UNIVERSITÀ DELLA CALABRIA

BUSINESS INTELLIGENCE AND ANALYTICS (Data Warehouse Module)

CORSO DI LAUREA MAGISTRALE IN COMPUTER SCIENCE

Source selection

For this project, the "Kickstarter Projects" (https://www.kaggle.com/kemical/kickstarter-projects) dataset has been chosen, that contains all the projects launched on the Kickstarter platform from 2009 to 2018.

Each project is identified by various attributes, listed below.

Nome	Descrizione
ID	Project identifier
Name	Project name
Main Category	Main category of the project (15 values)
Category	Subcategory of the project (159 values)
Currency	Currency used for fundraising
Launched	Start date
Deadline	End date of the campaign
Goal	Total goal to achieve (original currency)
Pledged	Total currently raised (original currency)
State	Project status, it can be: - canceled - failed - live - successful - suspended - undefined
Backers	Number of users who supported the campaign
Country	Country of project (23 values)
USD_Pledged	USD conversion of the 'Pledged' field (done by Kickstarter)
USD_Pledged_Real	USD conversion of the 'Pledged' field (done by Fixer.io)
USD_Goal_Real	USD conversion of the 'Goal' field (done by Fixer.io)

Data Cleaning

The dataset was analyzed using **Tableau Prep**, then modified using **Pentaho Data Integration** (Kettle).

In general, the data is quite consistent, except for some issues:

- **Name**: 5 records have null values.
- **Launched**: 7 records have an impossible date, before Kickstarter opened.
- **State**: 3562 records have an undefined value (*undefined*). Validity: 99.99%.
- **Country**: 3797 have an unknown value (*N,O"*). Validity: 99.99%.
- **USD Pledged**: 3797 have a null value. Completeness: 99.99%.

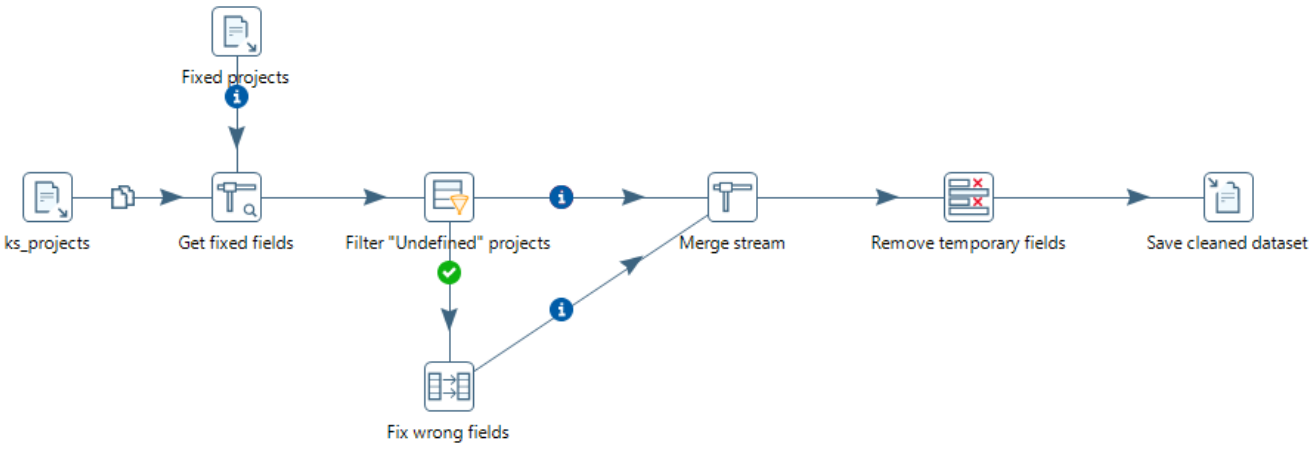
It was noticed that the unknown values of *Country* correspond 1:1 with the null values of *USD Pledged*, and the undefined values of *State* are a subset of them.

The records to be deleted would therefore be about **4,000** (about 1% of the total). Since only a couple of fields are corrupted, rather than deleting entire rows, it was preferred to try to save them by integrating the incorrect values, extracting them directly from Kickstarter.

However, as Kickstarter does not seem to offer public APIs, it was necessary to write a Python script that, given the list of incorrect records using Pentaho, made a request to the server for each record through **a name search**. Each request returns the response in JSON format, which is processed to create a new *CSV* file with the reference **project ID and the correct fields**.

The absence of a search by ID prevented the correct extraction of all incorrect projects, while the absence of official APIs led to the activation of **temporary access blocks** after too many consecutive requests (about 200), slowing down the data extraction phase. In total, **2953 projects out of 3562 (about 83%) were integrated**.

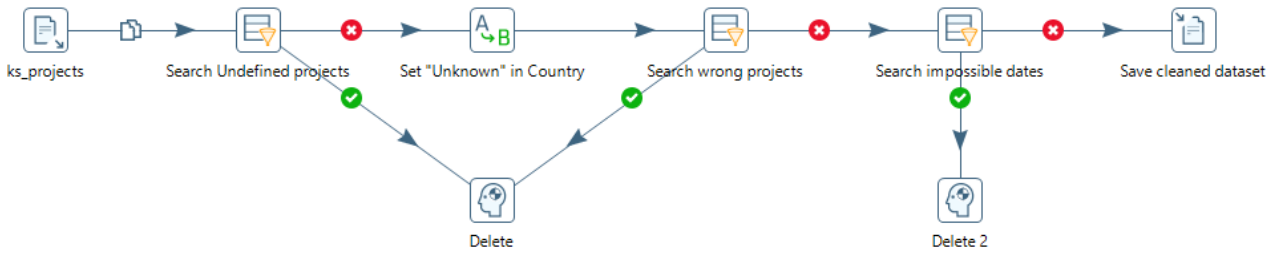
The retrieved projects were then finally merged into the original dataset using Pentaho.



In the Cleaning part, the following operations were carried out:

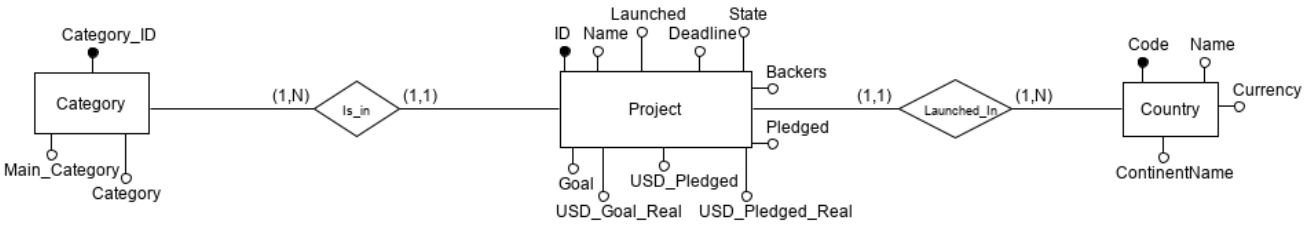
- Deletion of records with **undefined** status: being an important field, it was deemed necessary to completely delete the records (about 600).
- Replacement of values *N, O"* with *Unknown* in the **Country** field.
- Elimination of **logically impossible** records, *e.g.* projects with zero backers but with an amount of funds raised greater than zero, or projects with a start date later than the end date, etc.

In total, **820** records were deleted (about 0.2% of the total), compared to the initially expected 4,000.



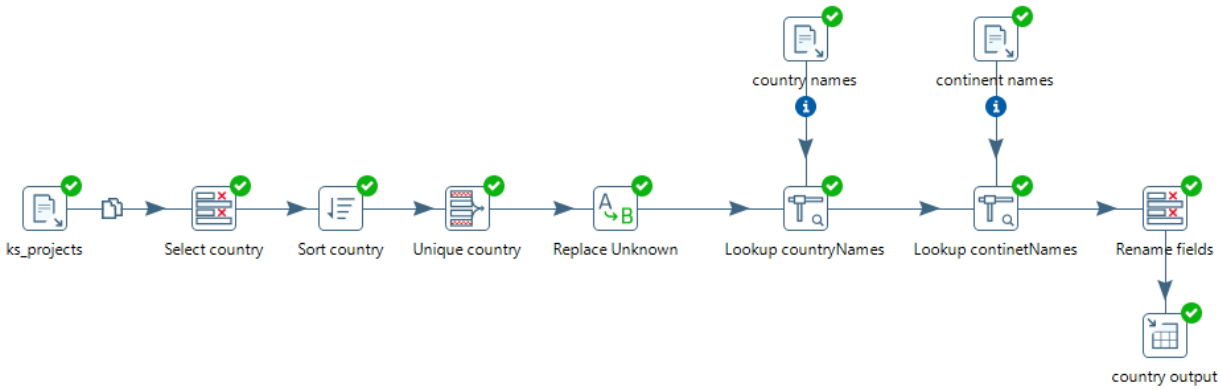
Once the dataset was cleaned, Pentaho was used to define a **MySQL** database.

Schema E-R:



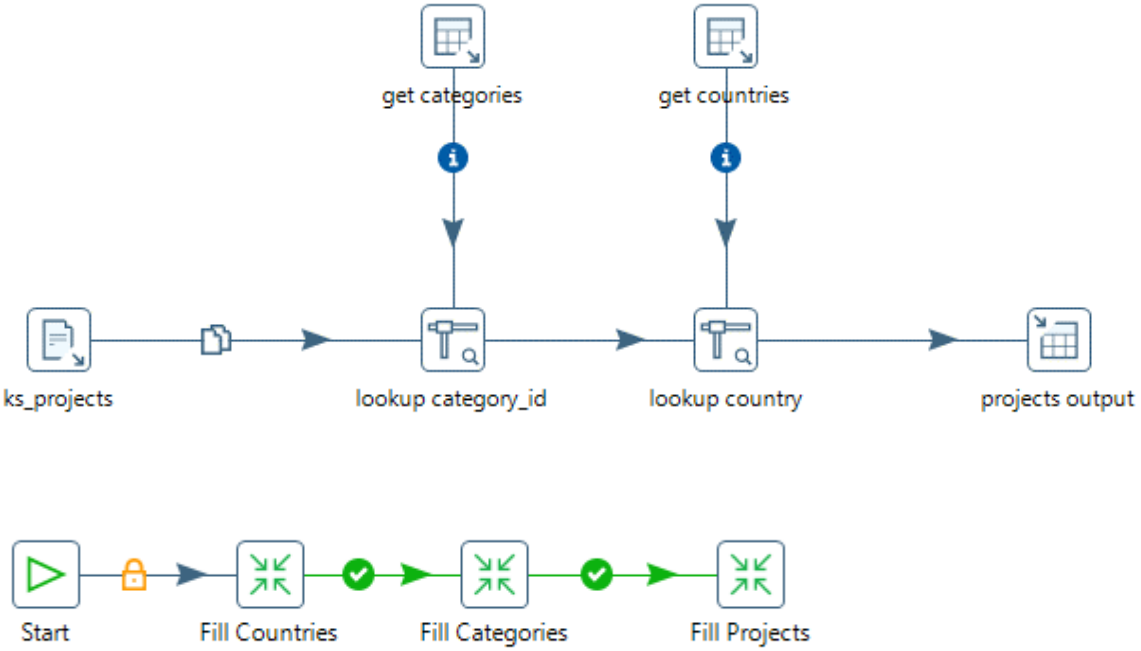
Pentaho transformations:

1. A table was created to insert all the countries of the projects.
   From the dataset, the country code and the currency used were selected, all duplicates were removed, and for each country, its full name and reference continent were added, through a mapping with additional *CSV* files.

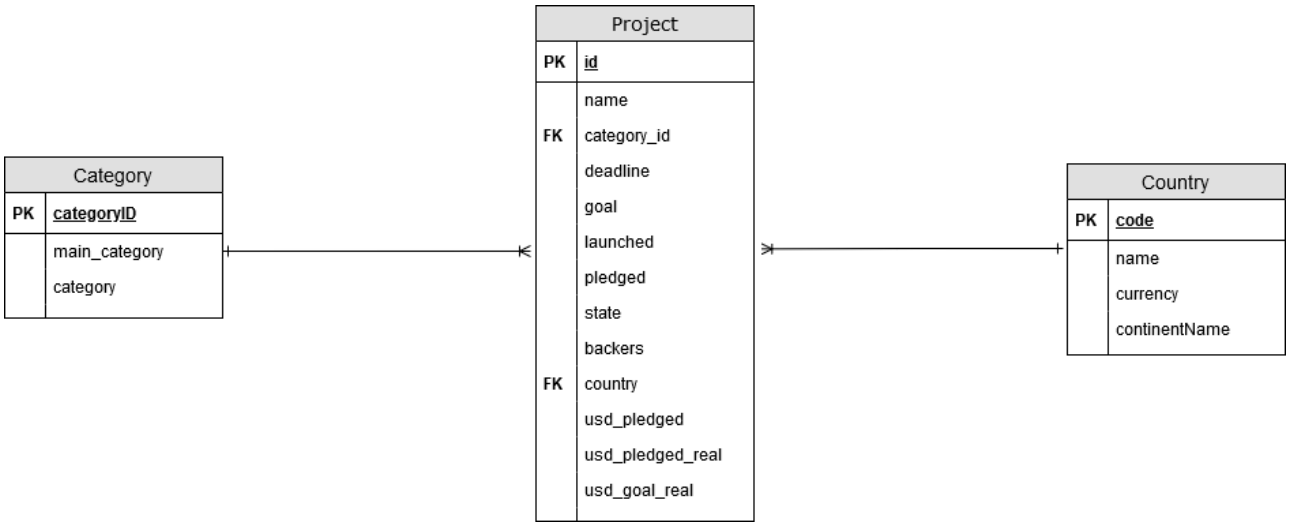


2. A table was created to insert all the combinations of "*category*" and "*main category*".
   The two attributes were selected from the dataset, duplicates were removed, and an incremental ID was inserted.
   The table has an incremental ID because it was observed that the "*category - main category*" correspondence is, in some cases, many-to-many rather than one-to-many as one might intuitively think.
   (e.g., the Anthologies category is present in both Comics and Publishing)

3. A third table was created to insert all the projects.

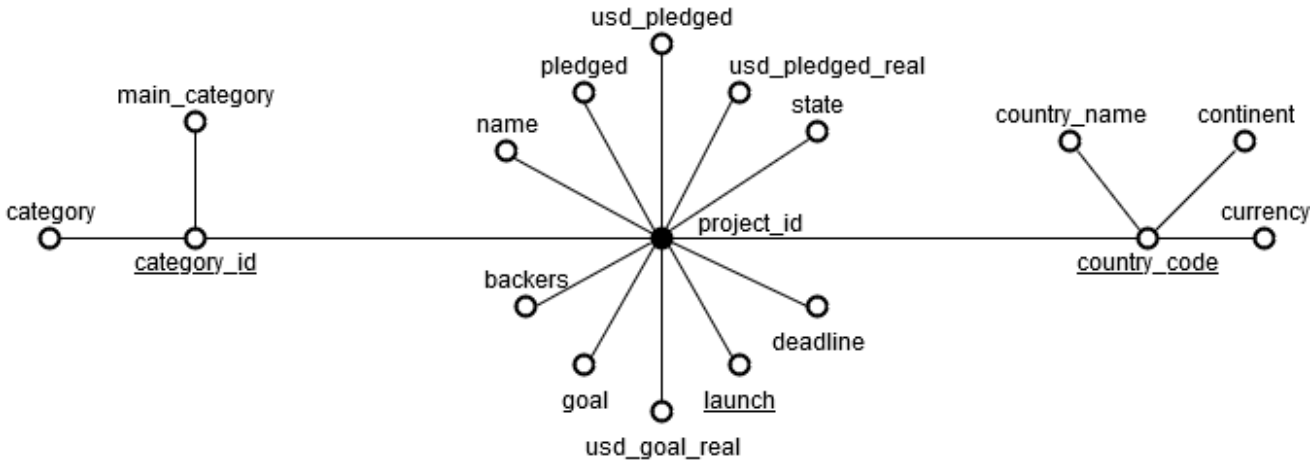


Therefore, the generated database is the following:



Conceptual Design

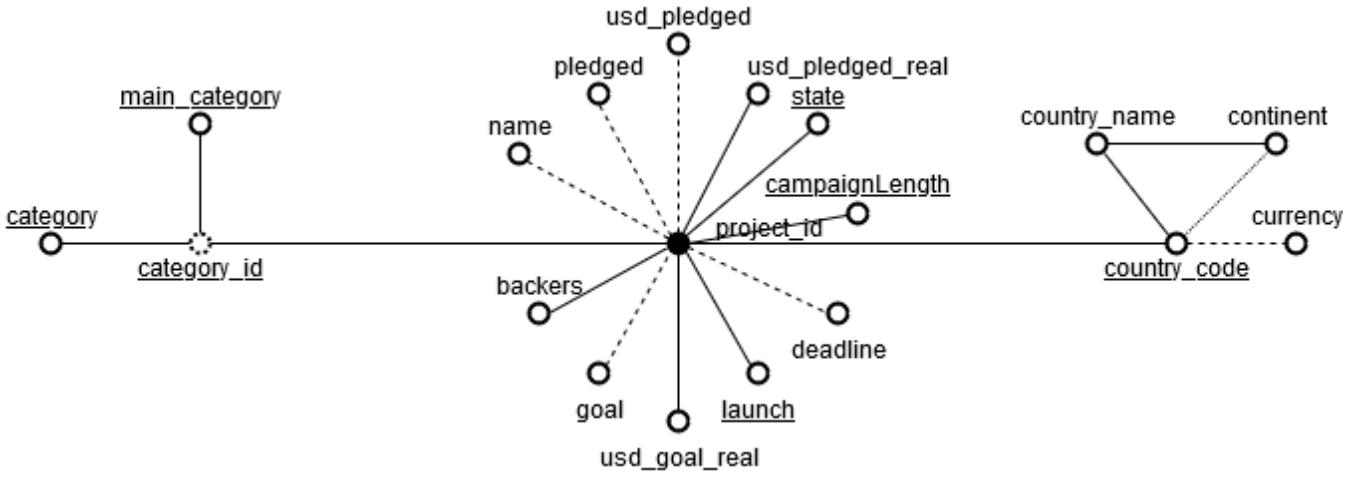
Starting from the previous E-R schema, it was possible to define the attribute tree, as depicted below:



To the above attribute tree, some modifications were made:

1. **Removed** "*name*", "*pledged*", "*usd_pledged*", "*deadline*", "*goal*", and "*currency*" as redundant or less useful fields.
2. Made "*category*" and "*main_category*" **direct children** of the root, in order to use these attributes as dimensions.
3. Changed the connection to "*continent*" **from direct to indirect** (through "*country_name*").
4. Calculated the "*campaign length*" field, **discrete at intervals** (Short, Medium, or Long).

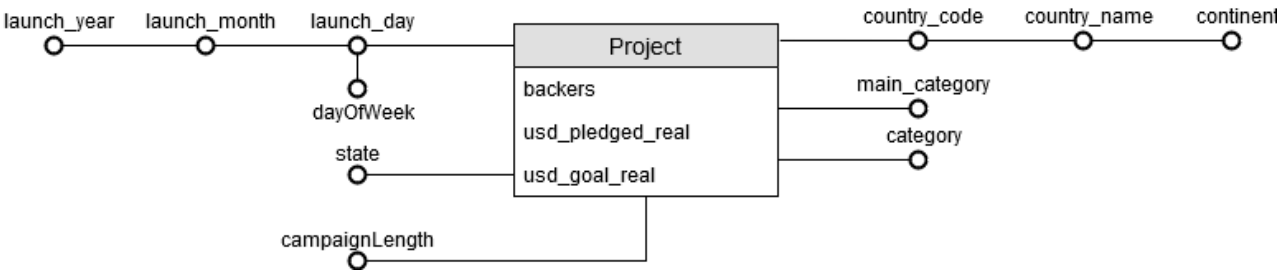
The result is the following:



The chosen **dimensions**, therefore, were:

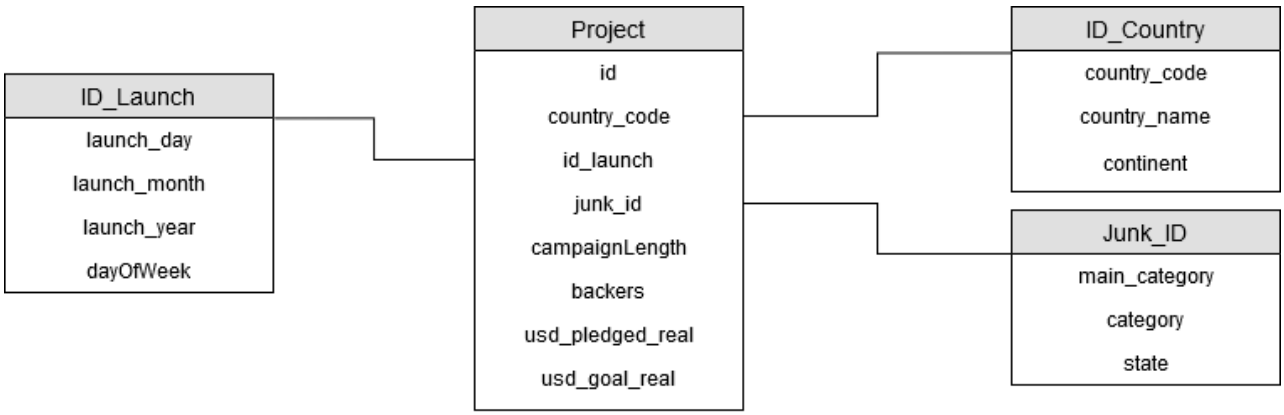
- main_category
- category
- state
- launch
- country_code
- campaignLength

Based on the attribute tree, the Dimensional Fact Model (**DFM**) was defined, to which the attribute "*dayOfWeek*" (indicating the day of the week the project was launched) was added:



Logical Design

Subsequently, it was possible to determine the schema of the datamart using a "Star Schema", as depicted below:



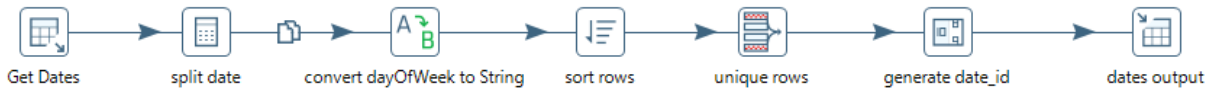
The degenerate dimensions were inserted into a single *junk table*, so that queries will be more efficient since various strings will be replaced by an integer that identifies the combination.

For the definition and filling of the database, Pentaho was used again, through the following transformations:

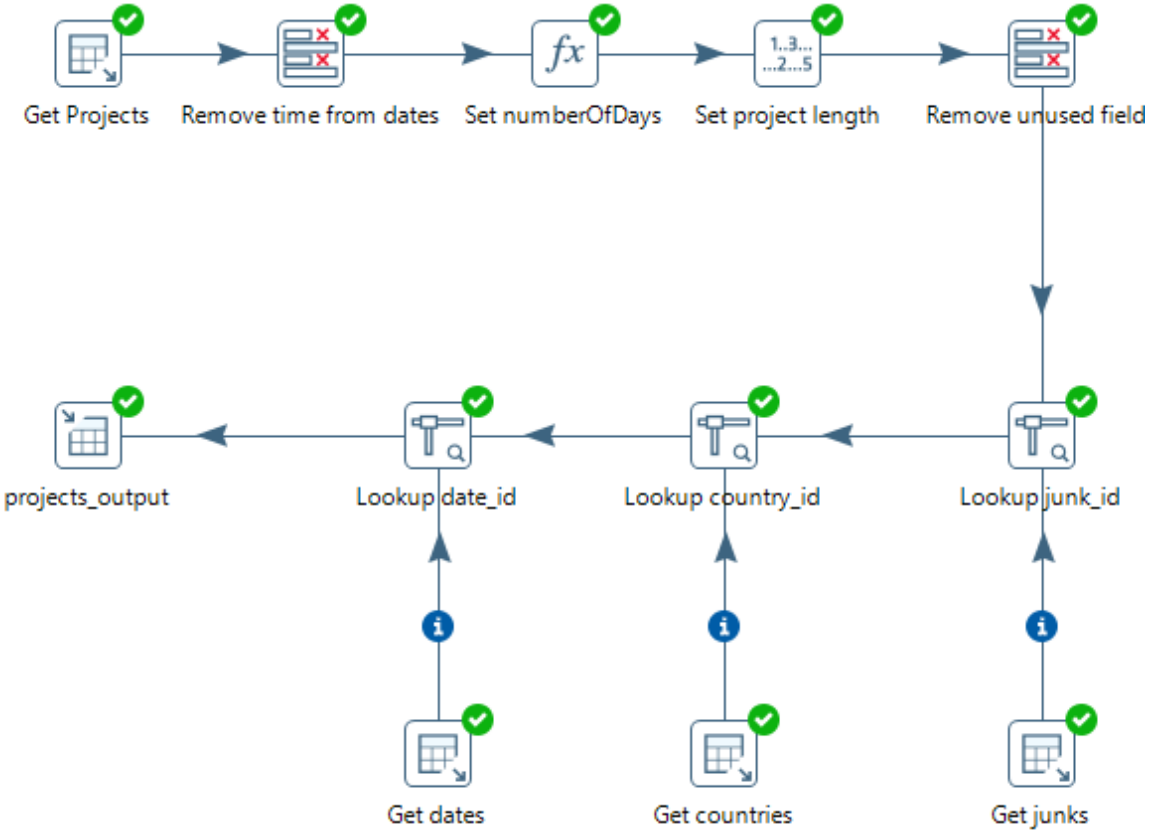
1. A table containing the countries was created, removing the currency and generating an incremental ID.

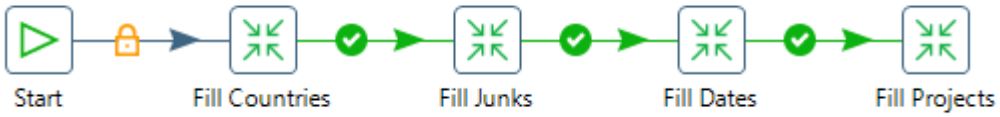
2. The *junk table* was created by taking all the degenerate dimensions from the dataset and generating an incremental ID.

3. The table of launch dates was created, calculating the day of the week and generating an incremental ID.



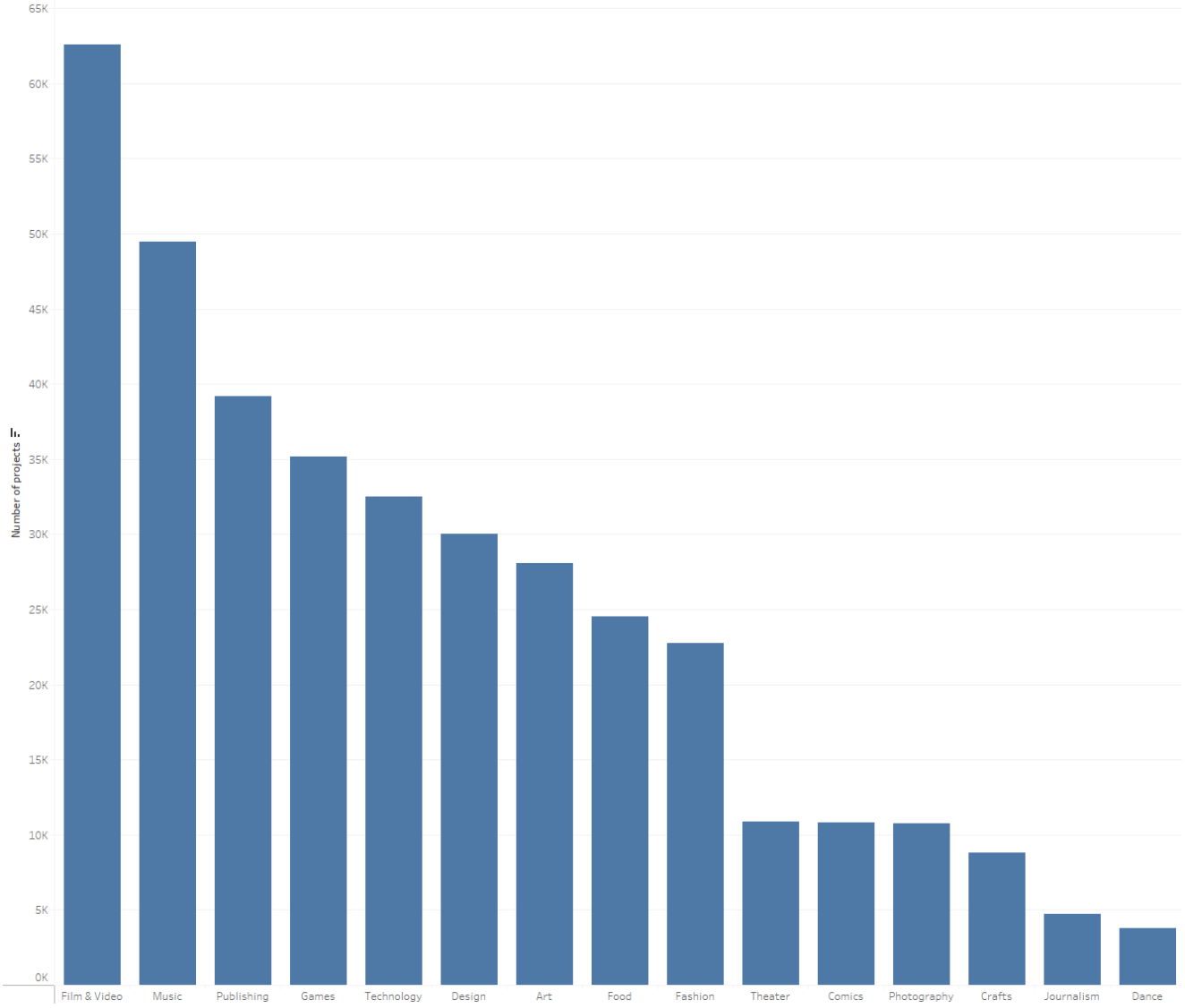
4. The table to insert all the projects was created.





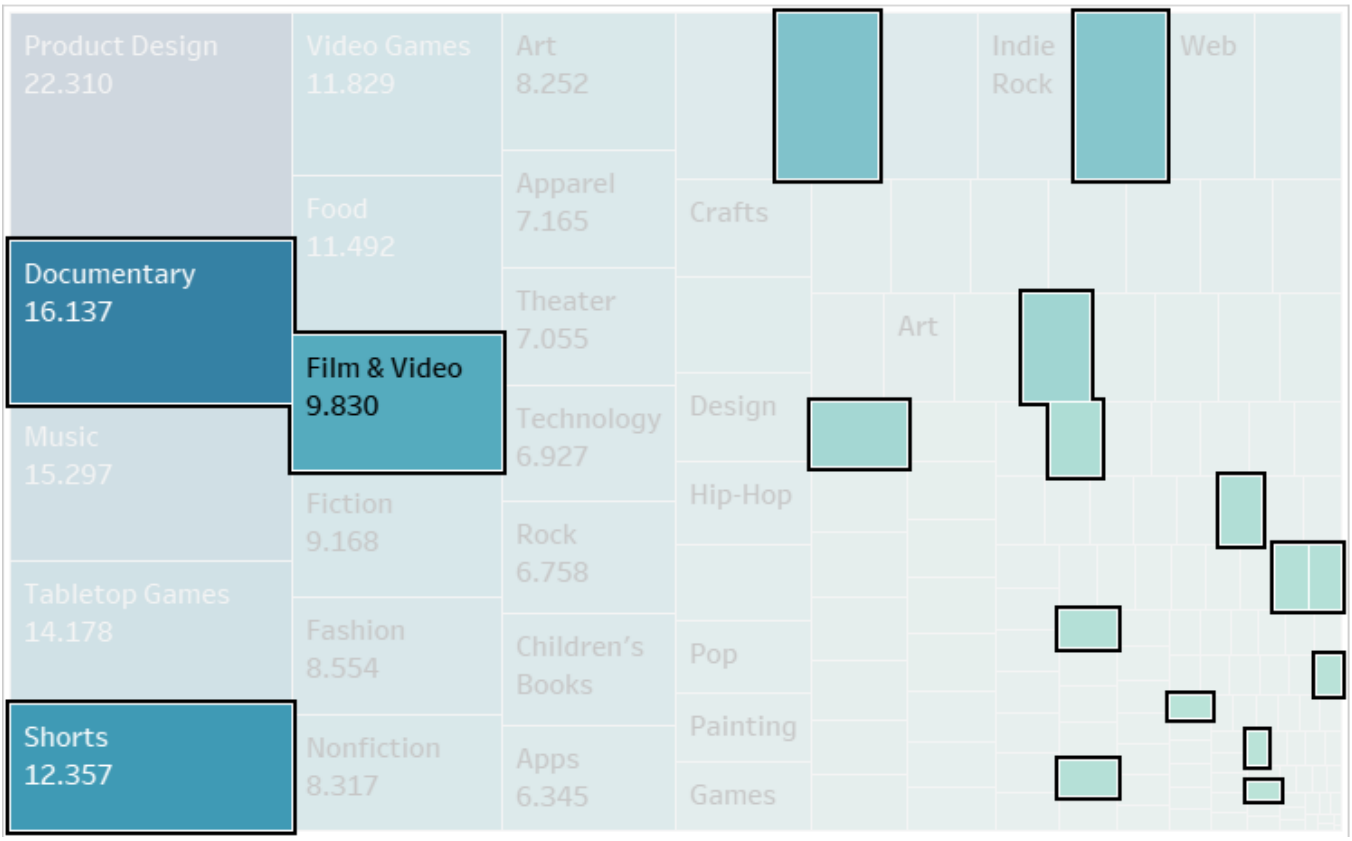
Analysis

Most Popular Categories

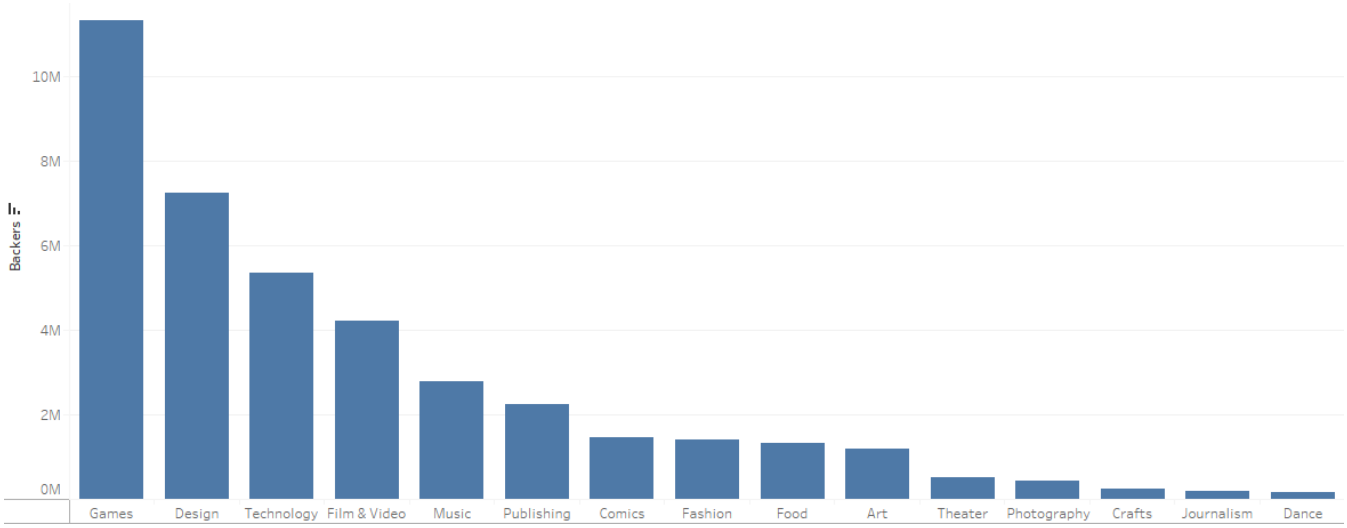


As can be seen, the main category for which Kickstarter is mainly used is "*Film & Video*", followed by "*Music*" and "*Publishing*."

In particular, the most popular subcategory in Film & Video is documentaries, followed by shorts.

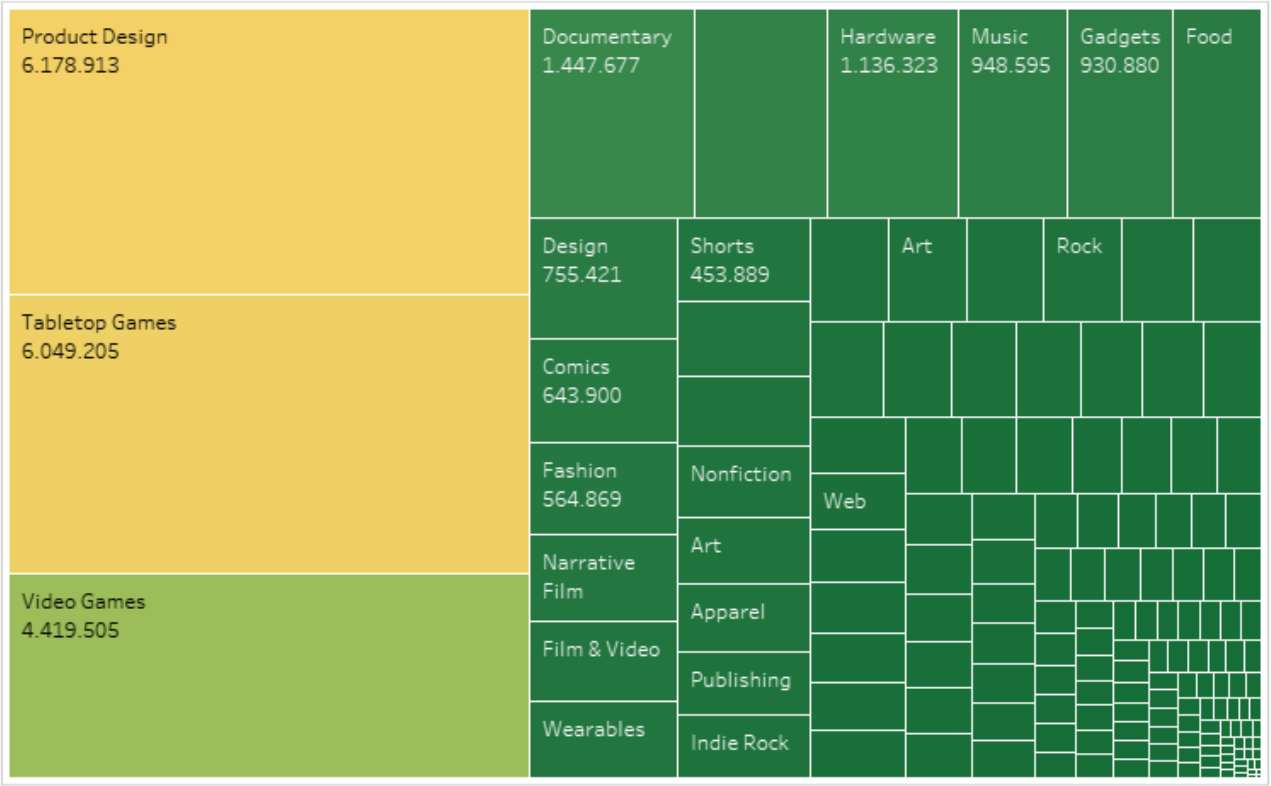


Most backed categories



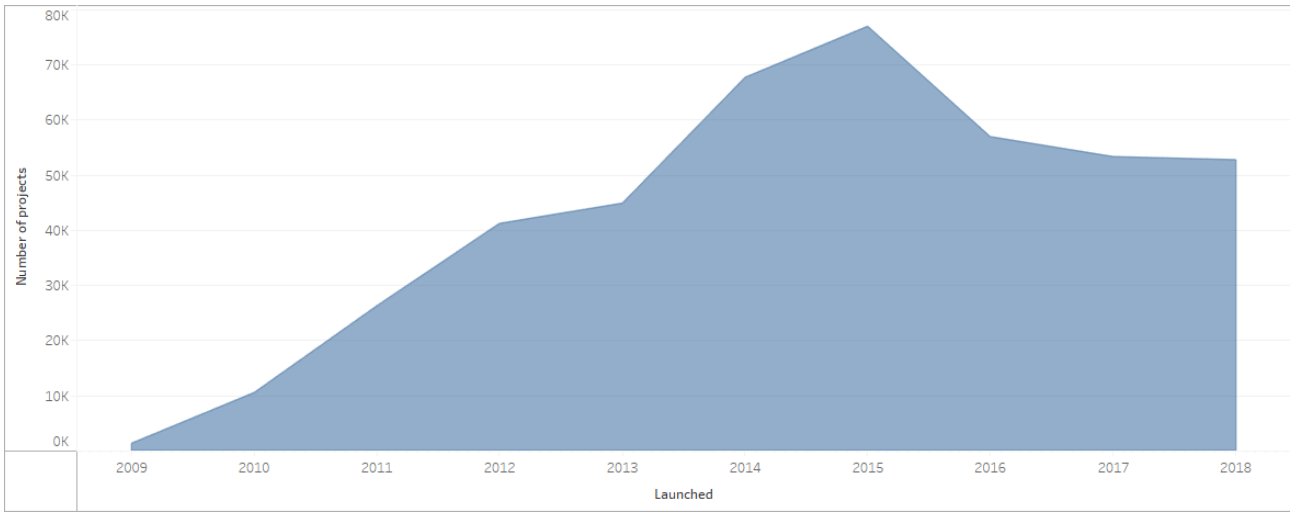
Although the most popular category is "*Film & Video*", the one with the most backers is the "*Games*" category with **11,336,829 backers**, followed by "*Design*" with a difference of about 4 million, while "*Film & Video*" is in fourth place with about one-third of the supporters of the first category.

We can therefore assume that, for the average Kickstarter user, projects related to gaming are preferable.



However, "*Games*" category is fragmented, divided into about **60% board games** and **40% video games**, while the "*Product Design*" subcategory (part of the "*Design*" category) alone manages to surpass all the other subcategories.

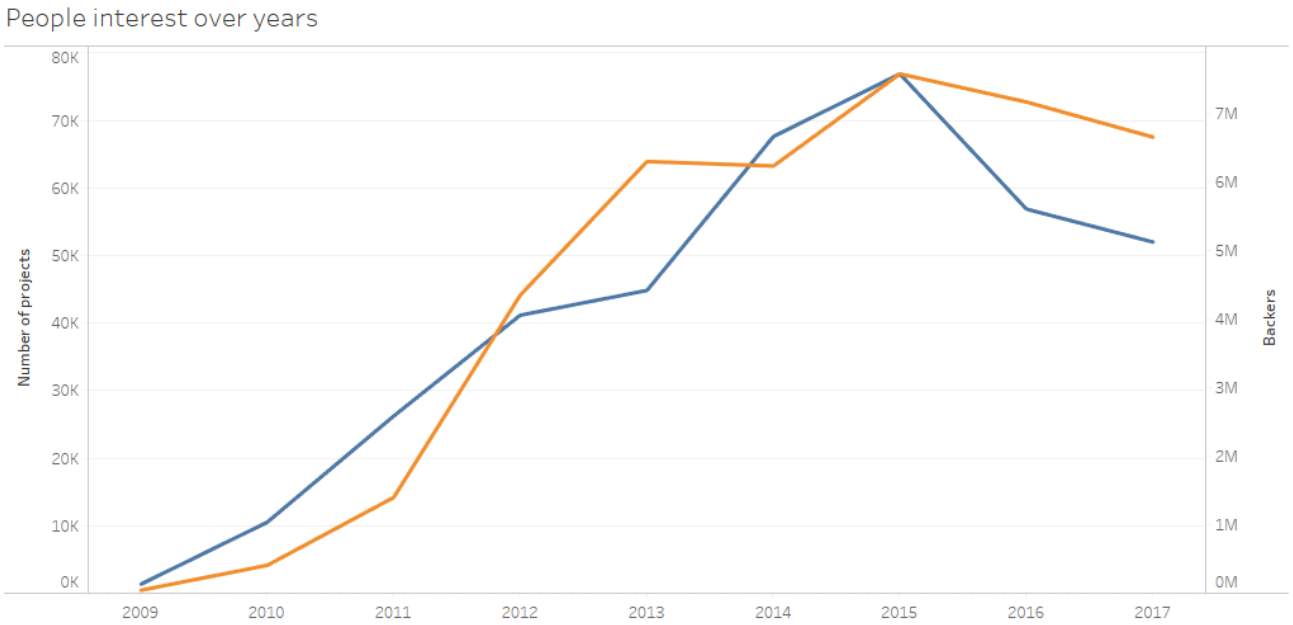
Annual Project Numbers



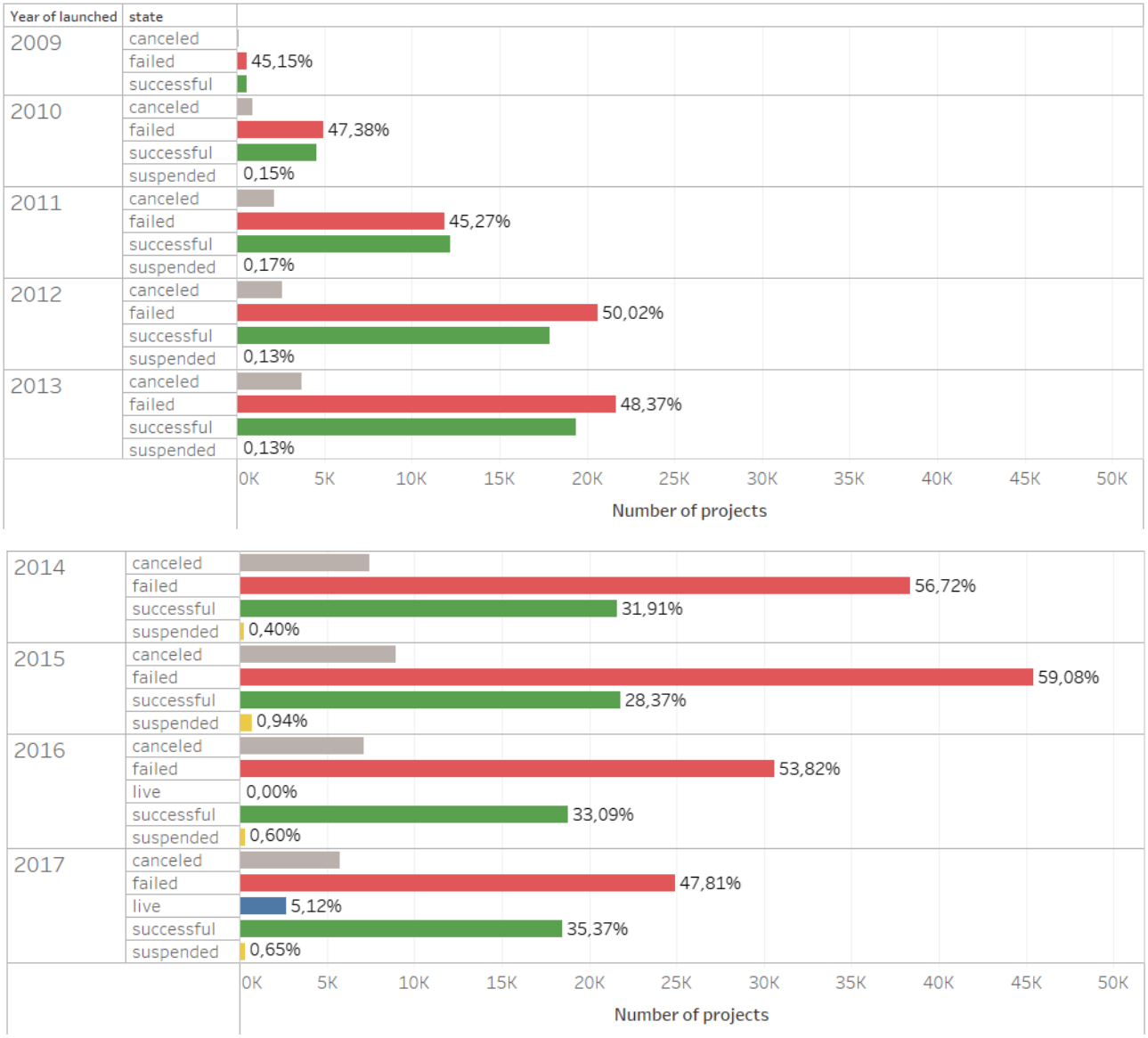
As can be easily seen from this graph, the number of projects launched on Kickstarter has been steadily increasing from 2009 until 2015, the year it reached its peak.

Subsequently, there was a sharp decline, maybe because the crowdfunding system started to be abused over the years.

The last year (**2018**) is a forecast calculated with Tableau.

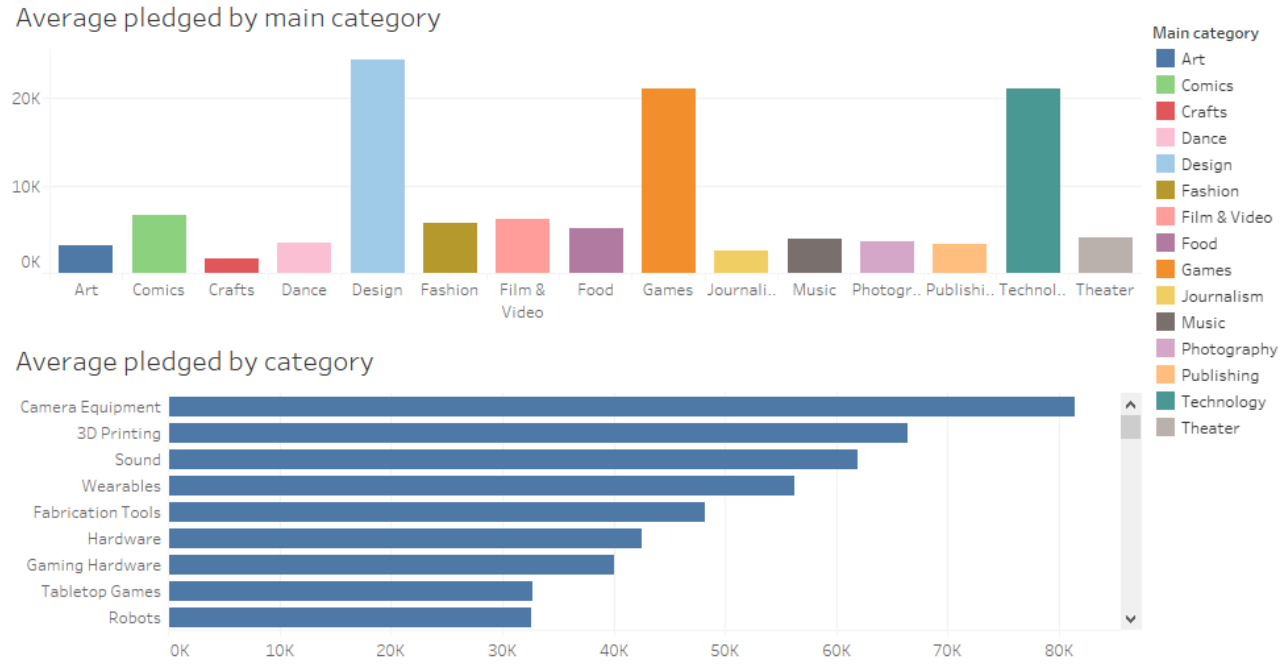


The trend of the number of backers is quite similar to the trend of the number of projects on the platform, with almost constant growth until 2015.

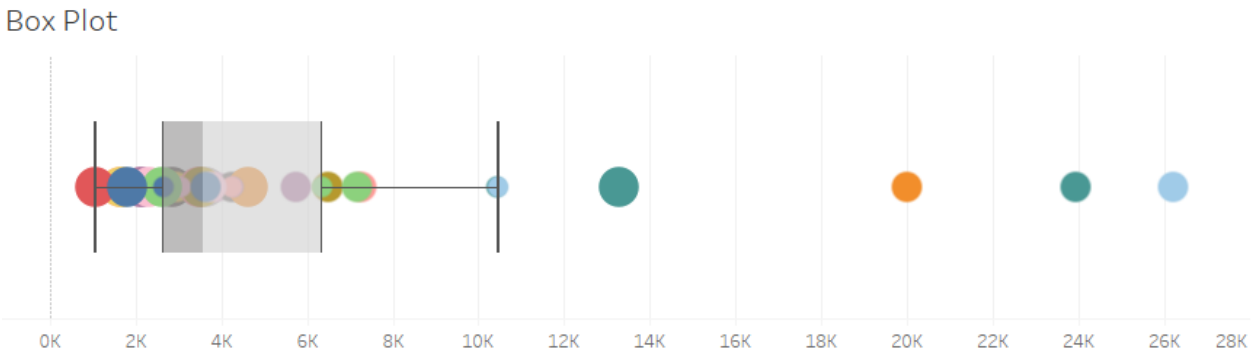


It can be noted that, over the years, the number of failed projects exceeds (sometimes widely) the number of successful projects, **except** for the year 2011, in which there were 12,171 successes against 11,878 failures.

Average Investment



According to the data, *Design*, *Games*, and *Technology* are the three categories that can aim for an average raising of **$20,000**, while the others barely exceed $5,000. This can be useful for determining the target to aim for your own project.

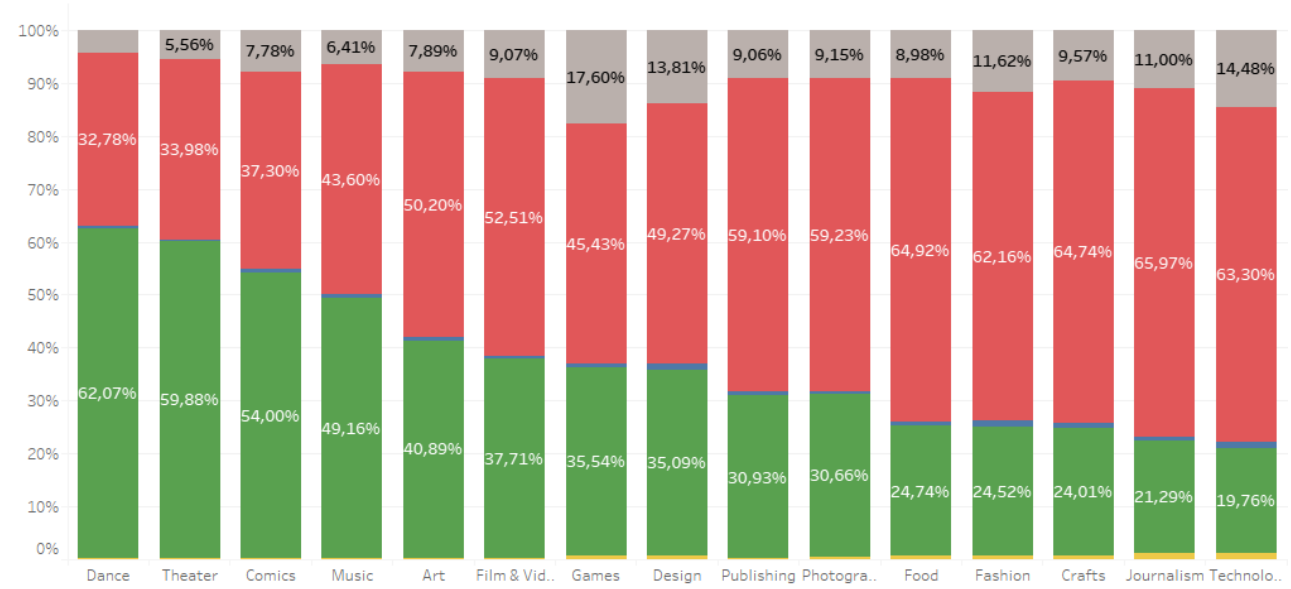


Things change slightly if you consider the duration of a project. In this case, a **medium-duration** Design project (between 30 and 59 days of fundraising) averages $26,000.

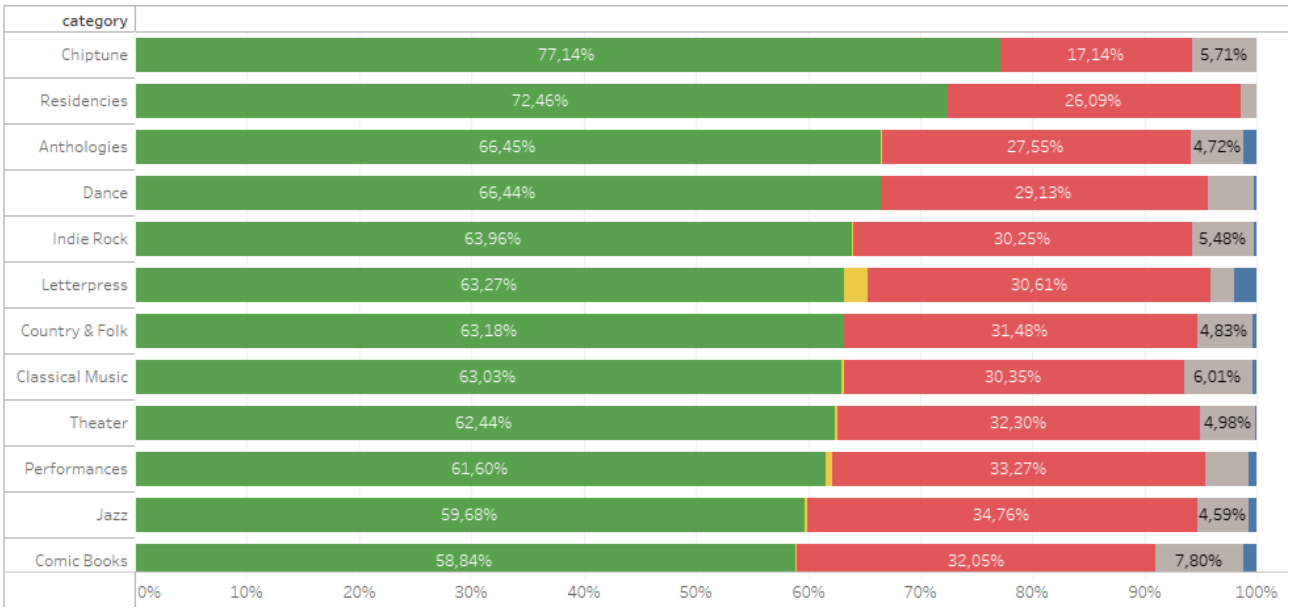
Contrary to what one might think, usually **long-term** projects (60 days or more) seem to raise much **less** than others in the same category.

Success Rates

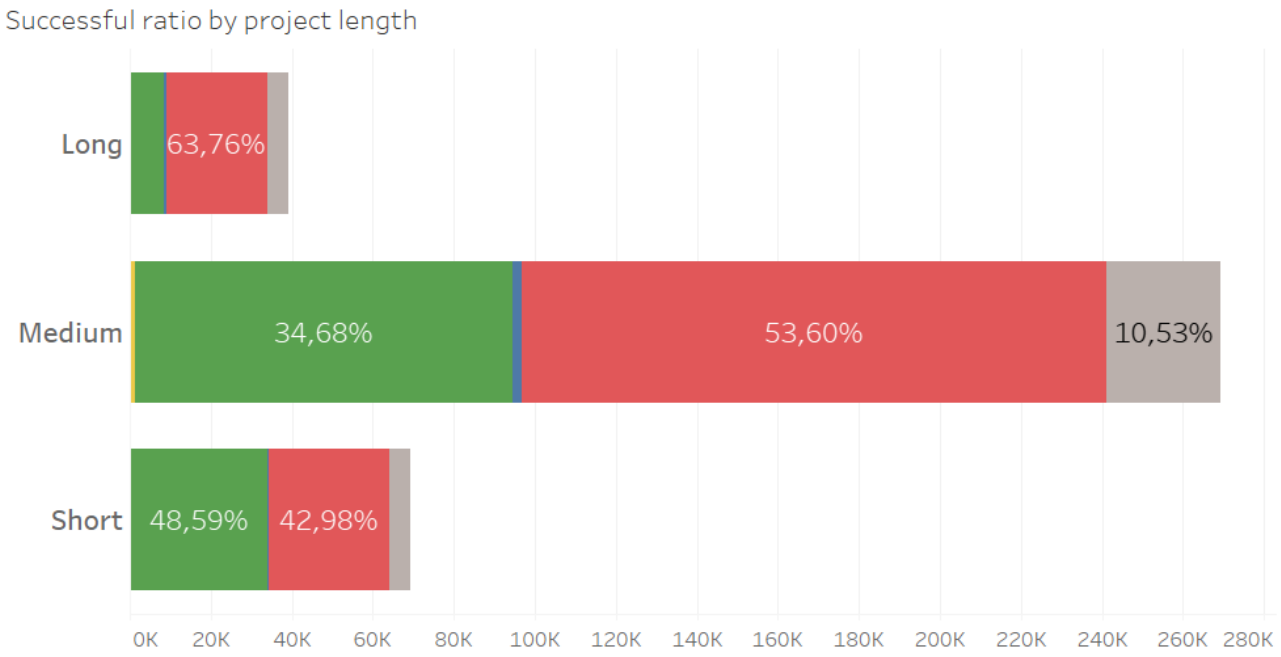
Successful ratio by main category



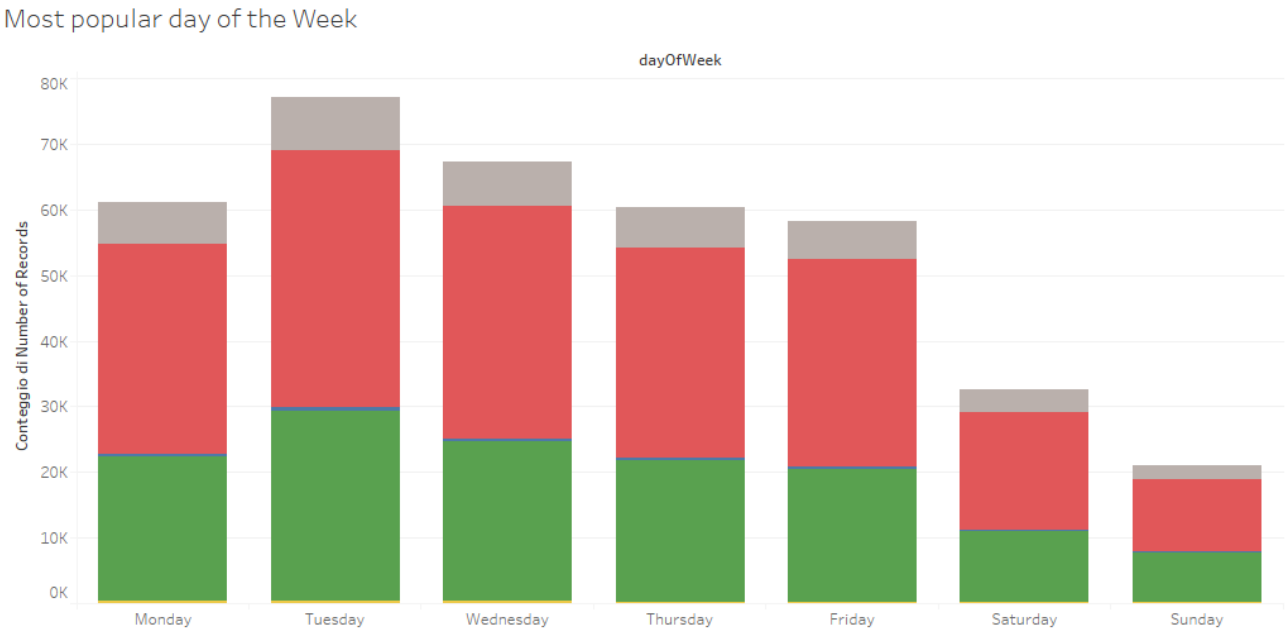
Successful ratio by category



The category with the most successes overall is *Dance*, followed by *Theater and Comics*. Maybe because they are categories that, although less supported than others, require on average less investment.

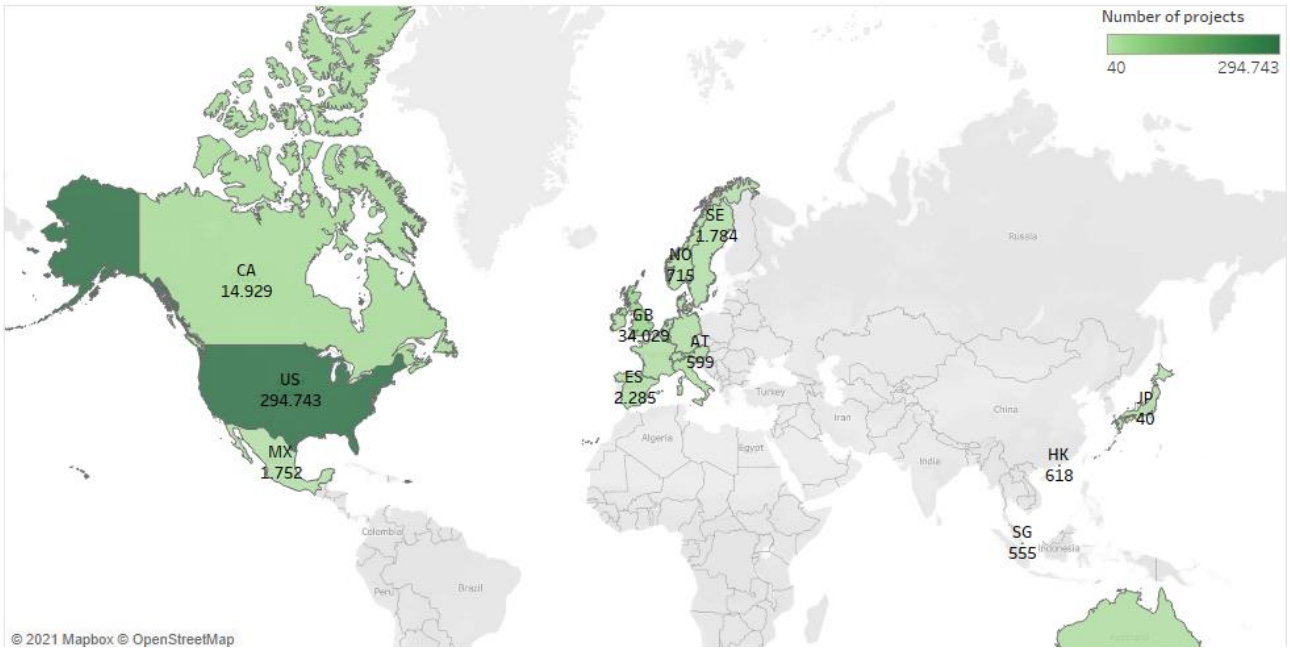


Finally, most projects (about 270,000) have an **average** length, but **more than half** fail or are canceled, despite raising more money. **Short** projects have the **highest** success rate, while about **eight out of ten long projects** never see the light of day.



World Distribution

Where is Kickstarter most used?



Projects distribution by State

	state					
Continent	canceled	failed	live	successful	suspended	Grand Tot..
Asia	162	553	72	401	25	1.213
Europe	6.709	30.639	666	17.387	374	55.775
North Amer..	30.665	162.411	1.979	115.024	1.345	311.424
Oceania	1.216	5.491	81	2.507	99	9.394
Unknown	12	22	1			35

Finally, as can be easily seen, the distribution of projects in the world is **extremely** in favor of North America, where **82%** of Kickstarter projects reside. This could be due to both a higher number of users in the American territory and any legislative regulations governing crowdfunding in various countries.

Successful ratio by State

	state				
Continent	canceled	failed	live	successful	suspended
Asia	13,36%	45,59%	5,94%	33,06%	2,06%
Europe	12,03%	54,93%	1,19%	31,17%	0,67%
North America	9,85%	52,15%	0,64%	36,93%	0,43%
Oceania	12,94%	58,45%	0,86%	26,69%	1,05%