, rating, mean duration) about the movies broadcast on a network/channel;
- Calendar Heatmap: to compare a channel and a network by highlighting the advertising;
- Chord: to see the different genres of the movies;
- Scatterplot: to plot the dimensionality reduction (see in details **Dimensionality Reduction** section 4.1).

All these visualizations are interactive and coordinated with each other. This makes the overall system very complex in terms of connections and logic for a user who is simply a movie enthusiast and wants to learn more. Clearly, all this work was envisioned and then implemented in our design specifically for an experienced user as is, indeed, the TV manager. For this reason, the application is not intended to run effectively on low-end PC, due to the high real-time computation.

2 Related Works

How can be this work located with respect to existing literature about similar topics? To answer this question an analysis of related scientific papers is going to be executed. The interesting part of this research was to discover that there are many theoretical analyses on the sharing comparison between Sky and Mediaset and in general papers describing some characteristics of the main Italian TV channels; however, we did not find a real in-depth analysis on the Italian TV about movies. In addition, in these papers the description of the visual part is very limited as well as the part of analytics. This lack of prior literature and a tool used by an expert user spurred us to complete the project.

Nevertheless, we have found some papers that deal with some of the points we have analyzed with our tool. For instance, in the paper [3] is conducted an analysis on the sharing of the Italian TV with the goal of forecasting the share. Obviously, it is not one of the aspects we focused on, but if we have had sufficient data on *Mediaset* and *Sky* sharing we could have done a deeper analysis as the authors of this paper. Indeed, they used an accurate dataset from Auditel, having the daily sharing at each hour available. So, the authors could analyze the sharing of each TV program from that channel according to different scheduling, fixing day and hour. The results then are displayed on a mosaic plot and on a scatterplot.

Another interesting paper to analyze is [1] that shows a relation between the success in cinemas and TV audience. The analysis performed by these authors allowed us to confirm some of our theses, and as a consequence we can say that our analysis was successfully completed. First of all, the authors state that: *"In the area of pay television, movies - along with sports - are one of the main components of programming"*. This is also reflected in our dataset. Indeed, we can see that Sky, being a pay television, has a number of thematic channels in which the programming schedule consists mainly of movies. There are, however, some exceptions, like *Mediaset's Cine34*. But we will discuss this aspect in connection with another paper.

Next, the authors state that: *"[...] In particular movies that had a scanty success in cinemas reach high audience performance if broadcast on hi-share TV networks"* and go on saying that: *"A movie, especially if it is a success, can constitute a breakthrough product for television programming and in some cases is managed as an event (possibly with advance publicity in the normal television programs)"*. Given this statement, we were curious to verify it and we found that this is true. Many films, even with low ratings, have been very successful among the audiences of the channels that broadcast them. Taking advantage of this, the network inserts more advertising precisely because of the high sharing (see in more detail the **Insights** section 5.6.1).

Finally, the authors state that *"In Italy Rete 4, a television network of the Mediaset group, has for several years programmed a movie every day in the late-night timeslot"*. This can also be found in our tool; but not only. This statement is also reiterated in another paper that is [2] which states that : *"[...] Italia1, RAI1 and Rete4 are the channels with the most Italian feature films broadcast in the night time slot [...]"*.

Another important aspect to highlight in this paper is the following: *"The channel which pays most attention to Italian cinema is Rete4 (445 films/year), followed by Italia1 (237) and Rai3(218)"*. This thesis is supported by the authors by collecting a large amount of data and then representing the latter on a histogram. Indeed, we have also seen this claim to be true through our tool (see in more detail the **Insights** section 5.6.1).

3 Dataset

The final datasets are composed by multiple data coming from different sources: TV data from an Italian TV guide; movie information from Kaggle; sharing data from Auditel (Italian company for TV audience information). The study is only focused on the movies broadcast in 2022 by different channels, thus all the information are about movies. The part described here was done in Python by using the *Pandas* library.

3.1 TV Guide Data

The main information comes from the following *website* which displays the TV schedule of all the Italian channels. We retrieved the data through web scraping; indeed, we extracted the following information via a Python script:

- *day* (string): the day in which the movie has been broadcast.
- *day_number* (integer): the day number associated with the *day*.
- *month* (string): the month in which the movie was shown.
- *daytime* (string): the moment of the day (*morning*, *afternoon*, *evening*, *night*) in which the movie has been broadcast.
- *hour* (time): the exact starting time of the movie.
- *title* (string): the title of the movie; in most of the case, it is the Italian translation of the original title.
- *duration_with_advertising* (integer): the entire duration in minutes of the broadcast movie, also including the advertising time.
- *channel* (string): the channel in which the movie has been broadcast.

All the information can be find in the TSV file “dataset/TV/movies_without_december.tsv”. It does not contain data of movies displayed in December, since the TV guide was incomplete at the time we started the project and because the sharing data was not available for that month. A row in this dataset is for example:

day day_number month daytime hour title duration_with_advertising channel
Martedì 25 gennaio notte 01:10 Match Point 125 Sky Cinema Due

3.2 Sharing Auditel

The sharing audience data was taken from the *Auditel Public Data*. The *share* is a percentage showing the ratio between the viewers of a TV channel and the total viewers watching any program on any channel.

At the time we started the project, there were available only the first 11 months of 2022, therefore as previously mentioned we do not have any data about December. Differently from the other information we used, the sharing data is not related to the sharing in a particular day, but about an entire month. For this reason, this type of information must be read carefully; indeed, if we see a high sharing percentage in a particular month of a channel, it does not necessarily imply that broadcast movies of that channel were watched by a bigger audience, since our sharing data also includes the audience about other TV shows or news. A reasonable approximation can be made on the *Sky Cinema* channels, as they mainly present movies, but for *Italia 1* or *Rete 4*, for instance, it is a very rough approximation.

Nevertheless, we decided to maintain this type of data because could be really informative to the user, in case we could obtain the real one.

We integrated the sharing data with the TV guide dataset, by building a dataset containing for each channel and each month, the number of movies broadcast by that channel in that month with the respecting sharing. The file is called “channel_month_count_sharing.csv” and can be found in the “dataset” folder.

3.3 Kaggle IMDB

All the details about a movie were retrieved thanks to the Kaggle set of datasets called *IMDb Dataset - From 1888 to 2023*, which is based on info from *IMDB*. Since the majority of the movies we retrieved from the TV guide were Italian movies, we had to choose a dataset also containing titles in the original language, so that we could merge the TV guide dataset and the Kaggle one. For this reason, we used the CSV file “ImdbTitleBasics.csv” of the Kaggle set, as it contains the attribute *originalTitle*.

However, before the actual join between the two datasets, we noticed that the Kaggle dataset presents movies with same title (duplicates). Therefore, we could not decide which was in reality the movie in the TV guide dataset, as we had no other information for distinguishing between duplicates. Thus, we decided to remove all duplicates from the Kaggle dataset.

3.3.1 Merging the Datasets

After the duplicates removal, we could join the TV guide dataset and the Kaggle one on the *originalTitle* attribute, and obtained a new dataset containing information about movies whose original titles are in Italian and some international movies, whose titles have been kept in their original language on the TV guide.

One important point to mention is that we cannot be sure that the broadcast movies are exactly the same of the Kaggle movies, again for the problem of duplicates. For instance, we might have a broadcast movie X that is in the TV guide dataset but not in the Kaggle dataset; yet it may happen that the join is successful because there exists Y in the Kaggle dataset having the same title of X.

3.3.2 Channel Selection

At this point, we decided to keep only some channels, so the ones having more than 90 movies and that may be relevant in our study. We kept: 5 channels from the private network *Sky* (*Sky Cinema Drama, Sky Cinema Due, Sky Cinema Suspense, Sky Cinema Comedy, Sky Cinema Action*); 4 channels from the public network *Mediaset* (*Italia 1, Iris, Rete 4, Cine34*); the last one is *Cielo*. In the analysis *Cielo* is considered in the *Other* network.

After this selection, we remained with 1751 different movies subsequently to the join, starting from 7749. In some cases, it might also happen that some movies did not join, due to little differences between the title in the TV guide dataset and in the Kaggle one, such as dashes or apostrophes.

Moreover, since in our study we needed the advertising time of a particular channel, we directly computed it by subtracting to the *duration_with_advertising* attribute the *duration* one. Also here we removed other tuples, since in some cases the *advertising* resulted negative: there could be some errors in the TV guide reporting the entire duration, or errors as a consequence of the merging with the Kaggle dataset.

3.3.3 Rating

In the end, we retrieved the users' rating from the "ImdbTitleRatings.csv" CSV file. At this point, the attributes added to the original dataset are:

- *duration* (integer): the duration in minutes of the movie.
- *year* (integer): the release year of the movie.
- *genres* (string): the genres of the movie. A movie may also have 3 genres, according to the Kaggle dataset.
- *rating* (float): the rating of the movie on the IMDB site.

After the merging with the rating dataset, we further cleaned it by removing tuples with missing values.

3.3.4 Managing Genres

In the Kaggle dataset we have movies of various genres. We collapsed some of them into one single genre, e.g., *Mystery* into *Crime*, or simply removed genres appearing in very few movies, such as *Musical* that appears in only 7 movies. After the cleaning, we obtained 12 genres: *Documentary, Western, Adventure, Fantasy, Horror, Sci-Fi, Comedy, Drama, Thriller, Action, Romance, Crime.*
By keeping the information of the Kaggle dataset, we have that a movie can have at most 3 genres.

3.4 Final Dataset

In the end, the final dataset contains 5718 tuples with 1171 different movies, having the following attributes: *day*, *day_number*, *month*, *daytime*, *title*, *duration*, *duration_with_advertising*, *advertising*, *channel*, *year*, *genres*, *rating*.

As mentioned in the **Sharing Auditel** section, the *sharing* information is present in the sharing dataset. Instead, for what concerns the movie dataset, we decided to remove the *hour* attribute as too precise in this type of analysis where we mainly look at periods of time. The attribute *advertising* is derived from *duration_with_advertising* and *duration*, but we decided to keep it, as it can save a little bit of computation in the analytical part. A tuple in the final dataset is for instance:

day day_number month daytime title duration duration_with_advertising channel year genres rating
Tuesday 25 january night match point 124 125 Sky Due 2005 Romance,Thriller,Drama 7.6

For what concerns the computational part, it can be experimented some lag if the PC is low-end. We decided to maintain the entire dataset, so the user could analyze different channels and networks at the same moment, without changing every time to the desired dataset. Thus, we expect that a TV manager has the sufficient computational power to run the entire application effectively without too much delay.

4 Dimensionality Reduction

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset, allowing to summarize the original n-dimensional data in a lower k-dimensional component.

This type of technique was really suited in our analysis, as it allows to compute dissimilarities on customized distance functions between categorical data. Indeed, our dataset contains many categorical values (*day*, *month*, *daytime*, *title*, *genres*). The following analysis was done in Python by using libraries like *Pandas*, *NumPy*, *SciPy* and *Sklearn*.

4.1 Distance Functions

MDS is based on the computation of dissimilarity matrices which allow to define a metric of similarity between data points.

The **Euclidean distance** was used for the numerical attributes (*day_number*, *duration*, *year*, *rating*, *duration_with_advertising*, *advertising*). Instead, for all the categorical attributes we defined a specific distance function.

The **Day distance** takes into consideration the day (i.e, Monday, Tuesday, etc.) in which movies have been broadcast, and the fact if it was a public holiday (i.e, Easter, New year, etc.) or Sunday. The distance between two days follows this strategy:

- If one of the two days is a public holiday or Sunday, the distance is **2**.
- Same days (e.g., both Monday) have distance **0**, if both are normal days or holidays.
- Different days (e.g., one Monday and the other Tuesday) have distance **1**, if both are normal days or holidays.

Therefore, we mark the distinction between a normal day and a holiday, as the general audience may change his behaviour in the holiday periods. For instance, it may happen that on Sunday people watch more TV instead of on a Wednesday.

The **Month distance** is based on the following strategy. Each month is mapped to its integer value (e.g., January → 1). The distance is the absolute value of the difference between the month integers, but with the caveat that the maximum distance is 6. If the result is greater than 6, then the distances become: $7 \rightarrow 5; 8 \rightarrow 4; 9 \rightarrow 3; 10 \rightarrow 2$. This result can be obtained by applying the formula when the distance is greater than 6:

$$\textit{absolute value difference} \text{ - } (2 \text{ x } (\textit{absolute value difference} \ \% \ 6))$$

This type of distance is justified by the fact that we wanted to stress that months are cyclic, thus *January* is "closer" to *November* than to *April*, even if the absolute difference between to former is 10 (11 - 1) and the latter is 3 (4 - 1).

The **Daytime distance** simply consists in the absolute difference between the daytime associated integers. We mapped: morning → 0; afternoon → 1; evening → 2; night → 3. In this case we do not consider the cyclic aspect of the day as for the month, because movies broadcast in the morning are generally really different from the ones showed in the night, even if morning and night are "closer" in the day.

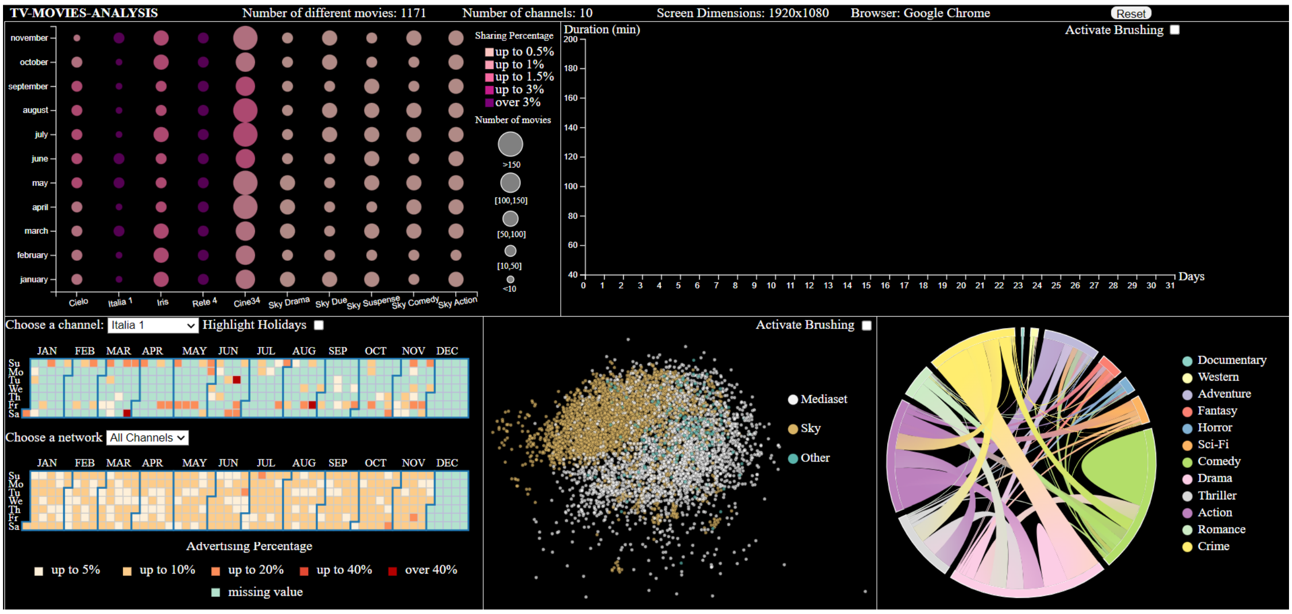
The **Titles distance** is the general **Edit distance** (or *Levenshtein distance*) between strings.

Finally, the **Genres distance** is the **Jaccard distance** where each genre is considered as a element in a set.

The final dissimilarity matrix was obtained by summing up all the dissimilarity matrices computed with the distances mentioned above. However, we put different weights on each distance in the sum in order to give more or less importance to different aspects. For instance, we weighted more the distances of attributes that could be of bigger interest to the user, such as, the matrices of *advertising*, *ratings*, *days*.

The pre-processing computation does not take too much time (about 3 minutes), with the exception of the **Titles distance** (about 30 minutes). Instead, the MDS computation takes about 25 minutes to be computed, and this is the reason why we could not integrate it as a runtime visualization.

5 Visualizations, Interactions and Analytics



This is our TV-Movies-Analysis web page built through the *D3.js* Javascript library and CSS. All the colors were carefully selected from the *ColorBrewer* website. Each chart presents a tooltip to display specific information relevant to it. In addition, on the top of the page, we leave space for a toolbar that contains the following information:

- the number of different movies in the dataset;
- the number of channels considered;
- the optimal screen size to use our tool;
- the browser in which it is desirable to use the tool and in which we test it;
- the reset button.

In general, the user can conduct an in-depth analysis for a particular channel and month, starting from the Stacked Bubble chart in the top left corner, by clicking on a specific bubble. This will change all the other graphs showing the relevant information to sustain the analysis for that channel in that month with respect to the network of the channel. Otherwise, the user can make his considerations by looking at the other charts without starting from the Stacked Bubble chart.

5.1 Stacked Bubble

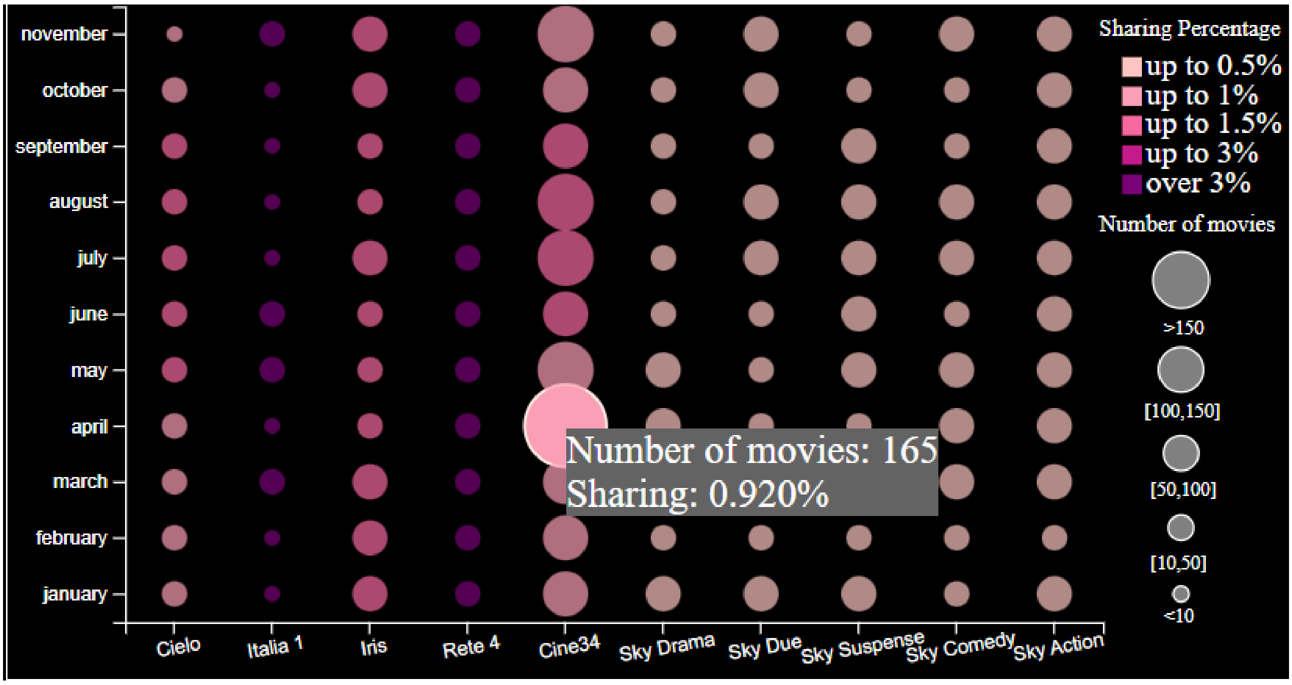


Figure 1: Stacked Bubble chart

Visualization

In this chart it is possible to see the sharing and the number of broadcast movies for each channel during the year. Indeed, the X axis represents all the channels present in our dataset; on the Y axis, instead, are displayed all the months; as mention in the **Dataset** section 3.4, December was excluded.

The other information we can see are:

- the radius of the bubble which represents the **number of movies** that have been broadcast in that channel for that specific month;
- the **sharing percentage** of the hovered channel in that month, showed by a discrete color scale.

Interaction

The user can hover on the circles. As a consequence, their radius will increase and highlighted by a white stroke. In this way, a tooltip will display the number of movies and the sharing percentage about that channel in that month. Obviously, this tooltip will be hidden as soon as the user does not hover anymore on the circle. As we said before, clicking on a specific circle of the Stacked Bubble chart, all the other charts will be updated:

- The Calendar Heatmap on the top will display the advertising trend of that channel during the year, while the one on the bottom will display the advertising trend of the network the channel belongs to. This allow the user to immediately compare the two trends. Notice that it is possible to select different channels/networks to analyze.
- The MDS chart will show all the movies related to the network of the selected channel, differentiating each of them with colors: one for the the channel and the month previously selected; another for the channel in the other months; the last color for the other channels of the network.
- The Bubbleplot will display all the movies of the network the channel belongs to with a green line representing the mean duration of all the movies.
- The chord diagram will display all the genres of the movies broadcast in the network the channel belongs to.

The deselection of a clicked bubble will reset entirely all the other graphs as at the beginning.

5.2 Scatterplot

Visualization

In the Scatterplot it is possible to visualize the results of Dimensionality Reduction applied to our dataset through the Multidimensional Scaling technique (MDS). Every movie is plotted as a circle where its *cx* and *cy* attributes are coherent with the *mds_x* and *mds_y* columns respectively. So, it is possible to see that the closer two movies are, the more similar they are according to the defined distance functions (see **MDS** section 4.1).

Moreover, the chart can be zoomed and the user can brush a specific cluster in order to obtain some information about it. In this case, the tooltip shows the title and the year of the movie that is hovered.

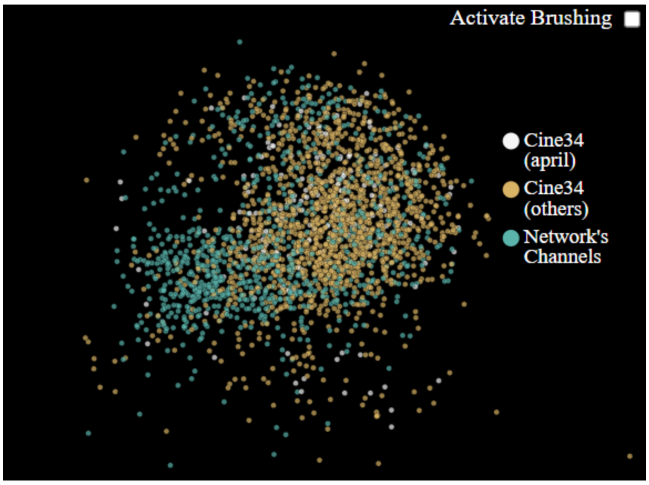


Figure 2: Scatterplot MDS after clicking on the *Cine34 April* bubble on the Stacked Bubble chart

Interaction

At the beginning, it is possible to select only the movies of a specific network by clicking on the related legend, and this click will rebuild the Bubbleplot and the Chord diagram according to the information of the movies of the selected network. Moreover, in this chart it is possible to brush thanks to the **d3.brush** method in order to select only a cluster of our interest. The user can activate the brushing via the checkbox that appears in the top right corner of the area related to this chart. If the user clicks on a bubble on the Stacked Bubble chart, the points in the Scatterplot will be distributed in such a way that only the movies of the network the selected channel belongs to are displayed.

After the selection the following will happen:

- the Calendar Heatmap on the bottom will show the advertising percentage trend using as dataset the movies selected by the brush.
- The Bubbleplot is recreated using now only the movies selected.
- The Chord is rebuilt showing only the genres of the movies selected, but the selection on the paths or on the legend is now disabled.

5.3 Bubbleplot

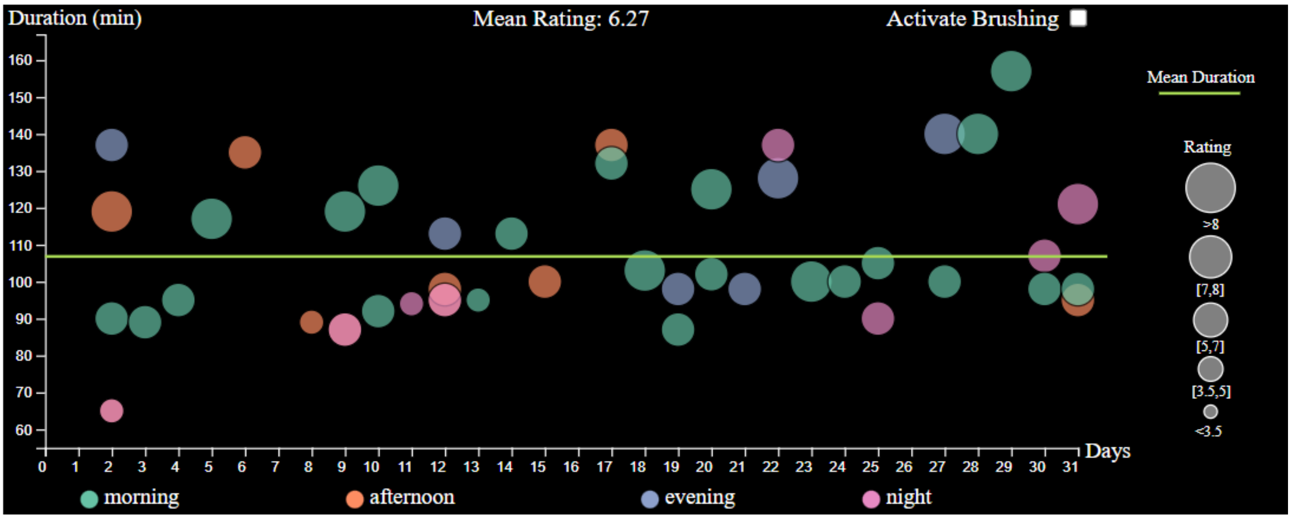


Figure 3: Bubbleplot of *Iris* in *May*

Visualization

The Bubbleplot is a scatter plot with a third dimension added. Indeed, the usual X and Y dimensions represent, respectively, the day in which the movie has been broadcast and its duration; the third dimension is the radius of the bubble which represents the rating of the movie according to the IMDB votes. So, with the use of **d3.axis()** and **d3.circle()** we have plotted the movies chosen by the user, adding a tooltip to every movie. The user can hover on the circles and the tooltip will show the title, the year, the genres, the rating, the channel, and the month related to the movie. Moreover, each circle has been filled with a color that depends on which part of the day the movie has been reproduced.

A green line shows the mean duration of the displayed movies and on the top we can also see the mean rating of the same movies.

Interaction

This chart is two-way connected with the Chord diagram, thus every change in it will implicate a change in the Chord chart. By ticking the checkbox related to the brush, it is possible to select a specific cluster of interest that will rebuild the Chord using only the movies selected, but the selection on the paths or on the legend is now disabled.

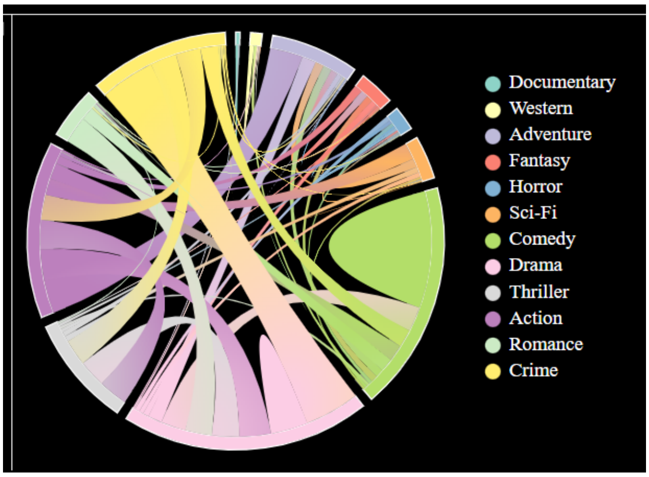


Figure 4: Chord diagram considering all movies

5.4 Chord

Visualization

A Chord diagram represents connections between several entities where each one is represented by a fragment on the outer part of the circular layout. Arcs are drawn between each entities and their size is proportional to the importance of that connection. In our case, it displays the movies' genres. A connection between genres expresses that in the dataset a movie has both the genre. Movies with only genre contributes to a self-connected arc (like an hill). In the case in which a movie presents three genres, we will have an arc for all the three possible combination of the genre. For instance, if a movie has genres A, B and C, then in the Chord it will contribute to the following arcs: $A \leftrightarrow B$, $B \leftrightarrow C$ and $A \leftrightarrow C$.

The legend displayed on its right is interactive, so clicking on a specific genre, the paths containing that genre as source or as target will be highlighted.

Interaction

Clicking one of the paths (that represents two genres or only one if it's a self-connected arc) implicates the re-computation of the Bubbleplot and the Scatterplot by only considering movies having those genres. Clicking on it will highlight that path and its deselection will bring the system to the previous state. The same effect can be obtained by clicking on the genres on the legend.

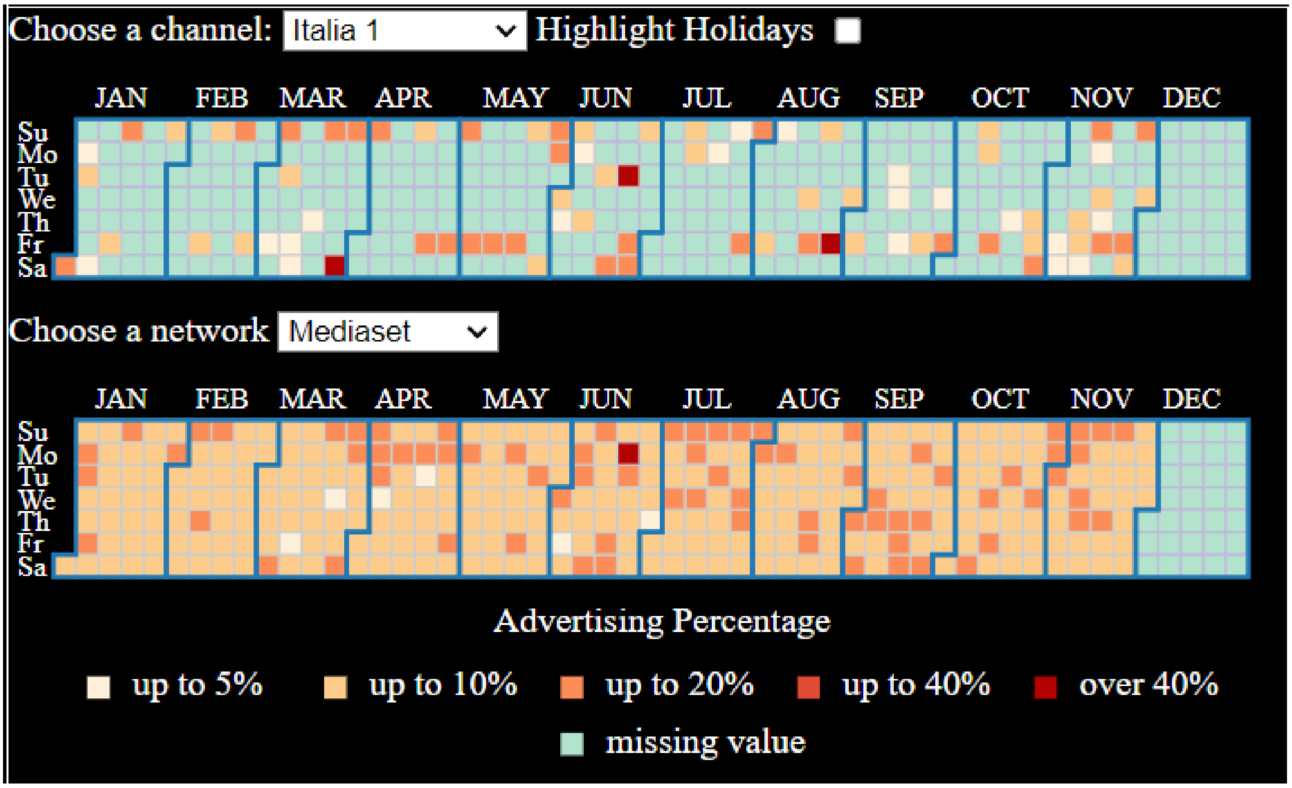
5.5 Calendar Heatmap

Visualization

The Calendar Heatmap is a variant of the Calendar chart used to show time series data using color gradients. In the chart on the top, it is possible to visualize the advertising trend of a specific channel during the year and obtain the specific advertising's percentage in that day pointing the mouse over the cell related to that day. In particular, the **advertising's percentage** for a cell is computed as the ratio between the sum over all the advertising of one day and the total duration (advertising included) of the movies in that day. Each cell is related to all days of the year 2022 except for December's days that are not present in our dataset.

Through the drop-down menu it is possible to select the specific channel to analyze and clicking the near checkbox *Highlight holidays* the user can highlight the holidays in order to conduct a deeper analysis.

On the chart on the bottom, instead, it is possible to select between a specific network or the totality of the channels, in order to compare the advertising trend between a channel (using the first chart) and the network it belongs to (using the second chart).



Choose a channel: Italia 1 Highlight Holidays ☐

Choose a network Mediaset

Figure 5: Calendar Heatmap comparing *Italia 1* and *Mediaset*

Interaction

In this chart, the user can select the specific channel and the specific network he wants to compare in terms of advertising percentage.

As described in the **Scatterplot** section, the Calendar on the bottom will be recreated accordingly to the brushed data points.

5.6 Insights

This section gives some intuitions about the capability of the system to answer to the questions it is created for. Here, it is important to stress that our results depend on the dataset we obtained. Since our data comes from different sources, we cannot be really sure that the following observations are completely coherent with the true original data.

Is there a channel that broadcasts fewer movies but with a high sharing audience?

In the Stacked Bubble we can notice that channels like *Italia 1* and *Rete 4* reached a higher audience but with a few number of broadcast movies. Nevertheless, it is important to highlight that this result comes because of our initial data pre-processing and filtering. Indeed, those two channels may have broadcast many more movies, but we could have removed most of them.

What are the most screened genres in a network?

Chosen a specific network, clicking on a circle related to a month and a channel belonging to the network, the chord will show all genres of the movies broadcast by the network the channel belongs to. Analyzing the size of the paths and of the hills (self-connected arcs), it is possible to detect the most watched genres. It is interesting to see that Mediaset offers a large number of Comedy movies while for Sky the majority is composed by Drama and Action genres.

What are the most screened genres in a channel in a particular month?

Chosen a channel and a specific month, we can click on the related circle on the Stacked Bubble chart. Then the Scatterplot related to the MDS will show all the movies related to the network of the selected channel. At this point, it is possible to click on the legend related to the month of interest and both the Bubbleplot and the Chord will be updated. Finally, on this one it is possible to catch the most screened genre according to the size of the paths of the genres. For instance, *Cine34* in *May* mainly broadcast Comedy movies (105 out of 158).

What is the mean duration time of broadcast movies by different channels during a particular month?

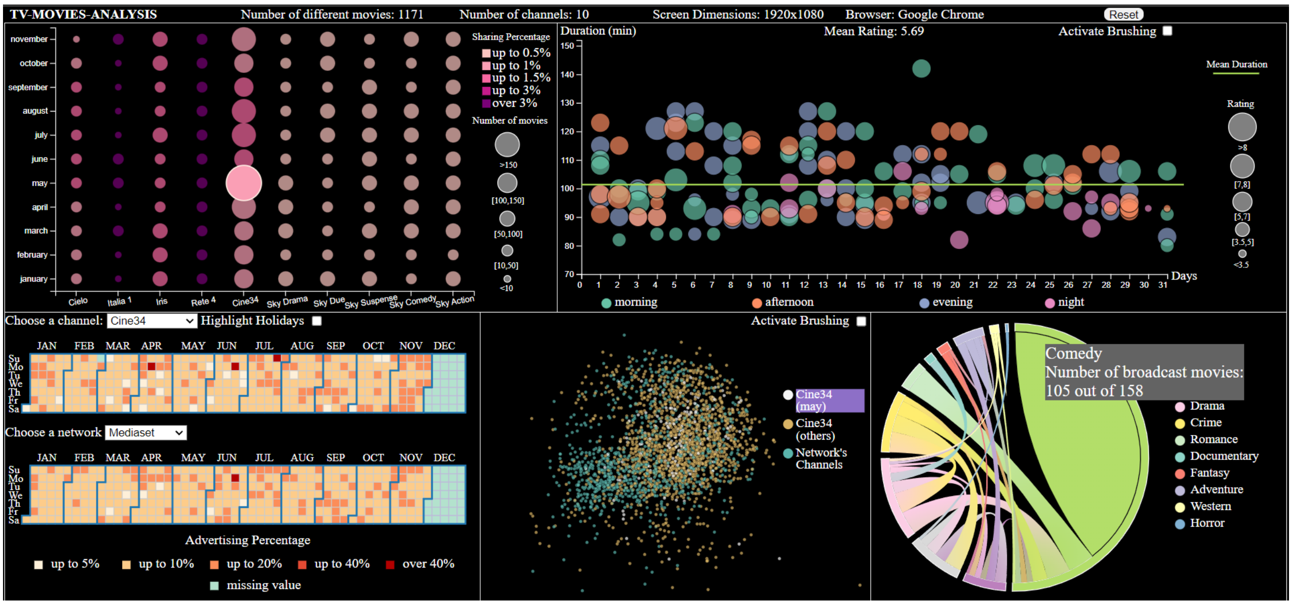


Figure 6: System state after clicking on the *Cine34 May* bubble

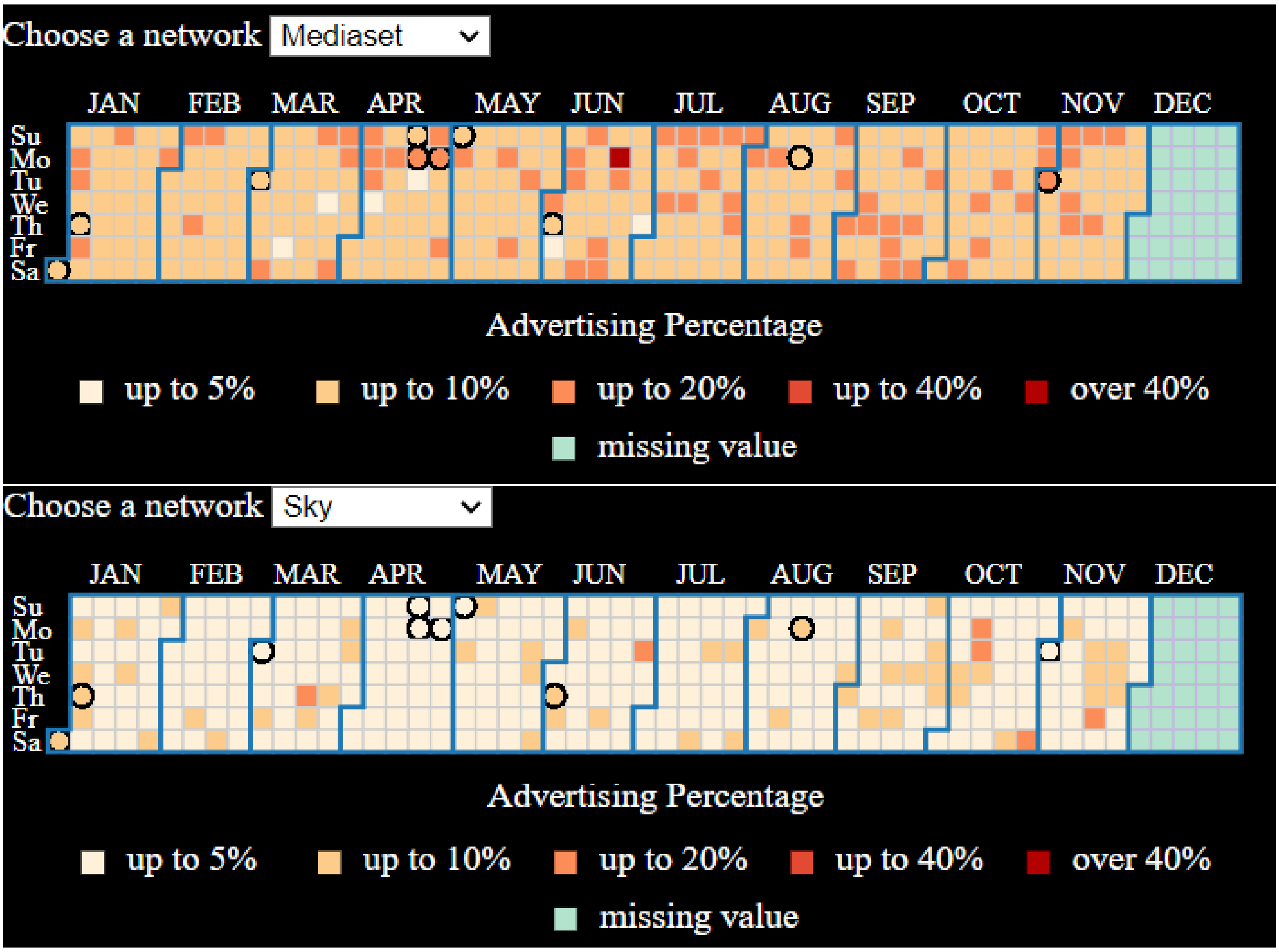


Figure 7: *Sky* and *Mediaset* Calendar Heatmap compared

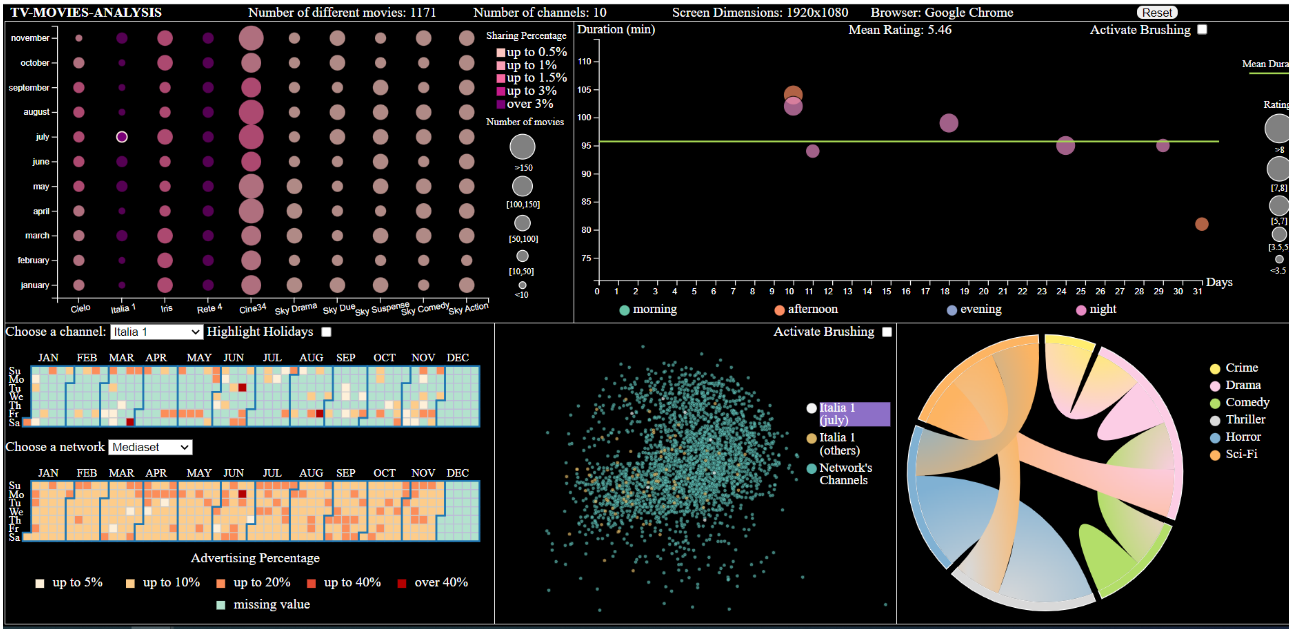
First of all, chosen a channel and whatever month and clicking on the related circle of the Stacked Bubble chart, the other charts will be updated. Then, if we want to see the mean duration for a particular channel we can click on the legend of the MDS, selecting the month of interest, that will modify the Bubbleplot accordingly. Now, the green line represents the mean duration of the movies broadcast during that month.

Another interesting point can be the analysis on the advertising percentage during the holidays, i.e., the period in which there could be more audience. We have found that in the April holidays, the Mediaset advertising percentage reaches about 30%; instead the Sky one remains almost constant. Moreover, it is possible to see that Mediaset has obtained a high percentage of advertising in all the Sundays. Again considering Mediaset, if we look at the sharing percentage in the Stacked Bubble chart, we can notice that in the summer period (June to August and in part September) its channels reached the highest sharing scores. For instance *Italia 1* reached the highest sharing percentage among all channels, namely 5.034%. Furthermore, in this period the advertising percentage of the channels was also pretty high, as we have most of the days with a percentage higher than 20%.

5.6.1 Related Works

In [1] the authors conclude that *"[...] In particular movies that had a scanty success in cinemas reach high audience performance if broadcast on hi-share TV networks"*. Through our tool the thesis affirmed in [1] can be visually confirmed. Now, let us consider the *rating* attribute as a measure of the movies success. If we look at *Italia 1* or *Rete 4* we can notice that the mean rating of their broadcast movies during the entire year is no more than 6.5, although their sharing percentage is always pretty high. For instance, *Italia 1* in *July* had a mean rating of the movies of 5.46 (fig. 8), with the highest sharing percentage among our data (5.034%).

However, as pointed before, this type of information must be read carefully; indeed, *July's* high sharing



Choose a channel: Italia 1

Choose a network Mediaset

Figure 8: Mean Rating Italia 1 July

percentage does not necessarily imply that *Italia 1's* broadcast movies were watched by a bigger audience, since our sharing data also includes the audience about other TV shows or news.

In [2] the authors conclude that *"[...] The channel which pays most attention to Italian cinema is Rete4 (445 films/year), followed by Italia1 (237) and Rai3 (218).".* Even if this thesis was related to 2008, the same trend can be partially confirmed nowadays. Through our tool this affirmation is highlighted simply visualizing the radius of the circles related to *Rete 4* in the Stacked Bubble Chart which are bigger than *Italia 1's* ones. The novelty that can be seen from our tool is the introduction of a new channel which is *Cine34*. The latter is entirely dedicated to the reproduction of films of different genres through the day, as can be seen from the size of the circles on the Stacked bubble chart. Consequently, it can be said that nowadays *Cine34*, in terms of broadcast, movies has reached and surpassed *Rete4*, even if it remains a precious point of reference.

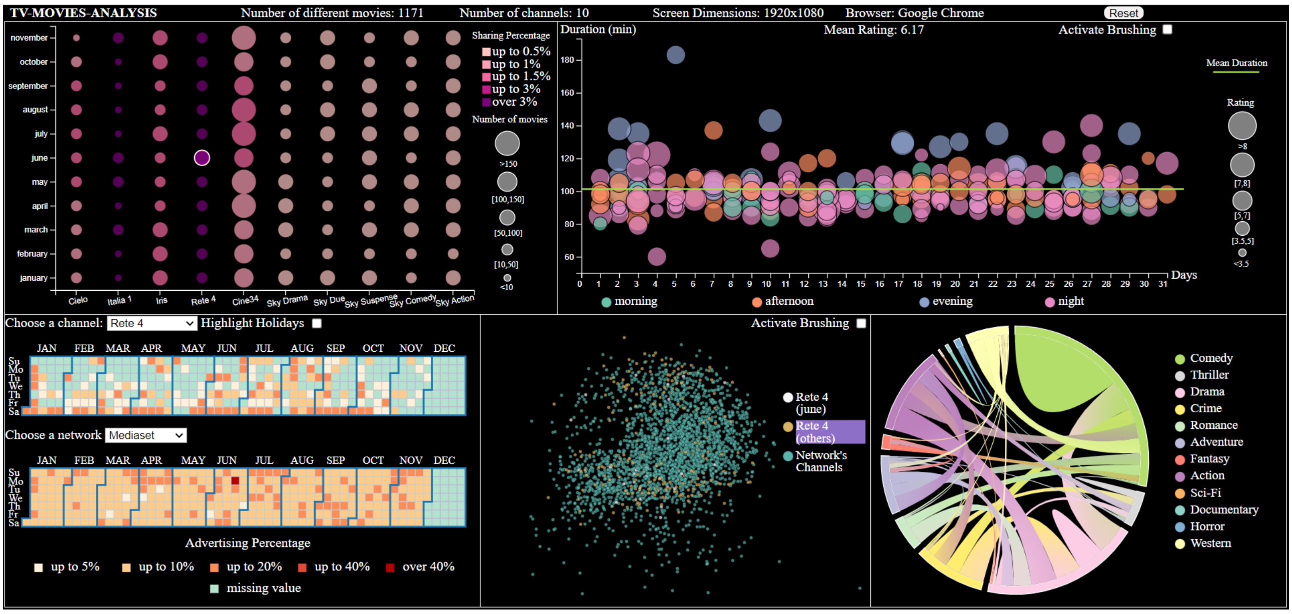
Moreover in [2], another conclusion is that *"[...] Italia1, RAI1 and Rete4 are the channels with the most Italian feature films broadcast in the night time slot [...]".* Also this trend can be visually confirmed. Indeed, choosing *Italia 1* or *Rete 4* in whatever month, it is possible to notice that the most colors present in the Bubbleplot belong to the *evening* or the *night* daytime (see the fig. 9).

6 Conclusions and Future Works

The final visual system allows for an interactive tool for TV guide analysis. It allows our intended user to make a survey on different channels of various networks using the different tools we have provided. He can make comparisons, select genres, focus on only some of the movies in the dataset, compare the sharing and duration of advertisements for each movie. This analysis allows him to understand how a channel of the network he manages could be improved or suggest him opening a specific new channel by putting together all the benefits that came out of this analysis.

Our project, as seen in the previous paragraphs, despite the complexity in interactions, manages to be user-friendly even in the color scheme that was chosen. Moreover, this project can be extended given the lack (according to our research) of a similar system and of data. In particular, if we had the possibility to use a complete dataset with more information, we could carry out a more accurate analysis. For instance, the audience age or a fined grained sharing data.

Finally, this project can be expanded by inserting other networks with their own channels (for example *Rai*) to be analyzed in the "Other" category which is currently scarce.



Choose a channel: Rete 4

Choose a network Mediaset

Figure 9: Movies broadcast mainly during the evening and night (Rete 4)

References

[1] M.Gambaro (2004), *The Relationship between Different Distribution Channels for Movies: Some Lessons from the Case of Free Television*

[2] J.Candeloro, M.Cucco (2008), *Italian Feature Films on National Public and Private Broadcasting Networks*

[3] M. Gasparini, D. Imparato (2007), *Forecasting TV audience: a consulting project with the Italian public television*

[4] `https://www.kaggle.com/datasets/komalkhetlani/imdb-dataset`

[5] `https://www.laguidatv.it/tutti-i-mesi`

[6] M. Angelini, G. Santucci Material of Visual analytics course, 2022/2023

[7] D3.js documentation `https://github.com/d3/d3/wiki`