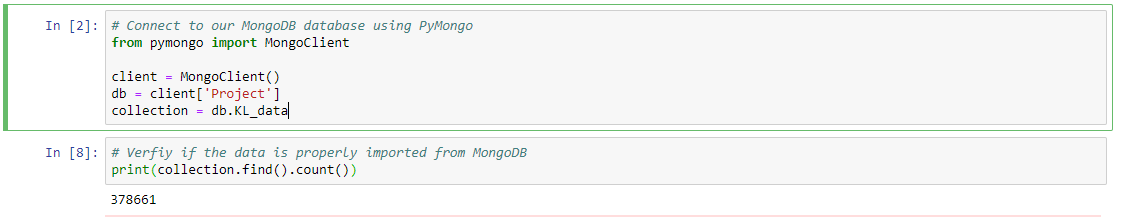
Data Analysis using PyMongo:

Methodology:

We use the 2018 Kickstarter data for analysis and further predictions. The data we receive is in the CSV format, we load this data into our local MongoDB server using the mongoimport utiltity. The following command helps import our data into our local MongoDB server.

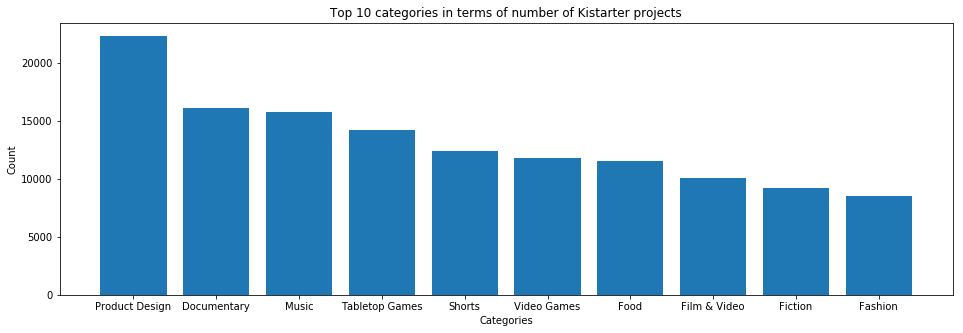
**mongoimport -d Project -c KL\_data --type CSV --file ks-projects.csv –headerline**

After loading our data into MongoDB we use Jupyter notebooks which is a Python based interface having coding and markup functionalities in the form of cells. The data is loaded in the database “Project” and the collection is “KL\_data”. We use the PyMongo library to fetch/query data from our local MongoDB server and get it in our Jupyter notebook. We first create a MongoClient object and then by default connect to the default local port of MongoDB. We then get our collection and then query our data in mongodb. Example code is as below.

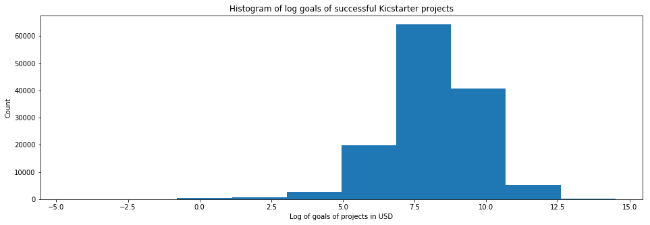


The data which we receive as response from the database after running the query is in JSON format. We convert this data into a pandas dataframe using the from\_records() function which is a part of the pandas library which we use for data manipulation. After getting the results in the form a pandas dataframe we use matplotlib to plot the results. Following are the analytics performed by us.

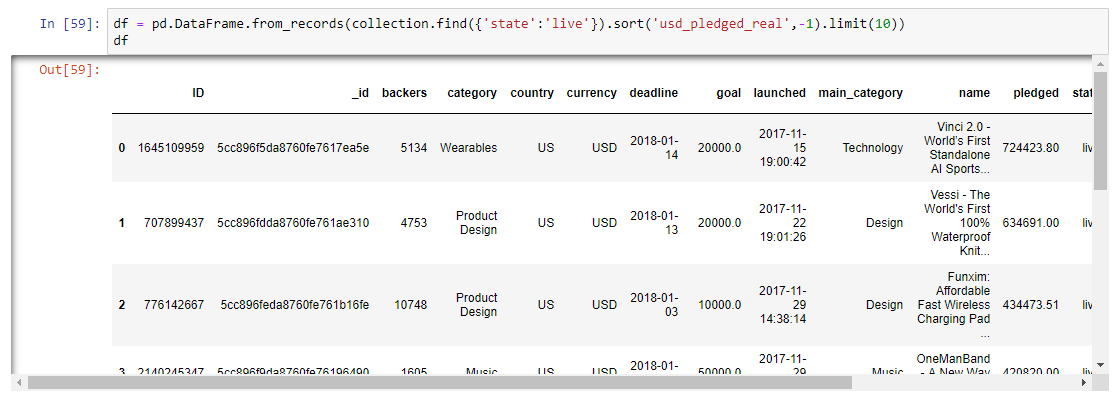
1. **Number of projects in the top 10 popular categories**

****

1. **Distribution of the target goals of successful projects after taking log transform as the target goals distribution is very right skewed**

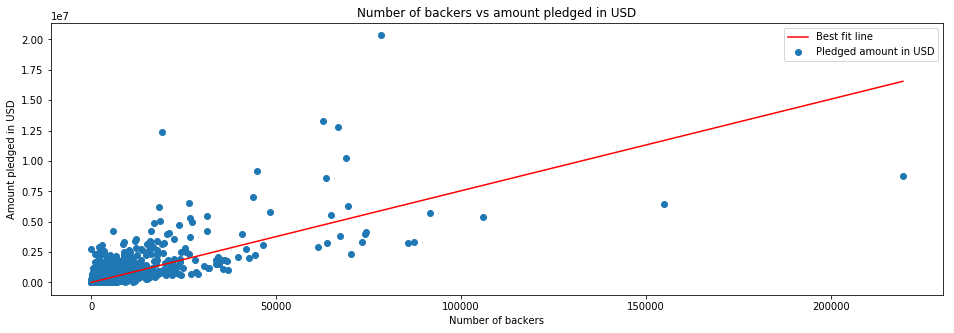


1. **Get a list of the top 10 projects which were still live when the data was recorded and had the highest amount pledged by the community**

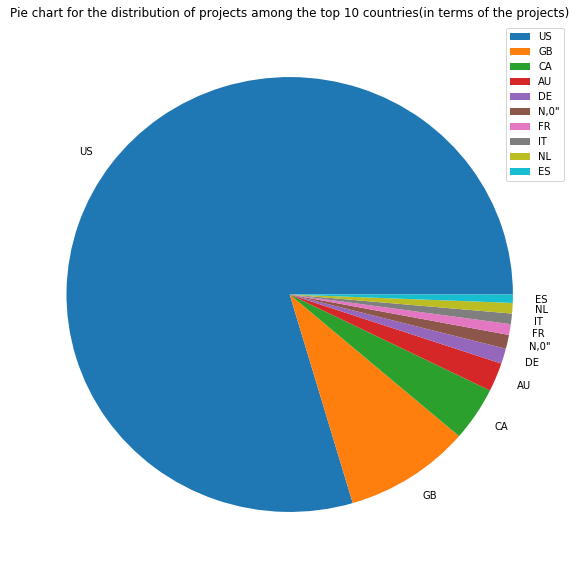


1. **Get the relationship between the number of backers and the amount pledged in USD**

We can see that the relationship has linear characteristics but also there are many outliers far from the line of best fit.

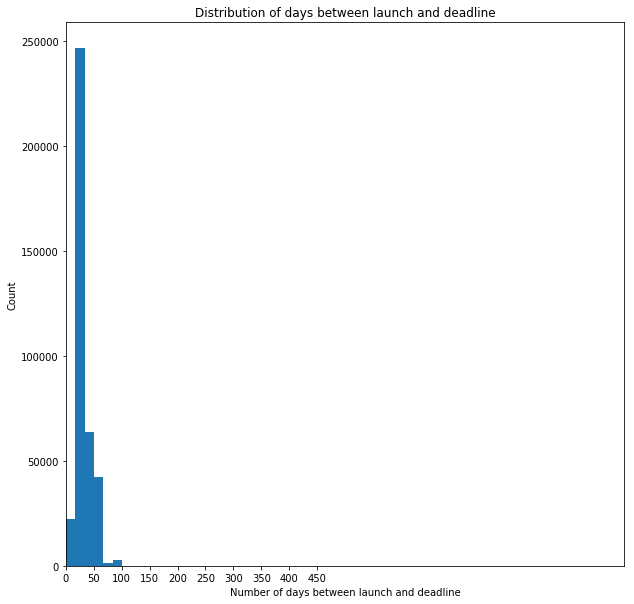


1. **Pie chart of the number of projects from the top 10 countries**

****

We can see that the majority of projects are from the US with GBR in second place, the number of projects from the other countries are much smaller in proportion.

1. **Distribution of the time between launch and deadline of projects**

****

We can see that almost all projects have about 30 to 40 days from launch to deadline. Although this distribution too is very right skewed.

**Implementing Map-Reduce in PyMongo:**

Map reduce in MongoDB needs a mapper and a reducer results. We can define these functions directly in PyMongo as they cannot be treated as python objects/functions. So to do this we use the BSON library and send the function as codes.

