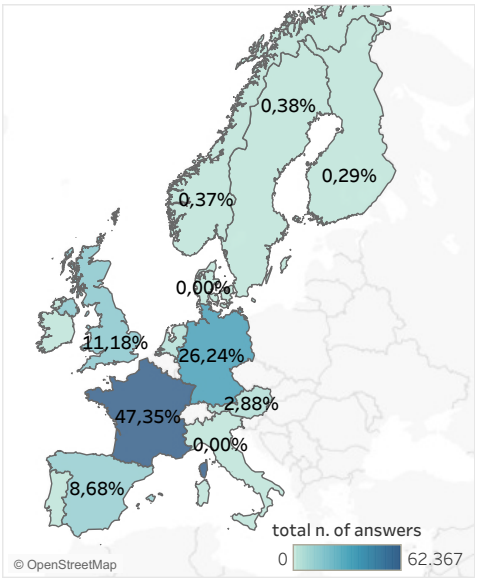
Predictive Analytics Challenge

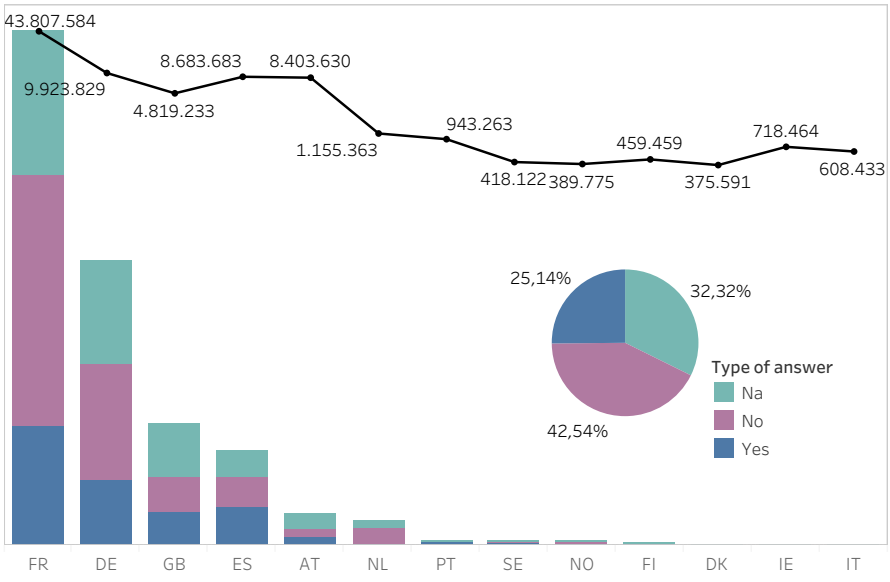
Task 1: Exploratory Data Analysis and Predictive model Ideas + Bonus Question

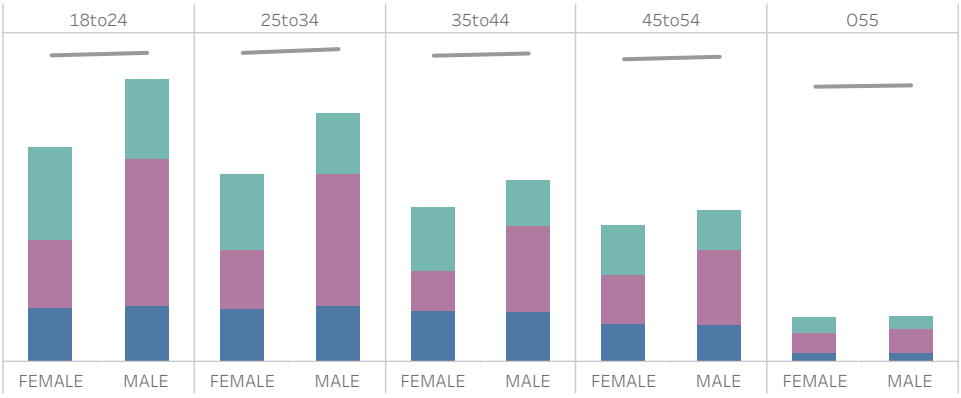
Claudia Stangarone

Results of the survey based on the demographics

The dashboard is interactive. Click or hover on each element to get more insights.

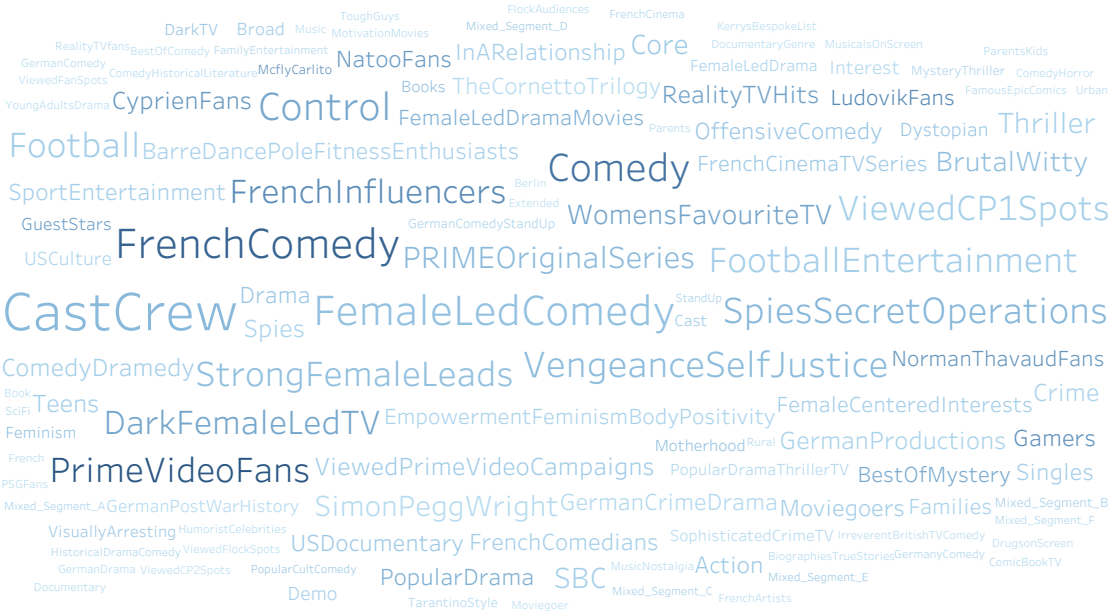






Word Cloud of all the taste audience

The size is given by the number of Yes at the survey the hue by the Impression



Filter the word cloud by selecting:

Gender
Tutti

Campaign Country
Tutti

Age (group)
Tutti

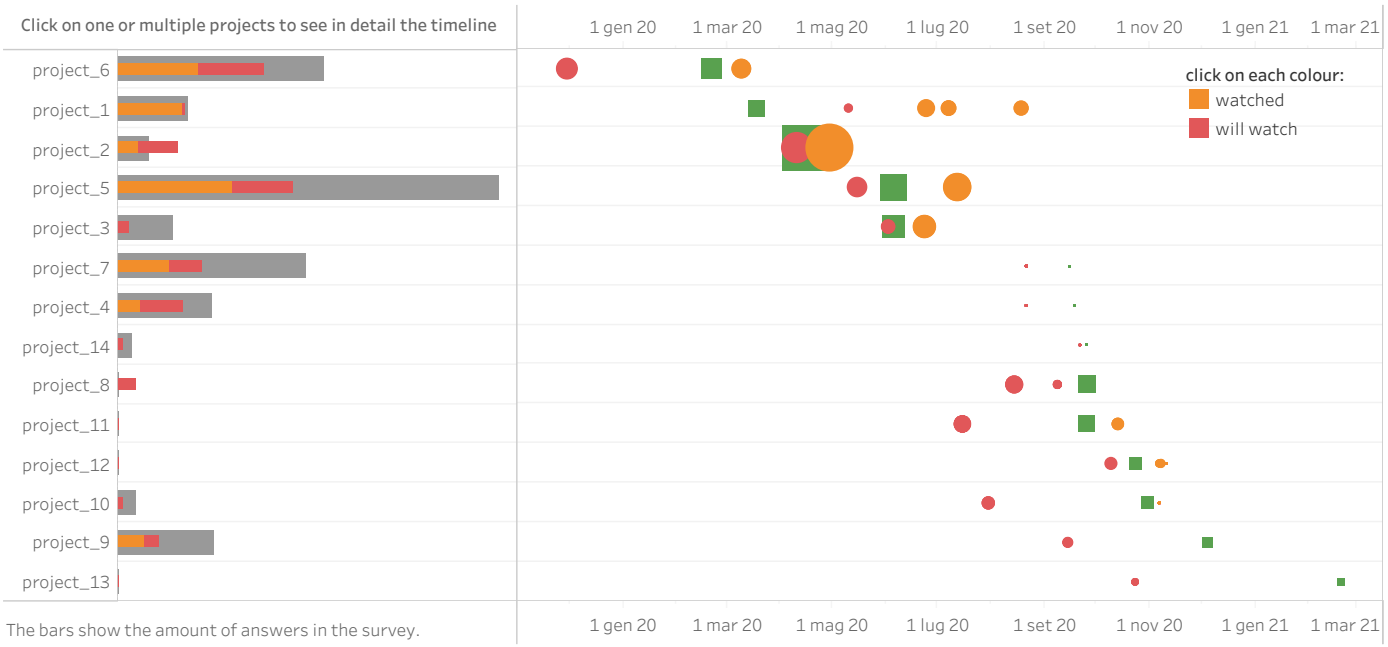
Intent
Tutti

Project Id
Tutti

Impressions
39 1.438.808

Timeline of the surveys and the release dates

Survey (circle) & Release (square). The size is given by the total amount of answers received on the survey date. Projects are ordered cro..



The bars show the amount of answers in the survey.
The **thick gray** bar are all the **NO** answers
The **thin line** are the **YES** answers.
Note: Looks like some projects didn't have any answer.

It looks like that intet surveys done few days before the release date, will most likely convert into viewership.

Surveys timeline: is there a seasonality in the rate of answers?

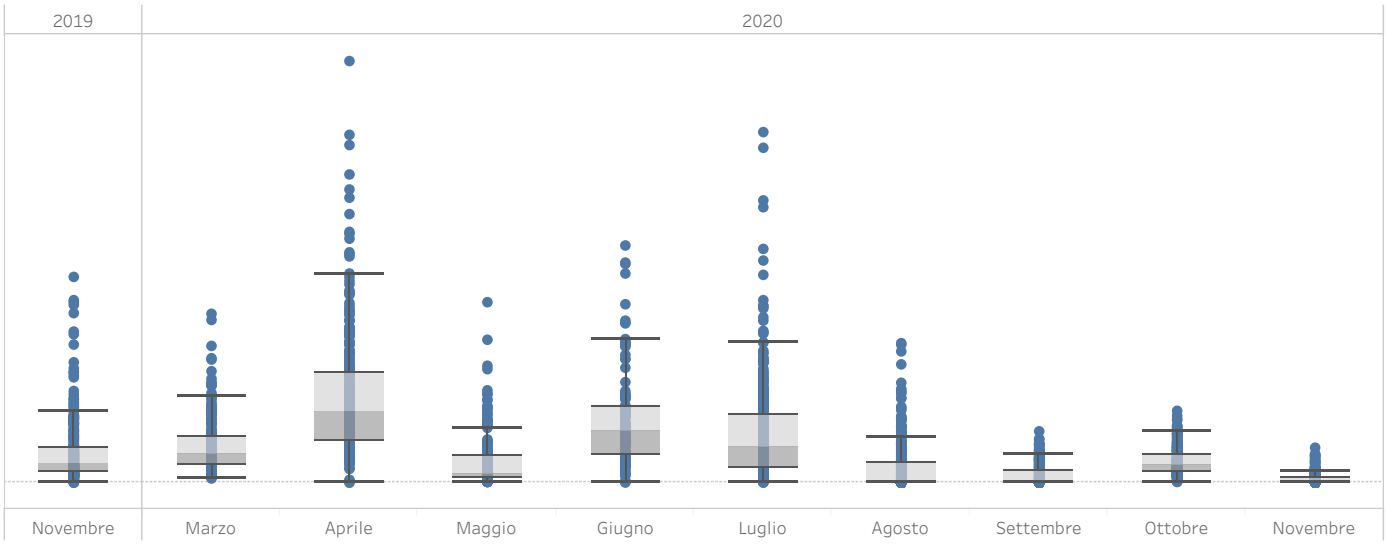
Filter the total amount of answers by the view intent, the age group the project id and the country, to see if there is any connection between the month of the survey and the response rate.

Intent
Tutti

Age (group)
~~All~~

Project Id
Tutti

Campaign Country
Tutti



Ideas for a *viewership conversion* prediction models

What are the potential prediction tasks that you can define for a given dataset and what are the problems to deal with?

Most likely I would employ a supervised machine learning model to predict the viewership conversion using linear regression or logistic regression-like models.
- I could imagine using a **linear regression model to predict the amount of yes to a survey that checks for the intent** (watched or will watch).

I would train my model on the historical data like the one provided, using as features the previous answers, impressions, audience taste, demographics, the country of the survey and also details if the title has been released or not (after the results of the EDA performed on the dataset provided, it looks like the time between the survey and the release date have somehow some influence on the viewership conversion) and the clustering results.

I would also use some real-time data (social media API, weather, news) as they may have an influence on the answer.

- **Logistic regression-like, or all those model that can provide a binary prediction (will watch or not)** based on the data provided (see the previous point). These models can be used as batch to determine the accuracy of each of them to the specific case.

Just to cite a few: KNeighbors Classifier, Gaussian Process Classifier, DecisionTree Classifier,Random Forest Classifier, Ada Boost Classifier, Quadratic Discriminant Analysis, Gaussian NB.

- The challenges I can imagine are:
If the datasets are sufficiently large and comprehensive datasets
The adaptability of models
Data cleaning/mining/wrangling
Data privacy and security

Bonus Questions:

Assuming that you have control over the incoming data integration, what would you propose improving/changing in dataset(s) in order to unlock more options for analysis/prediction?

I would collect the data per single answer, rather than clustering the answers per audience taste. This would allow to perform a better classification model in order to have a better train/test dataset to predict if the watching intention would convert into viewership.