**Title**: “Landscape of Funding in Kickstarter Projects”

**Authors**: Ibrahim Taher, Anant Jain, Sarthak Kothari, Forrest Hooton

**Summary**:

We propose the exploration and analysis of Kickstarter Campaigns using a combination of scraped datasets. Kickstarter is a popular online platform for crowdsourced funding. People post about their new ideas or products, and others browse through and have the opportunity to help fund projects they find interesting. Each project has a funding window that a user sets (maximum of 60 days), and this funding window is called a “Kickstarter Campaign.”

The Kickstarter dataset comes from a set of scraped json files, and contains around 200,000 rows and 109 columns. The columns provide general information about the specific campaign of a kickstarter project. For instance, if the same project had two campaigns (two funding windows), then each campaign would have a different row with the specifics for that campaign. Examples of column names are:

"ID" "name" "category" "currency" "deadline" "goal"

Our goal is to better understand the landscape of funding for Kickstarter campaigns. We will begin with questions, and use those inquiries to guide our data exploration. Examples of questions we’re asking are:

* What effect does a campaign's category have on its funding?
* Are there noticeable trends for projects that are fully funded?
* How do different categories affect success across different regions of the world?

**Proposed plan of research**:

We will focus on merging, cleaning, and an exploratory analysis in the first stage of our project. As stated previously, the data we are using exists across multiple json files that are webscrapings from kickstarter and will need to be merged using the jsonlite package. Furthermore, once the files are merged the documents need to be cleaned and irrelevant columns removed. When this is complete we will finally be able to do some general exploratory analysis to get a high level understand the kickstarter data. We plan on using R and methods learned during class throughout this process (i.e. tidyverse and stringr packages).

After the exploratory analysis, we are interested in further investigating what features are most predictive of whether or not a project is funded. Logistic regression will be the most appropriate method to try and predict whether or not a project is funded, and we can tweak certain variables to see how it affects the accuracy. We plan on measuring accuracy as the percentage of funded projects it predicts correctly from a test data set, and additionally the percentage of false positives/negatives.

Lastly, we will integrate the data into a shiny app. This will be an interactive way to present the project, so that we can walk others through our data exploration during our presentation. Additionally, the shiny app itself will be the product of our project, demonstrating proficiency in producing an intelligence tool (similar to a Tableau dashboard).

**Preliminary results**:

While our preliminary research has not been particularly in depth, it has yielded some insight into the dataset. Below are some of the early findings:

* The number of projects are most frequently handled in USD currency by far, which makes sense considering kickstarter began in the United States.
* Projects fail (do not meet funding goal) more often than they succeed (meet funding goal).
* “Music” and “Film & Video” have the highest number of projects over the history of our dataset.
* “Dance” and “Theater” are the categories with the highest proportion of successful projects.

**References**:

Kickstarter Website: <https://www.kickstarter.com/>

Data Sources: <https://webrobots.io/kickstarter-datasets/>