Mark Levine

UCI Data Analytics Bootcamp

July 26, 2018

# Unit 1 | Assignment - KickStart My Chart

## What are three conclusions we can make about Kickstarter campaigns given the provided

## data?

1. From a pure volume perspective, theatre (1,393 projects) , music (700) , technology (600) and film & video (520) are the most popular ventures with theatre’s volume considerably higher than the rest. Further, music-based projects yielded the highest success rate (77.1%) calculated by dividing successful projects by total project counts; theater (60.2%) and film & video (57.7%) were the second and third most successful project categories, respectively. On the other side of the coin, the least successful as measured by failed % included food (70.0%), games (63.6%), and publishing (53.6%).
2. Geographically, the United States (3,038 projects comprising 73.9% of total) was the most popular country likely due to the fact KickStarter is a US founded business entity with significant resources and advertising in the US; Great Britain (604 projects at 14.7% of total) and Canada (146 projects at 3.6% of total) came in second and third, respectively. Theater and music were the most popular in the US, GB and Canada. Music had the highest success rate in all three of the US, Great Britain and Canada.
3. The average donation amount from backer’s is considerably skewed as 89.8% of the average donation amounts were less than $155;

## What are some of the limitations of this dataset?

1. Considerable skewness as it relates to geolocation: the US (73.9% of all projects) dominates the data set which is likely due KickStarter being a U.S. based entity and the fact the “starving artist” persona is most popular in the United States compared with the rest of the world.
2. The data set lacks information about the backers: what industry are the backers from? What are the reasons the backers are investing? What is the distribution of the investments from the backers as opposed to a simple average (i.e. is it one backer giving 90% or even split 2-3% from each backer?)
3. The data set lacks a key which describes in detail each field’s purpose and source
4. While the blurb field may have useful information, it is highly inconsistent. However, it may be used for sentiment analysis (good, bad, neutral) and other Natural Language Process analyses.
5. The data set could be improved by adding demographic fields for the project leaders (age, background/industry, how many projects that individual has made to date, etc.)

What are some other possible tables/graphs that we could create?

1. Histogram of the average investment per Backer:
2. Geographic distribution of country and status/success rates
3. Histogram of the pledged amount per project: reveals significant right skew (left heavy) whereby the most frequent investment pledge amounts were ~1-$2,000. However, the average pledge amount was $11,223 which differs significantly from the median of $1,562. The maximum pledge amount, quite the outlier, was $2.3 million.