Importing Data:

As the dataset is a .csv file, used pandas pd.read_csv to import data into a variable called kickstarter_raw with parse_date parameter set to 0. The result is a pandas Dataframe.

Initial Data Assessment:

Used the attribute of file .info() to find the information about the data: there are total 15 columns, 5 columns has float data type, 1 column has int data type, 9 columns has object data type. 13 of 15 columns has 378661 non-null entries, the remaining two columns has less than this optimal number of entries is an indication of missing values in those two columns (**name** and **usd pledged**). Used the attribute of file . head(10) to check the first 10 items in data frame. Noticed that column **deadline** and **launched** should be date and time type. Columns **category**, **main_category**, **currency**, **state** and **country** are all category data type.

Data Cleaning:

Drop duplicates: used drop_duplicate attribute of the file to drop duplicated rows in data frame.

Convert data into their appropriate data format: convert **deadline** and **launched** into datetime data type via pd.to_datetime. Convert **category**, **main_category**, **currency**, **state** and **country** all into category data type for faster processing time. Used .astype() attribute of file to execute the conversion. Rerun the .info() confirmed that the following data type shows up in dataframe: 5 category, 2 datetime, 5 float, 1 int, 2 object, memory reduced to 34.0 MB from 43.3 MB.

Count unique values of country column: used value_counts attribute of country column to show number of projects for each country. The top three countries are US, UK, and Canada. The result table contained an abnormal country name = N0” with 3,797 entries.

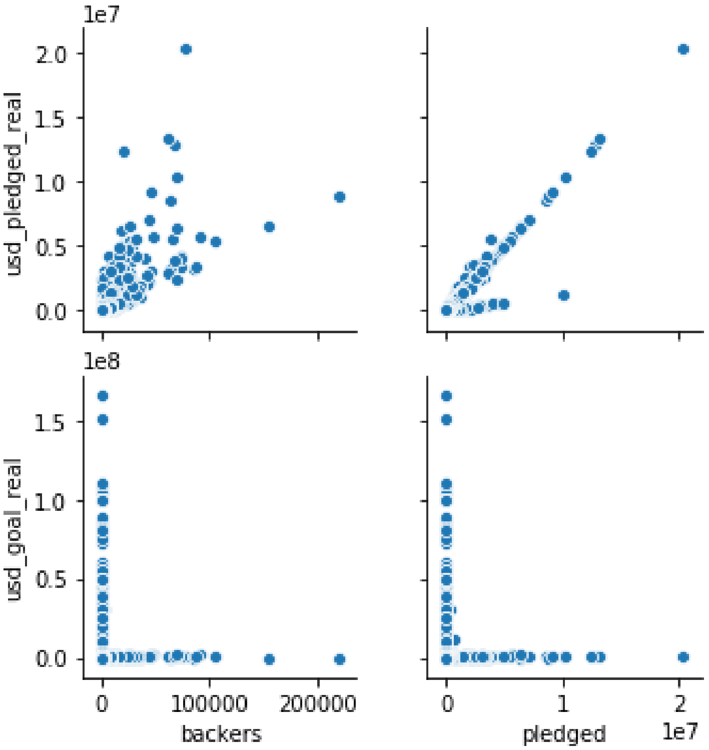
Extracting rows with this abnormal country name: from the resulting table, it shows rows with country = N0” all have 0 **backers** even most of them has an pledged amount. In addition to 0 backers, the **state** column of those rows are all undefined. From the first 10 rows in raw data frame, we cannot determine the **state** just from the fact that the pledged amount equal or greater than the goal. Noticed that some project still got cancelled even the pledged amount exceeded the goal. Since our project goal is to find out the factors that impact whether the project is successful or not, then the rows with **state = undefined** should be eliminated. There are 3,797 rows which is only 1% of the data set and should be safe to exclude them.

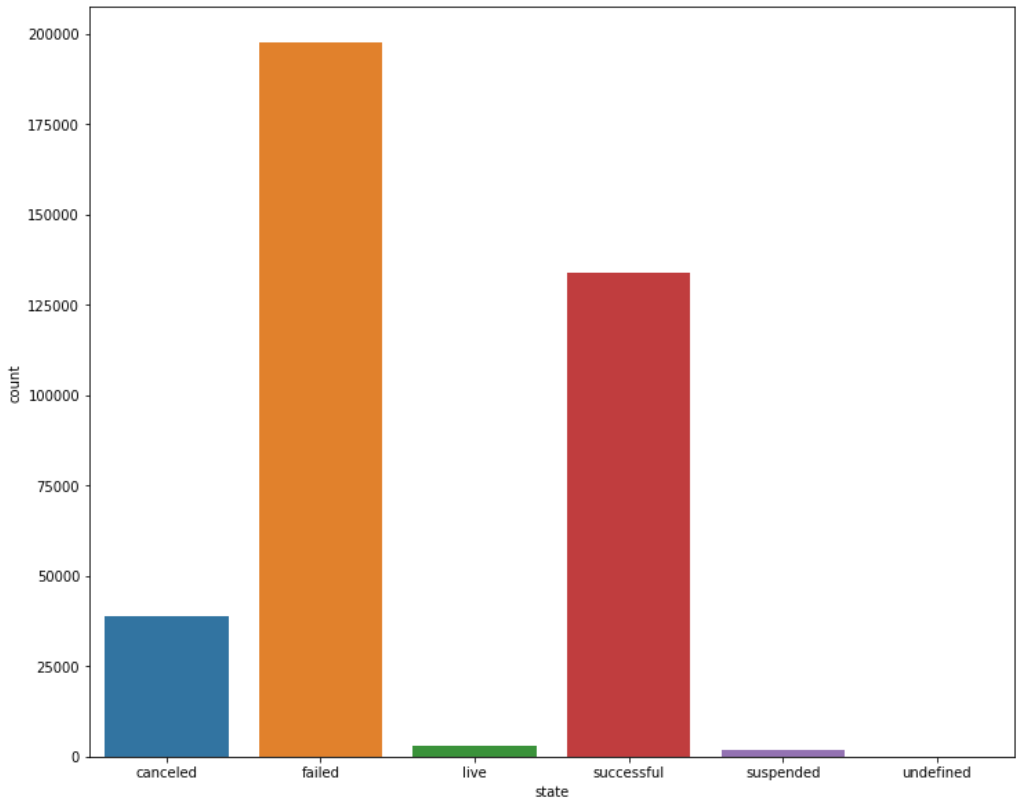
Drop rows with NaN values: used .dropna() attribute of the file to drop those 3,797 rows and assign the resulting dataframe to a new variable called: kickstarter_clean to indicate it is cleaned dataset. Then apply .isna().sum() to check whether there are more row with missing values. The results confirmed that kickstarter_clean dataframe does not have any missing values in all its columns.

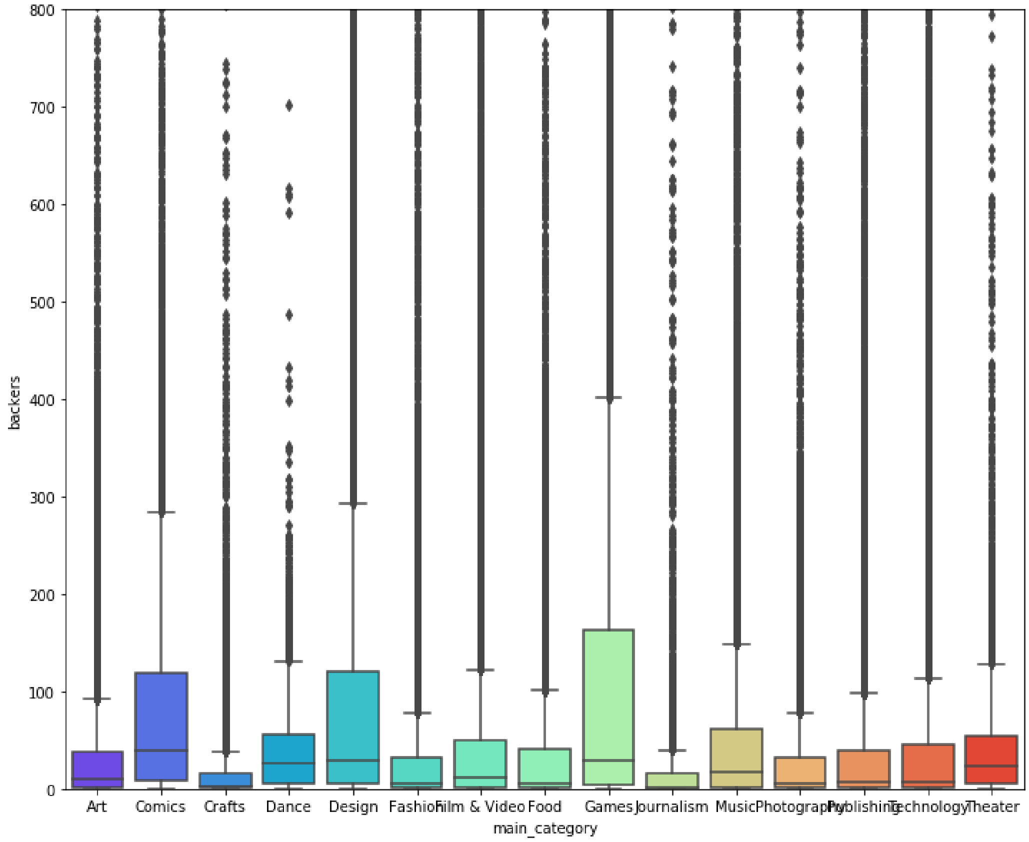
Getting to know more about the cleaned dataset: used value_counts on **main_category** find out the top 3 main categories are: Film & Video, Music and Publishing. Used value_counts on **state** find out the top three states are: failed, successful, and cancelled.

Identify outliers: applied .describe() to kickstarter_clean to get the statistical summary of the numerical data columns. Three main columns are **goal**, **pledged**, and **backers**. Since goal and pledged are the money amount, some of project needs more fund than other projects, it is possible to have a big amount in both columns. Like the money amount, number of backers could also be huge as more people wanted to pledge for certain projects. Or some less popular project only has few number of backers. None of those number should be considered as outliers. Will leave the those number as it is.

Initial data visualizations: perform several different kinds of data visualizations with matplotlib.pyplot, seaborn. Selected results are shown below.




200000
175000
150000
125000
100000
75000
50000
25000
0
count
canceled
failed
live
successful
suspended
undefined
state


800
700
600
500
400
300
200
100
0
backers
Art
Comics
Crafts
Dance
Design
Fashion
Film & Video
Food
Games
Journalism
Music
Photography
Publishing
Technology
Theater
main_category