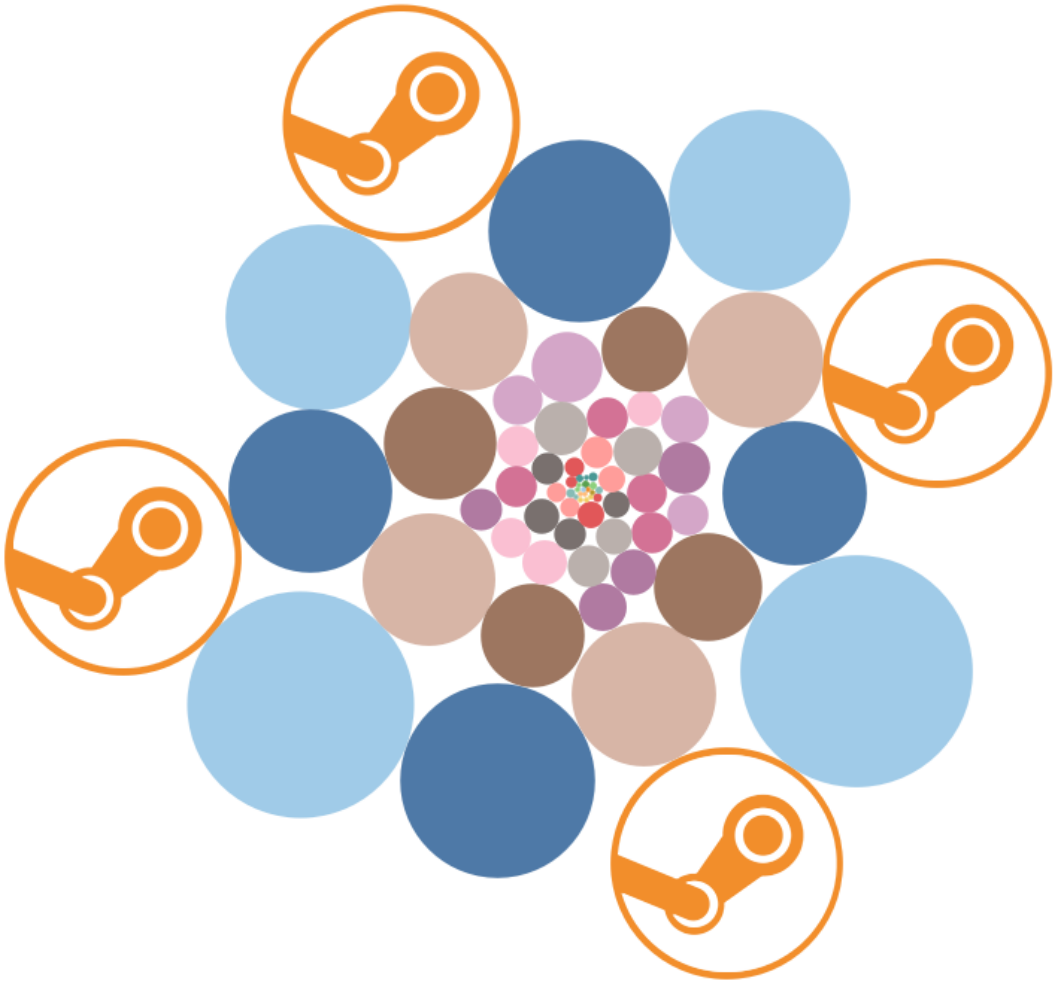
UNIVERSITÀ DELLA CALABRIA DIPARTIMENTO DI MATEMATICA E INFORMATICA

Course of Data Analytics (Data Warehouse and Visualization)

PUBLISHING OF STEAM STORE GAMES

An in-depth analysis of genre and category trends over the years and months to help developers release new games strategically over time.



Davide Gena

231873

INTRODUCTION

Steam is a video game digital distribution service and storefront by Valve. It was launched as a software client in September 2003 as a way for Valve to provide automatic updates for their games, and expanded to distributing and offering third-party game publishers' titles in late 2005.

Steam offers various features, like digital rights management (DRM), game server matchmaking and anti-cheat measures, and social networking and game streaming services. It provides the user with automatic game updating, saved game cloud synchronization, and community features such as friends messaging, in-game chat and a community market.

Offering all of these services, Steam is the store of choice for third-party developers to distribute their video games, in fact the Steam platform is the largest digital distribution platform for PC gaming, estimated around 75% of the market share.

About Dataset

Using data gathered from the Steam Store and SteamSpy APIs, this dataset provides information about various aspects of games on the store, such as its genre and the estimated number of owners.

Gathered around May 2019, it contains most games on the store released prior to that date. Unreleased titles were removed as well as many non-games like software.

The steam.csv file collects unique rows about 27033 video games. There are 18 columns described in the following table:

ATTRIBUTE NAME	DESCRIPTION	TYPE
appid	Unique identifier for each title	Numerical
name	Title of app (game)	String
release_date	Release date in format YYYY-MM-DD	Date format
english	Language support: 1 if is in	Boolean

	English	
developer	Name (or names) of developer(s). Semicolon delimited if multiple	String
publisher	Name (or names) of publisher(s). Semicolon delimited if multiple	String
platforms	Semicolon delimited list of supported platforms. At most includes: windows;mac;linux	String
required_age	Minimum required age according to PEGI UK standards. Many with 0 are unrated or unsupplied.	Numerical
categories	Semicolon delimited list of game categories, e.g. single-player;multi-player	String
genres	Semicolon delimited list of game genres, e.g. action;adventure	String
steamspy_tags	Semicolon delimited list of top steamspy game tags, similar to genres but community voted, e.g. action;adventure	String
achievements	Number of in-games achievements, if any	String
positive_ratings	Number of positive ratings, from SteamSpy	Numerical
negative_ratings	Number of negative ratings, from SteamSpy	Numerical
average_playtime	Average user playtime, from SteamSpy	Numerical
median_playtime	Median user playtime, from SteamSpy	Numerical

owners	Estimated number of owners. Contains lower and upper bound (like 20000-50000). May wish to take mid-point or lower	String
price	Current full price of title in GBP, (pounds sterling)	Numerical

GOAL

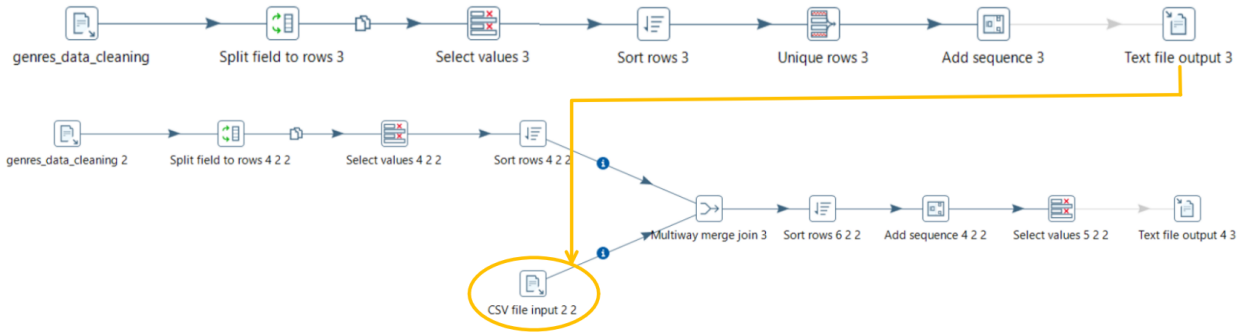
The main goal is to analyze genres and categories trends over the years and months in order to help third-party developers strategically release their products over time.

In addition, we want to analyze for each category and genre the rankings of developers, publishers and video games.

PRELIMINARY PHASE

An important preliminary phase is necessary before proceeding with the presentation of a relational scheme. As we can see in the previous attribute descriptions, the attributes `genres`, `categories`, `platforms`, `developer` and `publisher` can store multiple values in the same row delimited with a semicolon. This highlights many-to-many relations that can be modeled with the notion of multiple arcs, using special tables called **bridge tables**, that link the main table to a new table containing all the unique values. Then the attribute `genres` becomes a new table and so on.

For this purpose Pentaho was used, a powerful software to perform data transformation. Following the two groups of steps to obtain, for example, the table containing all the genres with an unique ID and then, the bridge table for the associations game - genre:



A bridge table, moreover, is characterized by a **weight** for each value that is important to perform weighted queries on values that have a different importance than others.

In this case the Pandas library was used for the creation of a script, called *MultipleArcsWeightGenerator.py* that take in input a bridge table without weights and calculates all weights as *1/n*, with n that in our case is the number of genres of a game, for example.

RELATIONAL DATABASE

Fixed this important issues we represent the dataset with a relational schema as a relational database:

steam(appid, name, release_date, required_age, achievements, positive_ratings, negative_ratings, average_playtime, median_playtime, owners, price)

steam_categories(category_id, category_name)

steam_genres(genres_id, genre_name)

steam_platforms(platform_id, platform_name)

steam_developers(developer_id, developer_name)

steam_publishers(publisher_id, publisher_name)

appcategory_associations(appcategory_id, appid: steam, category_id: steam_categories, weights)

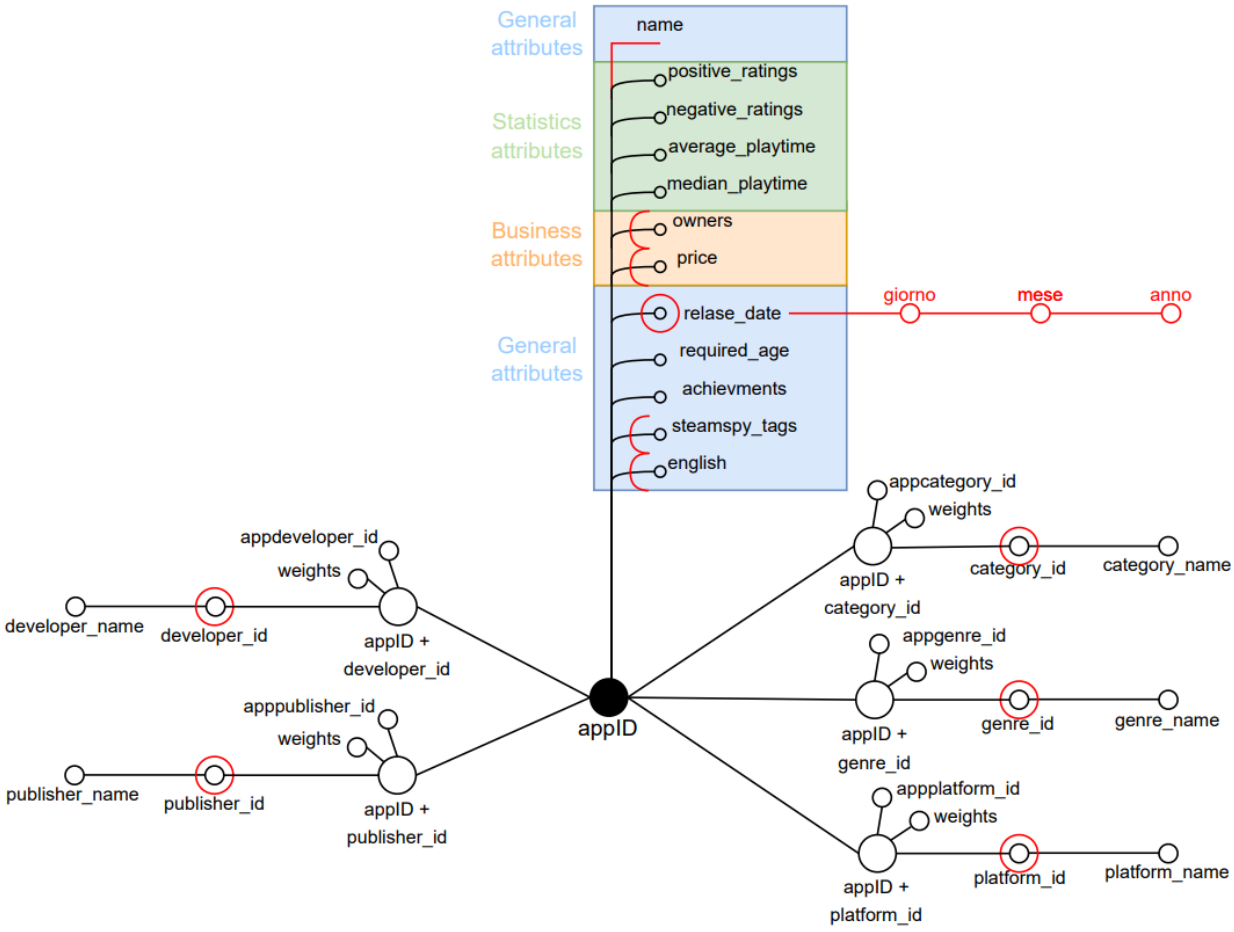
appgenre_associations(appgenre_id, appid: steam, genres_id: steam_genres, weights)

appplatform_associations(appplatform_id, appid: steam, platform_id: steam_platform, weights)

appdeveloper_associations(appdeloper_id, appid: steam, developer_id: steam_developers, weights)

apppublisher_associations(apppublisher_id, appid: steam, publisher_id: steam_publishers, weights)

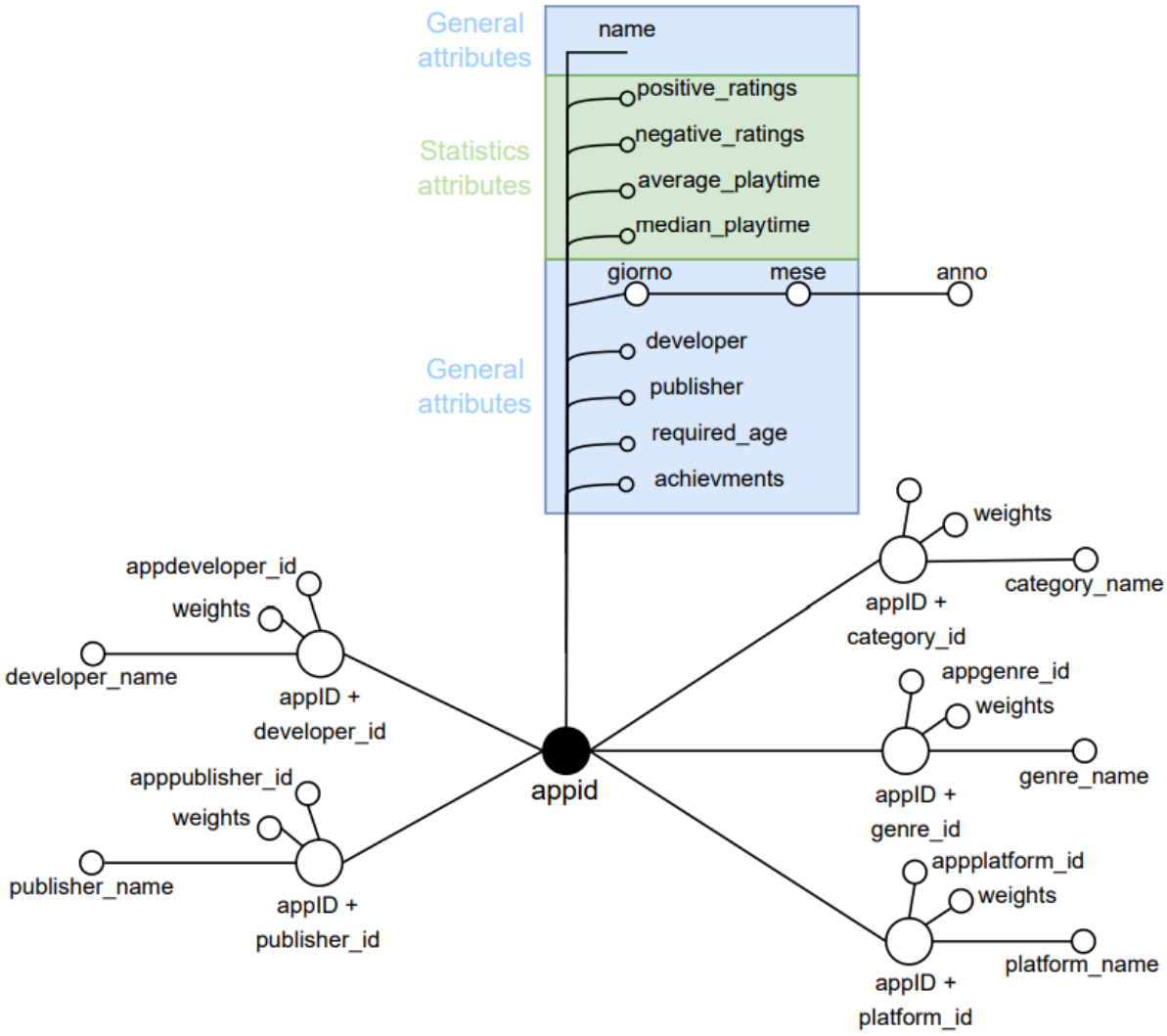
ATTRIBUTE TREE



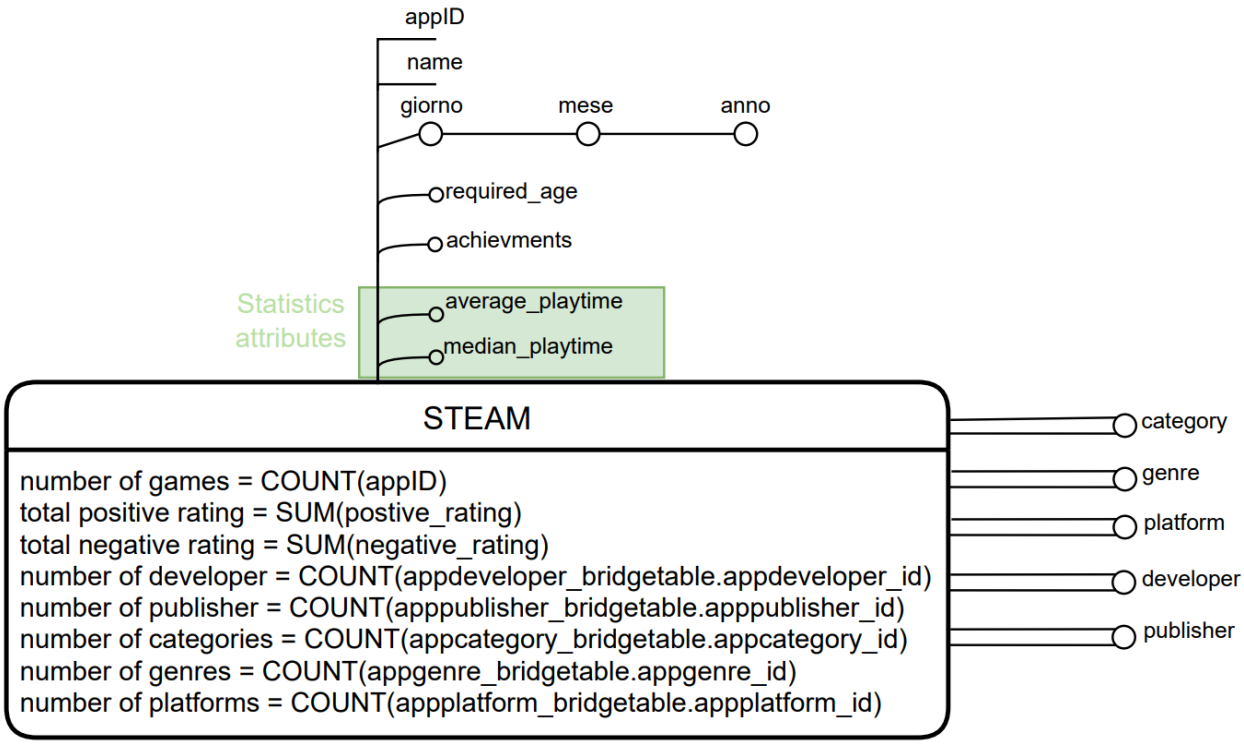
According with the goal explained previously, some attribute are useless for our analysis:

- The attribute **english**, because is just a boolean value equal to 1 if a game supports the english language and in most cases it happens.
- The attribute **steamspy_tags**, because is a clone of the attributes *genre* and *category*.
- The attribute **price** because we are not interested in a business analysis.
- The attribute **owners** for the same reason of the attribute *price*.

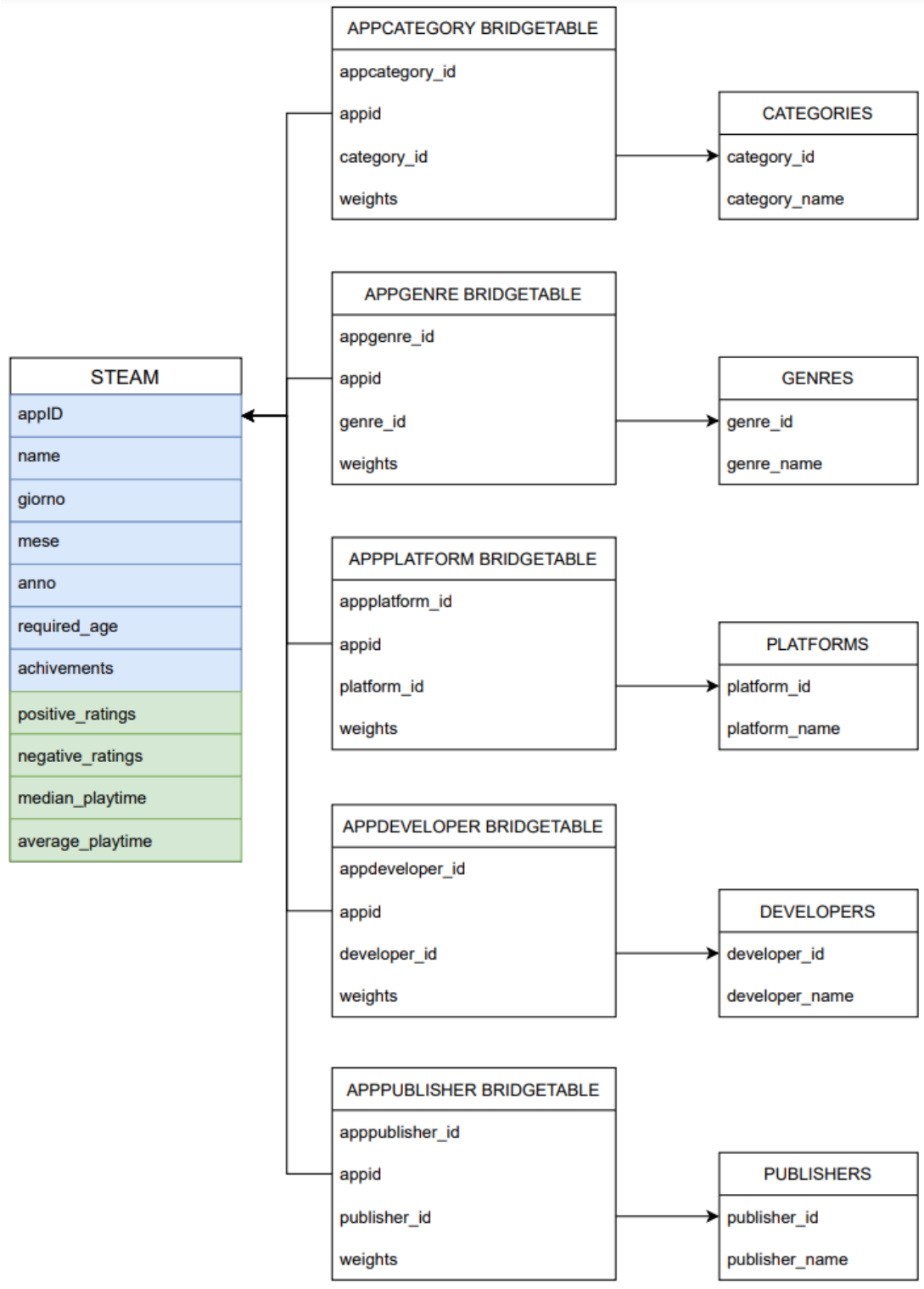
EDITED TREE


General attributes
name
Statistics attributes
positive_ratings
negative_ratings
average_playtime
median_playtime
giorno
mese
anno
General attributes
developer
publisher
required_age
achievments
appdeveloper_id
weights
developer_name
appID + developer_id
apppublisher_id
weights
publisher_name
appID + publisher_id
appid
weights
category_name
appID + category_id
appgenre_id
weights
genre_name
appID + genre_id
appplatform_id
weights
platform_name
appID + platform_id

FACT SCHEMA


appID
name
giorno
mese
anno
required_age
achievments
Statistics attributes
average_playtime
median_playtime
STEAM
number of games = COUNT(appID)
total positive rating = SUM(postive_rating)
total negative rating = SUM(negative_rating)
number of developer = COUNT(appdeveloper_bridgetable.appdeveloper_id)
number of publisher = COUNT(apppublisher_bridgetable.apppublisher_id)
number of categories = COUNT(appcategory_bridgetable.appcategory_id)
number of genres = COUNT(appgenre_bridgetable.appgenre_id)
number of platforms = COUNT(appplatform_bridgetable.appplatform_id)
category
genre
platform
developer
publisher

SNOWFLAKE SCHEMA



DATA QUALITY CHECKS

Using another Pandas script, called *DataQualityChecker.py,* we are able to see information about the quality of the dataset. In particular we know that there are no missing values, no duplicate rows and all rows are complete:

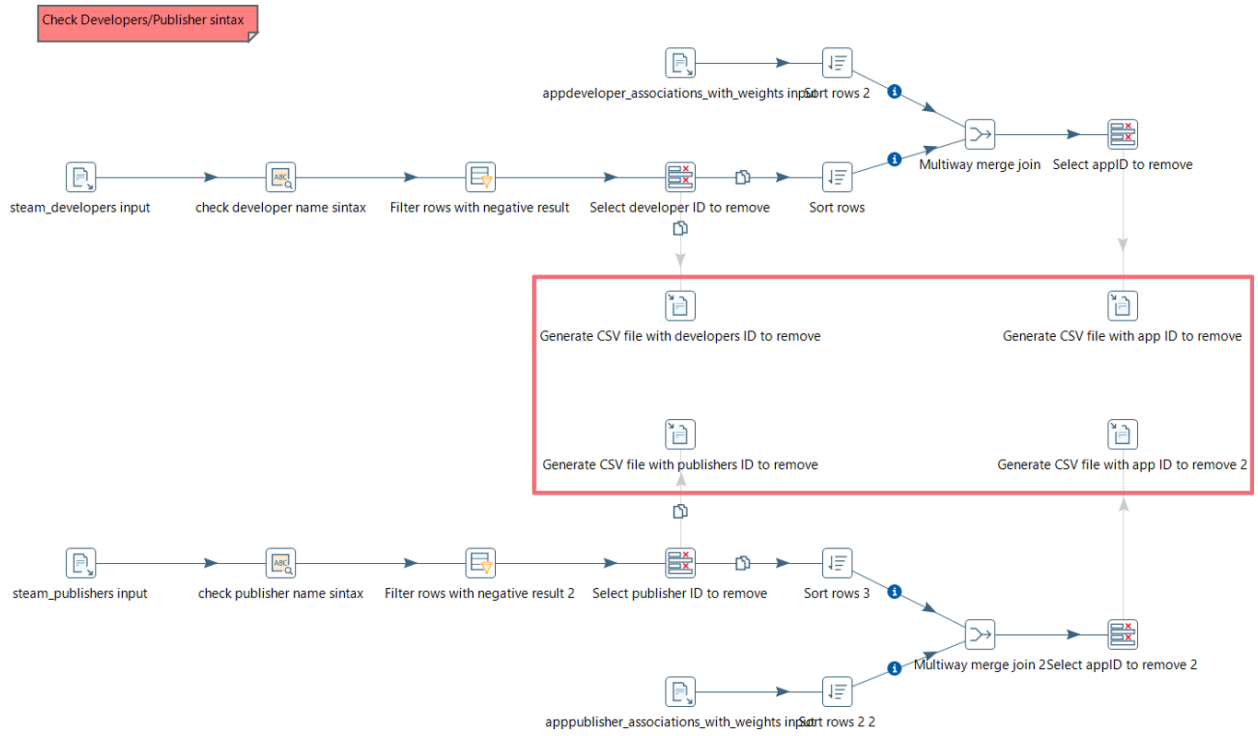
```
Percentage of missing values:
appid                0.0
name                 0.0
release_date         0.0
english              0.0
developer            0.0
publisher            0.0
platforms            0.0
required_age         0.0
categories           0.0
genres               0.0
steamspy_tags        0.0
achievements         0.0
positive_ratings     0.0
negative_ratings     0.0
average_playtime     0.0
median_playtime      0.0
owners               0.0
price                0.0
dtype: float64

Number of rows with at least one missing data:
0

Percentage of duplicates rows (False if unique, else True):
False    100.0
dtype: float64
```

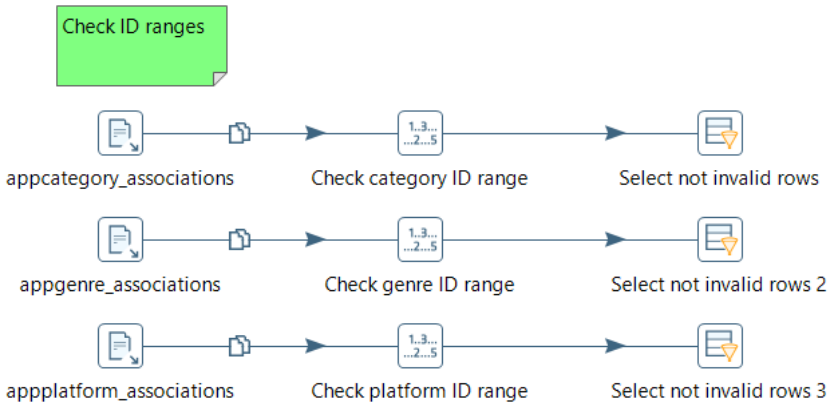
If there are missing values, the script will automatically replace them with valid values and if there are duplicate rows these will be dropped!

Moreover, with the following Pentaho steps, the syntax of the attributes `developer` and `publisher` are checked:



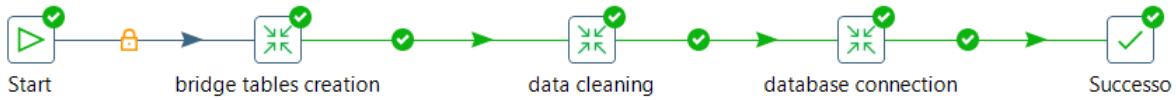
This Pentaho procedure generates 4 new .csv files that contain all the app ID, developer ID and publisher ID with wrong syntax of the corresponding names. Then, a Pandas script, called *InconsistentRowsDestroier.py*, takes in input these .csv files and drops all the rows with these IDs from the original tables.

In addition there are other Pentaho steps to check the ranges of the genre IDs, for example, of the corresponding bridge table to prevent future wrong rows addition:



The last Pentaho transformation loads all the original tables of the relational database and the cleaned ones in a **PostgreSQL** server.

All these transformations can been finally collect into a job:



ANALYSIS SHEETS AND DASHBOARDS

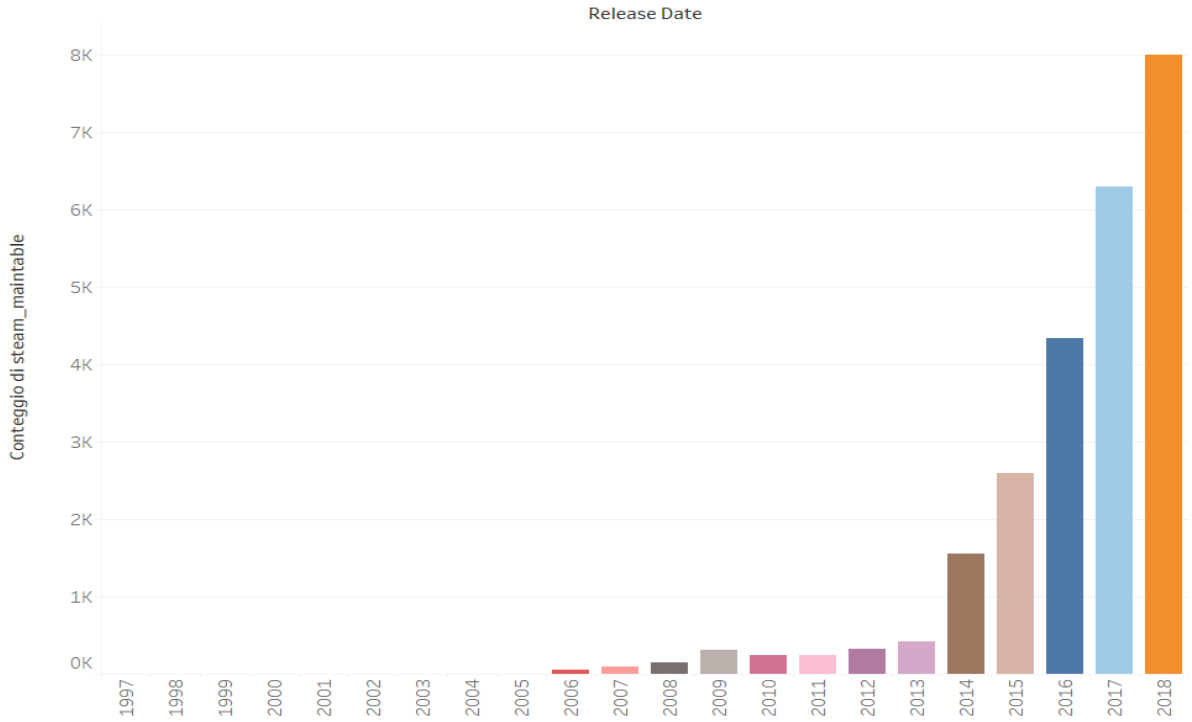
In this last, but not least, step, we use Tableau sheets to build plots in order to analyze our goal. Then sheets are collected into dynamic dashboards that permit final users to interact with more sheets in a unique place. To define an order of sheets and dashboards we can finally use a story.

SHEETS

In this first part of the analysis just some important sheets are described, because they are used as controllers of the dashboards or not included in them.

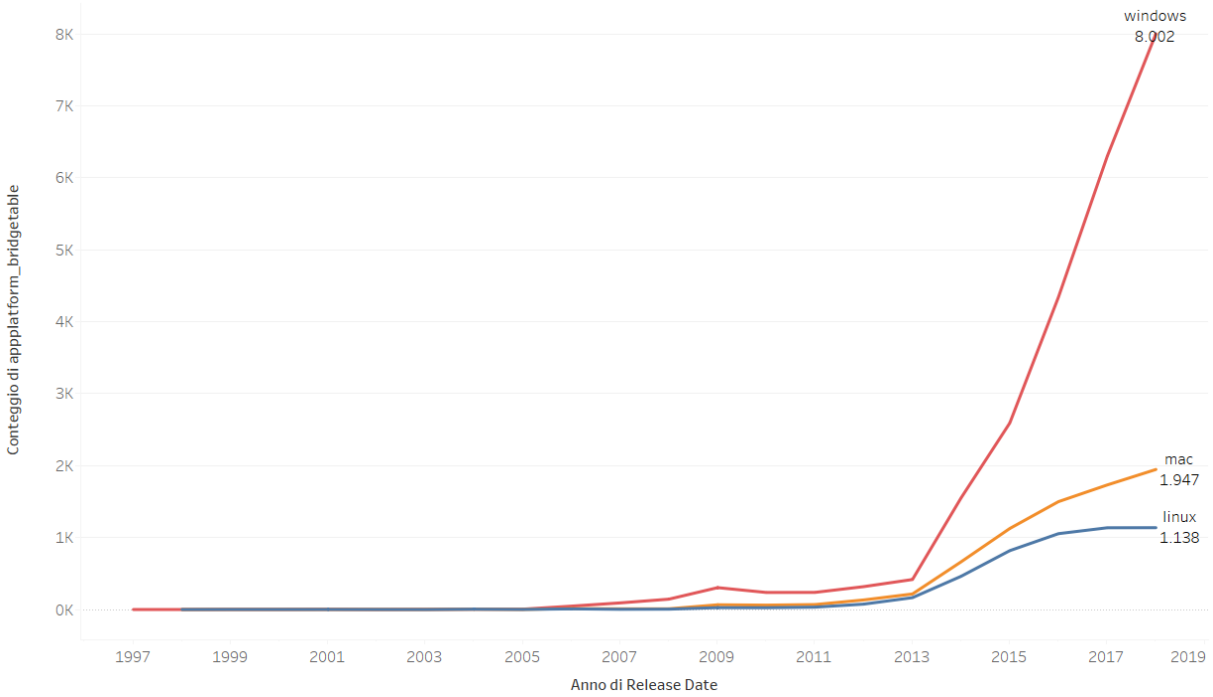
Video game development over the years

A simple and intuitive plot to understand how the video game industry is booming. This sheet will be used as a controller in several dashboards that we'll see later.



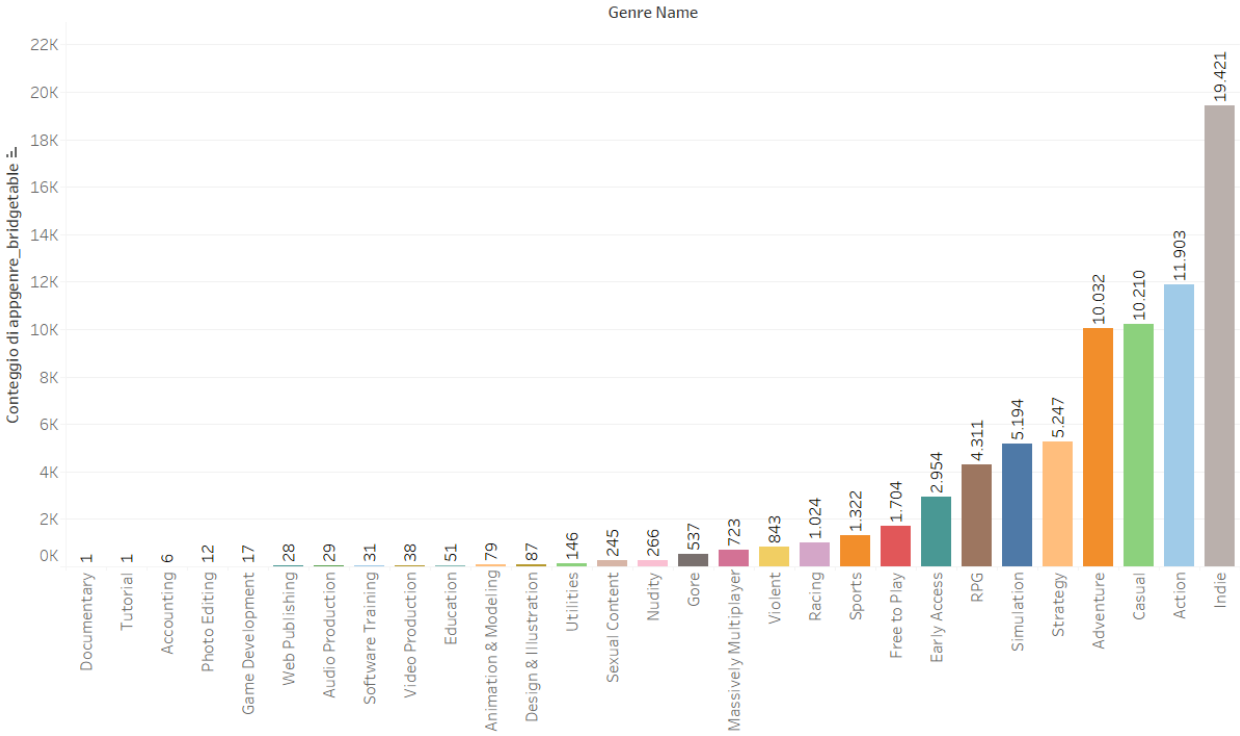
Platforms support over year

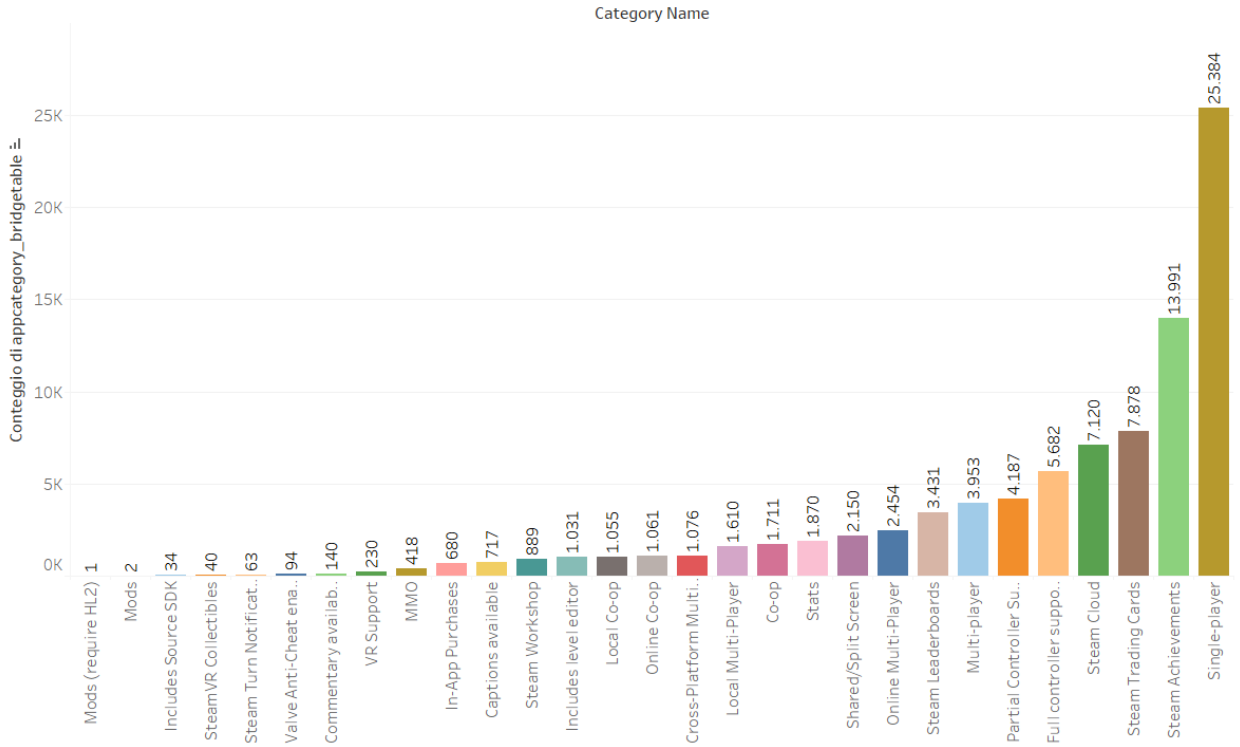
Thanks to this intuitive sheet we can easily confirm what we already knew and in the final chapter we can draw our own considerations on this.



Most developed genres and categories

Two other simple sheets are used as controllers in the dashboards and represent the most developed genre and category in all the years and hosted in the Steam store.



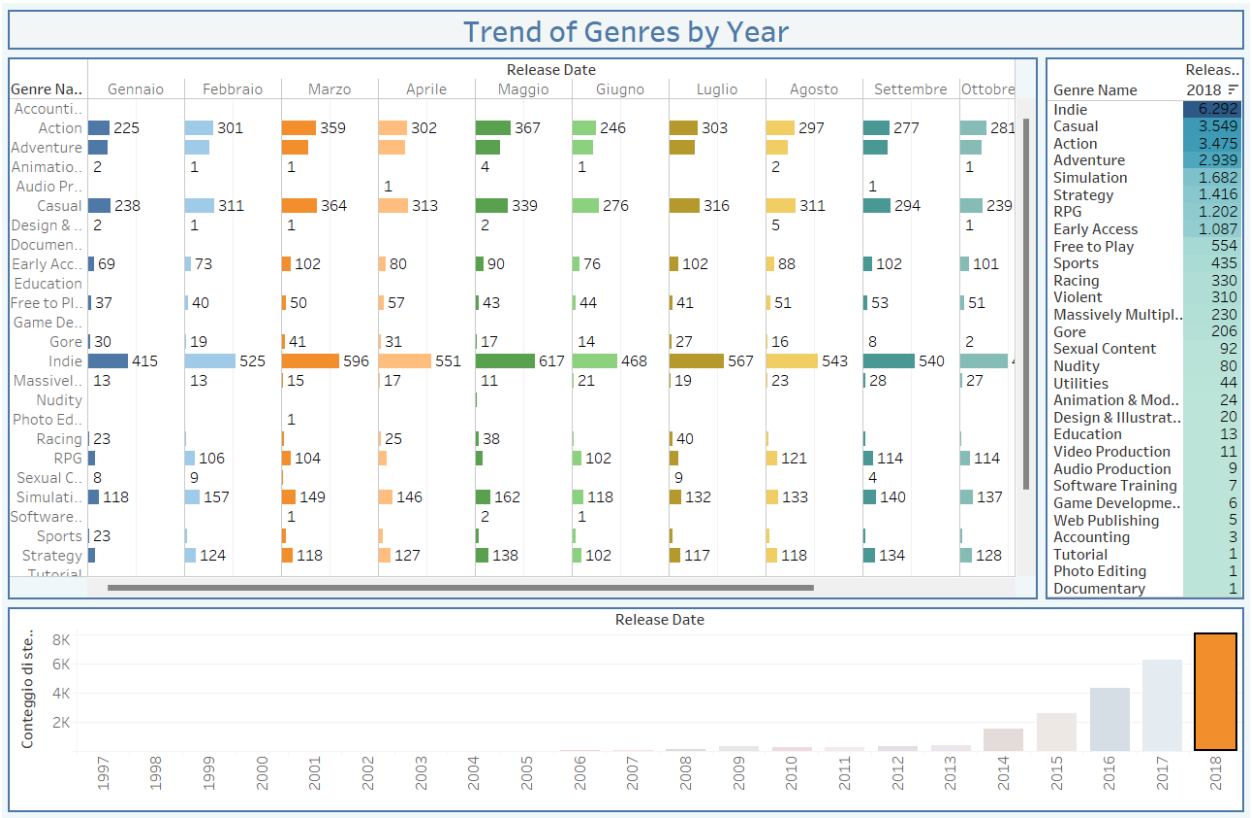


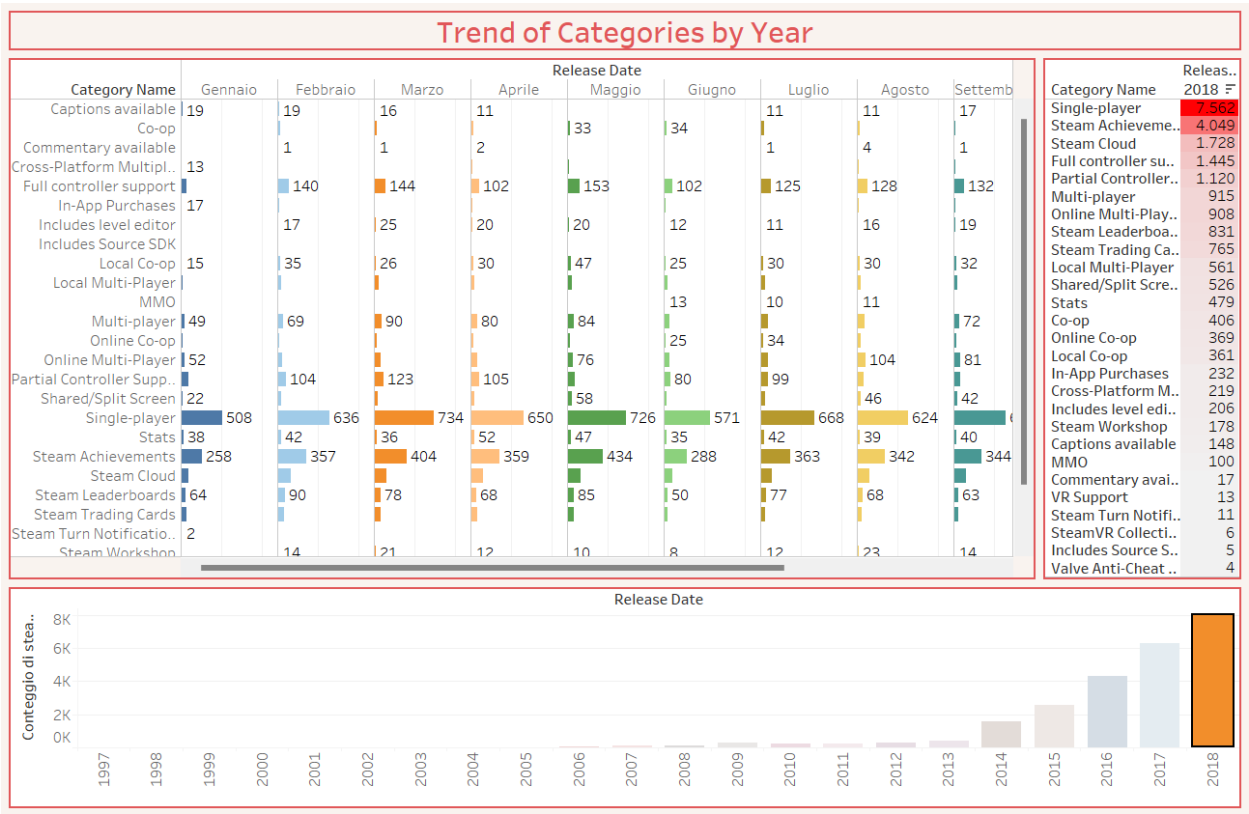
DASHBOARDS

Now these simple sheets will be integrated in dashboards with more detailed sheets to analyze trends and ranks.

Trends of Genres and Categories by year and month

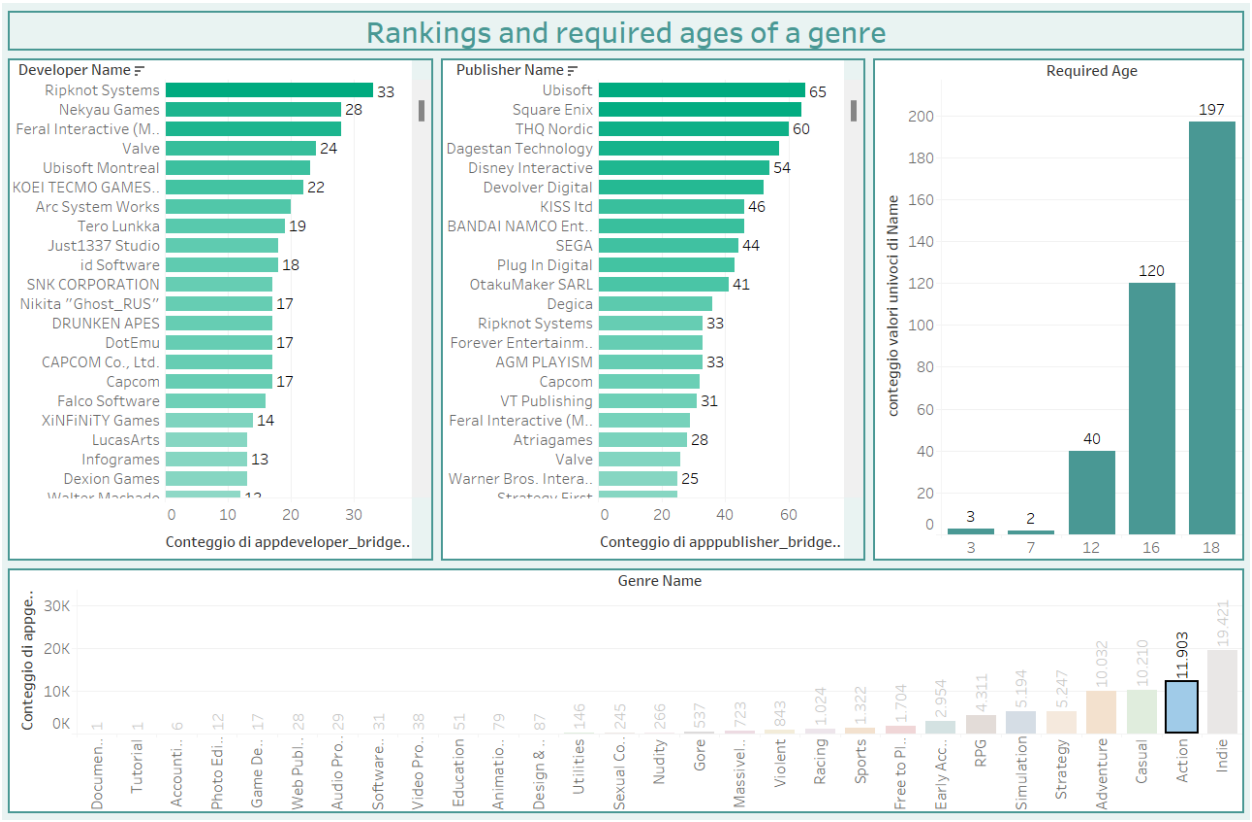
Thanks to the two following dashboards, selecting a year will show the trend of genre and category on the left panel, while on the right it will be possible to deepen the trend month by month of the same selected year.

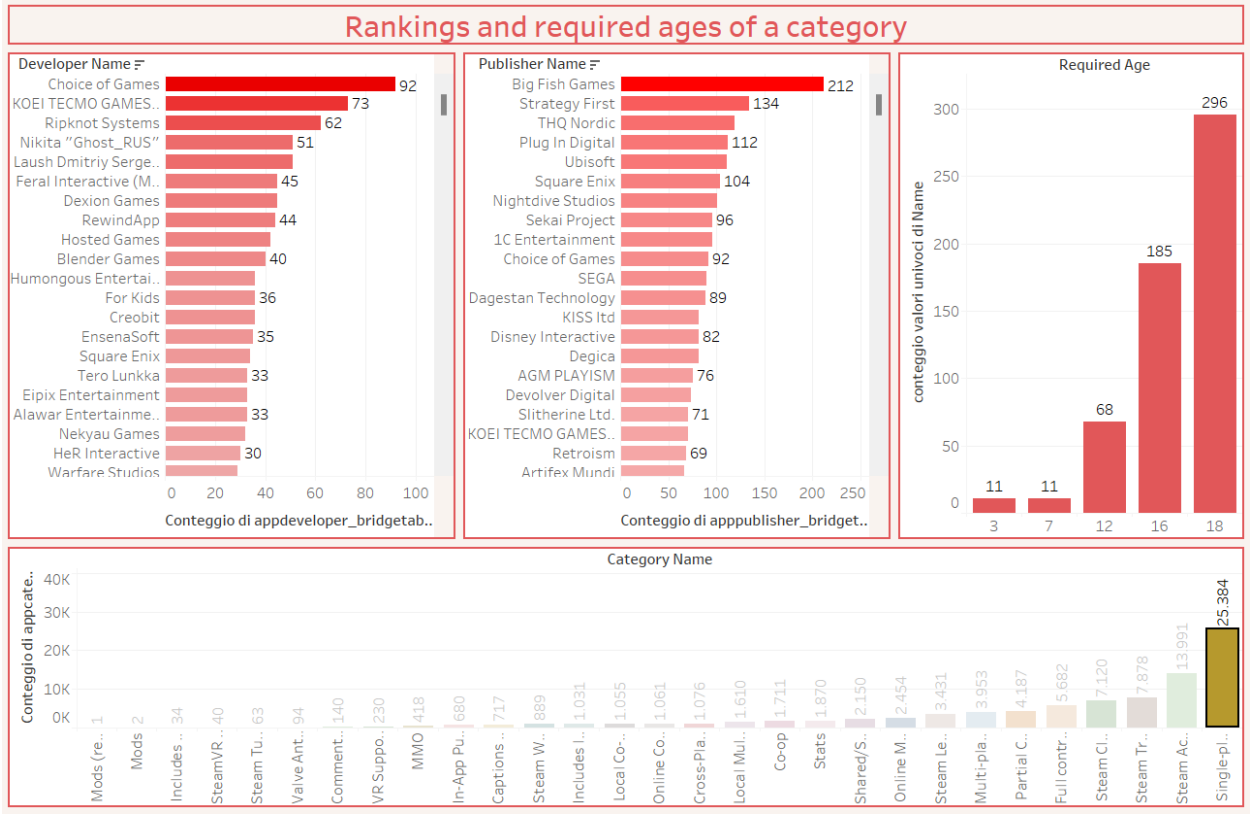




Rankings and required ages of genres and categories

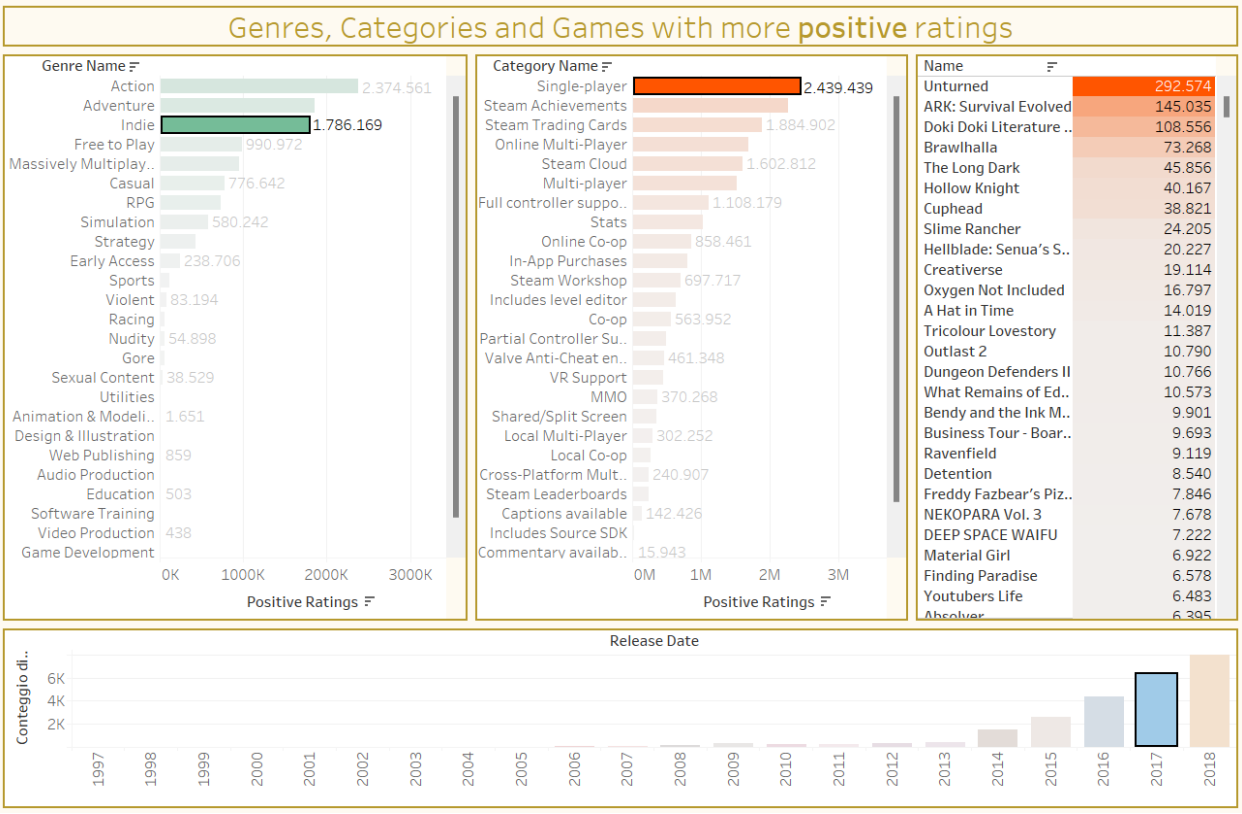
With the following two dashboards it is possible to select a genre or a category and see ranks about the developers and publishers that most have developed video games with that genre or category. Moreover it is possible to read information about the most applied minimum required age for that genre or category.

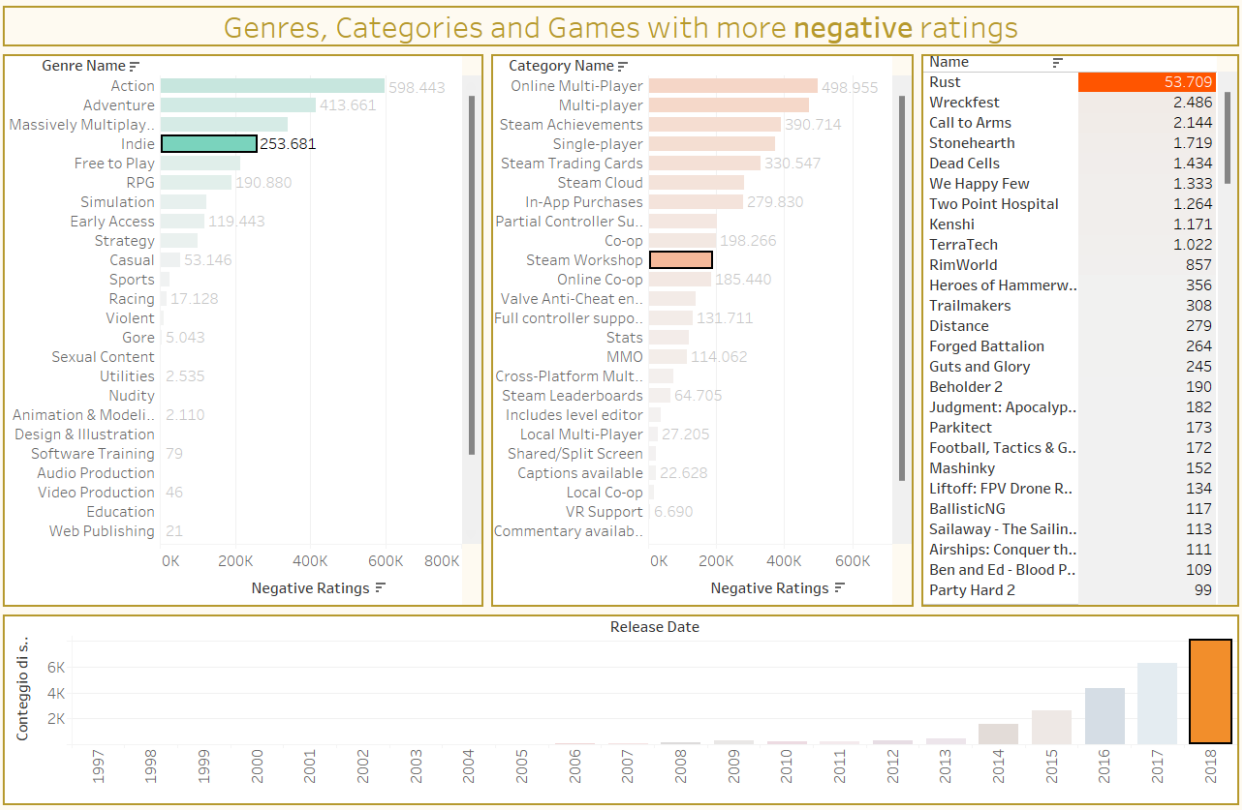

Rankings and required ages of a genre
Developer Name
Ripknot Systems 33
Nekyau Games 28
Feral Interactive (M..
Valve 24
Ubisoft Montreal
KOEI TECMO GAMES.. 22
Arc System Works
Tero Lunkka 19
Just1337 Studio
id Software 18
SNK CORPORATION
Nikita "Ghost_RUS" 17
DRUNKEN APES
DotEmu 17
CAPCOM Co., Ltd.
Capcom 17
Falco Software
XiNFiNiTY Games 14
LucasArts
Infogrames 13
Dexion Games
0 10 20 30
Conteggio di appdeveloper_bridge..
Publisher Name
Ubisoft 65
Square Enix
THQ Nordic 60
Dagestan Technology
Disney Interactive 54
Devolver Digital
KISS ltd 46
BANDAI NAMCO Ent..
SEGA 44
Plug In Digital
OtakuMaker SARL 41
Degica
Ripknot Systems 33
Forever Entertainm..
AGM PLAYISM 33
Capcom
VT Publishing 31
Feral Interactive (M..
Atriagames 28
Valve
Warner Bros. Intera.. 25
0 20 40 60
Conteggio di apppublisher_bridge..
Required Age
conteggio valori univoci di Name
0 20 40 60 80 100 120 140 160 180 200
3 2 40 120 197
3 7 12 16 18
Genre Name
Conteggio di appge..
0K 10K 20K 30K
Documen.. 1
Tutorial 1
Accounti.. 6
Photo Edi.. 12
Game De.. 17
Web Publ.. 28
Audio Pro.. 29
Software.. 31
Video Pro.. 38
Education 51
Animatio.. 79
Design &.. 87
Utilities 146
Sexual Co.. 245
Nudity 266
Gore 537
Massivel.. 723
Violent 843
Racing 1.024
Sports 1.322
Free to Pl.. 1.704
Early Acc.. 2.954
RPG 4.311
Simulation 5.194
Strategy 5.247
Adventure 10.032
Casual 10.210
Action 11.903
Indie 19.421


Rankings and required ages of a category
Developer Name
Choice of Games 92
KOEI TECMO GAMES.. 73
Ripknot Systems 62
Nikita "Ghost_RUS" 51
Laush Dmitriy Serge..
Feral Interactive (M.. 45
Dexion Games
RewindApp 44
Hosted Games
Blender Games 40
Humongous Entertai..
For Kids 36
Creobit
EnsenaSoft 35
Square Enix
Tero Lunkka 33
Eipix Entertainment
Alawar Entertainme.. 33
Nekyau Games
HeR Interactive 30
Warfare Studios
0 20 40 60 80 100
Conteggio di appdeveloper_bridgetab..
Publisher Name
Big Fish Games 212
Strategy First 134
THQ Nordic
Plug In Digital 112
Ubisoft
Square Enix 104
Nightdive Studios
Sekai Project 96
1C Entertainment
Choice of Games 92
SEGA
Dagestan Technology 89
KISS ltd
Disney Interactive 82
Degica
AGM PLAYISM 76
Devolver Digital
Slitherine Ltd. 71
KOEI TECMO GAMES..
Retroism 69
Artifex Mundi
0 50 100 150 200 250
Conteggio di apppublisher_bridget..
Required Age
conteggio valori univoci di Name
0 50 100 150 200 250 300
11 11 68 185 296
3 7 12 16 18
Category Name
Conteggio di appcate..
0K 10K 20K 30K 40K
Mods (re.. 1
Mods 2
Includes .. 34
SteamVR.. 40
Steam Tu.. 63
Valve Ant.. 94
Comment.. 140
VR Suppo.. 230
MMO 418
In-App Pu.. 680
Captions .. 717
Steam W.. 889
Includes I.. 1.031
Local Co.. 1.055
Online Co.. 1.061
Cross-Pla.. 1.076
Local Mul.. 1.610
Co-op 1.711
Stats 1.870
Shared/S.. 2.150
Online M.. 2.454
Steam Le.. 3.431
Multi-pla.. 3.953
Partial C.. 4.187
Full contr.. 5.682
Steam Cl.. 7.120
Steam Tr.. 7.878
Steam Ac.. 13.991
Single-pl.. 25.384

Rankings and required ages of genres and categories

These are the most dynamic dashboards. Selecting just a year it is possible to see the ranks of genres and categories with more positive/negative ratings and the games with more positive/negative ratings in that year. In addition, selecting also a genre and/or a category it is possible to filter the video games list with that genre and/or category.





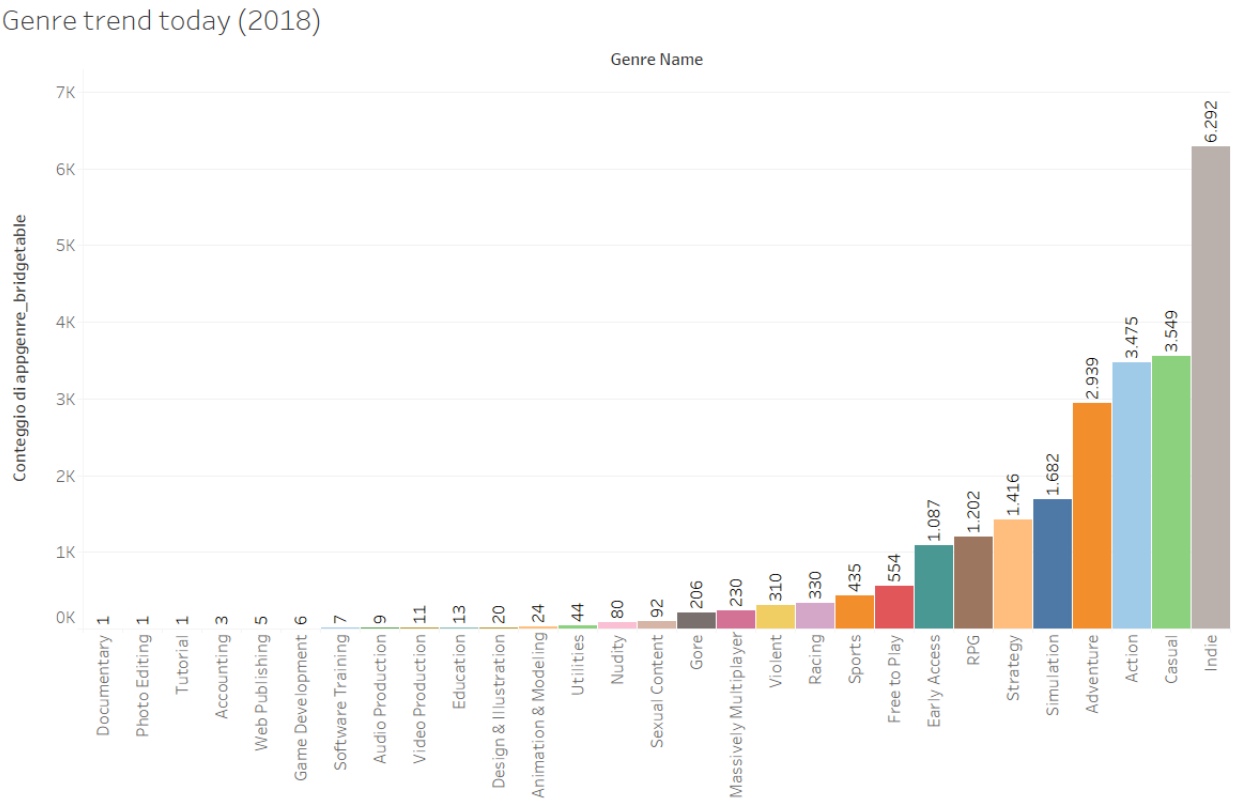
STORY

Designed and created all the needed sheets and dashboards for our analysis, we can collect them inside [Tableau Stories]. In our case is sufficient one story that summarizes all the analysis paths seen previously:

CONCLUSIONS

After the ETL process made with Pentaho software and Pandas scripts, and after the Tableau analysis we can take our conclusions related to different parts of the topic:

- First of all we can talk about the **OS support** of the released softwares. In most cases developers release video games with Windows support only and this is what we expected. Today we know that Steam Corporation released, on 25 february 2022, a new portable console with a proprietary Linux based OS, so it will probably encourage the development of video games on this platform. Moreover, many other third-party consoles are made using Linux based OS, but in this case we are considering just video games hosted on the Steam store.
- Considering the **trend of genres** of the 2018 (last complete year of the dataset), we notice that the Indie games, so the third-party games, are the most developed:



This is an important result that highlights the great growth of the video games sector. Not too many years ago this was impossible, because the small developers did not have the technologies suitable for development, they were not free to use the graphic engines that were often of private use of large companies and they also did not have the right tools to publish and distribute the finished product.

So this is good news, because large companies can take note of these results and maybe finance small future projects, as is already happening.

On a strategic point of view, "*Indie*" means just that the game is not a so-called Triple-A, so a game developed by a big company, but is developed by one or a little group of developers. In the dataset *Indie* is often accompanied with other genres, so we can also take a look at the other genres, like *casual*, *action*, *adventure* and so on.

REFERENCES

1. "Steam Store games" Dataset (on Kaggle platform)
2. Pandas Library documentation
3. Data Warehouse course book (By Matteo Golfarelli)
4. Tableau Software Documentation